Predictive Modelling of Auto Loan Defaults within European Asset Backed Securities (ABS)

A master's thesis project in fulfillment of the degree of Master of Science in the program of IndustrialEngineering and Management with the specialization of

Financial Engineering and Management

December
22nd 2021

Author:
Derek Rodink

Supervisors
University of Twente:
Berend Roorda dr.
Reinoud Joosten

Hypoport B.V:
Ketie Feaseler

University of Twente
Drienerlolaan 5
7522 NB Enschede

Hypoport B.V.
Gustav Mahlerplein 90
1082 MA Amsterdam

UNIVERSITY OF TWENTE.

HYPOPORT

List of abbreviations

AIC	Akaike Information Criterion
AUC	Area Under Curve
ABS	Asset Backed Security
ALI	Annualized Loss Index
APS	Absolute Prepayment Speed
CDO	Collateralized Debt Obligation
CMBS	Consumer Mortgage Backed Security
CRA	Credit Rating Agency
DNB	De Nederlandsche Bank
ECB	European Central Bank
EDW	European Data Warehouse
ESMA	European Securities and Markets Authority
EU	European Union
FTE	Full Time Employee
HL	Hosmer Lemeshow Test
GDP	Gross Domestic Product
GFC	Global Financial Crisis
ND	No Data
PoP	Priority of Payments
PRoMMiSe	Portfolio Risk and Mortgage Management System
ROC	Receiver Operating Characteristic
RMBS	Residential Mortgage Backed Securities
RWA	Risk Weighted Assets
SME	Small- and medium sized enterprises
STS	Simple, Transparent and Standardized
SPV	Special Purpose Vehicle

Preface

Enschede, December 22nd 2021

It has been a long and bumpy ride. In total, over two years have passed since I started the internship at Hypoport as conclusion for my master's program of Industrial Engineering and Management at the University of Twente. In February 2020 I started full time working for Hypoport, where this thesis was continued on a part-time basis.

The past two years have made quite an impact on my personal life as well. Countless times have I travelled from Enschede to Hypoport in Amsterdam in the early morning to arrive home in the late evening. The first nine months with a full-time focus on this research project, and then on a working base regarding day-to-day activities within the organization primarily focused on lease clients. During this time, I have had the opportunity to meet and learn frommany great people who I can now call my colleagues after being employed at Hypoport for two years.

From the start there have been some struggles with the subject and analysis part of this topic. During this journey there were times where I would have liked to throw in the towel, but the persistent support of friends and family have helped me through. A special thanks goes to everyone who has helped me along the way. I would like to thank all supervisorsinvolved in this project, especially Ketie and Berend for their great ideas and feedback throughout the project. Also, Reinoud for his valuable input during the final phase of this research.

A very special thanks goes out to my father-in-law Ge, who sadly passed away during the pandemic due to COVID and always believed in me. Lastly, I would like to thank my girlfriend Marcella who missed out on a lot of time together due to me working on this project on the side but always encouraged me to continue.

Derek

Table of contents

1. Thesis introduction

1.1 Transparency is key

The securitisation market will always be remembered for its role in the Global Financial Crisis (GFC) starting in 2007 (Shin, 2009). Securitisation is the packaging of assets into a financial product, such as a mortgage-backed security (MBS) or collateralized debt obligation (CDO). This can be done for several types of underlying collateral if it has a steady and predictable cash flow. Examples of these are credit card loans, mortgage payments and auto loans. Banks who issued the loans were no longer at risk in case of defaults since the default risk could be transferred to the investors. This resulted in a major drop in lending standards and therefore to an increase in sub-prime mortgages. These mortgages where paying higher interest rates and therefore investors would receive higher interest rates as well. This all went well, until the housing bubble in the United States collapsed. A lot of transactions and tranches with high credit score ratings turned out to be way more risky than their credit ratings suggested, resulting in major losses throughout the market.

The aftermath of this financial crisis had a high impact on the structured finance market. It required a transformation from all market participants. There was a strong need to improve transparency in the market. Transparency is necessary for a liquid market because investors need to know what they are investing in and must be able to perform their due diligence accordingly. Solely relying on the ratings of credit rating agency's is not enough, and investors need to do their own duediligence before investing in securities of any sort. As often is the case, financial innovations are driven by regulation (Calomiris, 2009), and the two main examples of this in this market are the ECB- and STS tapes.

The European Central Bank (ECB) started with mandatory requirements for loan-level data reporting of Asset Backed Securities in 2012. In today's market it is unheard of to think that less than a decade ago it was not required to report on the level of an individual contract. The ECB introduced a standardized way of loan level reporting of deals that can be accessed through the centralized European Data Warehouse (EDW) portal in 2012. Depending on the type of transaction issued, there is a template which specifies which fields are required to be reported (mandatory) or are optional. The template consists of fields related to assets, security/bond level information, contact level information and tranche level information.

In more recent years new regulation came into place, the so-called STS notifications, that is used for Simple, Transparent and Standardizes transactions. For a traditional securitisation to be considered STS, it must fulfil the STS requirements. This new legislation came into effect as of the 23rd of September 2020. With increased transparency in the securitisation market the goal is that investors in these types of bonds can do a better due diligence and analysis on the transactions they are investing in. This should enhance the issuance in Asset Backed Securities, which has been on the decline ever since the global financial crisis in terms of volume issued in these products. Until now, the impact of this on the volume issued in transactions is marginal. The question rises if the market is ever getting close to pre GFC standards.

1.2 What is securitisation?

Securitisation is the pooling of various types of contractual debt and selling their related cash flows to third party investors as bonds. By doing so, the risk of loan defaults is spread from the issuer to investors. Selling a single contract with the cash flows of a mortgage with the property as the underlying collateral is difficult. However, if there are a lot of these types of loans with a stable and predictable incoming cash flow of loans with different characteristics in a pool they can be grouped together and sold toinvestors.

When a financial institution creates securities backed by the cash flows from those assets it is said to have a securitized pool of assets. A pool of loans may be transformed in asenior and junior tranche. The senior tranche is first in line to receive the cash flows and the junior tranchethe last in line. The bonds can be

structured to the needs of the investor, depending on the risk appetite. Structuring the tranches is one of the essential parts for the issuer since it must attract the interest of different types of institutional investors.

Generally, the assets generate two types of cash flow for the investor: revenue (interest income) which goes to the revenue account/ledger and redemption (principal repayment) which goes to a principal account or ledger. From these accounts the funds will be paid out in accordance with the Priority of Payments (Popp), which is specified in the transaction documentation.

On each note payment date, the cash flows are paid out to the investors who are the owners of the bond. The order of payments sequence is specified in legal documentation. First, directors and counterparties in the transaction are paid. Any cash that is left is used to pay Class A investors. If there is any shortfall from the previous period, this is covered next. Then, this process is repeated for Class B and potentially the other remaining classes. Finally, if any cash is left this is the profit for the seller which is the so-called Deferred Purchase Price. The backbone of the transaction is the waterfall structure. The higher an investor is in the waterfall, the more likely the investor is to get paid. Losses are automatically absorbed by the lowest tranche giving investors in a more senior tranche protection (but also a lower interest rate). In a lot of transactions, the issuer is obliged to have an economic interest of 5% in their own transaction. Also, structures exist in which there is a subordinatedloan present in the transaction as a type over overcollateralization. This is the second type of protection mechanism.

1.3 Why securitisation?

There can be several motives for securitisation. By separating out receivables from its other assets, the originator may be able to reduce its cost of funding. Also, borrowing costs can be reduced by substituting one credit risk for another one, this may lead to credit arbitrage.

Secondly, financial institutions are required to hold a specific minimum amount of capital, depending on their risk-weighted assets (RWA). By removing assets from their balance sheets, securitisation reduces the amount of regulatory capital that required to hold. It also provides an alternative source of funding that may allow the originator to access to a new or more diversified pool of investors. Securitisation also provides liquidity to the market, by structuring a product that attracts the risk appetite of investors.

1.4 Special Purpose Vehicle (SPV)

When structured properly, the pool of financial assets can achieve higher ratings than the originating company. The main reason for this is that the assets are being sold to a bankruptcy remote Special Purpose Vehicle (SPV) (Na'im, 2006). Investors are thereby not investing in the company originating the securities, but in the underlying pool of assets. Therefore, higher ratings can be achieved than investing in the company of the originator. Credit Rating Agencies play an important role in the structuring of a transaction.

Loans of varying amounts, loan types and geographic areas are pooled together and sold to the SPV. The SPV has the sole purpose of purchasing assets and paying for them by issuing Asset-Backed Securities (ABS). The assets in the issuing trust generate enough cash flow to pay principal and interest to bondholders

and to pay a servicing fee to the seller/servicer. The payment stream to the bondholders is funded by the payment of principal and interest on the underlying pool of consumer loans. The creator of a securitisation transaction expects profit because the weighted average return on the assets in the underlying portfolio is greater than the weighted average return offered on the tranches.

1.5 The role of securitisation in the Global Financial Crisis (GFC)

Securitisation played a part in the creation of the housing bubble in the United States which led to the global financial crisis starting in 2007. Since mortgage originators knew that the mortgages would be securitized, there was evidence found from sub-prime loans that this led to lax screening of customers applying for a mortgage. Because banks relaxed their standards and the credit quality of the instruments being originated declined sharply, this led to a severe credit crisis. The GFC exposed investors who had based their due diligence mainly on the ratings that were given by the credit ratings agencies (CRAs) and faith in an ever-increasing value of real estate prices. During the GFC prices declined, illiquidity in the market was a result and there were a lot of rating downgrades for these transactions.

As a result, investors had lost confidence in the securities that had been created, and the liquidity in the market for this type of products dried up. The products related to this were mainly RMBS, Consumer Mortgage-Backed Securities (CMBS), and mortgage-backed Collateralized Debt Obligations (CDOs).

1.6 Securitisation market post GFC

After the GFC the securitisation market recovered, but never to the levels of before the crisis (Blommestein, 2011). Up and till 2010 many investors had lost appetite for investing in these products. Although the returns on the products post the GFC were good, many investors were hesitant to invest in these types of products.

Table 1 shows the issuance of European Historical Issuance post GFC. Although the market started to recover in 2014, the current levels in issuanceare still below the crisis levels. In terms of issuance volume, the market hit the bottom in 2013. Then the market for securitisation recovered in the next years.

Table 1: Issuance of European Historical issuance in billion €

Year	Q1	Q2	Q3	Q4	Total
2009	131.0	83.8	113.3	95.8	423.9
2010	75.5	32.6	110.7	159.2	378.0
2011	115.2	67.3	57.1	137.2	376.8
2012	64.3	67.7	62.0	63.9	257.8
2013	32.8	53.2	38.4	56.4	180.8
2014	20.0	99.5	37.8	59.8	217.1
2015	35.7	50.3	57.8	72.8	216.6
2016	57.0	75.8	46.6	60.1	239.6
2017	40.2	73.0	49.1	74.1	236.5
2018	58.5	68.1	54.5	88.4	269.4

Different types of collateral can be securitized. Table 2 shows the different types of collateral per quarter in 2018 issued in Europe. It illustrates that (RMBS) have the largest market share, followed by ABS and Collateralized Debt Obligations (CDO). Within the category of ABS there are several sub classes. These can be categorized in the following categories: auto loans, credit cards, equipment leases, student loans and others. Even cash flow streams such as royalties from future music revenues can be securitized, as illustrated by the famous Bowie Bonds in 1997.

Table 2: European issuance by collateral. Source AMFE (2019). Note a = asset-backed securities, b= collateralized debt obligation/collateralized loan obligation, c= commercial mortgage-backed securities, d=residential-mortgage backed securities, e= small and medium sized enterprises, f = whole business securities/project finance initiatives.

