Predicting Customer Churn

Tanishq Tiwari
2021496

Deepak Thappa
2021319

Madhav Krishan Garg
2021333

Om Garg
2021481

Naman Rawat
2021071

Abstract

This report explores machine learning techniques for predicting customer churn in the telecom sector. It leverages extensive preprocessing, including SMOTE and SMOTEENN for class imbalance, feature selection using Chi-Square and ANOVA, and scaling continuous variables. Seven classification models—XGBoost, Random Forest, SVM, and more—were compared using accuracy, F1 Score, and ROC-AUC metrics. XGBoost emerged as the most effective, achieving 95.93% accuracy. Insights into key customer behaviors and strategic retention recommendations emphasize ML's potential to enhance revenue, reduce churn, and optimize customer satisfaction. GitHub Link

1. Introduction

Churn in business refers to the number of customers who stop using a service after a certain period of time. This phenomenon is common in subscription-based services, particularly in the bussiness sectors, where users tend to switch operators frequently. Customers may switch for a variety of reasons.

1. The domain is highly competitive, with numerous players providing comparable services. The intense competition leads to promotional offers and aggressive pricing, attracting customers to switch providers.
2. Poor experiences like call drops, internet outages, and poor customer service can also significantly contribute to churn.
3. The latest technological advancements like 5G, WiFi calling, and eSIM can make older plans feel outdated. Customers may churn to access newer features or faster speeds offered by competitors.

For a telecom company, high churn rates mean loss of revenue as the company loses customers, replacing those churned customers with new ones is expensive, which leads to higher acquisition costs reducing the margins and high churn rates damage the company's reputation in the market. Predictive churn rates can help companies proactively address this issue and improve retention rates. This can also help them to know their customer needs better, and the information can then be used to implement targeted retention strategies, such as offering personalised promotions,improving customer service, or introducing new features that meet the specific needs of at-risk customers. By predicting churn and implementing effective retention strategies, companies can enhance customer satisfaction, improve revenue generation, and maintain a competitive edge in the market.

2. Literature Review

This literature review examines three recent research papers focused on churn prediction in the telecom sector, highlighting the use of machine learning and big data platforms to tackle this problem.

Ahmad et al. (2019) [1] explores churn prediction in SyriaTel (arabic telecom company), utilizing a large dataset of customer information spanning over nine months. The research utilizes a big data platform to handle the data's volume and variety, also integrating Social Network Analysis (analysing social structures through networks and graph theory) features for enhanced prediction accuracy. The study demonstrates the effectiveness of XGBOOST algorithm, achieving an AUC value of 93.3%. This work highlights the potential of big data platforms and SNA for churn prediction in telecom, offering valuable insights into customer behaviour.

Prabadevi et al. (2023) [2] focuses on analyzing the performance of various machine learning algorithms for

early churn prediction in the telecom industry. The study compares Logistic Regression, KNN, Random Forest, and Stochastic Gradient Booster, concluding that SGB performs best among all the models, achieving an AUC score of 0.84. This research emphasizes the importance of selecting the appropriate algorithm based on the dataset and goals.
Srinivasan et al. (2023) [3] studies the effectiveness of machine learning models in predicting churn on, employing Decision Tree and Random Forest. While both models initially show poor results due to data imbalance, implementing SMOTE-ENN sampling techniques significantly improves performance, with Random Forest having 95% accuracy score. This study underscores the importance of addressing data imbalance for effective churn prediction.
It has been evident from the survey that machine learning and artificial intelligence play a wider role in customer churn analysis, several opportunities for further research remain:

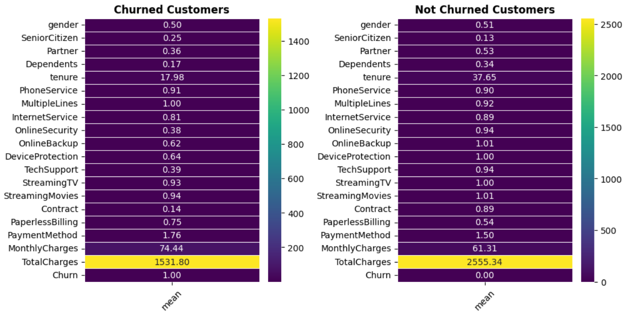
1. More sophisticated feature engineering and data processing techniques could further enhance churn prediction models.
2. Creating an ensemble of machine learning models could potentially lead to more accurate and robust churn prediction systems.
3. Building models that are more interpretable and explainable, enabling businesses to understand the underlying drivers of churn, is also a critical requirement.

