# Maximising independent model validation effectiveness for Aegon

**PUBLIC VERSION**

CONTAINS BLURRED/SHADED FIGURES

August 23, 2019


UNIVERSITY OF TWENTE.

**Author**
Gijsbert van den Engh

**Supervisors**
dr. Berend Roorda (UT)
dr. Reinoud Joosten (UT)
Bart Rikkert, Msc (AEGON)

**University of Twente**
Drienerlolaan 5
7522 NB
Enschede

**AEGON N.V.**
Aegonplein 50
2591 TV
Den Haag

# Executive summary

Model risk management is experiencing an increase in focus from financial services regulators. Setting up an independent model validation process enables firms to manage model risk (when applying sufficient follow-up measures). Aegon, an insurance company with its headquarters in The Hague, has developed a model risk management framework and is looking for ways to increase the model validation time effectiveness and/or decrease the model validation time in a proper way for models holding a 'medium model risk'.

We start with analysing comparisons/differences between the Aegon model risk management framework versus regulatory requirements and industry's best practices. The outcome is a list of similarities and differences, together with an identification of alterations to the Aegon model risk management framework possibly reducing the model validation time. In order to highlight the most relevant issues, we collected and analysed the data of historical model validation outcomes including the model characteristics. This resulted in five possible methods that can reduce the model validation time or increase the validation time effectiveness.

The first method, decreasing the model validation scope, reduces validation time – but can increase the model risk exposure of an individual model. However, time reductions can be used to tackle the backlog of models due for validation, resulting in a possible decrease of model risk exposure enterprise wide. We estimate that this results almost 5.5% more model validations per year (medium and high risk models).

We identify the second method, altering the model risk classification, as a short and long term solution. Our data analysis shows that the current risk classification can be improved, especially the variable that indicates the probability of a model deficiency ('complexity'). We suggest to redesign the Aegon model risk classification in a more data-dependent manner using supervised learning techniques. When implementing this, we found that the current available model characteristics are insufficient for reaching short term implementation. However, we listed possibilities and steps to take for long term implementation.

For this, we first listed several model characteristics that are tracked in the model inventory of Aegon the following years. We also found variables proposed in research papers. Especially continuous variables are needed for improving the prediction of model risk. Using these variables for estimating the likelihood of a model error can result in improved model validation prioritisation, and provides insight in risky model variables. We propose that future model validations use the latter outcome for model complexity reductions (Method 4). Overall, we believe this will improve the model validation effectiveness of all models, reducing the model risk exposure of Aegon.

Thirdly, we found that current model validation reporting includes more topics than the guidelines of the De Nederlandsche Bank suggest. We therefore propose that Aegon summarizes one of their validation reporting topics to reduce reporting time for medium risk models. This leads to a reduction in reporting time.

Our fourth method, decreasing model complexity, is expected to reduce model validation time in the long run. But first, model validation should include an examination if a model's complexity is reasonable or if a model can/should be simplified. Method 5, performing validations concurrent to the model development process, has been used in the past by Aegon, and was experienced as inefficient. We therefore do not recommended implementing this method.

Finally, we analysed possible process improvements that do not involve changes to the Aegon model validation policy. We suggest that a standardised model validation process workflow accompanied with standardised emails and checklists are created. We furthermore recommend using the available unique model ID's, as well as regular updates on the planning sheets (also retrospectively). Maintaining a high level of data tracking can be achieved by creating a culture of feedback regarding these efforts.

# Preface

I am very proud for presenting you this thesis report, and am thankful for all the support I received from my supervisors and the Aegon model validation team. In the beginning of 2019, I was still living in Augsburg busy with preparations for project deadlines and exams. I successfully reached the second round of interviews at Aegon and was invited to visit the headquarter in The Hague. I still remember travelling more than eight hours by train the day before the interview, and the same amount the following day right after the interview. Even though I had to invest a lot of time during a stressful period, I believe it was more than worth it.

Before my first day of work, I expected to work in an environment of work-pressure, direct feedback, and a top-down hierarchy – mostly due to prejudices obtained from books and movies. I experienced the opposite: I encountered a lot of friendly people at Aegon, the work environment is very healthy, and goals are realistic and achievable. I am very thankful for the way that Bart Rikkert has supervised me at Aegon. Bart, thank you for your patience, having trust in my capabilities, and also for the conversations about topics not involving my thesis. Also thanks to Amrit, Ian, Justin, Ling, Michael, Pier, Rebecca, Sunny, and Wim for always being available to answer questions, and discussing work and/or non-work related matters.

Before I could start with this master thesis project, the University of Twente provided the basis with the newest insights in Industrial Engineering & Management. I am very grateful for broad possibilities the Dutch university system provides. I experienced that everything is possible, as long as you put time and effort into it. Dr. Roorda, thank you for your guidance throughout this project and for teaching me during all the interesting lectures over the last five years. Your interest in decision making processes, valuation issues, and insights to the financial sector's characteristics have inspired me greatly. Dr. Joosten, thank you for your support and lectures. Your down to earth view and mentality has inspired me, especially the course microeconomics has made me a more pro-active student.

Friends and family have supported me during my current and previous study. Mom and dad, thank you for being patient and never doubting my decisions. Especially when I told you after three years of studying applied physics, that I was going to switch to IEM. Mascha, thank you for your support over the last years and for all the great times we experienced – and I am looking forward to all the moments still to come. Willem, Cornelis, Roderick, friends from Oldenzaal, my rowing squad, and 'Magnaten' – thanks for all the fun and unforgettable times!

Gijsbert van den Engh

*The Hague, August 2019*

# Contents

# Chapter 1
# Introduction

With this report we present the practical difficulties of managing model risk, which can be viewed as less prominent or common as other types of risk such as credit and market risk. Model validation is, when applying sufficient follow-up measures, a way to get insight into and allow conscious control over a firm's model risk exposure. While maintaining a proper level of model risk exposure, the model validation team of Aegon is looking for ways to maximize model validation effectiveness. This paper elaborates on model risk and steers towards a proposed strategy to improve the model validation effectiveness at Aegon.

We present the research topic, problem holder, and research goals. This chapter ends with the introduction of a leading example used throughout the report for illustrating practical descriptions and implications of topics mentioned.

## 1.1 Model risk management

Before elaborating on model risk management, it is important to consider how a model is defined. We take the definition presented by The Board of Governors of the Federal Reserve System (2011):

> "the term model refers to a quantitative method, system, or approach that applies statistical, economic, financial, or mathematical theories, techniques, and assumptions to process input data into quantitative estimates. A model consists of three components: an information input component, which delivers assumptions and data to the model; a processing component, which transforms inputs into estimates; and a reporting component, which translates the estimates into useful business information" (p. 3).

The board continues by stating that using models invariably holds model risk, formulating this as "the potential for adverse consequences from decisions based on incorrect or misused model outputs and reports." (p. 3). Since the use of models in the financial services industry is heavily increasing due to digitisation, regulatory measures, and decision making innovations - model risk management (in short: MRM) is becoming an important topic for the risk governance of a financial services company.

Effective MRM can result in avoiding losses, but can also result in capital improvements or enlarge profits by achieving cost reductions (Crespo et al., 2017). The occurrence of losses due to model risk does not have to be the fault of model developers per se, but can be caused by poor governance - for instance if changes in the market environment have not been detected in time (Whittingham, 2018).

One possible role within the MRM framework is the independent[1] validation of models (in this research paper addressed as 'model validation'). MRM requires that roles and responsibilities concerning these validations are documented and satisfied, as well as the validation process is being substantiated. More information on model risk, MRM and the role of model validation, as well as an in depth overview of studies and best practices in the field of MRM and model validation, will be presented in Chapter 2.

## 1.2 Research organisation

Aegon is an international life insurance, pensions, bank, and asset management group. AEGON N.V. is the publicly listed holding company with the head office stationed in The Hague (AEGON, 2018). The head of risk governance within Aegon corporate centre reports to the CRO. The risk governance group is divided into several divisions, one of those is 'Operational & Model Risk'. The model validation team is a separate team within this division.

The aim of model validation is to provide assurance on the integrity of models used across Aegon. Model integrity implies that models are fit for purpose and produce reliable results that management can use for decision making. It further implies that the model is based on approved methodologies, has been well developed and tested, and that the model is monitored correctly. The validation team is involved in the group-wide model validation policy, model change policy, and the centralised principles underlying expert judgement, and assumption logging.[2]

## 1.3 Research goal

The model validation team of the corporate centre, and the model validation teams of strategic business units (in short: SBUs) (or country units) are facing capacity challenges. The current goals for model validation include a mandatory validation frequency for risk models classified as having medium risk, being at risk given the high number of medium risk models and the time consuming practice of validation. On top of that, the Aegon model validation team has limited insights in historical validation performances and time allocations.

[1] Independent means that the validation is performed by a separate team within a company.
[2] This passage is based on the information presented in the internal report of Rikkert (2019).

The model validation team is looking for process optimisations or new methodologies to make "medium risk" model validations more efficient to ensure all medium risk models can be validated timely. It is however important that the quality of these validations remains sufficiently high to assure key issues are still found such that country units can bridge gaps and retain confidence in models and keep improving their models.

In order to optimise the medium risk model validations, research is necessary in order to clearly define the current validation process, an ideal validation scenario, and to detect possible gaps between the two. Gathering and analysing insights concerning the historical validation performances and effectiveness in time allocation will help fulfil this goal. The corresponding main and sub research questions are defined in the next two sub-sections.

### §1.3.1 Main research question

| How can the time effectiveness of the model validation process, specifically when focusing on the medium risk models, be improved while maintaining a desired level of quality? |
|---|

### §1.3.2 Sub-questions

1. Current situation
   - 1.1 Do regulators give guidance for model validation in MRM frameworks and proposed model validation policies, and what information and/or recommendations are available from research/financial institutions/consulting firms?
   - 1.2 What current policies are in effect for defining the criteria and requirements for model validation at Aegon, how does this correspond to or deviate from the information found in Sub-question 1.1, and which alterations to the Aegon policies can decrease the time necessary for validations?
   - 1.3 What factors influence the model validation outcome, on what factors is the time spent dependent, and what are the current performance measures?

2. Towards a (set of) proposal(s)
   - 2.1 What methodologies could possibly result in decreasing the model validation time of medium risk models and how will these affect the quality of model validations?
   - 2.2 How can these possible alterations to the model validation process be implemented, and what added/replacing tool(s) can be used to support these measures?
   - 2.3 What other recommendations can be found outside the Aegon model validation policy that could benefit the time necessary for model validations?

## 1.4 Deliverables

The final goal is to propose a set of suggestions involving at least one alteration to the model validation policy regarding medium risk models. On top of that, we state how this alteration affects the model validation quality. A description of a tool to further improve the model validation process is also provided. Finally, further recommendations beyond the scope of the Aegon model validation policies complete the deliverables.

## 1.5 Leading example

Since several topics described in this paper are of an abstract nature, we use a leading example throughout several chapters so that topics, suggestions, findings, etc. are assisted by practical illustrations. We now introduce the example itself, and the layout in which it is presented.

**Leading example: Endowment insurance (Part I)**

An endowment insurance combines the benefits of life insurance and life annuity. The *policyholder* pays a premium during the term $n$, and the policyholder's age is denoted by $x$. With this contract, the holder receives a death benefit ($DB$) in case of death during the term and an endowment payment ($EP$) at the end of the term if the holder is still alive.

For the insurer it is important to calculate the *net premium*, which is required for providing the contract benefits. Next to this, the insurer has to determine the *gross premium*, which is charged in practice. Following the definition of a model, the calculations for the net premium have the following inputs, assumptions, calculation process, and outputs.

**Net premium:**

- Main inputs: Policy term, DB, EP, customer characteristics.
- Main assumptions: Discount factor, mortality rates, surrender rates (cancellations).
- Computations: Expected present value of future income and benefits.
- Output: Net premium used for determining the gross premium for individuals.

The modelling of net premiums and reserves can be considered as an usage of assumptions and calculation techniques on the given input data, resulting in quantitative estimations – and is thus by definition a model.

# Chapter 2
# Model risk management frameworks

In this chapter, we provide information on model risk, model risk mitigation, MRM insights, and validations in order to find an answer to Sub-question 1.1. We start with elaborating more thoroughly on model risk in the financial services industry (Section 2.1), and on model risk mitigation possibilities (Section 2.2). A firm's freedom in adapting model risk mitigation frameworks however is bounded by legislation, therefore Section 2.3 focuses on MRM guidance proposed by regulators. With Section 2.4, we provide insights from advisory organisations and researchers. Section 2.5 ends this chapter with a summary of findings, and directly answers Sub-question 1.1.

## 2.1 On model risk

A straightforward method does not exist for managing model risk. Just as other types of risks present in the financial services industry, uncertainty cannot be mitigated completely. However, realising how and why decisions and assumptions have been made, can help firms in becoming aware of the risks they are facing. For most firms operating in the financial services industry, model risk is a relatively new concept in risk management. It gains more attention due to digitisation, and especially the increasing reliance on models.

Just like operational risk, it is hard to determine the right capital reserve for having a buffer in case of loss events due to the incorrect or misused model outputs. When managing a simple example of credit risk with a large pool of customers that have, in an estimated worst case scenario, a 30 percent probability of going into default (with a complete write-off) - it is logical for a firm to reserve at least 30 percent of the outgoing credit as a capital reserve. With model risk it is harder to determine what the probability of an incorrect or misused model output is, and determining its potential impact and required capital to hold for the effect is even more difficult.

For instance, in case of a loss event, it does not have to be apparent that the event occurred because of an incorrect or misused model output. Rather, the possibility exists that several types of uncertainties leading to the event were present, resulting in labelling the event as a loss due to only one or several other types of risk. The endowment insurance example illustrates the difficulties in determining the risk origin of a loss event (leading example Part II).

**Leading example: Endowment insurance (Part II)**

For determining the *gross premium*, the insurer has to calculate the *net premium*. As introduced in Part I of the leading example, the model has the following main inputs: Policy term, death benefits (DB), endowment payment (EP), customer characteristics. The main assumptions have to be formed for: Discount factor, mortality rates, surrender rates.

When calculating the net premium for each individual policyholder, the model user is able to evaluate the policies' risks - resulting in an advise to accept or decline the risk. Now let's assume that insurance benefits have exceeded the premium incomes five years after the policies have been sold.

In those five years, the insurer has developed new premium calculation and risk evaluation techniques. It is likely that the insurance company labels the loss event as occurred due to underwriting risk (i.e. an inaccurate assessment of the policyholder's risk profile and/or benefit probabilities) - especially when knowledge of the relatively old model is not present in the company anymore, or perhaps when the right documentation and program files are missing. This illustrates a scenario where thorough analysis of the risk origin is very time and cost intensive.

Possible causes of this loss because of model risk exposure are the following.

- In coding, the model developer used a '1' for males and a '2' for females, the model user switched these when calculating net premiums.
- The model developer made a slight computational mistake when discounting future value which led to lower estimated net premiums.
- Insufficient monitoring of model forecasts versus new available data.

## 2.2 Model risk mitigation

There are several ways in mitigating model risk, of which model validation is a possibility (when assumed that deficiencies found are resolved after the validation). Before listing several methods for mitigating model risk, it is worth rereading the exact definition of a model presented in Section 1.1.

As the definition states, a model uses theories, techniques, assumptions, and input data for creating a quantitative estimate. Model risk thus includes any uncertainty in the appositeness of these theories, techniques, assumptions, and input data. Mitigating this risk starts with becoming aware of the presence of each of these elements when developing a model, and is followed by realising that any element may be faulty, misused, or incomplete. This results in knowing what the model boundaries (i.e. limitations) are, and what key assumptions have been made.

The previous passage elaborates on the possibility of having an incorrect model output. The definition of model risk, however, also covers misused model outputs and model reports (Board of Governors of the Federal Reserve System, 2011). Therefore, mitigating model risk also requires focus on what the model outputs are used for and whether the model outputs fulfil the initial necessity for a model. Additionally, a model report insufficient in addressing the relationship between the model- inputs and outputs, including the known limitations and assumptions, can amplify model risk.

Since the definition of a model states that the model output is an estimate, total mitigation of model risk is by definition impossible since a model is ultimately incorrect at some level (Derman, 1996)[3]. Model validation as a means to mitigate model risk should thus have the aim to prove that a model is incorrect (since full validity of correctness is unobtainable) (Popper 1959). When the model validation results in no severe inconsistencies, the final goal of increasing confidence in the model is fulfilled (Robinson, 1997).

In practice, when resources are limited, an optimal resource allocation for increasing confidence in models is not evident. Also, it can be hard for firms to succeed in detecting models are likely to be faulty, and which models deal with relatively high materiality. In coping with this validation quandary, firms can manage model risk by determining their model risk appetite and create awareness of model risk amongst model developers, model output users, and controllers.

A direct mitigation technique can be found when one thinks of the following classification for model risk: Likeliness to be faulty versus materiality. A method in mitigating model risk can simply be lowering the likeliness to be faulty, namely by decreasing the complexity of models. This view is supported in the report of Haldane and Madouros (2012), which focuses on regulators' efforts. The authors argued that complex regulatory measures do not by definition lead to more reliable predictions. Rather, complex models yield high costs for information gathering and processing and contain the danger of being 'over-fitted'.

In the report of the CRO Forum (2017) the concept of model risk is split up into two parts. The first part of model risk is structural risk and is mainly present due to the complexity of models, which can be mitigated by performing the previously mentioned model validations (by the model owner and/or by independent validations). The second part, namely operational risk, can be mitigated with "appropriate process controls and adequate communication of model results" (CRO Forum, 2017, p. 7). Before a model risk appetite can be determined, firms must be aware of the legal boundaries or guidelines that regulators have defined.

[3] Here it must be noted that the possibility of fully mitigating model risk is achievable when one excludes the use of models, although one can hold the view that the exclusion of a model and relying on expert judgement is also a type of model risk.

## 2.3 Model risk management from a regulatory point of view

We now present the views of regulators, published in openly accessible documents. We select these outlooks since they prominently cover model risk. Still, reports/guidelines exist that are not mentioned in this research paper. The reports covered in this section can be seen as the starting point in managing model risk for firms, since the main incentive for managing model risk has been regulatory pressure (Chartis Research Ltd, 2014).

### §2.3.1 The United States regulators

As introduced in Section 1.1, model risk management is emphasised by the Federal Reserve System (in short: Fed) in a guidance report originating from 2011. However, model validation has been on the radar of governmental institutes for a longer period of time. For instance the Office of the Comptroller of the Currency (in short: OCC) published a guidance report for validating and testing models including the office's views on sound model validation processes (Office of the Comptroller of the Currency, 2000). The OCC stated at the time that it is expected of companies to set up formal policies for model validation in order to achieve an independent validation process from model constructors.

The Board of Governors of the Federal Reserve System (2011) expands this regulator's view on model risk and asks banks to have policies in place that correspond with requirements towards the broader aspect 'model risk management'. Next to model validation, the Fed gives guidelines for model development, implementation, and use, as well as governance, policies, and controls. On model validation, the Fed addresses the importance of the independence and skilfulness of validators. It further provides three core elements, which an effective validation framework should include (The Board of Governors of the Federal Reserve System, 2011, p. 11):

- "Evaluation of conceptual soundness, including developmental evidence".
- "Ongoing monitoring, including process verification and benchmarking".
- "Outcomes analysis, including back-testing".

