IE 7275 DATA MINING IN ENGINEERING

Case Study: An In-Depth Evaluation of Data Mining Methods in Customer Churn Prediction

Name: Pramoth Guhan

Email: guhan.p@northeastern.edu

Contact Number: +1 (857) 891-6677

Submission Date: 02/21/2024

Case Study: An In-Depth Evaluation of Data Mining Methods in Customer Churn Prediction

Introduction:

Customer churn, also known as customer attrition, is a critical challenge faced by businesses across various industries. It refers to the phenomenon where customers discontinue their relationship with a company or switch to a competitor's product or service. Customer churn can have significant negative impacts on a company's revenue, profitability, and long-term sustainability. Therefore, accurately predicting and preventing customer churn is a top priority for businesses seeking to maintain customer loyalty and maximize customer lifetime value.

Data mining techniques, particularly supervised learning algorithms, have emerged as powerful tools for predicting customer churn and identifying factors influencing customer retention. By analysing historical customer data, including demographics, transaction history, usage patterns, and interactions with the company's products or services, businesses can uncover valuable insights into customer behaviour and preferences. These insights enable companies to proactively identify customers at risk of churn and implement targeted retention strategies to mitigate the loss of valuable customers.

Problem Statement:

Customer churn, defined as the discontinuation of service and migration to competitors, poses a significant financial challenge for telecom companies. Acquiring new customers is inherently more expensive than retaining existing ones, making accurate churn prediction crucial for sustainable business growth. The competitive landscape of the telecom industry intensifies the impact of churn.

Losing customers directly affects revenue and can create a domino effect, with dissatisfied customers influencing others to switch providers. Traditionally, statistical models were used for churn prediction, but their limitations have driven the adoption of more advanced machine learning techniques.

Literature Review:

This research delves into five key papers that explore various aspects of customer churn prediction through machine learning:

1. "Customer Churn Prediction" by Senthilnayaki B et al. examines the feasibility of employing machine learning algorithms to identify customers at risk of churn. [1]
2. "Customer Churn Analysis Using Machine Learning" by Ritika Tyagi and K. Sindhu addresses the need for improved churn models and tackles the difficulty of accurately predicting churn rates. [2]
3. "Customer churn prediction in telecom using machine learning in big data platform" by Abdelrahim Kasem Ahmad et al. leverages a big data platform to develop a churn prediction model, highlighting the increasing importance of handling large data volumes in this domain. [3]
4. "Improving Customer Retention with Churn Prediction Models" by Vishwajyothi Reshmi and Krutika Kulkarni focuses on the economic importance of churn prediction and the evolution of methods from simple models to complex algorithms. [4]
5. "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees" by Arno De Caignya et al. proposes a novel hybrid approach combining the strengths of decision

trees and logistic regression to overcome their limitations and enhance predictive performance, particularly relevant in saturated markets. [5]

Data Considerations for Customer Churn Prediction Papers:

Customer churn prediction, the ability to identify customers at risk of discontinuing service, hinges on robust data considerations.

[1] starts with data collection and preprocessing to clean the raw data. The cleaned data is then subjected to feature selection to identify significant features. It utilized customer-related features like usage patterns, billing information, and demographics, emphasizing feature selection for model accuracy. Exploratory data analysis (EDA) and outlier treatment suggest data visualization played a role in understanding the data and ensuring its quality. The balanced class is divided into a 70% training and 30% testing dataset, upon which PCA is applied for dimensionality reduction and feature selection, choosing 60 components for model building based on the screen plot.

[2] analyses a Kaggle dataset with 21 features including gender, tenure, and telecom services. Extensive data analysis involved checking and addressing missing values by converting data types and removing rows with missing values, feature encoding, and correlation analysis to identify churn-related features. segments the customer history dataset, derives attributes and usage patterns, splits the data into training and testing sets (with 20% allocated for testing), builds models, and tests them for churn prediction accuracy. The importance of feature analysis and transformation is highlighted. A bar graph showing the correlation of features in a dataset, useful for identifying key factors influencing customer churn. Another chart illustrates that as monthly charges increase, the density of customers who churn also increases, indicating a correlation between higher fees and customer attrition.

