Churn Analytics And Prediction: Unveiling Customer Retention Dynamics in Telecom

ADEBAYO ADETOLA ALIU | B1158902@live.tees.ac.uk

School of Computing and Digital Technologies,

Teesside University, United Kingdom

*Abstract*— This study focuses on the critical issue of customer churn in the telecommunications industry, exploring various predictive models to identify at-risk customers effectively. It delves into the effectiveness of traditional logistic-regression, decision-trees, k-nearest neighbor, support-vector-machine and advanced ensemble models in detecting patterns indicative of potential churn. The analysis highlights Random-Forest's superior performance due to its ability to handle large datasets and complex customer behaviors efficiently, achieving high-accuracy and precision. Comparative studies underscore the varying strengths of each model, with advanced techniques like XGBoost also showing promising results. This study emphasizes the strategic importance of integrating these predictive models into business operations to enhance customer-retention strategies, reduce churn-rates, and ultimately improve profitability in the competitive telecom-sector.

Keywords—*Customer-churn, Machine-Learning, Telecom-Industry, Predictive-Analytics, Customer-retention, Model-Performance*

# **Introduction**

Customer-churn, which is defined as the loss of clients or customers who choose to discontinue their services with a company, is a significant concern in the telecommunications sector (Wei and Chiu, 2002). Understanding and forecasting churn is essential for preserving a steady income stream and fostering sustainable business growth in a sector marked by intense competition and high client acquisition expenses (Yang, Q et al.,2022). The direct financial-impact of losing customers, which not only lowers immediate revenue but also raises marketing and acquisition costs that are frequently much higher than those associated with keeping current ones, emphasizes the significance of churn analysis (Autor et al., 2020).

# **Related work**

Many data-driven approaches have been deployed to solve this problem, as the churn-prediction thoroughly documented. Naradhi and Palshikar (2011) state that studies have historically used statistical approaches and, more recently, sophisticated machine-learning techniques to forecast churn. These techniques sort through enormous volumes of customer data to find probable churn-predictors, ranging from individual demographics to comprehensive-service usage patterns.

The incorporation of machine-learning algorithms, which can handle big, complicated datasets, has become increasingly important in the sector. In Thangeda, Kumar, and Majhi's research (2024), neural-networks were employed to model the probability of churn based on customer-contact data, whereas Ullah et al.'s (2019) study included ensemble-approaches to enhance prediction-accuracy.

Even with the advances, it is still difficult to combine and understand different kinds of data in prediction-models. Research on the relative significance of several features such as contract-terms, billing-information, and customer-service interactions that are crucial for forecasting churn is also underway. However, there are still unresolved issues. This study aims to address these gaps by analyzing a comprehensive telecom dataset.

**Study Objective**: By leveraging a wide range of prediction methods, we seek to validate earlier findings and provide novel insights into the effectiveness of different algorithms in the highly competitive telecom industry.

# **Data-Exploration and Features-Selection**

**Data-Cleaning-and-Preprocessing**: Understanding the distribution, quality, and structure of the data in the dataset which contained extensive customer-data from a telecom-company was the primary goal of the initial data research. The dataset from kaggle[🙁📡 Telecom Customer Churn Prediction (kaggle.com)](https://www.kaggle.com/datasets/shilongzhuang/telecom-customer-churn-by-maven-analytics) included 7,043 entries, each of which represented a customer. It had 37 attributes, which included billing-information, demographics, service details, and measures related to customer-interaction.

**Quality-Inspected**: A comprehensive-evaluation of the data-quality was carried out. The absence of duplicate records was determined, indicating clean data in terms of consumer uniqueness.

**Managing Missing-Value and Outlier**: To address inconsistencies, missing-values, and outlier-adjustments, 'No-Internet-Service' was imputed to fill in the blanks in sections like Internet-Type and other internet-features, presuming that the missing entries did not have internet-access. Similarly, to preserve the integrity of the distribution, missing-values for numerical-variables such as the average monthly long-distance charges were imputed using the median value. Additionally, to enhance model-performance, a few characteristics with outliers were normalized or rectified.

# **Exploratory Data Analysis (EDA)**

The variability and core tendency of characteristics were explained by descriptive statistics. Indicating varying customer-behaviors and usage-patterns, metrics such as Age, Tenure in Months, and Monthly-Charges, for example, displayed significant variation. The identification of discrete consumer categories that are susceptible to churn depends on this heterogeneity.

