Churn Analysis Using Deep Learning: Customer Classification from a Practical Point of View

Tobias Albrecht and Daniel Baier

Abstract The business relevance of customer churn analysis is increasing due to the growing availability of corresponding data and intensifying competition. Here, especially the predictive accuracy of modeling approaches is in the focus of researchers and practitioners alike, with deep neural networks recently becoming an attractive method due to their high performance in a variety of fields. However, from a practical point of view, other factors such as the ease of application and model interpretability are also to be considered. These aspects are generally viewed as shortcomings of deep neural networks. Therefore, a novel framework for the application of deep learning in churn analysis is developed and tested in a practical setting. It is shown, that a less complex application procedure and more easily interpretable prediction modeling can be achieved.

Tobias Albrecht · Daniel Bayer
University of Bayreuth, Chair of Marketing and Innovation
Universitätsstraße 30, 95447 Bayreuth, Germany
✉ tobias.albrecht@uni-bayreuth.de
✉ daniel.baier@uni-bayreuth.de

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 6, No. 2, 2020

DOI: 10.5445/KSP/1000098012/04
ISSN 2363-9881

1 Introduction

In competitive markets with innovative technologies and global competition, the existing customer base is considered one of a company's most important assets (Kumar and Reinartz, 2016). Accordingly, customer retention is a determinant of persistent economic success (Larivière and van den Poel, 2005). In that matter, the early identification and forecast of customer churn is of key relevance (Hung et al., 2006). In the course of the digitization of customer relationships and the resulting increase in data volume and complexity, currently, the demands placed on the analytical methods used as well as on their forecasting quality are rising (Keramati et al., 2014). At the same time, developments in the area of artificial intelligence (AI) entail profound economic and societal changes, particularly related to the increasing automation of information processing and decision-making, supported by an almost unlimited flow of information (LeCun et al., 2015). A central technique to the progress in a variety of practical applications is deep learning based on artificial neural networks (Zhang et al., 2018). However, in customer data analysis, the exploration and practical integration of deep learning as an innovative forecasting method is still in its infancy (Najafabadi et al., 2015).

The purpose of this paper lies in the investigation of the potential of deep learning in customer churn analysis from a practical point of view by identifying key model requirements for the users as well as examining propositions for their implementation. Its main contribution is the development of an analytical framework to increase the applicability of deep learning models with regard to procedural complexity and interpretability. The remainder of this paper is structured as follows: The next section gives a brief overview of deep learning and discusses its current application potential in customer analysis. Next, Section 3 provides an introduction to customer churn analysis and its methodological challenges. Section 4 then defines the experimental design of the study and introduces an analytical framework for the implementation of deep neural networks in churn analysis with special focus on practical applicability through reduced procedural complexity and increased model interpretability. Subsequently, in Section 5, empirical findings are presented and applied methods are assessed from a user perspective. Finally, a conclusion is drawn and practical implications are derived in Section 6.

2 Deep Learning and its Application in Customer Analysis

2.1 Background and Functionality

Deep Learning, in the literature also referred to as deep structured learning or hierarchical learning, has emerged since 2006 as a new area of machine learning research (Hinton et al., 2006). The research field in its present form builds on the neuroscientifically influenced development of plain linear models and the subsequent evolution to simple artificial neural networks through the structured interconnection of neurons and associated training methods (Rumelhart et al., 1986). The aforementioned advances have enabled deep learning methods, as a class of machine learning techniques, that use multiple layers of information processing and abstraction to effectively exploit complex, compositional nonlinear functions for supervised or unsupervised learning of underlying feature representations (LeCun et al., 2015). While different definitions for these techniques belonging to the category of deep learning exist, two core elements of all approaches are identified by Deng (2014):

1. The structured composition of several layers or phases of nonlinear information processing, and
2. the hierarchical way of learning from feature representations of different levels of abstraction.

Accordingly, the characterizing depth of such models results from either a greater amount of computational stages or of learned concepts. However, so far no consensus among researchers on a quantifiable minimum depth for the inclusion in this model category is reached (Schmidhuber, 2015).

2.2 Application Potential

The most distinguishing feature and strength of artificial neural networks in general is their ability to automatically extract distinctive information by uncovering underlying patterns in available data representations, which most notably results in their excellent modeling power (Keramati and Ardabili, 2011). When it comes

to the efficiency in putting this potential to use, especially for the representation of complex functions, deep network architectures have an advantage over shallow ones. According to Bengio and LeCun (2007) this is achieved by overcoming the limitations of shallow models regarding the compactness of representation with respect to the overall required number of computational units (depth-breadth tradeoff), the amount of necessary training examples with increasing input dimensionality (curse of dimensionality) and the computational cost of learning with high volume of data. The resulting superiority of deep architectures in the representation of abstract functions ultimately makes them applicable for highly complex tasks (LeCun et al., 2015). To this end, various new architectures beyond basic feedforward concepts, including recurrent networks (Hochreiter and Schmidhuber, 1997), convolutional networks (LeCun et al., 1989) and deep autoencoders (Hinton and Salakhutdinov, 2006) have been developed.

In the field of customer analysis, the strengths of deep learning algorithms identified by literature are most recently viewed as a chance to cope with the high demands caused by the accumulation of large amounts of customer data (Najafabadi et al., 2015). Due to the associated rising data complexity, besides classical data mining techniques, deep learning algorithms are increasingly being used for classification and regression tasks to accurately predict customer behavior and, therefore, optimize resource allocation and the companies' responses to customer needs (Wedel and Kannan, 2016).

