How can machine learning help in Churn prediction for DeJong and Laan?

Author: Aryan Telang
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT,
This report is of a study conducted at the company DeJong and Laan and deals with applying machine learning to their customer data to predict customer churn. We begin this process by understanding churn exists in parts and is a dynamic presence in the customer journey. It is also a key performance indicator of a business's health and in this specific case, it is a function of customer satisfaction and loyalty as well as several constraints. The operationalisation of partial churn is done using several variables measuring demographics and the degree of recency, frequency, and significance[monetary] of customers' purchasing behaviour at the company. The created data set is run under a binary logistic regression model with customer exit acting as the Boolean variable. The attempt is to identify which customers are likely to leave to further strategize their retention.

Graduation Committee members:

First supervisor – dr. M.L. Ehrenhard (Michel)
Second supervisor – dr. M. de Visser (Matthias)

Keywords
Customer journey, partial churn, customer satisfaction, Key performance indicators, Logistic regression

1. INTRODUCTION

1.1 Outline

This chapter seeks to introduce the reader to the direction of this paper as well as provide them with relevant background information. It will also highlight the structure of the report and the objective behind making it. This chapter is further divided into six sections starting with the background information and focusing on the research question.

1.2 Academic Relevance

This report deals with the application of artificial intelligence technologies in a business discipline like marketing. With the advent of industry 4.0, a widespread need to incorporate the latest technology has been observed to ensure a business's ability to deliver high-quality products at a lower cost to safeguard its viability in the long run. This report specifically deals with churn- which refers to a customer no longer employing the services of a business. It is very relevant to the disciplines of business administration as it indicates the end of a customer journey and is an opportunity for a business to evaluate its service. Albeit the thesis is going to feature some technical aspects, its focus is set on one of marketing's most core concepts: The product and customer fit.

1.3 Basic concepts

1.3.1 Churn analysis

Customer churn or attrition is a measurement of the percentage of customers that cancel or choose not to renew their subscriptions at a business (Xie et al., 2009). Simply put, it refers to the state of departure from the customer journey and denotes the end of the business and customer relationship. Broadly speaking, churn is a function of the number of customers that exist and can be defined as follows (Campbell, 2020):

$$\frac{\text{Number of customers churned in a period}}{\text{Number of customers at the beginning of the period}} * 100$$

Churn analysis is a metric that provides an answer to the question "Are we losing customers?" It refers to the evaluation of the company's churn, and brand loyalty and is the study of the factors leading to churn with the natural intention of reducing it and maintaining some level of MRR [monthly recurring revenue] (Campbell, 2020).

1.3.2 Importance of Churn analysis

Churn analysis essentially is the study of customer retention and by extension, is the study of predicting demand along with future cash flows. For any company, acquiring new customers requires the expenditure of a lot of time and resources and it is in the company's interest to retain its existing customers to ensure and predict some level of demand (García et al., 2016). This cost of acquiring new customers and the incentive to achieve high rates of customer retention increases as we shift from B2C towards B2B and from product-based businesses to service-based businesses (Koeslag, 2016). Furthermore, given that solving a customer's problem is the function of a business, a customer leaving might indicate incompetence or dissatisfaction with the business's product and can signal a need for change. Essentially, churn analysis indicates possible vulnerabilities of the product. Therefore, a customer's exit from a business serves as an evaluation of their product and by extension their USP (unique selling point). A business's entire value also stems from the output it can generate; a high churn rate indicates a need to innovate. It leads to better understanding of customer segmentation, higher profits by providing enhanced customer experience.

What we can infer from citing the importance of churn analysis is that churn can be said to be a KPI (key performance indicator) for the overall health of the business and can be a tool used for growth (Tueanrat et al., 2021). Churn analysis deals with customer product fit which could help define the customer segment that is best suited for the business and can yield the most growth. Also in the financial services industry, for a company like DeJong and Laan which features a wide product portfolio with a lot of B2-B services, customer retention is of prime importance.

1.3.3 Machine learning and data mining

Machine learning and data mining are application practices for the large amounts of data which can be gathered today. A business has considerable amounts of data about its customers which could potentially generate a lot of value. The utilization of this data to generate valuable insights are broadly labelled as data mining (Hung et al., 2006). More specifically it involves analyzing large amounts of data for discovering patterns by applying relevant algorithms. Machine learning is a branch of artificial intelligence (AI) and one such process of data mining which focuses on using the data to imitate the way that humans learn, aiming for a gradual increase in accuracy. In this thesis, I will attempt to implement a model that conducts ML on DeJong and Laan's customer data for churn prediction. Fundamentally, a logistic regression model will be used which when computed by a powerful ML system can produce advanced results.

1.4 Description of the situation at De Jong and Laan.

DeJong and Laan is a privately owned accountancy firm based in the Netherlands, headquartered in Vroomshop. They operate in the financial administration industry and are a large firm with between 500 – 1000 employees. They are accountants, financial operators, tax advisers and so on. They have 23 branches from Groningen to Elst and from Harderwijk, via Apeldoorn, to Enschede.

DeJong and Laan has a wide range of clients with a focus on B2B services. Their products can be categorized as premium indicating high quality at a higher price point. Also considering that it is a service-based company i.e. low marginal costs, whose products can be consumed periodically, it stands to reason that it is in the business's interest to retain its existing customers. Literature tells us that the significance of churn prediction is directly proportional to the marginal value brought by a customer. Customer acquisition in a premium segment is difficult and resource-consuming. Churn is a very relevant metric for a business that seeks to retain its existing customers to be viable. For a company like De Jong & Laan, being able to predict that a customer is looking to leave, can have a significant impact on profits. Through this study, they seek possible insights into their customer churn rate based on the data they have already collected. They can potentially launch investigations into the cases of high value clients to strategize their retention and discern whether there is an underlying problem in their service which is causing the churning. This iterates the importance of predicting churn for them as it serves as a measure of the quality of their product service and whether they can satisfy their customers. Ultimately should help them deliver better products.