On the validation frequency, the Fed proposes that banks should, at least annually, perform a periodic review on a model to check whether a model is working as intended and if follow-up validation activities are necessary.

### §2.3.2 The EU regulators

The regulators of the European Union placed their focus on model risk in 2013. The directive does not go beyond stating that institutions should "implement policies and processes to evaluate and manage the exposure to operational risk, including model risk…" (European Parliament, 2013, article 85).

Considering insurance and reinsurance regulation, the Solvency II regulation of the European Parliament (2009), which came into force in 2016, mandates companies to have regular cycles of model validation for "monitoring the performance of the internal model, reviewing the ongoing appropriateness of its specification, and testing its results against experience." (article 124). It adds that effective statistical processes should be in place for validating internal models used for capital requirements and that an analysis of the stability of internal models must be included.

### §2.3.3 De Nederlandsche Bank

De Nederlandsche Bank (in short: DNB) is the Dutch central bank as well as the main supervisor of Aegon[4]. Preliminary to the Solvency II regulation going into force, De Nederlandsche Bank (2013) published optional guidance and defines what the authority sees as good practices for model validation.

Just as the Fed stated, DNB suggests a company to set up an independent model validation process, and adds a proposal for the model validation organisation structure as well as what requirements can lead to independent model testing. In the third chapter of the guidance, DNB states that outsourcing can be used in order to validate topics as "governance, data accuracy and completeness, or IT" (De Nederlandsche Bank, 2013, p. 5). DNB continues with listing a proposition of the minimum requirements a model validation report should require:

- A management summary stating the overall opinion, key messages, and important limitations.
- The scope of the validation.
- What activities were performed and how were these performed.
- Limitations to the validation activities.
- Additional activities required, and why these were necessary.
- A full set of findings and scoring.

## 2.4 Model risk management research

After regulators have increased their focus on MRM, risk management services and consultants performed empirical research. Mainly on the priority that financial institutions have, and the progression made on effective model risk management. Next to that, model risk practitioners and researchers give their views on model risk and propose best practices or examples for managing model risk. This section summarises what advisory firms and researchers say about a firm's model risk appetite, and what steps should be taken in order to mitigate model risk.

[4] Since AEGON N.V. is operating in several countries, national regulators cooperate in supervising the company. In this cooperation, De Nederlandsche Bank is the lead regulator of Aegon.

### §2.4.1 Empirical research

Chartis Research Ltd (2014) states that, at the time of the study in 2014, few (12%) organisations classify their MRM program as comprehensive. And, just as mentioned in Section 2.3, firms see regulatory pressure as the main driver for MRM. Given those responses, an area for improvement has been identified by the authors - and they advise enterprises to include model risk into their risk management framework strategically instead of responsive. Chartis Research adds that awareness of model risk and an effective MRM framework goes beyond a reserve of capital, but should improve the enterprise risk management as a whole. In other words, firms should go beyond the vague and broad boundaries regulators are placing and define a risk appetite.

Whittingham (2018) also addresses the current responsive nature of model risk management and states that this is caused by the focus of Solvency II requirements on internal model validation of models used for regulatory measures, excluding focus on managing other internal models with comparable or higher materiality. The author advises firms to classify models according to their model risk, and adds that firms should decide per model class or type what controls should be used in order to mitigate model risk. He explicitly states that using independent model validation does not have to be a prerequisite in order to mitigate model risk when taking into account the model type or materiality (see page 27 of Whittingham (2018)).

From the results of the survey and interviews with the respondents, Chartis Research Ltd (2014) proposes a mapping of a MRM framework which can be assessed in Appendix A.1. An interesting point of view is the model inventory filtering in which Chartis proposes organisations to determine whether models should still to be used in future decision making.

Whittingham (2018) puts more emphasis on the relation between the model developer, model inventory, and model validation. The proposed MRM cycle of is presented in Appendix A.2. He states that firms are able to define their risk appetite by identifying and inventorying models, determining their model risk class, model type, and stating which controls should be used (of which one is an independent validation). Thereby, a distinction is made between initial- and subsequently follow-up validations. The CRO Forum (2017) believes that validation, monitoring, and reporting efforts should be proportional to the materiality of the model.

Next to the papers mentioned and evaluated, are several research reports worth mentioning. The first is the paper by HM Treasury (2013), which addresses several social and cultural aspects of model development. Another report worth mentioning is the one of Lloyd's (2017), which gives guidance on the model validation of internal models for solvency capital reserves.

### §2.4.2 UK actuaries on model risk – Actuaries' Model Risk Working Party

Aggarwal et al. (2016) ("phase I") elaborates broadly on the philosophy of model risk within an organisation, and proposes a MRM framework for organisations such that model risk can be managed enterprise-wide. The authors propose a policy in which a relatively high responsibility is given to the model owner before delivering their models for validation: "The primary responsibility of the model owner is to ensure that the model complies with the requirements of the Model Risk Management Framework" (p. 242). Also, it is notable that the authors advise to have the scope and intensity of the reporting proportional to the materiality and use of a model. The MRM framework can be found in Appendix A.3, including a description of the blocks and proposed model inventory entries.

Black et al. (2018) wrote a follow-up paper ("phase II") in which more focus is given to specific areas. The authors suggest that models with a so called 'high control level' should be validated every three years minimally, although no argument for this certain frequency is given. Next to that, the validations should be evidenced and inventoried in a central model inventory, of which the key data points are summarised in §4.4.3 of the report (Black et al., 2018, p. 18).

The methodology of the 'control level' that Black et al. (2018) uses is described by the model risk assessment in Section 4.6 of the paper (p. 27), where it is stated that the MRM effort should be in contrast to the model risk. The assessment of model risk is referred to as 'triage' and has a partly means to "reduce the amount of effort for the less material models" (Black et al., 2018, p. 21). A process leading to a high triage accuracy is the use of data gathering for approximating the likelihood of a model error. Black et al. (2018) propose to use meta data for model risk classification, for instance by using machine learning techniques. An example for such a methodology is given in Appendix A.4.

### §2.4.3 Validation research of simulation models

Sargent (2013) researched the verification and validation of simulation models. Although the research focuses on one type of modelling technique, namely simulation models, methods proposed for simulation model validation can give beneficial insights for the model validation in the financial services industry. Also in the area of simulation model validation, time allocation versus validation effectiveness is a concern. This becomes clear when the author addresses that a thorough validation of a model can be too costly and time-consuming for firms.

Sargent (2013) advises a manner of validation where the validator is independent, especially when the associated problem has a high cost and/or risk, and when public acceptance plays a role[5]. Other possibilities are validations that are performed by the model developer or model user. Furthermore, Sargent (2013) names two common approaches how validations can be conducted; either parallel to the model development process, or after model development.

[5] A recent development indicating the importance of this statement, is the Boeing 737 MAX validation commotion. See Gates (2019).

Sargent (2013) mentions that a positive aspect of the concurrent way is that the model developer can provide model verification results to the validator and can further develop the model by implementing the feedback received. Also, the model developer is focusing on the model continuously, without a time gap between the moment of model development and model validation. Validation after the model development process is not preferred by both Sargent (2013) and Wood (1986), of which the latter concludes that such a validation is costly and time consuming. The author also gives an overview of simulation model validation techniques. We list the relevant validation in Appendix A.5.

## 2.5 Main findings (answer Sub-question 1.1)

This section gives an overview of Sections 2.1 – 2.4. It furthermore gives direct answer to Sub-question 1.1: *Do regulators give guidance for model validation in MRM frameworks and proposed model validation policies, and what information and/or recommendations are available from research/financial institutions/consulting firms?*

Model risk can be divided into structural model risk and operational model risk (CRO Forum, 2017). Structural model risk can be partly mitigated by performing model validations, which should have the aim to prove the incorrectness of a model (Robinson, 1997). The operational risk factor can be comprehended by implementing process controls and correct model output use (CRO Forum, 2017). A basic way of mitigating structural model risk is lowering the complexity of models, which does not per se lead to lower predictive performances for regulatory means (Haldane and Madouros, 2012).

From the publicly reported guidelines of bank- and insurance regulators, it seems that main focus is laid on models used for capital requirements. Therefore, they require firms to have an independent model validation policy in place and ask firms to regularly monitor outputs and changes made to models. The topics that a validation report should require, have been listed by DNB and can be found in Section 2.3.3. Besides, regulators do not give concrete or specific guidance in publicly available documents for managing model risk - especially concerning models used for other means than calculating capital requirements. It can therefore be concluded that regulation is principle based.

Research/advisory reports state that mitigation efforts for model risk should not be responsive, but should be implemented strategically. Creating a model inventory will give firms a holistic view of possible model risk impacts. Research reports propose several MRM frameworks: Chartis Research Ltd (2014) suggests an MRM framework that uses a model inventory to allocate validation effort to each model after the inventory is filtered. Aggarwal et al. (2016) expand upon this method and address the responsibility of model developers. Black et al. (2018) and Whittingham (2018) advise to allocate model validation efforts proportional to a model's risk. Black et al. (2018) elaborate further on a model's risk, and highlight the role of data gathering for a continuous classification.

On model validation, Sargent (2013) lists several possibilities and model validation methods. A firm can choose to validate certain models during the development stage, or after the model is used. The study also lists several model validation techniques, where it is addressed that effort and scope for these model validations should be in line with the model risk of a model. Also, firms should create a risk appetite, such that the appropriate controls are used to mitigate model risk in line with the materiality of the risk. Such a methodology is proposed by Whittingham (2018), Black et al. (2018), and HM Treasury (2013). Furthermore, the effort and extent of the validation report should be in line with the materiality or model risk of a model (Aggarwal et al., 2016), although standardised reports hold advantages (Chartis Research Ltd, 2014).

# Chapter 3
# The model validation process at Aegon

This chapter provides insight into the model validation- policies, standards, and governance at Aegon - which are documented for internal use. Also, further information on the broader aspect of MRM strategies of Aegon is provided. Throughout the first three sections of this chapter, possibilities for decreasing the model validation time are highlighted such that a link with Sub-question 1.2 is made. After that, Section 3.5 answers Sub-question 1.2 by listing these methods and comparing results of this chapter with the answer to Sub-question 1.1 (see Section 2.5).

## 3.1 Model development standards and model change policy[6]

The process of developing a model starts with the need for one, after which a model concept is formed. The report of Van Roon and Van de Kraats (2015), which represents the Aegon model development standards, assists model owners in ensuring model integrity. The report sketches a list of steps that need to be fulfilled before the actual development of a model, listed below.

1. Initiation: Risk assessment of boundaries, data availability, and required assumptions.
2. Development of model concept: Definition of the purpose, scope, boundaries, and feasibility.
3. Planning: Schedule of time planning and resources necessary.
4. Requirements analysis: Determination of business and regulatory requirements.
5. Model design: Transformation of requirements into design including an analysis of possible use of other models or external data.

These 5 steps complete the initiation of the model development, the following act is model development in line with the model design (Step 6). Thereafter, baseline tests need to be performed by the model developer/owner before the model is implemented in practice (Step 7). The following tests can be used before implementation.

[6] This section is based on the information presented in the internal reports of Van Roon and Van de Kraats (2015) and Jadnanansing (2018a).

- Model development tests: The main goal is to compare model design with model development.
- Data testing: For judgement of integrity of the internal and external data used for the model.
- Sensitivity and stress tests: Assessment of changes in assumptions and adverse events.
- Output tests: Testing the stability of the outputs.

Together with the test plan, the model owner/developer has to document assumptions and expert judgements that are used for the model development. Expert judgements should be made by persons with relevant knowledge experience and understanding of risks within the field of insurance or reinsurance, according to the Solvency II regulation[7]. When the tests give a positive result, and the documentation is according to the standards - the model is implemented. Phases 1, 2, and 3 of the model life cycle, as presented in the figure below, have then been completed.

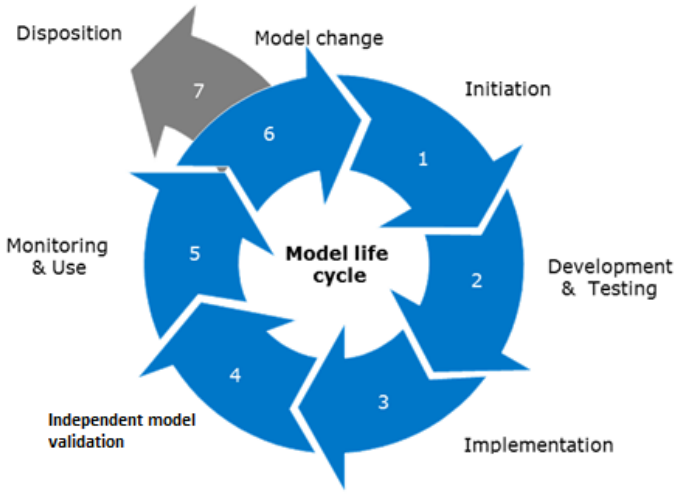



Figure 1: Model life cycle.[8]

In this research paper, we focus on Phase 4 of Figure 1. Independent model validation is further addressed in Section 3.3. Note that the phases in Figure 1 do not have to be completed consequently, but that processes can overlap. Also, models developed in the past have not been subject to current Aegon policies, and therefore are validated in a later phase. Phase 5 of the model life cycle (Figure 1) covers model testing in the operational environment performed by the model owner, mainly by comparing projections with actual data. Also back-tests are performed to empirically validate assumptions and expert judgements made during the model development phase.

[7] For more information on the use, validation, and regulation of expert judgements see the paper of Ashcroft et al. (2016).
[8] As presented in Jadnanansing (2019). Note that phase 4 has been renamed from 'Review' to 'Independent model validation' such that the terminology of this research paper is used.

The model life cycle restarts when a model change is adapted (Phase 6). Aegon internally defines model changes as "any major or minor change to models that are in scope of the Model Validation Policy and have been approved for production purposes" (Jadnanansing, 2018, p. 7), and classifies these as either minor or major. Model change triggers are for example regulatory changes, new methodologies, model gaps (identified by monitoring or model validation). The Aegon model change policy describes the reporting and documentation efforts that need to be fulfilled by the model owner. Model validation efforts have to be performed if the change is made to a model previously validated.

## 3.2 Model validation policy and model review standards[9]

The independent model validation team ensures model integrity and therefore assesses if a model is fit for purpose. This is done by reviewing the following aspects, which will be referred to as the **model validation topics** in the remainder of this paper:

1. Methodology. (in short: Meth)
2. Model development & testing. (in short: Mdev)
3. Data.
4. Assumptions. (in short: Assu)
5. Preparing for and validating model runs. (in short: Prep)
6. Reporting and use of results. (in short: Report)

There is a difference in the focus of the validation if the model is validated a first time, or if the validation is a follow-up to a previous validation. If the validation is an initial validation, more focus will be given to the following model validation topics: Methodology, model development & testing, data, and assumptions. A follow-up validation focuses more on the preparations for and validation of model runs, and reporting and use of results. A practical illustration on what is assessed exactly per model validation topic is given using the endowment insurance leading example.

**1:** A possibility for decreasing the model validation time is decreasing the model validation scope (i.e. take out one or several model validation topics).

[9] This section is based on the information presented in the internal reports of Van de Kraats and Rikkert (2015) and Rikkert (2019).

**Leading example: Endowment insurance (Part III)**

When validating the endowment insurance model, we assume that the model is already in use and has not been validated before. Therefore the independent model validation is taking place concurrent to the 'monitoring & use' phase (Phase 5 in Figure 1). The validation efforts will now be addressed per model validation topic:

1. **Methodology:** The model validator will examine the methodology used for the model development. These will be compared with internal standards if available, or external (actuarial) standards if internal documents are not present. Also, a risk assessment which identifies all risks relevant for the model is performed, and professional judgement is given on the actuarial and statistical techniques.

2. **Model development & testing:** The topic validation is mainly focused on the actual model development & testing of the endowment model versus the model concept and design – and if documentation is complete. Also, the model (pre-implementation) testing is assessed: Does the testing cover the full model and have relevant tools been used.

3. **Data:** Validation efforts focus mainly on the use of internal and external data. For example, for the net premium calculation, publicly available mortality tables are likely to be used. The validator determines whether these data are accurate, complete, relevant, and used appropriately. Furthermore, the validator will check if the data is documented well (e.g. source, year, version).

4. **Assumptions:** The validator investigates if the determination and documentation of the assumptions (see Part I) are in line with the corporate's- or local policy.

5. **Preparing for and validating model runs:** Efforts are focused mainly on the documentation and completeness of tests performed in Phase 5 of the model life cycle (Figure 1). An example could be a simulation run of premiums received versus benefit payments. Validation efforts then include analyses of model runs, spot checking, and plausibility checks.

6. **Reporting and use of results:** The validator mainly checks whether the net premium outputs are realistic. Also, are the results used for the net premium calculations of the right target group?

As seen in the previous leading example, documentation is a recurring matter throughout the model validation topics. The Aegon model review standards explicitly states that it is likely that models developed in the more distant past contain less documentation. This holds risk for third party validations and internal cross usage. Also, it is likely that only a small group of employees carry in depth knowledge of the model which is risky for the company. These risk indicators are used when proposals are presented in this research paper.

For determining which models have the highest validation priority, Aegon uses classifications for a model's risk that can either be *low*, *medium*, or *high*. The gradation of the model risk is dependent on two different factors: **Complexity** and **materiality**. Both metrics are also scored as *low*, *medium*, or *high*.

Model complexity is judged by the model developer/owner since it requires in depth knowledge of the model. In the Aegon policies there is no standardised quantitative or qualitative analysis given to determine the complexity of a model. Only examples of high complexity characteristics are given, such as: Low frequency & high severity claims, heterogeneous portfolios, low predictability, etc.

The model materiality is dependent on maximum thresholds set by the group management board for each model type. After the gradations for complexity and materiality are determined, the model risk classification is determined by using Table 1.

Table 1: Model risk classification.

| Complexity / Materiality | **High** | **Medium** | **Low** |
|---|---|---|---|
| **High** | High | High | Medium |
| **Medium** | High | Medium | Low |
| **Low** | Medium/Low[10] | Low | Low |

The **frequency** of the model validation is dependent on the model risk. High risk-, and medium risk models are validated every three and four years respectively. This validation frequency is increased to annual or biannual after a validation resulted in a 'unsatisfactory' or 'requires attention' opinion respectively (please see Section 3.1.3 for information on the validation opinion). The validation frequency is also increased if one or more major changes to the model have been made since the last validation. Low risk models are mainly being self-assessed by the model owner once per five years.

**2:** Another possibility for decreasing the model validation time is reconsidering the classification as presented in Table 1 and/or adjusting the corresponding validation frequencies.

[10] The model validation policy is currently being changed. The plan is to classify models having low materiality always as a low risk model.

**Leading example: Endowment insurance (Part IV)**

Basic net premium calculations for the endowment insurance can assumed to be classified by the model owner as *low* **complexity**. The model does not include stochastic variables and does not include embedded option characteristics or other complex details. An example for a *high* complex model would be a more broad pricing model which includes Solvency II capital requirements, insurance benefits linked to many policy holder characteristics (e.g. income, living circumstances), insurance benefits including other scenario's than only death (e.g. physical disability), and detailed strategic projections.

