|  |
| --- |
| DocuLex Classification and Summary using NLP and RAG |
|  |  |  |
| Submitted by: Aditi Gupta, July 2024 |
| Table of Contents  [Introduction 2](#_Toc167807232)  [Dataset 2](#_Toc167807233)  [Problem Statement 4](#_Toc167807234)  [Data Wrangling 4](#_Toc167807235)  [Exploratory data Analysis 5](#_Toc167807236)  [Preprocessing and Training 10](#_Toc167807237)  [Modelling 11](#_Toc167807238)  [Model Evaluation 11](#_Toc167807239)  [Conclusion 14](#_Toc167807240)  [Recommendations 14](#_Toc167807241)  [Future scope of work 14](#_Toc167807242) |

# Introduction

**What is customer churn?**

Customer churn is defined as when customers or subscribers discontinue doing business with a firm or service.

Customer churn is a critical challenge faced by the telecom industry. As customers switch from one service provider to another, telecom companies experience revenue loss and increased customer acquisition costs. To address this issue, we embarked on a project to develop machine learning models that can predict the likelihood of customer churn.

* Gather insights from the data to understand what is driving the high customer churn rate.
* Develop a Machine Learning model that can accurately predict the customers that are more likely to churn.
* Prescribe customized actions that could be taken to retain each of those customers.

# Dataset

This dataset([link](https://www.kaggle.com/datasets/johnflag/jb-link-telco-customer-churn)) if of JB Link a small size telecom company located in the state of California that provides Phone and Internet services to customers in more than a 1,000 cities and 1,600 zip codes.

|  |  |
| --- | --- |
| Column Name | Description |
| Churn Value | 1 = the customer left the company this quarter. 0 = the customer remained with the company |
| Customer ID | A unique ID that identifies each customer |
| Referred a Friend | Indicates if the customer has ever referred a friend or family member to this company |
| Number of Referrals | Indicates the number of referrals to date that the customer has made |
| Tenure in Months | Indicates the total amount of months that the customer has been with the company by the end of the quarter specified |
| Offer | Identifies the last marketing offer that the customer accepted, if applicable |
| Phone Service | Indicates if the customer subscribes to home phone service with the company |
| Avg Monthly Long-Distance Charges | Indicates the customer’s average long-distance charges, calculated to the end of the quarter |
| Multiple Lines | Indicates if the customer subscribes to multiple telephone lines with the company |
| Internet Service | Indicates if the customer subscribes to Internet service with the company |
| Internet Type | Indicates the type of Internet service the customer subscribes |
| Avg Monthly GB Download | Indicates the customer’s average download volume in gigabytes, calculated to the end of the quarter |
| Online Security | Indicates if the customer subscribes to an additional online security service provided by the company |
| Online Backup | Indicates if the customer subscribes to an additional online backup service provided by the company |
| Device Protection Plan | Indicates if the customer subscribes to an additional device protection plan for their Internet equipment |
| Premium Tech Support | Indicates if the customer subscribes to an additional technical support plan from the company with reduced |
| Streaming TV | Indicates if the customer uses their Internet service to stream television programing from a third-party provider |
| Streaming Movies | Indicates if the customer uses their Internet service to stream movies from a third-party provider |
| Streaming Music | Indicates if the customer uses their Internet service to stream music from a third-party provider |
| Unlimited Data | Indicates if the customer has paid an additional monthly fee to have unlimited data downloads/uploads |
| Contract | Indicates the customer’s current contract type |
| Paperless Billing | Indicates if the customer has chosen paperless billing |
| Payment Method | Indicates how the customer pays their bill |
| Monthly Charge | Indicates the customer’s current total monthly charge for all their services from the company |
| Total Regular Charges | Indicates the customer’s total regular charges, excluding additional charges |
| Total Refunds | Indicates the customer’s total refunds |
| Total Extra Data Charges | Indicates the customer’s total charges for extra data downloads above those specified in their plan |
| Total Long Distance Charges | Indicates the customer’s total charges for long distance above those specified in their plan |
| Gender | The customer's gender |
| Age | The customer’s current age |
| Under 30 | Indicates if the customer is under 30 years old |
| Senior Citizen | Indicates if the customer is 65 or older |
| Married | Indicates if the customer is married |
| Dependents | Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc. |
| Number of Dependents | Indicates the number of dependents that live with the customer |
| City | The city of the customer’s primary residence |
| Zip Code | The zip code of the customer’s primary residence |
| Latitude | The latitude of the customer’s primary residence |
| Longitude | The longitude of the customer’s primary residence |
| Population | A current population estimate for the entire Zip Code area |
| CLTV | Customer Lifetime Value. A predicted CLTV is calculated using corporate formulas and existing data. The higher the value, the more valuable the customer |
| Churn Category | A high-level category for the customer’s reason for churning |
| Churn Reason | A customer’s specific reason for leaving the company |
| Total Customer Svc Requests | Number of times the customer contacted customer service in the past quarter |
| Product/Service Issues Reported | Number of times the customer reported an issue with a product or service in the past quarter |
| Customer Satisfaction | A customer’s overall satisfaction rating of the company from 1 (Very Unsatisfied) to 5 (Very Satisfied) collected on customer service requests |