1.5 Research objective

The research objective of the thesis is to identify key metrics and performance indicators that could signal churning in customers of DeJong and Laan. Research will be done in analysing churn, operationalising the previously mentioned performance indicators, and modelling them for their prediction.

1.6 Research Question

How can machine learning help in Churn prediction for DeJong and Laan?

1. What are the Key performance indicators that highlight customer loyalty?
2. What are the different kinds of churn and their significance?
3. How do you measure customer satisfaction?
4. How to operationalise customer loyalty and predict churn?
5. How can machine learning be applied to customer data?

1.7 Structure of the report

This section of the report served to provide the reader with the necessary background information to proceed. The next section features the analysis of theoretical resources to break down what churn is and how it can be measured. Concepts of customer satisfaction and loyalty will be introduced which will then be operationalized in the next section of methodology. The Methodology will also feature a description of the parameters i.e., the dataset that will be modelled as well as the model selection and the machine learning code.

2. THEORETICAL FRAMEWORK

2.1 Outline

This chapter explores academic literature surrounding the topic of churn. The search strategy is to predominantly find papers via google scholar to ensure the reliability of information. Initially, the strategy for searching literature will be mentioned after which relevant theories will be mentioned and inferences will be drawn. We start with discussing customer journey and what partial churn is, after that, we will analyze different types of churn and customers and their implications to a business- like DeJong and Laan. A section will also feature some analysis of customer loyalty and expectations and another section will discuss churn specific to the accountancy industry.

2.1.1 Customer journey and partial churn

Based on the available literature on churn, we can comment that is not a static occurrence but is a dynamic presence throughout the customer journey. Customer journey is defined as the path of interactions an individual has with a business including all cognitive, affective, sensory, and behavioural consumer responses during all stages of consumption (Lemon & Verhoef, 2016). Simply put, it refers to the entirety of the customer experience with the business. Understanding the definition of the customer journey is essential to this report as it demonstrates that an *exit* or *churning* does not exist outside a customer's journey with a business but instead is a part of it (Tueanrat et al., 2021). Furthermore, mapping the customer journey will help indicate the stage that features early signs of churning which is the focus. See figure 1, Adapted from "Going on a journey: A review of the customer journey literature" by (Tueanrat et al., 2021)

Analysis of the customer's journey map yields some interesting results. A customer's engagement with the business is volatile and follows some cyclical consumption trends. It can be divided into many phases; it begins with customer acquisition after which a customer is actively engaging with the business's products (Tueanrat et al., 2021). After that, the amount of engagement and utility decreases. This point can be recognised as a junction where a customer will either re-engage with the business and become an active customer again or will exit the customer journey or churn. This helps iterate the fact that on most occasions, churn is not an independent event but is a function of the events subject to the customer till it reaches the said junction. We can refer to this phenomenon as partial churn. Partial churn, therefore, refers to the state where a customer is on the path to churning and not on the path of re-engagement. It can be inferred that a customer is exhibiting lower levels of engagement than ideal when they are an active customer so operationalising partial churn and measuring it will be the key when looking to measure churn (Miguéis et al., 2012).

2.2 CLASSIFICATION OF CHURN

In this section, we will use literature to classify churn and analyze whether all types of churn are equally relevant to a business. The purpose of this is to make inferences surrounding which aspects of churn are significant when operationalizing it later in the report. Churn can broadly be categorized into the following categories. Involuntary and voluntary: Involuntary or Delinquent churn

Involuntary refers to a situation where a customer exits a business due to circumstances beyond their control (Kirui et al., 2013). Literature suggests a myriad of causes including the shutting down of a business, death, and various other reasons. Predicting involuntary churn is beyond the scope of this report and since retaining customers in such a situation is not feasible, we can deem involuntary churn as not significant. For the

purposes of this report, involuntary churn is not relevant as predicting it is beyond the scope of the report and retaining the customers is not feasible. Delinquent churn is a form of involuntary churn where a customer is lost because of a trivial matter surrounding the payment method. This may include card expirations, processor problems and so on. Given that the subject of the report is an accountancy firm, delinquent churn is not relevant either. Voluntary churn refers to a situation where a customer actively chooses to disengage from the services of a business (Kirui et al., 2013). Voluntary churn usually refers to when a customer either no longer requires the service being provided or when they switch to a competitor. The important thing to note is that in this situation is that the customer still has more utility to the business. Efforts in prediction analysis of churn lost to a competitor and its prevention is likely to lead to customer re-engagement.

Beyond the broad categories of voluntary and involuntary churn, churn can also be divided based on measurement i.e. in terms of revenue (Koeslag, 2016). This is known as revenue churn. There may be a circumstance where a business is retaining exiting customers, but the revenue is decreasing which could indicate low engagement or high levels of partial churn or even demonstrate the need to upgrade the product portfolio. We can infer that revenue churn is perhaps a more significant metric to a business as it not only indicates partial churn but also demonstrates a loss in the Monthly Recurring Revenue which is crucial to a business's health. Equally relevant is classifying customers. DeJong and Laan has a wide range of customers that exists in both the B2B and B2C segments. It stands to reason that when classifying customers, to account for their value to the business. Given that the company's USP is premium services accompanied by a higher price point, the B2B segment is more valuable to it. It is important to note that the value of churn is equally high as the value of customer loyalty. The extent of customer classification is limited to this in this sub-section as it was to iterate the importance of accounting for revenue achieved from a customer when operationalizing partial churn whilst optimizing the model for value.