The **materiality** is dependent on the local materiality thresholds. For instance, the reserve calculations for the endowment insurance could represent 24% of the local economic capital (EC). When (fictitious) thresholds suggest that model materiality is *low* if the capital reserves represent a percentage between 0 and 10% of the EC - *medium* between 10% and 20% of the EC - and *high* when it represents 20% of the EC or more. The materiality class for the endowment insurance model is therefore *high*. The **model risk** is determined using Table 1. Since the complexity is *low* and the materiality is *high*, the model risk is classified as *medium*.

Since the model has a *medium* risk class, the goal is to validate the model every four years. This **frequency** is however increased to every two years if a previous validation has given an alarming result, and increased to annually if the previous validation has given an 'unsatisfactory' opinion. If a major change is made to the model, such as the use of an updated mortality table, the frequency is increased to annual.

## 3.3 Model validation reporting[11]

When reviewing the model validation topics, the validator may come across certain model issues. To ensure cross-model comparability, validation findings are reported in standardised manner. Aegon is using the term 'model validation gap' for an issue, if the issue is "related to the model integrity component of model risk" (Van de Kraats and Rikkert, 2015, p. 18). If not, the issue is not considered as a model validation gap and is not reported.

Gaps are either **financial- or control gaps**. If a gap embodies a direct, or severe indirect concern to the model results, the gap is considered a financial gap. If not, the gap is considered a control gap. The presence of control gaps increases the probability of incorrect model use and model errors. The severity of financial- and control gaps is either *low*, *medium*, or *high*, and are adjudged to one of the six model validation topics. This classification is linked to a certain percentage of the model materiality for financial gaps, whereas classifying control gaps is of a more arbitrary nature. In the (excel) reporting scorecard, a high control gap is said to be "extremely serious", a medium control gap "important to resolve", and a low control gap a "minor control issue" (Lloyd, 2019).

[11] Just as the previous, this section is based on the information presented in the internal reports of Van de Kraats and Rikkert (2015) and Rikkert (2019).

Next to financial- and control gaps, **simplifications** and **limitations** are also reported per model validation topic. A simplification is defined as "a deliberately chosen shortcut by the model owner to cope with data, systems and other constraints" and a limitation as a "noted matter that cannot be resolved by its nature" (Rikkert, 2019, p. 11). Both findings are also classified as either *low*, *medium*, or *high* relative to the model materiality threshold. Simplifications and limitations are logged by the model owner during model development. Limitations are relevant to the model accuracy and its applicability but will not influence the model opinion judgement during the model validation. Limitations can however result in gaps when the validator questions the completeness of the limitation log, or does not agree with certain limitations.

All four types of findings are reported in a standardised spreadsheet scorecard and standardised text file document. This report will refer to these document as **scorecard** and **MV report** respectively. The scorecard calculates a **topic opinion** per model validation topic. This outcome depends on the number of- and severity of gaps and simplifications found. The topic opinion can either be one of the following:

- Green (effective) – fit for use.
- Yellow (further improvements are recommended) – acceptable.
- Amber (requires attention) – can only be used with mitigating actions in play.
- Red (unsatisfactory) – only acceptable (short term under strict requirements) with special dispensation from management.

The topic opinion is created by summing up the total impact of financial gaps and simplifications. Just as a financial gap or simplification itself, each topic has financial threshold limits which result in an initial topic opinion. After this, control gaps possibly shift this opinion to a more severe level. This leads to a mechanic topic opinion. The worst performing topic opinion will be a first comparison for determining the total model opinion. The second comparison is created by summing all financial gaps and simplifications found in the entire model. The pre-determined materiality thresholds will give an initial result which can be shifted to a higher level when taking into account all control gaps found in the model.

The worst of both comparisons results in the **model opinion**, which uses the same colour classifications as the topic opinions (*green*, *yellow*, *amber*, or *red*). Of course, not every model validation is the same, and a standardised validation scoring method sometimes cannot reflect nuances. A model validator can therefore override the mechanical scoring results. When doing this, the validator has to justify why this override is performed. Please see Figure 2 for the process flow leading to the model opinion.

The MV report gives an overview of the scorecard outcomes with a short summary per model validation topic and an explanation for a score override if applicable. Also, more model background information is given, and the report includes a description of the validation work. Furthermore, a broad description is given per finding, including a justification of the categorisation and classification. This is accompanied by a response of the model owner and an overview of the outstanding gaps.

**3:** Simplifying the reporting template in general will decrease the model validation time. This can be applied to the scorecard, MV report, or both.

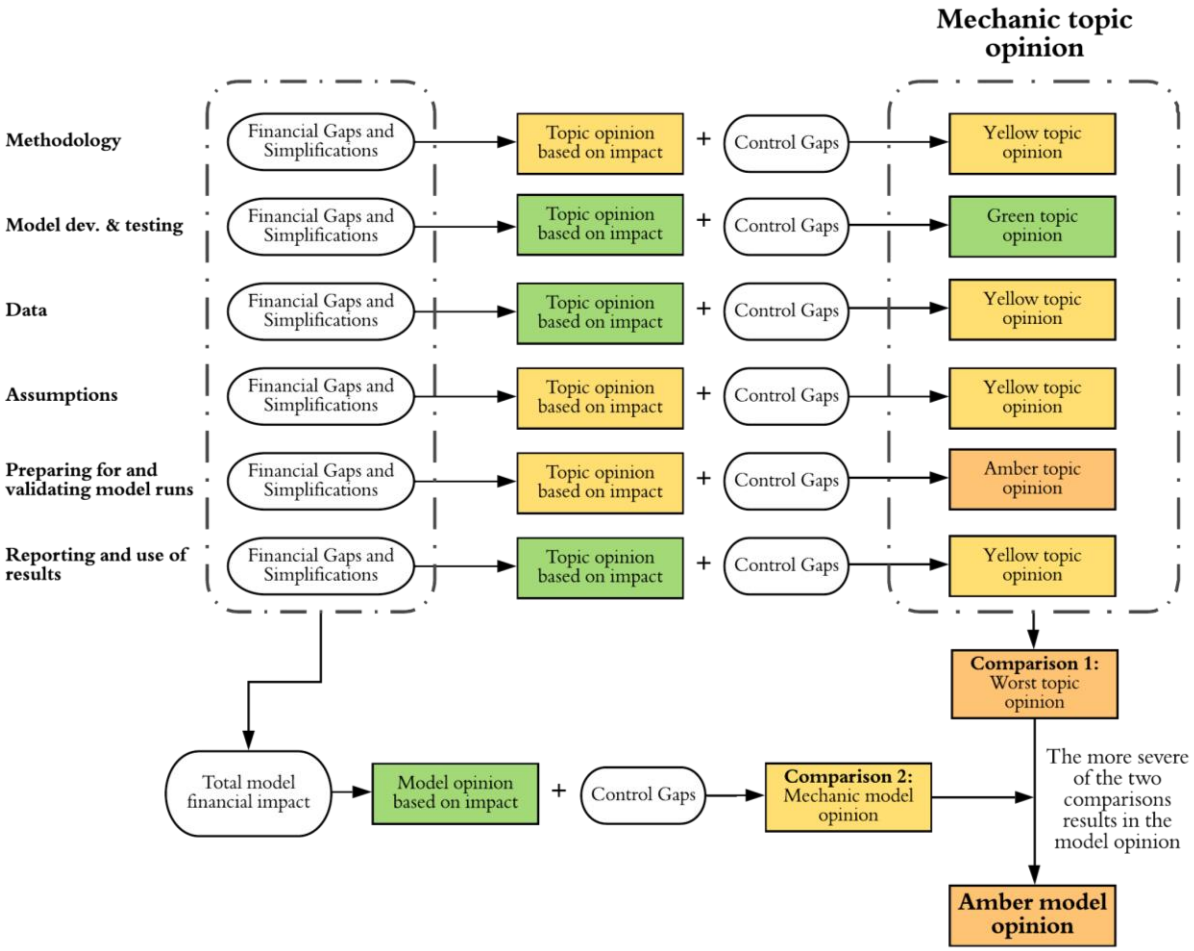



Figure 2: Model opinion process flow.[12]

[12] This process flow is slightly different for IFRS models. For these models, simplifications will not affect the topic opinions and model opinions - and are thus taken into account the same way as limitations.

**Leading example: Endowment insurance (Part V)**

The model validation policy states that materiality thresholds for classifying **financial gaps, simplifications, and limitations** found in insurance products are linked to the local economic capital (EC). The (fictitious) boundaries are as follows:

- *Low* if finding is below 1% of EC.
- *Medium* if finding represents 1% of EC or more, and less than 2% of EC.
- *High* if finding represents 2% of EC or more.

We take the following fictitious thresholds per mechanical **topic opinion** are:

- *Green* if finding is below 1% of EC.
- *Yellow* if finding represents 1% of EC or more, and less than 2% of EC.
- *Amber* if finding represents 2% of EC or more, and less than 3% of EC.
- *Red* if finding represents 3% of EC or more.

These boundaries are multiplied by a factor of 2 for finding the mechanical **model opinion**.

The validation resulting in four findings. The first is that only mortality tables dating from five years ago are available. By looking at more recent mortality tables of surrounding countries, the validator estimates that a worst scenario shift in probabilities has an effect of 5% in net premium calculations. Since the model represents 24% of EC, the total effect is between 1% and 2% of the total EC. The finding is therefore classified as a medium limitation under the 'data' model validation topic.

Secondly, the validators found that no discrimination between men and women is made for net premium calculations. This affects the outcome, since death probabilities are different. Examining the mortality table, the validators conclude that this has an effect of 5% in net premium calculations (and 1.2% of the unit's EC). And thus classifies this finding as a medium simplification under the methodology.

The validators also found that computational mistakes in discounting have been made, totaling up to 0.9% of the unit's EC. This relates to the methodology, and is classified as a low financial gap. Lastly, the validators detect that the methodology documents are lacking references and are outdated compared to today's actuarial standards. A high control gap is therefore added.

All in all the mechanical topic opinions are green for all topics but the methodology (since limitations are not taken into account). The financial gap and simplification for the methodology topic add up to 2.1% of the unit's EC, and therefore result in an amber topic opinion. This opinion is mechanically raised to red because of the high control gap.

The entire model total sum over all financial gaps and simplifications also represents 2.1% of the EC and therefore gives a yellow model opinion. The high control gap increases the opinion to amber. The final model opinion is the worst result of this opinion versus the worst of all topics: Amber versus red. The result is a red model opinion.

## 3.4 Aegon MRM

Model validation and model risk management are separated in the organisational structure of Aegon. MRM activities are described as ongoing efforts to ensure a proper level of control and oversight of model risk in-between model validations.

The basis of MRM is the model inventory, which gives an overview of all models that are in use. The model risk manager is responsible for the classification of the model complexity and materiality. This is done by constant communication efforts between the corporate centre and the country units (partly because the country units are responsible for assigning the model complexity). The model inventory tracks the following entries per model.

- Model name & ID.
- Model risk, complexity, and materiality.
- Calculation kernel.
- Brief description and model purpose.
- Model developer and model user.
- Last model opinion (if applicable).
- Year of last model validation and year of next model validation.

After a model validation has taken place, the model risk manager documents the outstanding model gaps and tracks whether these model gaps are being solved as agreed upon. The model risk manager is not authorised to validate whether the model gaps are closed sufficiently, this is only done by a model validator.

Current strategic efforts are mainly focused on data gathering and dashboard visualisation for the model inventory. Since the policies concerning model risk are relatively new at Aegon, not all models have had an initial validation. The model risk manager therefore has insight per country unit what amount of the present models have been validated, and what the model opinions are for the validated models. This is a necessary step for gaining insight in the model risk exposure.

When one realises that Aegon has offices in over more than 25 countries spread over the world, it is not hard to imagine that cultural differences and different perceptions of model robustness and model risk play a role in enlarging data gathering endeavours - especially when model inventories have to be updated frequently. The model inventory states that currently a total number of 715 models are in use at Aegon with a medium- or high model risk classification. The number of medium risk models is close to the number of high risk models enterprise wide, and this distribution is also present for most country units.

## 3.5 Main findings and comparison (answering Sub-question 1.2)

This section gives answer to Sub-question 1.2 - *What current policies are in effect for defining the criteria and requirements for model validation at Aegon, how does this correspond to or deviate from the information found in Sub-question 1.1, and which alterations to the Aegon policies can decrease the time necessary for validations?*

Alterations to the Aegon policies decreasing model validation time.

1. Decreasing the model validation scope.
2. Reconsidering the classification as presented in Table 1 and/or adjusting the corresponding validation frequencies.
3. Simplifying reporting templates of the scorecard, MV report, or both.

Correspondences between Aegon policies and Sub-question 1.1 findings.

4. Just as Aggarwal et al. (2016) proposes, the Aegon model development standards gives model developers responsibility in following guidance with the means to limit model risk.
5. The Aegon model validation framework meets the (broadly defined) requirements of the FED, Solvency II, and DNB. Regarding the DNB, Aegon reports the minimum requirements a model validation report should require as listed in Section 2.3.3.

Additions or differences found between Aegon policies and Sub-question 1.1 findings.

6. Aegon uses model validation processes and follow-up gap closures to mitigate the structural- and operational model risk aspects, as identified by the CRO Forum (2017). Mitigation efforts are however not directly focused on lowering complexity for over-complex models as Haldane & Madouros (2012) suggest.
7. Aegon is expanding its model inventory of which the current entries per model are listed in Section 3.4. Further possible additions from Aggarwal et al. (2016) are: Model use frequency, an overview of how the model works, key assumptions and/or inputs, model hierarchy and interdependencies (as presented in Appendix A.3 in this paper).
8. The exact model risk classification of Aegon, as shown in Table 1 of Section 3.2, is also found in the report of Whittingham (2018). The report of Black et al. (2018) proposes an assessment of model risk which is more reliant on data (Appendix A.4).
9. Whittingham (2018), Black et al. (2018), and HM Treasury (2013) suggest that appropriate controls are used to mitigate model risk in line with the model materiality (regarding the model inventory and validation efforts). Aegon does not apply this strategy.
10. Aegon's reporting efforts are standardised for all model validations. Aggarwal et al. (2016) however suggests to report in proportion to a model's materiality or model risk class.
11. Model validations at Aegon take place after model developed, it could be considered to have a validation concurrent to the model development as proposed by Sargent (2013).

These twelve main findings are addressed in the remainder of this research paper as 'Chapter 3 finding §x', with x representing the finding number.

# Chapter 4

# Aegon model validation data study

In order to focus on the right process alterations, it is important to investigate what the current performance measures of the model validation team of Aegon are and what dependencies exist. Therefore, we have gathered information of model validations performed in 2017 and 2018. In this chapter, we present our KPI's to determine the current model validation performance. In Section 4.1, we present the entries and variables we registered, and we give insight in the data quality. Section 4.2 elaborates on factors that mostly influence the model opinion with emphasis on deterministic model validation topics. Section 4.3 focuses on the time spent per model validation, and we investigate what this allocated effort is dependent on. Section 4.4 presents current model validation performance measures, by analysing the number of model deficiencies found during validations. Finally, we answer Sub-questions 1.3, 2.1, and 2.2 in Section 4.5.

## 4.1 Data collection

For the data gathering of performed model validations, we have examined planning sheets of the model validation team giving insight in the validations performed and the hours spent. The library of model validation reports and scorecards have additionally been used to extract information on the model validation findings (model opinion, gaps, simplifications, and limitations). An overview was not available beforehand, which made this process time intensive. We used the model inventory for finding the model purpose, model risk, materiality, and complexity. Table 2 gives an overview of the variables gathered per model.

We have however experienced that planning sheets have not always been updated and hours spent have thus not always been found. Also, we found that the planning sheets did not always use an unique model ID, and that model names did in some cases not match the model name stated in the model inventory. This caused that the data gathering process was time intensive.

Table 2 also gives the completeness of each variable, that is: Which percent of the entries in the 2017 and 2018 planning sheets have data available for the variable? We have gathered a total number of 250 entries. We see that for roughly one sixth of the entry points, a model validation report or scorecard were not available. Most of these reports are not available because of business units Aegon sold (for example the operations in Ireland). Furthermore, for roughly one third of the entry points an hour-spent quantity is missing. The reason for this is that hours spent have not been tracked and updated regularly.

Table 2: Data variables.

| Source | Variable | Form | Accuracy |
|---|---|---|---|
| Planning | Model name | String | 100% (250) |
| Planning | Validation year | 2017/2018/2019 | 100% |
| Planning | Region model use | Group, NL, US, etc. | 100% |
| Planning | Planned hours | Quantity in hours | 57.4% |
| Planning and check | Hours spent | Quantity in hours | 62.2% |
| Planning | Lead validator | Validator name | 96.4% |
| Model inventory | Model risk | High, medium, low | 98.4% |
| Model inventory | Complexity | High, medium, low | 94.0% |
| Model inventory | Materiality | High, medium, low | 94.0% |
| Model inventory | Model purpose | IFRS, pricing, Solvency II, etc. | 91.6% |
| Model inventory | Kernel | Excel, C++, SQL, etc. | 90.8% |
| Reports and scorecards | Model opinion | Green, yellow, amber, red | 85.5% |
| Reports and scorecards | Topic opinions | Green, yellow, amber, red | 83.1% |
| Reports and scorecards | Findings | High, medium, low; per topic | 83.1% |

## 4.2 Model opinion analysis

For the data-analysis in this section, we evaluate two outputs. The first is the **model opinion**, which is either green, yellow, amber, or red. For the model opinion analysis it is interesting to look for influences that impact the model opinion. The goal is to find indicators that predict if a model is likely to be faulty or not. For the evaluation of the second output, the **topic opinion**, we give analyses on the influence on the model opinion, and the effectiveness of review in Section 4.2.3.

### 4.2.1 Model opinion overview

Of all entry point, 212 entries contain a model opinion. The distribution of green, yellow, amber, and red opinions is found in Table 3. How the model opinion is formed is described in Section 3.3 with a process visualisation in Figure 2.

Table 3: Distribution of opinions.[13]

| Opinion | Count | Percentage |
|---|---|---|
| Green | | |
| Yellow | | |
| Amber | | |
| Red | | |
| **Total** | | |

[13] An aware reader might notice that the total count of model validations of which a model opinion is present should be 213 (looking at the percentage mentioned in Table 2). However, one model validation has not led to a model opinion since too many limitations were found.

The distribution shows that the most frequent model opinion is the yellow opinion, which is described as 'further improvements are recommended'.

### 4.2.2 Model risk, complexity, and materiality

Figure 3 gives insight in the model opinion per model risk class. We see that a low model risk class usually leads to a relatively good model opinion, while the medium model risk class gives the most severe model opinions in comparison. A possible reason for a better model opinion outcome for high risk models is that model developers might tend to put more effort into creating a robust model, since they realise that validation efforts are going to be relatively intense.