Looking at the application of deep learning models for customer analysis in business practice however, there are still reservations and inhibiting factors affecting the practical implementation. First, the complexity of model configuration and optimization is affected, as it is perceived as an unstandardized and time-consuming process (Paliwal and Kumar, 2009). Another unappealing characteristic trait of deep neural networks to practitioners involves the lack of interpretability. By hiding their internal logic to the user, they fall in the category labeled as black box models (Guidotti et al., 2019). This term characterizes systems, that do not provide meaning in understandable manner to humans. This implies in particular not enabling explanations for its reasoning, from which e.g. inferences regarding the significance of certain variables could be drawn (Doshi-Velez and Kim, 2017).

To overcome this weak point of low interpretability in many machine learning systems, two general model properties worth striving for are proposed by Lipton (2016). Accordingly, interpretability in supervised machine learning models

can be achieved through transparency at the level of individual components (simulatability), the entire model (decomposability) or the learning algorithm (algorithmic transparency) as well as through post-hoc interpretations in the form of text explanations, visualization, local explanations or explanation by example. In this context, applied research in various scopes has, amongst others, been provided by Baehrens et al. (2010) and Ribeiro et al. (2016b).

3 Customer Churn Analysis

3.1 Methodology and Central Challenges

Customer churn prediction as the early identification and forecast of threatening, customer-initiated termination of the business relationship is a crucial part of the broader field of churn management in customer relationship management (CRM) (Lejeune, 2001). Its importance arises from its role as the analytical basis for customer retention strategies implemented by companies to directly deal with churn and its consequential influence on the profitability of businesses (Ganesh et al., 2000). The economic value of long-term customers is widely recognized in literature (Rosenberg and Czepiel, 1984). Desired targeted and proactive concepts of churn management in that matter are characterized by the aim of identifying customers with high inclination to abandon the company at an early stage to specifically address them with customer retention programs and incentives in time (Burez and van den Poel, 2007). For the purpose of a sound analytical basis for this approach, researchers examine customer churn prediction and potential strategies for improvement from two different perspectives. While descriptive studies focus on understanding the underlying factors and main drivers of customer churn (e.g. Ahn et al., 2006; Keaveney, 1995), predictive research aims at improving churn prediction results by developing and enhancing prediction models and classification algorithms (e.g. Verbeke et al., 2012).

Customer churn analysis has been described in literature as a five-step process, that is closely linked to the stages of the general data mining procedure (Datta et al., 2000; Hadden et al., 2007). It consists of the following phases:

1. Data selection, which is about identifying the optimal customer data with regard to relevant data sources and input volume to fit the predefined problem statement and methodological intent (Hung et al., 2006),

2. Data semantics, where the focus is on understanding and interpreting the data in the given context for a correct usage in the following phases (Datta et al., 2000),
3. Feature selection as the process of subsetting the customer feature space by removing redundant, irrelevant or noisy variables to decrease the computational complexity of the prediction model (Huang et al., 2010),
4. Model development, where a predictive model is built to forecast the customers' churn behavior based on underlying patterns in the input data (Neslin et al., 2006; Verbeke et al., 2011),
5. Validation of results, that includes validating the model performance and evaluating its prediction results (Verbeke et al., 2012).

For the application of recent advances beyond traditional data mining like machine learning methods in this context however, adaptions and alterations seem indispensable.

Along with the different angles in research to churn analysis as a managerial or statistical prediction problem, specific challenges to the overall procedure can be identified. From a predictive point of view, class imbalance stemming from low churn rates is a central problem, as the objects of interest are located in the minority class and, therefore, pose a challenge to model learning as well as performance evaluation (Zhu et al., 2017). Moreover, apart from structural aspects of the data, the higher relative economic significance of churners compared to non-churners results in asymmetric misclassification costs (Weiss, 2004). When it comes to the managerial and practical application aspects of churn analysis, the need for a probabilistic classification output as the basis for subsequent customer retention activities is to be considered (Burez and van den Poel, 2007). Besides, not only the segmentation of customers as a feature of the prediction output, but also insights in the drivers of customer churn as a result of that process are of crucial importance (Ahn et al., 2006). Therefore, according to Verbeke et al. (2011), churn prediction models have to be interpretable and overall comprehensible to fulfill the demands of practical use.

3.2 Prediction Modeling

The selection of the optimal classification technique can be identified as one of the most widely studied fields of churn analysis and is therefore credited with high relevance for its outcome. The present churn prediction task consists of the binary classification of the future behavior of, by then, unknown customers as churners or non-churners based on historical data that is used for model training (Lima et al., 2011). For that purpose, various categories of prediction models such as logistic regression (Kumar et al., 1995), decision trees and random forests (Lemmens and Croux, 2006) as well as neural networks (Huang et al., 2012) have been applied and tested in the literature so far.

While the first two techniques are considered as the most popular due to their fast and robust functioning, results regarding the model performance in churn prediction are ambiguous and commonly depend on the research setting and the methodological focus (Buckinx and van den Poel, 2005; Verbeke et al., 2012). Exemplarily, Neslin et al. (2006) and Huang et al. (2012) found logistic and tree approaches to perform best in the early identification of churning customers, while Mozer et al. (2000) and Hung et al. (2006) describe simple neural networks as superior to the aforementioned techniques. In addition, innovative machine learning methods such as deep learning algorithms are increasingly included in the discourse. Deep feedforward neural networks are found to outperform flat architectures as well as standard linear techniques by Castanedo et al. (2014) and Keramati et al. (2014). Furthermore, architectural evolutions like recurrent and convolutional neural networks are gaining in importance due to the expansion of customer churn analysis from established industries such as telecommunications, finance and retail to online services and platforms (e.g. Sung et al., 2017). Thus, excellent predictive performance of deep learning is particularly achieved for churn problems characterized by data-related challenges like multidimensionality, time dependencies or generic input complexity (Zhang et al., 2018).