3. Dataset description

3.1. EDA

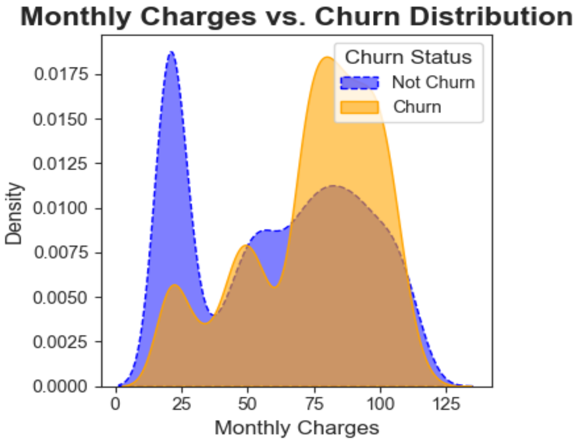
Statistical Analysis: Customers who churned exhibit significantly lower mean tenure (17.98 months) compared to those who remained (37.57 months), alongside shorter contract durations and higher mean monthly charges (74.44 vs. 61.27). Features like Online Security, Online Backup, Device Protection, and Tech Support are more prevalent among retained customers, who also have higher total charges (2555.34 vs. 1531.80). The presence of numerous categorical features leads to concentrated mean values around 0.

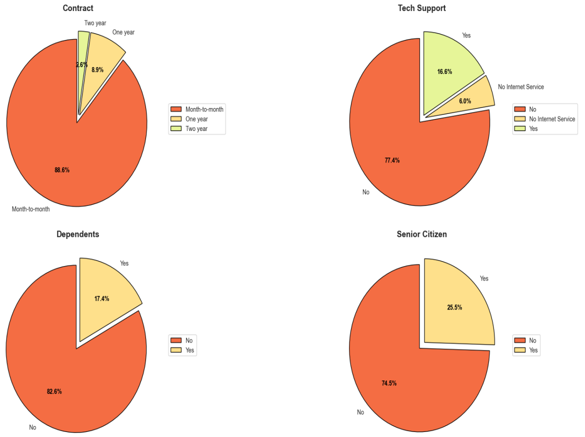
Categorical Feature Analysis: Male customers are more likely to churn than females, while Senior Citizens exhibit lower churn rates. Not having a partner or dependents increases churn likelihood. Regarding services, having phone service and multiple lines correlates with lower churn, and fiber optic internet users are more prone to churn than DSL users. The absence of online security, online backup, etc., indicates higher churn. Contract-wise, month-to-month contracts show higher churn than one-year and two-year contracts. PaperlessBilling customers are more likely to churn. In payment methods, electronic checks users have a higher churn risk, followed by mailed checks, bank transfers (automatic), and credit cards (automatic). Understanding these hierarchies is essential for devising effective retention strategies.



Class Imbalance: The dataset exhibits an imbalanced distribution, with a ratio of approximately 3:1 for Not-Churn to Churn customers. This imbalance introduces a bias in predictions, where the model may lean towards accurately predicting Not-Churn instances.

Numerical Features Distribution The tenure distribution exhibits a bimodal pattern with peaks at 0-70, indicating the presence of two distinct groups within the customer base. MonthlyCharges create a bimodal distribution with peaks at 20-80, suggesting the existence of two prevalent cost structures or service tiers. Total Charges displays a positively or rightly skewed distribution, indicating a concentration of lower values with a gradual tapering towards higher values.





