Customer Churn Prediction- Telco Customer Analysis

Akshada G Malpure agmalpur@ncsu.edu
Rucha M Kulkarni rkulkar5@ncsu.edu
Rishi A Dange radange@ncsu.edu

1 Background and Introduction

Customer retention has become a crucial focus in subscription-based companies where competition is fierce. A depletion in retention may be caused due to lack of customer engagement and support, lower prices offered by competitive companies, etcetera. Losing customers, referred to as churn, directly impacts profitability. Predicting churn can help companies take proactive steps to retain customers by identifying patterns in their behavior, such as service usage and payment habits.

In this report, we aim to develop a predictive model to identify the likelihood of a customer churn, utilizing the Telco Customer Churn dataset. Churn prediction, through the help of data science and machine learning, is considered an effective solution to this problem. By analyzing customer data, the business team can intervene early, ensuring customer satisfaction and long-term loyalty.

1.1 Problem

The central problem explored in this project is predicting telecom customer churn, a complex challenge that depends on several factors like demographics, service providers, contract types, and so on. The objective is to build a machine learning model capable of accurately predicting which customers are at risk of leaving the service and understanding the key factors why customer churn, thus enabling the company to take timely action to retain these customers.

1.2 Literature survey

Several machine learning techniques have been applied to predict customer churn in the telecommunications sector. Wagh et al. (2024) highlight the importance of demographic and usage features in their analysis using methods like logistic regression and random forests. They emphasize how feature engineering enhances model performance [1]. AlShourbaji et al. (2023) demonstrate that combining gradient boosting machines with optimization techniques leads to more accurate churn predictions, especially with feature selection and hyperparameter tuning [2]. Additionally, Jain et al. (2021) review ensemble methods, such as XGBoost, and underscore the need for balancing imbalanced datasets and extracting domain-specific features to improve results [3].

This project builds on these studies by applying modern machine learning techniques to the Telco Customer Churn dataset, focusing on feature selection and predictive modeling for churn detection.

2 Method

This research employs a structured approach involving data preparation, feature selection, and model training to effectively manage the class imbalance issue and enhance predictive accuracy. The key components are as follows:

1. **Data Preprocessing:** This involves understanding the data, handling missing values, normalizing numerical features, and encoding categorical variables.
2. **Data Visualization:** Through this process, we apply various visualization techniques to further understand the underlying patterns in the data, and to analyze or optimize them further.
3. **Handling Bias in the Data:** SMOTE is a technique used to handle an imbalance in the data.

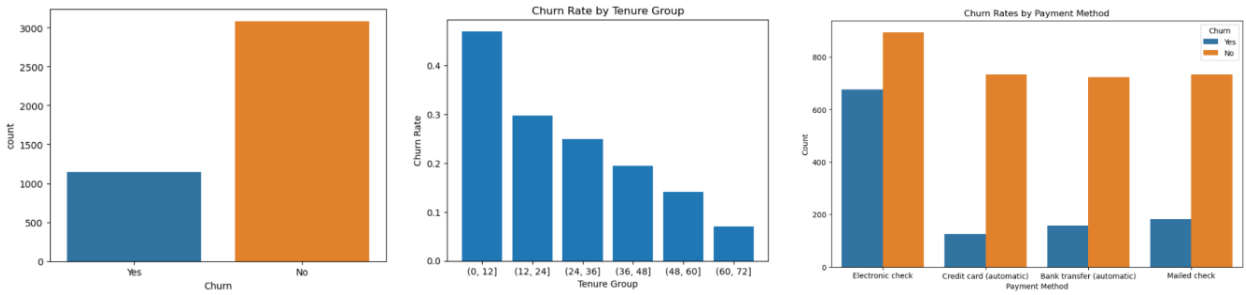

Figure 1: Churn analysis across categories: Distribution of churn vs. non-churn, churn rate by tenure groups, and churn patterns by payment method

4. **Model selection:** Training the preprocessed data using various predictive models, including Logistic Regression, K-NN Classifier, Naive Bayes Classifier, and Random Forest was utilized.
5. **Performance metrics:** The effectiveness of the models was assessed using accuracy, precision, recall, F1 score, and AUC-ROC. To further enhance the accuracy of the resultant model, techniques like hyperparameter tuning have been implemented.

