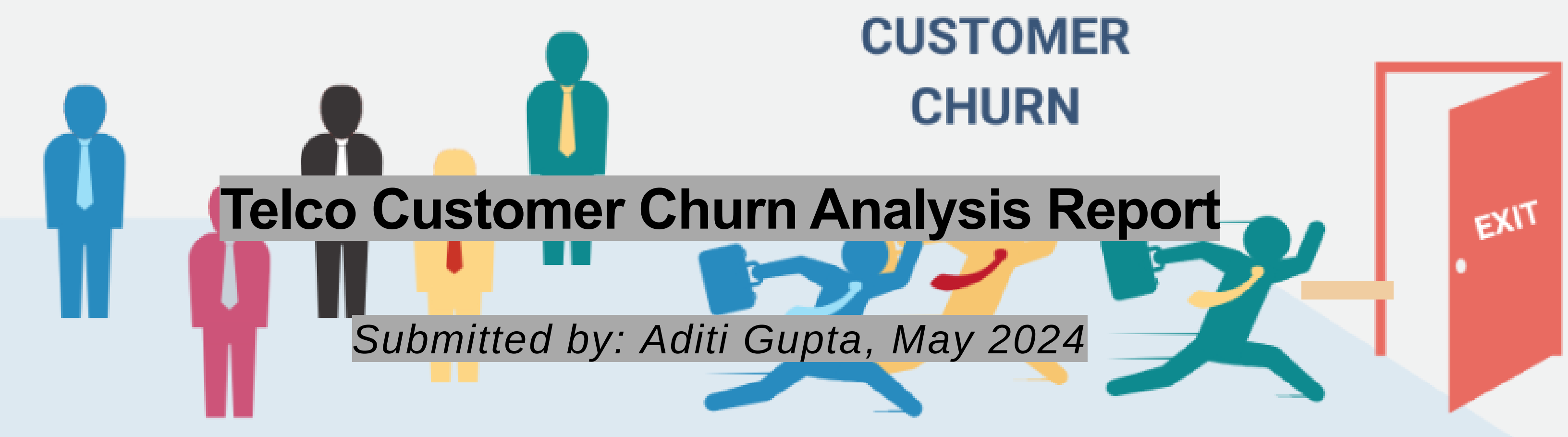

Telco Customer Churn Analysis Report

Submitted by: Aditi Gupta, May 2024

INTRODUCTION

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) 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

Our dataset includes several important values like the total vertical drop

1. Imported data into a dataframe
2. Explored data columns using head(), info(), shape, counting and visualizing missing values.
3. Cleansed data after we counted null values and unique values per column. Also converted the 'Yes' 'No' values columns to 1 and 0 so that these can be included as numeric features for statistical analysis in next step.
4. In end removed the unnecessary columns.

EXPLORATORY DATA ANALYSIS

Our initial step was to perform exploratory data analysis (EDA) to gain insights into the dataset. We cleaned the data, handled missing values, and engineered relevant features. Additionally, we conducted a descriptive statistics summary, distribution analysis, and correlation analysis. EDA provided us with valuable insights into the relationships between different variables and helped us identify potential patterns and trends. Below are some of the questions that helped in the analysis.

Hypothesis

1. Customers with longer tenure are less likely to churn than those with short tenure.

2. Customers with lesser income are likely to churn than those who have higher

3. Customers are more likely to switch to a network that offer better call plan to call other networks.

4. Customers who patronize a particular plan or service are most likely to churn.

Business Questions:

Here are seven potential business questions that can be answered using the telecom churn data:

1. What is the overall churn rate for the telecom company during the observed period?

This question aims to provide an understanding of the churn rate as a baseline for further analysis and decision-making.

2. Are there any specific regions or geographic areas with a higher churn rate compared to others?

By analyzing churn rates across different regions, the telecom company can identify areas that require targeted retention strategies or improved service quality.

3. Do customers who have been with the network for a longer tenure exhibit lower churn rates?

This question explores the relationship between customer tenure and churn rate, helping the company understand the impact of customer loyalty on churn.

4. Is there a correlation between top-up amount (MONTANT) and churn rate?

This question investigates whether customers with higher or lower top-up amounts are more likely to churn, providing insights into the relationship between spending behavior and churn.

5. Are customers who frequently activate specific top pack packages (TOP_PACK) less likely to churn?

This question examines the influence of top pack usage on churn rate, helping the company identify which packs contribute to customer retention and can be promoted further.

6. Are customers who have a higher number of on-net calls (ON_NET) less likely to churn?

This question seeks help to help the company to assess how ON_NET or inter expresso calls contribute to the churn rate of the company.

7. Do customers who regularly refill their accounts (FREQUENCE_RECH) have lower churn rates compared to those who refill less frequently?

- These questions can provide valuable insights into churn patterns, customer behavior, and factors contributing to customer retention.

Analyzing the telecom churn data in relation to these questions can help the company make informed decisions and develop effective strategies to reduce churn, improve customer satisfaction, and enhance business performance.

These are some of the visualizations we had from the analysis:

To capture relevant state data related to our interests.

