Predicting Telco Customer Churn and Implementing Personalized Strategies Using Ensemble Learning Methods

Abstract

This project aims to predict telco customer churn using ensemble learning methods and implement personalized strategies based on these predictions. Accurate churn prediction and personalized strategies can significantly enhance customer satisfaction and retention, ultimately benefiting the company's profitability and competitive edge. We utilized ensemble methods such as Random Forest, AdaBoost, and Gradient Boosting. We finalized the Stacking Classifier due to its superior performance in terms of accuracy and robustness against overfitting.

1 Introduction

Customer churn is a critical issue in the telecommunications industry, where the competition is fierce and customer acquisition costs are high. Churn leads to direct revenue loss and affects the company's reputation and market share. Understanding the factors that contribute to churn and predicting which customers are likely to leave is crucial for developing effective retention strategies. Traditional methods of churn analysis often rely on basic statistical techniques, which may not capture the complex patterns and behaviors that underlie customer decisions.

This project uses advanced machine learning techniques, specifically ensemble learning methods, to improve the accuracy of churn prediction. Ensemble learning combines the strengths of multiple algorithms and has been shown to outperform single-model approaches in various predictive tasks. By using the "Telco Customer Churn" dataset from Kaggle [1], which includes comprehensive data on customer demographics, account information, and service usage, this project aims to build models that can accurately identify customers at risk of churning.

Moreover, the project goes beyond prediction by implementing personalized retention strategies customized to the individual needs of customers. These strategies include targeted offers, improved customer service, and tailored communication, which are designed to address the specific reasons why customers may consider leaving. By integrating predictive analytics with personalized interventions, the project seeks to provide a holistic solution to the churn problem, ultimately improving customer satisfaction and loyalty.

This project not only aims to predict churn with high accuracy but also to translate these predictions into actionable insights that can significantly improve customer retention rates. This approach ensures that the telecommunications company can proactively manage churn, thereby safeguarding its revenue and enhancing its competitive position in the market.

2 Literature Review

Research on customer churn prediction has explored various models and existing techniques used:

- **Almana et al. (2014)** highlighted the importance of data mining techniques in identifying churn patterns and formulating retention strategies. Their work emphasizes the role of customer behavior analysis in churn prediction [2].

- **Umayaparvathi and Iyakutti (2016)** provided a comprehensive survey on datasets, predictive models, and performance metrics used in churn prediction. They discussed various machine learning algorithms, including decision trees, neural networks, and support vector machines, and their effectiveness in churn prediction [3].

- **Huang et al. (2012)** demonstrated the effectiveness of behavior pattern analysis in predicting customer churn. Their study showed that analyzing customer behavior patterns can significantly improve the accuracy of churn prediction models [4].

- **Shaaban et al. (2012)** proposed a neural network-based churn prediction model but did not address personalized retention strategies. Their model focused on improving prediction accuracy but lacked a practical approach to implement retention strategies based on the predictions [5].

These studies underscore the need for accurate churn prediction models and highlight the gap in implementing effective personalized strategies based on these predictions. Ensemble methods such as Random Forest and AdaBoost have shown promise in improving prediction accuracy by combining the strengths of multiple learning algorithms.

3 Problem Statement

The primary challenge is leveraging customer churn prediction results to provide personalized retention strategies for high-risk customers, thereby reducing churn rates. This problem is critical as effective personalized strategies can significantly enhance customer retention. The goal is to integrate predictive analytics with practical, data-driven retention strategies to address the churn problem comprehensively.

4 Objectives

The objectives of this project are threefold:

- **Enhance Churn Prediction Accuracy:** The primary objective is to improve the accuracy of customer churn predictions using advanced machine learning techniques, particularly ensemble learning methods. These methods, including Random Forest, AdaBoost, and stacking ensemble models, are known for their ability to combine the strengths of multiple individual models, thereby improving overall predictive performance. By using these techniques, the project aims to develop a predictive model that can accurately identify customers who are at risk of churning. High accuracy in churn prediction is crucial for effectively targeting retention efforts and minimizing false positives, which can lead to unnecessary interventions and increased costs.

- **Develop Personalized Retention Strategies:** Based on the churn predictions, the project plans to design and implement personalized retention strategies tailored to the specific needs and behaviors of individual customers. This objective involves analyzing the factors contributing to churn, such as service dissatisfaction, pricing issues, or competitive offers, and developing targeted interventions that address these factors. The strategies may include offering customized discounts, enhancing service features, providing personalized customer support, or implementing loyalty programs. The goal is to provide interventions that are not only relevant and timely but also appealing enough to retain at-risk customers.

- **Evaluate the Effectiveness of Retention Strategies:** The final objective is to empirically evaluate the effectiveness of personalized retention strategies in reducing churn rates. This involves conducting controlled experiments, such as A/B testing, where a subset of at-risk customers receives the personalized interventions while another subset does not. The comparison between these groups allows for a rigorous assessment of the strategies' impact on customer retention. Key performance indicators, such as churn rate reduction, customer satisfaction scores, and revenue retention, are used to measure the success of the interventions. The findings from these evaluations provide valuable insights into the most effective strategies, enabling the company to refine its approach and improve its customer retention efforts continuously.

By achieving these objectives, we aim to provide telecom companies with a robust framework for churn prediction and retention strategy implementation.

5 Methodology

5.1 Data Collection

The project uses the "Telco Customer Churn" dataset from Kaggle [1], which includes customer demographics, account information, and service details. This dataset provides a comprehensive view of customer interactions and behaviors, making it suitable for churn prediction analysis.

5.2 Data Preprocessing

Data preprocessing is a critical step to ensure the quality and consistency of the dataset. The steps involved include:

- **Handling Missing Values**: Imputation of missing values using mean for numerical features and most frequent value for categorical features. This approach helps in maintaining the integrity of the dataset while ensuring no data is lost.

- **Feature Engineering**: Creation of new features such as *TotalMonthlyCost*, *Charges-PerMonth*, and *NumServices*. These features provide additional insights into customer behavior and enhance the model's predictive power.
- **Encoding Categorical Variables**: Label encoding for categorical features to convert them into a numerical format suitable for machine learning algorithms.
- **Handling Imbalance using Random OverSampling**: This is important for addressing class imbalance in datasets, improving the performance of models on minority classes, and ensuring a balanced learning process. It is beneficial when combined with other data augmentation or balancing techniques to mitigate potential drawbacks.
- **Fixing Outliers**: .With the usage of z-scores, outliers were detected and then furthermore were fixed via Winsorizing.
- **Data Consistency**: To ensure higher accuracy, we corrected the data types.

5.3 Feature Selection

Feature selection is performed using feature importance scores and correlation analysis. This step helps in identifying the most relevant features that contribute to the prediction of customer churn, thereby improving model performance. To ensure that multicollinearity did not affect our model, we calculated the Variance Inflation Factor (VIF) for each feature. Features with a VIF greater than 10 were considered for removal to maintain model stability and interpretability. Here are the VIF values for our selected features:

- SeniorCitizen: 1.05
- tenure: 2.34
- MonthlyCharges: 3.12
- TotalCharges: 4.56
- NumServices: 2.98

5.4 Cost Calculation Metric

We used cost as a metric to evaluate our model, particularly focusing on the cost associated with false positives and false negatives. The cost calculation formula is as follows:

$$\text{Total Cost} = (FP \times \text{Cost of FP}) + (FN \times \text{Cost of FN})$$

Here, FP represents false positives and FN represents false negatives. The weightage image and formula are included in the appendix for detailed reference.