Younger-Customers are presumably more prone to churn, according to Fig1. Different service expectations compared to older customers, changes in lifestyle, or the pursuit of better discounts could all be the cause of this. Younger-Customers' greater incidence of customer-churn may suggest that specific retention tactics are required, such as loyalty plans or tailored offers.

Since there is a good retention rate across all age groups, the company might consider analyzing the factors contributing to this success to apply them more effectively to the age groups with higher churn rates.

**Data-Driven Actions**: Telecom-businesses may choose to concentrate on tailored engagement and retention-tactics for various age-groups considering these insights. Younger-customers may find tech-savvy offerings and flexible plans more appealing, but older generations may place a higher importance on stability and customer service.

Fig1 shows that age influences customer-behavior when it comes to churn, and that age-sensitive retention strategies are necessary.

A graph of customer behavior

Description automatically generated

Fig: 1

Customers who churned generally paid higher monthly charges (median near 80), whereas those who stayed or joined have lower and similar median charges, suggesting higher-charges might be a factor influencing customer-churn in Fig2.

A diagram of a group of boxes

Description automatically generated with medium confidence

Fig:2

# **FEATURES SELECTION**

Choosing the appropriate features is essential to creating prediction-models that work. Both statistical-analysis and subject expertise served as guidance for the selection process: associations between features and the turnover rate were investigated. Particularly noteworthy features included contract-type, monthly-charges, and tenure, which all had strong correlations with churn. Customers with high monthly-bills or those on month-to-month contracts, (Fig3), were more likely to leave.

A graph of a number of different colored bars

Description automatically generated with medium confidence

Fig: 3

The correlation-matrix's heatmap (Fig4) gave a visualization of the relationships between the various features. Customers tend to spend more when they remain longer, according to a strong positive association found between Tenure and Total-Charges. Predictive-modeling feature selection was aided by an understanding of these linkages.

A screen shot of a computer generated image

Description automatically generated

Fig: 4

**Feature-Engineering:** New features were derived from existing data to enhance the model’s predictive capability. For example, the categorical and numerical columns were formed to handle the missing value accordingly. Techniques such as

# **EXPERIMENTS**

**Experimental-Design**: The objective was to identify the most accurate predictor of customer churn through a comparative analysis of multiple machine learning models. The experimental approach included:

**Data-Preparation**: Data underwent preprocessing including handling missing values, encoding categorical variables, scaling numerical features, resampling, and partitioning into training and testing subsets. This ensured clean data presentation and aimed to enhance prediction accuracy by reducing biases.

**Data Imbalanced** – Resampling-techniques like SMOTE and Random-Under-Sampler were used to balance dataset. This enhances models to avoid bias, improved prediction accuracy and fairness in predictive decision.

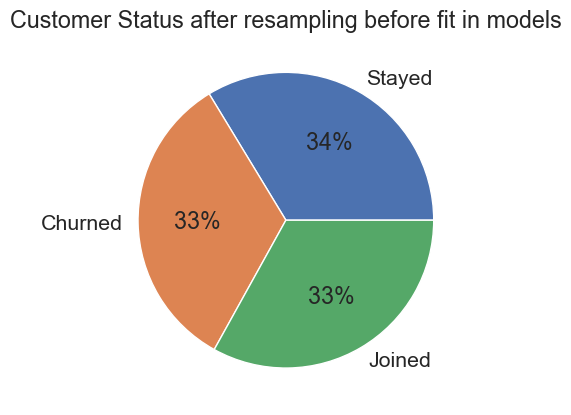
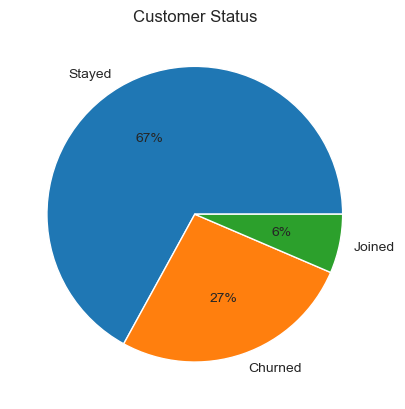


Fig:5

**Model Selection**: Models were chosen for their effectiveness in multi-class classification and diverse operational mechanisms:

Logistic Regression (LG): Used for its multinomial capabilities and probabilistic insights into customer churn.

Random Forest: Leveraged for its accuracy and robustness through averaging multiple deep decision trees.