Consequently, the addition of machine learning methods to the range of churn prediction techniques leads to a higher variety and concentration of similarly strong methods and subsequently to more complicated model comparison and selection problems. Therefore, not only from a managerial standpoint, regarding the trust in and further usability of results, but also from a predictive research perspective, a need for additional measures beyond mere performance figures

for the expanded evaluation of churn prediction models is determined. Taking on this issue for classification models in general, Martens et al. (2011) identify three key requirements that are supposed to be fulfilled from a user perspective and apply them to a churn scenario (see also Verbeke et al., 2011): Besides being able to predict correctly, models should be comprehensible and justifiable to achieve greater acceptance in practical settings.

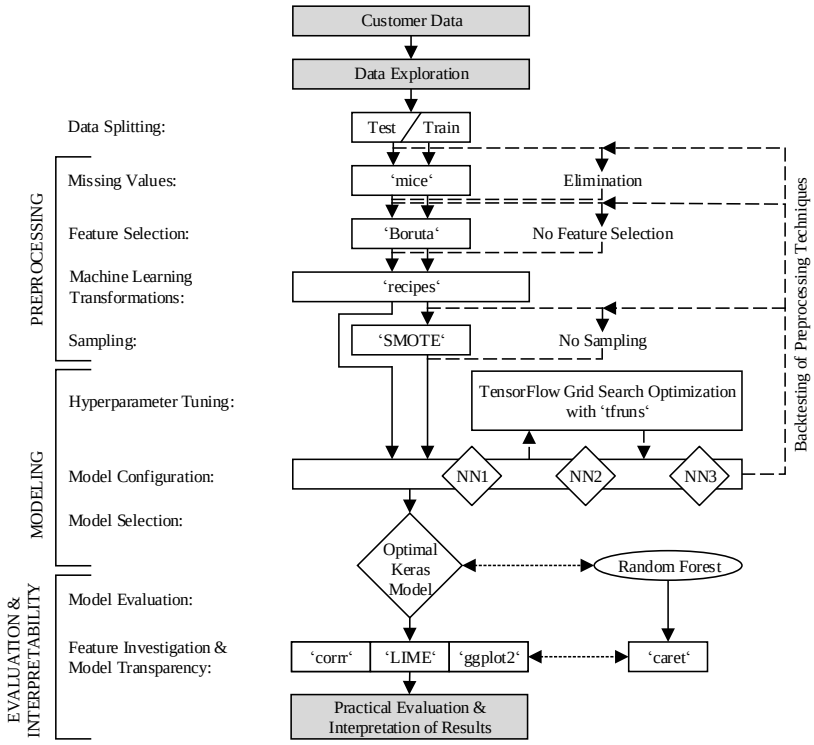

Figure 1: Analytical framework for the implementation of deep learning in churn analysis. R-packages and functions used: mice, Boruta, recipes, SMOTE, corrr, LIME, ggplot2, and caret.

In this context, the comprehensibility of a model considers how well it is understood and approved by the user and is thus synonymous with its interpretability (Freitas, 2014). It is frequently deployed as a subjective measure for the mental fit of a model. At the same time, making model predictions and the logic behind

them understandable to the user also facilitates examining whether the provided results are in line with domain knowledge and hence justifiable. To obtain more comprehensible classification models, three different strategies are proposed by Martens et al. (2011):

1. Building rule-based models or using rule extraction,
2. combining output types, and
3. visualization.

Hence, regarding the interpretability issue of deep learning models as well as their postulated strengths in providing accurate predictions for classification and regression tasks, the churn literature offers promising approaches. Thereby, further research, especially regarding the practical application from a user perspective, is encouraged.

4 Experimental Design

4.1 Data and Analytical Framework

The experimental design of this study is chosen to address the inhibiting factors of the practical application of deep learning models in churn analysis. To pursue the goal of lower complexity in usage and higher overall interpretability and to assess the progress made in that matter, a well-documented data set in customer churn research is selected. Moreover, to ensure clarity through unambiguous classification of customers as churners, a contractual churn setting is examined (Ascarza et al., 2018). The Cell2Cell data set used in this research was originally published by Duke University's Fuqua School of Business as a case study on customer churn of a real US telecommunications provider. The data contains 51,047 individual customer objects, which are described by 56 potential explanatory variables from the fields of demographics, service usage and previous contact. In addition, a target variable indicating the future churn behavior of the customers in the period of 31 to 60 days after the elicitation of the predictor variables is provided in binary form. The corresponding churn rate in the data amounts to 28.8 %. Extensive research using the present data or conformant customized subsets has e.g. been done by Neslin et al. (2006), Lemmens and Croux (2006), Verbeke et al. (2012) and de Caigny et al. (2018).

According to these studies, best prediction results can be achieved with decision tree-based methods. Nevertheless, the seemingly unfavorable data conditions for high neural network performance are accepted in the present study in favor of the comprehensive documentation of data set properties.

The following churn analysis is based on a specifically developed analytical framework aggregating the acquired knowledge on procedural as well as practical application aspects from machine learning and churn literature as shown in Figure 1. In the process, particularly the perceived complexity of model configuration and optimization as well as the lack of interpretability are addressed by proposing a structured course of action through the example of deep feedforward neural networks. Implemented exemplary methods and functions can be adapted according to individual prerequisites and needs along the suggested steps. The starting points for their implementation are derived from the identified relevant properties of interpretable machine learning as well as the proposed strategies to obtain comprehensible classification models. Conforming to the relevance of these areas to churn prediction and considering the previous extent of research, of the suggested means in literature, increasing the model transparency at an individual component level and providing post-hoc interpretations through local explanations and visualizations are selected as the most promising approaches. A detailed explanation of the proposed procedure and applied techniques is given in the subsequent Sections 4.2, 4.3 and 4.4. As a whole, preprocessing of the raw data to ensure the facilitated processability by and trainability of deep learning models is implemented as well as a modeling phase, where churn prediction models are built, trained and optimized for the present classification task, and an evaluation and interpretation phase using test data. As a first step towards an increase in practicability, the schematic integration and guideline itself is supposed to provide a clearer and more standardized application process and thus make the use of deep learning methods for churn prediction less complex and time-consuming for practitioners.