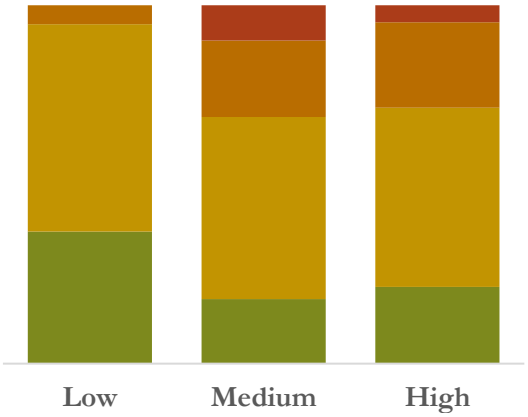



Figure 3: Model opinion per model risk class.

The relation between the model opinion versus complexity class or materiality class can be seen in Figure 4. Before making statements about the differences between Figure 3 and Figure 4, we should be aware of the correlation between the model risk class versus the complexity & materiality. When evaluating complexity individually in Figure 4, we see no clear difference in the severity of the model opinion for different complexity levels. This is quite remarkable, since we can assume that Aegon uses the complexity levels to determine model risk.

Contrarily, an increase in the materiality class holds a positive correlation to the severity of the model validation opinion, although this effect decreases when comparing medium and high materiality. We expect no clear correlation, since the amount of money that a model represents should not influence the model validation opinion. Still, a possible explanation is the thresholds for the severity of financial gap and simplification findings. These thresholds are percentages of the model materiality, but also contain a limit in absolute terms for becoming a high severity finding. Gaps found for high materiality models have a higher probability of reaching this limit, which thus influences the model validation opinion.

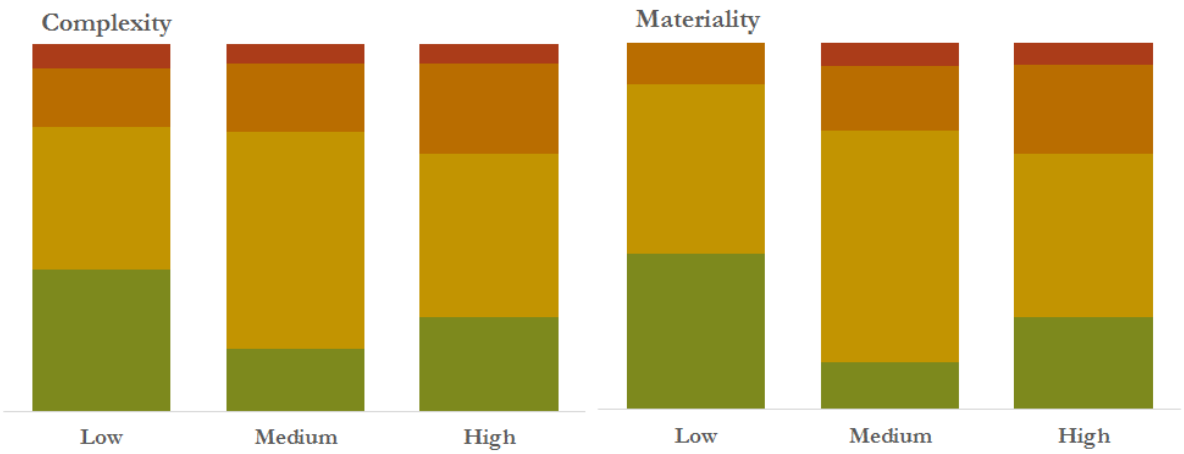



Figure 4: Model opinion versus complexity class (left) and model opinion versus materiality class (right).

### 4.2.3 Model validation topic

We now analyse non-green model validations in this section. It is interesting to see which (set of) model validation topic(s) has/have been determining the model validation opinion. We define this deterministic power as the occurrence when a single or set of model validation topics has/have the most severe topic opinion, and thus individually or as a set influence the model opinion. The model validation topics have been addressed in Section 3.2.

In Table 4 we see that the second model validation topic – 'model development and testing' – has the largest influence in determining the model opinion, followed by the first model validation opinion – 'methodology'. The last two validation topics – 'preparing for and validating model runs' and 'reporting and use of results' respectively – have the smallest influence.

But, as we described in Section 3.3 - a yellow model opinion is still acceptable, and the model validation frequency is not increased. Therefore, it is also interesting to analyse the deterministic power only when the validation has resulted in amber or red model opinions. For this, see Table 5. We again see that the last two validation topics have small influence, but also the third topic – 'data' – has a relatively small stake.

Table 4: Model validation topic influence on yellow, amber, and red model opinions.

| MV Topic | Single | Duo | Trio | Quadruple | Weighted score[14] |
|---|---|---|---|---|---|
| **Meth** | 18.4% | 55.6% | 45.0% | 75.0% | **10.33** |
| **Mdev** | 52.6% | 59.3% | 75.0% | 100% | **18.69** |
| **Data** | 7.9% | 31.5% | 45.0% | 62.5% | **5.88** |
| **Assu** | 9.2% | 27.8% | 55.0% | 87.5% | **6.20** |
| **Prep** | 5.3% | 20.4% | 50.0% | 37.5% | **4.28** |
| **Report** | 6.6% | 3.7% | 30.0% | 37.5% | **2.62** |
| *Weights (I)* | *0.4375* | *0.2625* | *0.175* | *0.125* | |
| *Total # (II)* | *76* | *54* | *20* | *8* | *158* |

Table 5: Model validation topic influence on amber and red model opinions.

| MV Topic | Single | Duo | Trio | Weighted score[13] |
|---|---|---|---|---|
| **Meth** | 19.4% | 52.4% | 66.7% | **12.18** |
| **Mdev** | 54.8% | 66.7% | 66.7% | **22.18** |
| **Data** | 3.2% | 23.8% | 33.3% | **7.27** |
| **Assu** | 16.1% | 38.1% | 66.7% | **11.27** |
| **Prep** | 0% | 14.3% | 0% | **6.00** |
| **Report** | 6.5% | 4.8% | 66.7% | **8.55** |
| *Weights (I)* | *0.5* | *0.3* | *0.2* | |
| *Total # (II)* | | | | |

### 4.2.4 Model purpose and kernel

The variety of programs used for modelling across Aegon is very broad. But for only three kernels, we deem the number of validation entries large enough (more than ten) to taken into account for researching the kernel influence on the model validation output: MS Excel, Moses[15], and ALFA. We find the full kernel analysis result in Appendix B.2, which shows that models built in MS Excel and Moses have a high percentage of amber and red model opinions.

Regarding the model purpose, Figure 5 shows that for Solvency II, IFRS, MVN, and MTP purposes, the amount of amber and red opinions is the smallest. Purpose types 'banking' and 'pricing' have a relative high percentage of these opinions.

[14] We calculate the weighted score by using two weights. The first weight is found in the second row from the bottom, and is used since a topic that individually affects the model opinion should be weighted heavier than being part of a group of topics. The second weight is based on the number of singles, duos, triples, and quadruples relatively to the sum over all (bottom row).

[15] Including Risk Agility FM software.

Figure 5: Model opinion per model purpose type.

## 4.3 Validation time spent analysis

Now that more information is available on the dependencies for the model opinion, the **time spent** for model validations can be evaluated. The total distribution function of time spent can be found in Appendix B.1, which indicates that a model validation usually does not take more time than 420 hours (6 outlier model validations took longer). The overall average amount of time spent on a model validation is equal to 185 hours.

Table 6: Average hours spent per model opinion.

| Model opinion | Hours (avg.) |
|---|---|
| Green | 181 h |
| Yellow | 172 h |
| Amber | 216 h |
| Red | 224 h |

The first evaluation is an illustration of hours spent per model opinion (Table 6). It becomes clear that a model validation leading to a green model opinion takes on average slightly more than 180 hours (22.5 working days). Surprisingly, a model validation leading to a yellow opinion requires less time on average. A model validation leading to an amber model opinion takes more time on average, and the process resulting in a red model opinion takes almost 25% more time than the amount of a 'green validation'. The remainder of this section will focus on factors identifiable beforehand that influence the model validation time.

### 4.3.1 Time dependency - model risk, complexity, and materiality

Table 7 shows that there is a positive correlation between the average model validation time and the model risk. This is not surprising, since the model validation team is allocating validation efforts in proportion to the model risk class. We also find that there is no clear difference in model validation time for low and medium complexity & materiality models. Again, the complexity variable shows a surprising result as well as the materiality variable. We can assume that increasing complexity means that it takes longer to understand and validate a model.

Table 7: Average hours spent per severity class.

| | Complexity | # | Materiality | # | Model Risk | # |
|---|---|---|---|---|---|---|
| **Low** | 159 h | 12 | 134 h | 12 | 113 h | 8 |
| **Medium** | 158 h | 33 | 135 h | 15 | 170 h | 32 |
| **High** | 226 h | 54 | 221 h | 69 | 219 h | 69 |

### 4.3.2 Time dependency - model purpose and kernel

The time spent on a model validation is dependent on the model purpose, see Table 8. In this table, solely model purposes which are registered for more than 10 models are taken into account. Furthermore, it must be noted that a model can have several model purposes.

Table 8: Validation time spent per model purpose.[16]

| Model purpose | Hours (avg.) |
|---|---|
| Solvency II | 175 h |
| IFRS | 229 h |
| MVN | 166 h |
| Pricing | 157 h |
| MCVNB | 200 h |
| MTP | 198 h |

When considering the model kernel, only MS Excel is used by more than 10 models for which sufficient validation data is available. The average amount of time spent for this is 175 hours with a total number of 89 models which have been validated. Historical validations that involve MS Excel thus on average take less time than the overall average.

[16] Model purpose abbreviations are as follows: IFRS – international financial reporting standards, MVN – market value numbers, MCVNB – market consistent value new businesses, MTP – medium term plan.

## 4.4 Validation performance measures

In this section we investigate what the model validation performance measures are. We do this by evaluating the number of model deficiencies found per evaluated model validation topic per time unit. This can either be measured in absolute or relative terms. **Absolute** means that low severity and high severity findings have equal weight. With the **relative** measurement, we scale down high and medium severity measures to low severity measures. The difference between these measurements is addressed in the sixth part of the leading example.

The translation factor for financial gaps is based on the materiality thresholds. A translation factor for scaling different control gap severities is defined in the Aegon policies but will not be mentioned in this research paper due to confidentiality. Since simplifications and limitations are logged by the model developer before the validation takes place - and since these findings do not influence the model validation opinion for certain types of model; simplifications and limitations are out of scope for this analysis.

**Leading example: Endowment insurance (Part VI)**

As the previous part of the leading example series has stated, the boundaries for determining the severity for financial gaps are, as follows:

- *Low* if finding is below 1% of EC.
- *Medium* if finding represents 1% of EC or more, and less than 2% of EC.
- *High* if finding represents 2% of EC or more.

The methodology for determining the **absolute** measure of summed financial gaps is straightforward. A low financial gap is namely weighted the same as a high financial gap. For the **relative** measure, this is done differently. High and medium financial gaps are expressed in a number of low financial gaps according to the following methodology:

Since a low financial gap has a minimum impact of 0 and a maximum impact of 1% of EC, it is assumed that the average impact is 0.5% of EC. For a medium financial gap, this average is 1.5% and thus 3 times as severe as a low financial gap. Since there is no maximum to the impact of a high financial gap, it is assumed that the average impact is somewhat above the minimum of 4%.

A medium financial gap is therefore translated to 3 low financial gaps. And a high financial gap is assumed to be equal to 8 low financial gaps. The translation of high severity control gaps to low severity control gaps is done the same way, however, with more arbitrary (confidential) boundaries.

### 4.4.1 Performance measure per model risk class

Since Aegon is classifying each model according to its assumed model risk, and has policies and planning goals dependent on this model risk class – it is interesting to see how many model deficiencies are found. Thereby, we must realise that a higher model risk class means that the model validation time allocated is increasing such that more deficiencies can be found (Table 7). Still, a clear distinction of the number of deficiencies found would be expected when evaluating the model risk class.

Table 9 therefore shows the number of model deficiencies found without the time spent for validation taken into account, whereas Table 10 does include the time spent. Both tables also evaluate the severity of the findings in the two most right columns ('relative'). To pinpoint the difference between these measures, financial and control gaps measured in absolute terms are presented in capital letters (FG and CG), and the relative measured amounts are denoted using the opposite (fg and cg).

Table 9: Number of financial gaps and control gaps found per validated topic.

| | *Absolute* | | | *Relative* | |
|---|---|---|---|---|---|
| **Model risk** | **FG+CG** | **FG/Topic** | **CG/Topic** | **fg/Topic** | **cg/Topic** |
| Low | 1.176 | 0.148 | 1.028 | 0.287 | 2.380 |
| Medium | 1.416 | 0.326 | 1.090 | 1.529 | 3.480 |
| High | 1.871 | 0.489 | 1.382 | 1.838 | 4.145 |

Table 10: Number of financial gaps and control gaps found per validated topic - **time spent included**.[17]

| | *Absolute* | | | *Relative* | |
|---|---|---|---|---|---|
| **Model risk** | **FG+CG/t** | **FG/Topic/t** | **CG/Topic/t** | **fg/Topic/t** | **cg/Topic/t** |
| Low | 0.1782 | 0.0210 | 0.1572 | 0.0488 | 0.3832 |
| Medium | 0.1523 | 0.0314 | 0.1209 | 0.1487 | 0.3347 |
| High | 0.1955 | 0.0354 | 0.1601 | 0.1630 | 0.3965 |

There is no horizontal relationship between the absolute and relative measures. The multiplication factor between absolute and relative measures only indicates how severe the findings are. For instance, concerning medium risk models, we see in Table 9 that in absolute terms 0.334 financial gaps have been found per model validation. In relative terms, 1.577 low severity financial gaps are found. This means, that on average, a financial gap finding for the medium risk models holds the severity of almost five low severity financial gaps.

We see in Table 9 that the number of findings (FG+CG, fg, cg) per evaluated topic increase as the model risk increases, which is to be expected since more validation time is allocated. Considering the time effectiveness, Table 10 shows a spread in effectiveness for finding control gaps – both in absolute and relative measures. The effectiveness for finding financial gaps only increases slightly when a medium or high risk model is validated in comparison to a low risk model.

[17] The time spent is in terms of 100 hours.

### 4.4.2 Performance measures per validated topic

As we have seen in Section 4.2.2, certain model validation topics have a higher influence in determining the model validation outcome than others. We can thus expect that the number of findings for high influential model validation topics are higher, but what is the time effectiveness of the validation of each six topics? Before researching this, we have asked seven Aegon validation seniors to state how much time they spent on each model validation topic. The averages are given in the following figure.


Figure 6: Time stake per model validation topic.

We use this validation effort allocation for adjusting the number of model deficiencies found per 100 hours according to the time spent on each validation topic. We illustrate this by analysing the absolute and relative measurement in Figure 7 and Figure 8 respectively. In these figures, the red and blue line represent the average amount of findings found over all six validation topics. We can thus easily see if a the validation of a certain topic holds an above-average or below-average effectiveness.

Figure 7 shows that using the absolute measure, the model validation topics 4, 5, and 6 provide fewer model gaps found per time measure when combining the findings. In relative terms, Figure 8 shows that the same topics are underperforming considering control gaps. Topics 2, 5 and 6 are below average regarding finding the relatively scaled financial gaps.

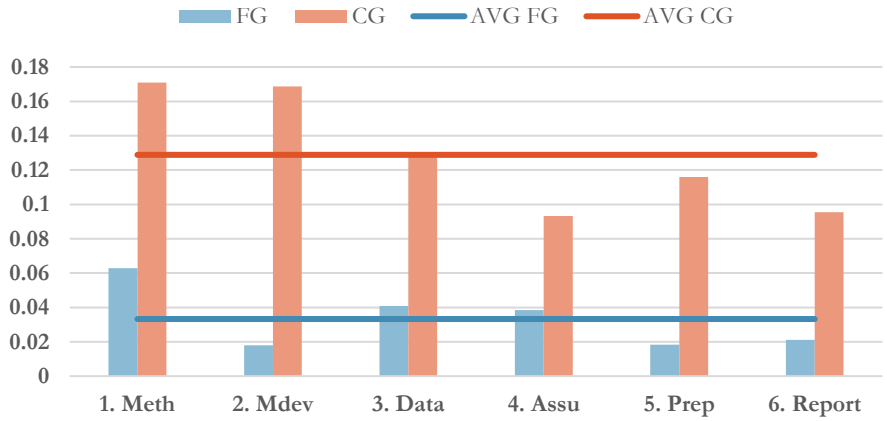



Figure 7: Average number of financial and control gaps found per 100 hours (absolute).

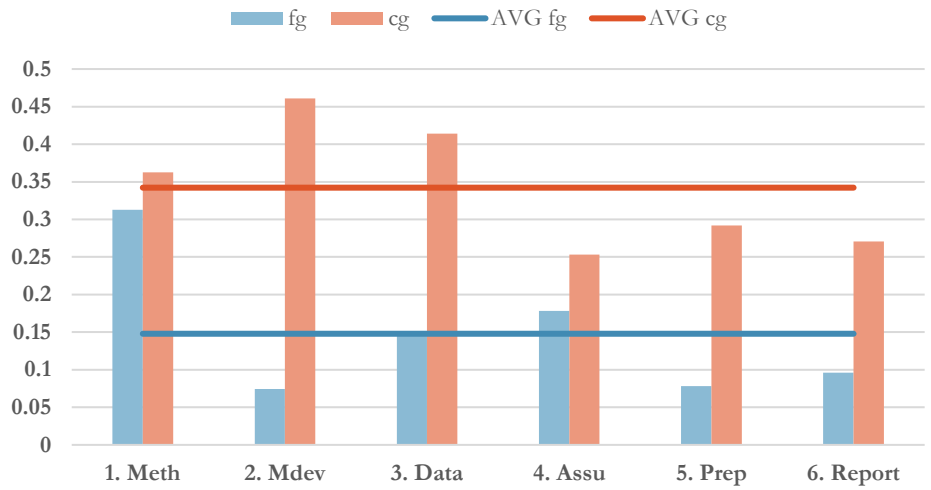



Figure 8: Average number of financial and control gaps found per 100 hours (relative).

## 4.5 Data study conclusion and summarising possible process optimisations

This section summarises the findings presented in this chapter and directly answers Sub-question 1.3: *What factors influence the model validation outcome, on what factors is the time spent dependent, and what are the current performance measures?* Furthermore, Section 4.5.2 will highlight the relationship to the main findings of Section 3.5.

### 4.5.1 Chapter 4 findings

Factors influencing the model validation outcome

- The *low* **model risk** class has a relatively small group of model validations leading to a severe model opinion. Surprisingly, the *medium* class leads to the highest number of yellow and red opinions (Figure 3). The same conclusion can be drawn considering the model **complexity** (Figure 4).
- **Model validation topics** 1, 2, and 4 (meth, mdev, data) are most influencing the model validation opinion. The time effectiveness of all six model validation topics is later addressed in this section.
- Regarding the model **kernel**, models developed in MS Excel and Moses have a relative high number of model validations leading to an amber or red model opinion in comparison to ALFA. This also counts for models with a banking or pricing **purpose**.

Factors affecting the validation time spent

- There is a positive correlation between the **model risk** class and the time spent for a model validation. This is straight forward, since the model validation is allocating planned validation time according to this variable.
- Validations leading to a yellow **model opinion** take the least time on average, whereas coming to an amber or red result take far above average time.

