Proceedings of the 3rd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/154/2024.19543

Improving Predictive Accuracy of Customer Churn Models Through Data Cleaning in the Telecommunications Industry

Jiaheng Zhang[1,a,*]

[1] *Wenzhou-Kean University (WKU), Wenzhou, Zhejiang Province, 325060, China*
a. zhanjiah@kean.edu
**corresponding author*

Abstract: Customer churn poses a significant challenge in the telecommunications industry, leading to substantial financial losses. Retaining customers is widely recognized as more cost-effective than acquiring new ones. However, existing churn prediction models often lack accuracy due to poor data quality, including issues like missing values and inconsistencies. This study explores the impact of data-cleaning techniques on improving the predictive accuracy of churn models, specifically focusing on telecommunications. Using the IBM Telco Customer Churn dataset, it applied methods such as imputation, one-hot encoding, and outlier removal to preprocess the data, followed by testing logistic regression, decision trees, and gradient boosting models. Our analysis revealed that data cleaning notably enhanced model performance, with accuracy improvements ranging from 78.4% to 89.4%, alongside increases in precision, recall, and AUC scores. These results underscore the importance of high-quality data in churn prediction, suggesting that telecommunications companies can benefit from implementing robust data-cleaning processes. The findings contribute to the development of reliable churn prediction models that support effective customer retention strategies.

Keywords: customer churn, data cleaning, churn prediction models, telecommunications, machine learning.

1. Introduction

Customer churn remains a significant issue within the telecommunications industry, with substantial impacts on profitability and growth. Retaining existing customers is often more cost-effective than acquiring new ones. Recent studies suggest that the application of advanced machine learning techniques can greatly enhance the accuracy of churn predictions, thereby supporting more effective customer retention strategies [1-3].

Regardless of the claim regarding the usefulness of computers in forecasting the future, the reality is that two-thirds of businesses fail due to model risk. For almost all the predictive models, there are mostly issues related to the data that affect adversely the model's efficacy. Acknowledging that there are missing values, inconsistencies, and noise in the input data, it is not unreasonable to expect that future users of the model will fail to detect all at-risk customers in the organization which as a result compromises the model's usefulness in a business setting. In this regard, appropriate data scrubbing is another of the main tasks that one needs to carry out to increase the productivity and efficiency of the machine learning models. The reason is that appropriate data cleansing will remove all unwanted interferences and bugs from the raw data, thus making it ready for analysis and making the model more efficient.

There has been a lot of progress in the use of machine learning applications, however, there are still existing gaps in the knowledge surrounding the most effective data-cleaning strategies that would improve the predictive accuracy of churn prediction in the telecommunications industry. However, the improvement of data quality measurement application is claimed to be 20% in some situations. This brings us to a rather unfortunate conclusion, and that is, the identification and execution of the most effective data cleaning techniques have been ignored, including the telecoms sector that has utilized churn datasets for empirical investigation. Such an absence of evidence further indicates the necessity of investigating questions about the efficiency of different data cleaning approaches, not least to substantiate theoretical claims but also to enable the telecommunications industry in enhancing the forecasting strength of churn models.

2. Theoretical Framework

2.1. Data Cleaning Conceptualization

Effective data cleaning is crucial for the accuracy of machine learning models, addressing issues like missing values and inconsistencies which can degrade model performance [4].

2.2. Machine Learning for Churn Prediction of Customers

Techniques such as decision trees, gradient boosting, and ensemble methods have been effectively used for churn prediction, benefiting from their ability to reduce errors and handle complex datasets [5, 6].

Nonetheless, the deployment of regression logs occasionally proves ineffective in capturing the nonlinear correlation between variables, whereas datasets that are too small may lead to overfitting when decision trees are employed. Following these are random forests and gradient boosting, which reduce errors by combining predictions from multiple models, leveraging their strength as ensemble methods. Random forests construct several decision trees on randomly selected subsets of the data and use the average prediction from these trees to increase accuracy and prevent overfitting. This method efficiently manages large and complex datasets such as those with high-dimensional data and important features, although it can be computationally intensive and less interpretable due to its complexity. Gradient boosting builds models sequentially, with each new model addressing vulnerabilities in the previous ones, forming a robust model composed of many layers. However, this requires meticulous tuning to prevent overfitting and is highly sensitive to data quality, emphasizing the necessity of thorough data preprocessing to maximize effectiveness.

2.3. The Quality of a Dataset as a Determinant of Predictive Performance Capability

The quality of a dataset is a critical determinant of the predictive performance of machine learning models. The success of these models often correlates with the quality and quantity of the data used, with numerous studies demonstrating that data quality significantly impacts predictive outcomes. Effective data cleaning practices, such as one-hot encoding of categorical variables and z-score normalization of numerical data, are crucial in ensuring algorithms are trained correctly. These practices maintain the diversity of the data, preparing machine learning models to be trained on specific datasets and exposed to new data, thereby mitigating significant risks of underfitting or overfitting, particularly in the telecommunications sector's churn predictions.