2.3 Churn in the accountancy industry

As inferred, delinquent churn does not exist in the accountancy industry given that issuers of financial services have control of their client's finances. According to research, the most common causes for churn in the accountancy industry are linked to three causes: a problem with the product, a change in the customer business or the advent of a capable competitor (*IBISWorld - Industry Market Research, Reports, and Statistics*, 2020). In the first scenario, a customer is either unhappy with the quality or price of the service. In such a situation, a customer demands more utility from the accountant where a negative sentiment may have risen under the perception of having been profited from. Another situation that causes churn is when the client business is negatively impacted by the pandemic economy or is he relocating. Another situation is where a rival firm is employing newer technologies and delivering a better product or a similar product at a lower price point. In both of those scenarios, a need to re-engage the customer can be seen. The inferences that can be drawn in this section have to do with customer perception. Customer perception refers to the cognitive response of the customer surrounding your brand and it deals with their emotions, opinions, feelings, and beliefs. Research suggests that building awareness during the initial stages of the customer journey is the most significant when affecting perception (*IBISWorld - Industry Market Research, Reports, and Statistics*, 2020). Also, customers perceive a large array of emotions when interacting with a business, customers that feel pleased and in control are more likely to remain loyal to the business. This helps us in making some inferences about the accountancy industry. Accountancy is a very technical and tedious subject which is the first barrier as it is hard to understand. Unfortunately, it must also deal with having negative certain connotations like other money-focused industries. Also, customers do not generally have a lot of control in this industry.

The purpose of this subsection is to analyse the reasons for churn to help in the operationalisation of partial churn as well as in the strategizing of churn prevention. In the financial services industry, utility is of importance to customers and customer relationship management is of great significance. The sentiment around accountancy is not very positive and service differentiation is difficult. In such a situation, the relationship between the customer and business can prove the factor that affects churn the most. This will be further explored in the next subsection.

2.4 Analysis of customer satisfaction and loyalty

In the literature review till this point, we have noted a need to focus on customer engagement to ensure valuable churn prevention. Customer engagement is a function of satisfaction and loyalty and is a potential indicator for churn which needs to be analysed (Tueanrat et al., 2021). A lot of Researchers like Meyer and Schwager have highlighted how customers are continuously adjusting and evolving their perceptions of a brand at every point throughout the journey and that increasingly increased saturated markets have made satisfaction and loyalty harder to achieve (Koeslag, 2016). The competition for customer acquisition grows fiercer as technology innovates and the marginal costs decrease, making services more scalable.

Customer satisfaction can be conceptualised as the alignment between the service delivered and the customer's expectations (Tueanrat et al., 2021). It stands to reason that the customers are satisfied when their expected utility of a product or service is achieved. Research has established that maintaining a high level of customer satisfaction is a critical indicator of the company's health and churn. Furthermore, it can be inferred that the gap between customer expectations and satisfaction is what leads to churn and a business should strive to minimise it to enhance its competitive advantage. A reference to the previous analysis of customers is important as customers gain satisfaction because of the various interactions throughout the whole journey process and this further iterates the point of importance of Customer Relationship Management in accountancy (Ahn, J et al., 2006). It is important to note that firstly the scope of relationship management goes beyond addressing complaints. The business must maintain a good and a non-burdensome customer experience, especially in accounting. Secondly, if we analyse a situation where there is a complaint with a product, we can observe a momentary increase in customer expectations which increases the gap between expectations and experience which in turn relates to churn. It is also important to note that while variables like customer satisfaction and experience are an indicator of the business's performance, they are also a function of the customer and their relationship with the company. We can use this to further classify customers.

2.4.1 Classification of customers

Research suggests that customers can be classified based on the extent of their loyalty versus satisfaction (Koeslag, 2016). See table 1, here we can see the classification of customers based on satisfaction versus loyalty. The simple purpose of demonstrating varying degrees of loyalty is twofold- firstly to again highlight

the importance of customer satisfaction and secondly to link these abstract concepts to output. We can infer that a customer with higher levels of loyalty is likely to give the business more value and the opportunity to bridge a momentary gap between their experience and expectation, meaning that partial churn could theoretically be said to have begun even in the first stage of the customer journey- a point which will be explored in the methodology section.

2.5 Analysis of customer purchasing behavior

We have discerned that in order to maximize churn prediction for a company like DeJong and Laan i.e. a premium B2B financial services company, we must focus on voluntary churn while emphasizing on revenue. This helps set the direction for the methodology which is to be carried out in section 3. A focus on voluntary churn essentially means a focus on the decisions of a customer. Thus, by analyzing the purchasing behavior of a customer, we can gather insight into their level of loyalty which is a function of customer engagement throughout the customer journey. As is already established, we are looking to measure the levels of partial churn throughout the customer journey and hence must measure the purchasing behavior of customers against time. This means that the predictor variables for partial churn will focus around the monetary, recency and frequency of the customers' purchasing habits.

2.6 Antecedents to customer loyalty

This study aims to identify certain customer traits that can indicate the extent of loyalty. Before we begin the methodology, it will prove useful to establish a qualitative analysis of the antecedents of customer loyalty. According to the literature available online, these antecedents are broadly of two types: situational and demographic (Ahn, J et al., 2006)..

1. Demographic factors

Demographic factors refer to those that help us measure differences between customers based on their various attributes. The purpose of these variables is to gather insight upon the value of each specific client for the purposes of analysing revenue churn. These attributes affect the relationship between a customer and the business as well as the customer's purchasing habits. We have already analysed the customer in favour of DeJong and Laan but an emphasis of purchasing habits can be noted. Demographic factors are easy to identify and measure and are often correlated with service usage and are thus of significance.

2. Situational factors

Situational factors are events that can bring about a change. These can include a variety of things including the recent onset of inflation because of the pandemic, technological changes, incentives from competitors, switching costs and so on. Out of these factors, the barrier to entry and exit is an important antecedent to loyalty. For a customer, switching would require time and effort. This leads to the argument that loyalty can be the result of not only a strong customer-business relationship but can be a result of constraints. We can infer that the customer's loyalty is a result of the extent of customer satisfaction along with certain situational and demographic factors.

