Table of Contents

Churn Prediction Project Report

1. Introduction

1.1 Objective: Developing an Effective Churn Prediction Model

The primary objective of this project is to construct a robust and precise churn prediction model for a telecommunications company. The aim is to empower the company with the capability to implement proactive customer retention strategies. The significance of this objective lies in the dynamic and competitive landscape of the telecommunications industry, where customer churn can have substantial financial implications. By accurately identifying customers at risk of churning, the company can strategically allocate resources, tailor retention efforts, and ultimately enhance customer satisfaction and loyalty.

1.2 Dataset: Kaggle Telco Customer Churn Dataset

1.2.1 Dataset Origin and Composition

The dataset employed for this project is sourced from Kaggle and is known as the "Telco Customer Churn" dataset. This dataset encapsulates a rich array of customer-centric attributes, service-related information, and churn indicators. The dataset's composition is instrumental in capturing the multifaceted nature of customer behavior within the telecommunications domain. Key attributes include demographic details, types of services subscribed, contract specifics, and a binary churn indicator, making it a comprehensive foundation for predictive modeling.

1.2.2 Dataset Significance

- **Demographic Attributes:** Capture customer-specific details such as gender and senior citizen status.

- **Service Information:** Encompasses a variety of services, including internet service, online security, and tech support.
- **Contract Details:** Provides insights into contract duration, payment methods, and billing preferences.
- **Churn Indicator:** The binary churn indicator is the focal point of the analysis, serving as the target variable for the predictive model.

1.2.3 Diverse Dimensions

The dataset's diversity in terms of dimensions allows for a nuanced exploration of customer behavior, contributing to a comprehensive and accurate churn prediction model. This depth is crucial for understanding the intricate interplay of various factors that contribute to customer attrition.

In essence, the choice of the Kaggle Telco Customer Churn dataset aligns with the project's objective, providing a multifaceted lens through which to explore and model customer churn within the telecommunications industry. This depth ensures that the predictive model is not only accurate but also capable of capturing the underlying complexities of customer dynamics that influence churn.

2. Data Collection and Preprocessing

2.1 Data Source

The dataset, obtained from Kaggle, provides a comprehensive view of customer-related variables. It encompasses a range of information crucial for our churn prediction model, including demographic details, services subscribed, contract specifics, and the churn status of customers.

2.2 Preprocessing

2.2.1 Addressing Missing Values through Imputation

Missing values in the dataset were handled meticulously through imputation techniques. For numerical features, appropriate statistical measures such as mean, median, or mode were employed based on the nature of the variable. Categorical features underwent imputation with the most frequent category.

2.2.2 One-Hot Encoding and Label Encoding for Categorical Variables

To effectively utilize categorical data in our model, a combination of one-hot encoding and label encoding techniques was applied. One-hot encoding was employed for categorical features with no inherent ordinal relationship, creating binary columns for each category. Label encoding was selectively used when the categorical data exhibited ordinal relationships.

3. Exploratory Data Analysis (EDA)

3.1 Key Findings

3.1.1 Churn Rate Observation

The analysis reveals a noteworthy churn rate of approximately 26.54%, indicating that more than a quarter of the customers in the dataset experienced churn. Understanding this baseline churn rate provides context for evaluating the impact of different factors on customer retention.

3.1.2 Contract Terms and Churn Correlation

Upon investigation, a strong correlation emerges between shorter contract terms and higher churn rates. Customers with month-to-month contracts are more likely to churn compared to those with longer-term contracts. This finding suggests that there may be an opportunity to improve customer retention by incentivizing longer-term commitments.

3.1.3 Fiber Optic Internet Service and Churn

The analysis identifies a significant correlation between the use of fiber optic internet service and higher churn rates. This suggests that customers with fiber optic internet service are more prone to churning. Further exploration into the reasons behind this correlation is warranted. It could be related to service quality, pricing, or other factors that impact customer satisfaction.