[3] incorporates a vast array of data, including customer information, network logs, call records, and mobile IMEI data. This data, amounting to 70 TB and covering 10 million customers over nine months, is processed using a big data platform. The inclusion of Social Network Analysis (SNA) features significantly enhances the performance of the model. The data preparation phase addresses the challenges of large volume, variety, and imbalance in the dataset through techniques such as oversampling, undersampling, and no rebalancing. The research handles data collected from various systems and databases, each producing different file types: structured, semi-structured (XML-JSON), and unstructured (CSV-Text). The big data platform processes these diverse data types without requiring any modifications or transformations, thereby eliminating issues related to the size and format of the data and enabling seamless data handling and analysis. The secondary class, representing churn customers in the SyriaTel dataset, constitutes only about 5% of the total data. This imbalance can lead to poor results from machine learning algorithms that don't account for class balance. However, the research effectively addresses this issue, ensuring the reliability and accuracy of the churn prediction model.

[4] The dataset used in the study is a large customer dataset from 2015, with 20,469 records and 26 attributes, including customer demographics, usage details, and 'churn' as the key attribute. The data, collected over two months, provides a comprehensive view of customer behaviour and is instrumental for churn analysis. Emphasizes algorithm suitability based on factors like dataset size, interpretability, and computational efficiency. In Azure, the training and testing procedure is streamlined by Analysis Services, which divides the dataset into training and testing data using a default 70:30 ratio. After the model is created, it's processed with the training dataset and tested against the test data. It discusses decision trees, support vector machines, and neural networks, highlighting performance metrics like accuracy, precision, recall, and AUC.

[5] utilizes datasets from 14 industries for customer churn prediction, employing features not detailed in the provided excerpt. A 5x3 cross-validation design was used, dividing the data into training, selection, and validation sets, each containing one-third of the data. Data preprocessing involves imputing and flagging attributes with over 5% missing values, removing instances with under 5% missing values, adjusting outliers to within three standard deviations from the mean, and balancing classes in churn prediction through random under sampling. Feature selection through Fisher selection indicates a method for optimal model performance. While the LLM's performance is compared against traditional models, data visualization techniques are not specified.

Comparative Insights: Comparing these papers showcases diverse data considerations for customer churn prediction. [3] stands out with its big data approach, potentially offering richer insights but requiring significant resources like processing power. Others utilize smaller, readily available datasets like Kaggle sources. Feature engineering strategies also vary, with [3] incorporating unique Social Network Analysis (SNA) features. While visualization plays a role in some studies, like using AUC and ROC curves in [3], its extent and specific techniques (e.g., heatmaps, time series plots) differ. Overlooked features like social media activity and external economic factors highlight areas for further exploration across all papers.

Methodology Comparison in Customer Churn Prediction:

Examining the five papers, we delve into their diverse approaches to model selection and comparison:

[1]: This study utilizes Stratified k-fold cross-validation to evaluate the performance of three established machine learning algorithms: Random Forest, KNN, and Logistic Regression. The cross-validation graph, a visual

representation of the results, reveals that Random Forest outperforms the others with a score of 96.3%, indicating its superior predictive capability, particularly for non-linear data like churn. KNN shows reasonable performance but falls short compared to Random Forest. Logistic Regression, on the other hand, demonstrates the least effectiveness among the three. While the graph guides the selection of the most effective model, it's crucial to consider that model performance can vary based on the specific characteristics of the dataset and the problem at hand. Thus, the selection process should be comprehensive, considering various factors beyond cross-validation scores.