3. METHODOLODY

This section of the report deals with the methodology and will feature the application of the previously conducted theoretical research. Initially, a deductive approach will be used where the theoretical framework will be used to identify relevant topics of interest to operationalise customer satisfaction and partial churn independent of the data available at DeJong and Laan. Next, an inductive approach will be used where practical operationalisation will be done in accordance with what data is available. The attempt will be to create a functional data set using the inferences form the deductive approach. Next will be some data analysis will be done and model for ML will be chosen and carried out.

3.1 Deductive approach.

As mentioned in the outline, the purpose of this subsection is to operationalise customer satisfaction and partial churn using the theoretical framework to draw inferences. Note that this will be done independent of the data available at DeJong and Laan. In the theoretical framework, we initially noted that churn is a part of the customer journey and can be predicted by detecting partial churn. We then noted that partial churn is a result of lower levels of customer engagement. We then deduced that lower levels of customer engagement are a function of expectations versus delivery and can be measured as customer satisfaction along with situational and demographic factors. Hence, to predict churn- we will have to operationalise customer satisfaction and measure the contextual factors.

Before we operationalise customer satisfaction, it is important to consider DeJong and Laan and the financial services industry. We noted that for a company like DeJong and Laan, not all kinds of churn are relevant. Revenue churn- which is a result of the customer voluntarily choosing to switch from the company's services is of the utmost importance. This highlights the need to measure variables indicating the purchasing behaviour along with monetary influence (Miguéis et al., 2012). Another very important distinction in churn is customer re-engagement because of loyalty and customer re-engagement because of constraints. In the accountancy industry, customer's relationship with the company bounds clients to remain loyal as a result of constraints. This is due to the costs that will be incurred to switch and will be directly proportional to length of the relationship the customer has had with DeJong and Laan, recency and how frequently it avails its services. It will also be influenced by the employees that are responsible for the relation. Also, some contextual variables are of significance including the type of product being consumed, industry, region, company size and other relevant demographic factors. Thus, predicting churn will involve the measurement of variables that measure four kinds of parameters: Frequency, Recency, Monetary and demographic. Note that buying behaviour is being measured in frequency (Miguéis et al., 2012)

.

Here is a possible list of variables that can be measured (Tueanrat et al., 2021), (Koeslag, 2016).

:

1. Parameters relating to Frequency
    1. Number of transactions of the given period
    2. Number of transactions in the 6 months last
    3. Calibration period and related variables
    4. Relative change in the number of transactions or calibration period
    5. Length of the relationship
2. Parameters relating to Recency
    1. Number of days since last purchase
    2. Number of days between transactions

3. Standard deviation of the inter-purchase time

3. Parameters relating to Monetary
    1. Total amount of monetary spending
    2. Monetary spending in the last six months
    3. Relative change in the monetary spending pattern
4. Parameters relating to demography
    1. Location of customer
    2. Industry size
    3. Company size, in terms of revenue
    4. Company size in terms of growth potential

3.2 Inductive approach.

3.2.1 Data Selection and acquisition

The first step to the inductive methodology is to gather data (Hung et al., 2006). The data is obtained from the consolidated Power-BI data system at DeJong and Laan. For the data selection, the parameters linked to churn have already been identified in the deductive approach. Variables will be selected based on their influence on partial churn. These variables could be of demographic, monetary, recency and frequency nature. Relevant variables will be selected and exported for churn analysis. The limitations with the data will be discussed later in the report.

See table 2, displaying the selected data set. This comprises a list of predictor variables gathered from DeJong and Laan's data system. The first variable is Account_name or the name of the customer which will serve as an identification variable. The next two variables are account_type and Client_size which will be used in the next subsection to select high value customers. The other predictor variables include Project_count and registered_hours measuring the length and degree of purchasing behavior of customers. Also being measured is the sales_price which indicates the monetary significance of the customers' purchasing behaviour. The next variable being measured is the length of the relationship as well as invoice status to establish the recency of the customer-business relationship. The next variables are of general category with the ability to influence purchasing behavior and these include the industry of the client and the type of service purchased.

3.2.2 Data Preparation

Before the data set can be operated upon, the data needs to be prepared in a particular fashion which involves further selection and cleaning (Hung et al., 2006). This preparation was done using the filter function in Power-Bi as well as SPSS. The steps taken are as follows:

1. All parameters with null values for variables need to be cleaned except for the variables: Afmeldingsreden/ Reason_churn, Msdyn_sales price, Count_projecten and Gescheven_uren as their null values indicate that churn has occurred and is not a sign of data impurity.
2. All variables must also be converted to scalar or ordinal for processing.
3. Churn is being measured by variable relatie_type which is converted into a Boolean variable i.e values are on or off. Here customers are classified as either an active customer or Klant or an ex-Klant signalling that the customer has churned.
4. As was covered in the theoretical framework, the report is focusing on modelling churn for high value customers of DeJong and Laan- hence, the account_type is filtered as businesses and the Klant_grootte or client size is filtered either as medium or large, in order to focus on high value customers.

