**Journal of Supply Chain Management Science**

https://journals.open.tudelft.nl/jscms

http://dx.doi.org/10.18757/jscms.2021.6125

# Forecasting customer churn: Comparing the performance of statistical methods on more than just accuracy

Ronan Duchemin [a*], Ricardo Matheus [b]

[a] *Rotterdam School of Management, Erasmus University, The Netherlands*

[b] *Department Engineering Systems and Services, Faculty of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands*

Article history: received 01-11-2021, accepted 10-12-2021, published 31-12-2021

**Abstract** – There is a consensus that the best way to forecast customer churn is by statistical methods. It is, however, unclear when which statistical method is more appropriate. This study aims to provide a set of guidelines to data scientists and researchers who are interested in optimizing statistical methods. A systematic literature review revealed six most promising methods for churn forecasting and a selection of metrics which can be used to evaluate the performance of the methods. The six statistical methods are evaluated on five metrics of performance. The best-worst method (BWM), a multi-criteria decision-making method, is used to elicit the relative importance of the performance metrics. Based on the relative importance of the metrics and the performance of the methods, using additive value function, we find an overall value for each method based on which the forecasting methods can be ranked. Experimental analysis reveals that finding an overall value for each statistical analysis leads to a different ranking than when we use a single performance metric like accuracy or AUC. We argue that relying on an aggregated value, like the one we propose in this study, is more reliable than considering only one metric (a common practice).

**Keywords:** Best-Worst Method (BWM); Churn Forecasting, Machine Learning; Artificial Intelligence; Decision-Making; Business Analytics; Data Science; Statistical Methods; Performance Evaluation

## 1. Introduction

With the increase in transparency to -and expectations of- the customer, a decrease in customer loyalty is visible (Eskildsen and Kristensen 2007). This has forced companies to look for new methods to lock in their customers. One of the recent developments is the rise of the subscription economy in which consumers are picking fixed price ‘all you can eat’ options over pay-per-use (Weinman 2018). Customer retention is becoming a primary focus for most companies. One of the most important obstacles for customer retention is churn management (Evans 2002). Adequately, the scientific and professional field of marketing is recognizing the importance of forecasting churn. Churn can be defined as a customer abandoning its subscription (Coussement and Van den Poel 2008), closing its account (Larivière and Van den Poel 2004), or not making a purchase or showing any activity for a predefined period of time (Yu et al. 2011).

One of the crucial components of churn management is to forecast churn. It enables companies and institutions to act beforehand to persuade the customer or member to stay with the service. There exists a diverse set of forecasting tools and methods. However, there seems to be no clear best practice in the statistical methods and previous research has shown varying results with similar methods (Mahajan et al. 2015, Sharma et al. 2013). Furthermore, the comparison of the methods and the evaluation of the performance is often based on one metric of performance (Chu et al. 2007). There is little research done that successfully combines more than one metric of performance into a meaningful overall value. Therefore, data scientists are always having to deal with the limitations of their selected performance metric. This leads us to formulate the following main research question:

[*] Corresponding author. Email address: ronan.duchemin@gmail.com

*How can decision-makers (data scientists, analysts, managers) find the most appropriate statistical method for forecasting customers churn considering a number of relevant performance metrics?*

Although there are significant advances in both the fields of machine learning and expert decision making (Nguyen et al. 2018, Beliën and Forcé 2012), current research has missed to combine these two for the selection of statistical methods in the field of data science. Taking into consideration this research gap, this paper aims to provide a methodology to data scientists selecting the most appropriate churn forecasting methods, taking into consideration both expert opinion and data metrics. To answer the research question and achieving the aim we propose a generic methodology which is demonstrated in a real-word case study. Although, the scope of the case study is limited, the methodology can be used in other similar predictive analysis to identify the best performing statistical method. The methodology can be adopted to mitigate between stakeholders with different expectations of a statistical method. This can be realized by gathering all stakeholders' opinion before evaluating what method best fits the requirements. Furthermore, this study provides statistics on the performance of six statistical methods on five different performance metrics.

This introduction is followed by the theoretical background in Section 2. Experimental analysis is presented in Section 3 where we discuss our experiment plan, statistical methods, performance metrics, and the data collection. The results are presented in Section 4 which is followed by a discussion in Section 5. The paper is concluded in Section 6 with some future research directions.

## 2. Theoretical Background

For this study, a structured and systematic literature review (Jesson et al. 2011) was used to find and discuss existing literature on churn forecasting. This study followed eight consecutive steps as depicted in Figure 1. The protocol serves as a plan that needs to be carried out to complete this literature review successfully and transparently. Defining the plan beforehand will increase both the objectivity and the reproducibility of this research (Jesson et al. 2011).

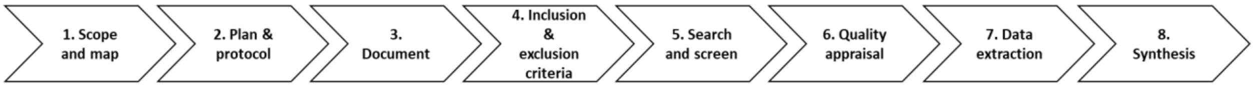


Figure 1. Review protocol (Jesson et al. 2011)

### 2.1. Statistical Methods Applied for Churn Forecasting

We reviewed the literature of churn forecasting and identified the employed methods which are presented in Table 1 and discussed in this sub-section. As depicted in Table 1, a wide variety of methods are applied to churn problems in the scientific literature. In order to provide a clear overview, a few of the worst performing methods that only came forward once have been left out of Table 1.

The methodology that was used most frequently for churn forecasting is Logistic regression (Ha et al. 2005). The most prominent reason that is given for picking this method is its simplicity (Tamaddoni et al. 2016). In a hackathon style tournament the Logistic regression was the most used method with 44% (Neslin et al. 2006). Additionally, the logistic method performs well on the longevity of the method its performance without having to retrain it (Risselada et al. 2010). The method prevails in situations where churn is uncommon.

The second most used method is the Decision tree (Lemmens and Croux 2006). It is praised for its high interpretability (Barfar et al. 2017) and ease of understanding (Neslin et al. 2006) while still showing competitive results (Lemmens and Croux 2006). This allows for humans to understand the reasoning of the method and explain the decisions to other stakeholders (Lemmens and Croux 2006). They often come out as the worst performing statistical method when compared to others, but in Risselada et al. (2010), the tree method shows true potential when combined with the bagging ensemble method. It is argued that because of its ease of interpretability and the traceability, the decision tree should still be considered as a viable method for churn forecasting.

Table 1. Statistical methods applied from the literature review

| Statistical Method | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **SVM** | **Boosting** | **Bagging** | **KNN** | **Logistic Regression** | **Neural Network** | **Decision Tree** | **PLS** | **Naive Bayes Classifier** | **HMM** | ***Scientific Sources*** |
| 2 | 1 | | | 3 | | | | | | Tamaddoni et al. (2016) |
| | | | | X | | | | | | Wieringa and Verhoef (2007) |
| | | | | X | | | | | | Moser et al. (2018) |
| | | | | | | | | | X | Ascarza (2018) |
| | | | | X | | X | | | | Neslin et al. (2006) |
| | X | X | | | | X | | | | Lemmens and Croux (2006) |
| | | 1 + tree<br>3 + logit | | 2<br>3 + bagging | | 1 + bagging<br>4 | | | | Risselada et al. (2010) |
| 1, 2, 3 | | | | 4 | | | | | | Hyoung-joo Lee et al. (2010) |
| | | 1 + NN<br>3 + LR | | 3 + bagging | 1 + bagging<br>2 | | | | | Ha et al. (2005) |
| | | | | | | | | | X | Mestre and Vitoria (2013) |
| | | | | X | | | | | | Coussement et al. (2017) |
| X | | | | X | | | | | | Moeyersoms and Martens (2015) |
| | | | | X | | | | | | Dierkes et al. (2011) |
| | | | X | | | | | | | Lee et al. (2012) |
| | | | | 2 | 3 | 4 | 1 | | | Hyeseon Lee et al. (2011) |
| | | | | X | | X | | | | Barfar et al. (2017) |
| | | | | X | | | | | | Prinzie and van den Poel (2006) |
| 2 | | | | | | | | 1 | | De Cnudde and Martens (2015) |
| 1 | | | | 4 | 3 | 2 | | | | Delen (2010) |
| | | | | 2 | 1 | 3 | | | | Zhang et al. (2012) |
| | | | | | X | X | | | | Chu et al. (2007) |
| X | | | | | | | | | | Chen and Fan (2012) |
| 6 | 2 | 3 | 1 | 15 | 5 | 8 | 1 | 1 | 2 | **Total** |

Abbreviations: SVM = Support Vector Machine; KNN = k-Nearest neighbour; PLS = Partial least squares; HMM = Hidden Markov method
Note 1: The numbers (1 to 4) are used to indicate the rank of the method relative to the other methods applied in the same article based on the performance criteria that has been used in this article.
Note 2: Some methods have been used in combination with an ensemble method (boosting or bagging) and without the ensemble method. For these articles the respective ranks have been given in both columns but in separate lines of text.
Note 3: The Random Forrest is an example of combining Bagging and the decision tree; the Gradient Boosted Tree is an example of Boosting and the Decision tree.

The next most used statistical method is the Support Vector Machine (SVM) (see Table 1). Like Decision trees, it is a classification algorithm. However, it uses different methods to classify the data. This makes it more difficult for users to explain the outcome of the method. SVM is considered as one of the better performing methods (Delen 2010, Lee et al. 2010). However, some studies (De Cnudde and Martens 2015, Tamaddoni, Stakhovych, and Ewing 2016) criticize SVM performance.

There seems to be a wide variety in the types of Neural networks that are used. To exemplify, the Self-Organizing Map algorithm provides desirable characteristics like stability and flexibility (Chu et al. 2007). Delen (2010) and Ha et al. (2005) use a Multi-layer perceptron algorithm for which Ha et al. (2005) have used the Bagging ensemble method to account for overfitting and instability. Lee et al. (2012) stay with the ‘basic’ three-layered Neural network. The Neural network shows varying performance results. This could be explained by the different types of architecture used and differences in the datasets.

Next up are bagging (Risselada et al. 2010) and boosting (Tamaddoni et al. 2016, Lemmens and Croux 2006). They are both ensemble techniques that try to improve the performance of a method by combining it with other (often, of the same type) methods, decreasing the variance and the bias of the methods. The ensemble methods mostly improve the performance of the statistical methods when compared in the same study. Although Boosting was only quantified once, it performed better compared to the SVM and Logistic regression (Tamaddoni et al. 2016).