[2]: This study broadens the algorithm comparison, including Logistic Regression, ADA Boost Classifier, XGBoost Classifier, Decision Tree Classifier, Random Forest, and SVC. Models are progressively developed using cross-validation and hyperparameter tuning. An ensemble model, comprising Linear Regression, Ada Boost Classifier, XG Boost Classifier, and SVC, is constructed and its performance is assessed through hard voting. The Customer Churn Rate is calculated by dividing the total number of customers lost during a specific period by the total number of customers at the beginning of that period.

[3]: The study focuses on tree-based methods and experiments with several algorithms including Decision Tree, Random Forest, Gradient Boosted Machine Tree (GBM), and Extreme Gradient Boosting (XGBOOST). The Decision Tree algorithm is a simple yet powerful machine learning technique that uses a tree-like model of decisions and is easy to understand and interpret, making it a popular choice for many predictive modelling problems. Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the class that is the mode of the classes of the individual trees. It's known for its robustness and ability to prevent overfitting. GBM is another ensemble machine learning algorithm that constructs new predictors that aim to correct the residuals errors of the prior predictor, reducing

the errors gradually. XGBOOST, on the other hand, is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It produces an ensemble of weak prediction models, typically decision trees. In this study, XGBOOST outperformed the other algorithms, achieving the highest AUC (Area Under the Curve) value of 93.3%. This result underscores the effectiveness of XGBOOST for this specific problem of customer churn prediction in the telecom sector. The high AUC value indicates that the XGBOOST model has a high measure of separability, meaning it's capable of distinguishing between customers who will churn and those who won't with high accuracy.

[4]: This review emphasizes algorithm suitability based on factors like dataset size, interpretability, and computational efficiency. It discusses decision trees, support vector machines, and neural networks, highlighting performance metrics like accuracy, precision, recall, and AUC. The chart, "Four Step Implementations," outlines a strategy for customer-related business processes. It includes steps for product development, data-driven customer journey analysis, churn driver analysis, and churn prediction, all aimed at enhancing customer retention and satisfaction.

[5]: This study introduces a novel hybrid approach, the Logit Leaf Model (LLM), which combines decision rules and logistic regression to capture both linear and interaction effects. The LLM outperforms its building blocks, logistic regression and decision trees, and matches the performance of advanced ensemble methods like Random Forests and Logistic Model Trees. In terms of comprehensibility, the LLM shows key benefits over decision trees and logistic regression. Performance is measured using the area under the receiver operating characteristics curve (AUC) and top decile lift (TDL). Despite their strong predictive performance, decision trees struggle with linear relations, and logistic regression has difficulties with interaction effects. Advanced ensemble methods

offer strong predictive performance, but the LLM matches their performance while providing better comprehensibility.

Comparative Insights: The papers showcase a fascinating array of methodological approaches. While some studies opt for established algorithms like Logistic Regression [1] [2], others explore more complex ensemble methods like Random Forest and XGBOOST [1] [3] [4]. The hybrid LLM approach in [5] offers a unique perspective, aiming to bridge the gap between interpretability and performance. Notably, the choice of algorithm significantly impacts results, as evidenced by XGBOOST's superior performance in [3]. Understanding these diverse approaches, their underlying assumptions, and potential trade-offs empowers telecom companies to select the most effective methodology for their specific churn prediction needs.

Performance and Evaluation:

Beyond methodology, effectively gauging the performance of churn prediction models is crucial for making informed decisions.

[1]: This study aims to accurately identify customers who are likely to leave a service, known as churners. They use a method called Random Forest, which gives them a high accuracy score of 96.3%. Comparatively, other methods like KNN and Logistic Regression score lower, at 88.8% and 81.72% respectively. This means Random Forest is really good at spotting customers who might churn.

[2]: This study looks at different ways to measure how well their model predicts customer churn. They consider more than just accuracy, also looking at things like F1 score, precision, and recall. For example, they found that XGBoostClassifier had the highest accuracy at 82.51%, while Decision Tree had the lowest at 74.69%. They also tweak their model to make it even better.