3.2.3 Inferential statistics

Using SPSS, inferential data analysis will be performed on the forged data. The purpose of this is to gain a fundamental understanding of the nature of the data and to perform exploratory analysis. Using Pearson's correlation, an initial assessment of the dataset becomes possible, a high correlation could indicate a higher level of health. Also, logistic regression(explained in the next subsection) can be carried out without incorporating any machine learning to highlight the difference in results

3.2.4 Machine learning application

The identification of the appropriate Machine learning technique depends on the goal of the study. The most fundamental ML models are of two types: Association and classification (Kirui et al., 2013). Association involves investigating the relationship between items and classification involves predicting future behaviour through classifying database records. Given that we are trying to predict churn, or classify customers as potential churners, we must use a classification ML model (Kirui et al., 2013). Furthermore, there are techniques within this function which include:

1. Logistic regression
2. Neural networks
3. Decision tree
4. Genetic algorithm
5. K-nearest neighbor

Given that the output variable is a Boolean, the most sensible approach will be to use binary logistic regression (Kirui et al., 2013).

$$P(Y) = \frac{1}{1 + e - (b0 + b1x1i + b2x2i)}$$

Here, P(Y) is the probability that the event occurs. Y- the dependant variable which is to occur. x1, x2 are the dependant variables. So P(Y) is the probability that a customer will churn, Y is the variable which indicates either a customer has churned or not and the x variables are the parameters that are being operationalised in order to predict churn. Note that Y is a Boolean variable with the values (0,1) where 0 indicates customer loyalty of the highest level and as the value tends to one it indicates that a customer will exit DeJong and Laan.

3.2.5 The code

See figure 2 in the appendix, the machine learning is being conducted by Googles TensorFlow 2.0 using the Keras API powered by Python(being coded in IDLE- development environment for python). Here, tensorflow refers to the Artificial intelligence module developed by google which can be coded using an API called Keras. The first step is to load the data set onto IDLE. This can be done using the pandas framework which can read csv files. Next the dataset is prepared by setting the delimiter. The next step would be to format the dataset according to a 0.1f float format and then convert it into a single Numpy array. This step basically converts the data set into a single readable array for the model. The next step is to divide the array into two: the predictors and the output variable denoted as X and Y respectively. The next step is to perform machine learning.

This is done importing the sequential module from keras which allows for association models like logistic regression. This model can be complied with the Dense module from tensorflow layers. Dense allows for neural networking where we have called 1 neuron and allowed for loss calculation and optimization as well given a command to measure accuracy. The final step is to run the model for a certain epoch count i.e. run the model a set number of times with each iteration of the model learning from the previous one.

4. RESULTS

4.1 Inferential statistics

Before the machine learning results are analyzed, some exploratory data analysis can be performed on the available data. Based on the current literature available, measuring Pearson's correlation appeared to be the wisest choice. This can be viewed in the table in Figure 3 in the appendix for all fifty-seven thousand customers. Any significant relationships are highlighted by "**".

The purpose of measuring Pearson's correlation is to do a preliminary assessment of the dataset's health. We can note that with respect to the output variable churn, the pearson's correlation coefficient is significant in the relationships with Registered_hours at 0.02, Reason_churn at -0.87 and Invoice at 0.02. This preliminary analysis shows that this dataset is fairly healthy and has the potential to yield exciting results. The relationship of other variables is not as significant to churning as can be seen in the table.

Also featured in the appendix, in figure 4 and figure 5, is a basic logistic regression being run on the data set [Block wise method] using SPSS. The purpose behind this is to highlight the difference in results after implementing machine learning and neural networks. This will also help iterate the significance of machine learning as we will also be conducting binary logistic regression then only with neural networks. As far as this logistic regression is concerned, the results can be seen in the output where the significance levels for the regression tend to 0 indicating poor health of the model. High standard error can also be noted for the variables, especially for reason_churn and invoice_status which were of significance according to pearson's correlation.

In conclusion, as can be seen throughout a lot of literature surrounding big data, performing inferential statistics does not necessarily yield significant results. This also is due to the fact that not all variables are scalar, this can be seen in the high standard error or lack of correlation coefficient. This highlights the need for machine learning, a model that is able to imitate human thinking can find links in the data which simple descriptive statistics cannot.

4.2 Machine learning results

As stated in the code explanation, the model as an epoch value which makes it run for a specific number of times. With each iteration it performs neural networks to learn and produces higher accuracy and lower loss. This can be seen in Figure 6 in the appendix. The code is running 256 times which is the industry practice and by the 256th iteration its producing results with 100% accuracy.

Here, the model trains and applies itself within the same dataset. It trains from the cases where the output variable is 0, indicating that the customer has churned and performs the regression on the cases where the output variable is 1, i.e. the customer is still present with the company. The output the model produces are predictions which refer to the probability of the instance. What this means is that the model gives a probability score to each case. For customers that have churned, the precicted values should tend to 0, this can be seen in figure 7 in the appendix. Here the 'Y' array shows the output variable in the data set- which are all 0s given that this section was used to train the data and the 'predictions' array is displaying the probabilities of these customers having partial churned which is also all 0s indicating that the model trained.

Similarly values of customers that have not churned can be displayed against their partial churn score, which is being generated by the model, this can be seen in figures 8 and 9. Figure 8 shows the data set values and figure 9 shows the corresponding predicted values. In figure 9, we can see that most values tend to 1 but there are also values of 0.6 denoting that that particular customer is showing high levels of partial churn.

4.3 Validity of results

In order to ascertain whether the results are of any real significance, the model can be extrapolated on a new data set. The new dataset contains 100 cases, 50 churners and 50 non churners. As the present model is extrapolated on the new data set, it measures an accuracy of 53.1%. This was measures using a confusion matrix, see in appendix. What this means is that, if we treat the initial big data set as a training data set and run the model on the new sample data set, it has an accuracy of 53.10%. This accuracy can increase if machine learning is performed again on the model. This means that the model is not yielding highly significant results in terms of accuracy indicating that more data is required to perform this analysis. The data is however significantly precise to some extent that it can be used by the company to start investigations into certain clients. What we can also note by the results is that the model is significantly more adept at calculating low partial churn scores i.e., churners than it is at predicting non-churners.