3. Research Design and Methodology

3.1. Dataset Description

The IBM Telco Customer Churn dataset stands as a cornerstone in telecommunications research, specifically in the predictive analytics of customer retention. This dataset is a primary tool in the Telecommunication Companies Churn Prediction Model Competition, showcasing a rich array of variables critical for analysis. The dataset encapsulates a range of data points including tenure, contract specifics, and billing intervals, which have been identified through prior studies as pivotal in forecasting customer churn. The comprehensive nature of this dataset allows for nuanced analysis, providing insights into customer behavior and retention patterns that are vital for developing targeted strategies in telecommunications services.

3.2. Data Preparation Procedures Implemented

In the realm of churn prediction for telecommunication companies, the data preparation process is critical for developing accurate and reliable predictive models. This process typically involves three key steps: Missing Data Imputation, Data Transformation, and Outlier Detection and Removal.

Missing Data Imputation

The completeness and accuracy of the data are foundational for robust predictive modeling. For the 'Total charges' field in our dataset, which often lacks data primarily due to new customers who haven't completed a billing cycle, missing values are imputed with zeros. This approach aligns with industry standards and helps maintain the integrity of the dataset by ensuring consistent data across all entries, thereby minimizing any potential biases in the predictive analysis.

Data Transformation

Transforming data into a format suitable for machine learning is crucial. Techniques such as one-hot encoding are used to convert categorical variables like 'Contract' and 'Internet Service' into a binary format. This method eliminates any ordinal assumptions that could potentially skew the model's interpretation. For high-cardinality variables such as 'Payment Method,' target encoding is applied to reduce dimensionality while preserving important information, thereby enhancing the model's ability to generalize from training data to unseen data effectively.

Outlier Detection and Removal

Financial metrics such as 'Monthly Charges' and 'Total Charges' are particularly scrutinized for outliers, which can significantly distort predictive accuracy. Using the z-score method, outliers—data points that lie beyond three standard deviations from the mean—are identified and removed. This step ensures that the predictions are realistic and grounded in typical customer behavior, avoiding the influence of anomalous data.

These preprocessing steps are essential for preparing the dataset for subsequent analysis using sophisticated machine learning models such as logistic regression, decision trees, and gradient boosting. These models are enhanced with ensemble techniques to improve their reliability and accuracy in predicting customer churn. The comprehensive data preparation process not only supports the robustness of the analytical methods but also ensures that the predictive models developed can deliver insights that are both statistically valid and highly relevant to the telecommunications industry.

The rigorous approach to data preparation and model selection, including the handling of missing data, transformation of categorical data, and removal of outliers, forms the backbone of effective churn prediction. This methodology is crucial for enabling telecommunications companies to implement more effective customer retention strategies based on data-driven insights. The integration of these advanced techniques ensures that the models are not only accurate but also applicable to real-world scenarios, facilitating strategic decisions that enhance customer satisfaction and reduce churn.

For further details on these techniques and their applications in predictive modeling, references such as Ahmad et al. and publications by the IEEE provide extensive insights and case studies, highlighting the importance of these methods in the telecommunications sector [7].

3.3. Machine Learning Models

The analytical framework incorporated three predominant machine learning models, each selected for their relevance to the binary classification task inherent in churn prediction. The logistic regression model offers a baseline with its simplicity and interpretability, while decision trees provide deeper insights into the significance and interactions of various predictors. Gradient boosting, known for its precision and ability to handle diverse datasets, complements this array by focusing on improving predictions iteratively. Each model is calibrated to optimize data quality, a critical factor in enhancing overall performance and ensuring robustness in predictions across various customer scenarios.

3.4. Evaluation Models Metrics

To accurately assess model performance, a suite of metrics was employed, each chosen for its ability to reflect different aspects of model effectiveness in the context of churn prediction:

Accuracy: This metric provides a straightforward measure of the model's overall effectiveness by quantifying the proportion of correct predictions made across all categories.

Precision and Recall: These metrics offer a nuanced view of model performance, with precision measuring the accuracy of positive predictions (i.e., correctly predicted churn cases) and recall assessing the model's ability to identify all potential churn cases. High precision reduces false positives, while high recall ensures minimal missed churn opportunities.

Area Under Curve (AUC): AUC evaluates the model’s discriminative ability, i.e., its capacity to distinguish between churners and non-churners at various threshold levels. This metric is particularly valuable in operational settings, where it guides the setting of decision thresholds that align with business objectives.

Collectively, these metrics facilitate a comprehensive evaluation of the predictive models, guiding strategic decisions in customer relationship management and helping telecom companies implement effective retention strategies informed by robust data analysis [8].