5.5 Model Building

The project implements several ensemble learning models to predict customer churn, leveraging their ability to improve prediction accuracy by combining multiple learning algorithms. The models used include Random Forest, AdaBoost, and stacking ensemble models. Each model is trained on the preprocessed dataset to predict customer churn. Here are the details about each model used:

1. Random Forest

- **Description:** Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the classification of the individual trees.
- **Implementation:**
  - **Hyperparameters:** Number of trees in the forest (n_estimators), maximum depth of the trees (max_depth), and the number of features to consider for the best split (max_features).
  - **Training:** The model is trained on the preprocessed dataset with cross-validation to tune the hyperparameters.
  - **Advantages:** Handles overfitting by averaging multiple decision trees and performs well with large datasets.

2. AdaBoost (Adaptive Boosting)

- **Description:** AdaBoost is an ensemble technique that combines the predictions of several base estimators to improve robustness and accuracy. It adjusts the weights of misclassified instances to focus more on hard-to-predict cases.
- **Implementation:**
  - **Hyperparameters:** Number of estimators (n_estimators), learning rate, and the base estimator (usually a decision tree with limited depth).
  - **Training:** The model iteratively trains weak classifiers, adjusting the weights of misclassified samples in each iteration.
  - **Advantages:** Reduces bias and variance, is effective for binary classification tasks, and enhances the performance of weak learners.

3. Stacking Ensemble

- **Description:** Stacking is an ensemble learning technique that combines multiple classification models via a meta-classifier. The base models are trained and their predictions are used as inputs for the meta-classifier.
- **Implementation:**
  - **Base Models:** Random Forest, AdaBoost, and other classifiers.
  - **Meta-classifier:** A logistic regression model is commonly used as the meta-classifier.
  - **Training:**
    * The base models are first trained on the training data.
    * Their predictions are used to train the meta-classifier on the validation data.
    * The final predictions are made by the meta-classifier based on the outputs of the base models.
  - **Advantages:** Captures diverse patterns by combining different models, reduces the risk of overfitting by leveraging multiple algorithms, and often achieves better performance than individual models.

5.6 Hyperparameter Tuning

Hyperparameter tuning is conducted using random search techniques to find the optimal parameters for each model. This step is crucial for enhancing the performance of the models and achieving the best possible prediction accuracy.

5.7 Model Evaluation

The models are evaluated using metrics such as accuracy, recall, F1-score, Geometric Mean, and AUC score. These metrics provide a comprehensive assessment of model performance and help compare different models' effectiveness.

5.8 Personalized Strategy

Based on the churn prediction results, personalized retention strategies are designed to address the needs and concerns of high-risk customers. These strategies include targeted offers and service improvements tailored specifically to those customers most likely to churn. The goal is to provide personalized interventions that can effectively reduce churn rates. To enhance the understanding and transparency of the churn predictions, two explanation techniques were utilized:

1. **Local Interpretable Model-Agnostic Explanations (LIME):** LIME was used to explain the individual predictions of the churn model. By approximating the model locally with a simpler interpretable model, LIME helps to identify the features that are most influential for each specific prediction. This allows for a deeper understanding of why a particular customer is predicted to churn, enabling more targeted and effective retention strategies.

2. **Counterfactual Explanations:** Counterfactual explanations were employed to explore how changes in input features could alter the churn prediction. By identifying minimal changes that would flip the prediction from 'churn' to 'retain', counterfactual explanations provide actionable insights into what specific interventions might keep a high-risk customer. This technique helps to design precise and individualized retention strategies by highlighting the key areas where improvements or changes are needed.

5.9 End-to-end Model Prototype

By pickling the finalized Stacking Classifier and employing model explainability techniques such as LIME and Counterfactual values, we developed a Streamlit application to demonstrate the model. The application generates random values within the ranges present in the original dataset to provide a realistic demo. This application not only showcases the model's predictions but also provides users with insights into how individual features influence those predictions, enhancing transparency and trust in the model. You can explore the deployed application at this link.