4.2 Preprocessing

First, after performing an 80-20 split of the data to obtain training and test sets and to ensure the individual application of the intended data preprocessing steps, missing values are dealt with by multivariate imputations via chained

equations (MICE) to maintain the full extent of the size-wise limited data set (van Buuren and Groothuis-Oudshoorn, 2011). Another way of ensuring stability of test results in this case, besides a validation split, is k-fold cross validation. However, its implementation along the proposed framework has not yet been developed and tested. For values missing at random, several imputations are generated to account for uncertainty by a series of regression models for continuous and categorical variables, conditional upon their distribution and dependence on other variables (Schafer, 1997). Next, feature selection is performed using the random forest-based wrapper method Boruta (Kursa and Rudnicki, 2010). By reducing the feature space based on variable importance regarding predictive power and considering underlying inter- and multivariable relationships, higher model stability through elimination of collinearity and noise in the data and better generalization ability of deep learning models due to reduced risk of overfitting are strived for (Dash and Liu, 1997; Guyon and Elisseeff, 2003). Moreover, a limited number of input variables makes the model easier to grasp for the users and, therefore, produces higher interpretability (Piramuthu, 2004). Subsequently, variable transformations are carried out for the remaining features to improve learning by a deep neural network. Depending on the individual necessity for each feature, the addressed transformations include appropriate data type conversions, reduction of the number of categories and one hot-encoding of categorical variables as well as logarithmic transformation, discretization and normalization of numerical variables (Kotsiantis et al., 2006; Sola and Sevilla, 1997). The variable specific, sequential preprocessing steps are pre-defined in what might be referred to as a custom preprocessing function, that allows for future single-step application. This not only ensures consistent transferability of variable-level preprocessing between training and test set, but also straightforward reproducibility of data preparation for future data. As a result, data preparation in the form of variable transformations for machine learning is supposed to be less unwieldy and time-consuming and hence less complex in practical application.

The final preprocessing phase is formed by oversampling the training data to deal with class imbalance to the disadvantage of churner as the objects of interest for an effective training of the deep learning algorithm. The synthetic minority oversampling technique (SMOTE) is applied to this end (Chawla et al., 2002). As a variant of regular oversampling, here, instead of exact replications, new objects belonging to the minority class are artificially created by interpolating randomly

selected samples. Finally, the effectiveness of the implemented preprocessing stages regarding positive influence on predictive power of various deep neural networks with initial as well as optimized hyperparameter configurations is verified by a back test.

4.3 Network Modeling and Hyperparameter Tuning

The modeling phase of churn analysis is about building, optimizing and selecting the optimal prediction model. In this study, three different feedforward neural networks with one to three hidden layers are optimized and contrasted with respect to their classification performance to confirm the predictive impact of neural network depth on churn prediction. Since, according to the literature, the stage of prediction modeling plays a crucial role in the perception of complexity and comprehensibility by the users, a more well-arranged procedure for the configuration and optimization of deep learning models in this context is implemented. A human user should be able to contemplate the coherence between the parameters of the model and its prediction in reasonable time (Lipton 2016). This capability is moderated by the size of the model and the computational effort for inferences. But in that matter, for neural networks in particular, it can be argued, that also the structured mapping of hyperparameters and the consequential clarification of their effects on the prediction output can play a crucial role in an increase of transparency at the model component level (simulatability) from a practical user perspective.

Accordingly, the key hyperparameters of the networks are predefined in a lucid register and represented by placeholders in the actual code for model definition. This not only enables convenient access to the core of the configuration of deep learning models and ease of application to the user, but also lays the foundation for the following optimization procedure. Instead of manual, iterative adaptation and evaluation of hyperparameter settings, automatic grid search within specified value limits is applied. For better traceability of the effects of individual configuration changes in the hyperparameter space and, therefore, maximum transparency at the model component level, this method is chosen over random search in the present analysis (Bergstra and Bengio, 2012).

The investigated hyperparameters include standard compositional details to improve prediction performance (Zhang, 2000) as well as regularizers of

layer parameters (Nowlan and Hinton, 1992) and dropout rate (Srivastava et al., 2014) to prevent overfitting. Subsequently, the best neural network is selected regarding the classification performance measured via a validation split by the area under the receiver operating characteristics curve (AUC or AUROC) as most commonly used performance criterion in churn prediction and in logical continuation of the applied hyperparameter optimization objective (Fawcett, 2006). An extensive debate about appropriate evaluation criteria is conducted in the literature (e.g. Burez and van den Poel, 2009). To benchmark the quantitative results and provide a broader context, the deep learning models are additionally compared with a fully optimized random forest classifier.

4.4 Evaluation and Interpretability of Results

On the basis of the final churn predictions on the test set, a conclusive evaluation of the predictive performance of the investigated models is carried out. In addition, the results provided by the best deep neural network are interpreted with particular consideration of the special requirements of customer churn analysis regarding the informative value and interpretability of results for practical application. The proposed solutions in the literature to a higher comprehensibility of classification models built for tasks like churn prediction are brought together with the suggested approaches for the increase of interpretability of black box models by the means of post-hoc interpretations. The concept of post-hoc interpretations describes the extraction of useful information, such as the variable importance of customer features for a model's classification decision in churn prediction, from learned models for practitioners and end users. One emerging approach in that matter is the use of local explanations for the decision-making behavior of neural networks (Lipton, 2016). Comparably, the extraction of symbolic rules from trained models, rather than directly from the data, is proposed to make classification decisions more comprehensible (Martens et al., 2011).