3.1.4 Impact of Monthly and Total Charges

Monthly charges and total charges emerge as influential factors in customer churn. Higher monthly charges and total charges are associated with increased churn. Understanding the threshold at which these charges significantly impact churn is crucial for pricing strategies and customer retention efforts.

3.1.5 Effect of Online Security and Tech Support

Interestingly, the presence of online security and tech support features is associated with reduced churn. Customers with these services are less likely to churn. This finding underscores the importance of these features in enhancing customer satisfaction and loyalty.

3.2 Visualizations

3.2.1 Bar Charts

Utilized bar charts to visualize the distribution of categorical variables such as contract types and internet services. These visualizations provide a clear representation of the frequency of each category, aiding in the identification of patterns and trends.

3.2.2 Histograms

Histograms were employed to illustrate the distribution of continuous variables like monthly charges and total charges. Understanding the distribution of charges helps in identifying common price points and their association with customer churn.

3.2.3 Correlation Matrices

Correlation matrices were generated to quantify the relationships between variables. These matrices visually represent the strength and direction of correlations, helping identify potential multicollinearity and key drivers of churn.

3.2.4 Distribution Analysis

Explored the distribution of contract types, internet services, and monthly charges. This involved creating visualizations such as pie charts for categorical variables and kernel density plots for continuous variables. These visualizations contribute to a more comprehensive understanding of the dataset's composition.

3.3 Key Takeaways

The in-depth exploration of key findings and visualizations in the EDA phase provides valuable insights into the factors influencing

customer churn. These insights lay the foundation for informed decision-making in subsequent stages of the project, guiding feature engineering and model development.

4. Feature Engineering

4.1 Feature Creation: 'Family' Feature

In the realm of feature engineering, the creation of the 'Family' feature represents a deliberate effort to capture nuanced dynamics that might influence customer churn. This feature is constructed based on the presence or absence of a partner and dependents.

4.1.1 Motivation

The rationale behind introducing the 'Family' feature stems from the acknowledgment that customer behaviors and decisions are often intertwined with family considerations. The hypothesis posits that customers with familial ties may exhibit different churn patterns compared to those without such connections.

4.1.2 Methodology

The 'Family' feature is a binary indicator, taking the value of 1 if the customer has a partner or dependents, and 0 otherwise. This binary representation encapsulates the simplicity of family structure while retaining its insightful nature.

4.1.3 Potential Impact on Churn Prediction

- **Family Stability:** Customers with partners or dependents might demonstrate greater stability, potentially reducing the likelihood of churn.
- **Responsiveness to Services:** Family-oriented customers may have different expectations and requirements regarding the services offered, influencing their decision to churn.

4.1.4 Considerations and Limitations

- **Assumption of Family Influence:** The feature assumes that family dynamics play a significant role in churn decisions. This assumption may not hold true for all customer segments.
- **Binary Representation:** The simplicity of the binary representation may overlook nuances in family structures.

Future iterations might explore more detailed family-related features.

4.1.5 Visualization

Visualizations, such as bar charts illustrating the distribution of churn rates among customers with and without the 'Family' feature, can provide a deeper understanding of its impact.

4.1.6 Future Iterations

Continued refinement and exploration of the 'Family' feature's impact on churn prediction, potentially considering different representations or aggregations of family-related attributes, remain avenues for future development.

In summary, the introduction of the 'Family' feature represents a nuanced approach to feature engineering, aiming to capture the intricate dynamics associated with familial relationships and their potential influence on customer churn decisions. Further analysis and refinement will shed light on the true impact of this feature in the context of the churn prediction model.

5. Churn Prediction Model

5.1 Model Selection

In the pursuit of developing a robust churn prediction model, three distinct algorithms were chosen to offer diverse perspectives on the dataset.

5.1.1Random Forest

- **Introduction:** Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes as the prediction.
    - **Key Features:**
    - Ensemble of Trees: Reduces overfitting and increases accuracy.
    - Feature Importance: Provides insights into influential features.
- **Considerations:** Random Forest is resilient to outliers and noise, offering a robust approach to classification tasks.