Additionally, the Hidden Markov Method (HMM) was employed in two articles. The Hidden Markov method is characterized by its changing behaviour with unobserved states that may be influenced by earlier states. The method can account for different types of customer churn; namely silent churners and overt churners. Furthermore, it allows to estimate the future state of a customer in a more iterative way (Mestre and Vitoria 2013).

Finally, a few methods were used in only one study, but are still worth naming. They were either the only method used by the authors, or they performed similar or better than the other methods used in the study. The k-Nearest Neighbours algorithm is a relatively simple machine learning technique that shows decent results on classification problems (Lee et al. 2012). The Partial Least squares method is applied because of its minimal demands on measurement scales, residual distributions, and sample size (Lee et al. 2012). The Naive Bayes classifier is able to handle the big dimensions of a data set (De Cnudde and Martens 2015). However, the Naive Bayes classifier outperformed the popular SVM in their study.

The literature review reveals what the preferred statistical methods are in existing churn forecasting research. However, it failed to reveal how these researchers have come to their method selection. This paper will explore how method selection can be more transparent and standardized whilst respecting stakeholders' input.

## 2.2. Performance Metrics Applied for Churn Forecasting

This section provides a definition and explanation for the 12 different performance metrics that came forward during the literature review. Additionally, it will elaborate on the popularity of these metrics and their appropriateness for a churn forecasting problem. A schematic overview of the results is given in Table 2.

Performance metrics can have two types of goals. One type being focused on the direct performance of a specific method on a specific dataset and the other type being focused on the staying power and reproducibility of the method. The definitions of some important terms included in the confusion matrix are provided below (Provost and Fawcett 2013).

***True positive***: A test result for which the method detects the condition, and the condition is actually present.
***True negative***: A test result for which the method detects no condition when the condition is actually absent.
***False positive***: A test result for which the method detects the condition, but the condition is actually absent.
***False negative***: A test result for which the method detects no condition, but the condition is actually present.

**Accuracy** refers to the ability of the statistical method to predict the correct value. Accuracy is the most used performance metric. Accuracy is considered to be an appropriate benchmark when comparing methods' performance to each other. A limitation that comes forward is that focusing on accuracy as the only performance metric ignores customer profitability (Tamaddoni et al. 2016). To clarify, the formulas are provided in Eq. (1), the notations used in these formulas are coming from Table 2.

$$\begin{cases} Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \\ Error\ rate = \frac{FP+FN}{TP+TN+FP+FN} \\ Accuracy = 1 - Error\ rate \\ Error\ rate = 1 - Accuracy \end{cases} \quad (1)$$

**Precision** refers to the closeness of the measurements with respect to the observed values. Precision is independent of accuracy; the findings of a method could be very precise, but inaccurate. Variance is the counterpart of precision. Variance leads to classification error, and will therefore reduce the accuracy of a statistical method (Ha et al. 2005). It is measured by dividing the amount of correct predictions (true positives) by the total amount of predicted positive results (true positive + false positive) (Dierkes et al. 2011) (see Eq. 2).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Table 2. Metrics of performance

| **Accuracy** | **AUC** | **EVF** | **Lift** | **(R)MSE** | **Gini** | **Precision** | **TPR** | **FNR** | **FPR** | **Fastness** | **Ease of interpretation** | ***Reference*** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X | X | X | | | | | | | | | | Tamaddoni et al. (2016) |
| X | X | | | X | | | | | | | | Ascarza et al. (2016) |
| X | | | X | | X | | | | | | | Neslin et al. (2006) |
| X | | | X | | | | | | | | | Lemmens and Croux (2006) |
| | X | | | | | | | | | | | Ascarza (2018) |
| | | | X | | X | | | | | | | Risselada et al. (2010) |
| X | X | | | | | X | | | | | | Ha et al. (2005) |
| X | | | | X | | | | | | | | Mestre and Vitoria (2013) |
| X | | X | | | | | | | | | | Kitchens et al. (2018) |
| | X | | X | | | | | | | | | Coussement et al. (2017) |
| | X | | X | | | X | X | | | | | Moeyersoms and Martens (2015) |
| X | | X | X | | | | | | | | | Dierkes et al. (2011) |
| | | | | | | | | X | X | | | Lee et al. (2012) |
| X | | | X | | | | | | | X | X | Hyeseon Lee et al. (2011) |
| X | X | | | | | | | | | | | Barfar et al. (2017) |
| | X | | | | | | | | | | | Prinzie and van den Poel (2006) |
| | X | | | | | | | | | | | De Cnudde and Martens (2015) |
| X | | | | | | | | | | | | Delen (2010) |
| X | X | | X | | | | | | | | | Zhang et al. (2012) |
| X | | | | | | | | | | | | Chu et al. (2007) |
| X | X | X | | | | | X | | | X | | Chen and Fan (2012) |
| 14 | 11 | 4 | 8 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | ***Frequency*** |

Abbreviations: AUC = Area Under ROC Curve; EVF = Expected Value Framework; (R)MSE = (Root) Mean Squared Error; TPR = True Positive Rate; FNR = False Negative Rate; FPR = False Positive Rate

The **True Positive Rate (TPR)**; also referred to as recall or sensitivity, is defined as *"the proportion of positive examples which are predicted to be positive"* (Lee et al. 2015) (see formula in Eq. (3)).

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

The **False Negative Rate (FNR)** is defined as the percentage of false negatives of the total amount of negative predictions (Lee et al. 2012) (see formula in Eq. (4)). The costs of acquiring new customers in the telecom industry is significantly higher than the cost of retaining existing customers, it is important to keep the FNR as low as possible while respecting other costs.

$$FNR = \frac{FN}{FN + TN} \tag{4}$$

Lee et al. (2012) also refer to the **False Positive Rate (FPR)** as the false alarm rate. According to the authors, a high FPR increases unnecessary costs because you will be attempting to retain customers that are already loyal to your brand (see Eq. (5)).

$$FPR = \frac{FP}{FP + TP} \tag{5}$$

According to Ha et al. (2005) the **Receiver Operator Characteristic (ROC)** curve plots one minus the specificity against one minus the sensitivity on the other axis. Sensitivity measures the proportion of true positives that have been correctly identified as true positives (Dierkes et al. 2011) (see Eq. (6) and Eq. (7)). Figure 2 provides an example of an ROC curve.

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{TN+FP} \tag{7}$$

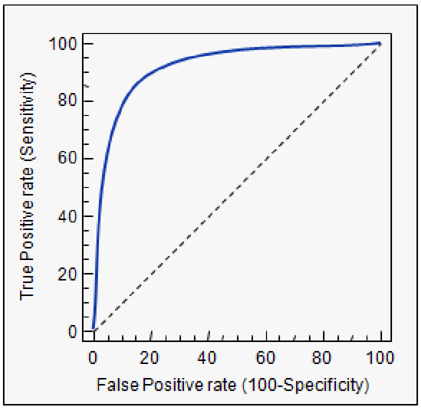


Figure 2. ROC curve (Schoonjans 2018)

The **Area Under the ROC Curve (AUC)** is an estimation of the probability that a randomly selected churner has a higher posterior probability of churn than a randomly selected customer who did not churn (Coussement et al. 2017). Furthermore, the ROC curves address the trade-off between the true positive rate and the false positive rate (Barfar et al. 2017). AUC summarizes the effect of the sensitivity and specificity in one value (Prinzie and van den Poel 2006, Chen and Fan 2012). Moreover, multiple studies seem to find advantages of using the AUC as a measurement over accuracy (Ascarza 2018, Moeyersoms and Martens 2015, De Cnudde and Martens 2015).

The **expected value framework** includes the corresponding costs of true positives, true negatives, false positives, and false negatives to get an estimated cost or benefit from running a retention campaign (Provost and Fawcett 2013). A connection to Customer Lifetime Value is indicated, a company with access to such a number can use this as a more accurate representation of the benefit of a successful retention offer than the profit of one additional year (Kitchens et al. 2018, Tamaddoni et al. 2016).

The expected benefit of targeting is defined as follows in Eq. (7), where $p(R \mid x)$ is the probability that a targeted customer will respond, $v_R$ is the expected benefit of this customer responding, $1 - p(R \mid x)$ is the likelihood that a customer will not respond to the offer, and $v_{NR}$ is the cost associated with not getting any response (Provost and Fawcett 2013).

$$Expected\ benefit\ of\ targeting = p(R \mid x) \cdot v_R + 1 - p(R \mid x) \cdot v_{NR} \tag{8}$$

**Easy to interpret** refers to the ability of the data scientist or manager to interpret the results, explain how the method came to a decision, and work with the results of the method. Forecasting methods should aim to have predictive and comprehensive power (Lee et al. 2011). This metric can be of the utmost importance in some of the professional fields. Additionally, scientists should consider the ‘Occam’s razor’ principle that states that if all other things are equal, the simpler method will generalizes better and should be preferred (Lee et al. 2011).

**Fastness** of the method is about the time required to build and train a model. It also considers the processing power involved to run the method. This can be done by calculating the CPU time requirement (Lee et al. 2011, Chen and Fan 2012). Often, for two similar methods, a faster runtime will indicate that the parameters are optimized better, leading to a faster learning process.

Top decile **lift** has managerial value due to its focus on the likely churners and arranging the customers by the predicted risk of churning (Coussement et al. 2017). A top decile lift of two would imply that the top 10 percent of most likely churners are twice as likely to churn as the remaining 90 percent of the dataset. Lift can be a beneficial metric to companies that are limited in the number of people they can reach out to. Lift is a good comparison tool to quantify the improvement with a random method. For a method to perform better than a random classifier, lift should be greater than one (Moeyersoms and Martens 2015 , Neslin et al. 2006). Lift is one of the most popular evaluation metrics in churn forecasting (Lemmens and Croux 2006, Tamaddoni et al. 2016, Neslin et al. 2006), which is formulated as follows (Neslin et al. 2006).

$$Fraction\ of\ customers\ that\ churn = Lift * fraction\ of\ targeted\ customers\ that\ churn \qquad (9)$$

The **Gini coefficient** looks at the area between the cumulative Lift curve and the curve of a random prediction (Lift = 1) (Neslin et al. 2006). The combination of Lift and the Gini coefficient is a good addition to accuracy; specifically when forecasting rare events, for which the accuracy sometimes is really inappropriate (Neslin et al. 2006).