- Validations performed on models having Solvency II, pricing, or MVN **purposes** clearly take less time than average, whereas IFRS models take significant more time.

Validation performance measures

- When evaluating the **model risk classes**, Aegon is effective in pre-determining the model risk since more model deficiencies are found as the model risk class becomes more severe. This is due to the influence of the financial gaps, the control gaps show more spread. However, when taking time effectiveness into account, the findings per time unit do not increase as the model risk becomes more severe. When evaluating the financial gaps in absolute or relative measures, we can conclude that there is no clear difference in time effectiveness between a medium risk model or high risk model.
- Regarding the findings found for each of the **model validation topics**, when taking the relative measure, we conclude that historical validations show that model validation Topics 5 and 6 (prep, report) yield less efficiency in finding model deficiencies (for both absolute and relative measures). Additionally, finding severe financial gaps takes on average more time for validation Topic 2 (mdev), and control gaps for Topic 4 (data) as well.

### 4.5.2 Model validation time reduction possibilities

Up until this point, we analysed gathered information in order to detect points for improvement. We now give an overview of subject discussed so far, and how these are used for finding process improvements. We refer to the main findings of Section 3.5 as 'Chapter 3 finding §x'. This section thereby gives answer to Sub-question 2.1: *What methodologies could possibly result in decreasing the model validation time of medium risk models and how will these affect the quality of model validations?*

**Method 1: Limiting the model validation scope**

Historical validation analysis shows that certain validation topics are more deterministic and time effective as others. Validation topics 'data' (3), 'preparing for and validating model runs' (5), and 'reporting and use of results' (6) have relative small deterministic power. This effect can be partly explained by the findings of Figure 7 – namely that less gaps are found. The historical analysis also indicates that Topics 5 and 6 lack time effectiveness, further research emphasis will be assigned to these validation topics in this research paper. For this, Chapter 3 findings §1 and §9 will be used.

This method will however affect the quality of model validations. If the scope is to be decreased, model validators will decrease the model validation time by not validating certain parts intensely. The number of deficiencies found for those parts will decline. On the other hand, model validators can increase the number of models validated per year and can reallocate their time on validation topics with higher efficiency – which possibly reduces the model risk of Aegon as a whole. The effect on the average model validation time can be found in Section 6.1.

**Method 2: Reconsidering the model risk classification and prioritisation**
We believe that the model risk classification currently operational does not flag the actual model risk or likeliness to find deficiencies detected after the validation. Moreover, the time effectiveness does not fully support the current classification of model risk. These findings support Chapter 3 findings §2 and §8, which state that reconsidering the model risk classification and validation frequencies can decrease the overall model validation time, or that detecting model risk can be done by a more data analytical approach. Chapter 3 finding §8 shows that literature gives examples for expanding the model inventory such that more characteristics of models are stored. In Section 4.5.1 we have listed several model variables influencing the model validation opinion, and we therefore believe that these can be used for predicting model risk.

This method can lead to an increase in the quality of model validations. The process of validating models will not necessarily be altered, and is therefore not a straightforward solution to our research goal. However, our view is that this method can be used for increasing the time effectiveness and decreasing model risk of Aegon. We further elaborate in Chapter 5.

**Method 3: Increasing the effectiveness of the model validation reporting template**
We have not analysed this method in Chapter 4, but we mentioned the possibility of altering the reporting template in Chapter 3 - see Chapter 3 findings §3 and §10. Also, when reconsidering the reporting template, we must take Chapter 3 finding §5 into account since it is advised to meet the minimum requirements for reporting proposed by DNB (Section 2.3.3).

Applying this method does not affect the quality of model validations, but influences the reporting quality. If Aegon model validators succeed in choosing which information is essential, the effect of decreasing reporting quality can be minimised.

**Method 4: Validation concurrent to model development (according to Sargent (2013))**
This method is solely based on Chapter 3 finding §11 and an initial pitch regarding this method is given to the Aegon model validation team. The response was that this methodology was used in the past, but was experienced as ineffective. Therefore, we do not further discuss this method.

**Method 5: Mitigating model risk by decreasing complexity**
There is an indirect effect between decreasing complexity of models and decreasing the model validation time. We can be assume that less complex models take less time to validate, but decreasing the complexity of models cannot be done on short notice. Decreasing complexity can be done in many ways, depending on the design and purpose of a model. Since we find a high variety in both these measures, we do not be further discuss this method in detail – but mention this in Section 6.4.

# Chapter 5
# Redesigning the model risk classification

Our theoretical, company, and empirical analyses show that there is room for improvement regarding the Aegon model risk classification (Table 1). In this chapter, we shortly summarise the current situation, and after that we present insights from the financial services industry as well as self-crafted ideas to improve the classification. The main focus is thus finding an answer to Sub-question 2.2 for the proposed 'Method 2' of Section 4.5.2. This means that we design a tool, in order to show how this alteration of the Aegon model validation policy is supported. We highlight the requirements the tool should embody in Section 5.2 when introducing the new 'likelihood of model error' metric. In Section 5.3 we present the model design, and elaborate on the results obtained from a single model run. Lastly, we discuss the tool performance in Section 5.4.

## 5.1 Aegon model risk classification challenges

Our goal of redesigning the Aegon model risk classification is to better prioritise model validation efforts on various models of Aegon. That means that ideally a model's risk class has a positive correlation with the negativity of the model validation output. This does not directly meet our main goal of this research paper, but it can increase the effectiveness of the Aegon model validation team – namely by decreasing the model risk exposure. Summarising, redesigning the Aegon model risk classification should at least aim to provide a solution to the following challenges we found in the previous chapters.

1. The model risk classification currently depends on the metric 'complexity', which is judged by the model developer. Although this hypothesis has not been confirmed, the model developer might have some incentive to underestimate the model complexity to obtain a lower model risk class. A lower model risk class holds a lower validation frequency cycle, and thus decreases the effort necessary for providing validation support.
2. The metric 'complexity' can also be seen as vague: As Section 3.2 shows, the model validation policy contains no standardised quantitative or qualitative analysis given to determine the metric.

3. Empirical analysis shows that both metrics 'model risk' and 'complexity' do not have a strong positive correlation to the model validation opinion (see Figure 3 and Figure 4). The metric 'materiality' however does have a positive relation. Also, validation performance yields improvement in line with model risk (see Table 9 and Table 10).
4. The categorisation of models is broad. There are three classes and a validation frequency is assigned. After pitching the idea for redesigning the model risk classification to Aegon, we concluded that it could be beneficial for Aegon to have a larger scale; i.e. have a prioritisation score over all models that indicates which models need to be validated first (because they hold a high model risk).

To take on these challenges, we use the insights provided by Black et al. (2018) as a starting point - previously highlighted in Section 2.4.2. The authors propose to use meta data – known attributes before a model is validated – for determining the likelihood of a model error. Examples for these meta data attributes are found in Appendix A.4. Our view is that using this methodology can provide the following solutions to the challenges previously listed.

1. The methodology is less likely to be used by model developers to hypothetically influence the model validation schedule, since it is dependent on more attributes.
2. Determining the likelihood of a model error becomes more transparent, especially when historical validation results are used.
3. Since historical validation data are used, it becomes possible to use an optimisation algorithm to increase the validation efficiency.
4. The likeliness of a model error can use a larger scale. This gives the possibility for the Aegon model validation team to prioritise the validation schedule each moment in time.

## 5.2 From 'complexity' to 'likelihood of a model error'

The challenges we listed coming along with the use of the metric 'complexity', do not mean that the metric cannot be of value for estimating the new metric 'likelihood of a model error' (in short: LME). The variable 'complexity' has an advantage, namely that the model owner can give an indication to what extent the model development and model methodology was challenging. We believe that this could to some extent correspond with the likeliness of an error made. But as we stated in Section 5.1, solely depending on this variable brings challenges.

### 5.2.1 Available building blocks

We let the LME depend on several different variables. Using variables that are already recorded in the model inventory will increase the possibility for implementation at Aegon. We listed the model metrics currently stored in the model inventory in Section 3.4, of which the following metrics are likely to influence the LME.

- Complexity.
- Calculation kernel.
- Model purpose.
- Model developer (possible to categorise per team, country unit, or region).
- Initial (full) or follow-up validation.
- Last model opinion (if applicable).
- Time since last validation (if applicable).

Indirectly, the bottom two of this list are used for the model validation planning. The report of Rikkert (2019), the Aegon model validation policy, states that if a model opinion is amber or red, the model is scheduled to be validated earlier than the validation frequency of 3, 4, or 5 years. These frequencies related to the model risk class are chosen to limit the model risk exposure. It can however be questioned if the order of actions is correct. Does the time since the last validation influence the LME, or should the time gap between validations be chosen based on the LME?

The answer to this last question is very important for modelling the LME. A model can be created for a deterministic classification of model risk, similar to the current risk classification Aegon is using. We can then assign validation frequency targets to each model risk class. This methodology is preferred if one believes that the time since the last validation has a relative weak influence on the LME. On the other hand, we can decide to model a LME score such that in any point in time it becomes clear which model should yield the priority to be validated next. The problem is however, that we did not track the time since the previous validation during the data gathering in Chapter 4. Collecting these data can result in a valuable future building block for model risk managers.

When using this building block in the future, we must take into account the following: The Aegon model validation policies set goals for validation cycles. The validation cycle is three years for high risk models, four years for medium risk models, and 5 years for low risk models. This cycle is increased to every single year, or every two years, when a previous model opinion was red or amber respectively. It is likely that a model which obtained a negative model opinion is rapidly revaluated and improved between validations, which could result in a **negative correlation** between the time since last validation and severity of the model validation outcome. This factor might be disruptive in determining the LME, and might only be beneficial is the previous model opinion is either green or yellow. We therefore suggest to split initial (full) and follow-up validations when the amount of data is sufficient.

### 5.2.2 MRM future building blocks

As we mentioned in Section 3.4, the Aegon model risk managers aim to expand the model inventory. A pilot is currently running involving one country unit, where the following information (additional to the metrics of Section 5.2.1) is tracked:

- Model user.
- Expert judgements.
- Simplifications and limitations.[18]
- Models used for input.
- Dependent models.
- Time since last validation (if applicable).[19]

Next to these metrics, we can think of other candidates. As Appendix A.4 shows, variables such as the number of the number of people familiar with the methodology, the number of developers, the code coverage, etc. can also be tracked (Black et al., 2018). But, collecting more model metrics is time intensive, and asking model developers to manually enter an increasing number of entries can be discouraging.

### 5.2.3 The unit of LME

Now that we listed the available and possible future predictive variables in Sections 5.2.1 and 5.2.2, we need to decide what the possible quantities of the LME are. The possibilities for the predictive measure are listed below.

1. Model validation opinion – {green, yellow, amber, red}.
2. Model validation opinion (grouped) – {green or yellow, amber or red}.
3. Absolute number of findings expected – {0, 1, 2, …}.
4. Relatively measured number of findings expected - $\{fg \in \mathbb{R}_{>1}\}$ and/or $\{cg \in \mathbb{R}_{>1}\}$.

Each of these predictive measures has advantages, but also disadvantages. The variety in possible outcomes is low for the second measurement. With this, we mean that a slight amber model opinion is threated very different as a severe yellow model opinion. The advantage for the first two measurements is that the categorisation does not alter the definition of model risk in the current Aegon policies. The bottom two have a broader scale, and enable to focus on financial gaps or control gaps together or separately. Also, the performance measures (as introduced in Section 4.4.1) can be taken into account when prioritising validation efforts.

| Ideally, the tool contains the possibility to predict each of these four listed measures such that the user can prioritise on his or her liking. |
|---|

[18] As mentioned in Section 3.3, simplifications and limitations are logged during model development.

[19] Please read the second passage of this page for the reasoning why the 'time since last validation' is considered to be a future building block, as well as an available building block (see Section 5.2.1).

### 5.2.4 The role of materiality

The materiality is still an important variable, it namely determines the severity of model errors. If the model validation prioritisation tool wants to state which models have the highest priority for validation effort, this severity should be taken into account as a multiplier to the LME. To increase the short term applicability for the LME tool for Aegon, we let the current classification of materiality remain the same. Table 1 indicates that materiality and complexity (i.e. probability of model error) have an equal weight. The multiplication of materiality on the LME should thus be adjusted to the full scale of LME values.

> The materiality should be taken into account by the tool when determining the model validation prioritisation. The current classification of materiality will be used, and will have the same weight as the LME.

## 5.3 Modelling the LME

The programming environment of Aegon MRM is Power BI, a business analytics service of Microsoft. It is possible to implement R scripts in Power BI. On a first glance, we see that a supervised learning algorithm using a classification or regression analysis (dependent on the unit of LME) for the building blocks mentioned in Section 5.2.1 can be used. This means that initially the LME is estimated using the database of analysed historical validations of Chapter 4, and the influence of variables is recalculated each time a new validation is added to the database. In this section, we present how the input variables are structured and we give an elaboration of the classifier algorithms used. We give the results of one model run as an example.

### 5.3.1 Structuring input variables

Looking at the current building blocks, we see that all variables but one (time since last validation) are categorical. For supervised learning purposes, these variables need to be reformed to binary categorical variables. For instance, initially the complexity is denoted as either low, medium, or high. For modelling, we transform this variable into three binary variables – 'LowComplexity', 'MediumComplexity', and 'HighComplexity' – which either hold the value *0* or *1*. Having a high number of binary variables brings challenges when trying to predict variables with a continuous scale using regression analyses. The input variables for predicting the LME can be found in Appendix C.1.

As previously said, the only variable with a continuous scale is the **time since last validation**. However, we find this variable challenging to use for predictive purposes due to the current model validation frequencies for models who have obtained a severe model validation outcome. Also, there is a lack of available data since roughly a third of the historical validations were follow-up validations and the data set itself is small.

Since no continuous variable is available for the LME (using available building blocks), it is not possible to predict the absolute number of findings expected, nor the relatively measured number of findings expected (listed third and fourth in Section 5.2.3). Our current efforts for determining the LME will therefore only target predicting **the individual or grouped model validation opinion**. For predictive purposes, we evaluate a classification tree and random forest tree algorithm. For applying these, we split the dataset randomly in a training set and test set using a split ratio of 0.75. For comparison purposes, we use a constant seed value for obtaining this random number.

### 5.3.2 The decision tree classifier

The decision tree is a supervised learning algorithm that is used for regression or classification purposes. A decision tree classifier is easy to visualise and explain, but may lack predictive accuracy in comparison to other algorithms (Le, 2018). The dataset on which the algorithm is applied is split into smaller 'branches' or 'sub-trees'. Splits are performed by 'decision nodes'. Splitting is applied until the subset is divided far enough such that it contains a large percentage of entries with the same remark ('terminal nodes'). For predicting the LME, this means that the algorithm tries to form groups of entries with a green (or yellow, or amber, or red) model opinion.

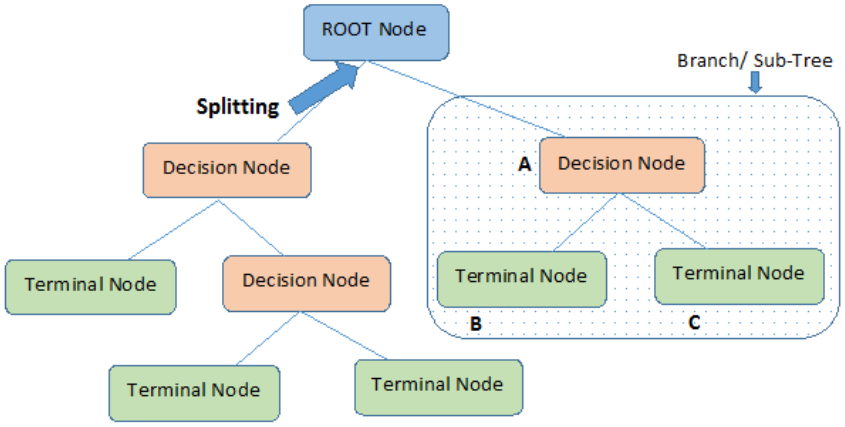



Figure 9: Decision tree methodology and terminology, from Le (2018).

We apply the decision tree classifier to the training sets for predicting the individual or grouped model validation opinion. To optimise the algorithm output, we use resampling parameters to streamline the classification (via the 'caret' package in R). We use the 'repeatedcsv' method for resampling, using fifteen sets of parameters ('tunelength'), which we repeat three times. R-project (2019) broadly elaborates on the caret functionalities. See Appendix C.2 for our R-script.

The decision tree we programmed for predicting the individual model validation opinion is found in Figure 10. The tree classifies models - having a Solvency II purpose, no green previous opinion, not developed by the US country unit, and not modelled in Excel - as most likely to obtain a yellow model validation opinion. The tree is not suitable for predicting red model validation opinions, possibly due to the small number of models having a red opinion in the training set.

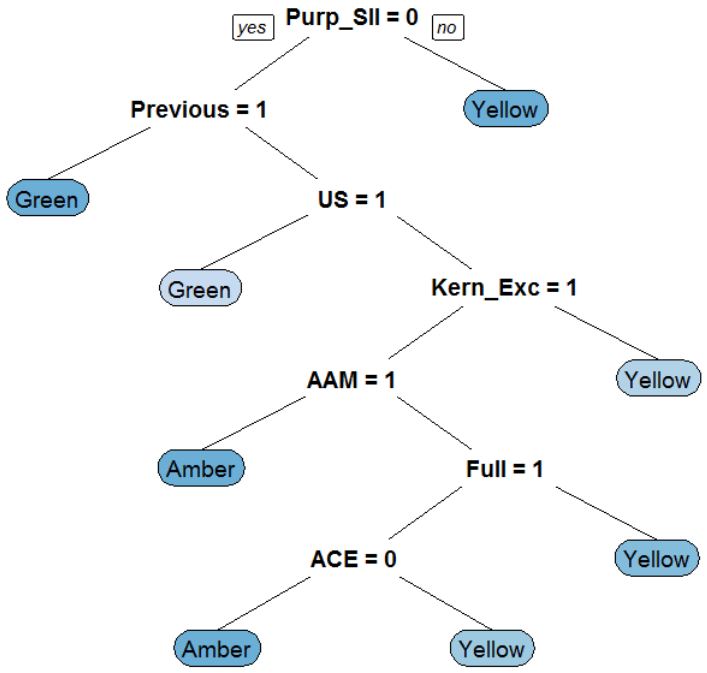



Figure 10: Decision tree classifier for model opinion prediction (seed = 407).

We now use this tree for predicting the model validation opinion of the models in the test set. We then compare the prediction to the actual model validation opinion, and the number of correct and incorrect predictions are shown in Table 11. The **accuracy** of the decision tree classifier is calculated as the sum of the diagonal divided by the sum of all, and gives a percentage of **53.8%**. This result shows that the algorithm is unsuccessful in predicting validations resulting in an amber or red model opinion. To be certain of this conclusion, we must perform several model runs with different train and test sets. These results are given in Section 5.4.