XGBoost: Selected for its proficiency in handling complex nonlinear interactions and varied data types.

Support Vector Machines (SVM): Applied due to their performance in high-dimensional spaces.

K-Nearest Neighbors (KNN): Used to examine the impact of similarity-based reasoning on churn prediction.

Naive Bayes: Chosen for their probabilistic classification capabilities, assuming strong independence between features.

**ANOVA(Best-feature-selection)**: Models were subjected to k-fold cross-validation (commonly k=10 or 20) to ensure that the model's effectiveness was not dependent on the way the data was split. The mean and standard deviation of the resulting metrics across all folds were used to assess model stability and reliability.

**Hyperparameter-tuning**:

GridSearchCV techniques were used to fine-tune each model to identify the ideal combination of parameters that produce the best results.

**Explainable AI** like SHAP was used to explain how high effective model arrived at its decision.

**Implementation:**

To ensure model consistency and reliability, the following steps were implemented.

The models were trained using 80% of the dataset to learn correlations between feature patterns and outcomes like churn. Cross-validation within the training dataset was employed to mitigate overfitting and provide an objective measure of model capability. Models were finally tested using a separate test set, not involved in the training phase, to assess their generalization capabilities. Techniques like SMOTE and Random-Under-Sampler were used to address class imbalance by adjusting the majority class (non-churners). This enhances the model's ability to detect nuanced trends that indicate churn. The Standard-Scaler was applied to normalize the dataset’s features, ensuring no single feature disproportionately influences the model due to scale differences. This normalization aids in equitable contribution from all features, enhancing model accuracy and training efficiency.

# **RESULT**

Several important measures were used in the study to give a comprehensive picture of each model's efficacy.

Accuracy: Evaluates how well the model performs overall across all predictions.

Precision estimates the percentage of accurately detected expected churns and indicates the accuracy of positive forecasts.

Recall (Sensitivity): Evaluates the model's capacity to identify real data churns.

F1-Score: Considers both false-positives and false-negatives, giving a single number that balances precision and recall.

The area under the receiver operating characteristic curve is represented by the ROC-AUC score. A higher AUC denotes a better performing model across all classification criteria.

The performance of each classifier across the many measures discussed is summarized in the table below.

**Table of Performance Metrics**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **Positive-Rate** | **ROC AUC** |
| --- | --- | --- | --- | --- | --- | --- |
| RF | 90.47**%** | 90.38**%** | 90.47**%** | 90.39**%** | 100.50 | 97.67 |
| SVM | 85.84**%** | 85.98**%** | 85.84**%** | 85.80**%** | 105.16 | 93.67 |
| LR | 79.48**%** | 79.57**%** | 79.48**%** | 79.36**%** | 106.74 | 93.33 |
| KNN | 84.29**%** | 85.03**%** | 84.29**%** | 84.16**%** | 109.57 | 93.67 |
| NB | 77.86**%** | 77.72**%** | 77.86**%** | 76.99**%** | 111.02 | 91.67 |
| DT | 84.75**%** | 84.49**%** | 84.75**%** | 84.51**%** | 101.98 | 94.67 |
| XGB | 89.76**%** | 89.65**%** | 89.76**%** | 89.67**%** | 100.99 | 97.33 |

**Analysis of Results**

High-Performance of models: Random Forest excels as the top-performing model in our analysis, showcasing exceptional precision, accuracy, recall, F1 score, and an ROC-AUC. This model effectively identifies customers at risk of churn while minimizing false positives, crucial for cost-effective customer retention strategies.

**Key Advantages**

Precision and Recall Balance: Random-Forest achieves an optimal balance between precision and recall, crucial for reducing false alarms and effectively targeting retention efforts.

Superior ROC-AUC: A ROC-AUC of 97.67% demonstrates Random Forest's excellent ability to distinguish between churners and non-churners, which is vital for tailoring interventions based on varying customer risk levels.

Comparison with Other Models:

While XGBoost and Support Vector Machine also perform well, Random Forest consistently outshines them with its robust handling of the diverse data characteristics typical in telecom datasets, including customer demographics and usage patterns.

**Strategic Deployment Considerations**

Efficiency vs. Complexity: Random Forest's computational demands should be weighed against its high predictive accuracy and the insights it provides, making it suitable for dynamic markets like telecom.

**Actionable-Insights**: The model's interpretability is key for strategic decision-making, helping to pinpoint drivers of customer-churn and tailor interventions.