In this connection, the local interpretable model-agnostic explanations (LIME) algorithm proposed by Ribeiro et al. (2016b) is applied in the present study to interpret the deep learning churn classifier by giving faithful characterizations of the underlying mechanisms of the black box model on an individual instance level through local approximations with a separate sparse linear model. At the

same time, this explanation technique provides a way of graphically representing the latent operations leading to the results and thus also complies with another proposition made to achieve greater comprehensibility of the displayed model output for practitioners. The visualization of data in a more interpretable graphical or tabular format also entails higher intelligibility as well as easier validation and plausibility of information for the user and therefore also leads to corresponding justifiability of the model (Martens et al., 2011). To pick up the justifiability issue and to enhance the users' trust in the explanatory value of local findings, correlation analysis and scatter plots are then implemented for specific variables as supplementary graphical visualization elements.

5 Results

5.1 Preprocessing Effectiveness

First, the impact of the implemented preprocessing techniques is assessed. For this purpose, four different preprocessed data sets were created, three of which are adapted forms of the proposed complete procedure, where the stages of missing value treatment, feature selection or over-sampling of training data are left out. Then, the influence of the individual preprocessing phases on prediction performance of various neural networks with different numbers of hidden layers and diverse levels of hyperparameter optimization is evaluated. An overview of the obtained results is provided exemplarily for the network with three hidden layers with and without hyperparameter optimization in Table 1. Overall, for the majority of the investigated deep learning models, the best classification performance regarding the AUC is achieved through the implementation of the full range of preprocessing stages. Only the elimination of missing values instead of their imputation is found to lead to slightly better results in certain cases. This is attributed to the relatively low proportion of missing values in the data and the corresponding small impact of the deletion of customer objects. With regard to the magnitude of influence, the implemented sampling of the training data via SMOTE is detected to be the most influential preprocessing stage in the present case. This adds to existing research on dealing with class imbalance in churn analysis (Burez and van den Poel, 2009; Zhu et al., 2017).

When it comes to the applicability of the process, especially the pre-definition of variable transformation steps is to be highlighted as an advancement towards a less complex and less time-consuming procedure of data preparation for deep learning in the eyes of the users. In the context of churn prediction, greater transparency of the requisite steps for practitioners and higher reproducibility of feature transformations for newly incoming customer data is achieved. Additionally, the crucial challenge of leakage of information between training and test data for legitimate and utilizable results is dealt with by strict data separation and individual application of the preprocessing methods (Zhang, 2007).

Table 1: Influence of preprocessing steps and hyperparameter tuning on model performance for the test data.

AUC	NA Elimination	No Feature Selection	No Sampling	Full Process
NN3 (default)	.617	.617	.615	.622
NN3 (optimized)	.638	.641	.614	.645

5.2 Model Performance

The comparison of the implemented neural networks with regard to their classification performance on the test data shows, that the deep learning model with three hidden layers, and therefore with the greatest depth among the investigated networks, achieves the biggest AUC. An overview of the obtained results is provided by Table 2. In practical terms, the likelihood of assigning a randomly selected churning customer a higher churn probability than a randomly selected non-churner is 64.5 % for the best deep learning classifier. Considering the given data conditions explained above, the classification results of the best deep neural network are very convincing and the model even outperforms previous predictions results achieved by various decision tree-based methods found in literature (see e.g. de Caigny et al., 2018).

Moreover, the value of these results is confirmed as it is the first analysis to obtain good prediction results using deep neural networks on this data set and, at the same time, provide insights on the model selection and optimization process. For the favorable data case at hand, the optimized random forest, included as

a comparison method, even beats the network's prediction performance, but does not show as significantly better results as in similar studies before. The maximum attained AUC of 0.645 means an increase in test data performance of 3.698 % through the hyperparameter optimization process.

To factor in class imbalance and asymmetric misclassification costs in churn analysis, the weighting of classes within the target variable is considered as a cost-sensitive component in the learning algorithm. The setting determining these weights is identified as a salient network argument.

Table 2: Overview of quantitative prediction model performances for the test data.

Model	**NN (1 hidden layer)**	**NN (2 hidden layers)**	**NN (3 hidden layers)**	**Random Forest**
AUC	.636	.644	.645	.678
Precision	.436	.419	.401	.457
Recall	.232	.378	.484	.423

In this connection, for practical utilization purposes, the recall value, influenceable by the selected classification threshold, is highlighted as an important performance measure. The proposition to approach the issue of high complexity and insufficient interpretability by adding structure and tidiness and therefore more transparency to the individual component level of model configuration and optimization is assessed as highly promising. On the one hand, the ease of application is increased by a more user-friendly access to and clear arrangement of network hyperparameters, on the other hand the optimization procedure gains in interpretability. The latter is achieved by making the individual effects of hyperparameter adjustments traceable over the course of implemented grid search, as exploited for the class weight argument. Moreover, the accomplished high transparency and accessibility allow for better control over the tendency to overfit during network training for the practical user.

5.3 Interpretability of Results

The application of the LIME algorithm gives insights in the local feature importance of the neural network's classification decision for single customer objects, first of all, in form of a bar graph as shown in Figure 2. This includes information regarding the predicted class, the determined churn probability, as well as the most influential customer variables and the extent and direction of impact on the prediction made for every customer investigated. Moreover, a condensed visualization for a relatively bigger subset of customer objects can be examined by the means of a heatmap. Results for the present Cell2Cell data show, for the local window of 16 classified customer objects, four variables standing out as most relevant for model predictions. Those are the features "CurrentEqDays", "MonthlyMinutes", "MonthlyRevenue" and "PeakCallsInOut", which are associated with the service usage of the customers as well as with the up-to-dateness of the telephone equipment used as a trigger of customer switching intentions.