The **Mean Squared Error (MSE)** can be used to aggregate prediction errors over multiple groups in data and is, therefore, best applicable when clustering customers before running the forecasting method (Ascarza 2018). The Root Mean Squared Error (RMSE) is suitable for clusters; they describe the RMSE as an error metric based on the centres of the clusters that can assist them in forecasting future customer segments (Mestre and Vitoria 2013).

From the literature in churn forecasting, it becomes evident that using one performance metric to evaluate the performance of a statistical method will always come with its limitations. Moreover, in complex corporate environments there may be various requirements of the performance of the statistical methods. This demonstrates the need for a method providing data scientists with the option to evaluate their methods on multiple performance metrics.

## 2.3. Multi-Criteria Decision-Analysis

As discussed before in this study we aim to evaluate several statistical methods with respect to several performance metrics. As the performance metrics are of different importance, the problem is inherently a multi-criteria decision analysis (MCDA) problem. In a typical MCDA problem, we have $m$ alternatives ($i = 1, 2, \ldots, m$) where each alternative can be represented by its performance with respect to $n$ criteria ($j = 1, 2, \ldots, n$). If $a_{kj}$ indicates the performance of alternative $a_k$ with respect to criterion $j$ and $v_{kj}(a_{kj})$ be the normalized value of $a_{kj}$, and $w_j$ shows the relative importance (weight) of criterion $j$, we could find the overall value of alternative $a_k$ or $v(a_k)$ using the additive value function, among other approaches, proposed by Keeney and Raifa (1976), presented as follows.

$$v(a_k) = \sum_{j=1}^{n} w_j v_{kj}(a_{kj}) \qquad (10)$$

The overall value of alternatives can then be used to compare and rank the alternatives. Alternative $a_k$ is considered to be preferred to alternative $a_z$ if and only if $v(a_k)$ is greater than $v(a_z)$ or:

$$a_k \succ a_z \Leftrightarrow v(a_k) > v(a_z) \qquad (11)$$

The normalized value of $a_{kj}$ or $v_{kj}(a_{kj})$ can be obtained using several different approaches (see, for instance, (RL and Raiffa 1976, Dyer and Sarin 1979, Rezaei 2018). In this study, we use the following approach.

$$v_{kj}(a_{kj}) = \begin{cases} \frac{a_{kj} - \min\{a_{ij}\}}{\max\{a_{ij}\} - \min\{a_{ij}\}}, \text{for benefit criteria} \\ \frac{\max\{a_{ij}\} - a_{kj}}{\max\{a_{ij}\} - \min\{a_{ij}\}}, \text{for cost criteria} \end{cases} \qquad (12)$$

Another component of the additive value function is the weights of the criteria ($w_j$). There exist several methods to find the weights of the criteria such as Analytic Hierarchy Process (AHP) (Saaty 1977), Simple Multi-Attribute Rating Technique (SMART) (Edwards 1977), Trade-off (RL and Raiffa 1976), Direct Rating method (Bottomley and Doyle 2001). In this study we use one of the latest major developments in the field, the Best-Worst Method (BWM), which is developed in 2015 by Rezaei (2015). BWM is a pairwise comparison-based method which relies on decision-makers (DMs)/experts to elicit the weights of the criteria (and alternatives).

The structured pairwise comparison of BWM brings about several interesting features (Rezaei 2020) which is why we chose this method for our study: (i) identifying the best and the worst criteria before conducting the pairwise comparisons, informs the DM/expert on the range of evaluation. When a DM/expert knows that A is the most important and B is the least important, the biggest score goes to comparing A and B and all the other comparisons fall somewhere in between. This feature is the main reason why BWM leads to more consistent pairwise comparisons compared to other methods such as AHP (Rezaei 2015). (ii) in the optimization model of BWM which is used to find the weights, two vectors are used as input (Best to others; others to worst). As the two vectors present pairwise comparisons against two opposite reference points (best and worst), the findings (weights) are less vulnerable against possible anchoring bias. (iii) BWM stands in the middle of a continuum of MCDA methods in terms of data efficiency (methods like SMART use a single vector while methods like AHP use a full pairwise comparison matrix). BWM uses only two vectors which means it is more data efficient than methods like AHP and less data efficient than methods like SMART. However, a method like SMART could not systematically check the consistency of a DM/expert who has provided the data while BWM does. This implies that BWM is a more data-efficient method which could at the same time checks the consistency of the pairwise comparison data. For more features of the method a reader might refer to Rezaei (2020). Here we describe the steps we need to follow when implementing BWM (Rezaei 2016).

The Bayesian BWM also generates a Credal Ranking which describes the relation (> or <) of each pair of criteria with a confidence level. The latter “represents the extent to which one can be certain about the superiority of a criterion over one another” (Mohammadi and Rezaei 2019), which can significantly improve the DM’s decisions.

**Step 1:** Determine a set of criteria $\{C_1, C_2, \dots, C_n\}$ by the DMs/experts
**Step 2:** Determine the best (most important, or most contributing) and worst (least important or least contributing) criteria by the DMs/experts.
**Step 3:** Determine the preference of the best criterion over the other criteria by the DMs/experts using a number between 1 and 9, where 1 is equal importance, 9 is extremely more important, resulting in a Best-to-Others vector as follows in Eq. 13.
$A_B = (a_{B1}, a_{B2}, \dots, a_{Bn})$
**Step 4:** Determine the preference of the criteria over the worst criterion by the DMs/experts using a number between 1 and 9, resulting in an Others-to-Worst vector as follows in Eq. 14.
$A_W = (a_{1W}, a_{2W}, \dots, a_{nW})$
**Step 5:** Find the optimal weights.

According to the linear model of BWM (Rezaei 2016), the optimal weights $(w_1^*, w_2^*, \dots, w_n^*)$ are identified where the maximum of $\{|w_B - a_{Bj}w_j|, |w_j - a_{jw}w_W|\}$ for all $j$, is minimized. Considering the normality and non-negativity conditions of the weights we need to solve the following optimization problem to get the weights.

$$\min \max_j \{|w_B - a_{Bj}w_j|, |w_j - a_{jw}w_W|\}$$
s.t.
$$\textstyle\sum_{j=1}^{n} w_j = 1 \qquad (15)$$
$$w_j \geq 1, for\ all\ j$$

This problem is transformed to a linear programming problem as follows.

$$\min \xi^L$$
s.t.
$$|w_B - a_{Bj}w_j| \leq \xi^L, for\ all\ j$$
$$|w_j - a_{jW}w_W| \leq \xi^L, for\ all\ j \qquad (16)$$
$$\sum_{j=1}^{n} w_j = 1$$
$$w_j \geq 1, for\ all\ j$$

Solving this linear programming problem, we find the optimal weights $(w_1^*, w_2^*, \dots, w_n^*)$ and $\xi^{L*}$. A value of $\xi^{L*}$ close to zero indicates a high consistency and therefore a high reliability. For the linear BWM we could use the input-based consistency ratios and thresholds (see Table 3) to check the acceptability of the provided pairwise comparisons by the DMs/experts (Liang et al. 2020).

Table 3. Input-based consistency ratio thresholds (Liang et al. 2020)

| | Criteria | | | | | | |
|---|---|---|---|---|---|---|---|
| **Scales** | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *3* | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| *4* | 0.1121 | 0.1529 | 0.1898 | 0.2206 | 0.2527 | 0.2577 | 0.2683 |
| *5* | 0.1354 | 0.1994 | 0.2306 | 0.2546 | 0.2716 | 0.2844 | 0.2960 |
| *6* | 0.1330 | 0.1990 | 0.2643 | 0.3044 | 0.3144 | 0.3221 | 0.3262 |
| *7* | 0.1294 | 0.2457 | 0.2819 | 0.3029 | 0.3144 | 0.3251 | 0.3403 |
| *8* | 0.1309 | 0.2521 | 0.2958 | 0.3154 | 0.3408 | 0.3620 | 0.3657 |
| *9* | 0.1359 | 0.2681 | 0.3062 | 0.3337 | 0.3517 | 0.3620 | 0.3662 |

## 3. Experimental Analysis

This section will elaborate on the method that has been followed for our research. First, the experiment plan will be described as well as the data. Hereafter, the selection of the statistical methods and performance metrics will be discussed. Finally, this section will explain how the BWM is being used.

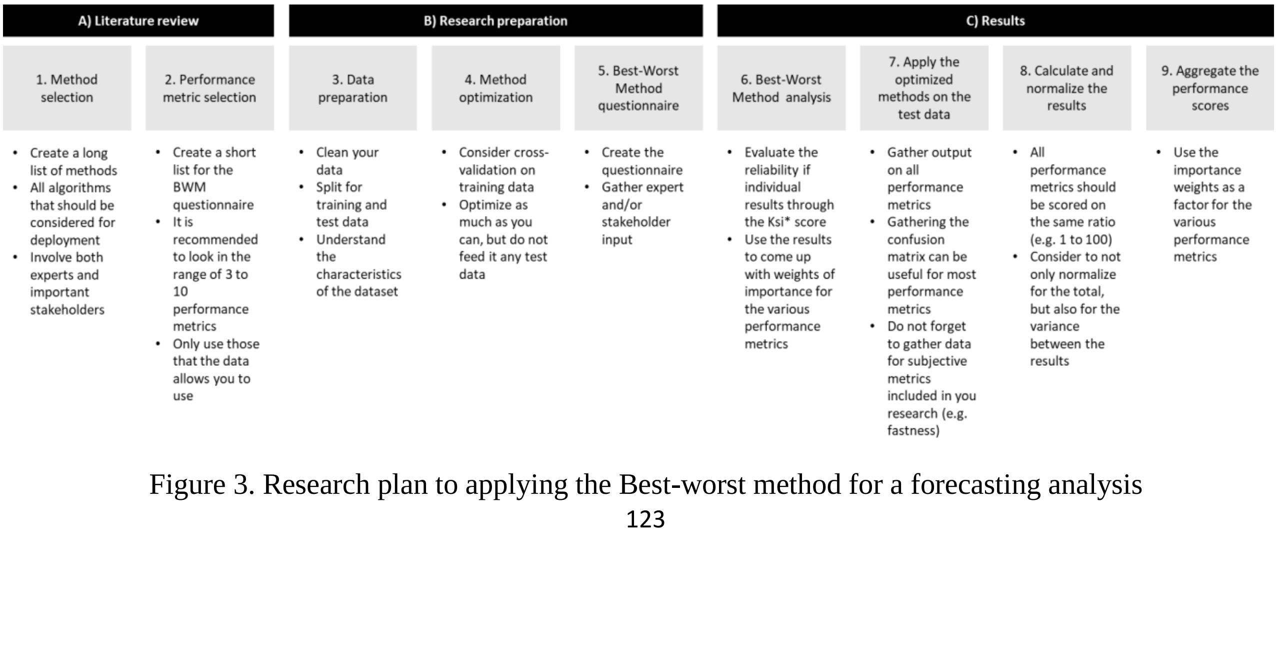
### 3.1. Experiment Plan

For the experimental analysis we selected six different statistical methods: Decision tree, Logistic regression, SVM, Random forest, Gradient Boosted Trees (GBT), and a Multi-layer perceptron neural network. To have a fair comparison between the methods, these methods are first optimized for forecasting customer churn.