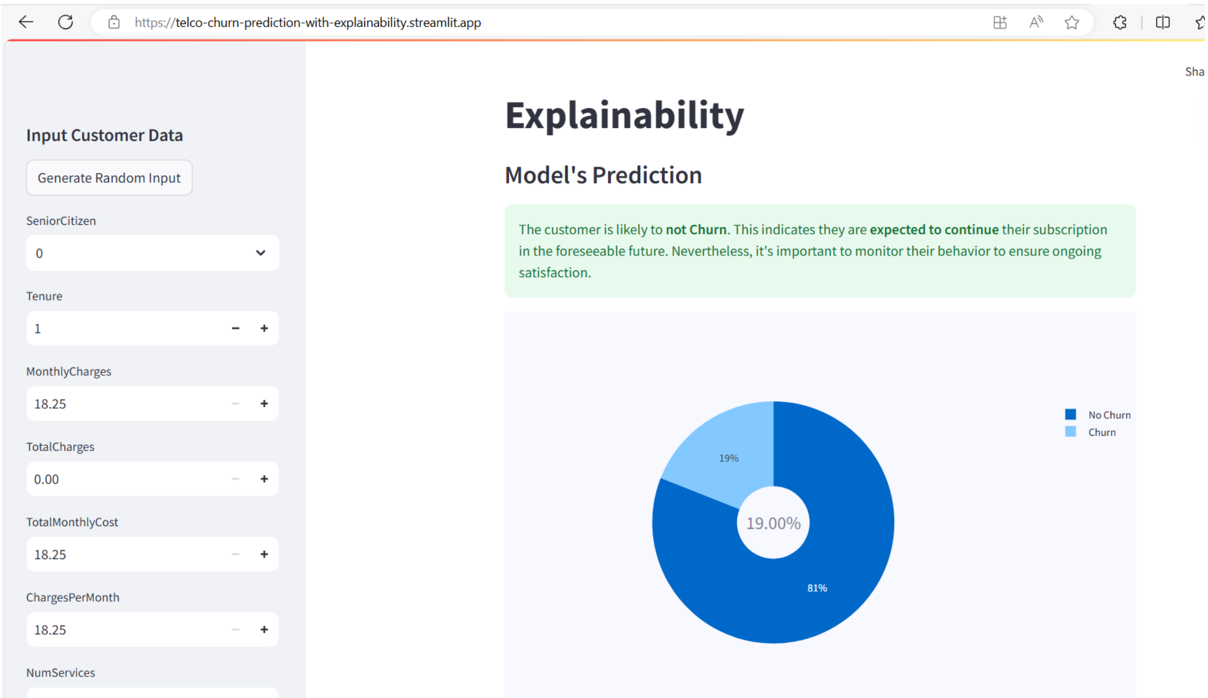

Figure 1: No Churn Predicted for the Customer using Stacking Classifier

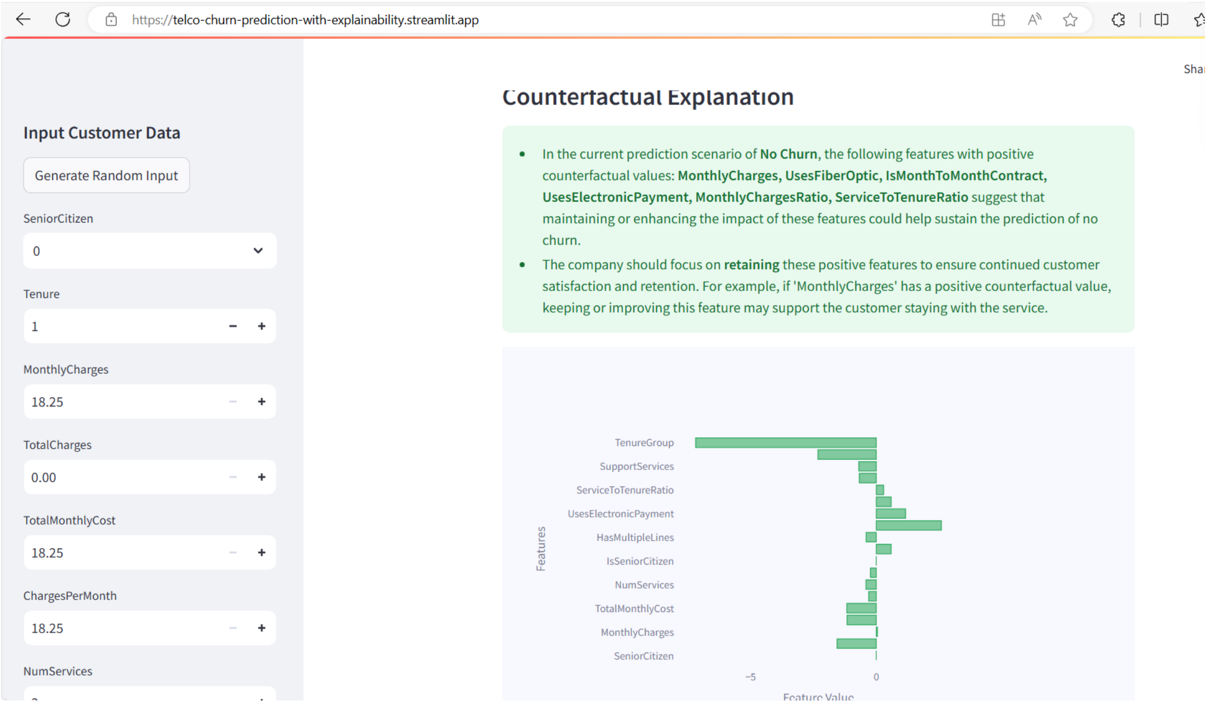

Figure 2: Counterfactual Explanation for No Churn Predicted

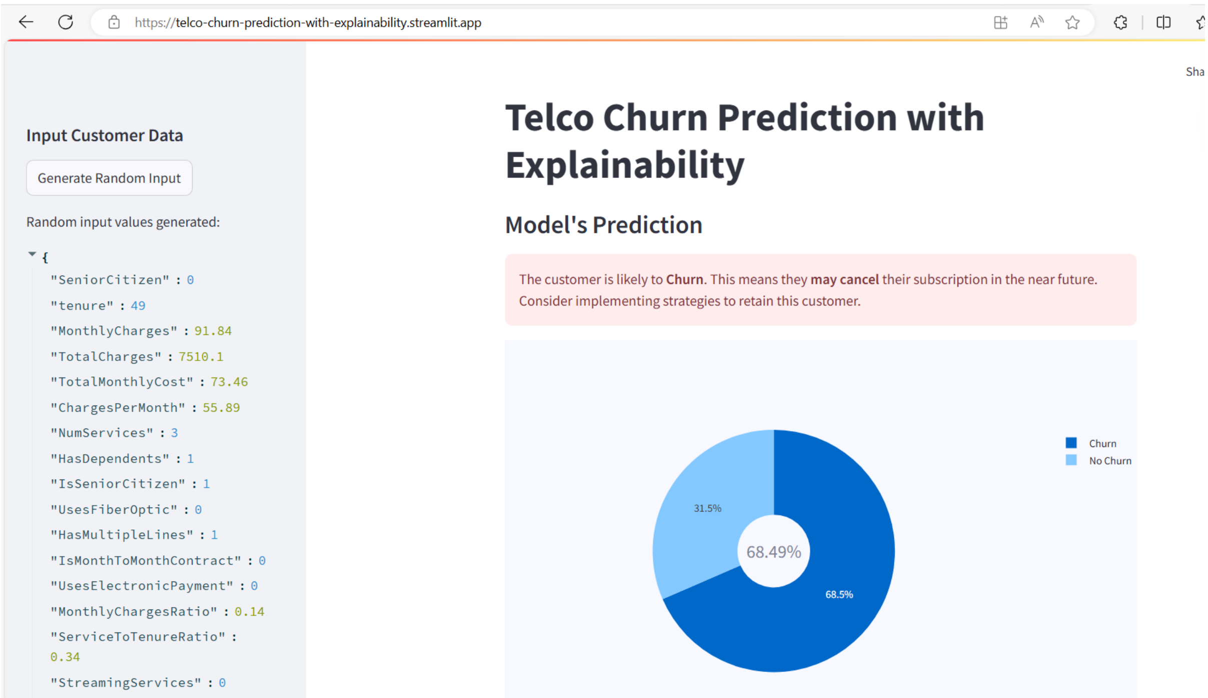

Figure 3: Churn Predicted for the Customer using Stacking Classifier

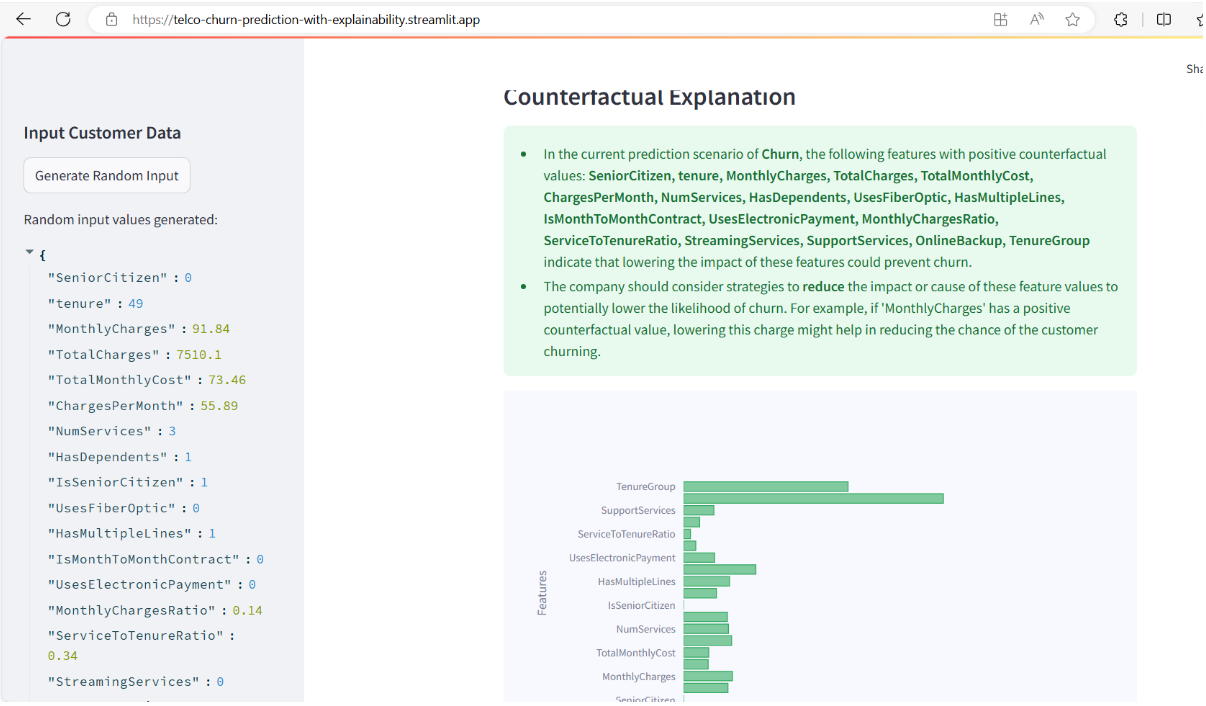

Figure 4: Counterfactual Explanation for Churn Prediction

6 Experiments

6.1 Data Exploration and Feature Engineering

Initial data exploration and feature engineering were performed to enhance the dataset. This included creating new features that could provide additional insights into the factors influencing