Collateral	2018: Q1	2018: Q2	2018: Q3	2018: Q4	Total
ABS_a	13.0	18.5	9.1	28.1	68.7
CDO/CLO_b	12.6	15.2	14.2	9.5	51.6
$CMBS_c$	0.4	2.4	1.1	1.85	5.8
$RMBS_d$	29.3	29.5	28.2	26.3	113.3
SME_e	3.1	2.5	1.9	22.0	29.5
WBS/PFI_f				0.6	0.6
Total	58.5	68.1	54.5	88.4	269.4

1.7 Asset class: Auto loans/leases

This research focus is on the market of Auto Loans and Leases. In these loans the vehicle itself is the underlying collateral. In this type of transactions, a pool cut is created from a sub-set of the total amount of assets available (on the back book or origination pool). The future payments of interest and principal on the loans are used to give out notes to investors. The more risk is involved with the tranche, the higher the interest rates will become.

The goal is to find out whether the data from the European Datawarehouse (EDW) can be used to analyse the default risk of in European Asset Backed Securities (ABS) transactions. Within this context it is tested if the increased transparency due to regulation has positively impacted the predictability of Auto Loan defaults. During the GFC the delinquency rates and default rates on the loans skyrocketed. Therefore, it would be interesting to determine using the ECB template to see whether predicting loan defaults using the ECB template is possible and if it gives new meaningful insights.

Auto loans were some of the earliest loans to be securitised in the ABS market and still form a large part of the ABS market. There are several reasons why investors are attracted to Auto ABS. The main reason being is that the underlying asset of the vehicle is often easy to sell. It is a tangible asset in the case the obligor of the loan defaults since the underlying exposure can be sold in the secondary market. Cars are often essential purchases and have a relatively short loan exposure (typically between three and five years). The short duration of the loan is a disincentive to refinance, therefore there is no real prepayment structure, which is better for the duration of the transaction. Prepayments can heavily impact the duration of the loan and therefore also the duration of a transaction. The prepayment speed for Auto loans is known to be quite unaffected by interest rates. This is different from the RMBS market, where refinancing is much more common due to changes in interest rates. A typical down payment of the loan comes from both cash and trade-ins of former cars. Typically, they are in between 10-20% of the vehicle's value.

Auto ABS securitisations are structured similarly to other ABS transactions: the sponsor creates a SPV to isolate the auto loans from its bankruptcy or insolvency. The assets are isolated from the seller-servicer to achieve bankruptcy-remote status. The two most widely used structures for auto ABS are grantor trust and owner trust. Depending on the structure, multiple classes of securities, fixed- or floating rate bonds, varying amortising schedules, maturities and ratings can be tailored to meet investor appetite.

There are major differences between the industry environment of Auto ABS in Europe and the United States. The main difference is the type of customers being targeted. In Europe there are mainly prime borrowers, while in the U.S. there is a mix between prime and subprime borrowers. Prime borrowers are considered borrowers with a below average credit risk,whereas subprime borrowers are borrowers that are considered to have a relatively large credit risk. Credit risk is commonly defined as the risk of default on a

debt that may arise from the borrower failing to make the required payments over a certain amount of time.

The European transactions typically have simpler deal structures with less tranches. Both markets securitise both new and used vehicles. The distinction between the new and used cars is relevant since the residual value (value of the car after the contract has terminated) is often included in the securitized value of a transaction. New cards tend to have a higher residual value than used cars. The Loan-to-Value (ratio of the loan amount and the value of the vehicle) of the assets are typically higher in the U.S compared to Europe.

1.8 Historical performance of Auto ABS in Europe

Figure 1 shows the historical performance of Auto ABS based on the Annualised Loss Index (ALI) and delinquency indices based on figures from credit rating agency Fitch. It is important to note that in Auto ABS transactions a loss occurs when a lessee has not made its payments timely for a period of more than 90 days. This corresponds with the lessee failing to meet 3 consecutive payments. Therefore, the definition of default is a contract being in arrears for a period over 90 days.

Figure 1 shows that the losses are relatively small for European Auto ABS. Although the GFC resulted in a peak of the losses between 2008 and 2010. The ALI index in Europe peaked to 1% in 2010, but then steadily declined to former levels.

Figure 1: Annualised loss index of European Auto ABS. Source: Fitch.

Besides the annualised loss index (ALI), there are two other metrics that are relevant in the historical performance; these are the delinquency rates of 30+ days and 60+ days. The reason why these numbers of arrears are important, are because they are well known indicators of default. There is a strong correlation between these metrics. If the 30+ delinquency rates and 60+ delinquency rates are rising, it is likely that the default rates will also increase. The reasoning behind this that if an obligor has not made timely payments for a period of 60 days, it is more likely he will also not make the next payment than that the previous and current arrears will be paid in this month. The graphs in Figure 2 illustrate this relationship.

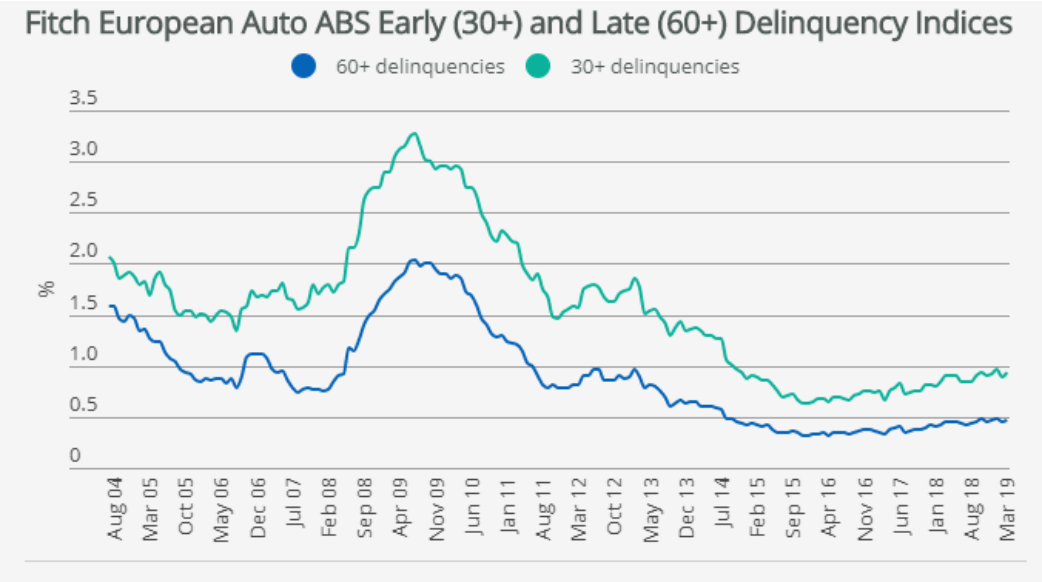

Figure 2: Early (30+) and late (60+) delinquency indices over time. Source Fitch 2019.

1.9 Company introduction

The master thesis assignment was conducted at the company of Hypoport B.V. The company is a subsidiary of Hypoport SE, which is the holding of a network of technology companies for the credit, real estate, and insurance industries. Hypoport is a financial technology firm specialized in providing software and consultancy services to the financial industry. The main business of Hypoport is related to PRoMMiSe, a software platform that is used by clients to structure and administrate financial transactions of clients. PRoMMiSe is an abbreviation of Portfolio Risk and Mortgage Management System. This system is used by their clients (banks, pension funds, asset managers, insurance companies, real estate finance companies, trust companies and car lease companies) to manage structured finance and fixed income products.

PRoMMiSe allows clients to manage their portfolios on both the assets and liabilities. The software is tailor made for each asset class. Mainly the underlying collateral of the loans are mortgages (Residential Mortgage-Backed Securities – RMBS), but also other asset classes such as corporate loans, auto leases and loans to the small and medium sized enterprises (SME).

PRoMMiSe enables clients to monitor and optimizetheir asset pools, extract investor reports, and perform discounting calculations on both the assets and the liabilities. The company supplies an administrative solution for every part of the structured finance process. The software enables day-to-day operations regarding valuation, risk management, administration, compliance, data analysis and reporting on multiple asset classes. Every month a total amount 1000 billion Euro's is processed by PRoMMiSe. Hereby Hypoport is the main master servicing provider in the Benelux.

2. Alternative methods

Throughout this research the method of logistic regression was performed on several datasets for predicting loan defaults on auto loan clients. Based on a set of features (variables) we try to identify the probability of a customer being a defaulted customer or performing customer. These two types of classes are being distinguished with the method of logistic regression. However, there are also noteworthy alternative algorithms that could have been investigated. In fact, each algorithm that is being able to predict the outcome of different classes could have been used as an alternative. However, the main categories of algorithms can be divided in the following four classes: tree-based methods, traditional statistical methods, neural networks and support vector machines and k-nearest neighbors. There is no single alternative being superior. Choosing a method depends on the problem at hand and the skills and familiarity with the other techniques as well. In this section other methods besides logistic regression are introduced including the advantages and disadvantages of these methods as well.

The first category of methods is tree-based methods. These include decision trees, random forests, and extreme random forests. The goal of using a decision tree is to train a model in such a way that it can predict the class or value of the target variable by learning simple decision rules inferred from the training data (Myles, 2004). Suppose we have a dataset with the variables outlook, humidity, windy and would like to predict if we aregoing to play golf that day. Based on a dataset with twenty-two observations the following decision tree can be constructed. Apparently, the node starts with the variable outlook and branches to the other variables. The algorithm is using the Gini coefficient to determine which variable to start. This is a cost function, which is being minimized. This approach is often named a greedy approach since it is solving the heuristic problem by making an optimal decision at each node. Choosing the optimum decision at each node gives the approximate solution on a global level. In this example the cost function of thevariable Outlook is Lower than the other variables.

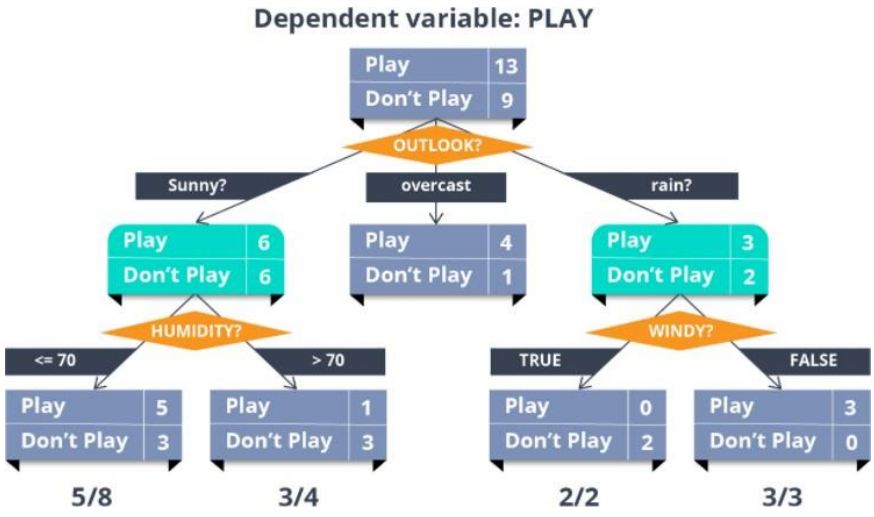

Figure 3: Decision tree to play golf or not to play golf.

The main benefit of decision trees is that they generate an understandable output. Contrary to logistic regression it is easy to see the results of the model and it is easy to visualize the results. Decision trees don't require much computation and are therefore quite fast in producing output. They can handle both categorical data and continuous data and they provide a clear indication of which variables are most important for the prediction of the target variable. More important variables are higher up in the decision tree.