[3]: This study focuses on dealing with datasets where there are a lot more non-churners than churners. They use a method called AUC, which helps them see how well their model works in this situation. XGBOOST came out on top, with a score of 93.3%, showing it's really good at spotting churners. On the other hand, Random Forest and Decision Trees didn't perform as well.

[4]: This review talks about how to measure how well a churn prediction model works. They say it's not just about accuracy, but also other things like precision and recall. They also mention special metrics for when there are way more non-churners than churners. They stress the importance of picking the right measurement method to get the most accurate results. There is a confusion matrix and a multi-class ROC graph, which are used to evaluate the performance of a machine learning classification algorithm in predicting customer churn. These visualizations highlight the balance between various performance metrics.

[5]: This study tries out a mix of different methods to see which one is best at spotting customers who might leave. They use measures like AUC and Top Decile Lift to see how well their model works. Their hybrid method, called LLM, did really well, almost as good as Random Forest. They say using both AUC and Top Decile Lift gives a better picture of how good the model is.

Comparative Insights: The papers showcase a range of evaluation approaches. While some prioritize churner identification accuracy [1], others acknowledge the need for broader metrics like F1 score and confusion matrices [2] [4]. The use of AUC for imbalanced datasets is highlighted in papers [3] [4] [5] further incorporating TDL for practical relevance. This comparative analysis underscores the importance of selecting appropriate metrics based on dataset characteristics and business goals. By understanding the trade-offs of different metrics, telecom companies can gain deeper insights into model performance and make informed decisions for effective churn prediction and customer retention strategies.

Conclusion:

In conclusion, the challenge of customer churn, particularly in the telecom industry, remains a significant concern for businesses aiming to maintain sustainable growth and profitability. Through the literature review and comparative analysis of five key papers on customer churn prediction, several important insights have emerged.

Firstly, the adoption of machine learning techniques, especially in conjunction with big data platforms, has revolutionized the field of churn prediction. Papers such as Ahmad et al. (2019) highlight the importance of leveraging large volumes of diverse data sources to develop robust churn prediction models, thereby enhancing the accuracy and effectiveness of customer retention strategies.

Secondly, the methodological diversity observed in the reviewed papers underscores the importance of selecting appropriate algorithms and evaluation metrics based on the specific characteristics of the dataset and business objectives. While some studies favour ensemble methods like Random Forest and XGBOOST for their superior performance, others explore hybrid approaches combining decision rules and logistic regression to balance predictive accuracy and model interpretability.

Furthermore, the evaluation of churn prediction models extends beyond traditional accuracy metrics, with an increasing emphasis on broader performance indicators such as precision, recall, and area under the receiver operating characteristic curve (AUC). This holistic approach to model evaluation, as demonstrated in papers like Reshmi and Kulkarni (2021), enables businesses to gain a comprehensive understanding of model performance and make informed decisions regarding customer retention strategies.

Overall, the findings from this literature review emphasize the critical role of data-driven insights and advanced analytical techniques in mitigating customer churn. By leveraging machine learning algorithms, businesses can proactively identify customers at risk of churn and implement targeted retention initiatives, thereby maximizing customer lifetime value and fostering long-term competitiveness in the telecom industry.

Citations:

[1] Senthilnayaki, B., Swetha, M., & Nivedha, D. (2023). CUSTOMER CHURN PREDICTION. Unpublished manuscript.

[2] Tyagi, R., & Sindhu, K. (2022). Customer churn analysis using machine learning. International Journal of Innovative Technology and Exploring Engineering, 11(8), 557-563.

[3] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 2(1), 1-18.

[4] Reshmi, V., & Kulkarni, K. (2021). Improving customer retention with churn prediction models. International Journal of Research in Engineering and Science (IJRES), 10(8), 59-63.

[5] De Caigny, A., Coussement, K., & De Bock, K. W. (2015). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. Expert Systems with Applications, 42(1), 337-347.