# Problem Statement

JB link telco company is encountering a problem of a high 27% customer loss leading to a 12% drop in our customer numbers. And urgently need to forecast which customers are prone to churn and recommend tailored strategies to retain customers.

# Data Wrangling

Data wrangling involved the following steps:

* **Importing the dataset**: Loaded the dataset into a pandas data frame.
* **Exploring data columns**: Used functions like head(), info(), and shape to understand the structure and size of the dataset.
* **Handling missing values**: Identified and visualized missing values, then decided on appropriate strategies for handling them.
* **Converting categorical data**: Converted categorical 'Yes'/'No' columns into binary (1/0) for easier analysis.

# Exploratory data Analysis

Exploratory Data Analysis (EDA) helped in uncovering insights and patterns in the data:

* **Customer Statistics**: Analyzed overall customer demographics and service usage.

A graph showing the distribution of a number of different colored bars

Description automatically generated with medium confidence

* **Churn Reasons**: Examined the reasons why customers are churning, such as better service or pricing from competitors.

A graph with different colored bars

Description automatically generated

* **Key Insights**:
  + **Contract Type**: Customers with month-to-month contracts have a significantly higher churn rate. Approximately 54% of customers with month-to-month contracts churned, compared to only 11% with one-year contracts and 3% with two-year contracts.

A chart with green and red squares

Description automatically generated

* + **Customer Satisfaction**: Customers with satisfaction scores below 3 are more likely to churn. Visualization of customer satisfaction distribution highlighted that lower scores correlate with higher churn.

A graph with green and red bars

Description automatically generated

* + **Internet Type**: Customers with cable or DSL services show higher churn rates compared to those with fiber optic services. This insight suggests that service quality differences impact customer retention.

A graph of a customer service distribution

Description automatically generated

* + **Geographic Analysis**: Los Angeles has the highest number of customers, but San Diego has the highest number of churned customers, with many citing better offers from competitors. This indicates a regional disparity in competitive pressure and customer retention.

A screen shot of a graph

Description automatically generated

# Preprocessing and Training

Pre-processing and training data development involved the following steps:

* **Feature Selection**: Removed unnecessary columns and retained features with significant correlation to churn. Created a correlation matrix to identify these features, setting a threshold of 0.2.

A graph with green and blue bars

Description automatically generated with medium confidence

* **Data Transformation**: Converted categorical variables into numerical ones using one-hot encoding.
* **Data Splitting**: Split the data into training (80%) and testing (20%) sets to prepare for model training and evaluation.

# Modelling

The objective of this step was to build predictive models that accurately identify customers who are likely to churn. We evaluated three different machine learning models: Random Forest Classifier, Logistic Regression, and Histogram-based Gradient Boosting Classification Tree (HistGradientBoostingClassifier). Below are the detailed steps involved in the modeling process:

**Model Selection**

We chose the following models due to their distinct characteristics and potential effectiveness in classification tasks:

1. **Random Forest Classifier**:
   * An ensemble method that combines multiple decision trees to improve prediction accuracy and control over-fitting.
   * It can handle large datasets with higher dimensionality and is robust to missing values and outliers.
2. **Logistic Regression**:
   * A simple yet powerful model for binary classification problems.
   * Provides interpretable coefficients that can help in understanding the impact of each feature on the churn probability.
3. **HistGradientBoostingClassifier**:
   * A high-performance implementation of gradient boosting suitable for large datasets.
   * It builds trees iteratively, focusing on correcting errors made by previous trees, leading to improved accuracy.

**Model Training**

Each model was trained on the standardized training dataset. We used cross-validation to tune hyperparameters and prevent overfitting. Here are the steps for training each model:

1. **Data Preprocessing**:
   * Numerical features were standardized using StandardScaler.
   * Categorical features were converted to numerical values using one-hot encoding.
   * Data was split into training (80%) and testing (20%) sets.

# Model Evaluation

After training the models, we evaluated their performance using a set of standard metrics: Accuracy, F1 Score, Precision, Recall, and Confusion Matrix. These metrics provide insights into different aspects of model performance and help in selecting the best model.