The main disadvantages of decision trees are that they are prone to overfitting. In a fully worked out decision tree the model might overfit the data. Another disadvantage is that it may lead to nodes with very few final observations in the final model. To overcome this issue, it is required to either prune the tree or move to random forests for analysis. When pruning the tree, several sections of the decision tree that are non-critical to classify instances are removed. Pruning reduces the complexity of the model, which generally leads to improving the accuracy of the model. However, it is difficult to tell the algorithm when to stop.

Random forests consist of many individual decision trees that operate as an ensemble. Each tree gives out a class prediction and the class with the most votes become the model's prediction. The forest is random, to have a low correlation between the models (Pal, 2005). Small changes to the training result can result in significantly different tree structures. Randomforests take advantage of this by allowing each individual tree to randomly sample from the dataset with replacement. In a regular decision tree, we consider every possible feature and pick the feature that producesthe most separation between the observation in the nodes. In a random forest each tree can only pick froma random subset of features. This enhances the variation in the decision trees, but overall lowers the correlation between the decision trees and increases diversification. The result are decision trees that are trained on different subsets of the data, but also use different features to make decisions.

The K-nearest neighbors (KNN) algorithm is an interesting method for classification problems in which a new data point is being classified based on it is surrounding (similarly) data points also known as their neighbors (Peterson, 2009). The value of K is the number of neighbors closest to the data point where the classification is being based on. If K = 1, then the object is simply assigned to the class of that single nearest neighbor. The optimal value of K can be optimized using trial and error. A general rule, the starting point for K is to take the square root of the number of datapoints in the test data set and go from there.

The algorithm works by finding the Euclean distance between two points in a dimension. Within this method it is important to normalise the features first, to create a level playing field for all variables that are of importance. Normalizing the data between 0 and 1 can be done using the following function: (x-min(x))/max(x) – min (x)). The KNN algorithm works well with numerical variables, however it requires more effort when combining categorical with numeric variables. Therefore, it would not be suited optimally for the problem at hand.

Neural Networks are a set of algorithms that are designed to recognize patterns, modeled after the human brain. They interpret data through labelling or clustering the raw input. They help to group unlabeled data according to similarities on the inputs and classify data when they have a labeled dataset to train on (Gurney, 2018). Like decision trees and random forests, neural networks can be stacked, creating a deep learning network. Deep learning is the name for stacked neural networks that consist of several layers. Each layer is made of nodes, this is a place where a calculation happens, patterned to the neuron inthe human brain does when it encounters stimuli. All a note does is combine input from the data inputs witha set of weights or coefficients, assigning significance to the inputs. The algorithm is trying to learn whichinput is most helpful in classifying the data without error. The sum of the input and weight products are summed and passed through an activation function. This function determines to what extent the signal should progress through the network. The figure below gives a diagram of how the process in a neuron works. A layer of nodes is simply a row of these neurons that switch on or off as the input is being put through the network. Figure 4 shows the representation of a neural diagram.

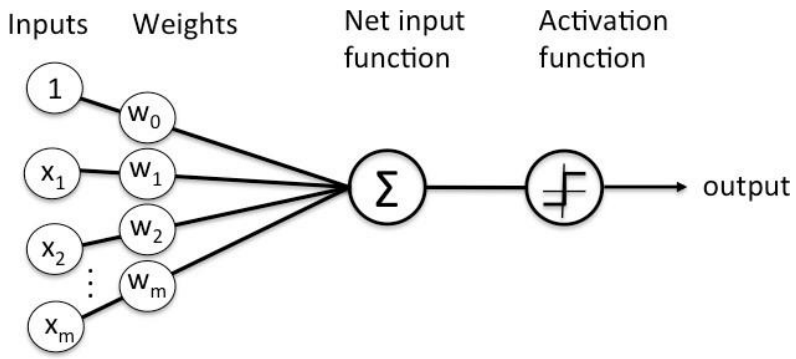

Figure 4: Representation of neural diagram

3. Research questions

This research focus is on transactions of Asset-Backed Securities with Auto Loans and Leases being the underlying exposure type. The goal is to find out if the data from the EDW can be used to analyze the default risk of in European ABS transactions. The ECB introduced a standardized way of loan level reporting of deals that can be accessed through the centralized EDW portal since 2012. The main research question is to answer the following question:

What are the determinants for auto loan defaults in Auto ABS transactions based on the framework set by the ECB in the context of the regulatory requirements?

To answer this research question several sub-questions have been formulated:

1. *What data are required to be reported by the issuers of Auto ABS on a loan level and how to prepare a representative data set?* This question will be answered after a study to the regulatory requirements from the ECB. Hypoport has granted access to a database of the European Data Warehouse, that has monthly information on all Auto ABS transactions registered. The answer to the first question gives an indication of the number of loan level fields (parameters) that are required to be reported on a loan level basis.

2. *What fields are relevant in the determination of default?* When modelling a specific transaction to identify the key indicators of default, it is important to analyse the independent variables first. Therefore, it is important to do a first analysis on the data fields present in the ECB tape. Fields that have no added value can be deleted from the list. A literature study to the topic of loan default parameters gives direction to which variables to include in the a priori analysis. The answer to this research question will yield a breakdown of the template to a smaller sub-set of variables.

3. *What transactions are selected for further analysis?* The database is enormous and contains so much information regarding the ABS transactions for Auto loans, making it not feasible to analyse each individual transaction. Based on objective criteria a data set from a transaction will be included or excluded to the study. The starting data set has 223 transactions in 11 different European countries. However, not all transactions have the same data quality. In some transactions some fields might be not reported (ND1-ND6 codes) that are marked potentially impactful on the probability of default from research sub-question 2. In other transactions, the number of defaults present, are exceptionally low which creates a major imbalance problem of the data set. After this research question is answered we have created a scope for ourselves. The research is limited to the variables that are of interest and the transactions that are of interest.

4. *How to model the default determinants within a transaction?* In the current era of advanced statistics, machine learning and artificial intelligence there is a sprawl of analysis methods and programming languages to choose from. Given the specific problem and the size of the dataset, some methods or languages can be preferred over others. The answer to this research question provides direction to the analysis part of this thesis, which is to model the ABS transactions in such a way that we get an understanding of the contributing factors of Auto Loan defaults are in European ABS transactions.

5. *What modelling approach is desired?* Based on the results from sub-question 4 the decision was made to use the software package R for data analysis and logistic regression as the method for analysis. The reason for logistic regression is rather simple. Instead of for simple linear regression, the outcome variable is dichotomous. Either is contract is in arrears or a contract is not in arrears. Within logistic regression there are two paths to choose: forward selection and backward selection.

In backward selection the initial model is going to select all independent variables. Step by step the variable is removed that contributing the least to the overall performance of the model, until the point is reached where it is not beneficial anymore to remove more variables. This is the final solution to the problem at hand, and the variables remaining in the model are the ones impacting the defaults. Similarly, forward selection is using one variable as a starting point for the model. It is first going to try to predict defaults on each single variable. The variable that scores the best, will be used as a starting point. A univariate analysis is performed before this step, to only include variables that score higher than 25%. Both methods do not have to yield the same results. The indicator that will be used predominantly to determine if a model with independent variables (X,Y,Z) is "better" than the model with parameters (X,Y) is the Akaike Information Criterion (AIC).

6. *How well can we predict defaults from the logistic regression model?* The answer to this question is twofold since it is based on several factors. First, the quality of the data is relevant. Without good data quality it much harder to create an accurate model. Also, using a lot of independent variables can lead to overfitting the model, which is a well-known undesired phenomenon. To determine how well the model performs, we are going to use training data set (70% of the loans are randomly selected in this set) and a testing data set (the remaining 30% of the data is usedfor testing purposes). The model is fit on the testing data set using the results from the forward andbackward logistic regression. In logistic regression there are well known performance metrics thatindicate how good the model is. Since we know what the status of the loan is (defaulted or not defaulted), we can compare this to the predictions the model made on the testing and later the training data set. Comparing these data is possible through a contingency table. Based on this table a lot of metrics such as the accuracy, recall, precision, and sensitivity can be calculated. Other metrics that visualize the performance of the model are the ROC graph (Receiver Operating Characteristic) and the AUC metric (Area Under Curve).

7. *Do the results of different transactions align with each other and with previous studies performed?* This question sounds easy to answer, however this is not the case. There have been no scientific papers written on the Auto loan sector using the data from the EDW. However, the independent variables that are influential in one transaction might not be significant in another transaction. Also, there might be contradictory results; in one transaction the car being 'Used' might be an indicator of default, whereas in the next transaction a 'New' car has a positive impact on the probability of default.

8. *What do the results of the modelling exercise imply for regarding the increased transparency in the market?* Based on the results of the models, a conclusion can be made on what the default determinants within a transaction are. The results might also indicate that it is not possible to predict defaults from a large data set using the method applied. In this case, recommendations can be made regarding improvement of the ECB template being used.

4.Methodology

The methodology of this research consists of several steps. A literature study on the topic of Auto Loan defaults is performed. Although there is no real data gathering required, the data that are available is required to be studied. Hypoport has a license fee which grants access to the database of the EDW containing information from all types of Auto loan transactions. In Europe, there are 223 transactions from 11 European countries in the database. A multi criteria analysis is performed to determine which contracts are of interest in the analysis.

After we know which transactions to study, an analysis is performed regarding the importance of the independent variables in the dataset. Not all fields are relevant, because some contain dummy data, such as typical identifiers for example. For logistic regression, it is required that the variables are independent of each other (Wright, 1995). If some variables have a high correlation, it is required to remove a variable from the analysis. Some transactions might have missing data, since not every field is mandatory to be filled. If there are too manymissing values, this could be a potential problem for the analysis.

After the transactions are selected and it is known which variables are included in this study, a hypothesis is formed. For some variables this is trivial. When just looking at the primary income of the obligor, a lower income is associated with a higher probability of default. However, for other variablesthis relationship is not always clear. Are obligors more likely to default on a new car or on a used car? Is the size of an upfront down payment for the loan an indicator for default or not? Is data cleansing required?Are there independent variables which need to be altered?

There are multiple ways of doing these kinds of analysis. However, what is the best method for the problem at hand? What kind of language/application is required to perform the analysis? If this question is answered, there are also some modelling decisions that are required to be answered. What is the split between the training and the test population? How are we building are model with significant variables that predict defaults of auto loans? How do we determine to remove or keep a variable in our model? How good is the model in predicting defaults, and how is this measured? How can we avoid overfitting of the model? What variables are significant, and can we observe a pattern over multiple transaction? What variables have a positive impact on the probability of default and which variables have a negative impact on the probability of default? Are the results over multiple transactions consistent with each other?

5.Theoretical background

It is required to have a basic understandingof how what logistic regression models are, how to assess the performance of the model and their statisticalvalue and the practical value in doing so.

Logistic regression is one of the most standard models available. The foundation of the method to estimate theprobability of default has been laid oud over half a century ago. In the article he studies events with binary outcome dependent on multiple independent variables (Cox, 1958) The technique is nowadays still widely applied in the field of retail credit risk modelling. The method is about estimating the beta-coefficients in the following formula. The logistic regression equation is used to estimate the probability of default.

$$\log\left(\frac{P(x)}{1-P(x)}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \qquad (1)$$

The left-hand side is called the logit function, where P(x) is the probability of default as a function of risk driving variables x. The right-hand side of the equation is a linear combination of different risk driving

variables and their weights. For such a model to be qualified as a good model, a lot of assumptions must be satisfied. The most important assumptions that are required to be fulfilled are that the independent variables have an acceptable correlation, that there is no multicollinearity amongst the independent variables and that the logit of the continuous variables is linear.

A correlation matrix is used to evaluate what variables show correlation to the variable that we aim to predict. Variables with a high positive or negative correlation to the target variable have high predictive power, when the correlation is zero, the variable will (in general) not add predictive power to a regression model. The other (maybe even more important) reason for checkingthe correlation matrix is that a Logistic Regression model assumes independence between the model variables. When variables are correlated, assumptions and conclusions drawn from the fitted model might be misleading. Therefore, if there are multiple variables with a high correlation, one or more of them are to be removed.