We then use five performance metrics (Accuracy, Precision, AUC, Easy to interpret, Fastness) to evaluate the performance of the chosen forecasting methods. The BWM is used to find the relative importance (weight) of the metrics based on experts' opinion. The weights and the performance of the methods with respect to the performance metrics are then used in an additive value function. As a result, we can find the overall value of each method which is used to rank the methods and identify the best performing one.

Learning curves have been created to show the performance of the methods in situations where training data is available to a smaller extent. These learning curves have also been used to evaluate the effects of cross validation for the traditional methods and the effect of scaling the data manually for the Logistic regression. Although this research aims to compare the methods on multiple performance metrics, the decision has been made to only consider accuracy and AUC while optimizing the parameters and creating the learning curves. This is because there is not yet any package or method to optimize a statistical method on multiple performance metrics. For our best performing methods on the performance metrics accuracy and AUC, the other metrics of performance have also been scored.

A learning curve has been created for both the statistical methods and the ensemble methods mapping the performance of the method with various, logistically increasing, training sizes (Yelle 1979). To gather the input for this learning curves a starting number of 7 examples was used, which was multiplied by two repeatedly until it reached 14,336, after which the full training sample of 16,000 examples would be trained. A schematic overview of the research plan is provided in Figure 3.



A) Literature review

1. Method selection
- Create a long list of methods
- All algorithms that should be considered for deployment
- Involve both experts and important stakeholders

2. Performance metric selection
- Create a short list for the BWM questionnaire
- It is recommended to look in the range of 3 to 10 performance metrics
- Only use those that the data allows you to use

B) Research preparation

3. Data preparation
- Clean your data
- Split for training and test data
- Understand the characteristics of the dataset

4. Method optimization
- Consider cross-validation on training data
- Optimize as much as you can, but do not feed it any test data

5. Best-Worst Method questionnaire
- Create the questionnaire
- Gather expert and/or stakeholder input

C) Results

6. Best-Worst Method analysis
- Evaluate the reliability if individual results through the Ksi* score
- Use the results to come up with weights of importance for the various performance metrics

7. Apply the optimized methods on the test data
- Gather output on all performance metrics
- Gathering the confusion matrix can be useful for most performance metrics
- Do not forget to gather data for subjective metrics included in you research (e.g. fastness)

8. Calculate and normalize the results
- All performance metrics should be scored on the same ratio (e.g. 1 to 100)
- Consider to not only normalize for the total, but also for the variance between the results

9. Aggregate the performance scores
- Use the importance weights as a factor for the various performance metrics

Figure 3. Research plan to applying the Best-worst method for a forecasting analysis

### *Data description*

The dataset that has been used for this research is an academic dataset[*] which consists of 20,000 rows (which represent customers of a telephone company) and 12 variables. The data does not have any time or date information. The dataset is clean and shows no missing variables. The dataset is in the Attribute-Relation File Format (ARFF) which can be described as a Comma-Separated Value (CSV) file with additional information of the variables included. Out of the 20,000 examples, 9,852 have churned and the remaining 10,148 extended their contract. The 12 variables consist of seven numerical variables and five categorical variables; of which one is the binomial classifier that will be the target variable. A description of the twelve variables is given in Table 4.

Table 4. Variable descriptions

| Variable name | Variable description | Variable type | Variable content |
|---|---|---|---|
| College | Is the customer college educated? | Binominal | zero (no); one (yes) |
| Income | Annual income | Numerical | Number in euros |
| Overage | Average overcharges per month | Numerical | Number in euros |
| Leftover | Average % leftover minutes per month | Numerical | Percentage (number) |
| House | Value of dwelling (from census tract) | Numerical | Number in euros |
| Handset_price | Cost of phone | Numerical | Number in euros |
| Over_15min_calls_per_month | Average number of long (>15 mins) calls per month | Numerical | Real number |
| Avarage_call_duration | Average call duration | Numerical | Real number |
| Reported_satisfaction | Reported level of satisfaction | Categorical | very_unsat; unsat; avg; sat; ver_sat |
| Reported_usage_level | Self-reported usage level | Categorical | very_little; little; avg; high; very_high |
| Considering_change_of_plan | Was customer considering changing his/her plan? | Categorical | actively_looking_into_it; considering; perhaps; never_thought; no |
| Leave | Class variable: whether customer left or stayed | Binominal | LEAVE; STAY |

### *Data preparation*

The data preparation has been done in R Studio and has provided the master dataset that was used for all types of modelling to overcome variances. By using the same training data and the same test set for all statistical methods this research aims to provide a fair comparison of the performance of these methods.

For this research, the training dataset contains 80 percent of the original dataset resulting in 16,000 examples and the test dataset has 4,000. The sample function has been used to shuffle the dataset before splitting. This increases the likelihood of having representative subsets. Duplication of the same example was not allowed when picking the random samples; therefore, all 20,000 rows from the original dataset are included in either the training- or the test subset. Finally, the literature suggests that the Logistic regression (Knol et al. 2007) and SVM (Hsu and Lin 2002) perform best when the numeric variables are scaled. To have a fair comparison the scaling is done in advance of the modelling and the same scaled dataset is used for both approaches. To scale the numerical variables, a copy of the earlier created master datasets was made for both the training and the test subset.

## 3.2. Method Selection

The six selected statistical methods can be categorized into three types: traditional statistical methods, ensemble methods, and Neural networks. The literature review on churn forecasting has assisted in the selection of the traditional methods and has also guided the selection of the ensemble methods. This section will briefly explain how these methods have been selected.

[*] https://www.dropbox.com/s/h40f4y2a5wiue2u/churn.csv?dl=1

**Decision tree** has been selected because it is very intuitive to interpret and is therefore a very popular method in the field of management and decision making (James et al. 2013). The Decision tree consists of a number of binomial splits (Provost and Fawcett 2013). To decide where to make these splits the classification error rate is used. This is explained as the fraction of the observations in the training data that do not belong to the most common class (James et al. 2013).

**Logistic regression** was selected because it has been applied most frequently in the existing literature. It has proven to be very useful for binominal predictions such as with churn forecasting (Provost and Fawcett 2013). It is also a relatively simple method which keeps it easy to interpret the results. Log-odds can help statistical methods with allocating negative effects to unlikely (< 50%) events and positive effects to likely (> 50%) events (see Table 5) (Provost and Fawcett 2013). These odds can also be drawn out which shows the sigmoid curve.

Table 5. Log-odds (Provost and Fawcett 2013)

| Probability | Odds | Log-odds |
|---|---|---|
| 0.5 | 50:50 or 1 | 0 |
| 0.9 | 90:10 or 9 | 2.19 |
| 0.999 | 999:1 or 999 | 6.9 |
| 0.01 | 1:99 or 0.0101 | -4.6 |
| 0.001 | 1:999 or 0.001001 | -6.9 |

**Support Vector Machine** is known to be one of the best ‘out of the box’ classifiers, but is limited in the type of data(sets) it can process (James et al. 2013). It was not applied frequently, but it was often one of the better performing methods (Delen 2010). The SVM requires all data to consist of real numbers. Therefore, researchers should consider converting their categorical variables into numeric data. To achieve this, a categorical variable with three values included can be transferred into Boolean vectors. To illustrate this, an attribute with the options {red, blue, green} could be represented as (1,0,0) for red, (0,1,0) for blue, and (0,0,1) for green (Hsu, Chang and Lin 2010).

**Random forest** can be interpreted as an improvement of the ‘bagged tree’ by adding a small tweak that has a decorrelation effect (Provost and Fawcett 2013). This assists the algorithm in considering variables that typically would be dominated by stronger predictors; this is particularly interesting, because it can help with predicting the ‘hard to predict’ cases. The model ensembles a number of Decision trees, which is specified in the ‘number of trees’ parameter. The trees will be trained on a bootstrapped subset of the original training set. When provided with test data, each tree will make a prediction for the given example and majority voting will be used to select a prediction. Pruning could be leveraged to reduce the complexity of the model by replacing so called sub-trees that provide little predictive power.

**Gradient Boosted Trees** is one of the most effective methods to reduce bias (Ganjisaffar et al. 2011). A GBT model is an ensemble of tree models. It can be described as a forward-learning ensemble method that gradually improves prediction estimations. It uses a flexible nonlinear regression procedure to advance the accuracy of the trees (boosting). The improved accuracy of the model comes with a trade-off with computation time and human interpretability.

**Neural networks** are a type of generalized nonlinear and nonparametric model inspired by studies of the brain and nerve system (Alon et al. 2001). In comparison to the examined traditional models and other econometric models, neural networks have the ability to model complex -possibly nonlinear- relationships without requiring any prior assumptions (Hornik et al. 1990). A large problem is being tackled through the divide and conquer methodology (Karn 2016). A lot of deep learning practitioners now use the weights and parameters of this model as a basis of the model they are building. This will pre-train the weights of the model and allow the practitioner to build very decent models with a relatively small amount of data. In neuroscience, the brain is described as an implicit learning machine that evolves over millions of years (Marblestoneet al. 2016). This description also relates to the idea of using pre-trained weights and parameters for machine learning models. For example, Neural networks for natural language processing that where pre-trained on Wikipedia text data performed extraordinary well with small amounts of training data (Collobert and Weston 2008).

### 3.3. Selection of Performance Metrics

In this study we selected six performance metrics to evaluate the performance of the selected forecasting methods. Below, we discuss why we have chosen these metrics in our study.