Table 11: Decision tree classifier performance for model opinion (seed = 407).

| | **Actual →** | | | |
|---|---|---|---|---|
| **↓Predictor** | **Green** | **Yellow** | **Amber** | **Red** |
| **Green** | | | | |
| **Yellow** | | | | |
| **Amber** | | | | |
| **Red** | | | | |

Figure 11 presents the decision tree generated for model opinion severity prediction (a severity of 1 means an amber or red model validation opinion). For example, if a model is used for Solvency II purposes, is modelled in Excel, is not used for pricing purposes, and has not been validated before (thus a full validation), the tree predicts that the validation results in an amber or red model validation opinion. We apply this classifier to the test set, and compare the actual and predicted model validation opinion severity.

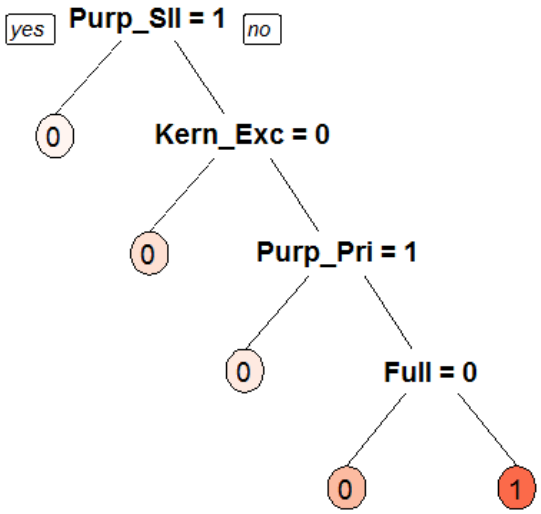



Figure 11: Decision tree classifier for prediction severe model opinion (amber or red) (seed = 407).

Table 12 shows that an **accuracy** of **77.4%** is obtained. However, we question if the performance shown above is applicable in practice. We namely see that the algorithm predicts ten models to obtain a severe model opinion – and this is correct in six of these cases. Being wrong in this case will not have any severe consequences for Aegon, since the goal is not to give an amber or red model opinion always. However, predicting non-severe opinion while the model risk is actually severe is not preferred.

Fawcett (2015) addresses the problem of a relatively high number of 'false negatives' in data studies, and recommends using the **recall** for analysing predictive performances. Recall is calculated by taking the number of true positives, and dividing these by the sum of true positives and false negatives – resulting in a value of only **42.9%**. As stated before, more model runs using different test and train sets are analysed in Section 5.4.

Table 12: Decision tree classifier performance for model opinion severity (seed = 407).

| | **Actual →** | |
|---|---|---|
| **↓Predictor** | **YES** | **NO** |
| **YES** | **6** | **4** |
| **NO** | **8** | **35** |

### 5.3.3 The random forest classifier

The random forest supervised learning algorithm can improve the predictive performance of the decision tree. It is capable of exploring the data set by treating missing values and outliers appropriately, and uses bootstrapping for random sampling (Le, 2018). The algorithm builds several decision trees as shown in Figure 9, but chooses a decision node splitting variable candidate by analysing a random set of candidates. The predictive power of splitting variables is determined by averaging over all individual trees grown. It therefore does not make sense to visualise a single decision tree, but Foreman (2013) suggests that it is beneficial to present the contribution to reducing 'node impurity' of each variable. We use the `varImpPlot` command in R. The results can be found in Appendix C.3.

Table 13: Random forest classifier performance for model opinion (seed = 407).

| | **Actual →** | | | |
|---|---|---|---|---|
| **↓Predictor** | **Green** | **Yellow** | **Amber** | **Red** |
| **Green** | | | | |
| **Yellow** | | | | |
| **Amber** | | | | |
| **Red** | | | | |

Table 13 and Table 14 give the results of one model run using the same train and test sets as we used for the decision tree classifier in Section 5.3.2. The random forest classifier for all possible model opinions obtains an accuracy of **53.4%**, and the classifier for predicting severity reaches an accuracy of **71.4%** and a recall of only **35.7%**.

Table 14: Random forest classifier performance for model opinion severity (seed = 407).

| | **Actual →** | |
|---|---|---|
| **↓Predictor** | **YES** | **NO** |
| **YES** | **5** | **5** |
| **NO** | **9** | **30** |

## 5.4 LME model results and discussion

For determining the performance of the two predicted results, we run the model eight times. When evaluating the performance for predicting the model opinion, we track the accuracy. For predicting the model opinion severity, we record both the accuracy and recall. We described in Section 5.3.1 how these measurements are calculated. We present the result in the table below.

Table 15: LME modelling results for predicting the model opinion (MO) and opinion severity (Sev).

| | Seed | 407 | 9 | 86 | 912 | 435 | 176 | 820 | 502 | Avg |
|---|---|---|---|---|---|---|---|---|---|---|
| **Dec. tree** | **MO acc.** | 0.538 | 0.404 | 0.481 | 0.462 | 0.577 | 0.462 | 0.462 | 0.538 | **0.49** |
| | **Sev. acc.** | 0.774 | 0.754 | 0.679 | 0.717 | 0.774 | 0.734 | 0.511 | 0.717 | **0.71** |
| | **Sev. recall** | 0.429 | 0.143 | 0.357 | 0.286 | 0.286 | 0 | 0.357 | 0.07 | **0.24** |
| **Rand forest** | **MO acc.** | 0.543 | 0.458 | 0.5 | 0.447 | 0.52 | 0.489 | 0.717 | 0.62 | **0.54** |
| | **Sev. acc.** | 0.714 | 0.771 | 0.75 | 0.708 | 0.7 | 0.792 | 0.688 | 0.660 | **0.72** |
| | **Sev. recall** | 0.357 | 0.308 | 0.182 | 0.308 | 0.308 | 0.308 | 0.357 | 0 | **0.27** |

From Table 15 we cannot clearly conclude that the random forest algorithm is performing better regarding the accuracy and recall. More importantly, we must conclude that the results for both algorithms are disappointing. The accuracy of predicting the model opinion is low, and the recall for predicting the opinion severity close to zero. Also, we detect a high variance, especially for model opinion predictions. This indicates that there is a high variance in the dataset itself, since the randomly chosen train and test set highly influence the model run results. We identify all variables being binary classifiers as the main cause. On the whole, using currently available model variables for estimating the likelihood of model error is unsuccessful. We did succeed in creating a starting point by restructuring model validation data and modelled a R-script for future use.

If Aegon decides to continue trying to estimate the likelihood of a model error, recording other variables is necessary. We recommend recording **continuous variables**. As listed in Section 5.2.2, the number of expert judgements, the number of simplifications, and the number of limitations are possible candidates – and are planned to be recorded in the model inventory. Also, the time since last validation can be used although the usefulness is debatable[20]. Black et al. (2018) gives other possibilities (also listed in Appendix A.4 of this research paper).

- Code coverage percentage (to be determined during a validation).
- Number of model developers.
- Number of trained users.
- Number of restatements of model results.

[20] See Section 5.2.1, negative correlation between the time since last validation and the severity of the model opinion is expected since the validation cycle frequencies are increased when the previous model opinion was either amber or red.

Since most variables are recorded in the model inventory, work load for the model risk management department will increase. Validation teams can however put effort into recording several of these variables per validation. The problem is however, that only a relatively small part of models is validated each year. Redesigning the model risk classification by disposing complexity, and estimating the LME can therefore not be achieved on short term notice. We should therefore regard using model data to estimate the LME as a long term project, for which each year new data must be recorded and analysed. We recommend starting with using variables that are planned for recording in the model inventory, and analyse their contribution to predicting model risk. If this is successful, recording new variables such as listed by Black et al. (2018) should be included in the model validation process.

# Chapter 6
# Proposals for process improvements

Up until this point, the problem description and research goals have been presented in Chapter 1. Chapter 2 has summarised the industry's best practices and regulatory requirements. The following chapter presents an overview of the Aegon model risk policies and ends with a comparison to the findings of the previous chapter. The data study of Chapter 4 succeeds in giving force to several differences between the Aegon model risk mitigation process and the industry's proposals, or to possible alterations of the Aegon policies – and concludes with an overview of five possibilities for reducing the model validation time. This chapter (6) concludes this research paper with a section assigned to each of these methods (excluding Method 4) with a proposal for the practical implementation. Section 6.5 presents the answer to Sub-question 2.3 by listing further recommendations not mentioned in the previous chapters of this paper.

## 6.1 Limiting the model validation scope

Chapter 4 holds the information necessary for the considerations regarding this first method for reducing the model validation time. As Section 4.2.3 shows, the historical analysis of validations indicates that the validation topics 'data', 'preparing for and validating model runs', and 'reporting and use of results' have little effect in determining the model opinion (since these topic opinions only have the most severe opinion in a small amount of the validations).

Limiting the scope will however affect the quality of model validations, since in absolute terms the number of model deficiencies found will decrease. Follow up measures for mitigating this model risk can therefore not be performed. Yet, when one takes a more holistic view on mitigating model risk at Aegon, it can be said that the decrease in validation effort originally reserved for less effective topics can be used for validation effort on the next model on the planning list. This results in a more effective way for finding model deficiencies (given that the current model validation schedule is not met because of too little time available).

Our **first proposal** for implementing this method is to limit the scope for medium risk models. Limiting the scope can be done by not validating the three model validation topics mentioned in the first passage of this section. Which topic is excluded should depend on the characteristics of the model. In the early stage of the model validation (after scanning the significant model documentation), the validator can request to rule out one (or perhaps two, or even three) topics.

We do however favour our **second proposal** for implementation, namely to first validate the most deterministic model validation topics before deciding to exclude the validation scope. The validation team can construct rules that are applied when a time reduction is desirable (during a medium model risk model validation). An example is to exclude one or several of the model validation topics mentioned in the first passage of this section when the three other topic opinions all result in green opinions. Since these three topics bear more deterministic power, the probability for the excluded topics having a more severe topic opinion is small.

Now we assume that this method leads to a decline of on average one model validation topic in the model validation scope for the medium risk models – and this decline in scope is equally divided between model validation topics 'data', 'prep' and 'report'. And we know that these topics on average account for 12.1% of the model validation time (Figure 6). Also, Section 3.4 states that the number of medium risk models is roughly equal to the number of high risk models. Furthermore, the validation frequency for medium risk models is four years, and three years for high risk model validations; low risk models are mostly self-assessed. The percentage of medium risk model validations is thus (3/7=) 42.9% of all validations. Altogether, this means that the overall average model validation time is reduced by roughly (12.1% * 42.9% =) 5.2% (for medium and high risk models together). The number of models validated per year will thereby increase with almost 5.5%.

## 6.2 Reconsidering the model risk classification and prioritisation

Chapter 5 focusses on the construction of the new metric 'likelihood of a model error' (in short: LME). This process was however unsuccessful. We assume that the main reason is the absence of continuous model variables. Section 5.4 concludes that estimating model risk cannot be achieved on short term notice. Model risk management efforts are currently directed towards expanding the model inventory. This is beneficial for future achievability of estimating the LME.

Our **proposal** is to use these newly recorded continuous variables for retrying to model the LME. The aim would be to do this within one or two years. If modelling the LME shows improvement, other variables such as the number of trained users, number of model developers, number of model result restatements, and the code coverage percentage (from Black et al. (2018)) can be recorded and used.

Being unsuccessful in modelling the LME does not change the lack of predictive power in determining model risk of the model metric 'complexity'. Section 4.2.2 and in particular Figure 4, show that historical model validation data indicates that the level of 'materiality' has a higher correlation to the severity of the model validation opinion. Therefore, for short term implementation, we **secondly propose** to consider classifying model risk solely based on materiality. Model risk and model validation prioritisation is then fully in line with the impact of a possible model error.

## 6.3 Increasing the effectiveness of the model validation reporting template

Aegon is currently using a reporting template that is around 7 pages (excluding the title page) long. This template is used for every validation and includes notes per section on what should be included in the report. Below is a list of subjects that are documented, of which the bold typed ones are required by the DNB as described in Section 2.3.3. Behind each subject are the number of pages expected as stated in the reporting template.

1. Cover page (1).
2. **Executive summary: overall opinion, key messages, and important limitations** (1-2).
3. **Validation scope, limitations, and approach** (1-2).
4. Overview of key findings per model validation topic **and scoring** (2-3).
5. Background (history, development, process of calculation, purpose, business) (1-3).
6. Validation work per topic (2-4).
7. **A full set of findings** (2-3).
8. **Description of findings,** closed gaps, proposed actions (2-4).

The expected number of pages of the validation report thus lies between 12 and 22 pages. When only focusing on the minimal reporting requirement of the DNB, the validation report could have a length of around 6 to 9 pages. We performed a quick analysis to give insight in the current report size, to see whether this differs from the expected number of pages stated in the reporting template. For this analysis, we selected 20 random validation reports, and the average number of pages are found in the following table.

Table 16: Average number of pages per topic listed at start of this section.[21]

| **Listing** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **Total** |
|---|---|---|---|---|---|---|---|---|---|
| Avg #pages | 1.0 | 1.1 | 1.5 | 2.1 | 1.3 | 3.8 | 1.5 | 2.9 | 15.1 |

Looking at these page amounts, we see that 'validation work per topic' holds the largest average number of pages for reporting. At the same time, according to the guidelines of the DNB, an extensive description should not be necessary. This is also the case for the reporting topic 'background'. But for this subject, the average number of pages is low. Furthermore, the information can be copied from the previous model validation report (after validating for correctness) in case of a follow-up validation. Alterations to this reporting topic will thus not be taken into account for the proposal.

[21] Note that some of the reports did not cover each of the reporting topics. For instance, only 7 of the 20 randomly selected reports did contain an executive summary. The average in those cases is calculated over the non-zero values. The total amount in the most right column is the sum over these non-zero value averages.

The **proposal** for increasing the effectiveness of reporting for low and medium risk class models, is to summarise the 'validation work' - and move this summary in parts to the 'executive summary' and 'validation scope, limitations, and approach' topics.[22] Also, for lower risk class models (or less material models), the other reporting topics can be written more concise. This decreases the report size and time necessary for reporting.

## 6.4 Mitigating model risk by decreasing complexity

As mentioned in Section 4.5.2, there is no short term effect in decreasing model complexities and decreasing the model validation time (or increasing the validation's effectiveness). An ad hoc analysis of all models, and evaluating whether the complexity can be decreased, is time intensive and unrealistic. Complexity can however be judged during a model validation, and is already done in practice. Still, it does not embody large emphasis during an independent model validation.

If the Aegon model validation were to keep the current model risk classification methodology as described in Section 3.2, the **proposal** would be to focus on a model's complexity for models with a higher complexity class than materiality class.[23] It can be assumed that a high complex model comes along with a relative long log of expert judgments, simplifications, and limitations. Also, it is likely that the model testing is relatively intensive. If a model developer would be able to decrease the model complexity, it is possible to limit these logs and model testing efforts, resulting in a decrease in validation efforts.

If Aegon were to implement the reclassification as described in Chapter 5, the proposal mentioned in the previous passage can also be used. Nonetheless, using the LME gives more possibilities. The LME algorithms give insight in factors that influence a model in the likeliness of containing an error. The validator can then evaluate high risk factors and validate if these are necessary and/or can be replaced by less risky characteristics. For example, let's say that a validation is taking place for a model programmed in R, and this programming language has a bad reputation when examining the historical validations. A validator can then validate if this is a necessity, and when this is not the case – propose to program the current or future models in a less risky (and probably less complex for the organisation/department) kernel.

[22] As stated, this proposal aims to increase the reporting effectiveness for low- and medium risk class models, since this is requested in the main research question. As CRO Forum (2017) suggests, Aegon can also choose to alter the reporting according to the materiality of the model.

[23] Aegon recently focused on the difference of complexity versus materiality. As the footnote in Section 3.2 in Table 1 states, high complexity – low materiality models are given a different model risk classification. They received a downgrade from the medium model risk to low. This decision does however not affect the model's complexity, even though it seems that Aegon classifies these models as 'over-complex'.

## 6.5 Further improvements for the model validation process

This section contains possible improvements regarding the independent validation process which do not require alterations to the Aegon model validation policy. Sub-question 2.3 will thus be answered: *What other recommendations can be found outside the Aegon model validation policy that could benefit the time necessary for model validations?*

### 6.5.1 Communication to model owners

Although we did not thoroughly analyse this, we experienced that the communication and model validation planning does not always go smoothly. The model validation planning is crafted in the ending months of a year for the following year. After that, the model owners are informed that a model validation is happening in a certain quarter/month. A model validation requires time and attention from model owners as they need to provide model files, documentation, and need to answer questions.

In practice, it can happen that a model owner is still caught by surprise, especially when the country unit is experiencing big workloads. Some model validators counter this by having their own process flow for validations (e.g. when to send a first email, checklist, etc.). We believe it is beneficial if a process flow, email templates, and standard documents (e.g. checklists) to have available for all model validators. The possible result is a decrease in validation idle time, together with a decrease of cancelled model validations.

**Improvement 1:** Create a standardised model validation process workflow, clearly indicating which process steps are necessary before a model validation starts and when these need to be fulfilled. Accompany this with standardised emails and checklists.

### 6.5.2 Philosophy towards data

Experiences of the data collection used for the analysis given in Chapter 4 indicate that the quality and availability of historical validation data can be improved. This does not include the information available of the performed validation, since this is stored well. The storage of validation data in the model validation planning sheets can be improved by regularly updating the days spent on a validation such that it can be seen if the start date of future validations are coming up sooner or later. Another metric that is missing often is a unique model ID. Since this is missing, the performed validations in the planning sheet are sometimes only linked to the model inventory by a name (or sometimes even a description of a model) that does not meet the notation in the model inventory.

**Improvement 2:** The planning sheets should be updated regularly (including retrospectively), and should include a unique model ID. A self-chosen model name should be discarded.

It must be said that this improvement is currently taken care of. We believe that the data collection for this research paper has accelerated this improvement. A pitfall would be that this focus on data collection decreases when this research paper is finalised. It is recommended that this should be avoided, by continuously pointing out to each other (in a friendly manner) when one detects that collection effort is lacking. More emphasis on collecting data seems to improve the model validation time, but this effect is countered when it is possible to better estimate a model's validation time or when historical data is accessible.

**Improvement 3:** Keep focus on collecting data and create a data collection culture of feedback regarding these efforts. Implement periodical checks.

### 6.5.3 Ease of validation work

The last improvement is more out-of-the-box and not very scientific. Still, the Aegon model validation team is currently using a company laptop together with a single monitor. Since the screen sizes are different, using both screens simultaneously (right next to each other) is not very user friendly. The validation work does require to open model documents, analyse models, and track model deficiencies at the same time – which indicates that a second monitor is beneficiary.

**Improvement 4:** Provide a second monitor at the model validation team workspaces.

## 6.6 Overall conclusion

The main research question of this research paper is as follows: *How can the time effectiveness of the model validation process, specifically when focusing on the medium risk models, be improved while maintaining a desired level of quality?*

To have all proposals in one place, all possible model validation process improvements are listed in Appendix D. Short term model validation process improvements, focusing on medium risk models, can be obtained by limiting the scope of model validations. We suggest to not include a validation on 'data', 'preparing for and validating model runs', or 'reporting and use of results'. Aegon must however consider that this possibly increases the model risk for individual models. However, when we realise that there currently is a backlog of models due for validation, reducing model validation time can increase the total number of models validated. We assume that this will reduce model risk exposure enterprise wide, and estimate models validated to increase by 5.5%.