5. DISCUSSION

5.1 Academic relevance

Most of the literature available online, deals with churn in a B2C setting in the secondary sector. As we shift towards a B2B setting, in the tertiary sector- we can note that a change is require in churn prediction models. This report also deals with churn in the financial industry thus eliminating delinquent churn which is a huge parameter in other industries. It also highlights the importance of measuring factors that influence churn outside of customer loyalty highlighting the impact of opportunity cost and constraints for a customer. For future research, constraint driven churn is a very relevant topic. This is a very evident and significant research gap. Churn as a function of opportunity loss is potentially a very significant research gap particularly in the accounting industry. This research could also be extended in analyzing labor turnover and its influence on churn given the significance of customer relationships. Because of new technologies like neural networks, it should become possible to perform datamining on predictor variables that have only a correlation and not causal relationship with the output variable.

Equally significantly, the report can also signal those new technologies like AI and machine learning have a huge application in business disciplines. Essentially, there is a solution for situations where a seemingly qualitative analysis would be required but is not feasible.

5.2 Practical implications

As the study was being performed at DeJong and Laan the results and inferences are specific to their organization. The purpose of

the study was to perform churn analysis to determine which customers at DeJong and Laan are potentially exiting and possible measures that can be taken to avoid it. Using the model developed according to the literature in this report, the company can generate the partial churning score of their existing customer as well as for new customers once their data becomes available. The model generates an accuracy score and is not fixed to the predictors meaning that the company can also try different datasets with different predictor variables in a quest for greater accuracy. Using the partial churning score they can measure lower levels of engagement which can be the starting point for investigations for product innovation.

Also, as we noted that in the financial industry, loyalty because of constraints is very significant. This highlights the importance of customer relationship management. More significantly for DeJong and Laan, it shows an opportunity to retain clients. Given that the business is a top 10 accountancy firm- higher considering most high-level firms do not target SMEs, creating a customer environment can heavily influence the level of constraints on a customer. The reason for negative sentiments against the financial services is a result of the highly technical nature of the business that bars outsiders from understanding it along with the notion that they are being charged for every little activity for an outsourced function. Creating a customer environment like amazon with prime* can certainly boost the business. Subscription models, recurring and free consultations are possible methods of creating a customer environment. This along with networking opportunities amongst SMEs should aid in increasing constraints

5.3 Limitations

There are several significant limitations to this report. It is important to realize that data is not readily available. Even for the theoretical framework, owing to the proprietary nature of customer data, not a lot of literature is available with respect to B2B businesses. Also, a lot of data at DeJong and Laan is segmented and cannot be compiled into a coherent dataset. Also, a very basic ML model was applied in this thesis, there are more complex machine learning techniques like clustering, forecasting and sequence discovery which could potentially yield exciting results but are beyond the current scope. Logistic regression estimated the probability of an event, like churning, occurring- based on a given data set of independent variables. Also In the output, we can note that as we progress through the epoch, i.e., that the number of times the model is run increases, the accuracy increases as well. This can very well indicate overfitting data. Another limitation is that involuntary churn is not being measured, situational factors are beyond the scope of this report. Every business has been hit by the pandemic and is being influenced by the promotional marketing of competitors. Measuring and operationalizing these variables is significant and can be a focus for future research as it has a direct influence on expectations and satisfaction levels in customers.

5.4 Ethical concerns

Ethics is a very broad and nuanced discipline which unfortunately does not yield very clear and definitive results. Analysis of ethical dilemmas generally have controversial results. When considering the ethical implications of machine learning, there are quite a few concerns. These concerns spread from discrimination to unintended consequences (Lo Piano, S., 2020). Equally valuable for society, are the significant opportunities that can be brought by technological advancements. As with any ethical dilemma, one must weigh the advantages and disadvantages of undertaking the said action. The concern here is about regulating innovation to ensure responsible technological practices that do not infringe societal rights.

First and foremost, the primary concern would arise surrounding the privacy of the individuals whose data is being analyzed. The standing requirement would be explicit permission when using personal data. From a legal perspective, an individual's right to privacy- safeguarded by the GDPR (within the EU)- is not breached when implementing machine learning although a need for newer guidelines can be theorized (Kritikos, M, 2020). Ethically, one can argue against analyzing large amounts of customer data in favor of a business and whether it is responsible and right to do so. One can argue in favor of this by stating that there is a need for innovation, growth, and employment, as well as the business's responsibility to deliver up-to-date products to its customers and argue against this motion by bringing up possible security issues, discrimination, and exploitation (Lo Piano, S., 2020). Within the scope of this report which will feature rather basic models, no such concerns are relevant

6. CONCLUSION

In conclusion, churn is a very significant metric that is essential to a business. It indicates the overall level of the business's health and by extension measures its ability to sustain itself. For this reason, churn prediction is of vital importance to a business. To predict churn, one must predict the levels of partial churn that occur in a customer through the customer journey as they experience lower levels of engagement. This iterates the fact that churn is not a definitive occurrence but occurs in stages and can be combatted in stages, throughout the customer journey. Measuring partial churn can be done by measuring differences in purchasing behaviour of customers especially pertaining to the frequency and monetary aspects of their demand. These factors along with various general variables affect the relationship of a business with its client. Using these variables, this report features machine learning being conducted to predict the levels of partial churn on the customers of Dejong and laan. We can observe that the quality of data available at DeJong and Laan is not satisfactory, and the accuracy of the results are sub-par although this highlights scope for future research.