**Accuracy** has been selected because it is known to be the most comprehensive metric to summarize the performance of the methods. The biggest limitation that came forward in the literature is that the accuracy does not perform well on datasets with only a small fraction of the target variable being positive or negative. In Section 3.1.1, it is shown that this is not the case for the dataset in this study with 49.3 percent of the customers churning such a method would perform similar to a random method.

**Precision** provides information about the fraction of the positive predictions that are positive results. There is a trade-off between the accuracy and precision, which is often referred to as the bias-variance trade-off (Moeyersoms and Martens 2015).

**AUC** is praised for its effectiveness as an optimizing metric for statistical methods (Delen 2010). By providing one meaningful number that summarizes all four values in the confusion matrix, it justifies the removal of the true positive rate and the false negative rate from the performance metrics.

**The expected value framework** has been included in this selection because this research aims to provide information valuable for both managers and data scientists. Using such a framework is one of the most important decision criteria for most managers. It allows them to justify expenditure on marketing campaigns and to quantify the results they could achieve. However, it was later removed for the analysis as the data did not allow for adding a (monetary) value to the outcome of a prediction.

**The ease of interpretation** has been selected with the decision maker in mind. If the method provides a manager with the tools to explain how a certain decision has been made, the method will score high in this metric.

**The fastness metric** has been included with a few points of evaluation that have been considered when scoring this metric. The computation time of running the final method has been measured in seconds. Furthermore, the time of building a working prototype of the method and the time spent on optimizing the parameters have been indicated on a five-point Likert scale. Finally, it has been indicated if it was possible to create a method that performs significantly better than a random model in RapidMiner which can reduce the technical knowledge required as well as the deployment time.

Finally, to get one score that expresses the performance of the method, all six metrics needed to be transferred into a number with the same ratio. Since most of our performance metrics are either a percentage or a ratio, we transform all metrics to a number ranging from 0 to 100. For the accuracy and precision, this would mean that the percentage would be expressed as a number. The AUC will be multiplied by 100. For the five-point Likert scale questions, a score will be allocated to all scores ranging from 0 to 100 (0: very low/very slow, 25: low/slow, 50: average, 75: high/fast, 100: very high/very fast). The computation times range from 1 second to 38 seconds, which is still a very short time for training a statistical method. Therefore, the choice has been made to calculate the ratio for computation time by subtracting the number of seconds from 100. Finally, the Boolean category that indicates if the method is available through RapidMiner has the following scores allocated to it: ‘yes’ gets the full score of 100 and ‘no’ gets a score of 60. This was chosen because it would give a more similar number to that of the other metrics. However, it is recommended to standardize all metrics to get similar standard deviations. This will be elaborated on in the discussion.

### 3.4. Method Optimization

The main purpose of the Decision tree in this research was to provide a benchmark to compare the more advanced models with. The parameters of the best performing decision tree can be found below in Table 6.

Table 6. Parameters of the Decision tree

| **Parameter** | **Chosen value** |
|---|---|
| criterion | information_gain |
| Maximal depth | 20 |
| Apply pruning | unchecked |
| Apply prepruning | checked |
| Minimal gain | 0 |
| Minimal leaf size | 30 |
| Minimal size for split | 4 |
| Number of prepruning alternatives | 3 |

The logistic regression is being presented as a ‘fast’ model; referring to both the time to build the model and the time to train the model. Therefore, experimentation with the parameters of the logistic regression was limited. The parameters that have been used for this research are shown in Table 7. The results show that there is no visible difference between the scale function in R and the ‘standardize’ parameter for the Logistic regression in RapidMiner.

Table 7. Parameters of the Logistic regression

| Parameter | Chosen value |
|---|---|
| solver | AUTO |
| reproducible | unchecked |
| use regularization | unchecked |
| standardize | Flexible → see below |
| non-negative coefficients | unchecked |
| add intercept | checked |
| compute p-values | checked |
| remove collinear columns | checked |
| missing values handling | MeanImputation |
| max iterations | 0 |
| max runtime seconds | 0 |

For the SVM. The ‘scale’ parameter will remain unchecked as the master dataset already provides us with a scaled dataset. An overview of the final parameters is provided in Table 8.

Table 8. Parameters of the SVM before and after optimizing

| Parameter | Before optimizing | After optimizing |
|---|---|---|
| Kernel type | dot | dot |
| Kernel cache | 200 | 200 |
| C | 0 | 0 |
| convergence epsilon | 0.001 | 0.005 |
| max iterations | 100000 | 32000 |
| scale | unchecked | unchecked |
| L pos | 1 | 1 |
| L neg | 1 | 1 |
| epsilon | 0 | 0 |
| epislon plus | 0 | 0 |
| epsilon minus | 0 | 0 |
| balance cost | unchecked | checked |
| quadratic loss pos | unchecked | unchecked |
| quadratic loss neg | unchecked | unchecked |
| **Performance** | | |
| Accuracy | 63.42% | 64.08% |
| AUC | 0.696 | 0.699 |

When the Random forest is provided with test data, each tree will make a prediction for the given example and majority voting will be used to select a prediction. Pruning could be leveraged to reduce the complexity of the model by replacing so called sub-trees that provide little predictive power. The settings for the parameters before and after the experimentation can be viewed in Table 9.

A gradient boosted model is an ensemble of tree models. It can be described as a forward-learning ensemble method that gradually improves prediction estimations. It uses a flexible nonlinear regression procedure to advance the accuracy of the trees; this process is also called boosting. The improved accuracy of the model comes with a trade-off with computation time and human interpretability. To minimize the negative effects from this trade-off, GBT generalizes tree boosting. Table 10 shows the parameters before and after the optimization process.

The main differentiator in the architectures of Neural networks can be found in the configuration of the layers. All Neural networks consist of an input layer, one or more hidden layers, and an output layer. A specific variable will contain all continuous variables and another variable will contain the categorical variables.

Table 9. Parameters of the random forest before and after optimizing

| Parameter | Before optimizing | After optimizing |
|---|---|---|
| number of trees | 20 | 100 |
| criterion | accuracy | information_gain |
| maximal depth | 20 | 20 |
| apply pruning | unchecked | unchecked |
| apply prepruning | checked | checked |
| minimal gain | 0 | 0 |
| minimal leaf size | 30 | 20 |
| minimal size for split | 3 | 8 |
| number of prepruning alternatives | 3 | 3 |
| random splits | unchecked | unchecked |
| guess subset ratio | checked | checked |
| voting strategy | confidence vote | confidence vote |
| use local random seed | 1337 | 1337 |
| enable parallel execution | checked | checked |
| **Performance (sample size = 7168)** | | |
| Accuracy | 0.6844 | 0.6978 |
| AUC | 0.757 | 0.767 |

Table 10. Parameters of the Gradient Boosted Trees before and after optimizing

| Parameter | Before optimizing | After optimizing |
|---|---|---|
| number of trees | 100 | 200 |
| reproducible | unchecked | unchecked |
| maximal depth | 10 | 8 |
| min rows | 10 | 35 |
| min split improvement | 0 | 0 |
| number of bins | 20 | 20 |
| leaning rate | 0.01 | 0.006 |
| sample rate | 1 | 0.83 |
| distribution | AUTO | AUTO |
| early stopping | unchecked | unchecked |
| max runtime seconds | - | - |
| **Performance** | | |
| Accuracy | 0.6864 | 0.688 |
| AUC | 0.764 | 0.773 |

Furthermore, a variable describes what the target variable is. Next, the test set needs to be defined before transferring the data.

A Neural network with no hidden layers would not be able to handle any irregularities in the data as it is not possible to work with nonlinearity. First, a function was built that can calculate the AUC corresponding to each epoch of the tabular learner. Furthermore, the tabular learner was set up with the required number of neurons in the hidden layers. The learning rate is selected based on an evaluation of the learning curve. For the selected configuration of 23 neurons in the first hidden layer and 500 hidden neurons in the second hidden layer the learning curve is shown in Figure 4. Selecting the learning rate is an intuitive process, aiming to take the lowest learning curve without skipping any moments where the loss is reduced a lot (where the line follows a steep curve) to maximize the performance (Howard and Thomas 2021). In Figure 4 it is a difficult decision between 1e-2 and 1e-3. In these cases, it is best to experiment with both, but the 1e-3 option is the safer and better choice in this case. The selection of the architecture was not only based on its high accuracy, but moreover, on its consistent high score over multiple epochs Table 11.

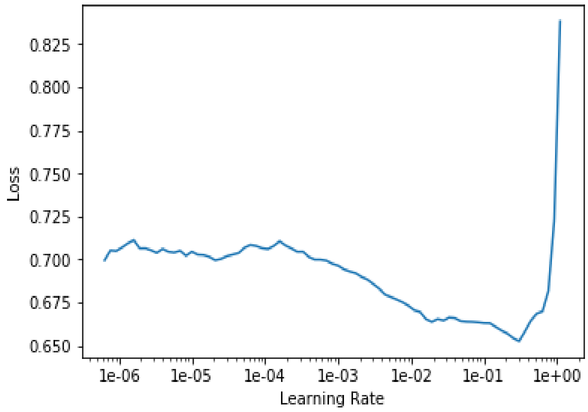


Figure 4. The learning curve

Table 11. The results of fitting the model

| epoch | train_loss | valid_loss | accuracy | AUROC | time |
|---|---|---|---|---|---|
| 0 | 0.595299 | 0.588706 | 0.693500 | 0.754750 | 00:03 |
| 1 | 0.594059 | 0.581007 | 0.686500 | 0.763900 | 00:03 |
| 2 | 0.585856 | 0.598321 | 0.682000 | 0.755847 | 00:03 |

### 3.5. Gathering Expert Opinion

In this section we discuss the questionnaire and respondents for eliciting the weights of the performance metrics using BWM.

***The questionnaire and respondents***

Since the content of the questionnaire is quite specific, even practitioners and experts may need additional information. Moreover, because it was important that all experts followed the same definition, this information was provided in the questionnaire. For the distribution of the survey Qualtrics has been the main contributor.

Due to the specificity of the questions it was important to gather respondents that fitted into a certain profile. Because of this, it was decided to personally reach out to as many data scientists, academics, and other experts. LinkedIn was the main platform for this. Moreover, a blogpost has been shared on the Data Science Foundation [66], which helped a lot with gathering valuable opinions. A total of 63 completed questionnaire were collected, of which 35 complied with the approximated consistency thresholds and could therefore be used for further analysis (Liang et al. 2020). The 35 valid respondents are divided into two groups of experts: professionals and academics. The professionals make up for the biggest portion with 22 respondents and the academics provided 13 results. An overview of the professional fields and location of the respondents can be found in Table 12.