3 Experiment setup

3.1 Dataset description

The Telco Customer Churn dataset, available on Kaggle, contains 7,043 customer records and 21 features. These features fall into three categories:

1. **Personal information** including details like customer gender, age, and dependents.
2. **Services subscribed** detailing services such as phone lines, internet, and security features.
3. **Billing information** which includes monthly charges, payment methods, tenure, and the overall cost of services.
4. The primary target feature in this dataset is **Churn**, a binary variable indicating whether or not a customer has discontinued their services.

The dataset contains both numerical and categorical variables, which will require pre-processing before model training. The dataset is split into a training set (60%) and a test set (40%). Out of the (60%) of train data, it was again split into an (80%) model training set and a (20%) model testing set to calculate the performance metrics.

3.1.1 Hypothesis

It is proposed that employing SMOTE for dataset balancing will lead to enhanced accuracy and recall in predicting churned customers. Furthermore, it is anticipated that more sophisticated models, such as Random Forest and Gradient Boosting, will yield better results compared to simpler models like Logistic Regression.

3.2 Experimental design

The experiment design focuses on how the study is structured to test the hypothesis effectively. It encompasses key stages like data pre-processing, feature selection and model training, ensuring a methodical approach to generating reliable results. The following subsections highlight the specific steps taken to prepare the dataset, identify relevant features, and build predictive models.

3.2.1 Data pre-processing

We now discuss how the initial data was prepared for analysis. It includes strategies for managing missing values, scaling, and encoding:

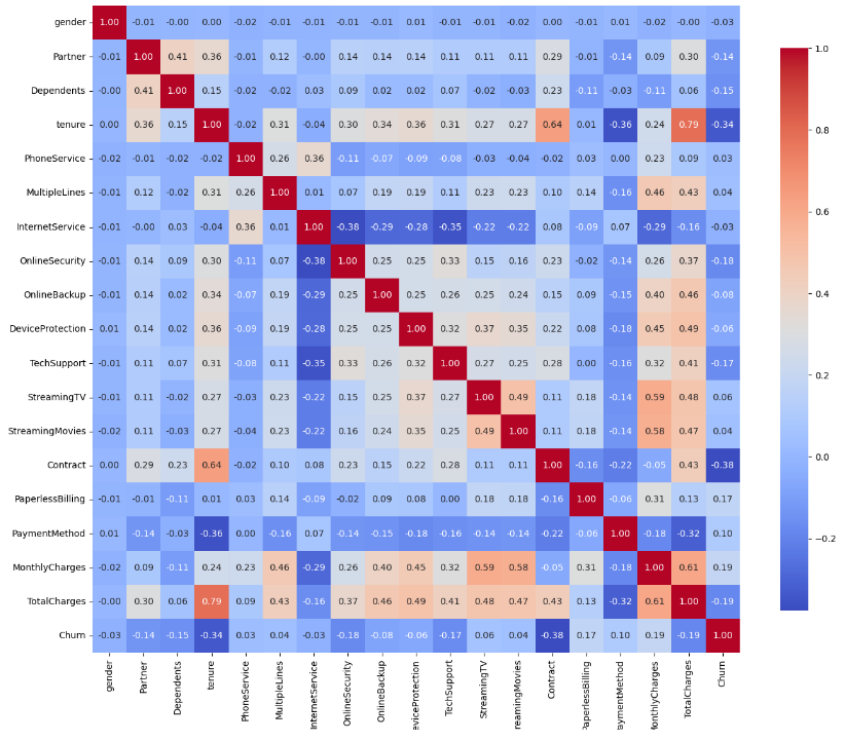

Figure 2: Feature heatmap

1. **Handling missing data:** Numerical features with missing values were filled with the mean, while categorical features were filled with the mode.
2. **Scaling:** Numerical variables were standardized using the StandardScaler to ensure they were of a comparable weightage.
3. **Encoding:** Categorical variables underwent One-Hot Encoding.