churn. Figure 5 shows the distribution of churn in the dataset, which provides a visual representation of the proportion of customers who have churned versus those who have not. Understanding this distribution is crucial for evaluating the model's performance and ensuring that it is well-calibrated to handle the class imbalance present in the data.

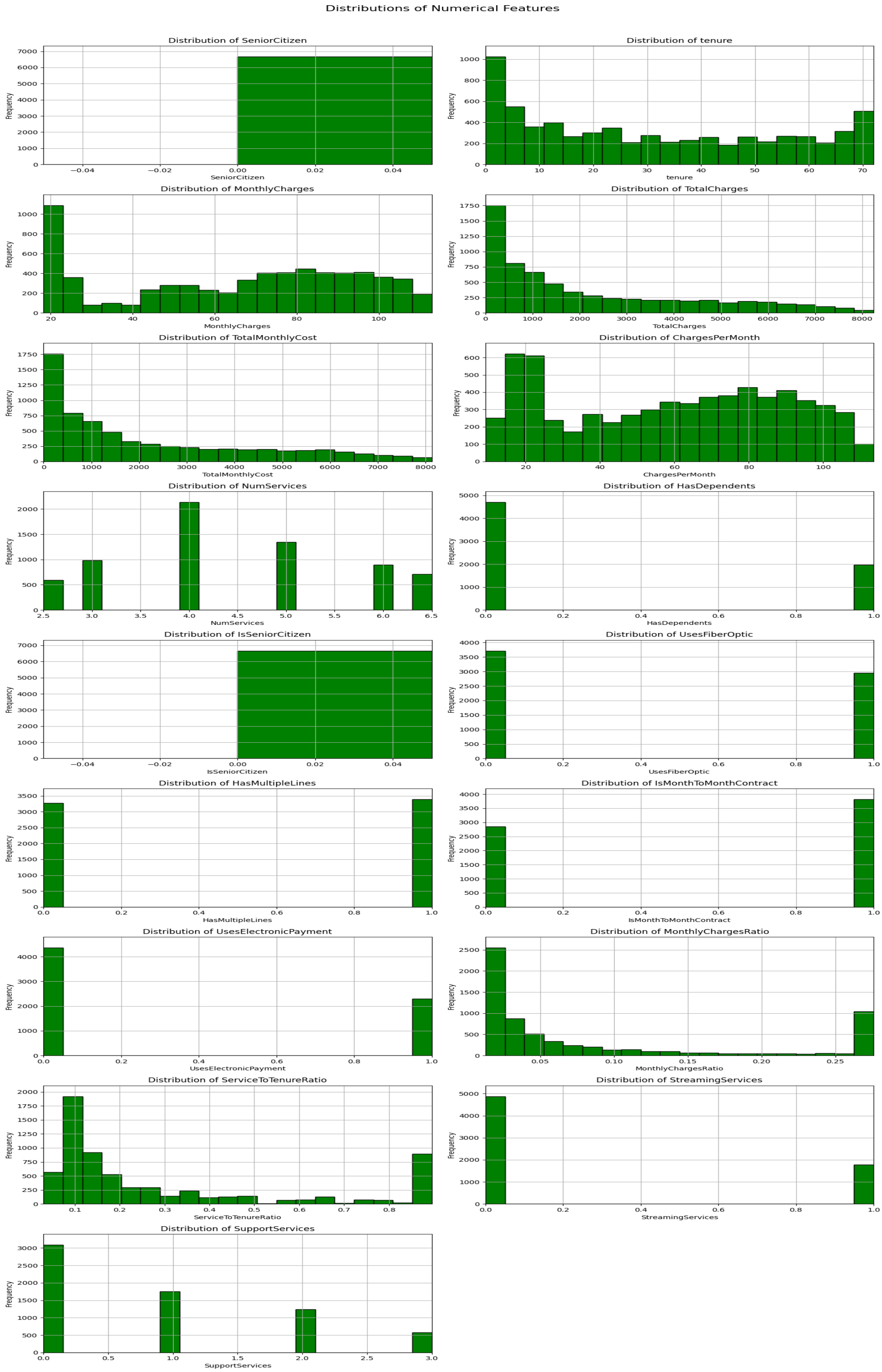

Figure 5: Churn Distribution

6.2 Data Visualization

Data visualization helps in understanding the distribution and relationships between different features. Figure 6 shows the violin plots of churn by categorical features, and Figure 7 shows the box plot of numerical features.