1. Analyzed the columns by looking at the pair plot to find columns of interests and found below columns those have good mix of both churned and non-churned customers - Number of Referrals, Avg Monthly GB Download, Avg Monthly Long Distance Charges, Total Refunds, Total Extra Data Charges, Number of Dependent, Total Customer Svc Requests, Customer Satisfaction
2. Analyzed the distribution of values and statistical values like min, max, mean, std, 25,50,75th percentiles.
3. Also looked at the churn category and reasons, these columns are not always filled but top two reasons mentioned for churn here are better competitor offer (device, more data, better coverage/speed) and customer service provided to individuals where customers were unhappy with the service provider (person whom they talked on phone or worked by logging a ticket). Further analysis of both these reasons are out of scope for our churn prediction because competitor offer needs an analysis of market analysis of competitor offers and customer service needs further analysis of data on how customer service is solving customer issues like how much time it takes to respond, to solve and what are their blockers. For these reasons, removing the columns Churn category, and Churn Reason. We will use other features to predict the churn for this project and in the next step we will use correlation matrix to select the features.

PREPROCESSING AND TRAINING

After identifying the four categories with the strongest correlation to price

In this step:

1. Created the correlation matrix to see the relationship of the features and removed the features where correlation is lower than 0.2 threshold.
2. Created the heatmap to visulaize the correlation of the remaining columns and found that Churn has highest correlation with columns Total Customer Svc Requests, Product/Service Issues Reported and Monthly charged.

3. Explored the data frame using pairplot again to see the column relationship. And this visualization also confirmed the strong relationship between churn and service request columns as well as monthly charge.
4. In this step converted categorical columns into one-hot encoding and splitted the data into 80:20 ratio of training and test.

MODELLING

Models Evaluated Random Forest Classifier Logistic Regression HistGradientBoostingClassifier Evaluation Metrics Accuracy: Represents the proportion of correctly predicted instances out of the total instances.

Random Forest Classifier: 90.13% Logistic Regression: 88.36% HistGradientBoostingClassifier: 91.34% The HistGradientBoostingClassifier has the highest accuracy among the three models, suggesting it performs best in terms of overall correct predictions.

F1 Score: The harmonic mean of precision and recall, which provides a balance between these two metrics. It's particularly useful when the classes are imbalanced.

Random Forest Classifier: 80.93% Logistic Regression: 76.90% HistGradientBoostingClassifier: 83.47% Again, the HistGradientBoostingClassifier has the highest F1 score, indicating the best balance between precision and recall.

Precision: The proportion of true positive predictions out of all positive predictions.

Random Forest Classifier: 89.67% Logistic Regression: 88.06% HistGradientBoostingClassifier: 91.12% The HistGradientBoostingClassifier shows the highest precision, suggesting that when it predicts a customer will churn, it is correct 91.12% of the time.

Recall: The proportion of true positive predictions out of all actual positives.

Random Forest Classifier: 73.75% Logistic Regression: 68.25% HistGradientBoostingClassifier: 77.00% The HistGradientBoostingClassifier also has the highest recall, meaning it identifies actual churns 77.00% of the time.

Confusion Matrix The confusion matrix provides a breakdown of the predictions made by the model compared to the actual labels. It is in the format:

mathematica Copy code [[True Negative, False Positive], [False Negative, True Positive]]

Random Forest Classifier:

True Negatives: 975 False Positives: 34 False Negatives: 105 True Positives: 295

Logistic Regression:

True Negatives: 972 False Positives: 37 False Negatives: 127 True Positives: 273

HistGradientBoostingClassifier:

True Negatives: 979 False Positives: 30 False Negatives: 92 True Positives: 308

Summary: HistGradientBoostingClassifier performs best across all metrics, making it the preferred model among the three evaluated. Random Forest Classifier comes second in terms of performance. Logistic Regression is the least accurate among the three models but still shows reasonable performance.

Evaluated three models in this step and picked HistGradientBoostingClassifier

1. Because HistGradientBoostingClassifier performs best across all metrics, making it the preferred model among the three evaluated. Random Forest Classifier comes second in terms of performance. Logistic Regression is the least accurate among the three models but still shows reasonable performance.
2. Explored the feature imprtance for top two models HistGradientBoostingClassifier and Random forest.
3. Evaluated the confusion matrix for HistGradientBoostingClassifier.
4. Visualized the predicted vs actual chun and not churn values.

FEATURE ENGINEERING

976 instances are True Negatives (TN): The model correctly predicted 976 customers as not churning out of the actual not churned customers.
33 instances are False Positives (FP): The model incorrectly predicted 33 customers as churning when they actually did not churn. These are potential false alarms where the model incorrectly flags a non-churning customer as churning.
124 instances are False Negatives (FN): The model incorrectly predicted 124 customers as not churning when they actually did churn. These are the cases where the model missed identifying a churning customer, which could lead to missed opportunities to retain these customers.
276 instances are True Positives (TP): The model correctly predicted 276 customers as churning out of the actual churning customers. These are the customers the model correctly identified as churning.

FUTURE SCOPE OF WORK

The price prediction model