Feature selection is essential for identifying the most impactful predictors in our dataset, streamlining the model, and enhancing overall performance. Visualizations and correlation analysis are used to gain insights into the relationships between features and churn.

1. **Correlation Assessment:** Features with a strong correlation to the target variable, such as tenure and internetServices and paymentMethods were retained.
2. **Variance Thresholding:** Features exhibiting minimal variance are excluded to reduce extraneous noise.
3. **Feature Importance from Tree-based Models:** Decision trees are employed to rank the importance of various features.

3.3 Model training

We now highlight the process of training various machine learning models to optimize predictive accuracy and assess their performance on the evaluation dataset:

1. **Implementing SMOTE:** SMOTE was implemented to balance the dataset before dividing it into training and testing sets.
2. **Cross-validation:** Cross-validation methods were used to optimize hyperparameters and enhance model generalizability.
3. **Model Comparison:** Different models, including Logistic Regression, Naive Bayes, K-NN Classifier and Random Forest were trained and evaluated.

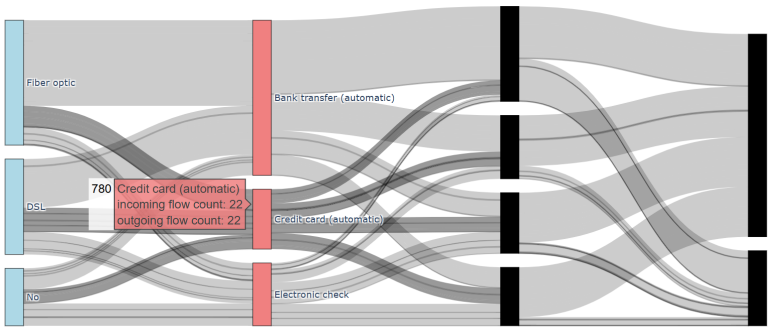

Figure 3: Sankey diagram between Flow of Internet Service, Contract Type, Payment Method, and Customer Churn

4 Results

The performance of each model is evaluated using metrics like accuracy, precision, recall, F1 score, and AUC-ROC. Based on the F1 score. The table below summarizes the performance metrics of each model tested:

	Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
0	Logistic Regression	0.746353	0.733129	0.774716	0.753349	0.746353
1	Random Forest	0.854943	0.882867	0.818476	0.849453	0.854943
2	Gaussian Naive Bayes	0.764992	0.748103	0.799028	0.772727	0.764992
3	K-Nearest Neighbors	0.769854	0.707865	0.918963	0.799718	0.769854

Figure 4: Accuracy metrics

The top two performing models are:

- Random Forest: F1 score of 0.849453
- K-Nearest Neighbours: F1 score of 0.799718

So far, the findings reveal that Random Forest, and K-NN exhibited superior accuracy and recall, indicating their reliability in predicting customer churn. However, by modelling the attributes further to optimize them, and through enhanced modelling and newer models, we can enhance the precision metrics even further.

5 Conclusion

Our study investigated the application of various machine-learning techniques for predicting customer churn within an imbalanced dataset. The use of SMOTE effectively addressed the imbalance, resulting in improved performance, especially with Random Forest and Gradient Boosting models, which achieved the highest metrics in terms of accuracy and recall. Future work may include further hyperparameter tuning, trying out new and more sophisticated models, and the exploration of alternative sampling methods to enhance model performance.

References

[1] Ibrahim AlShourbaji et al. *An efficient churn prediction model using gradient boosting machine and metaheuristic optimization.* 2023. DOI: 10.1038/s41598-023-41093-6.

[2] Hemlata Jain, Ajay Khunteta, and Sumit Srivastava. *Telecom churn prediction and used techniques, datasets and performance measures: a review*. 2021. DOI: 10.1007/s11235-020-00727-0.

[3] Sharmila K. Wagh et al. *Customer churn prediction in telecom sector using machine learning techniques*. 2024. DOI: 10.1016/j.rico.2023.100342. URL: https://doi.org/10.1016/j.rico.2023.100342.