These findings are in line with insights from the general literature on determinants of subscriber churn in the service provider and telecommunications industry as well as with previous research conducted on variable importance in the present Cell2Cell data (Ahn et al., 2006; Verbeke et al., 2012). This also applies to the two additional features "Handsets" and "TotalRecurringCharge", that are identified as crucial influencing factors for only certain customer objects via heatmap representation. Their scattered impact indicates, that selected variables only affect the classification results when specific parameter values are reached or certain interrelations with other feature characteristics exist. This exhibits a problem area of the algorithm in general and for the application in churn analysis in particular, that is traced back to the eminently local view on the decision-making behavior of the deep learning model (Ribeiro et al., 2016a).

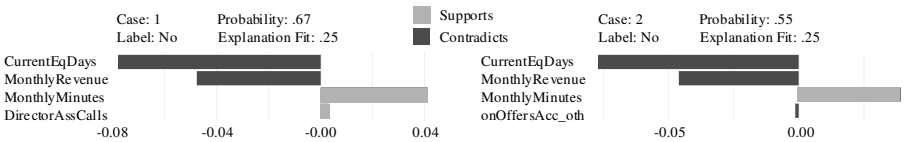

Figure 2: Excerpt of LIME bar diagram for local feature importance.

Thus, for practical implementation in churn analysis, user awareness of the exclusively local mechanisms is needed. To reflect the global value of customer features in the model's prediction as closely as possible, but at the same time limit the review effort in a practical setting, a trade-off between the investigation of a preferably large subset with a wide range of characteristics and the related time requirement of the analysts has to be found. To further illustrate the intuitiveness of the model's decision-making behavior with respect to certain crucial decision criteria and, therefore, to increase its justifiability, proceeding feature investigations and comparison with domain knowledge are proposed. In total, the implementation of algorithms like LIME in customer churn analysis can contribute to eliminate a key problem area of deep learning by increasing the interpretability of results through post-hoc interpretations. These interpretations, in the form of local explanations and corresponding visualizations, not only establish greater confidence in the models' churn prediction procedure, but also provide information on feature importance and churn drivers for subsequent customer retention activities. In the present case of the telecommunications provider, for instance, this could mean providing special offers for replacing equipment to customers who are using old equipment.

6 Conclusion and Practical Implications

Deep learning algorithms represent a powerful method for the modeling of complex problems, with particular strength in capturing non-linearities in intricate data. Beyond high prediction power however, the ease of application and the interpretability of the implemented model are identified as central requirements from a practical user perspective in customer churn analysis. So far, these necessities are considered to fall into the area of shortcomings of deep neural networks as black box models. To overcome this state of affairs, various strategies to enhance the interpretability of models are proposed by machine learning literature as well as classification and churn prediction research. Of the suggested means, increasing the model transparency at an individual component level and providing post-hoc interpretations through local explanations and visualization are selected as the most promising approaches for churn analysis.

This paper introduces an analytical framework for the implementation of deep learning in customer churn prediction, that defines a structured application procedure for the practical user and recommends approaches and methods in the

above-mentioned areas of improvement. Applied exploration of the framework and its components shows, that the perceived complexity of deep learning model configuration and optimization is reduced by deploying pre-defined variable transformation steps during data preprocessing as well as enabling more convenient access to and grid optimization of key network hyperparameters through scheduling and placeholders. Furthermore, higher interpretability is achieved by the means of feature selection, traceability of the individual impact of hyperparameter adaptation and, especially, through enhanced plausibility of churn classifications and the underlying decision-making process of the model by the LIME algorithm. At the same time, the enumerated techniques do not involve any loss in prediction accuracy, so that satisfying model performance is achieved in the empirical test of the present churn problem.

For the practical application of deep learning models in churn analysis, in conclusion, a high future potential can be attested. On the one hand, the digitization of customer relationships and the related increase in data volume and complexity requires the modeling strengths of deep learning algorithms, on the other hand recent and ongoing development of innovative methods successfully deals with reservations regarding their application complexity and interpretability from a user perspective. This trend offers new opportunities for research and practical use of deep learning in customer analysis.

References

Ahn JH, Han SP, Lee YS (2006) Customer Churn Analysis: Churn Determinants and Mediation Effects of Partial Defection in the Korean Mobile Telecommunications Service Industry. Telecommunications Policy 30(10-11):552–568. DOI: 10.1016/j.telpol.2006.09.006.

Ascarza E, Neslin SA, Netzer O, Anderson Z, Fader PS, Gupta S, Hardie BGS, Lemmens A, Libai B, Neal D, Provost F, Schrift R (2018) In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions. Customer Needs and Solutions 5:65–81. DOI: 10.1007/s40547-017-0080-0.

Baehrens D, Schroeter T, Harmeling S, Kawanabe M, Hansen K, Mueller KR (2010) How to Explain Individual Classification Decisions. Journal of Machine Learning Research 11(6):1803–1831. URL: http://www.jmlr.org/papers/v11/baehrens10a.html.

Bengio Y, LeCun Y (2007) Scaling Learning Algorithms toward AI. In: Large-Scale Kernel Machines, chap. 14. MIT Press, Cambridge (USA), Bottou L, Chapelle O, DeCoste D, Weston J (eds.). DOI: 10.7551/mitpress/7496.003.0016.

Bergstra J, Bengio Y (2012) 1. Journal of Machine Learning Research 13(2):281–305. URL: http://jmlr.csail.mit.edu/papers/v13/bergstra12a.html.

