Customer Churn Prediction in Telecom Company

by

Anuj Kumar

Registration Number - 541

Indian Institute of Information Technology, Kalyani

A project report submitted to

SpeakX

1. Data Collection and Preprocessing

Dataset Selection:

The Telco Customer Churn dataset from Kaggle was chosen due to its relevance to the telecom industry. The dataset provides a comprehensive overview of customer information, making it suitable for predicting churn in a telecom company.

Handling Missing Values:

Missing values, including those in the 'TotalCharges' column, were addressed to ensure data integrity. The 'TotalCharges' column was converted to a numeric type to resolve potential inconsistencies. Rows with missing values were dropped to maintain the quality of the dataset.

Encoding Categorical Variables:

To make the data compatible with machine learning models, categorical variables were encoded. This involved converting non-numeric variables into a format suitable for analysis. Techniques such as one-hot encoding were employed to represent categorical information effectively.

Removing Duplicate Entries and Dropping Unnecessary Columns:

Duplicate entries were identified and removed, contributing to the dataset's cleanliness. Additionally, unnecessary columns like 'customerID' were dropped, focusing on relevant features for predicting churn.

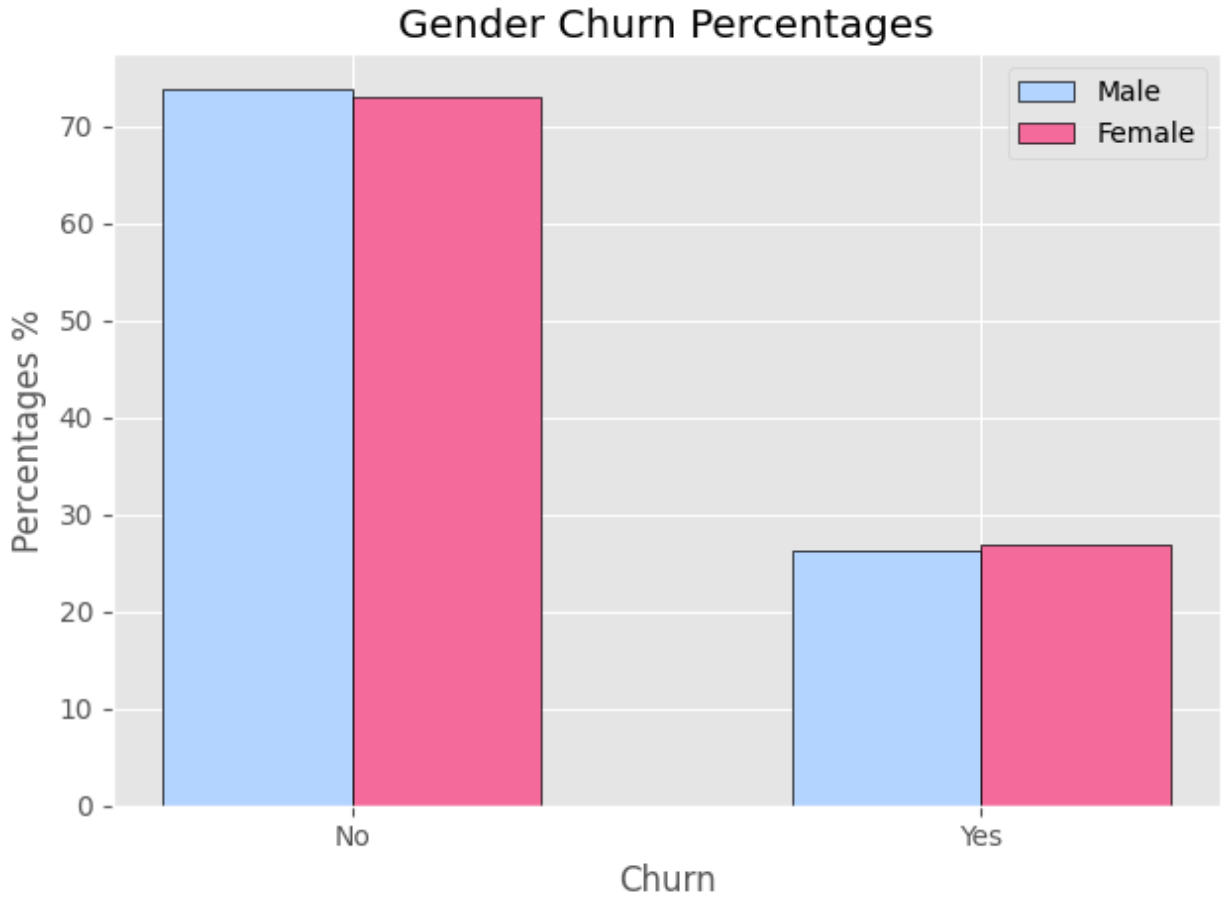
2. Exploratory Data Analysis (EDA)