4. Preprocessing

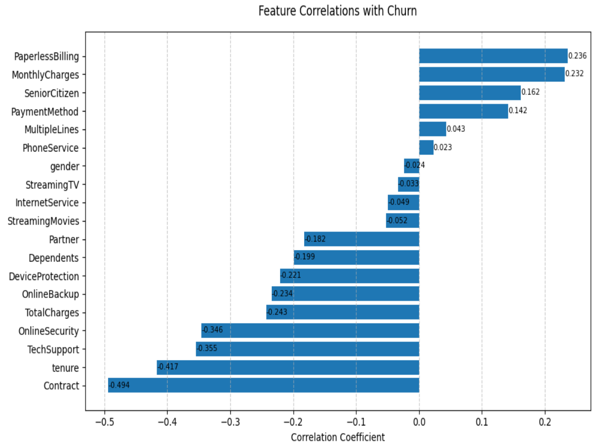
4.1. Addressing Data Imbalance with SMOTE

Class distribution before SMOTE: Counter(0: 5174, 1: 1869) Class distribution after SMOTE: Counter(0: 5174, 1: 5174)

4.2. Addressing Data Leakage

To safeguard against Data Leakage, a prudent approach involves employing a train-test-split prior to any transformations. This ensures that transformations are applied based on the training data, maintaining integrity for both training and test datasets.

4.3. Correlation Matrix and Feature Selection



4.4. Data Scaling

Applied to features with non-normally distributed data. Features like tenure, MonthlyCharges, and TotalCharges exhibit a right-skewed and bimodal data distribution, making normalization suitable. Typically used for features that display a normal (Gaussian) distribution. In the current dataset, none of the features undergo standardization.

5. Methodology

The dataset underwent extensive preprocessing to ensure data quality and suitability for analysis. Missing values were handled appropriately, data types were converted as needed, and categorical features were label-encoded. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was initially employed, enhancing the representation of the minority class. Continuous variables, including *Tenure*, *MonthlyCharges*, and *TotalCharges*, were scaled to standardize their ranges, while categorical features were selected using the Chi-Square test and numerical features based on their ANOVA scores. Features with minimal correlation to the target variable (*churn*) were excluded from the analysis to streamline model training.

Seven classification algorithms—XGBoost, Random Forest, Decision Tree, Logistic Regression, Support Vector Machine (SVM), Neural Network, and Naive Bayes—were trained using an 80-20 train-test split. Model evaluation employed metrics such as test accuracy, 5-fold cross-validation accuracy, standard deviation, ROC-AUC score, and F1 score to comprehensively assess performance. To further improve the dataset's quality and the models' predictive power, SMOTEENN (a combination of SMOTE and Edited Nearest Neighbors) was later applied, which not only oversampled the minority class but also removed noisy samples from the dataset.

Hyperparameter tuning was expanded for XGBoost, Random Forest, and SVM models, employing an exhaustive grid search across a broader range of parameters to optimize model configurations. Consistency in preprocessing and model fitting was ensured by implementing a standardized pipeline, which integrated scaling and model training steps uniformly across all algorithms. This systematic approach enabled a robust evaluation of model performance while addressing potential biases introduced during data preparation and training.

6. Results and Analysis

6.1. Results

Refer to Table 1 for results.

6.2. Analysis

The results indicate that XGBClassifier is the top-performing model, achieving the highest accuracy

Model	Accuracy	ROC	CV	F1
XG Boost	95.93	96.54	95.93	96.80
Random Forest	95.67	96.82	95.67	97.12
SVM	94.18	94.10	94.18	94.89
Decision Tree	94.01	93.77	94.01	94.47
Logistic Regression	92.26	91.93	92.26	92.78
Neural Network	88.95	88.48	91.67	90.37
Naive Bayes	88.95	88.48	88.95	90.37

Table 1: Performance metrics of various models

(95.93%), F1 score (96.8), and ROC AUC (95.54), making it the most reliable. Its performance aligns with expectations from literature, as gradient-boosted methods often perform well on structured data due to their ability to model complex relationships. However, gradient-boosting models tend to have high computational demands during training, which could limit their scalability in real-time or resource-constrained environments. RandomForest follows closely with solid accuracy (95.67%) and an F1 score of 97.12. High scores are potentially due to its ensemble mechanism. However, the lower ROC AUC indicates it may not differentiate well between classes compared to XGBClassifier.