7. REFERENCES

1. Kritikos, M. (2020). The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. Retrieved 13 May 2022, from https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf

2. Lo Piano, S. (2020)Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanit Soc Sci Commun* **7,** 9 (2020). https://doi.org/10.1057/s41599-020-0501-9

3. Ahn, J.-H., Han, S.-P., & Lee, Y.-S. (2006, October 16). *Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean Mobile Telecommunications Service Industry*. Telecommunications Policy. Retrieved May 14, 2022, from https://www.sciencedirect.com/science/article/pii/S0308596106000760

4. Miguéis, V. L., Van den Poel, D., Camanho, A. S., & Falcão e Cunha, J. (2012). Modeling partial customer churn: On the value of first product-category purchase sequences. Expert Systems with Applications, 39(12),

11250–11256.
https://doi.org/10.1016/j.eswa.2012.03.073

5. Hung, S.-Y., Yen, D. C., & Wang, H.-Y. (2006). Applying data mining to telecom churn management. Expert Systems with Applications, 31(3), 515–524. https://doi.org/10.1016/j.eswa.2005.09.080

6. García, D. L., Nebot, À., & Vellido, A. (2016). Intelligent data analysis approaches to churn as a business problem: a survey. Knowledge and Information Systems, 51(3), 719–774. https://doi.org/10.1007/s10115-016-0995-z

7. Campbell, P. (2020, May 27). How to Calculate Churn Rate: 4 Formulas For Calculating Churn. Www.profitwell.com. https://www.profitwell.com/customer-churn/calculate-churn-rate

8. IBISWorld - Industry Market Research, Reports, and Statistics. (2020, March 15). Www.ibisworld.com. https://www.ibisworld.com/netherlands/industry-statistics/accounting-auditing/3880/

9. Tueanrat, Y., Papagiannidis, S., & Alamanos, E. (2021). Going on a journey: A review of the customer journey literature. Journal of Business Research, 125, 336–353. https://doi.org/10.1016/j.jbusres.2020.12.028

10. Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. Expert Systems with Applications, 36(3), 5445–5449. https://doi.org/10.1016/j.eswa.2008.06.121

11. Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. Ch. (2015b). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, 1–9. https://doi.org/10.1016/j.simpat.2015.03.003

12. Koeslag, S. (2016). PREDICTION OF PARTIAL CHURNERS AND BEHAVIOURAL LOYAL CUSTOMERS THROUGH BEHAVIOURAL HISTORICAL CUSTOMER DATA. University of Twente.https://essay.utwente.nl/69651/1/Stef%20Koeslag_BA_BEHAVIOURAL%20MANAGEMENT%20AND%20SOCIAL%20SCIENCES.pdf

13. Lemon, K., & Verhoef, P. (2016). Understanding Customer Experience Throughout the Customer Journey. Journal of Marketing; Sage journals. https://journals.sagepub.com/doi/full/10.1509/jm.15.0420?casa_token=CEKnAEfq2SQAAAAA%3AoQ_7Uu_nCfZfY2OPh2MQ3wERYJE38ujYKqruxiA2Lp-LGbG6Mf0yt2wzvc-s1WTshq7hcwiAbXZFEEA

14. Kirui, C., Li, H., Cheruiyot, W., & Kirui, H. (2013). Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining. http://didawikinf.di.unipi.it/lib/exe/fetch.php/dm/churn-mobilephone.pdf

8. APPENDIX

1.

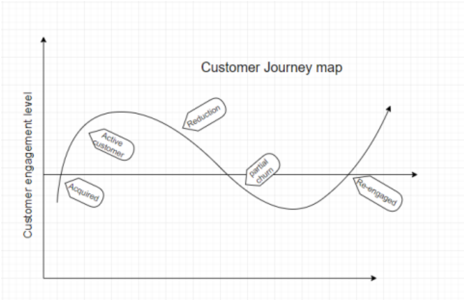

Figure 1. Customer journey map

2.

Customer Satisfaction			
		High	Low
Customer loyalty	High	Loyal customer	Hostage
	Low	Mercenary	Defector

Table 1. Customer classification.

3.

No.	Powe-BI field*	Variable	Description	Category
A.	Project account			
	1.	Account_name	Name of the customer	General
	2.	Account_type	Classification of customer as either a business or a person	General
	3.	Relatie_type*	Churn indication, relationship is either current or	CHURN
	4.	Afmeldingsreden/ Reason_churn	Reason to churn*	-
	5.	Relatie_sinds	Length of the relationship between the customer and DeJong and Laan	Recency
	6.	Klant_grootte	Size of the customer.	Monetary
	7.	Count_pojecten	Number of projects undertaken by the customer	Monetary
B.	SBI codes			
	8.	Letter	Industry of the customer	General/constraints
C.	Project goedkuringnen			
	9.	Geschevern_uren	Registered hours	Frequency
	10.	Msdyn_salesprice	Sales price	Monetary
D.	Taken			
	11.	Project_omschrijving	The type of project	General
E.	Werkelijke warden			
	12.	Facturering_Status	The status of the invoice	Recency

Table 2. Operational data set created using DeJong and Laan's data.

4.

Figure 2. Machine learning code in Python IDLE.

5.

		Churn	Count_project	Registered_hours	Reason_churnsc	Invoicesc	Sbi_sc	typesc
Churn	Pearson Correlation	1	.006	.020**	-.872**	.020**	.002	-.003
	Sig. (1-tailed)		.075	<.001	.000	<.001	.278	.213
	N	57795	57795	57795	57795	57795	57795	57795
Count_project	Pearson Correlation	.006	1	.084**	-.014**	-.024**	-.044**	.019**
	Sig. (1-tailed)	.075		<.001	<.001	<.001	<.001	<.001
	N	57795	57795	57795	57795	57795	57795	57795
Registered_hours	Pearson Correlation	.020**	.084**	1	-.017**	-.002	-.001	.211**
	Sig. (1-tailed)	<.001	<.001		<.001	.342	.432	.000
	N	57795	57795	57795	57795	57795	57795	57795
Reason_churnsc	Pearson Correlation	-.872**	-.014**	-.017**	1	-.010**	.001	.004
	Sig. (1-tailed)	.000	<.001	<.001		.006	.366	.153
	N	57795	57795	57795	57795	57795	57795	57795
Invoicesc	Pearson Correlation	.020**	-.024**	-.002	-.010**	1	.007*	.001
	Sig. (1-tailed)	<.001	<.001	.342	.006		.043	.413
	N	57795	57795	57795	57795	57795	57795	57795
Sbi_sc	Pearson Correlation	.002	-.044**	-.001	.001	.007*	1	-.007
	Sig. (1-tailed)	.278	<.001	.432	.366	.043		.059
	N	57795	57795	57795	57795	57795	57795	57795
typesc	Pearson Correlation	-.003	.019**	.211**	.004	.001	-.007	1
	Sig. (1-tailed)	.213	<.001	.000	.153	.413	.059	
	N	57795	57795	57795	57795	57795	57795	57795

**. Correlation is significant at the 0.01 level (1-tailed).

*. Correlation is significant at the 0.05 level (1-tailed).