Buckinx W, van den Poel D (2005) Customer Base Analysis: Partial Defection of Behaviourally Loyal Clients in a Non-contractual FMCG Rretail Setting. European Journal of Operational Research 164(1):252–268. DOI: 10.1016/j.ejor.2003.12.010.

Burez J, van den Poel D (2007) CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. Expert Systems with Applications 32(2):277–288. DOI: 10.1016/j.eswa.2005.11.037.

Burez J, van den Poel D (2009) Handling Class Imbalance in Customer Churn Prediction. Expert Systems with Applications 36(3):4626–4636. DOI: 10.1016/j.eswa.2008.05.027.

van Buuren S, Groothuis-Oudshoorn K (2011) mice : Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 45(3):1–67. DOI: 10.18637/jss.v045.i03.

de Caigny A, Coussement K, de Bock KW (2018) A New Hybrid Classification Algorithm for Customer Churn Prediction Based on Logistic Regression and Decision Trees. European Journal of Operational Research 269(2):760–772. DOI: 10.1016/j.ejor.2018.02.009.

Castanedo F, Valverde G, Zaratiegui J, Vazquez A (2014) Using Deep Learning to Predict Customer Churn in a Mobile Telecommunication Network. URL: https://www.wiseathena.com/pdf/wa_dl.pdf.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16:321–357. DOI: 10.1613/jair.953.

Dash M, Liu H (1997) Feature Selection for Classification. Intelligent Data Analysis 1(1-4):131–156. DOI: 10.1016/s1088-467x(97)00008-5.

Datta P, Masand B, Mani DR, Li B (2000) Automated Cellular Modeling and Prediction on a Large Scale. Artificial Intelligence Review 14:485–502.

Deng L (2014) Deep Learning: Methods and Applications. Foundations and Trends® in Signal Processing 7(3-4):197–387. DOI: 10.1561/2000000039.

Doshi-Velez F, Kim B (2017) Towards A Rigorous Science of Interpretable Machine Learning. URL: http://arxiv.org/pdf/1702.08608v2. ArXiv:1702.08608v2.

Fawcett T (2006) An Introduction to ROC Analysis. Pattern Recognition Letters 27(8):861–874. DOI: 10.1016/j.patrec.2005.10.010.

Freitas AA (2014) Comprehensible Classification Models. ACM Sigkdd Explorations Newsletter 15(1):1–10. DOI: 10.1145/2594473.2594475.

Ganesh J, Arnold MJ, Reynolds KE (2000) Understanding the Customer Base of Service Providers: An Examination of the Differences Between Switchers and Stayers. Journal of Marketing 64(3):65–87. DOI: 10.1509/jmkg.64.3.65.18028.

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2019) A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys 51(5):1–42. DOI: 10.1145/3236009.

Guyon I, Elisseeff A (2003) An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3(3):1157–1182. URL: `http://www.jmlr.org/papers/v3/guyon03a.html`.

Hadden J, Tiwari A, Roy R, Ruta D (2007) Computer Assisted Customer Churn Management: State-of-the-art and Future Trends. Computers & Operations Research 34(10):2902–2917. DOI: 10.1016/j.cor.2005.11.007.

Hinton G, Salakhutdinov R (2006) Reducing the Dimensionality of Data with Neural Networks. Science 28:504–507. DOI: 10.1126/science.1127647.

Hinton G, Osindero S, Teh YW (2006) A Fast Learning Algorithm for Deep Belief Nets. Neural Computation 18(7):1527–1554. DOI: 10.1162/neco.2006.18.7.1527.

Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. Neural Computation 9(8):1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Huang B, Buckley B, Kechadi MT (2010) Multi-objective Seature Selection by Using NSGA-II for Customer Churn Prediction in Telecommunications. Expert Systems with Applications 37(5):3638–3646. DOI: 10.1016/j.eswa.2009.10.027.

Huang B, Kechadi MT, Buckley B (2012) Customer Churn Prediction in Telecommunications. Expert Systems with Applications 39(1):1414–1425. DOI: 10.1016/j.eswa.2011.08.024.

Hung SY, Yen DC, Wang HY (2006) Applying Data Mining to Telecom Churn Management. Expert Systems with Applications 31(3):515–524. DOI: 10.1016/j.eswa.2005.09.080.

Keaveney SM (1995) Customer Switching Behavior in Service Industries: An Exploratory Study. Journal of Marketing 59(2):71. DOI: 10.2307/1252074.

Keramati A, Ardabili SM (2011) Churn Analysis for an Iranian Mobile Operator. Telecommunications Policy 35(4):344–356. DOI: 10.1016/j.telpol.2011.02.009.

Keramati A, Jafari-Marandi R, Aliannejadi M, Ahmadian I, Mozaffari M, Abbasi U (2014) Improved Churn Prediction in Telecommunication Industry Using Data Mining Techniques. Applied Soft Computing 24:994–1012. DOI: 10.1016/j.asoc.2014.08.041.

Kotsiantis SB, Kanellopoulos D, Pintelas PE (2006) Data Preprocessing for Supervised Leaning. International Journal of Computer Science 1(1):111–117.

Kumar A, Rao VR, Soni H (1995) An Empirical Comparison of Neural Network and Logistic Regression Models. Marketing Letters 6(4):251–263. DOI: 10.1007/BF00996189.

Kumar V, Reinartz W (2016) Creating Enduring Customer Value. Journal of Marketing 80(6):36–68. DOI: 10.1509/jm.15.0414.

Kursa MB, Rudnicki WR (2010) Feature Selection with the Boruta Package. Journal of Statistical Software 36(11):1–13. DOI: 10.18637/jss.v036.i11.

Larivière B, van den Poel D (2005) Predicting Customer Retention and Profitability By Using Random Forests and Regression Forests Techniques. Expert Systems with Applications 29(2):472–484. DOI: 10.1016/j.eswa.2005.04.043.