The method of logistic regression belongs to the family of the Generalized Linear Models (GLM)). GLM models are linear regressions where a link function is used to transform output to make sure it has the right characteristics (Nelder, 1972). When probability is estimated, the output must be in the interval[0,1] and to achieve this a logistic transformation is often used. When the logistic transformation is chosen, data can be transformed by means of Weight of Evidence (WOE). This method using the following formula and binned data to assign weights B_i in terms of the log odds ratio of the binary response variable. After this transformation, all variables have the property that a higher WOE bin value corresponds to a higher probability of default.

Goodness of fit

A logistic regression is said to provide a better fit to the data if it demonstrates an improvement over a model with less predictions. The likelihood ratio test compares the likelihood of the data under the full model against the likelihood of the data under a model with fewer predictors. Removing predictor variables from a model will always make the model fit less well but it is necessary to test whether the observed difference in model fit is statistically significant. The likelihood ratio test can beperformed in R using the lrtest() function from the lmtest package.

Higher values of the LR test statistic result into small p-values and provide evidence against the reduced model. However, when the test statistic is lower and the alpha level is greater than 0.05, we cannot reject the null hypothesis, and this provides evidence in support of the reduced model.

Pseudo R^2

In logistic regression there is no R^2 statistic which explains theproportion of variance in the dependent variable that is explained by the predictors. However, there are a number of pseudo R^2 metrics that could be of value. The most notable one is the McFadden's R^2.

Variable Importance

To assess the relative importance of individual predictors in the model, we can also look at the absolute value of the t-statistic for each model parameter. This technique is utilized by the varImp function in the caret package for general and generalized linear models in the R language. These variable importance tables inform us as to the strength of the effect associated with each explanatory variable in the model. One should carefully examine these table and use them to determine if there are predictors that could be thrown out of the regression model.

Forwards and backwards selection methods

There are two methods on how to select model variables to be included or excluded the model. These are the forward and backward logistic regression methods. Using the backwards logistic regression model, the first iteration of the model includes all the variables. It then proceeds by selecting the least contributing independent variable. This process is being repeated until an optimal model is reached. Forward selection is exactly the mirrored version of this process. Here is being started with selecting just one variable that on its own provides the most predictive value to the model. This variable can be selected after a univariate analysis. With this method you are adding variables to the model until an optimal model is reached. The results of the forward- and backward selected models does not have to correspond with each other in terms of the variables that are included in the final model.

AIC criterion

All combinations and thus iterations of the logistic regression models are compared based on the Akaike Information Criterion (AIC). The AIC measures how good a model is, relative to another model created in the same environment (Sakamoto, 1986). Models are punished for having unnecessary complexity, which is done with the following formula:

$$AIC = 2k - 2\ln(L) \tag{3}$$

Here k represents the number of estimated parameters and L represents the maximised value of model likelihood. The value of AIC itself has no meaning, but when we have models that are created in the same environment, we can select the best model by picking the one with the lowest AIC. This information is available after each model computation in the R software package. Thus, we are going to stop the backward selection process if the AIC value after deleting the next variableis increasing or in case of the forward selection method: if adding another variable is increasing the AIC.

Training vs testing

If the logistic regression model would be fitted on the whole data set, then the final solution is said to be overfitted. Therefore, it is a common approach to split the data into a training and a test set. Often a 70/30 split is performed on the total data set and each record is given a value at random (with a probability of 0.7 to get selected as a training record and a probability of 0.3 to get selected as a testing record). The logistic regression model was made based on the testing data. However, the testing data are going to be used to see how the model predicts the loans as either defaulted or non-defaulted loans. The model is then used to predict the probability of default for the testing data based on their parameter values of the independent variables that remain in the logistic regression model. This prediction can be analysed to see which loans of the

Contingency table

When developing models for prediction, the most critical metric regards how well the model does in predicting the target variable on out of sample observations. The process involves using the model estimates to predict values on the training set. Since we know the actual status of the loan (defaulted or not-defaulted) we can compare this with the predictions from the mode (defaulted or not-defaulted). By creating a two-by-two matrix of these we get a contingency table. This table is used for a lot of performance metrics regarding how well the model can distinguish defaulted loans compared to reality. One of the most common performance indicators is the classification rate. The classification rate is the percentage of that the modelled prediction and reality are equal. Besides the classification rate also other metrics can be calculated from the contingency table such as the sensitivity, specificity, and recall (Kateri, 2014).

ROC curve

Another performance indicator in logistic regression models is the receiver operating characteristic (ROC). This is a measure of the classifier performance. This graph is using the proportion of positive data points that are correctly considered as positive and the proportion of negative data points that are mistakenly considered as positive it is possible to generate a graph (Hanley, 1982). This graph shows the trade-off between the rate atwhich you can correctly predict defaults with the rate of incorrectly prediction defaults. With this graph, the metric is concerned with the area under the curve also named the AUC. The value of the AUC is in therange between 0.50 and 1.00. A value of 0.50 indicates that the model is not better than a random model inpredicting the defaults. It is said that the model has no predictive power in this case. If the value of the AUCis in between 0.50 and 0.70 it is advised to consider revisiting the individual predictors in the model and consider if any other explanatory variables should be included. AUC values over 0.80 indicate that the model does an exceptionally good job in discriminating between defaulted loans and non-defaulted loans.

5. Selection of transactions

This section answers the third research question in terms of which transactions are selected for analysis and which steps are taken in the model process of one complete transaction. The starting data set has 223 transactions in 11 different countries. However, the final list ended up with just two transactions.

However, not all transactions have the same data quality. In some transactions some fields might be not reported (ND1-ND6 codes) that are marked potentially impactful on the probability of default. These transactions are not selected. Furthermore, defaults tend to develop over the lifetime of a transaction. The average lease term of a contract is between 36 and 60 months. From a pragmatic perspective, only transactions with more than 24 months of data have been excluded from the analysis. In some transactions the currency of the reported deal was not in Euro's. This makes it difficult to compare the results with transactions denominated in Euro's, so therefore only transactions in EUR have been included. Since we are ultimately trying to classify defaults, a minimum requirement is that there are defaults present in the data set. If there are not enough defaults in a transaction, we can speak of rare events. In these kind of rare events other methods besides logistic regression are preferable (King, 2001). To avoid this problem, we set a minimum number of defaults to be present in the data for the transaction to be taken into consideration for analysis. The threshold was set at a default rate of 1 percent of the loans in the transaction.

After applying the criteria explained before there remained a list of 20 transactions that were deemed suitable for analysis. For each of these transactions a data quality check was performed. In some transactions important information was not available (credit risk score, income, Loan to Value (LTV). The final list contains 7 different transactions.

However, it is also necessary to verify that the assumptions for using the logistic regression methodology must hold (Starkweather, 2011). The major assumptions that must be met are that the errors must be independent, that there is anabsence of multicollinearity amongst the independent variables, there is a lack of strongly influential outliers and there is linearity in the logit of continuous variables.

6. Model steps

After the transaction selection was completed the modelling of the transaction could get started. The total process consists of a lot of different steps, divided over the following three main categories: Preparation, Modelling and Performance.

Preparation

Before getting started on the model itself some preparation had to be done. As a first step the data must be retrieved from the database and put in a format that can be handled by R. There was opted to safe the data from the SQL database as Excel files and to import them into the program R. The data was then cleansed, and the data types of the independent variable types were adjusted where necessary. A summary of the variables including the box plot statistics was made, graphs were plotted of the independent variables in relation to the default status of the loan. This gives a good overview of the dataset. Next, records with missing values in one of the fields were removed from the transaction. There was opted for this measure since the amount of loans was very large compared to the number of independent variables. Excluding missing values only resulted in minimal losses of data points compared to the transactions and will therefore have no impact on the overall model results. The last step in the preparation phase is to plot the logit against the values of the continuous independent variables. A linear trend must be observed to fulfill the linearity assumption that is required to perform the logistic regression.

The independent variables that ended up in the final evaluation of the transactions consist of the following list of 11 independent variables and their data types in R. A correlation matrix was created and observed that there was no presence of multicollinearity amongst the independent variables. Therefore, none of the following variables was excluded anymore. Important to note that one specific period was selected for all transactions. S

- Borrower Credit Quality (score): for some transactions there was a borrower credit quality score available. However, there was no legenda on the scores available and different transactions use different scoring systems. There was no credit scoring variable such as the FICO available.
- Employment Status (category score)
- Income (in EUR)
- Expected maturity of the loan (in months)
- Principal balance (in EUR)
- Down payment amount (in EUR)
- Loan to Value (in %)
- Interest rate on the loan (in %)
- Car manufacturer (brand)
- New or used vehicle (Boolean)
- The original vehicle value (in EUR)

Modelling

After the data has been prepared and cleansed it is ready to be modelled. The first step here is to split the data into a training and in a test set. The main idea here is that you would like to not only create a logistic regression model based on the default data of a transaction, but you would also like to see how well the model performs. To test this, the model must be tested on a different part of the dataset than the data it was fitted on. Thus, the full data set is split into a training data set and test data set with a 70/30 split at random.

There are two different methods in selecting which variables to include or exclude in the final model: backward selection and forward selection. Backward selection includes all variables in the first iteration of the model. This can be programmed like this in R:

```
Library(caTools)
Library(caret)
Library(HH)

# Logistic Regression Model 1 (all variables).

DefaultLogisticModel1 = glm(AccountStatus ~ BorrowerCreditQuality + EmploymentStatus + Income + ExpectedMaturity + PrincipalBalance + DownPayment + LoanToValue + InterestRate + CarManufacturer + NewUsed + VehicleValue, data=defaultTrain, family=binomial)

summary(DefaultLogisticModel1)
varImp(DefaultLogisticModel1)
vif(DefaultLogisticModel1)
```

Using the following commands in R it can be determined which of the variables is having the lowest importance to the model. A next iteration of the model is performed DefaultLogisticModel 2 that is executing the same statement. This process is repeated until the Akaike Information Criterion value (shown in the summary of the model iteration) is increasing. Remember that theAkaike Information Criterion punishes the model for having too many variables. After the breakpoint is reached where it is not beneficial anymore to remove additional variables we end up with our final model.Next, the goodness of fit test for the model is performed. These include the maximum likelihood ratio, the Hosmer Lemeshow test, and the Wald test.

The same steps are repeated for the forward logistic regression model. Instead of starting with all variables we are going to predict the defaults based on one variable first. This can be programmed similarly; in this example the independent variable Borrower Credit Quality is used. However, it is required to perform this for all 11 variables to determine which independent variable on its own is the best predictor for defaults in a logistic regression model. Similarly, to the backward method we are going to analyze the model coefficients and significance of the variables in the model, determine our decisions to add more independent variables to the model using the AIC criterion and the Variable Importance Factor.

```
DefaultLogisticModel1 = glm(AccountStatus ~ BorrowerCreditQuality, data=defaultTrain,
family=binomial)
```

It must be noted here that the field Account Status is the dependent variable. This field contains a 0 for loans that have not been defaulted, and a 1 for the fields that have been defaulted. Loans that were in arrears in the data field were removed from the dataset. The reason for this is that they are in the middle of the pack. The number of days the loan was in arrears was not taken into consideration.

Performance

Based on the backward- and forward selection during the modelling phase and several iteration steps in including and excluding the variables we end up with a final model based on the AIC value that has been minimized. The model that has been based on the training data is going to be tested by using the remaining 30% of the dataset. Using the characteristics of the loans and the coefficientsin the final model, a prediction can be made for each individual contract on the likelihood that the loan willdefault. Due to the nature of the logistic regression and the use of the Sigmoid function the output the prediction will always be scaled in the range of [0,1]. It therefore presents a likelihood of default instead of a probability of default. A loan with a likelihood of default of 0.30 is deemed more likely to default than a loan with the value of 0.03 for example.