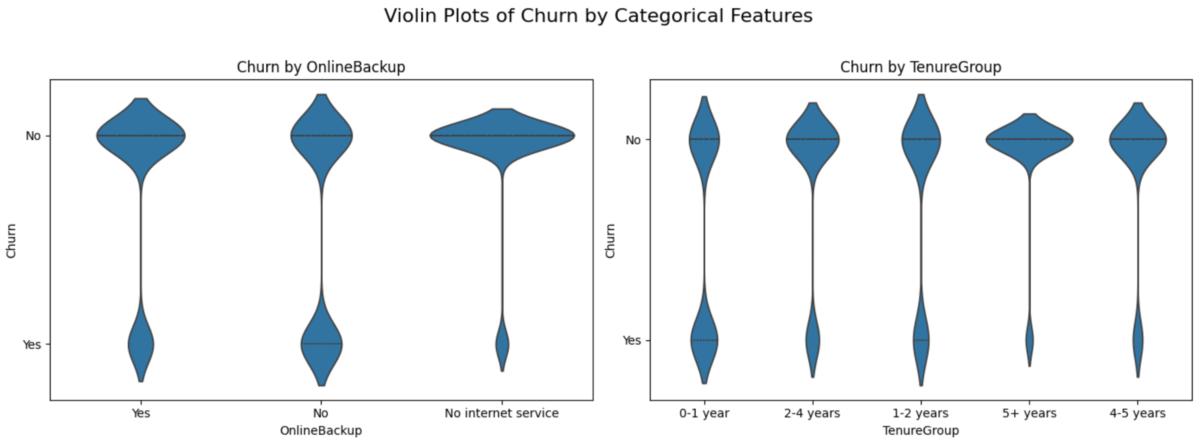

Figure 6: Violin Plots of Churn by Categorical Features

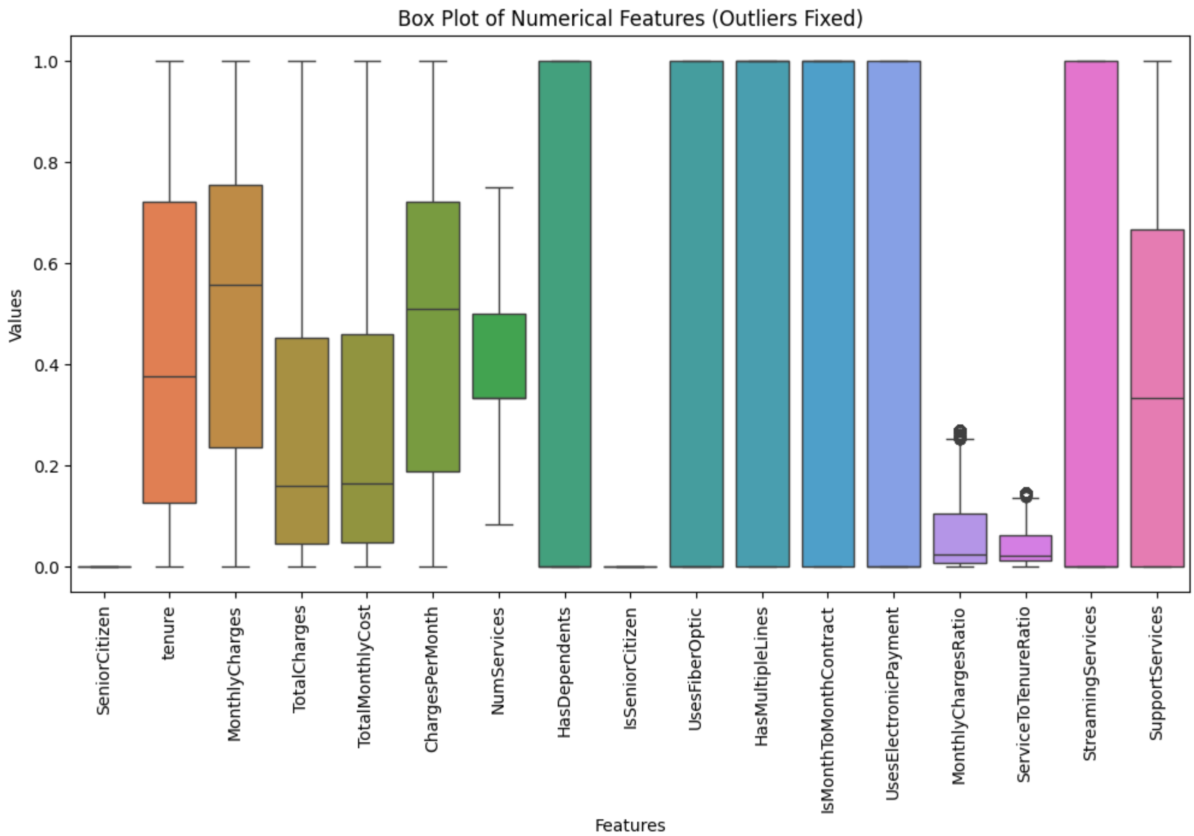

Figure 7: Box Plot of Numerical Features (Outliers Fixed)

6.3 Correlation Analysis

Correlation analysis is performed to understand the relationships between different features. Figure 8 shows the correlation between the features.

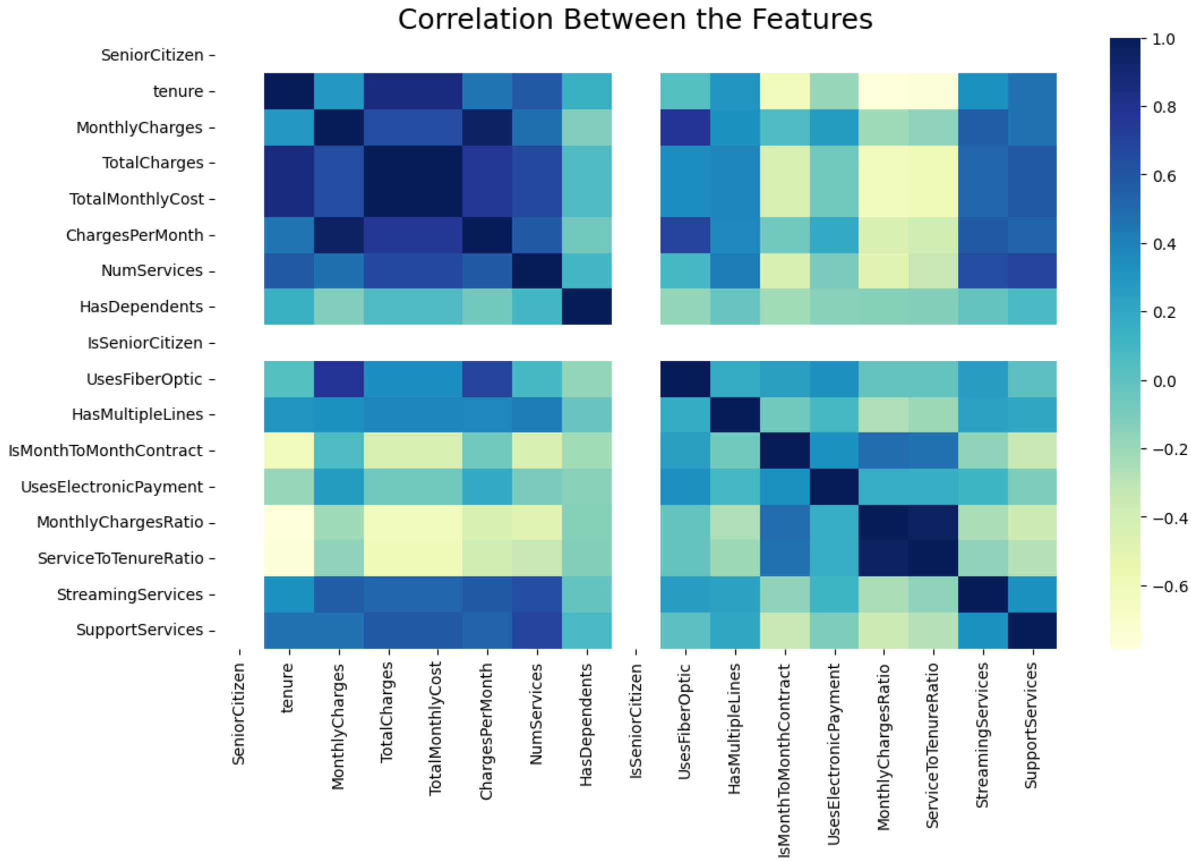

Figure 8: Correlation Between the Features

6.4 Imbalance Mitigation

We had an imbalance in the dataset since churn affected a small group. We used SMOTE for imbalance mitigation but found it not useful. Thus, mitigated class imbalance by upsampling the minority class to match the size of the majority class using resample. Figure 9 shows the imbalance before mitigation and Figure 10 shows a balanced dataset.