5.2 Training and Evaluation Metrics

The models were trained using preprocessing pipelines to ensure consistency and repeatability. Hyperparameter tuning was performed utilizing GridSearchCV to find the optimal combination of parameters for each algorithm.

5.2.1 Random Forest

5.2.1.1 Training Process

- Applied preprocessing steps, including one-hot encoding and standardization.
- Fitted the logistic regression model to the training data.

5.2.1.2 Evaluation Metrics

- **Accuracy:** 81.76%

- **Precision:** 67.79%
- **Recall:** 59.25%
- **F1 Score:** 63.23%

Confusion Matrix

[[931, 105],

[152, 221]]

- **Interpretation:** The confusion matrix details True Positives, True Negatives, False Positives, and False Negatives, providing a nuanced understanding of the model's performance.

6. Challenges Faced

6.1 Imbalanced Data

6.1.1 Problem Statement

The dataset exhibited a notable class imbalance with a higher prevalence of non-churn instances compared to churn instances. This imbalance raised concerns about the potential bias of the model towards the majority class and the necessity to devise strategies for improved handling.

6.1.2 Approach

6.1.2.1 Oversampling Techniques

Implemented oversampling techniques, such as Synthetic Minority Over-sampling Technique (SMOTE) and Random Over-sampling, to balance class distribution. These techniques involve generating synthetic instances of the minority class to mitigate the impact of class imbalance.

6.1.2.2 Alternative Evaluation Metrics

Explored alternative evaluation metrics beyond traditional accuracy to better assess model performance. Precision, recall, and F1 score became focal points, as they provide insights into the model's ability to correctly identify churn instances, which is particularly crucial in imbalanced datasets.

6.1.2.3 Model-Specific Considerations

Evaluated the performance of different models under imbalanced conditions. Investigated whether certain algorithms, such as ensemble methods or models inherently robust to class imbalance, could offer advantages in addressing this challenge.

6.1.3 Outcomes

The exploration of oversampling techniques and alternative evaluation metrics provided a nuanced understanding of the model's performance in the context of imbalanced data. This not

only facilitated better model training but also informed the interpretation of results and decision-making regarding trade-offs between precision and recall.

6.2 Feature Selection

6.2.1 Problem Statement

The dataset encompassed a plethora of features, necessitating a thoughtful approach to select the most relevant ones for model training. Feature selection aimed to enhance model interpretability, reduce computational complexity, and potentially improve predictive performance.

6.2.2 Approach

6.2.2.1 Correlation Analysis

Conducted correlation analysis to identify highly correlated features and eliminate redundant information. This step was crucial in ensuring that the selected features provided unique and valuable insights.

6.2.2.2 Model-Based Feature Importance

Leveraged model-based feature importance from tree-based algorithms to gauge the contribution of each feature to the predictive performance. This approach provided insights into the relevance of features within the specific context of the chosen models.

6.2.3 Outcomes

The meticulous approach to feature selection resulted in a streamlined set of features that not only contributed meaningfully to the model's predictions but also enhanced model interpretability. The adoption of diverse techniques ensured a comprehensive evaluation of feature relevance, paving the way for a more robust and efficient model.

7. Conclusion

7.1 Model Performance

7.1.1 Balanced Trade-off

The model exhibits significant promise, striking a balanced trade-off between precision and recall. This equilibrium is pivotal in ensuring the model's effectiveness in both correctly identifying customers at risk of churning (precision) and capturing the majority of actual churn cases (recall). The achieved precision of 67.79% and recall of 59.25% suggest a robust and well-calibrated model.

7.1.2 Role of the 'Family' Feature

The introduction of the 'Family' feature has proven instrumental in providing valuable insights into customer dynamics. This feature, derived from the presence of a partner or dependents, adds a nuanced dimension to the prediction model. Its contribution to the overall performance underscores the importance of considering familial relationships in understanding and predicting customer churn.

8. GitHub Repository

- Project Link