Based on a threshold value set in R the loans can either be classified as performing (non-defaulted) or non-performing (defaulted). Then the predictions made by the model can be checked against the actual state of the loans. By varying the threshold, we can observe whether the model is better at predicting the defaults considering a certain threshold. We already know the reality for a certain moment in time regarding to the actual account status. By comparing the results from the predictions in a 2x2 matrix (predicted vs actual outcomes), we receive a contingency table. Based on the contingency table some performance metrics of the model can be calculated such as the accuracy, precision,recall and the fest.

Using the relationship between the true positive rate and the true negative rate the performance of the model can be further analyzed. A Receiver Operating Curve (ROC) is plotted and the area under the curve is calculated to determine the score. A higher Area Under Curve (AUC) corresponds with a better model. After the performance evaluation of the model, it is also necessary to interpret the findings.

7. Hypothesis

For the independent variables that are used in the final analysis a hypothesis is made on the expected direction of the parameter in the logistic regression model. Table 3 shows these hypotheses for the variable as an impact on the Probability of Default (PD). The expected relationship is noted as either (+) or (-). A (+) indicates that a higher value of the independent variable corresponds with a higher expected probability of default and therefore a positive sign before the coefficient in the final model. Likewise, a (-) indicates that a lower value of the variable corresponds with a lower expected probability of default. To illustrate this, a higher loan to value (LTV) is an indicator of more risk and therefore a higher probability of default. Each row in the table shows the field name in the Model, the datatype, the hypothesis, and an explanation of the hypothesis in relation to the impact on default of the variable. The hypotheses of this section on a parameter level are validated against the actual results in the next sections of the modelled transactions to see whether there are any results that are counterintuitive, interesting, and provide room for discussion

Table 3: Hypothesis for each independent variable.

Field Name in R	Field Name	Data type	Impact on PD	Explanation Hypothesis
BorrowerCreditQuality	Borrower Credit Quality	Text/Numeric	Higher quality (-)	A higher borrower credit quality results in a less risky loan. Therefore, it is expected to have a lower probability of default. Therefore, we expect a negative relationship between higher quality and lower PD. However, the field BorrowerCreditQuality is different for every transaction. While some transactions use a numeric scale, others use categories [1,2] or [1,2,3] or colours to indicate the BorrowerCreditQuality score.The hypothesis is that a higher Borrower Credit Quality is associated with a lower probability of default. However, it is not feasible to derive from the data what is an indicator of a good or bad borrower credit quality.
ExpectedMaturity	Expected maturity	9(11).99	Higher maturity (+)	The maturity of the loan is the expected timing of the repayment of the principal + interest. Due to a longer repayment period, there are more installmentsto be made and more interest to be repaid on a loan. It also increases the likelihood for arrears to mount up to more than 90 days, since there are more overall payment moments compared to a shorter Expected Maturity. Due to prepayments the actual maturity of the loan can differ from the Expected Maturity date that is planned in the payment schedule of the loan.
EmploymentStatus	Borrowers Employment status	List	Employment (-) Unemployment (+)	Unemployment borrowers are expected to have a higher 9PD since they have no primary guaranteed income every month (+). Employment that leads to a primary income is expected to decrease the PD (-).
Income	Primary Income	9(11).99	Higher income (-)	The higher the income of the borrower, the more likely he is to make his payments. The relationship with the PD is therefore expected to be negative (-).
PrincipalBalance	Original Principal balance	9(11).99	Higher balance (+)	The principal balance is the amount of the loan that was borrowed. The principal is the part of the loan without any accrued interest that must be returned. The higher the principal balance is, the more a borrower has borrowed. A higher principal balance is associated with a higher risk of not making timelyand full payments and therefore is expected to increase the PD (+).

DownPayment	Down Payment amount	9(11).99	Higher down payment (-)	The down payment amount on a loan is a type of payment made on the onset of the purchase of an expensive good or service. The payment represents apercentage of the full purchase price. A typical down payment decreases the amount of interest paidover the lifetime on the loan. It also provides lenderswith a degree of security. Therefore, the higher the down payment amount, the less likely a borrower isto default (-).
LoanToValue	Original Loan to Value	9(3).99	Higher LTV (+)	The Original loan to value is a ratio calculated at thetime of origination for a loan. It reflects the loan to value ratio of the loan amount secured by the vehicleon the origination date of the underlying loan. Often lenders provide better loan terms to borrowers who have LTVs below a certain percentage, because the loan is less risky. The higher the original LTV the higher the expected PD becomes, especially for loans that are higher than the value of the underlying (>100%).
InterestRate	Current interest or discount rate	9(4).9(5)	Higher interest rate (+)	The interest rate/discount rate is charged to the borrowers, because of the time value of money that is decreasing over time. The discount rate also reflects the amount of risk that there is in the loan. Similarly, to the hypothesis of the BorrowerCreditQuality, a riskier loan is expected to have higher interest rate to compensate for this risk. Therefore, a higher interest rate is expected to have apositive effect on the PD (+).
CarManufacturer	Car manufacturer	Text	(+/-) depending on type	The purpose of the loan in Auto ABS transactions inthe first place is to be able to ride the vehicle. Unlikewith mortgages, vehicles have different types of brands. The research of Agarwal et. al (2008) suggests that there is information hidden in the choice for the car brand. Since there are a lot of brands, there are also a lot of components in the regression model. There is no hypothesis beforehand per car manufacturer made (+/-).
NewUsed	New or Used car	List	New (-) and used (+)	Due to the research of Agarwal et al, we expect that new cars have a lower PD (-) and used cars have ahigher PD (+).
VehicleValue	Car valuationat loan or lease origination	9(11).99	Higher car valuation (+)	The Vehicle value at loan/lease origination is the worth of the vehicle at origination in Euro's. We expect more valuable vehicles to be more expensiveto loan, increasing the risk in the loan. Therefore, it is expected that a higher vehicle has a positive relationship with the PD (+).
AccountStatus	Account Status	List	Dichotomous outcome variable	The account status is the field which states the current status of the account. According to the ECB template there are 10 possible statuses of an account.They can be either performing, restructured without or with arrears, defaulted, in arrears, repurchased by the seller, redeemed or other. The statuses we are interested in are the ones that are performing, in arrears, defaulted and restructured. These 5 give a complete overview of the performance of a transaction. This field also contains the dichotomousoutcome variable. A loan is either in default or not in default. The choice is made to exclude arrears from the table, since arrears can either go into default, or arrears could be transitioned into performing loans again after payments are made. Performing loans get a value of 0 in this field, whereas defaulted loans get a value of 1 in this field.

8. Results

Using the methodology described in Sections 3 and 4 the modelling of all seven logistic regression models has been performed. Due to the different data sets, the results are also different over the different transactions. Some transactions only have a few independent variables in the final iteration of the model, where others have over half of the initial parameters. Appendix A shows the results per parameter, for all the independent variables studied. Due to the repetitiveness of the results, transaction 5 modelling results have been included as an attachment in Appendix B.

The same methodology has been applied to six other transactions. The results in terms of the AUC values of the final model iterations on the training and testing data set are shown in Table 4 below. The results indicate that overall, the models are doing quite well in terms of their predictive value. The transaction being used in Appendix B is clearly showing the worst performing in terms of classifying defaults. Also, the difference between the training and test set is the greatest for that transaction, in the other transactions the AUC values are way closer to each other. The breakdown shows that there is one transaction in the 0.6-0.7 bucket, one transaction in the 0.7-0.8 bucket, two transactions in the 0.8-0.9 bucket and three transactions in the 0.9-1.0 bucket. Overall, the logistic regression has performed quite well for predicting defaults in these selected transactions except for transaction 5 that is further specified in appendix B. These results indicate that Logistic Regression is still working for modelling defaults, despite not being the most sophisticated algorithm available.

Table 4: AUC values for each transaction: training and testing data

Transaction	Training model	Testing model
#2	0,8597	0,8653
#5	0,7213	0,6564
#6	0,8816	0,916
#10	0,9837	0,9723
#11	0,8155	0,8171
#17	0,9846	0,986
#18	0,7994	0,7992

Which variables have no predictive value?

In terms of variability, there were some differences in the final outcomes of the models. Out of the 22 independent variables a total of 13 where present in the final iteration of the logistic regression model for one or multiple transactions. However, there were also some variables that never got to the final solution. The variables that were not selected in any of models are the following variables. These observations are highlighted in Appendix A.

- **Originator**: this one is self-explanatory. It is a dummy variable for the name of the transaction and does not provide any power in the model since it is the same for every loan within a transaction.
- **Regulated Loan**. If a loan is regulated or not was not an important indicator of default. For most transactions being research this variable had a very large part of regulated loans. This might be a reason why it did not end up in one of the transactions.
- **Income Verification**: unfortunately, the income verification did not impact the probability of default in one of the transactions significantly. Obligors that have their income verified when taking the loan instead of self-certifying the income would from a risk perspective lead to a lower probability of default. Like the regulated loan, most transactions studied had the incomes all verified.

- The **Geographic Region** was not successful in discriminating as a determinant of default in these transactions. There are too many options in most countries, creating a variable with a lot of factors. Although we have seen some regions with a significance (*) on the default, it was never included in the final model.
- The **Payment Frequency** is also one of the variables that did not end up in the final model. Reason for this is that in almost all studied transactions the payment frequency is monthly and does therefore not provide discriminatory power. Therefore, the hypothesis could not be tested.
- Surprisingly, the **Car Manufacturer** also never ended up in the final iteration in one of the models. The hypothesis here was that there is a difference in the so-called luxury brands and regular brans. Evidence for this was not found during this research. The research of Agarwal shows that there is hidden information revealed in the choice for the object type (Agarwal, 2008). One of the reasons that this researchdid not provide evidence for this might be a lack of luxury vehicles in the selected securitized portfolios.
- The **Vehicle Value at Origination** also surprisingly did not end up in any of the models. The main reason here is probably that there is a correlation with the Loan to Value, which is an indicator that ended up in multiple models. Other loan characteristics such as the Down Payment amount, principal balance and expected maturity did end up in the final model iterations.
- The **Customer Type** was of no large impact on predicting which customers would default.
- The **Payment Method** was of no large impact on predicting which customers would default.

Which variables did have predictive value in one or more final models?

To answer the main research question, it is necessary to answer what the determinations of default for Auto Loans in European ABS transactions. are What can be considered the determinants of default and how can these variables be off use in determining default risk in the transaction? The variables that have been included in one or more final iterations of the models the relationship with the hypothesis is explained.

- The **BorrowerCreditQuality** is an optional field that is causing the most issues to interpretate.The reasons for this are that every issuer is using their own internal rating system to classify obligorsin a certain way. This variable ended up being relevant in 5 of the 7 transactions. However, the meaning of the values in the source data have no meaning. Therefore, we can only conclude that this field is highly important for determining the probability of default. This variable is present in the following transactions: #6, #10,#11,#17 and #18. To improve our understanding of therelationship between the borrower credit quality score and the probability of default it is necessaryto streamline the method begin applied over issuers, or at least make a mapping to the same categories to enhance the comparability of this variable. The fact that it is optional, is also asking for problems. Issuers that don't provide data on this field prohibit the investors from doing a due diligence check on the data of the underlying contracts that they are investing in. This topic is further covered in the next section as well.

- The borrower **Employment Status** was found significant in transaction #11. There are two categorieswith employment types that had a positive impact on the probability of default, which are employment status pensioners and self-employed. These findings make sense, since the income forpensioners is generally lower, and self-employed obligors have more variability in their income than obligor who are employed by a boss. This is in line with what would be expected.