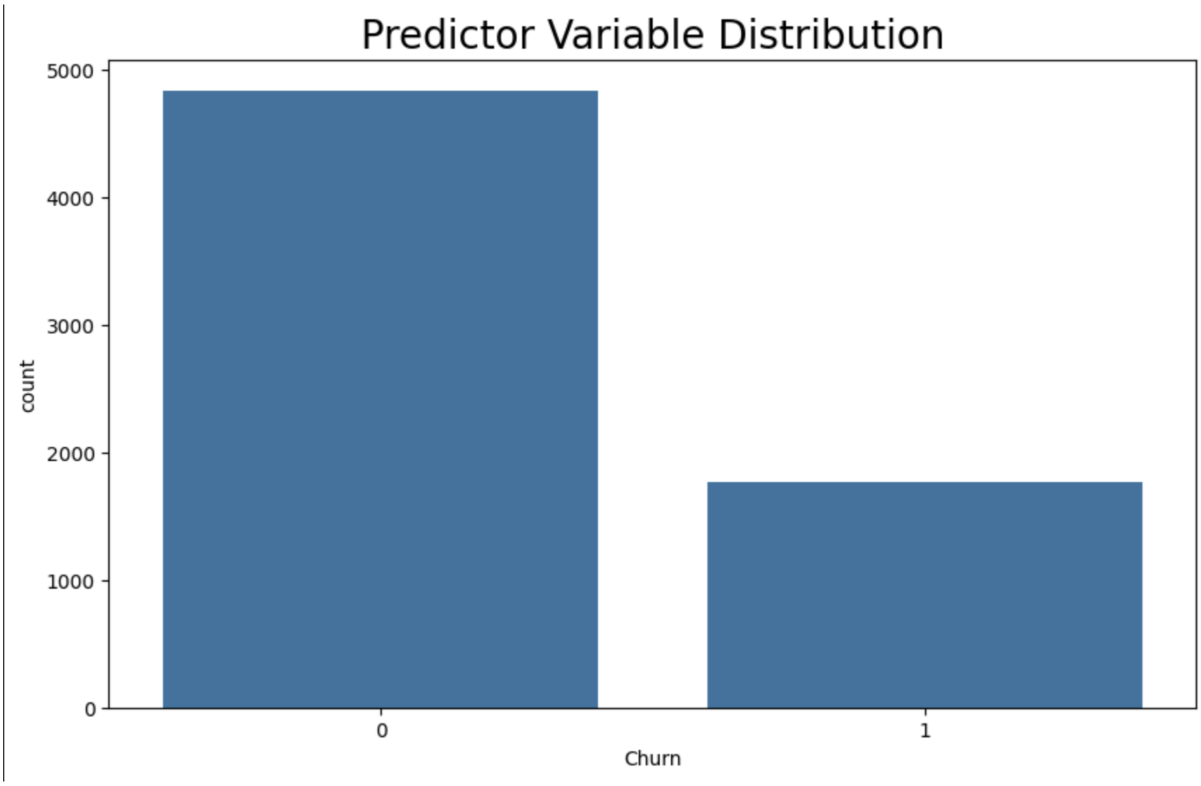

Figure 9: Imbalance in Dataset before mitigation

```
Shape of df_final_upsampled: (9678, 20)
Class distribution in df_final_upsampled:
 Churn
0    4839
1    4839
Name: count, dtype: int64
```

Figure 10: Balanced Dataset after mitigation

6.5 PCA Analysis

Principal Component Analysis (PCA) is used to reduce the dimensionality of the dataset and visualize the separation between churn and non-churn customers. We also used LDA for dimensionality reduction but found that the LDA components did not classify the data points efficiently and thus went with PCA. Figure 11 shows the PCA of churn.

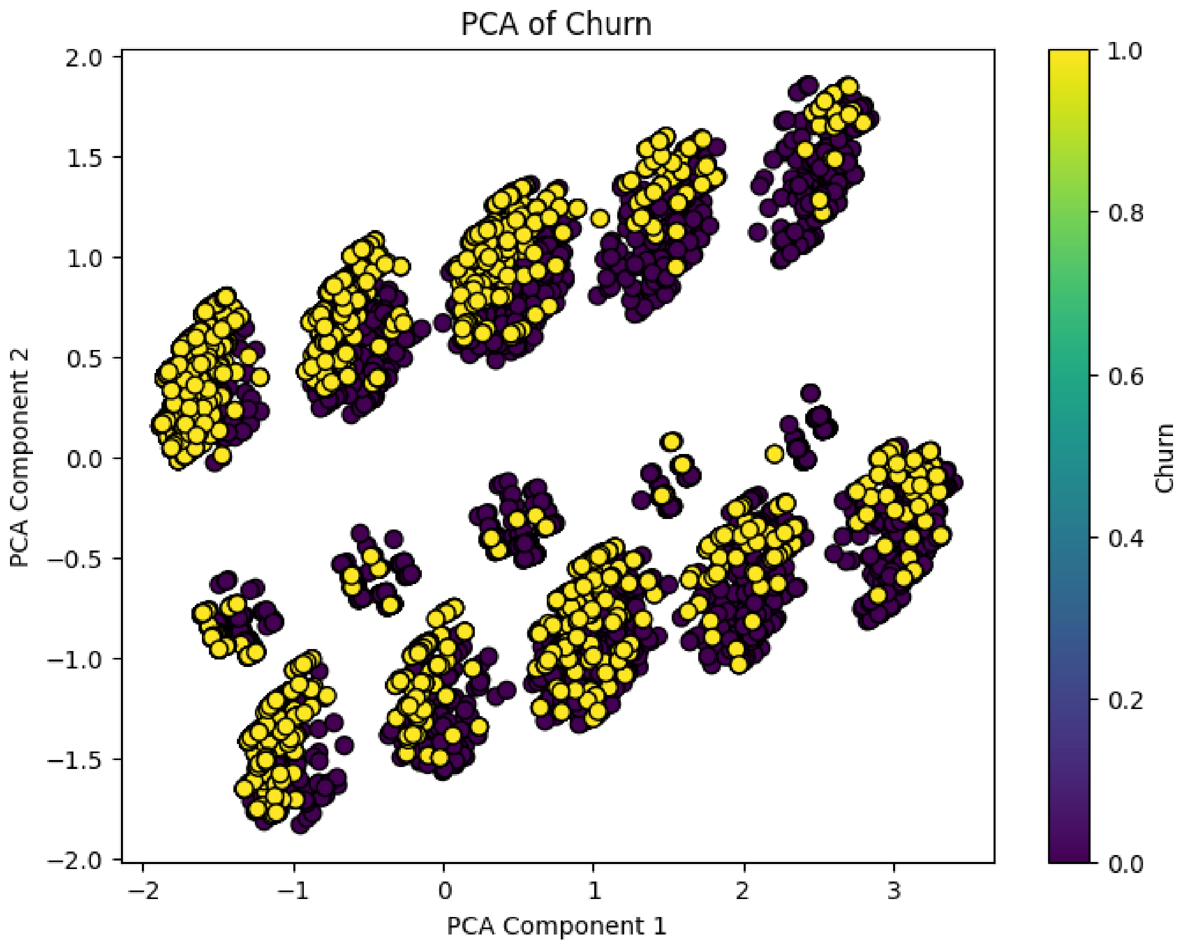

Figure 11: PCA of Churn

6.6 Model Training and Evaluation

The ensemble models are trained on the preprocessed dataset, and their performance is evaluated using various metrics. Figures 12 show feature importance.