**Evaluation Metrics**

1. **Accuracy**:
   * The proportion of correctly predicted instances out of the total instances.
2. **F1 Score**:
   * The harmonic mean of precision and recall. It provides a balance between these two metrics, especially useful for imbalanced classes.
3. **Precision**:
   * The proportion of true positive predictions out of all positive predictions. High precision indicates a low false positive rate.
4. **Recall**:
   * The proportion of true positive predictions out of all actual positives. High recall indicates a low false negative rate.
5. **Confusion Matrix**:
   * A table that describes the performance of a classification model by comparing predicted and actual values.

**Model Performance**

Here are the results for each model:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | F1 Score | Precision | Recall | Confusion Matrix |
| Random Forest Classifier | 89.85% | 80.59% | 88.13% | 74.25% | [[969, 40], [103, 297]] |
| Logistic Regression | 88.64% | 77.65% | 87.97% | 69.50% | [[971, 38], [122, 278]] |
| HistGradientBoostingClassifier | 91.62% | 84.05% | 91.47% | 77.75% | [[980, 29], [89, 311]] |

**Best Model Selection**

The HistGradientBoostingClassifier outperformed the other models across all metrics. It achieved the highest accuracy (91.62%), F1 score (84.05%), precision (91.47%), and recall (77.75%). Therefore, we selected the HistGradientBoostingClassifier as our final model.

**Confusion Matrix Analysis**

The confusion matrix for the HistGradientBoostingClassifier was as follows:

A blue squares with black numbers

Description automatically generated

* **True Positives (TP)**: 311
* **True Negatives (TN)**: 980
* **False Positives (FP)**: 89
* **False Negatives (FN)**: 29

**Key Findings**:

* **High True Positives and True Negatives**:
  + The model correctly identified a large number of both churned and non-churned customers.
* **Low False Negatives**:
  + Only 29 false negatives, indicating that the model missed few actual churn cases.
* **Moderate False Positives**:
  + 89 false positives, which is manageable but indicates some over-prediction of churn.

**5.5 Implications**

* **True Positives**:
  + Correctly identifying customers who are likely to churn allows the company to take proactive measures to retain these customers.
* **True Negatives**:
  + Correctly identifying customers who are not likely to churn avoids unnecessary retention efforts.
* **False Positives**:
  + These represent customers incorrectly predicted to churn, potentially leading to unnecessary retention offers.
* **False Negatives**:
  + These represent churn cases missed by the model, where no retention action would be taken.

# Conclusion

The HistGradientBoostingClassifier demonstrated the best performance, making it the recommended model for predicting customer churn. This model's high accuracy, precision, and recall ensure it can effectively identify customers at risk of churning, allowing JB Link Telecom to implement targeted retention strategies.

# Recommendations

1. **Customer Retention Strategies**:
   * **Customer Retention Programs**: Focus on customers with month-to-month contracts and low satisfaction scores by offering them incentives or improved service plans.
   * **Service Improvement**: Address issues with cable and DSL services to reduce churn in these segments. Invest in upgrading infrastructure to match the performance of fiber optic services.
   * **Targeted Offers**: Use the predictive model to identify customers at high risk of churning and provide them with personalized offers or loyalty programs to retain them.
   * **Customer Support Enhancement**: Improve customer service response times and effectiveness to enhance overall customer satisfaction and reduce churn related to service issues.
2. **Ongoing Model Maintenance**:
   * Regularly retrain the model with new data to ensure its predictions remain accurate.
   * Monitor model performance and adjust hyperparameters as needed.

# Future scope of work

To further improve the churn prediction model and customer retention strategies, consider the following future work:

* **Model Tuning**: Perform hyperparameter tuning on the HistGradientBoostingClassifier to further improve its performance, especially focusing on increasing recall if the cost of false negatives is high.
* **Feature Engineering**: Explore additional feature engineering techniques to create new features that might capture underlying patterns not evident in the current dataset.
* **Data Enrichment**: Integrate additional data sources such as social media interactions, customer feedback, and competitive market data to provide a more comprehensive view of customer behavior and churn drivers.
* **Real-Time Prediction**: Implement the churn prediction model in a real-time environment to provide immediate insights and allow for timely interventions.
* **A/B Testing**: Conduct A/B testing on retention strategies to evaluate their effectiveness and refine approaches based on empirical results.

By implementing these strategies and continuing to enhance the predictive model, JB Link can effectively reduce customer churn, retain more customers, and ultimately improve its revenue and customer satisfaction.