- The variable **Primary Income** ended up begin in only two of the seven transactions (#2 and #6). The impact of this variable was similar to expected. A higher income is associated with a lower

probability of default, a relationship that was not surprising to be found and that is in line with literature.

- The **Amortisation Type** ended up in one transaction, being #5. From the results of this logistic regression model having an amortization type being (5) increases the probability of default. There have not been made specific hypothesis on each category of amortisation types.

- The **Expected Maturity** equals the number of payments to be made, not taking into consideration prepayments on the loan. From the hypothesis of this field a higher expected maturity is associated with a higher probability of default. In the case of transactions #2, #5, #10, #17 and #18 the expected maturity was indeed in the final model. Interestingly in only three of these five transactions the direction of the sign is (+) as expected. There are two transactions where the hypothesis is off and shorter expected maturities are associated with a higher probability of default (-). This makes for an interesting finding.

- The higher the **Principal Balance** the greater the probability of default, according to the hypothesis for this field. And this is found to be true for all models where this variable was in the final iteration of the model selection (models of transactions #2, #5, #6, #11). The results are showing that regarding the principal balance, a higher balance is a clear indicator for more risk and thus a higher probability of default.

- The **Down Payment Amount** is included in three different logistic regression models, #5, #6 and #11. The hypothesisis that a larger down payment amount is associated with a lower probability of default. This hypothesis is supported by the findings of the models. Similarly, to the principal balance, there areno inconsistencies found, suggesting that the hypothesis is sound.

- The **Loan To Value** is an indicator of the ratio of the loan in relation to the value of the underlying of the loan. In this case the vehicle. A higher LTV is associated with more risk and therefore a higher probability of default. The variable ended up in three transactions: #2, #10, #18. Surprisingly in the results of the model for transaction 10 we find an inconsistency with our hypothesis. In this model a lower LTV is associated with a higher probability of default. For the other two transactions the results are in line with what is expected.

- The **Interest Rate** that is being asked by the issuer for the loan is present in four of the seven transactions (#2, #11,#17, #18). A higher interest rate is associated with a higher probability of default, because a larger part of the monthly installment consists of interest instead of redemption. The results support this hypothesis, since in all four transactions this relationship between a higher interest rate and higher probability of default is supported.

- **Registration Year** is an interesting variable. This variable is close to the NewUsed variable in teams of practical meaning, so these are discussed together. Interestingly, the year of registration is present in five of the seven transactions. In all cases there is found a negative relationship, indicating that newer cars are indeed less likely to default compared to cars with a higher vintage. However, when we look at the NewUsed variable we noticesomething unexpected. In three of the five transactions there is found that used cars have a higher probability of default, compared to two transactions the other way around. Therefore, it seems thatthe actual year of registration is a better indicator for the probability of default then the NewUsed variable.

Answering main research question

The key takeaway from modelling these transactions is that all transactions are different. It is not feasible to create a single model that covers all the transactions. Not a single final model iteration contained the same variables. This is a clear sign thatevery transaction is different, due to the unique characteristics in terms of the way a transaction is structured.

Therefore, investors are advised to perform their own due diligence as a complement to the reports that are available from CRAs. However, the results do provide direction for where to look like the determinants of defaults in the Auto ABS transactions studied in the context of the European market. Based on this research the independent variables that are considered as most promising determinants for ahigher probability of default on a securitized car loan are:

- a worse Borrower Credit Quality score
- a “bad” borrower credit quality score;
- a lower Primary Income of the obligor;
- a higher principal balance;
- a lower down payment amount;
- a higher LoanToValue;
- a higher interest rate;
- higher vintage of the car

9. Suggestions for further Research

There are several suggestions to be made for future research on the topic of transparency and default prediction in the (European) Auto ABS sector. These are the impact of STS reporting on the market, a time series analysis, macro-economic variables, standardisation efforts of the internal ratings scores, testing other alternative modelling techniques and sampling techniques.

Impact of STS reporting on transparency in the market

Since the time this research was executed, regulation has changed in this market. Most noticeably, the regulation has been impacted from the Securitisation Regulation. A general framework for securitisations is established for simple, transparent, and standardised (STS) transactions. The regulation is the cornerstone of the European Union's effort to establish a capital markets union, by creating a single market for investment services and activities and to insure a high degree of harmonized protection for investors in financial instrument.

Although later than expected, on the 23rd of September 2020 the technical standard on disclosure requirements underthe Securitisation Regulation were published in the official journal of the European Union. They came intoforce starting September 23rd, 2020. For each asset class a template has been created. For example, for theasset class of Auto loans, there are three templates mandatory to be reported; Annex 5 – underlying exposure automobiles, Annex 12 – Investor Report Non-ABCP securitisation and Annex 14: Inside information or significant event information Non-ABCP securitisation (esma.europe.eu, sd).

However, the first uploads of STS transactions were done in October 2020. Therefore, it was not feasible to incorporate this in the current research. For future research it would be interesting to do a similar analysis on data from transactions with the STS label, using the field from the new template. By doing such an analysis it can be tested if the STS requirements did make an impact on the regulation side. This research would be feasible starting from 2023 onwards, since only new transactions are required to fulfill these STS requirements and it generally takes several months or even years before defaults accumulate over a transaction. The impact of defaulted contracts would therefore only be visible starting in 2021/2022. Without enough data points of default, it becomes difficult to assess the predictability of defaults due to the imbalance of the dataset.

One of the goals of the European Securities and Markets Authority (ESMA) is to increase the transparency. It would be interesting to see whether due to the standardized approach default determinants over multiple transactions can be derived. The data for such an analysis could also be derived from the EDW. However, this would entail that only transactions can be considered that have data for all variables at hand. For some of the variables, the reporting entities are free to use the No Data codes as an option.

Time series analysis

Another factor that is interesting is the timing of the loans. In this research one period (with a large deviation from the origination of the transaction) was chosen based on the month that the research has started in. However, defaults are not random events. There is information in the prior months of data that might be valuable for predicting defaults. For example, the number of days in arrears in the previous period might be a good predictor of defaults. Currently we are looking at a picture that was taken at a certain moment in time, where it would be more interesting from an analysis point of view to look at the whole movie over the lifetime of a transaction.

Macroeconomic variables

A third recommendation is that in the current research only incorporates the data that is available from the database from the European Data Warehouse. This does not include other variables such as macro-economic variables. Defaults of customers are more likely to occur during a recession or crisis than during a period of economic prospect. Therefore, some macro-economic variables are interesting to consider when performing such an analysis. For example, the growth in the Gross Domestic Product (GDP), the unemployment rate in % or the consumer confidence rates are variables that are worth to be included in such an analysis. Including macroeconomic variables is suggested in literature as well ((Bellotti, 2009)

Standardize the internal rating score of good vs bad loans

A fourth recommendation is to dive more into the internal rating of the clients. When lessees apply for a loan the obligor is screened and put into risk categories by the risk department of the issuer. The variable that was used for this Research was the BorrowerCreditScore.

For research purposes the ratings that were currently available were unclear. Some transactions used the risk categories of 1 and 2; without giving meaning to this variable. Others provide a number-based score, such as a value between 0 and 1000, without giving meaning to that number. Although the internal rating does say something about the quality of the borrower of the loan, a standardized approach would help for research purposes. Now, only during the modelling phase and in case the variable is significant in the final model we get a grasp of what is intendedwith the borrower credit quality score of 1 or 2 for example. However, the originators have their own internal systems to classify the obligors into different risk categories. The United States already apply such as system commonly known as the FICO score. This score is determined by the Fair Iscaac Company and provides a credit score to a loaner based on several variables including historical credit evaluation. The values indicate the creditworthiness of an obligor, a metric that could be very useful when performing such an analysis, since it is the same for every transaction, unlike the BorrowerCreditQuality measure used.

There are other methods out there

The fifth recommendation to this research would be to apply other machine learning techniques and sampling methods to this problem to see whether there are alternatives that can be used for this classification problem. In this research the focus has been on one technique with a multitude of datasets with similar variables. However, it is also worth researching if there are any other machine learning methods that improve the performance of our classification models for the various transactions.

A comprehensive study to machine learning methodologies regarding credit default risk was recently performed (Stelzer A., 2019). This is a benchmark study that compares the performance of different kind of classification models and sampling methods. A total of 23 different machine learning techniques and 6 sampling methods are used over 4 datasets to score the performance of the classification models. The methods can be categorized in three different groups: individual models, homogenous ensembles and heterogenous ensembles. The market standard of logistic regression is outperformed by a lot of other methods, ranking just 18th, whereas the best six methods belong to the group of homogenous ensambling methods.

Try other sampling techniques

The final recommendation would be the sampling techniques being used in this type of analysis. The data set is very imbalanced since the number of events (defaults on the loans) is very low compared to the total number of loans. This has implications on the model because we are trying to predict the events with our classification model. Several strategies mitigate this imbalance in the dataset. The most well-known

strategies are up-sampling and down-sampling, where the test data set is manipulated. However, also other methods such as Random Over-Sampling Examples (ROSE) and Synthetic Minority Oversampling Technique (SMOTE) and Borderline Synthetic Minority Oversampling Technique (BSMOTE) are available. The results of Stelzer show that up-sampling is the most effective methodology to handle imbalanced data sets for logistic regression (Rahim, 2019)

References

Agarwal, S. A. (2008). Determinants of automobile loan default and prepayment. *Economic Perspectives*, 32(3).

Bellotti, T. &. (2009). Credit scoring with macroeconomic variables using survival analysis . *Journal of the Operational Research Society*, 1699-1707.

Blommestein, H. K. (2011). Outlook for the securitisation market. *OECD Financial Market Trends*

Calomiris, C. (2009). Financial Innovation, Regulation and Reform. *Cato Journal*, 65.

Cox, D. R. (1958). The regression analaysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 215-232.

esma.europe.eu. (n.d.). Retrieved from https://www.esma.europa.eu/policy-activities/securitisation/simple-transparent-and-standardised-sts-securitisation

Fagerland, M. &. (2012). A generalized Hosmer-Lemeshow goodness-of-fit test for multinomial logistic regression models. *The Stata Journal*, 447-453.

Gurney, K. (2018). *An introduction to neural networks.* CRC press.

Hanley, J. &. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiologoy*, 29-36.

Kateri, M. (2014). *Contingency table analysis. Methods and implementation using R (First edition).* Germany: Editoral Advisory Board.

King, G. &. (2001). Logistic regression in rare events data. . *Political analysis*, 137-163.

Myles, A. J. (2004). An introduction to decision tree modelling. *Journal of Chemometrics*.

Na'im, A. (2006). Special Purpose Vehicle Institutions: Their Business Natuers and Accounting Implications. *Gadjah Mada International Journal of Business*.

Nelder, J. &. (1972). Generalized linear models. . *Journals of the Royal Statistical Society: Series A (General)*, 370-384.

Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, 1565-1567.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing*, 217-222.

Peterson, L. E. (2009). K-nearest neighbor. *Scholorpedia*.

Rahim, A. H. (2019). *Smote approach to imbalanced dataset in logistic regression analysis.* Singapore: Springer.

Sakamoto, Y. I. (1986). Akaike Information criterion statistics.

Shin, H. S. (2009). Securitisation And Financial Stability. *The Econmic Journal*, 309-332.

Starkweather, J. &. (2011). Multinomial logistic regression.

Stelzer A. (2019). Predicting credit default probabilities using machine learning techniques in the face of unequal class distributions. *arXiv*.

Wright, R. (1995). Logistic Regression.

Appendix A: Results per variable

The following table shows the results of the variables included in this research compared. The priority column shows whether the fields are optional (yellow) or mandatory (green) to be delivered in thedata whereas the tag column shows if a data field is fixed (static) or can change (dynamic). The final column shows which of the variables is or is not used in the models.