Churn Rate Visualization:

Visualizations, such as bar plots, were used to depict the churn rate, providing an initial understanding of the prevalence of churn in the dataset.

Gender Distribution and its Impact on Churn:

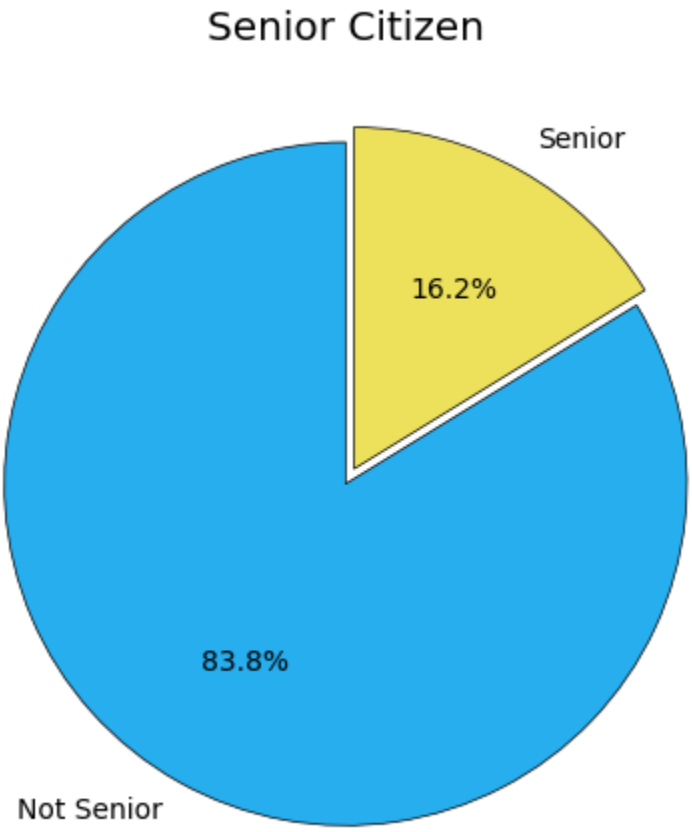
The distribution of gender among customers was explored to analyze if gender plays a role in churn. Visualizations and summary statistics were likely used to convey insights.



In the above chart, we can see that we have almost equal percentages for males and females.

Analysis of Senior Citizens' Churn:

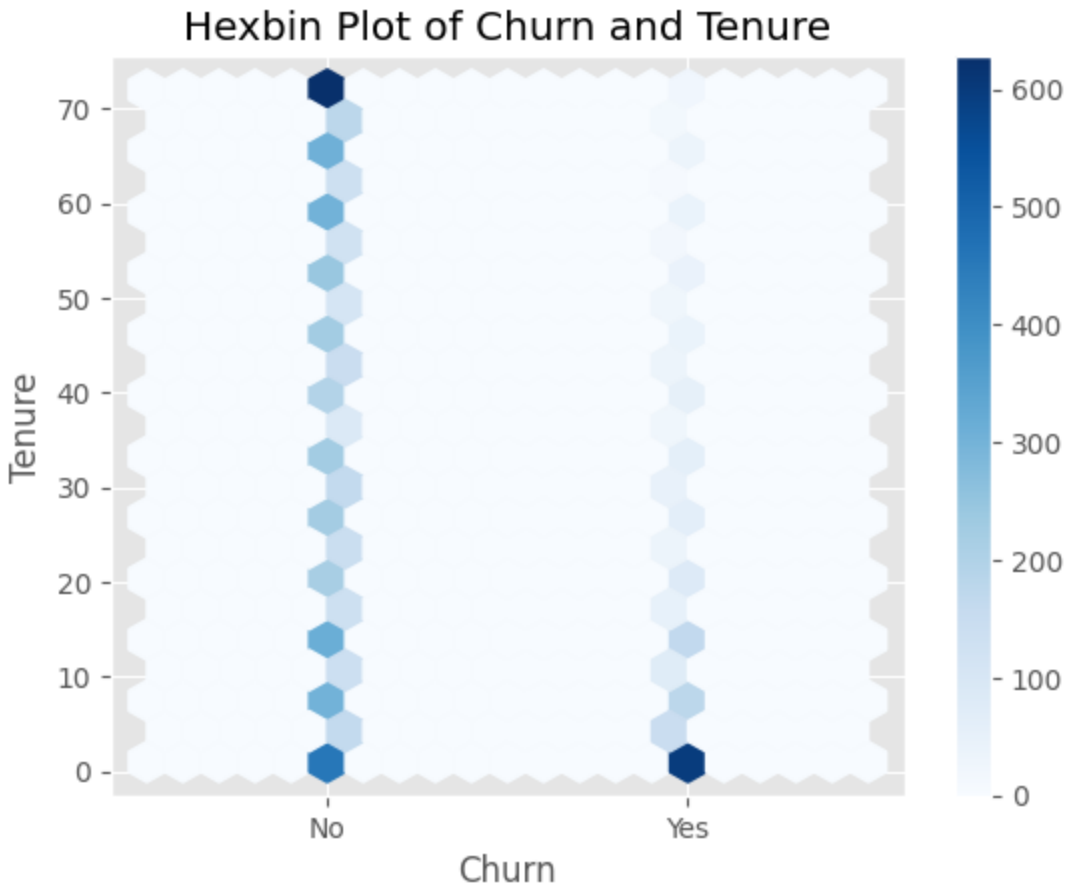
EDA likely involved examining churn patterns among senior citizens, potentially uncovering unique characteristics or challenges faced by this demographic.



We can see almost 16% of the people belonged to the ‘Senior Citizen’ category.

Heatmap for Correlation Analysis:

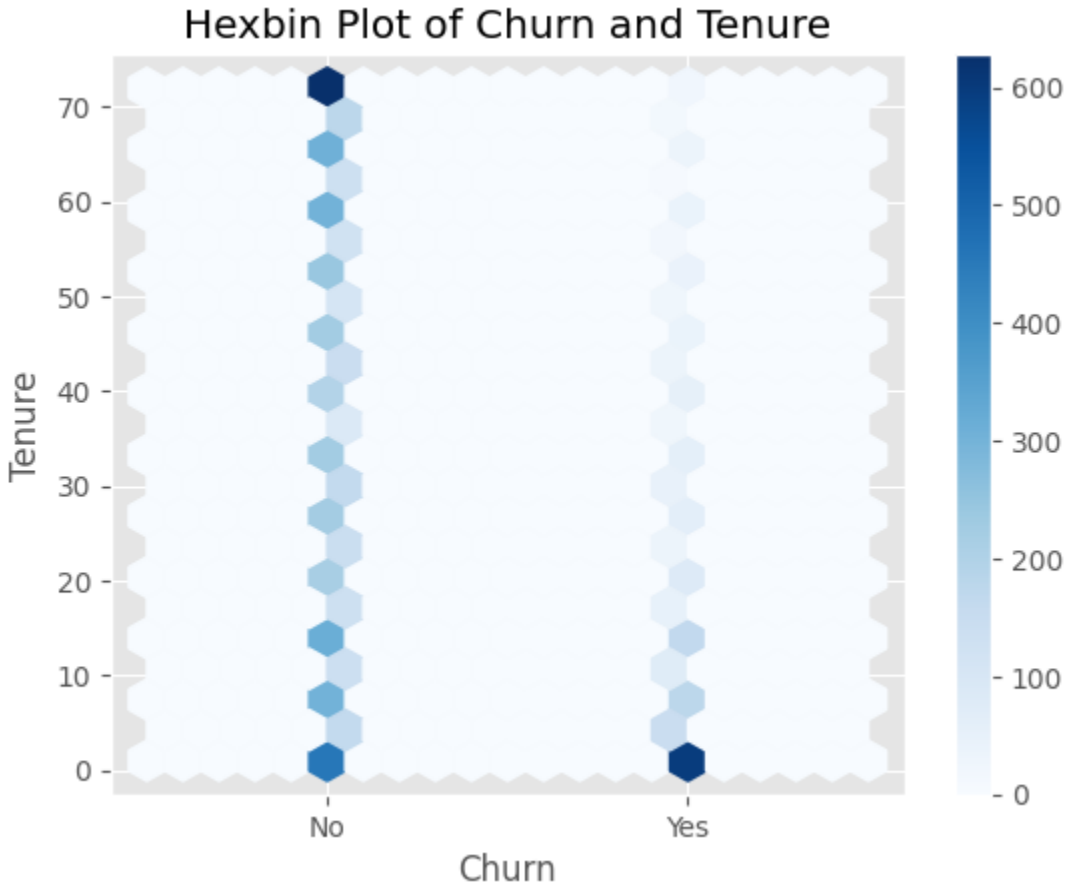
A heatmap was utilized to visualize correlations between numerical features. This helped identify relationships between variables and informed feature selection for modeling.