***Processing the data***

To come up with relative importance (weight) for all performance metrics, we need to aggregate the weights from different experts. To test for differences between the academics and professionals the scores have been examined in four ways; (i) the aggregate of all 35 respondents with equal importance, (ii) the aggregate of only the 22 professionals with equal importance, (iii) the aggregate of the 13 scholars with equal importance, and (iv) an aggregate of the score for the professionals and academics with both groups having a 50 percent weight. For the last option, the opinions of the academics weigh relatively more than those of the professionals. For our results, the option that provides equal importance to both groups (way iv) has been selected. This is chosen because this study aims to provide valuable insights to both groups.

## 4. Results

This section begins by presenting the results of the BWM. Subsequently, the results of the six statistical methods that have been selected for analysis are provided.

Table 12. An overview of the respondents

| Interviewee ID | Professional field | Job Title | Interviewee ID | Professional field | Job Title |
|---|---|---|---|---|---|
| 1 | Academic | PhD Student | 19 | Professional | Data Analytics Supervisor |
| 2 | Academic | Lecturer/ researcher | 20 | Academic | Assistant Professor |
| 3 | Academic | PhD student | 21 | Professional | Head of Data Science |
| 4 | Professional | Data Scientist | 22 | Academic | Professor |
| 5 | Professional | Data Scientist | 23 | Academic | Graduate Researcher |
| 6 | Academic | Professor | 24 | Professional | Data Scientist |
| 7 | Professional | Consultant | 25 | Professional | Data Scientist |
| 8 | Professional | Data Scientist | 26 | Professional | Analytics Director |
| 9 | Academic | Researcher | 27 | Professional | Freelancer |
| 10 | Professional | Business Analyst | 28 | Professional | Senior Business Intelligence Specialist |
| 11 | Academic | Researcher | 29 | Professional | Data Scientist & AI Lead |
| 12 | Academic | Lecturer/ researcher | 30 | Academic | Associate Professor |
| 13 | Professional | Data Scientist | 31 | Professional | CSO, R&D Director |
| 14 | Academic | PhD Student | 32 | Professional | Manager |
| 15 | Professional | CEO | 33 | Professional | Lead Data Scientist |
| 16 | Professional | Business Intelligence Specialist | 34 | Professional | Data Analytics Consultant |
| 17 | Professional | Data Scientist | 35 | Academic | Researcher |
| 18 | Professional | Data Scientist | | | |

### 4.1. The Performance Metric Weights

First, the weights of all 35 experts will be provided in a tabular format (see Appendix 1: Results of the Best-Worst Method). The results for the mean weights can be evaluated in Table 13.

Table 13. Aggregated weights of the Best-Worst Method

| | Accuracy | Precision | AUC | Easy to Interpret | Fastness | Expected Value Framework |
|---|---|---|---|---|---|---|
| Professionals | 0.22 | 0.13 | 0.25 | 0.14 | 0.11 | 0.15 |
| Academics | 0.21 | 0.19 | 0.16 | 0.15 | 0.14 | 0.15 |
| Mean (50/50) | 0.21 | 0.16 | 0.21 | 0.15 | 0.12 | 0.15 |

The results of the two separate groups will be compared first. The professionals selected AUC to be the most important performance metric and fastness was their least important metric. The academics show a more balanced distribution of the weights. The scholars selected accuracy to be the most important metric of performance and selected fastness as the least important metric. The results of the aggregate from all scores and the average with equal weights for the two groups only differ slightly.

Table 14 shows the transformation process of going from six to five performance metrics (removing expected value framework – EVF). This was done because the experiment did not have values connected to (in)correct predictions and thus, the expected value framework was not suitable. From the results of the BMW (see Table 14), it becomes clear that accuracy is the most important performance metric with a weight of 0.250. Very closely after the accuracy, the AUC follows with a weight of 0.241. The third place is taken by the precision metric that scores 0.191. Ease of interpretation follows with a score of 0.176. The least important performance metric is fastness with a score of 0.143.

Table 14. Removing the expected value framework metric

| Average (50/50) | Accuracy | Precision | AUC | Easy to Interpret | Fastness |
|---|---|---|---|---|---|
| Before normalizing | 0.212 | 0.163 | 0.205 | 0.149 | 0.121 |
| **After normalizing** | **0.250** | **0.191** | **0.241** | **0.176** | **0.143** |

Note: the selected group for further analysis is printed in **bold**

## 4.2. Results of the Churn Forecasting Methods

The methods that have been selected all proved to be appropriate methods for forecasting churn, which is demonstrated in Section 2.1. The results of the statistical methods are based on the final methods that have been selected following the description in Section 3.2. The performance of the methods is evaluated on five performance metrics: accuracy, precision, AUC, ease of interpretation, and fastness.

First, the confusion matrices of all six methods are displayed (see Table 15). These have been used to calculate the accuracy, precision, and the AUC of the methods. Next, the full results will be shown with the original data (see Table 15). Eventually, it is important that every metric is expressed with a number with the same range (between 0 and 100).

Table 15. Confusion matrices of the applied methods

| **Decision Tree** | | **Actual class** | | **Random forest** | | **Actual class** | |
|---|---|---|---|---|---|---|---|
| | | Positive condition | Negative condition | | | Positive condition | Negative condition |
| Predicted class | Positive prediction | 1264 | 447 | Predicted class | Positive prediction | 6645 | 2585 |
| | Negative prediction | 739 | 1550 | | Negative prediction | 3370 | 7400 |

| **Logistic Regression** | | **Actual class** | | **Gradient Boosted Trees** | | **Actual class** | |
|---|---|---|---|---|---|---|---|
| | | Positive condition | Negative condition | | | Positive condition | Negative condition |
| Predicted class | Positive prediction | 1374 | 776 | Predicted class | Positive prediction | 8000 | 4281 |
| | Negative prediction | 629 | 1221 | | Negative prediction | 2015 | 5704 |

| **Support Vector Machine** | | **Actual class** | | **Multi-layer perceptron Neural Network** | | **Actual class** | |
|---|---|---|---|---|---|---|---|
| | | Positive condition | Negative condition | | | Positive condition | Negative condition |
| Predicted class | Positive prediction | 1393 | 831 | Predicted class | Positive prediction | 598 | 198 |
| | Negative prediction | 610 | 1166 | | Negative prediction | 409 | 795 |

Table 16 shows all results from the final forecasting methods. For now, the focus will be on the ease of interpretation and the fastness of the methods, because they still need to be transformed. The process of transformation for all metrics of performance is described in Section 3.3.

Table 17 presents categorical variables transferred into ratios of 100. The column with the aggregate takes the average of the values in the columns before that. This is the value that will be used for further analysis. Using these new aggregate values, a new table with results is created which can be examined in Table 18.

Table 16. Summary of the results of the optimized methods

| | **Performance metrics** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Accuracy (%)** | **Precision (%)** | **AUC** | **Ease of interpretation** | | **Fastness** | | | |
| **Method** | | | | Able to explain decisions | Able to run additional analysis | Computation time | Building prototype | Optimizing parameters | RapidMiner |
| **Decision Tree** | 70.35 | 73.87 | 0.769 | Very high | Very high | 1 | Very fast | Fast | Yes |
| **Logistic Regression** | 64.88 | 63.91 | 0.702 | High | Very high | 1 | Very fast | Fast | Yes |
| **SVM** | 63.98 | 62.63 | 0.698 | High | High | 26 | Average | Average | Yes |
| **Random forest** | 70.22 | 71.99 | 0.771 | High | High | 30 | Very fast | Fast | Yes |
| **GBT** | 68.52 | 65.14 | 0.772 | High | High | 38 | Fast | Average | Yes |
| **Neural network** | 69.65 | 75.13 | 0.763 | Low | Average | 3 | Slow | Average | No |

Table 17. Aggregating the ease of interpretation and fastness into one score

| **Metric** | **Ease of interpretation** | | | **Fastness** | | | | |
|---|---|---|---|---|---|---|---|---|
| | Able to explain decision | Able to run additional analysis | **Aggregate** | Computation time | Building prototype | Optimizing parameters | RapidMiner | **Aggregate** |
| Decision Tree | 100 | 100 | **100.00** | 99 | 100 | 75 | 100 | **93.50** |
| Logistic Regression | 75 | 100 | **87.50** | 99 | 100 | 75 | 100 | **93.50** |
| SVM | 75 | 75 | **75.00** | 74 | 50 | 50 | 100 | **68.50** |
| Random forest | 75 | 75 | **75.00** | 70 | 100 | 75 | 100 | **86.25** |
| GBT | 75 | 75 | **75.00** | 62 | 75 | 50 | 100 | **71.75** |
| Neural network | 25 | 50 | **37.50** | 97 | 25 | 50 | 60 | **58.00** |

Note: the selected group for further analysis is printed in **bold**

Table 18. Results of the churn forecasting methods

| | **Accuracy** | **Precision** | **AUC** | **Ease of interpretation** | **Fastness** |
|---|---|---|---|---|---|
| **Decision Tree** | 70.35% | 73.87% | 0.769 | 100.0 | 93.50 |
| **Logistic Regression** | 64.88% | 63.91% | 0.702 | 87.5 | 93.50 |
| **Support Vector Machine** | 63.98% | 62.63% | 0.698 | 75.0 | 68.50 |
| **Random forest** | 70.22% | 71.99% | 0.771 | 75.0 | 86.25 |
| **Gradient Boosted Trees** | 68.52% | 65.14% | 0.772 | 75.0 | 71.75 |
| **Neural network** | 69.65% | 75.13% | 0.763 | 37.5 | 58.00 |

From these results we learn that the Decision tree scores the highest on the accuracy and ease of interpretation. Additionally, the Decision tree is tied with the Logistic regression for the best performance regarding the fastness of the method. The Neural network has the highest precision and the lowest score for ease of interpretation and fastness of the method. The GBT show the highest AUC followed closely by the Random forest. The SVM resulted in the lowest scores for the accuracy, precision, and the AUC.

In Table 19, we report the overall performance of the six statistical methods (see the additive value function, equation 10). This will be done using matrix multiplication of the weights on the ratios from our performance metrics. The highest overall performance is scored by the Decision tree with a total score of 81.11. The Random forest comes in second with a total score of 75.34. The Logistic regression follows with a score of 74.03. The GBT got a total performance score of 71.57. The SVM comes in fifth with a total score of 67.70. The lowest total score is realized by the selected Neural network with a total score of 65.00.