Field number	Priority	TAG	Field Name in R	Field Name	Data type	Impact on PD	Explanation Hypothesis	Used in one of the models ?
AA6	Mandatory	Static	Originator	Originator	Text	(+/-)	Different Originators of the transaction might have a different impact on the PD. Therefore the relationship is (+/-)	No
AA7	Optional	Static	RegulatedLoan	Regulated Loan	Y/N	Y(-) and N(+)	Regulted loans are expected to have better borrower credit quality and thus have a lower PD. Therefore the hypothesis is that the relationship is (-). For non regulated loans the hypothesis is that the borrower credit quality is worse.	No
AA15	Optional	Static	BorrowerCreditQuality	Borrower Credit Quality	Text/Numeric	Higher quality (-)	A higher borrower credit quality results in a less risky loan. Therefore it is expected to have a lower probability of default. Therefore we expect a negative relationship between higher quality and lower PD. However the field BorrowerCreditQuality is different for every transaction. While some transactions use a numeric scale, others use categories [1,2] or [1,2,3] or colours to indicate the BorrowerCreditQuality score. The hypothesis	Yes
AA16	Mandatory	Static	EmploymentStatus	Borrowers Employment status	List	Employment (-) Unemployment (+)	Unemployment borrowers are expected to have a higher PD since they have no primary guaranteed income every month (+). Employment that leads to a primary income is expected to decrease the PD (-).	Yes

AA17	Mandatory	Static	Income	Primary Income	9(11).99	Higher income (-)	The higher the income of the borrower, the more likely he is to make his payments. The relationship with the PD is therefore expected to be negative (-).	Yes
AA19	Mandatory	Dynamic	Amortisation	Amortisation type	List	(+/-) depending on amortization type	The type of amortisation determines the scheduled payment schedule for the borrower. Different amortisation types might impact the PD.	Yes
AA20	Mandatory	Static	IncomeVerification	Income verification	List	Verified (-) non-verified (+)	This variable indicates whether the primary income has been verified. If the income has not been verified, it is less reliable/trustworthy, and therefore it is expected that non-verified incomes have a higher PD (+).	No
AA21	Mandatory	Static	GeographicRegion	Geographic Region	List	(+/-) depending on location	The NUTS-3 regions is a result of the regional classification of the European Statistics Bureau Eurostat. The code consists of COROP regions. For example the Netherlands has 40 COROP regions. Within the data of a transaction the COROP regions of the borrowers are present. However, since there are so many corop regions within a transaction, only the letters of the code are used. This implies that all 'DE123' COROP regions are seen as DE. The relationship between the GeographicRegion and the PD is thus research on country level.	No
AA24	Mandatory	Static	ExpectedMaturity	Original Loan or lease term	Numeric	Longer term (+)	A loan is paid back over multiple periods of time. The longer the loan term is the later the expected maturity date of the loan equals. The hypothesis is that a longer loan/lease term results in a higher	Yes

							probability of default due to the longer payment obligations (+)	
AA26	Mandatory	Static	PrincipalBalance	Original Principal balance	9(11).99	Higher balance (+)	The principal balance is the amount of the loan that was borrowerd. The prinicpal is the part of the loan without any accured interest that has to be returned. The higher the principal balance is, the more a borrower has borrowed. A higher principal balance is associated with a higher risk of not making timely and full payments and therefore is expected to increase the PD (+)	Yes
AA29	Mandatory	Static	PaymentFrequency	Scheduled payment frequency	List	Higher frequency (+)	The payment frequency is the speed at which payments take place. It is common to do this on a monthly basis (thus a frequency of 12 times a year) but it can also differ. The higher the payment frequency, the less likely a borrower is to make timely payments and thus the higher the PD is (+)	No
AA34	Mandatory	Static	DownPayment	Down Payment amount	9(11).99	Higher down payment (-)	The down payment amount on a loan is a type of payment made on the onset of the purchase of an expensive good or service. The payment represents a percentage of the full purchase price. A typical down payment decreases the amount of interest paid over the lifetime on the loan. It also provides lenders with a degree of security. Therefore the higher the down payment amount, the less likely a borrower is to default (-)	Yes

AA35	Mandatory	Static	LoanToValue	Original Loan to Value	9(3).99	Higher LTV (+)	The Original loan to value is a ratio calculated at the time of origination for a loan. It reflects the loan to value ratio of the loan amount secured by the vehicle on the origination date of the underlying loan. Often lenders provide better loan terms to borrowers who have LTV's below a certain percentage, because the loan is less risky. The higher the original LTV the higher the expected PD becomes, especially for loans that are higher than the value of the underlying (>100%).	Yes
AA36	Mandatory	Static	ProductType	Product type	List	(+\-) depending on product type	There are different product types that can be arranged for a loan. It could be a contract purchase, hire, lease purchase, finance lease, operating lease, other or a loan. In the category of loans, there are amortising loans and balloon loans. The data from our transactions is related to amortising loans and balloon payments. Thus this field is also highly correlated with the field of amortisation type.	Yes
AA40	Mandatory	Dynamic	InterestRate	Current interest or discount rate	9(4).9(5)	Higher interest rate (+)	The interest rate/discount rate is charged to the borrowers, because of the time value of money that is decreasing over time. The discount rate also reflects the amount of risk that there is in the loan. Similarly to the hypothesis of the BorrowerCreditQuality, a more risky loan is expected to have higher interest rate to compensate for this risk. Therefore a higher interest rate is expected to have a positive effect on the PD (+).	Yes

AA44	Mandatory	Static	CarManufacturer	Car manufacturer	Text	(+/-) depending on type	The purpose of the loan in Auto ABS transactions in the first place is to be able to ride the vehicle. Unlike with mortgages, vehicles have different types of brands. The research of Agarwal et. al (2008) suggests that there is information hidden in the choice for the car brand. Since there are a lot of brands, there are also a lot of components in the regression model. There is no hypothesis beforehand per car manufacturer made (+/-)	No
AA47	Optional	Static	RegistrationYear	Year of Registration	YYYY	The higher YYYY (-)	The next characterisitc of the car being loaned is the year that the car is from. The year of registration of the car tells something about how old the vehicle is that is being securitized. It also says something about if the car is new or used (correlation with field NewUsed). The research from Agarwal et al suggests that there are Used cars have a higher probability to default. Used cars have a lower year of registration. Thus, the hypothesis is that if the year of registration increases the PD decreases (-)	Yes
AA48	Mandatory	Static	NewUsed	New or Used car	List	New (-) and used (+)	Due to the research of Agarwal et al, we expect that new cars have a lower PD (-) and used cars have a higher PD (+).	Yes
AA49	Optional	Static	VehicleValue	Car valuation at loan or lease origination	9(11).99	Higher car valuation (+)	The Vehicle value at loan/lease origination is the worth of the vehicle at origination in Euro's. We expect more valuable vehicles to be more expensive to loan, increasing the risk in the loan. Therefore it is expected that a higher vehicle has a positive	No

							relationship with the PD (+).	
AA54	Optional	Static	OrignationChannel	Origination Channel	List	(+/-) depending on origination channel	The origination channel is a field that describes where the loan is originated from. According to the ECB template for Auto ABS there are four main orignation channels: a dealer, broker, directly or indirectly. The fifth category is other. Depending on the origination channel the PD can be different. No hypothesis was made for this field per origination channel.	Yes
AA55	Mandatory	Static	CustomerType	Customer type	List	(+/-) depending on customer	The CustomerType field is the legal form of the customer. The customer can be either a public company, a limited company, a partnership, an individual, a government entity or other. No hypothesis was made about each individual customer type regarding the PD (+/-)	No
AA57	Mandatory	Dynamic	PaymentMethod	Payment method	List	(+/-) depending on payment method	The payment method is the way in which the customer is paying for the loan. This is a dynamic field that is based on the last payment received. There are various ways in which the loan can be paid, either through: direct debit, standing order, by cheque, cash, other or bank transfer (not direct debit). No hypothesis was made about each individual customer type regarding the PD (+/-)	No

AA74	Mandatory	Dynamic	AccountStatus	Account Status	List	Dichotomous outcome variable	The account status is the field which states the current status of the account. According to the ECB template there are 10 possible statuses of an account. They can be either performing, restructured without or with arreas, defaulted, in arrears, repurchased by the seller, redeemed or other. The statuses we are interested in are the ones that are performing, in arrears, defaulted and restructured. These 5 give a complete overview of the performance of a transaction. This field also contains the dichotomous outcome variable. A loan is either in default or not in default. The choice is made to exclude arreas from the table, since arrears can either go into default, or arreas could be transitioned into performing loans again after payments are made. Performing loans get a value of 0 in this field, whereas defaulted loans get a value of 1 in this field.	

18	Mandatory	Static	19
5	Optional	Dynamic	4
23	Total	Total	23

Appendix B: Results for transaction #5

This section is explaining the results of just one transaction due to the repetitive character of the results. The transaction selected for this sample is issued by Volkswagen by the originator Volkswagen Finance S.A. in Spain. The full models in R for all modelled transactions are available upon request by sending an enquiry to the author. This example is taken to illustrate the results from one transaction, followed by the overall results of the transactions and

Descriptive statistics

Before starting the modelling, we are having a glance at the descriptive statistics of the transaction. From Figure 6 below we quickly observe that there are 28.214 loans and data points. All the loans are regulated and have the same amortisation type, payment frequency, origination channel and payment method. Therefore, these variables will not be considered in creating the logistic regression model. Most of the obligors have an employment status of being employed (69%), which corresponds with EmploymentStatus = 1. However also other statuses are available in the data.

The median or mean income is roughly the same, around sixteen thousand Euro's. There are exactly 853 data points where the income is unknown and has also not been verified. The median expected maturity of the loans is exactly 5 years, and the principal balance is around seventeen thousand Euro's. There is a median down payment amount of 4500 for the vehicle.

The loan to value is quite high since the median value of the LTV is 95% and the 75th percentile of the data is over 100. Regarding the interest rate we observe very high interest rates throughout the portfolio of contracts. The median interest rate is 10%, whereas the 25th percentile already marks the 9.5%. The loans are mostly for new cars. There are only a few car manufacturers in this transaction, with the types of Seat (40%), Volkswagen (31%) and Audi (15%) together make up over 85% of the total pool. The median vehicle value of the objects is slightly lower than one yearly income.