Figure 3. Pearson's correlation.

6.

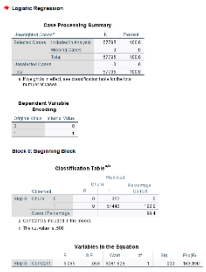

Figure 4. Logistic regression.

7.

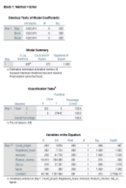

Figure 5. Logistic regression [Blockwise method].

8.

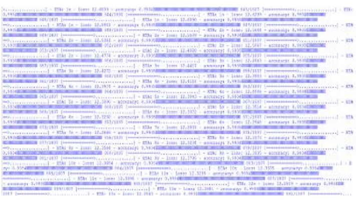

Figure 6. The machine learning model training.

9.

Figure 7. Machine learning output for training data.

10.

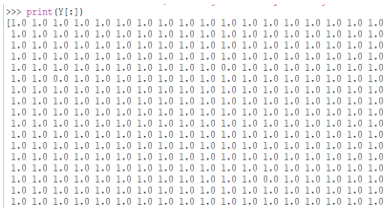

Figure 8. Machine learning dataset output.

11.

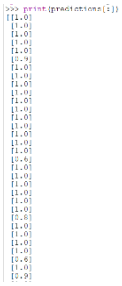

Figure 9. Machine learning data predictive output.

12.

Figure 11. Accuracy of model on new dataset.

13.

Figure 12. Predictive values of model on new dataset.

14.

		Predicted Churn score		
		Churned	Not churned	
Actual	Churned	40	8	
	Not churned	38	13	
			Accuracy	53%
	#Churned	49		
	#Predicted churned	78		

Figure 8. Confusion matrix on new dataset.

15. Figure 3 but readable

		Churn	Count_project	Registered_hours	Reason_churnsc	Invoicesc	Sbi_sc	typesc
Churn	Pearson Correlation	1	.006	.020**	-.872**	.020**	.002	-.003
	Sig. (1-tailed)		.075	<.001	.000	<.001	.278	.213
	N	57795	57795	57795	57795	57795	57795	57795
Count_project	Pearson Correlation	.006	1	.084**	-.014**	-.024**	-.044**	.019**
	Sig. (1-tailed)	.075		<.001	<.001	<.001	<.001	<.001
	N	57795	57795	57795	57795	57795	57795	57795
Registered_hours	Pearson Correlation	.020**	.084**	1	-.017**	-.002	-.001	.211**
	Sig. (1-tailed)	<.001	<.001		<.001	.342	.432	.000
	N	57795	57795	57795	57795	57795	57795	57795
Reason_churnsc	Pearson Correlation	-.872**	-.014**	-.017**	1	-.010**	.001	.004
	Sig. (1-tailed)	.000	<.001	<.001		.006	.366	.153
	N	57795	57795	57795	57795	57795	57795	57795
Invoicesc	Pearson Correlation	.020**	-.024**	-.002	-.010**	1	.007*	.001
	Sig. (1-tailed)	<.001	<.001	.342	.006		.043	.413
	N	57795	57795	57795	57795	57795	57795	57795
Sbi_sc	Pearson Correlation	.002	-.044**	-.001	.001	.007*	1	-.007
	Sig. (1-tailed)	.278	<.001	.432	.366	.043		.059
	N	57795	57795	57795	57795	57795	57795	57795
typesc	Pearson Correlation	-.003	.019**	.211**	.004	.001	-.007	1
	Sig. (1-tailed)	.213	<.001	.000	.153	.413	.059	
	N	57795	57795	57795	57795	57795	57795	57795

**. Correlation is significant at the 0.01 level (1-tailed).

*. Correlation is significant at the 0.05 level (1-tailed).

16. Figure 4 but readable

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	57795	100.0
	Missing Cases	0	.0
	Total	57795	100.0
Unselected Cases		0	.0
Total		57795	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

Block 0: Beginning Block

Classification Table[a,b]

			Predicted		
			Churn		Percentage Correct
	Observed		0	1	
Step 0	Churn	0	0	352	.0
		1	0	57443	100.0
	Overall Percentage				99.4

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	5.095	.053	9081.629	1	.000	163.190

17. Figure 5 but readable

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	4292.974	6	.000
	Block	4292.974	6	.000
	Model	4292.974	6	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000[a]	.072	1.000

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Classification Table[a]

			Predicted		
			Churn		Percentage Correct
	Observed		0	1	
Step 1	Churn	0	352	0	100.0
		1	0	57443	100.0
	Overall Percentage				100.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Count_project	-.003	3.654	.000	1	.999	.997
	Registered_hours	.002	7.724	.000	1	1.000	1.002
	Invoicesc	.026	41.720	.000	1	1.000	1.026
	Reason_churnsc	-19.414	594.488	.001	1	.974	.000
	Sbi_sc	.018	27.267	.000	1	.999	1.018
	typesc	.000	.380	.000	1	1.000	1.000
	Constant	40.166	1167.499	.001	1	.973	2.779E+17

a. Variable(s) entered on step 1: Count_project, Registered_hours, Invoicesc, Reason_churnsc, Sbi_sc, typesc.

18. Figure 6 but readable