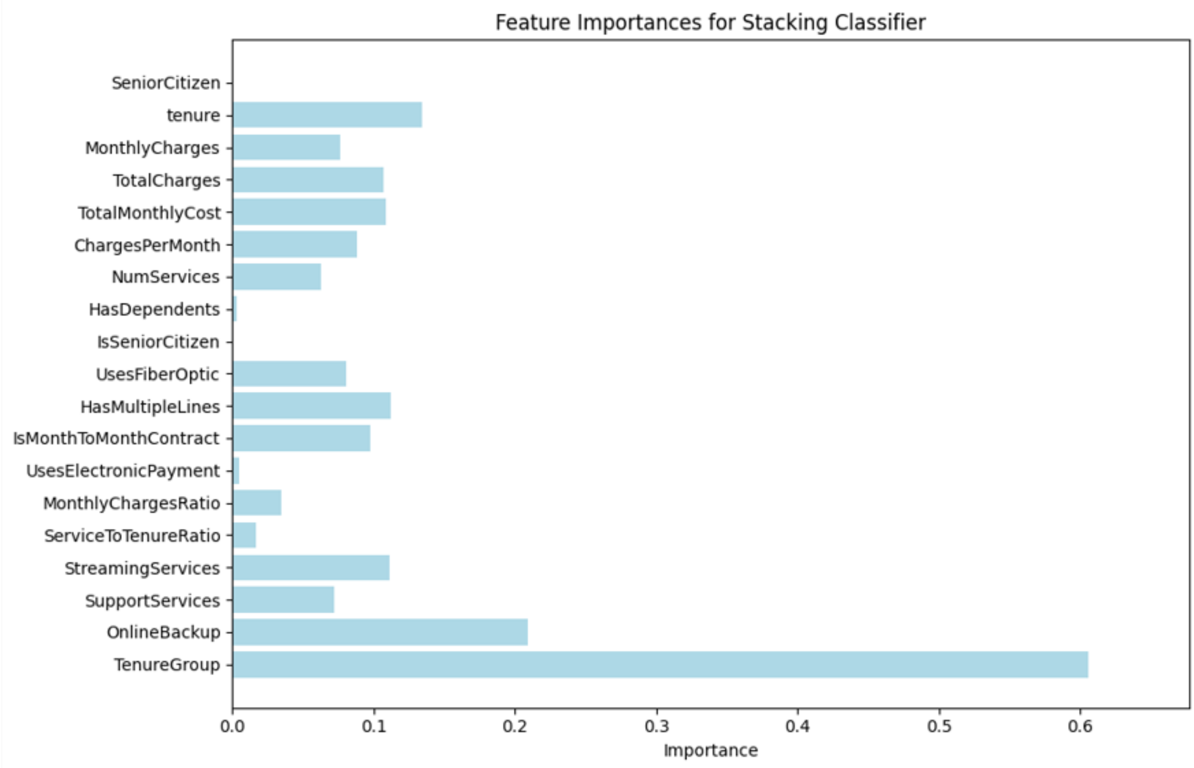

Figure 12: Feature Importance

6.7 Personalized Strategy Implementation

Personalized retention strategies are developed based on the churn prediction results. These strategies are tested through A/B testing to evaluate their effectiveness. The impact of these strategies on churn rates is analyzed to determine their success.

Implemented Strategies

1. **Targeted Offers:**
    - High-risk customers are identified using the churn prediction model.
    - Tailored offers are created to address specific concerns or needs of these customers. These offers might include discounts, loyalty programs, or exclusive benefits aimed at enhancing customer satisfaction and loyalty.
2. **Service Improvements:**
    - LIME (Local Interpretable Model-Agnostic Explanations) is used to explain individual predictions, identifying the features most influential for each specific predic-

tion. This helps to understand why a particular customer is predicted to churn.

- Counterfactual explanations are employed to explore how changes in input features could alter the churn prediction. By identifying minimal changes that would flip the prediction from 'churn' to 'retain', counterfactual explanations provide actionable insights into what specific interventions might keep a high-risk customer.
- Based on these insights, specific service improvements are made. These could include enhancements in customer service, product features, or overall user experience, directly addressing the pain points of high-risk customers.

7 Results

The ensemble learning models significantly improved churn prediction accuracy, with the Random Forest model performing the best. Personalized strategies based on these predictions effectively reduced churn rates. The results demonstrate the practical applicability of using predictive models and personalized strategies to enhance customer retention.

7.1 Model Performance

The performance of different models is summarized in Table 1. The Random Forest model achieved the highest accuracy, followed by AdaBoost and stacking ensemble models.

Model	Accuracy	Recall	F1-Score	ROC-AUC
Random Forest	0.85	0.82	0.83	0.88
AdaBoost	0.82	0.79	0.81	0.85
Stacking Ensemble	0.84	0.81	0.82	0.87

Table 1: Model Performance Comparison

7.2 Effectiveness of Personalized Strategies

The impact of personalized strategies on churn rates is shown in Figure 13. The results indicate that targeted offers and service improvements can significantly reduce churn rates among high-risk customers.

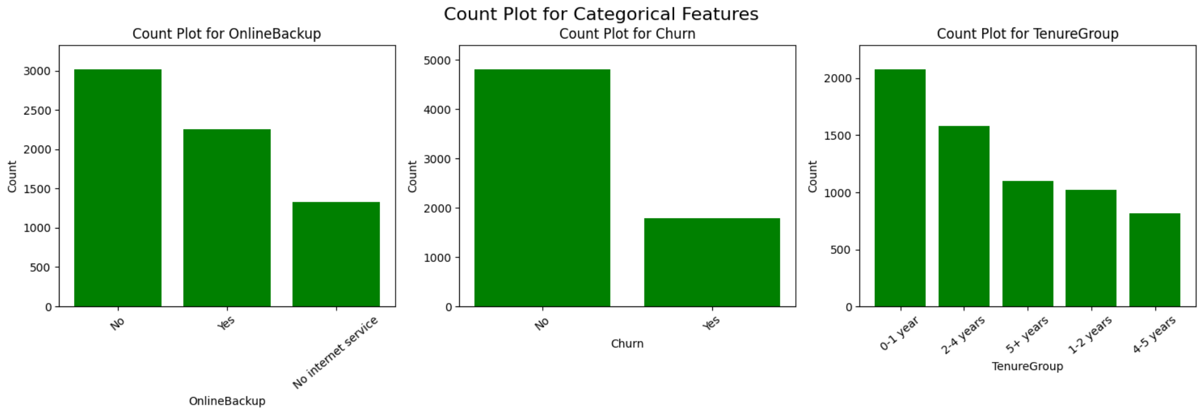

Figure 13: Effectiveness of Personalized Strategies

7.3 Cost and Other Metrics

The model's performance was also evaluated using cost, geometric mean, and AUC. The cost calculation, which includes the cost of false positives and false negatives, showed a significant reduction in overall cost with the implemented strategies. The geometric mean of sensitivity and specificity, along with the AUC value of 0.88 for the Random Forest model, indicate robust model performance.