The overall view from the descriptive statistics is that for the issuer there is probably a large margin on this transaction. The income of the obligors is relatively low compared to the principal balance of the loan. The LTVs are high to very high and the interest rate is extremely high. However, the statistics on the account status of these loans indicate that the default rate on this transaction is quite low: 0.88%. Defaults are being defined as arrears larger than 90 days in this transaction.

```
> summary(default)
  Originator        RegulatedLoan BorrowerCreditQuality EmploymentStatus     Income       Amortisation IncomeVerification
 Length:28214       Y:28214       1:12926               1:19456           Min.   :     0   1:28214      3   :27361
 Class :character                 2:15288               3: 3737           1st Qu.: 11000                NA's:  853
 Mode  :character                                       6:  853           Median : 16000
                                                        7:  703           Mean   : 16662
                                                        8: 2224           3rd Qu.: 21000
                                                        9: 1241           Max.   :120000
                                                                          NA's   :853
 GeographicRegion   ExpectedMaturity PrincipalBalance PaymentFrequency  DownPayment     LoanToValue     ProductType  InterestRate
 Length:28214       Min.   :35.00    Min.   :  4487   1:28214          Min.   :    0   Min.   : 20.00   8:28214     Min.   : 4.000
 Class :character   1st Qu.:48.00    1st Qu.: 13156                    1st Qu.: 2000   1st Qu.: 75.00               1st Qu.: 9.500
 Mode  :character   Median :60.00    Median : 16281                    Median : 4500   Median : 95.00               Median :10.000
                    Mean   :66.25    Mean   : 17366                    Mean   : 5792   Mean   : 94.02               Mean   : 9.894
                    3rd Qu.:84.00    3rd Qu.: 20361                    3rd Qu.: 8100   3rd Qu.:110.00               3rd Qu.:10.500
                    Max.   :96.00    Max.   :132757                    Max.   :81800   Max.   :145.00               Max.   :12.500

   CarManufacturer  NewUsed    VehicleValue   OriginationChannel CustomerType PaymentMethod AccountStatus
 SEAT      :11118   1:22650   Min.   : 5300   1:28214            1   :   51   1:28214       0:27969
 VOLKSWAGEN: 8813   2: 5564   1st Qu.:11700                      2   :  730                 1:  245
 AUDI      : 4168             Median :14900                      3   :   28
 SKODA     : 3259             Mean   :15826                      4   :27361
 VW LCV    :  849             3rd Qu.:18700                      NA's:   44
 FORD      :    2             Max.   :82900
 (Other)   :    5
```

Figure 1: Descriptive statistics of independent variables

Figure 2 provides a summary of the default variables as being recognized by R after loading in the data. The data is imported from an Excel file. The variables that have factor 1 have no predictive value (similar for all loans) in the model and are removed. These removed variables include: RegulatedLoan, Amortisation, IncomeVerification, PaymentFrequency, ProductType, OriginationChannel and PaymentMethod.

```
> str(default)
Classes 'tbl_df', 'tbl' and 'data.frame':       28214 obs. of  22 variables:
 $ Originator           : chr  "VOLKSWAGEN FINANCE S.A." "VOLKSWAGEN FINANCE S.A." "VOLKSWAGEN FI
 $ RegulatedLoan        : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...
 $ BorrowerCreditQuality: Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...
 $ EmploymentStatus     : Factor w/ 6 levels "1","3","6","7",..: 2 1 1 1 2 1 1 1 6 1 ...
 $ Income               : int  14000 14000 106000 11000 23000 13000 13000 0 0 22000 ...
 $ Amortisation         : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...
 $ IncomeVerification   : Factor w/ 1 level "3": 1 1 1 1 1 1 1 1 1 1 ...
 $ GeographicRegion     : chr  "ES432" "ES300" "ES532" "ES616" ...
 $ ExpectedMaturity     : int  60 48 66 72 60 84 60 84 60 60 ...
 $ PrincipalBalance     : int  11657 13038 29928 17209 11880 12600 11935 29478 18774 25207 ...
 $ PaymentFrequency     : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...
 $ DownPayment          : num  2900 1200 2800 1800 4500 1200 2500 4000 2500 4500 ...
 $ LoanToValue          : num  80 95 95 95 70 110 100 105 95 105 ...
 $ ProductType          : Factor w/ 1 level "8": 1 1 1 1 1 1 1 1 1 1 ...
 $ InterestRate         : num  9 10 10.5 8 9.5 9.5 10 9.5 9.5 9.5 ...
 $ CarManufacturer      : Factor w/ 11 levels "AUDI","CITROEN",..: 6 3 10 2 5 4 3 11 11 11 ...
 $ NewUsed              : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 1 1 1 ...
 $ VehicleValue         : num  11900 11500 24500 14800 13500 8700 9500 21400 16500 19300 ...
 $ OriginationChannel   : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...
 $ CustomerType         : Factor w/ 4 levels "1","2","3","4": 4 4 4 4 4 4 4 4 4 4 ...
 $ PaymentMethod        : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...
 $ AccountStatus        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

Figure 2: Summary of variables and types.

Figure 2 shows that there were several data points where not all the information was reported. For example, 853 loans had no income and no income verified. Therefore, the data points with missing values were omitted. Thus, the base model for this transaction starts with just the following 12 variables: BorrowerCreditQuality, EmploymentStatus, Income, ExpectedMaturity, PrincipalBalance, DownPayment, LoanToValue, InterestRate, CarManufacturer, NewUsed, Vehicle Value and Customer Type.

DefaultLogisticModelIteration1 = glm(AccountStatus ~ BorrowerCreditQuality + EmploymentStatus + Income + ExpectedMaturity + PrincipalBalance + DownPayment + LoanToValue + InterestRate + CarManufacturer + NewUsed + VehicleValue, + CustomerType, data=defaultTrain, family=binomial).

Overall, the AIC was optimized because of the process of eliminating/adding several variables using backwards and forward selection. The 9th iteration of this transaction yields the best results. In the final iteration of the model the AIC was minimal with 1773.6. Although the Akaike Information Criterion value does not provide any meaning, a lower value is preferred over a higher AIC.

In the final model there are just three independent variables: ExpectedMaturity, PrincipalBalance and DownPayment. Adding the next best independent variable leads to an increase in the AIC. Therefore, the decision was made to stick with model iteration 9: It can be observed that this is based on the training data set of this transaction (70% of the data).

DefaultLogisticModelIteration9 = glm(AccountStatus ~ ExpectedMaturity + PrincipalBalance + DownPayment, data=defaultTrain, family=binomial).

Figure 3 shows the summary of the final iteration for this specific logistic regression model on this transaction. This example has the least number of explanatory variables in its final solution. Judging from the signs of the estimators there are two variables that have a negative impact on the log odds of default (and therefore on the probability of default) which are the expected maturity and the principal balance. On the contrary, increasing the value for the down payment amount is decreasing the probability of default. Please note the significance values of the variables. The intercept and DownPayment are significant with code (***) whereas for the ExpectedMaturity (0.05) and Principal Balance (0.01) this is less significant.

```
> summary(DefaultLogisticModel9)

Call:
glm(formula = AccountStatus ~ ExpectedMaturity + PrincipalBalance +
    DownPayment, family = binomial, data = defaultTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3596  -0.1571  -0.1163  -0.0746   3.8409

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.309e+00  4.490e-01 -11.825  < 2e-16 ***
ExpectedMaturity  1.318e-02  5.895e-03   2.236   0.0254 *
PrincipalBalance  2.552e-05  1.318e-05   1.936   0.0528 .
DownPayment      -2.122e-04  3.104e-05  -6.834 8.27e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1868.9  on 19151  degrees of freedom
Residual deviance: 1765.6  on 19148  degrees of freedom
AIC: 1773.6

Number of Fisher Scoring iterations: 9
```

Figure 3: Summary of the final model for this transaction

The findings for this Volkswagen Transaction are consistent with the Hypothesis for the variables in the transaction. The higher the principal balance is, the more a borrower has borrowed. A higher principal balance is associated with a higher risk of not making timely and full payments and therefore is expected to increase the PD (+). Table 1 shows the results of the final model with the direction of the parameter compared to the hypothesis. For this transaction the hypothesis was correct for all three parameters in the final model.

Table 1: Check results vs. hypothesis

Field numbe	Priority	TAG	Field Name in R	Field Name	Data type	Hypothesis: impact PD	Type Variable	In Final Model	Impact on PD	Significance	Hypothesis Correct
AA24	Mandatory	Static	ExpectedMaturity	Original Loan or lease term	Numeric	Longer term (+)	Loan	WAAR	(+)		WAAR
AA26	Mandatory	Static	PrincipalBalance	Original Principal balance	9(11).99	Higher balance (+)	Loan	WAAR	(+)		WAAR
AA34	Mandatory	Static	DownPayment	Down Payment amount	9(11).99	Higher down payment (-)	Loan	WAAR	(-)		WAAR
3	Mandatory	Static	3								
0	Optional	Dynamic	0								
3	Total	Total	3								

Interpreting the impact of the logistic regression model might be a bit hard, but by taking the exponent of the coefficients we get a better intuitive picture of the impact of altering 1 unit of the independent with thelog odds of default. Figure 4 shows that loans having a higher principal balance or maturity the probability of default is increasing the probability of default, where a higher down payment amount decreases the probability of default.

```
> exp(coef(DefaultLogisticModel9))
     (Intercept) ExpectedMaturity PrincipalBalance      DownPayment
      0.00494558       1.01326827       1.00002552       0.99978787
```

Figure 4: Results VarImp function

The VarImp function tells us which of the variables in the model is the most important. For this specific transaction this is clearly the DownPayment variable as the Figure 5 illustrates. The principal balance is the least important variable remaining in the final iteration of the model.

```
> varImp(DefaultLogisticModel9)
                  Overall
ExpectedMaturity 2.235895
PrincipalBalance 1.936399
DownPayment      6.833753
```

Figure 5: Results VarImp function

The final model is tested for multicollinearity using the vif function in R. Values above 10 indicate problematic values. The values of all three variables are acceptable according to the results in Figure 6. This indicates that there is no multicollinearity amongst the three variables of the expected maturity, principal balance, and down payment amount.

```
> vif(DefaultLogisticModel9)
ExpectedMaturity PrincipalBalance      DownPayment
        1.323546         1.185319         1.131335
```

Figure 6: Results multicollinearity check

Performance

The final model is required to be tested to see if it can correctly predict defaults from the set of loans in this Spanish transaction. To test this the model has been used on the training data and on the testing data. Remember that a 70/30 split between the training and testing data was made before the forward- and backward modelling of the logistic regression.

The performance of the final model of this transaction is shown in Figures 7 and 8. The results indicate that the model has some distinctive power, since the AUC values are above 0.50. This line would be wherethe false positive rate equals the true positive rate with the area being 0.50. The model was fitted on the training data and then used to model the predictions of defaults on the remaining test data. The graphs indicate that the model has some predictive power, since the threshold value of 0.5 has at least been exceeded but there is room for improvement.

This transaction might suffer from a relatively low number of loans compared to the independent variables (factor 1:2500). This imbalance in the data was not accounted for in the modelling of the transaction. The model clearly performs worse on the testing data than on the training data, which might be an indicator of overfitting the model on the training data. The model does have predictive power on both the training and the testing data.

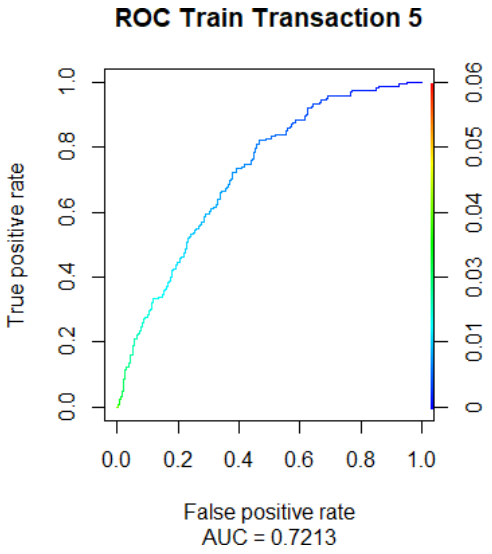

Figure 7: ROC/AUC model on training data.

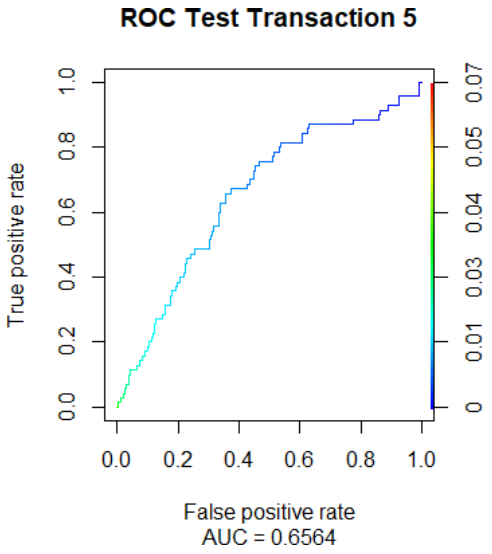

Figure 8: ROC/AUC model on test data.