On top of that, we have shown that there is room for improvement regarding the current model risk classification – specifically the 'complexity' metric. Our attempts to find an alternative for estimating the likelihood of a model error were unsuccessful, but we provide a basis for future use when more model data are available. On short term notice, we do propose to dispose the use of the metric 'complexity' for determining model risk – and solely focus on the amount of money a model represents ('materiality'). Historical data analysis namely shows that 'materiality' is a better predictor for the severity of a model validation outcome. On top of that, we listed several challenges that the use of 'complexity' for model risk classification brings – which are resolved when the metric is not used anymore for model risk classification.

Another way of reducing model validation time is adjusting the reporting template for low and medium risk model validations. Currently, the same template is used for all validations. The reporting topic 'validation work per topic' is not found essential by the Dutch regulator (De Nederlandsche Bank, 2013). For low and medium risk model validations, we propose to summarise this reporting topic, and direct this summary to the topics 'management summary' and 'validation scope, limitations, and approach'.

By our theoretical analysis, we found that mitigating model risk is achieved by reducing model complexity. Since we assume that a more complex model is more likely to contain model errors, we suggest that validation focus is directed to reducing complexity. This results in reduced model risk, and decreases model validation time needed for the next validation of that particular model (if we assume that lower complexity means that less validation effort is needed).

Further improvements regarding topics outside the Aegon model validation policy are about practical process improvements. We feel that the creation of a standardised validation workflow can increase time effectiveness. Also, regularly updating planning sheets and using unique model ID's will improve data availability. This is done by creating a "data collection culture of feedback" regarding these efforts.

**Future possibilities and potency**

When considering long term improvements, we see that the Aegon model inventory is becoming more and more complete in providing an overview of all models and their characteristics. This provides opportunities for predicting model risk, and can give insight in risk factors. We believe that tracking several continuous variables (see Section 5.4), and using these for the LME estimation (presented in Chapter 5) can improve model risk prediction.

As stated, being successful in predicting model risk also means that Aegon is able to identify risk factors. We believe that knowledge about risk factors can be used for a more strategic implementation of model validation. For instance, if using a "risky kernel" (let's say C++) is labelled as over-complex and unnecessary during a model validation – the validator can suggest to stick to a less "risky kernel" (e.g. Excel). Knowing what main risk factors are, and filtering which factors are unnecessary can give insights in new model development standards, reducing the model risk exposure of Aegon. We must however note, that such a conservative methodology can reduce innovations in model development.

The starting point is data collection, data quality, and awareness of data analysis potency. We suggest that this study is succeeded in one or two years when the model inventory is expanded, focusing on the prediction of the model validation opinion (LME). An important prerequisite is that continuous model variables are available at that time.

With this thesis, we have used input data and made estimations. We have used processing techniques to create estimates. After that, these estimates are used for producing valuable business information. By the definition mentioned in Section 1.1, we have thus created a model. We must thus realise that also this model is incorrect to some degree, especially since conclusions are based on data gathered from model validation efforts over a time span of two years. It is advised to reassess if new data study findings can be reproduced in future time when business decisions are made. Also, carefully take into consideration the limitations of the research, as well as the assumptions made.

# References


AEGON N.V. (2018). Annual Report 2017.

Aggarwal, A., Beck, M.B., Cann, M., Ford, T., Georgescu, D., Morjaria, N., Smith, A., Taylor, Y., Tsanakas, A., Witts, L., Ye, I. (2016). Model risk – daring to open up the black box. British Actuarial Journal, 21(2), 229-296.

Ashcroft, M., Austin, R., Barnes, K., MacDonald, D., Makin, S., Morgan, S., Taylor, R., Scolley, P. (2016). Expert judgement. British Actuarial Journal, 21(2), 314-363.

Black, R., Tsanakas, A., Smith, A.D., Beck, M.B., Maclugash, I.D., Grewal, J., Witts, L., Morjaria, N., Green, R.J., Lim, Z. (2018). Model risk: illuminating the black box. British Actuarial Journal, 23.

Board of Governors of the Federal Reserve System (2011). SR 11-7: guidance on model risk management.

Crespo, I., Kumar, P., Noteboom, P., Taymans, M. (2017). The evolution of model risk management. Retrieved from https://www.mckinsey.com/business-functions/risk/our-insights/the-evolution-of-model-risk-management

Chartis Research Ltd (2014). The risk enabled enterprise – model risk management. Retrieved from http://www.chartis-research.com/research/reports/the-risk-enabled-enterprise-model-risk-management

CRO Forum (2017). Leading practices in model management. Industry Paper

De Nederlandsche Bank (2013). Guidance for model validation under Solvency II. Retrieved from www.toezicht.dnb.nl/en/binaries/51-230187.pdf

Derman, E. (1996). Model Risk. RISK, 9(5), 139-145, 34-37

European Parliament (2009). Directive 2009/138/EC of the European Parliament. On the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II).

European Parliament (2013). Directive 2013/36/EU of the European Parliament and of the Council.

Fawcett, T. (2015). The Basics of Classifier Evaluation: Part 1. Retrieved 08-01-2019 from https://www.svds.com/the-basics-of-classifier-evaluation-part-1/

Foreman, J. W. (2013). Moving from Spreadsheets into R. In *Data smart: Using data science to transform information into insight*, 361-394. Hoboken, NJ: John Wiley & Sons.

Gates, D. (2019). Flawed analysis, failed oversight: How Boeing, FAA certified the suspect 737 MAX flight control system. Retrieved 03-21-2019 from https://www.seattletimes.com/business/boeing-aerospace/failed-certification-faa-missed-safety-issues-in-the-737-max-system-implicated-in-the-lion-air-crash/

Haldane, A.G., Madouros, V. (2012). The dog and the frisbee. Federal Reserve Bank of Kansas City's 36th economic policy symposium, "The Changing Policy Landscape".

HM Treasury (2013). Review of quality assurance of government analytical models. Retrieved from https://www.gov.uk/government/publications/review-of-quality-assurance-of-government-models

Jadnanansing, A. (2018). Model Change Policy v1.92. Internal AEGON N.V. report. Unpublished.

Jadnanansing, A. (2019). Model Risk Management Policy v1.0. Internal AEGON N.V. report. Unpublished.

Le, J. (2018). Decision Trees in R. Retrieved 08-01-2019 from https://datacamp.com/community/tutorials/decision-trees-R

Lloyd's (2017). Internal Model Validation. Model Validation Guidance. Retrieved from www.lloyds.com/market-resources/regulatory/solvency-ii/information-for-managing-agents/guidance-and-workshops/model-validation

Lloyd, I., (2019). 2019Q1 MVSA template. Internal AEGON N.V. file. Unpublished.

Office of the Comptroller of the Currency (2000). OCC 2000-16. OCC BULLETIN

Popper, K.R. (1959). The Logic of Scientific Discovery. New York, NY: Basic Books.

R-project (2019). A Short Introduction to the caret Package. Retrieved 08-01-2019 from https://cran.r-project.org/web/packages/caret/vignettes/caret.html

Rikkert, B. (2019). Model Validation Policy v4.0. Internal AEGON N.V. report. Unpublished.

Robinson, S. (1997). Simulation model verification and validation: increasing the users' confidence. Winter Simulation Conference (pp. 53-59).

Sargent, R.G. (2013). Verification and validation of simulation models. Journal of simulation, 7(1), 12-24.

Van de Kraats, W., Rikkert, B. (2015). Model Review Standards v4.5. Internal AEGON N.V. report. Unpublished.

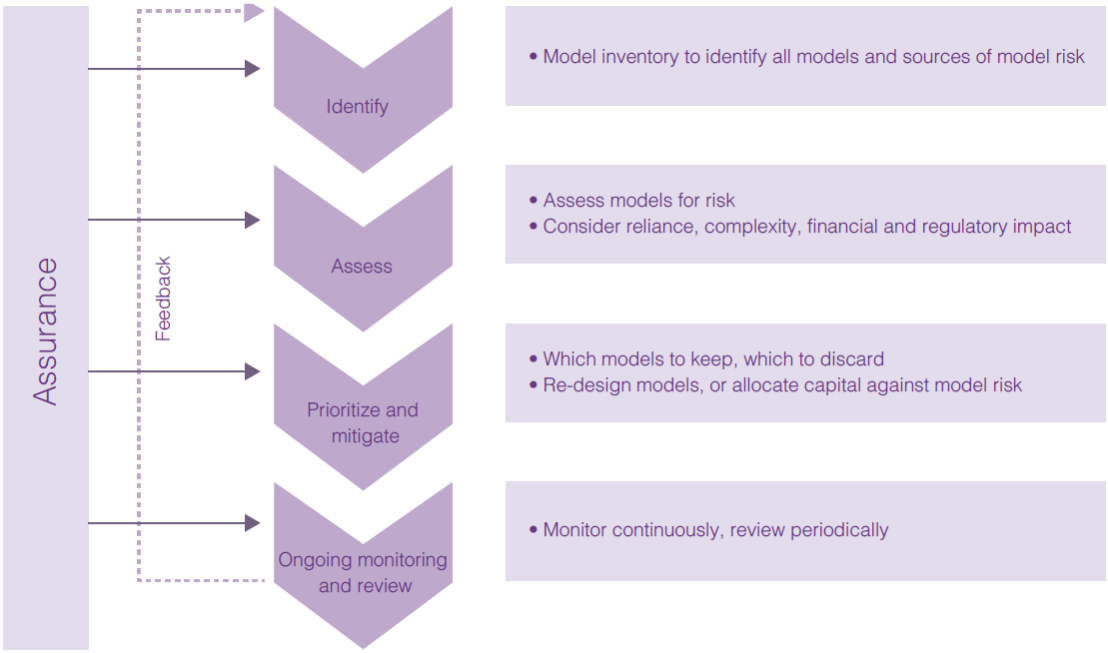
Van Roon, P., Van de Kraats, W. (2015). Model Development Standards v1.1. Internal AEGON N.V. report. Unpublished.

Whittingham, P. (2018). The Journey from Model Validation to Model Risk Management. Internal Model Industry Forum.

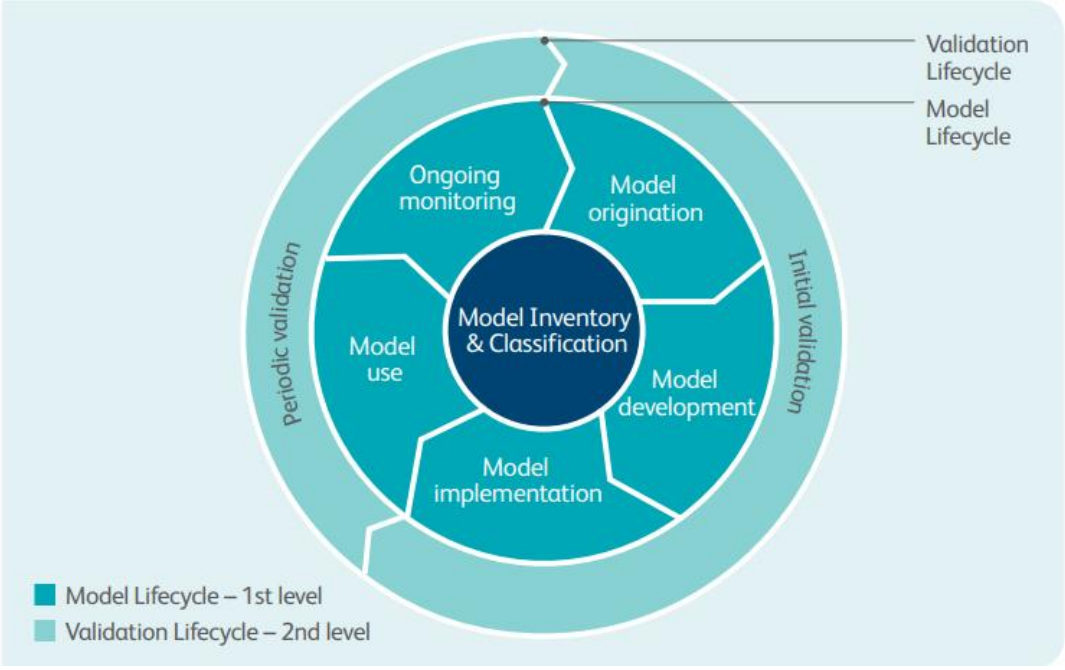
Wood, D.O. (1986). MIT model analysis program: what we have learned about policy model review. In Proceedings of the 18th conference on Winter simulation (pp. 248-252). ACM.

# Appendix A: MRM frameworks as discussed in Chapter 2

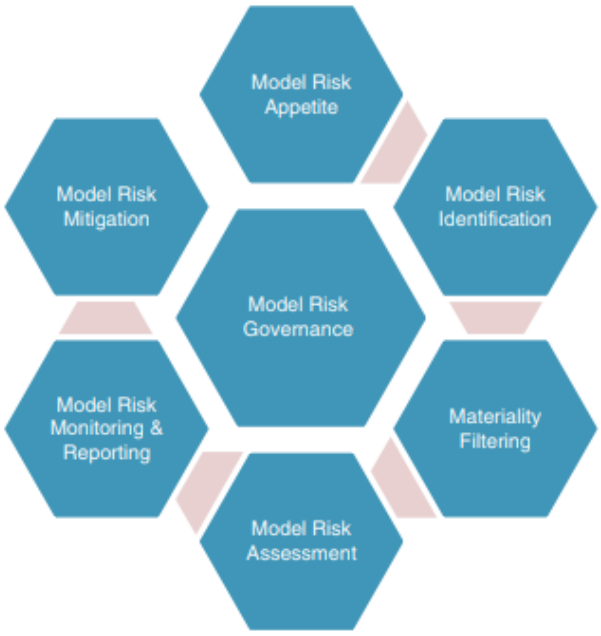
## A.1: MRM framework of Chartis Research Ltd (2014) (p. 26)




## A.2: "Ongoing MRM cycle" of Whittingham (2018) (p. 23)



## A.3: The MRM framework proposed by Aggarwal et al. (2016) (p. 240)



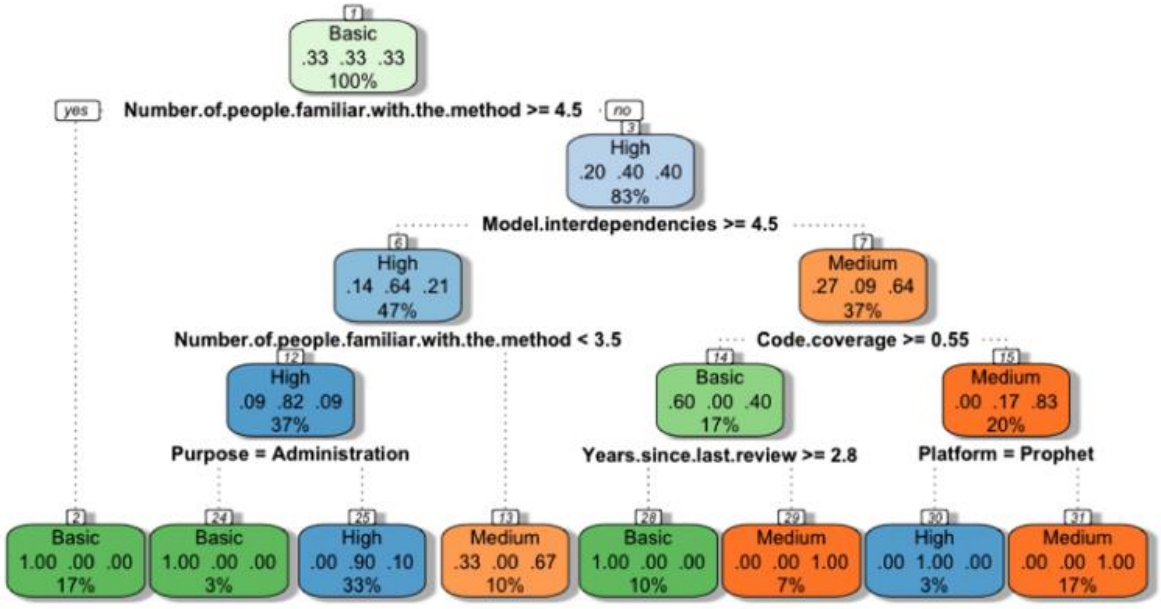

- Model Risk Appetite: Determination of the amount and types of risk an organisation is willing to take.
- Model Risk Identification: Identification of the model risk a company is exposed to.
- Materiality Filtering: Identification of the models that are material to a company as a whole.
- Model Risk Assessment: Quantitative or qualitative assessment of the model risk.
- Model Risk Monitoring & Reporting: Creating a model inventory covering a minimum requirement given by the authors and reporting "proportionate to the materiality of the model(s) and their use(s)" (Aggarwal et al., 2016, p. 250).
- Model Risk Mitigation: Determination by the governance body if the model risk is within the model risk appetite.

Aggarwal et al. (2016) list the following key features for each model to be captured in the model inventory:

- Model owner, model name, model platform.
- Storage location.
- Brief description.
- Overview of how the model works.
- Frequency of its use.
- Key assumptions and/or inputs.
- Model hierarchy and dependencies.

## A.4: Using meta data for model classification, as proposed by Black et al. (2018) (p. 22)

| Design Stage | Development Stage | Deployment Stage |
|---|---|---|
| Purpose, e.g. strategic/regulatory/forecasting/trading/ administration | Developer type, e.g. third party vendor/in-house | Number of approved/ trained users |
| Methodology, e.g. dictated by regulation/standard industry practice/adaptation of peer-reviewed method/cutting-edge | Number of developers (i.e. key man risk) | Model interdependencies |
| Number of people in the team/organisation familiar with the methodology | Platform, e.g. Excel/Prophet/R/ Python/.NET | Period since last review/ validation |
| | Automated regression testing | Number of restatements to published model result |
| | Code coverage, i.e. the degree to which the code of a model is executed | |
| | Automated version control system, e.g. Git/Mercurial | |



## A.5: Selection of validation techniques listed by Sargent (2013)

"Comparison to other models: Various results (eg, outputs) of the simulation model being validated are compared to results of other (valid) models" (Sargent, 2013, p. 16).

"Data relationship correctness: Data relationship correctness requires data to have the proper values regarding relationships that occur within a type of data, and between and among different types of data" (Sargent, 2013, p. 16).

"Degenerate tests: The degeneracy of the model's behaviour is tested by appropriate selection of values of the input and internal parameters" (Sargent, 2013, p. 16).

"Extreme condition test: The model structure and outputs should be plausible for any extreme and unlikely combination of levels of factors in the system" (Sargent, 2013, p. 16).

"Face validity: Individuals knowledgeable about the system are asked whether the model and/or its behaviour are reasonable" (Sargent, 2013, p. 16).

"Historical data validation: If historical data exist (eg, data collected on a system specifically for building and testing a model), part of the data is used to build the model and the remaining data are used to determine (test) whether the model behaves as the system does" (Sargent, 2013, p. 16).