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation 1(4):541–551. DOI: 10.1162/neco.1989.1.4.541.

LeCun Y, Bengio Y, Hinton G (2015) Deep Learning. Nature 521:436–444. DOI: 10.1038/nature14539.

Lejeune MA (2001) Measuring the Tmpact of Data Mining on Churn Management. Internet Research 11(5):375–387. ISSN: 1066-2243, DOI: 10.1108/10662240110410183.

Lemmens A, Croux C (2006) Bagging and Boosting Classification Trees to Predict Churn. Journal of Marketing Research 43(2):276–286. DOI: 10.1509/jmkr.43.2.276.

Lima E, Mues C, Baesens B (2011) Monitoring and Backtesting Churn Models. Expert Systems with Applications 38(1):975–982. DOI: 10.1016/j.eswa.2010.07.091.

Lipton ZC (2016) The Mythos of Model Interpretability. URL: `http://arxiv.org/pdf/1606.03490v3`. ArXiv:1606.03490v3.

Martens D, Vanthienen J, Verbeke W, Baesens B (2011) Performance of Classification Models From a User Perspective. Decision Support Systems 51(4):782–793. DOI: 10.1016/j.dss.2011.01.013.

Mozer MC, Wolniewicz R, Grimes DB, Johnson E, Kaushansky H (2000) Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry. IEEE Transactions on Neural Networks 11(3):690–696. DOI: 10.1109/72.846740.

Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E (2015) Deep Learning Applications and Challenges in Big Data Analytics. Journal of Big Data 2(1):1–21. DOI: 10.1186/s40537-014-0007-7.

Neslin SA, Gupta S, Kamakura W, Lu J, Mason CH (2006) Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. Journal of Marketing Research 43(2):204–211. DOI: 10.1509/jmkr.43.2.204.

Nowlan SJ, Hinton GE (1992) Simplifying Neural Networks by Soft Weight-Sharing. Neural Computation 4(4):473–493. DOI: 10.1162/neco.1992.4.4.473.

Paliwal M, Kumar UA (2009) Neural Networks and Statistical Techniques: A Review of Applications. Expert Systems with Applications 36(1):2–17. DOI: 10.1016/j.eswa.2007.10.005.

Piramuthu S (2004) Evaluating Feature Selection Methods for Learning in Data Mining Applications. European Journal of Operational Research 156(2):483–494. DOI: 10.1016/s0377-2217(02)00911-6.

Ribeiro MT, Singh S, Guestrin C (2016a) Model-Agnostic Interpretability of Machine Learning. 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016). URL: http://arxiv.org/pdf/1606.05386v1. ArXiv:1606.05386v1.

Ribeiro MT, Singh S, Guestrin C (2016b) Why Should I Trust You?: Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16) 8:1135–1144. DOI: 10.1145/2939672.2939778.

Rosenberg LJ, Czepiel JA (1984) A Marketing Approach for Customer Retention. Journal of Consumer Marketing 1(2):45–51. DOI: 10.1108/eb008094.

Rumelhart DE, Hinton G, Williams RJ (1986) Learning Representations by Back-propagating Errors. Nature 323:533–536. DOI: doi.org/10.1038/323533a0.

Schafer J (1997) Analysis of Incomplete Multivariate Data, 1st edn. Chapman & Hall. DOI: 10.1201/9781439821862.

Schmidhuber J (2015) Deep Learning in Neural Networks: An Overview. Neural Networks 61:85–117. DOI: 10.1016/j.neunet.2014.09.003.

Sola J, Sevilla J (1997) Importance of Input Data Normalization for the Application of Neural Networks to Complex Industrial Problems. IEEE Transactions on Nuclear Science 44(3):1464–1468. DOI: 10.1109/23.589532.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research 15(1):1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

Sung C, Higgins CY, Zhang B, Choe Y (2017) Evaluating Deep Learning in Chum Prediction for Everything-as-a-service in the Cloud. In: Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Institute of Electrical and Electronics Engineers, Piscataway (USA). DOI: 10.1109/ijcnn.2017.7966317.

Verbeke W, Martens D, Mues C, Baesens B (2011) Building Comprehensible Customer Churn Prediction Models with Advanced Rule Induction Techniques. Expert Systems with Applications 38(3):2354–2364. DOI: 10.1016/j.eswa.2010.08.023.

Verbeke W, Dejaeger K, Martens D, Hur J, Baesens B (2012) New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach. European Journal of Operational Research 218(1):211–229, Slowinski R, Billaut JC, Bomze IM, Dyson R, Peccati L, et al. (eds.). DOI: 10.1016/j.ejor.2011.09.031.

Wedel M, Kannan PK (2016) Marketing Analytics for Data-Rich Environments. Journal of Marketing 80(6):97–121. DOI: 10.1509/jm.15.0413.

Weiss GM (2004) Mining With Rarity: A Unifying Framework. ACM Sigkdd Explorations Newsletter 6(1):7. DOI: 10.1145/1007730.1007734.

Zhang GP (2000) Neural Networks for Classification: A Survey. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews) 30(4):451–462. DOI: 10.1109/5326.897072.

Zhang GP (2007) Avoiding Pitfalls in Neural Network Research. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews) 37(1):3–16. DOI: 10.1109/TSMCC.2006.876059.

Zhang Q, Yang LT, Chen Z, Li P (2018) A Survey on Deep Learning for Big Data. Information Fusion 42:146–157. DOI: 10.1016/j.inffus.2017.10.006.

Zhu B, Baesens B, vanden Broucke SK (2017) An Empirical Comparison of Techniques for the Class Imbalance Problem in Churn Prediction. Information Sciences 408:84–99. DOI: 10.1016/j.ins.2017.04.015.