4. Results and Discussion

4.1. Pre- and Post-Cleaning Performance Assessment of the Model

The importance of thorough data cleaning before the application of predictive models in churn analysis has been strongly evidenced by the results of this study. Our examination clearly shows that without the cleaning process, the models exhibited significantly lower accuracy. For instance, logistic regression, which is fundamentally sensitive to data quality, demonstrated an accuracy increase from baseline figures to between 78.4% and 84.6% post-cleaning. Similarly, controlled boosting, a method that relies heavily on the integrity of data, showcased a remarkable accuracy of 89.4%. These improvements were not limited to accuracy alone; enhancements were also observed in precision, recall, and the AUC score. These metrics collectively indicate not only a higher rate of correctly predicted churn events but also an improvement in the model's ability to distinguish between churn and no-churn instances effectively.

4.2. Application for Other Specific Data Cleaning Techniques

The application of specific data cleaning techniques such as one-hot encoding, target encoding, and outlier removal has proven pivotal in enhancing the stability and accuracy of the predictive models used in this study. One-hot encoding and target encoding have facilitated better management of categorical variables, which are plentiful in telecommunications datasets. These techniques help to convert categorical data into a numerical format that is more interpretable by algorithms, thus enhancing the logistic regression and decision tree models' performance significantly. Moreover, the strategic removal of outliers has been instrumental in increasing the generalizability of the models. By eliminating data points that are significantly distorted or extreme, the models can perform more consistently across various datasets, thereby reducing the likelihood of skewed predictions that could misinform strategic decisions.

4.3. Business Implications for the Telecom Industry

The findings of this study hold substantial implications for the telecommunications industry, especially in a competitive global market where customer retention is paramount. Enhanced predictive models, refined through rigorous data cleaning and sophisticated analytics, provide telecom providers with a sharper toolset to implement effective customer retention strategies. These strategies are crucial as they enable telecom companies to target potential churners with precision, offering tailored interventions that are likely to be more effective and cost-efficient. The ability to accurately predict and mitigate churn not only improves customer satisfaction and loyalty but also conserves resources that would otherwise be expended in broad, undirected efforts to retain the customer base. Thus, the application of advanced predictive analytics in churn management allows telecom companies to maintain a competitive edge by operating more strategically in customer engagement and retention.

5. Conclusion

In conclusion, the findings from this study demonstrate the transformative impact of advanced data cleaning and sophisticated machine learning techniques on the predictive accuracy of churn models within the telecommunications industry. The implementation of ensemble methods, as evidenced by the significant enhancements in model performance, supports the assertion that integrating diverse algorithms can lead to superior predictive outcomes. This is corroborated by recent scholarly discussions that emphasize the efficacy of these approaches.

Looking ahead, the field stands on the precipice of potentially groundbreaking advancements in predictive modeling. The automation of data cleaning processes presents a promising avenue for research, promising to elevate the efficiency and effectiveness of data preparation phases. Furthermore, the exploration and integration of cutting-edge machine learning techniques such as deep learning promise not only to enhance the granularity and accuracy of predictions but also to extend these models' applicability across varied datasets and contexts.

It is recommended that stakeholders within the telecommunications sector continue to invest in these technological advances. By doing so, they will not only maintain but enhance their competitive edge in a market where customer retention is increasingly pivotal. Continuous improvement in these areas, supported by ongoing research and development, will ensure that predictive models keep pace with the evolving dynamics of customer behavior and market conditions.

In essence, as it advances the methodologies and incorporates more sophisticated technologies, the horizon for what can be achieved with churn prediction models expands, bringing into view the potential for more robust, insightful, and actionable analytics that can drive strategic business decisions.

References

[1] Benson, V. (2024). Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models. Algorithms, 17(6), 231.

[2] Chang, V., Hall, K., Xu, Q. A., Amao, F. O., Ganatra, M. A., & Benson, V. (2024). Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models. Algorithms, 17(6), 231.

[3] Sikri, A. (2023). Enhancing customer retention in telecom industry with machine learning driven churn prediction. Electronic Commerce Research and Applications, 44, 101087.

[4] IEEE. (2024). Customer Churn Prediction Using Machine Learning Methods: A Comparative Analysis. IEEE Xplore.

[5] Karamollaoğlu, H., Yücedağ, İ., & Doğru, İ. A. (2021). Customer churn prediction using machine learning methods: A comparative analysis. In 2021 6th International Conference on Computer Science and Engineering (UBMK) (pp. 139-144). IEEE.

[6] Journal of Artificial Intelligence Research. (2023). Exploring Data Imputation Techniques for Better Predictive Models in Telecommunications. PubMed.

[7] Bruckner EP, Curk T, Đorđević L, Wang Z, Yang Y, Qiu R, Dannenhoffer AJ, Sai H, Kupferberg J, Palmer LC, Luijten E, Stupp SI. (2022). Hybrid Nanocrystals of Small Molecules and Chemically Disordered Polymers. ACS Nano. 16(6):8993-9003.

[8] Ijomah, T. I., Idemudia, C., Eyo-Udo, N. L., & Anjorin, K. F. (2024). The role of big data analytics in customer relationship management: Strategies for improving customer engagement and retention. World Journal of Advanced Science and Technology.