Table 19. Calculation of the aggregated performance score

| | **Accuracy** | **Precision** | **AUC** | **Ease of interpretation** | **Fastness** | **Total score** |
|---|---|---|---|---|---|---|
| Decision Tree | 70.35 | 73.87 | 76.90 | 100.00 | 93.50 | |
| Logistic Regression | 64.88 | 63.91 | 70.20 | 87.50 | 93.50 | |
| SVM | 63.98 | 62.63 | 69.80 | 75.00 | 68.50 | |
| Random forest | 70.22 | 71.99 | 77.10 | 75.00 | 86.25 | |
| GBT | 68.52 | 65.14 | 77.20 | 75.00 | 71.75 | |
| Neural network | 69.65 | 75.13 | 76.30 | 37.50 | 58.00 | |
| **Weights** | **0.25** | **0.19** | **0.24** | **0.18** | **0.14** | 1 |
| Decision Tree | 17.56 | 14.12 | 18.55 | 17.55 | 13.32 | **81.11** |
| Logistic Regression | 16.19 | 12.22 | 16.93 | 15.36 | 13.32 | **74.03** |
| SVM | 15.97 | 11.98 | 16.84 | 13.16 | 9.76 | **67.70** |
| Random forest | 17.53 | 13.77 | 18.60 | 13.16 | 12.29 | **75.34** |
| GBT | 17.10 | 12.46 | 18.62 | 13.16 | 10.22 | **71.57** |
| Neural network | 17.38 | 14.36 | 18.40 | 6.58 | 8.27 | **65.00** |

Finally, Table 20 presents the scores of each method with respect to individual performance metrics showing that the ranking of methods is different for individual metrics. It also shows that the decision tree is the best performing method when looking at both the accuracy and the aggregated overall value coming from BWM. The GBT followed by the Random forest score highest for the AUC metric.

Table 20. Performance of the methods measured by accuracy, AUC and overall performance

| **Rank** | **Statistical method** | **Accuracy** | **Statistical method** | **AUC** | **Statistical method** | **Best-worst method** |
|---|---|---|---|---|---|---|
| **1** | Decision Tree | 70.35% | Gradient Boosted Trees | 0.772 | Decision Tree | 81.106 |
| **2** | Random forest | 70.22% | Random forest | 0.771 | Random forest | 75.342 |
| **3** | Neural network | 69.65% | Decision Tree | 0.769 | Logistic Regression | 74.025 |
| **4** | Gradient Boosted Trees | 68.52% | Neural network | 0.763 | Gradient Boosted Trees | 71.565 |
| **5** | Logistic Regression | 64.88% | Logistic Regression | 0.702 | Support Vector Machine | 67.705 |
| **6** | Support Vector Machine | 63.98% | Support Vector Machine | 0.698 | Neural network | 64.998 |

## 5. Discussion

Using statistical methods to forecast customer churn is an interesting field for both academics and professionals. Companies can benefit immensely from slight improvements of the performance of a method. Best practices are constantly being outperformed by new methods. This research provides the tools to compare forecasting methods on multiple metrics of performance. Researchers and companies can choose to use the weights provided in this study to indicate the relative importance of the metrics. Moreover, the methods of this thesis provide practitioners with the tools to conduct the best-worst method themselves. Allowing them to involve all stakeholders in selecting a forecasting method. Although the focus of this research is on churn forecasting, the results and methods can be beneficial to other types of binary classifiers. However, practitioners should reconsider the performance metrics when this is done. The expected value framework can, for example, not be applied to a predictor that has no monetary outcomes.

The performance of the methods is somewhat different from our expectations. The assumption was that an advanced method like the Neural network would outperform the traditional statistical methods on the AUC and accuracy metrics. This is true regarding the Logistic regression and SVM. However, the Decision tree outperforms the two ensemble methods on both performance dimensions. Additionally, the ensemble methods were expected to improve the AUC and accuracy metric at the cost of the fastness and ease of interpretation metric. Overall, this seems to be the case in our research. This is suggested by the measurement of the AUC for which the ensemble methods come out as the best performing methods. However, due to the extraordinary performance of the Decision tree on the dataset for this research project, this did not end up being the case.

Conversely, it was expected that ensemble methods would outperform Neural networks on the ease of interpretation and fastness metrics. This shows in our research since the Neural network comes out last on both these dimensions.

Moreover, this research provides an overview of the statistical methods and performance metrics that are being applied for churn forecasting. This information can be used as a starting point for practitioners in selecting the appropriate tools for analysis. The selection of methods and performance metrics is based on the literature review and should not be treated as the exclusive best practice. Researchers and professionals are encouraged to experiment with other methods and performance metrics. Additionally, the performance of the methods might be completely different if there is a change in the size or cleanliness of the dataset that will be used for analysis. The dataset in this research had zero missing values. Neural networks are known to perform well on unstructured and uncleaned data. Moreover, more complex methods are more likely to benefit from additional training data as they have more relationship to understand and optimize. Future research may attempt to replicate this research with a larger dataset that includes missing values to test this.

Finally, it is important to understand that the BWM should only be applied if the criteria are independent of each other. Although our measurements of performance all could be considered independent. It should be noted that both accuracy and precision are related to the true positives of the statistical method. However, a statistical method could be precise, but not accurate. It could therefore be argued that the performance metrics are independent of each other. Nonetheless, it should be considered by future researchers when selecting their criteria for a MCDM method.

## 6. Conclusions

There seems to be no clear best practice in choosing statistical methods for churn forecasting and previous research has shown varying results with similar methods. Furthermore, the comparison of the methods of the performance is often based on just one performance metric. This paper shows that combining several metrics and involving the opinion of practitioners and scholars could lead to different results. We think that an aggregated value such as the one we developed in our study is more meaningful than relying on a single metric to select a method, especially in cases the performance of a method with respect to different metrics is varying.

Based on both a qualitative analysis of the literature and a quantitative comparative analysis of the statistical methods, it can be concluded that the best performing statistical method to forecast churn is the Decision tree. The results indicate that the method outperforms the other methods on the accuracy, ease of interpretation, and the fastness metric of performance.

The results of the literature review uncover that there are 12 types of statistical methods being applied to predict customer churn in the top-rated journals. The six statistical methods that were applied most frequently in the existing literature on churn prediction have been selected for further analysis. These methods are: the Decision tree, Logistic regression, SVM, Random forest, GBT, and Neural networks. Furthermore, the literature review helped in bringing to light what metrics of performance are being employed. In total, 12 metrics of performance have been identified. Six of these metrics have been selected for the BWM questionnaire. For the final evaluation of the methods, five metrics of performance have been used. Those five metrics are: accuracy, precision, AUC, ease of interpretation, and fastness.

The BWM focused on these six metrics. From the opinion of 35 industry experts this research has succeeded in weighting these metrics of performance. The performance metrics, from most important to least important, are accuracy, AUC, precision, expected value framework, ease of interpretation, and fastness of the method. The results from the BWM questionnaire also show that professional data scientists judge AUC as the most important performance metric, while academics judge that accuracy and precision to be more important.

The best performing statistical method measured by the accuracy and the BWM is the Decision tree. However, when looking at the AUC metric, the GBT and Random forest outperform the Decision tree. It is shown that the order of the best performing method is different over all three metrics of performance. This emphasises the usefulness of introducing a new metric that considers multiple metrics of the performance of a statistical method.

### 6.1. Limitations and Future Research

Here we discuss some of the limitations of our study. Firstly, the dataset used for analysis does not include any information about the time or date and is only focused on one contract period. Most customer databases that are used for churn forecasting consist of time-series data. The results of the statistical methods can be different when applied to such a dataset. Moreover, a Recurrent neural network architecture can be applied to this type of data, for which the literature suggests a good fit and performance. Another limitation also comes from working with a scientific dataset. As it is a theoretical dataset, there is no information available on the costs and benefits involved with correctly or incorrectly targeting a customer with a retention offer. Since this information is necessary to calculate the profit curve for the expected value framework, this metric of performance had to be excluded from the experiment.

As part of the BWM it is advised to give immediate feedback to the experts whose pairwise comparisons are not sufficiently consistent. In our research we did not provide feedback to those experts which was resulted in eliminating some of the data, which otherwise could have resulted in a larger sample size, hence more reliable conclusions.

There are several other ways in which future research can contribute to this work. First, a replication of this research on a real-life time-series dataset can contribute to the findings of this research by comparing the performance of the methods of the two studies. Moreover, it is recommended for any practitioner or researcher planning to do this to put some time and effort into identifying the involved stakeholders that will be working with the method, and to perform the best-worst method with the opinion of those people. Additionally, the comparison of the methods on multiple metrics of performance provides great additional value. An opportunity is identified to build loss functions and other optimization functions that use multiple metrics of performance when optimizing a method.

The fastness and ease of interpretation metric both were underrepresented in the existing literature. The articles that included these methods showed different interpretations of these metrics and different methods of calculating them. We plea for more qualitative research that could focus on identifying what factors influence these methods.

Furthermore, there was a significant difference in the ranges (or standard deviation) of the different metrics that have been measured on this research. This has not been taken into consideration for this study, but a study with a larger standard deviation shows characteristics of a metric with a larger weight. Therefore, future research should consider transforming the final scores of their metrics to scores that have similar standard deviations. Finally, the inclusion of lift and the Gini coefficient as performance metrics can be interesting for future research.

## References

Alon, I., Qi, M., and Sadowski, R. J. (2001) 'Forecasting Aggregate Retail Sales:: A Comparison of Artificial Neural Networks and Traditional Methods', *Journal of Retailing and Consumer Services*, 8, 147-56.

Ascarza, E. (2018) 'Retention Futility: Targeting High-Risk Customers Might Be Ineffective', *Journal of Marketing Research*, 55, 80-98.

Ascarza, E., Iyengar, R., and Schleicher, M. (2016) 'The Perils of Proactive Churn Prevention Using Plan Recommendations: Evidence from a Field Experiment', *Journal of Marketing Research*, 53, 46-60.

Barfar, A., Padmanabhan, B., and Hevner, A. (2017) 'Applying Behavioral Economics in Predictive Analytics for B2B Churn: Findings from Service Quality Data', *Decision Support Systems,* 101, 115-27.