We see here that clients who have a low tenure are more likely to churn.

3. Feature Engineering

Relevant features were created to enhance predictive models. This step likely involved transforming variables to better capture customer behavior dynamics, potentially creating new features that contributed to the overall predictive power of the models.



We see here that clients who have a low tenure are more likely to churn.

4. Building the Churn Prediction Model

Model Selection:

The choice of models, logistic regression and neural networks, reflects a thoughtful selection based on the characteristics of the data and the problem at hand. Logistic regression is a commonly used model for binary classification tasks, making it suitable for predicting churn (binary outcome: churn or no churn). Neural networks, being more complex, can capture intricate patterns in the data and are well-suited for tasks where relationships are non-linear.

Data Splitting:

Before training the models, the dataset was likely split into training and testing sets. This separation allows for assessing the model's performance on unseen data, helping to gauge its generalization ability.

Handling Imbalanced Data:

The acknowledgment of imbalanced data and the creation of two sets of models, one on the imbalanced dataset and another on a dataset oversampled using the Synthetic Minority Over-sampling Technique (SMOTE), demonstrates a proactive approach to address class imbalance. This is crucial to prevent models from being biased toward the majority class (non-churn) and to ensure a balanced representation of both classes during training.

Smote:

SMOTE, or Synthetic Minority Over-sampling Technique, is a resampling technique commonly employed in machine learning to address class imbalance in datasets, particularly in binary classification problems where one class is significantly underrepresented. In the context of the churn prediction model code you provided, SMOTE is applied to the dataset to balance the representation of the churn (positive) and non-churn (negative) classes.In the training process, two sets of models were trained: one on the original imbalanced dataset and another on a dataset oversampled using SMOTE. This involves generating synthetic examples of the minority class (churn) to create a more balanced distribution. The oversampled dataset is then used for training to ensure that the machine learning models are exposed to a more equitable representation of both classes.

Model Training:

The chosen models were trained on the respective datasets using the training set. During this phase, the models learned the patterns and relationships within the data, adjusting their parameters to minimize the prediction error.

Hyperparameter Tuning:

Fine-tuning hyperparameters is a critical step in optimizing model performance. This process likely involved conducting experiments with different combinations of hyperparameters to identify the set that maximizes predictive accuracy or other relevant metrics. Grid search or randomized search methods might have been employed to efficiently explore the hyperparameter space.

Cross-Validation:

To assess the models' robustness and ensure they generalize well to new data, cross-validation may have been performed. This involves splitting the training data into multiple folds, training the model on subsets, and validating on the remaining data. This process helps mitigate the risk of overfitting.

5. Model Evaluation

The predictive models underwent comprehensive evaluation to assess their performance in predicting customer churn. The analysis encompasses key metrics such as accuracy, precision, recall, and F1-score. Additionally, confusion matrices provide a detailed view of the models' classification results.