Support Vector Machines (SVM) and Decision Tree models deliver competitive results, with accuracies of 94.18% and 94.01%, respectively. SVM's strong performance in generalizability and its high F1 Score (94.89) make it a dependable choice. Logistic Regression, while still robust, shows a slightly lower accuracy (92.26%) and F1 Score (92.78). SVM's generalizability suggests it could be suitable for smaller datasets or cases where clear margins between classes exist, while Decision Trees might be preferable for their simplicity and interpretability.

The ANN Classifier having two hidden layers of size 32 and 16 with ReLU and regularisation, not leading in overall accuracy (88.95%), provides a balanced performance with an F1 Score of 90.37. This model can serve as a benchmark model due to its simplicity and low computational cost. However, its performance gap highlights the necessity of more complex algorithms. The performance reflects the importance of architecture optimization, data preprocessing, and potentially larger datasets for deep learning models.

The Naive Bayes model underperforms with an accuracy of 88.95%, suggesting limitations in handling complex data patterns, making it less suitable for high-performance requirements which are likely due to independence assumption inhereted in this model.

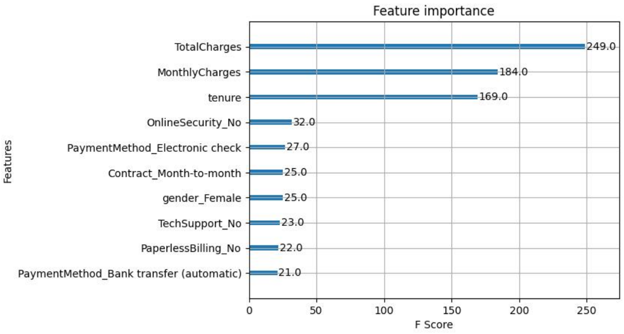
XGBClassifier emerges as the most effective model, combining high accuracy, F1 Score, and ROC AUC. However, practical deployment should consider resource constraints, model interpretability, and the specific application domain. Random Forest is a strong alternative, while SVM provides an option for simpler implementations with competitive performance. ANN and Naive Bayes, while showing limitations, could be explored further with additional tuning or in specialized use cases.

7. Model Explainability

The model shows the most important features and most sensitive to churn:

1. TotalCharges, MonthlyCharges, and Tenure are the three most contributing factors. This means customers with a higher total accumulated charges and monthly expenses or smaller tenures are at more risk of churn. It gets supported by the insights developed during doing EDA.

2. Other categorical factors, including OnlineSecurity (No) and PaymentMethod (Electronic Check), also seem to be significant so have some customer behavior and preference that would increase the chances of churning. Results validate the preprocessing steps comprising Chi-Square selection and ANOVA for prioritization of impactful variables.



8. Conclusion

This study demonstrates the effectiveness of machine learning models in predicting customer churn in the telecom sector. By addressing key challenges such as class imbalance using SMOTE/SMOTEENN and optimizing features through Chi-Square and ANOVA tests, the analysis highlights the significance of data preprocessing. XGBoost emerged as the most effective model, with a 95.93% accuracy and strong ROC-AUC, emphasizing its ability to capture complex patterns in the dataset. Strategic insights derived from the analysis, such as targeting senior citizens, enhancing payment options, and providing affordable entry-level plans, showcase how data-driven decisions can mitigate churn. These findings underline the value of ML-driven approaches in retaining customers, improving service quality, and strengthening competitive positioning in a highly dynamic market where costs of customer retention and acquisition are high.

9. Contributions

The project contributions are as follows: Tanishq handled EDA, decision tree, report and presentation. Deepak worked on preprocessing, random forest, and report. Madhav contributed to EDA, preprocessing, Naive Bayes, ANN, report. Om focused on EDA, SVM, XGBoost, and data augmentation. Naman contributed by working on logistic regression and the presentation.

References

[1] Abdelrahim Kasem Ahmad, Assef Jafar, and Kadan Aljoumaa. Customer churn prediction in telecom using machine learning in big data platform, 2019.

[2] B. Prabadevi, R. Shalini, and B.R. Kavitha. Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks*, 4:145–154, 2023.

[3] R Srinivasan, D Rajeswari, and G Elangovan. Customer churn prediction using machine learning approaches. In *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, Jan. 2023.