Beliën, J., and Forcé, H. (2012) 'Supply Chain Management of Blood Products: A Literature Review', *European Journal of Operational Research*, 217, 1-16.

Bottomley, P. A., and Doyle, J. R. (2001) 'A Comparison of Three Weight Elicitation Methods: Good, Better, and Best', *Omega*, 29, 6: 553-60.

Chen, Z. Y., and Fan, Z. P. (2012) 'Distributed Customer Behavior Prediction Using Multiplex Data: A Collaborative MK-SVM Approach', *Knowledge-Based Systems*, 35, 111-19.

Chu, B. H., Tsai, M. S., and Ho, C. S. (2007) 'Toward a Hybrid Data Mining Model for Customer Retention', *Knowledge-Based Systems*, 20, 703-18.

De Cnudde, S., and Martens, D. (2015) 'Loyal to Your City? A Data Mining Analysis of a Public Service Loyalty Program', *Decision Support Systems*, 73, 74-84.

Collobert, R., and Weston, J. (2008) 'A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning', *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 160-67.

Coussement, K., Lessmann, S., and Verstraeten, G. (2017) 'A Comparative Analysis of Data Preparation Algorithms for Customer Churn Prediction: A Case Study in the Telecommunication Industry', *Decision Support* Systems, 95, 27-36.

Coussement, K., and Van den Poel, D. (2008) 'Churn Prediction in Subscription Services: An Application of Support Vector Machines While Comparing Two Parameter-Selection Techniques', *Expert Systems with Applications* 34, 313-27.

Delen, D. (2010) 'A Comparative Analysis of Machine Learning Techniques for Student Retention Management', *Decision Support Systems*, 49, 498-506.

Dierkes, T., Bichler, M., and Krishnan, R. (2011) 'Estimating the Effect of Word of Mouth on Churn and Cross-Buying in the Mobile Phone Market with Markov Logic Networks', *Decision Support Systems*, 51, 361-71.

Dyer, J. S., and Sarin, R. K. (1979) 'Measurable Multiattribute Value Functions', *Operations Research*, 27, 810-22.

Edwards, W. (1977) 'How to Use Multiattribute Utility Measurement for Social Decisionmaking', *IEEE Transactions on Systems, Man, and Cybernetics*, 7, 326-40.

Eskildsen, J., and Kristensen, K. (2007) 'Customer Satisfaction–The Roleof Transparency', *Total Quality Management & Business Excellence*, 18, 39-47.

Evans, M. (2002) 'Prevention Is Better than Cure: Redoubling the Focus on Customer Retention', *Journal of Financial Services Marketing*, 7, 186-98.

Ganjisaffar, Y., Caruana, R., and Lopes, C. V. (2011) 'Bagging Gradient-Boosted Trees for High Precision, Low Variance Ranking Models', *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Beijing China, 85-94.

Ha, K., Cho, S., and MacLachlan, D. (2005) 'Response Models Based on Bagging Neural Networks', *Journal of Interactive Marketing*, 19, 17-30.

Hornik, K., Stinchcombe, M., and White, H. (1990) 'Universal Approximation of an Unknown Mapping and Its Derivatives Using Multilayer Feedforward Networks', *Neural Networks*, 3, 551-560.

Howard, J., & Thomas, R. (2021) *Practical Deep Learning for Coders Version 3*, available at https://course.fast.ai/.

Hsu, C. W., Chang, C. C., and Lin, C. J. (2010) 'A Practical Guide to Support Vector Classification', National Taiwan University, Taiwan.

Hsu, C. W., and Lin, C. J. (2002) 'A Comparison of Methods for Multiclass Support Vector Machines', *IEEE Transactions on Neural Networks*, 13, 415-25.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013) *An introduction to statistical learning*, 112, 18. Springer, New York.

Jesson, J., Matheson, L., and Lacey, F. M. (2011) *Doing Your Literature Review: Traditional and Systematic Techniques*, Sage publications, California, United States.

Karn, U. (2016) *An Intuitive Explanation of Convolutional Neural Networks.* available at https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

Keeney, R. L., and Raiffa, H. (1976) *Decisions with Multiple Objectives (Preferences and Value Tradeoffs)*. Wiley, New York, United States.

Kitchens, B., Dobolyi, D., Li, J., and Abbasi, A. (2018) 'Advanced Customer Analytics: Strategic Value through Integration of Relationship-Oriented Big Data', *Journal of Management Information Systems*, 35, 540-74.

Knol, M. J., Van der Tweel, I., Grobbee, D. E., Numans, M. E., and Geerlings, M. I. (2007) 'Estimating Interaction on an Additive Scale between Continuous Determinants in a Logistic Regression Model', *International Journal of Epidemiology*, 36, 1111-1118.

Larivière, B., and Van den Poel, D. (2004) 'Investigating the Role of Product Features in Preventing Customer Churn, by Using Survival Analysis and Choice Modeling: The Case of Financial Services', *Expert Systems with Applications*, 27, 277-85.

Lee, H., Lee, Y., Cho, H., Im, K., and Kim, Y. S. (2011) 'Mining Churning Behaviors and Developing Retention Strategies Based on a Partial Least Squares (PLS) Model', *Decision Support Systems*, 52, 207-16.

Lee, H. J., Shin, H., Hwang, S. S., Cho, S., and MacLachlan, D. (2010) 'Semi-Supervised Response Modeling', *Journal of Interactive Marketing*, 24, 42-54.

Lee, Y. H., Wei, C. P., Cheng, T. H., and Yang, C. T. (2012) 'Nearest-Neighbor-Based Approach to Time-Series Classification', *Decision Support Systems*, 53, 207-17.

Lemmens, A., and Croux, C. (2006) 'Bagging and Boosting Classification Trees to Predict Churn', *Journal of Marketing Research*, 43, 276-286.

Liang, F., Brunelli, M., and Rezaei, J. (2020) 'Consistency Issues in the Best Worst Method: Measurements and Thresholds', *Omega*, 96,102175.

Mahajan, V., Misra, R., and Mahajan, R. (2015) 'Review of Data Mining Techniques for Churn Prediction in Telecom', *Journal of Information and Organizational Sciences*, 39, 183-97.

Marblestone, A. H., Wayne, G., and Kording, K. P. (2016) 'Toward an Integration of Deep Learning and Neuroscience', *Frontiers in Computational Neuroscience*, 10, 94-107.

Mestre, M. R., and Vitoria, P. (2013) 'Tracking of Consumer Behaviour in E-Commerce', In *Proceedings of the 16th International Conference on Information Fusion*, Istanbul, Turkey, 1214-1221.

Moeyersoms, J., and Martens, D. (2015) 'Including High-Cardinality Attributes in Predictive Models: A Case Study in Churn Prediction in the Energy Sector', *Decision Support Systems*, 72, 72-81.

Moser, S., Schumann, J. H., Von Wangenheim, F., Uhrich, F., and Frank, F. (2018) 'The Effect of a Service Provider's Competitive Market Position on Churn among Flat-Rate Customers', *Journal of Service Research*, 21, 319-35.

Neslin, S. A, Gupta, S., Kamakura, W., Lu, J., and Mason, C. H. (2006) 'Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models', *Journal of Marketing Research*, 43, 204-11.

Nguyen, T., Li, Z., Spiegler, V., Ieromonachou, P., and Lin, Y. (2018) 'Big Data Analytics in Supply Chain Management: A State-of-the-Art Literature Review', *Computers & Operations Research*, 98, 254-64.

Prinzie, A., and Van den Poel, D. (2006) 'Incorporating Sequential Information into Traditional Classification Models by Using an Element/Position-Sensitive SAM', *Decision Support Systems*, 42, 508-26.

Provost, F., and Fawcett, T. (2013) *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media, Inc., USA.

Rezaei, J. (2015) 'Best-Worst Multi-Criteria Decision-Making Method', *Omega*, 53, 49-57.

Rezaei, J. (2016) 'Best-Worst Multi-Criteria Decision-Making Method: Some Properties and a Linear Model', *Omega*, 64, 126-30.

Rezaei, J. (2018) 'Piecewise Linear Value Functions for Multi-Criteria Decision-Making', *Expert Systems with Applications*, 98, 43-56.

Rezaei, J. (2020) 'A Concentration Ratio for Nonlinear Best Worst Method', *International Journal of Information Technology & Decision Making*, 19, 891-907.

Risselada, H., Verhoef, P. C., and Bijmolt, T. H. A. (2010) 'Staying Power of Churn Prediction Models', *Journal of Interactive Marketing*, 24, 198-208.

Saaty, T. L. (1977) 'A Scaling Method for Priorities in Hierarchical Structures', *Journal of Mathematical Psychology*, 15, 234-81.

Schoonjans (2018) *ROC Curve Analysis with MedCalc*. available at https://www.medcalc.org/manual/roc-curves.php.

Sharma, A., Panigrahi, Dr., and Kumar, P. (2013) 'A Neural Network Based Approach for Predicting Customer Churn in Cellular Network Services', *arXiv Preprint*, arXiv:1309.3945.

Tamaddoni, A., Stakhovych, S., and Ewing, M. (2016) 'Comparing Churn Prediction Techniques and Assessing Their Performance: A Contingent Perspective', *Journal of Service Research*, 19, 123-41.

Thornbury, J. R., and Fryback, D. G. (1992) 'Technology Assessment—an American View', *European Journal of Radiology*, 14, 147-56.

Weinman, J. (2018) 'The Economics of Pay-per-Use Pricing', *IEEE Cloud Computing*, 5, 101-133.

Wieringa, J. E., and Verhoef, P. C. (2007) 'Understanding Customer Switching Behavior in a Liberalizing Service Market: An Exploratory Study', *Journal of Service Research*, 10, 174-86.

Yelle, L. E. (1979) ''The learning curve: historical review and comprehensive survey', *Decision Sciences*, 10, 302-28.

Yu, X., Guo, S., Guo, J., and Huang, X. (2011) 'An Extended Support Vector Machine Forecasting Framework for Customer Churn in E-Commerce', *Expert Systems with Applications*, 38, 1425-30.

Zhang, X., Zhu, J., Xu, S., and Wan, Y. (2012) 'Predicting Customer Churn through Interpersonal Influence', *Knowledge-Based Systems*, 28, 97-104.

## Author Statement

**Ronan Duchemin**: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing – Original Draft, Writing - Review & Editing, Visualization, Project administration

**Ricardo Matheus**: Conceptualization, Methodology, Validation, Investigation, Resources, Writing - Review & Editing, Supervision, Project administration