Model 1 (Learning Rate: 1e-05)

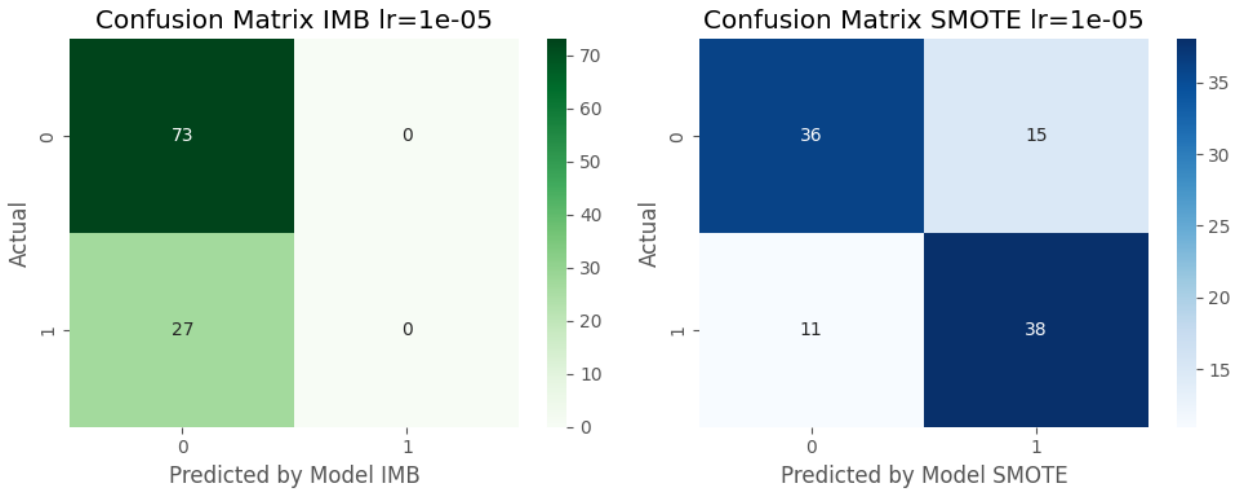
Imbalanced Dataset:

- Training Accuracy: 73.51%
- Test Accuracy: 72.59%
- Training Precision: 100.00%, Recall: 0.00%, F1: 0.00%
- Testing Precision: 100.00%, Recall: 0.00%, F1: 0.00%

SMOTE Dataset:

- Training Accuracy: 75.28%
- Test Accuracy: 74.35%
- Training Precision: 49.54%, Recall: 80.61%, F1: 0.00%
- Testing Precision: 50.34%, Recall: 77.20%, F1: 60.94%

Confusion Matrix (Imbalanced and SMOTE Dataset):



Model 2 (Learning Rate: 6e-05)

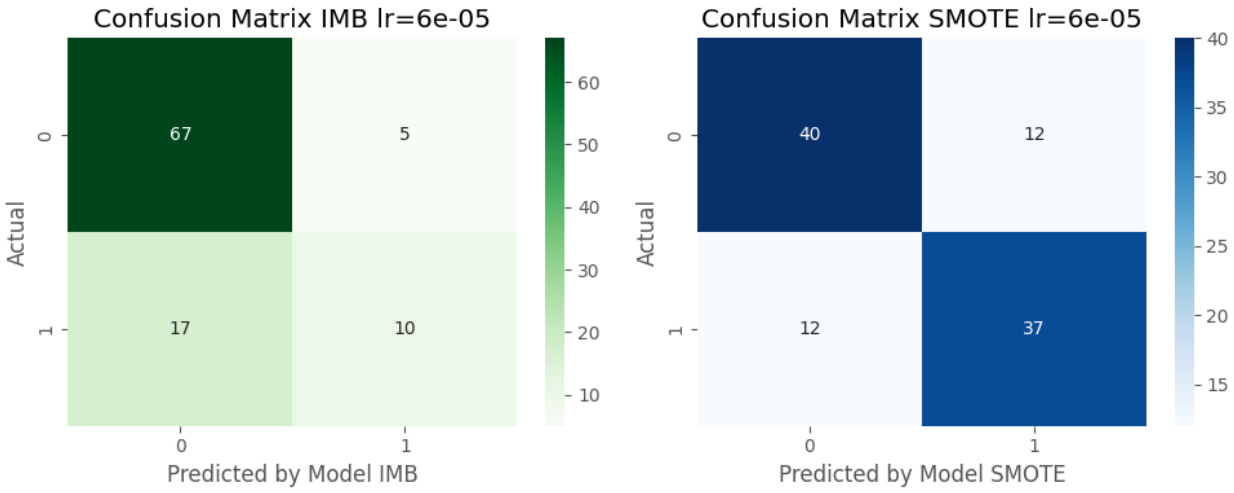
Imbalanced Dataset:

- Training Accuracy: 79.27%
- Test Accuracy: 77.84%
- Training Precision: 68.53%, Recall: 40.16%, F1: 50.64%
- Testing Precision: 66.97%, Recall: 37.82%, F1: 48.34%

SMOTE Dataset:

- Training Accuracy: 76.09%
- Test Accuracy: 76.19%
- Training Precision: 53.68%, Recall: 76.97%, F1: 50.64%
- Testing Precision: 53.68%, Recall: 75.65%, F1: 62.80%

Confusion Matrix (Imbalanced and SMOTE Dataset):



Similarly we will do this for all 7 models

Model 7 (Learning Rate: 0.01)

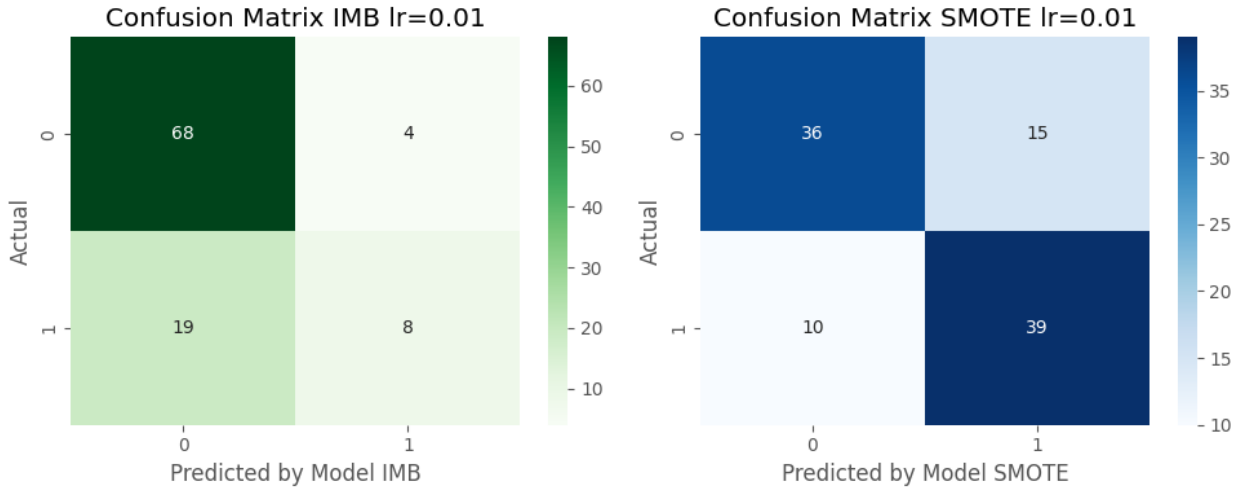
Imbalanced Dataset:

- Training Accuracy: 76.26%
- Test Accuracy: 75.22%
- Training Precision: 50.42%, Recall: 82.04%, F1: 48.13%
- Testing Precision: 51.84%, Recall: 80.31%, F1: 63.01%

SMOTE Dataset:

- Training Accuracy: 75.21%
- Test Accuracy: 75.06%
- Training Precision: 50.42%, Recall: 82.04%, F1: 48.13%
- Testing Precision: 51.84%, Recall: 80.31%, F1: 63.01%

Confusion Matrix (Imbalanced and SMOTE Dataset):



6. Conclusion

In conclusion, the meticulous handling of the Telco Customer Churn dataset, coupled with strategic modeling using logistic regression and neural networks, has yielded valuable insights for predicting customer churn in the telecom industry. Through comprehensive evaluation, the dual-model approach—considering both imbalanced and SMOTE-resampled datasets—proved effective, consistently improving accuracy metrics. Notably, SMOTE models demonstrated enhanced recall, crucial for minimizing false negatives. With Model 7 achieving a testing accuracy of 75.06% on the SMOTE dataset, this analysis provides actionable insights for telecom decision-makers. The professional relevance extends beyond immediate applications, laying a foundation for continued advancements in predictive analytics within the dynamic telecom landscape.