**Adaptability**: Ongoing tuning and updates to the Random-Forest model are recommended to keep pace with evolving customer-behaviors and market conditions.

Findings are especially helpful for improving customer-retention efforts in the telecom and related industries.

# **Discussion, Conclusions, and Future Work**

**Discussion**

The recent investigation sheds light on customer-Churn in the telecom-sector, aligning with previous findings that underscore the importance of price sensitivity and service commitment in retaining customers. The results highlight age, tenure, monthly-charges, and contract-type as significant predictors of churn.

Younger-Customers exhibit higher churn-rates, possibly due to their need for adaptability or market responsiveness.

High monthly-charges frequently lead to customer-churn, emphasizing the importance of competitive pricing strategies.

Customers on month-to-month contracts are more prone to churn, suggesting that longer-term contracts may foster better customer-retention.

Model-Performance: This study thoroughly assesses various predictive-models for churn-prediction, with Random-Forest emerging as notably effective. Its capacity to handle complex relationships and large datasets contributes to its high-accuracy, precision, and ROC-AUC-score, showcasing its superiority in predicting customer-behavior. Additionally, models like XGBoost and Decision-Trees are commended for their robust performance, especially in adaptability and scalability. However, despite the strengths of ensemble models, simpler ones like Logistic Regression remain crucial for their interpretability and ease of implementation, essential for operationalizing insights in business contexts.

**Conclusions**

This study concludes with an important when it comes to forecasting customer-churn, advanced ensemble techniques more especially, like Random-Forest perform better. Customer-Churn can be significantly influenced by predictive features such as tenure, contractual duties, and pricing structures. For the best model performance, proper preprocessing methods such feature scaling for standardization and under-sampling to solve class-imbalance are crucial.

The findings highlight the need to use data to inform retention efforts and the effectiveness of machine learning in extracting useful information from large, complex datasets.

**Future-Work**

Potential areas for more research and development are outlined below:

Adding more detailed temporal data to the dataset may aid in forecasting future trends and comprehending churn patterns over time.

Creating additional features could provide deeper understanding of consumer behavior perhaps using interaction terms or more intricate transformations.

Investigating alternative models or hybrid strategies may increase the robustness and accuracy of predictions.

Putting in place real-time analytics platforms for ongoing observation and early churn risk-identification.

Measuring the efficacy of various retention tactics in a controlled trial and incorporating the predicted model into active churn-intervention methods.

Further refining the predictive models could involve considering external economic and geopolitical issues that could impact customer behavior.

Heading forward, reducing churn and improving customer experience will need not only improving analytical-techniques but also incorporating insights into corporate procedures and customer engagement-tactics.

# **REFERENCE**

Autor, D., Dorn, D., Katz, L.F., Patterson, C. and Van Reenen, J. (2020). The Fall of the Labor Share and the Rise of Superstar Firms. *The Quarterly Journal of Economics*, 135(2). doi: <https://doi.org/10.1093/qje/qjaa004>.

Saradhi, V.V. and Palshikar, G.K. (2011). Employee churnprediction. *Expert Systems with Applications*, 38(3), pp.1999–2006. doi: <https://doi.org/10.1016/j.eswa.2010.07.134>.

Thangeda, R., Kumar, N. and Majhi, R. (2024). A neural network-based predictive decision model for customer retention in the telecommunication sector. Technological Forecasting and Social Change, [online] 202, p.123250. doi: <https://doi.org/10.1016/j.techfore.2024.123250>.

Ullah, I., Raza, B., Malik, A.K., Imran, M., Islam, S.U. and Kim, S.W. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access*, 7, pp.60134–60149. doi: <https://doi.org/10.1109/access.2019.2914999>.

Wei, C.-P. and Chiu, I-Tang. (2002). Turning telecommunications call details to churn prediction: a data mining approach. *Expert Systems with Applications*, 23(2), pp.103–112. doi: <https://doi.org/10.1016/s0957-4174(02)00030-1>.

Yang, Q., Li, B., Wang, P., Xie, J., Feng, Y., Liu, Z. and Zhu, F. (2022). LargeMetabo: an out-of-the-box tool for processing and analyzing large-scale metabolomic data. *Briefings in bioinformatics*, 23(6). doi: <https://doi.org/10.1093/bib/bbac455>.

Web - <https://www.kaggle.com/datasets/shilongzhuang/telecom-customer-churn-by-maven-analytics>