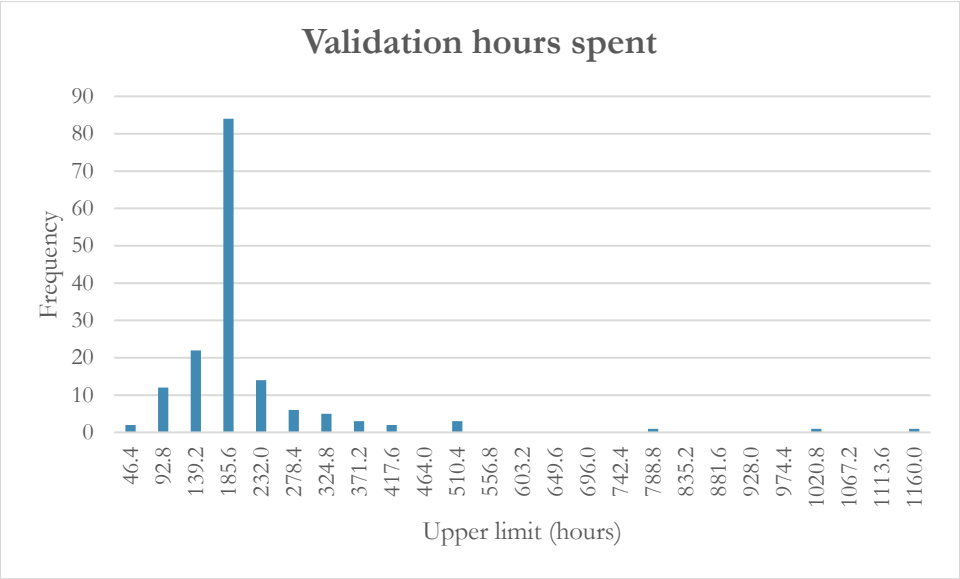
"Parameter variability-sensitivity: This technique consists of changing the values of the input and internal parameters of a model to determine the effect upon the model's behaviour or output" (Sargent, 2013, p. 17).

"Philosophy of science methods: The three philosophy of science methods are rationalism, empiricism, and positive economics. Rationalism requires a model to be logically developed (correctly) from a set of clearly stated assumptions. Empiricism requires every model assumption and outcome to be empirically validated. Positive economics requires only that the model outcomes are correct and is not concerned with a model's assumptions or structure (casual relationships or mechanisms" (Sargent, 2013, p. 17).

"Structured walkthrough: The entity under review is formally presented usually by the developer to a peer group to determine the entity's correctness" (Sargent, 2013, p. 17).

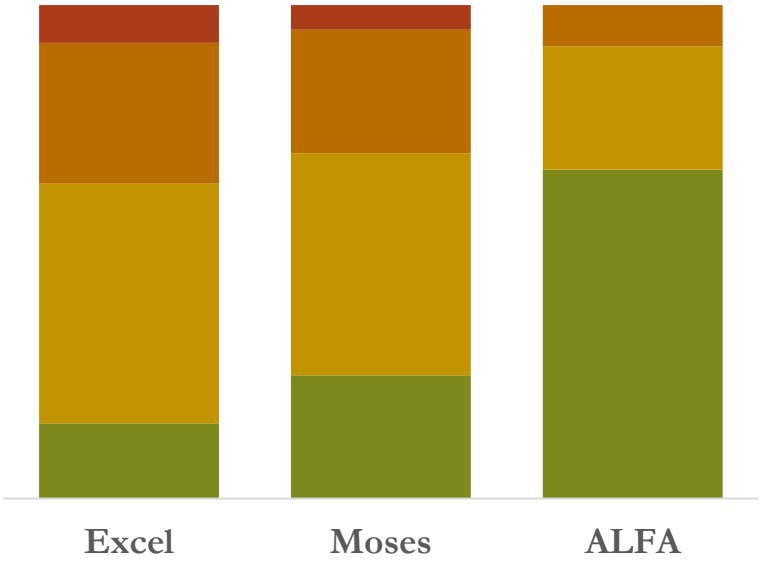
# Appendix B: Additional information and figures – Chapter 4

## B.1: Distribution hours for model validation




| *Descriptive* | |
|---|---|
| Mean | 184.5 |
| Median | 160 |
| Mode | 140 |
| Range | 1128 |
| Minimum | 32 |
| Maximum | 1160 |
| Count | 155 |

## B.2: Kernel influence on validation output



# Appendix C: LME modelling

## C.1: Input variables

Table C.1: Input variables.

| Variable | Values |
|---|---|
| LowComplexity | {0, 1} |
| MediumComplexity | {0, 1} |
| HighComplexity | {0, 1} |
| LowMateriality | {0, 1} |
| MediumMateriality | {0, 1} |
| HighMateriality | {0, 1} |
| Full | {0, 1} |
| PreviousGreen | {0, 1} |
| PreviousYellow | {0, 1} |
| PreviousAmber | {0, 1} |
| PreviousRed | {0, 1} |
| Region 1: US | {0, 1} |
| Region 2: Netherlands | {0, 1} |
| … | … |
| Region 10: SEE | {0, 1} |
| Purp_SII | {0, 1} |
| Purp_IFRS | {0, 1} |
| Purp_MVN | {0, 1} |
| Purp_Pricing | {0, 1} |
| Purp_MCVNB | {0, 1} |
| Purp_Banking | {0, 1} |
| Purp_MTP | {0, 1} |
| Kern_Excel | {0, 1} |
| Kern_Moses | {0, 1} |
| Kern_ALFA | {0, 1} |
| MO | {green, yellow, amber, red} |
| SevereMO (amber or red MO?) | {0, 1} |
| fg | $\in \mathbb{R}_{>1}$ |
| cg | $\in \mathbb{R}_{>1}$ |
| abs_findings | {0, 1, 2, …} |

## C.2: R-script

```
#clear all
rm(list=ls())

#Choose the working directory folder
setwd("X:/02. OMRM/02. Model Validation/07 - Thesis Gijsbert")
getwd() #check if previous was successful

#Load libraries - (install with command: install.packages("name")
library(caTools)
library(randomForest)
library(caret)
library(rpart.plot)

#variable s will be used for the seed value and t is the split factor
s <- 407
t <- 0.75

#Read data and split into training and testing set
#There will be two sets of each. One for predicting MO, and one for
predicting SevereMO
mydata <- read.delim("PowerBIextract.txt")
Modelopinion <- subset(mydata, MO != "")
Modelseverity <- subset(mydata, SevereMO != "")
##Set 1 for predicting MO
set.seed(s)
spl <- sample.split(Modelopinion$MO,SplitRatio = t)
TrainMO <- subset(Modelopinion, spl == TRUE)
TestMO <-subset(Modelopinion, spl == FALSE)
TrainMO$MO <- as.factor(TrainMO$MO)
TestMO$MO <- as.factor(TestMO$MO)
##Set 2 for predicting MOseverity
set.seed(s)
spl2 <- sample.split(Modelseverity$SevereMO, SplitRatio = t)
TrainSev <- subset(Modelseverity, spl2 == TRUE)
TestSev <- subset(Modelseverity, spl2 == FALSE)
TrainSev$SevereMO <- as.factor(TrainSev$SevereMO)
TestSev$SevereMO <- as.factor(TestSev$SevereMO)

#Now we use set 1, and we apply a decision tree classifier (DT) and random
forest (RF)
##Predictor DT: Model opinion (green, yellow, amber, red)
set.seed(s)
trctrl <-trainControl(method="repeatedcv",number = 15, repeats = 3)
dtree_fit <- train(MO~. -cg -fg -SevereMO -abs_findings -LowMateriality -
      MediumMateriality -HighMateriality, data=TrainMO, method="rpart",
      parms=list(split="information"), trControl=trctrl,tuneLength=10,
      na.action=na.omit)
prp(dtree_fit$finalModel, box.palette="Blues", tweak=1.2)
PredictTree <- predict(dtree_fit, newdata =TestMO, na.action=na.pass)
ConfMatrixDT <- table(TestMO$MO, PredictTree)
ConfMatrixDT
AccMoDT <- sum(diag(ConfMatrixDT))/sum(ConfMatrixDT)
```

```
##Predictor RF: Model opinion (green, yellow, amber, red)
set.seed(s)
SizeForest <-randomForest(MO~. -fg -cg -SevereMO -abs_findings -
      LowMateriality -MediumMateriality -HighMateriality,data= TrainMO,
      na.action=na.omit)
varImpPlot(SizeForest) #Variable importance -> Appendix C.3
PredictForest <-predict(SizeForest, newdata=TestMO)
ConfMatrixRF <- table(TestMO$MO, PredictForest)
ConfMatrixRF
AccMoRF <- sum(diag(ConfMatrixRF))/sum(ConfMatrixRF)

#Now we use set 2, and we again apply decision tree classifier (DT) and
random forest (RF)
##Predictor DT: Model opinion severity (amber or red model opinion?)
(1=yes)
set.seed(s)
trctrl2 <-trainControl(method="repeatedcv",number = 15, repeats = 3)
dtree_fit2 <- train(SevereMO~. -cg -fg -MO -abs_findings -LowMateriality -
      MediumMateriality -HighMateriality, data= TrainSev, method =
      "rpart", parms = list(split="information"),
      trControl=trctrl2,tuneLength=10, na.action = na.omit)
prp(dtree_fit2$finalModel,box.palette="Reds",tweak=1.2)
PredictTree2 <- predict(dtree_fit2, newdata =TestSev, na.action=na.pass)
ConfMatrixDT2 <- table(TestSev$SevereMO, PredictTree2)
ConfMatrixDT2
AccSevDT <- sum(diag(ConfMatrixDT2))/sum(ConfMatrixDT2)
RecallSevDT <- ConfMatrixDT2[2,2]/(ConfMatrixDT2[2,1]+ConfMatrixDT2[2,2])
##Predictor RF: Model opinion severity (amber or red model opinion?)
(1=yes)
set.seed(s)
SizeForest2 <-randomForest(SevereMO~. -cg -fg -MO -abs_findings -
      LowMateriality -MediumMateriality -HighMateriality,data= TrainSev,
      na.action=na.omit)
varImpPlot(SizeForest2) #Variable importance -> Appendix C.3
PredictForest2 <-predict(SizeForest2, newdata=TestSev)
ConfMatrixRF2 <- table(TestSev$SevereMO,PredictForest2)
ConfMatrixRF2
AccSevRF <- sum(diag(ConfMatrixRF2))/sum(ConfMatrixRF2)
RecallSevRF <- ConfMatrixRF2[2,2]/(ConfMatrixRF2[2,1]+ConfMatrixRF2[2,2])
```

**C.3: Output R-script for model opinion predictor (seed = 407)**

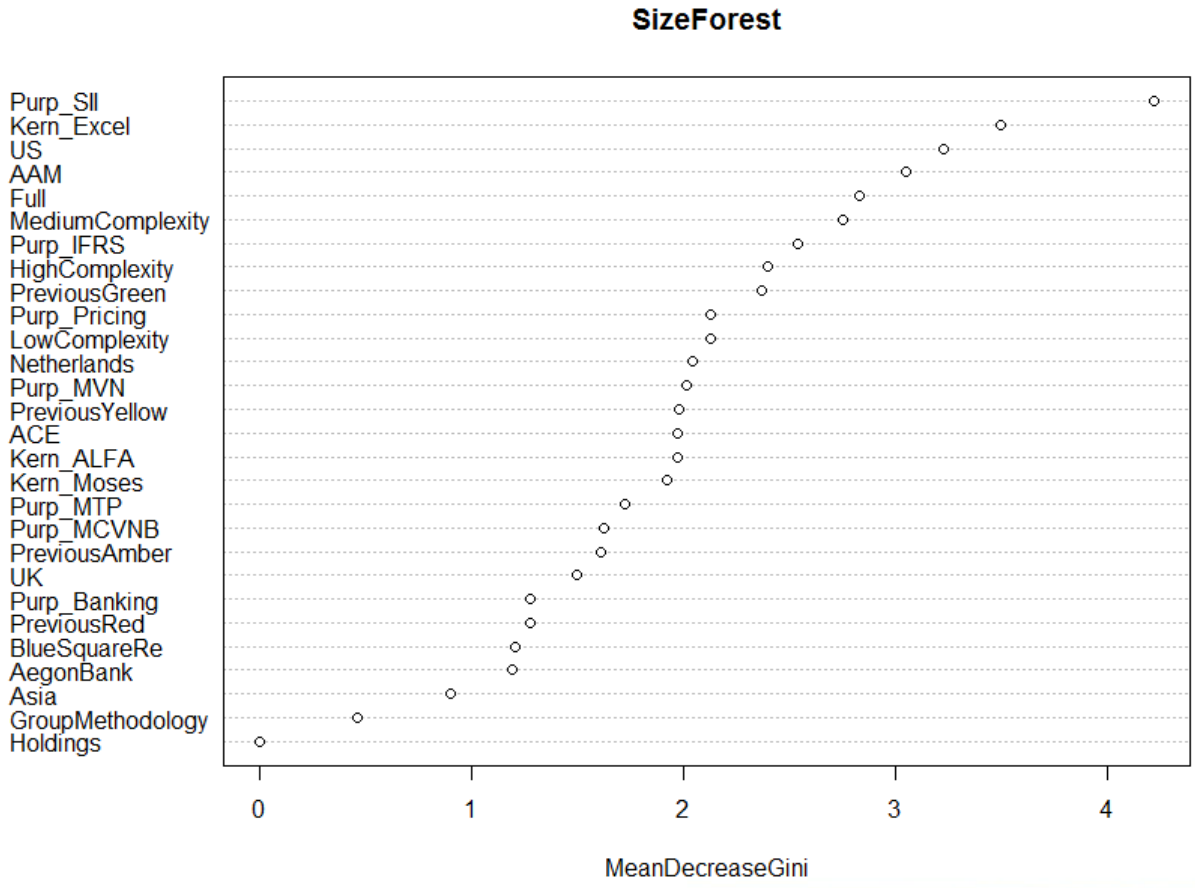



Figure C.1: Random Forest variable importance.

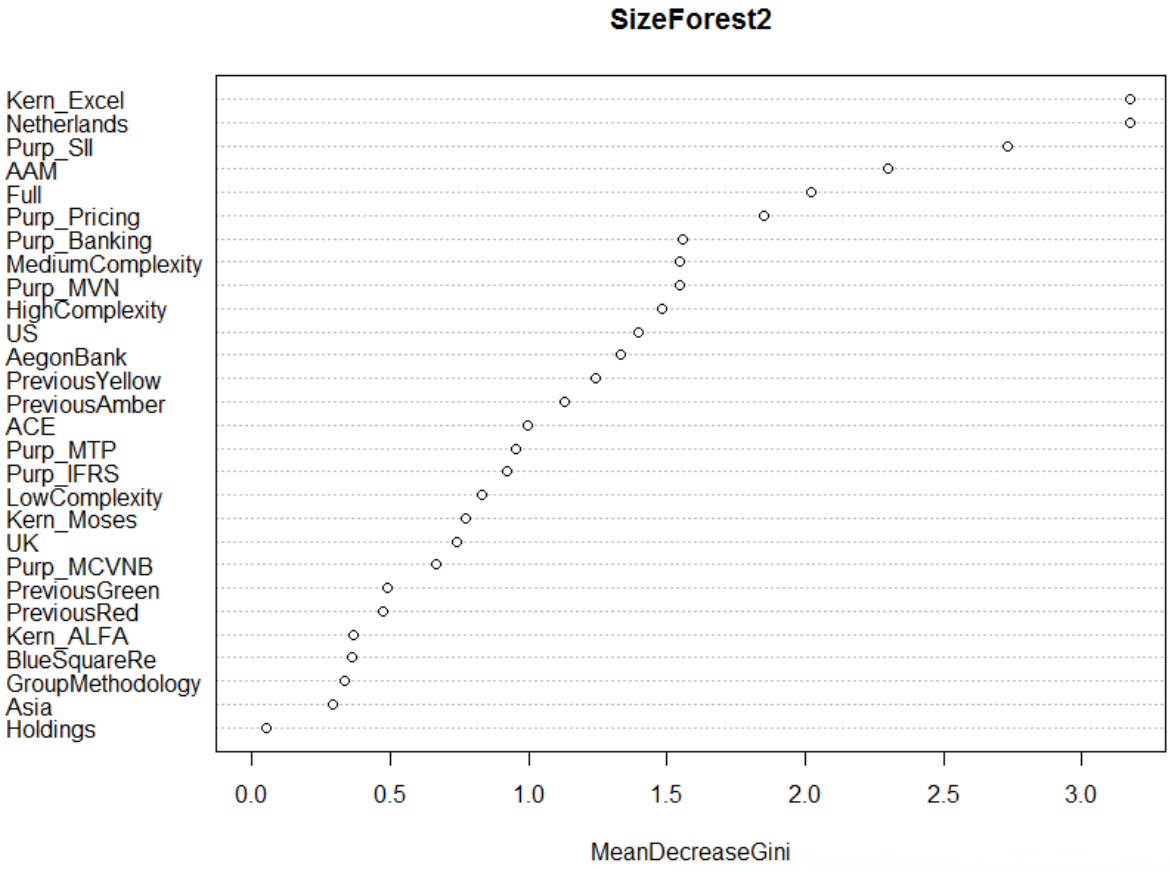



Figure C.2: Random Forest variable importance.

# Appendix D: Full list of proposals

**Section 6.1 – limiting the model validation scope.**

1. To reduce the model validation time for medium risk models, the scope should be decreased. Several model validation topics can be excluded, after analysing which topics are non-essential for the model.
2. Limiting the model validation scope can also be done in a different manner, namely by first validating the most deterministic model validation topics. If the result of these topic validations are all green, the validator can decide to stop there and finalise the validation on short term notice – redirecting efforts to the next validation on the planning.

**Section 6.2 – reconsidering the model risk classification and prioritisation.**

3. Using supervised learning for estimating the likelihood of model error, a newly crafted model variable that could replace complexity, was unsuccessful. We assume that this is (partly) caused by the absence of continuous variables. We propose to retry in one or two years with the new available variables obtained by model risk management. If the results are promising, more model metrics need to be recorded.
4. No alternative for using the metric 'complexity' is available. However, we propose that Aegon considers not using this metric for model risk classification, and fully directs model validation effort based on 'materiality' solely.

**Section 6.3 – Increasing the effectiveness of the model validation reporting template.**

5. Our proposal for increasing reporting effectiveness is focussing on reporting efforts for medium and low risk models. This is done by summarising 'validation work', and move this summary in parts to the two topics: 'executive summary' and 'validation scope, limitations, and approach'.

**Section 6.4 – Mitigating model risk by decreasing complexity.**

6. If the Aegon model validation were to keep the current model risk classification methodology: For models with a higher complexity class than materiality class, assign validation focus on a model's complexity and judge if the complexity is necessary.
7. After successful implementation of LME tool: The validator can evaluate high risk factors (as indicated by the LME tool) and validate if these are necessary and/or can be replaced by less risky characteristics.

**Section 6.5 – Further improvements.**

8. Create a standardised model validation process workflow, clearly indicating what process steps are necessary before a model validation starts and when these need to be fulfilled. Accompany this with standardised emails and checklists.
9. The planning sheets should be updated regularly (including retrospectively), and should include a unique model ID. A self-chosen model name should be discarded.
10. Keep focus on collecting data and create a data collection culture of feedback regarding these efforts. Implement periodical checks.
11. Provide a second monitor at the model validation team workspaces.