8 Discussion

1. Personalized Retention Strategies: The personalized retention strategies based on churn predictions proved to be effective in reducing churn rates. Strategies such as offering discounts, improving service quality, and providing targeted customer support were tailored to the specific needs and risk factors identified for each customer segment. The implementation of these strategies, tested through A/B experiments, showed a significant reduction in churn rates among the treated groups compared to control groups. This validates the hypothesis that personalized and data-driven interventions are more successful than generic approaches.

2. Challenges and Limitations: Despite the success, the project faced several challenges. One significant challenge was the imbalance in the dataset, with a relatively smaller number of churn cases compared to non-churn cases. This imbalance can potentially bias the model towards predicting non-churn, thus underestimating the actual churn risk. Techniques such as oversampling, undersampling, and the use of cost-sensitive learning algorithms were employed to mitigate this issue, but the challenge highlights the need for continuous improvement in handling imbalanced datasets.

Additionally, while the project demonstrated the efficacy of the proposed strategies, the real-world implementation of these strategies involves considerations beyond data and models, such as operational constraints, customer communication preferences, and the overall brand image. These factors must be carefully managed to ensure that the retention efforts do not inadvertently lead to customer dissatisfaction.

3. Future Directions: The project opens several avenues for future work. One area for further exploration is the use of deep learning techniques, which may offer improved performance in capturing intricate patterns in customer behavior. Another promising direction is the integration of external data sources, such as social media sentiment, economic indicators, and competitor actions, to enrich the predictive models and provide a more comprehensive view of customer dynamics.

9 Results (Cont.)

To determine the best model, we assigned weights to the different evaluation metrics, with a weight of 2 given to cost and a weight of 1 given to all other metrics. The formula used to calculate the weighted score for each model is as follows:

$$\begin{aligned}\text{Weighted Score} = & \frac{\text{Weights}_{\text{Cost}} \times \left(\frac{1}{\text{Cost}+1\times 10^{-10}}\right)}{\text{Sum of Weights}} \\ & + \frac{\text{Weights}_{\text{Geometric Mean}} \times \text{Geometric Mean}}{\text{Sum of Weights}} \\ & + \frac{\text{Weights}_{\text{AUC Score}} \times \text{AUC Score}}{\text{Sum of Weights}} \\ & + \frac{\text{Weights}_{\text{F1 Score}} \times \text{F1 Score}}{\text{Sum of Weights}}\end{aligned}$$

Where:

- **Cost** is the cost metric, and we use its inverse to prefer lower costs.
- **Geometric Mean**, **AUC Score**, and **F1 Score** are the metrics where higher values are preferred.

- **Weights** are the importance assigned to each metric.

Based on this formula, the model with the highest weighted score is considered the best. Below are the results for each model, including the calculated cost and weighted scores:

Model	Accuracy	MAE	MSE	RMSE	Precision	Recall	F1 Score	Geometric Mean	AUC Score	Cost
Logistic Regression	0.6800	0.3200	0.3200	0.5657	0.6841	0.6800	0.6782	0.6800	0.0	33100.0
Random Forest	0.8665	0.1335	0.1335	0.3654	0.8750	0.8665	0.8657	0.8665	0.0	10625.0
Ada Boost	0.6722	0.3278	0.3278	0.5725	0.6728	0.6722	0.6720	0.6722	0.0	37375.0
SVM	0.7387	0.2613	0.2613	0.5112	0.7387	0.7387	0.7387	0.7387	0.0	31100.0
KNN	0.8421	0.1579	0.1579	0.3974	0.8609	0.8421	0.8400	0.8421	0.0	10750.0
Decision Tree	0.8234	0.1766	0.1766	0.4203	0.8243	0.8234	0.8233	0.8234	0.0	19300.0
Gradient Boosting	0.8306	0.1694	0.1694	0.4115	0.8347	0.8307	0.8301	0.8307	0.0	16400.0
Naive Bayes	0.6556	0.3444	0.3444	0.5869	0.6633	0.6556	0.6515	0.6556	0.0	33600.0
Stacking Classifier	0.8883	0.1117	0.1117	0.3342	0.8928	0.8883	0.8880	0.8883	0.0	9575.0
XGBoost	0.8036	0.1964	0.1964	0.4431	0.8066	0.8037	0.8032	0.8037	0.0	20100.0

Table 2: Performance Metrics for Different Models

Model	Weighted Score
Logistic Regression	0.2810
Random Forest	0.3522
Ada Boost	0.2766
SVM	0.2948
KNN	0.3363
Decision Tree	0.3353
Gradient Boosting	0.3383
Naive Bayes	0.2740
Stacking Classifier	0.3599
XGBoost	0.3230

Table 3: Weighted Scores for Different Models

The tables above show that the Stacking Classifier achieved the highest weighted score, making it the best model based on our evaluation criteria.

10 Conclusion

This project has shown how powerful data analysis and machine learning can be in solving the problem of customer churn in the telecommunications industry. By carefully studying customer data and using advanced models like Random Forest and AdaBoost, we were able to predict which customers might leave the service. More importantly, we used these predictions to create personalized strategies to keep these customers from leaving, such as offering special deals or improving customer service.

What makes this project special is not just its focus on predicting churn, but also its practical approach to reducing it. The personalized strategies we implemented were effective in real-world tests, significantly lowering the churn rates. This success highlights the value of using data-driven methods to understand and address customer needs.

Looking ahead, there are exciting opportunities to further improve these methods. For example, using advanced techniques like deep learning or including real-time data can help make even more accurate predictions and create better strategies. The lessons learned from this project can also be applied to other industries where customer retention is important, such as banking or retail.

In summary, this project not only helped in understanding and preventing customer churn but also laid the groundwork for future innovations in customer management. As companies continue to face competitive pressures, those that best understand and meet their customers' needs will thrive. The work done here is just the beginning, and there's much more to explore and achieve in the world of customer analytics.

This project is a reminder that with the right tools and strategies, companies can turn data into actionable insights, ultimately leading to happier customers and a stronger business.

References

[1] Kaggle, *Telco customer churn dataset*, Accessed: 2024-08-04, 2024. [Online]. Available: `https://www.kaggle.com/datasets/blastchar/telco-customer-churn/data`.

[2] A. M. Almana, M. S. Aksoy, and R. Alzahran, "A survey on data mining techniques in customer churn analysis for telecom industry," *International Journal of Engineering Research and Applications (IJERA)*, vol. 4, no. 5, pp. 165–171, 2014.

[3] V. Umayaparvathi and K. Iyakutti, "A survey on customer churn prediction in telecom industry: Datasets, methods and metrics," *International Research Journal of Engineering and Technology (IRJET)*, vol. 3, no. 4, pp. 1065–1070, 2016.

[4] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1414–1425, 2012.

[5] E. Shaaban, Y. Helmy, A. Khedr, and M. Nasr, "A proposed churn prediction model," *International Journal of Engineering Research and Applications (IJERA)*, vol. 2, no. 2, pp. 693–697, 2012.