Telco Churn Analysis

Predict Customers who are likely to Churn

Saptarshi Syam

July 2023

Contents

1. Problem Statement

The telecommunications industry is characterized by intense competition and rapid technological advancements. With the rise of mobile devices and the increasing demand for high-speed internet, telcos continually face challenges to retain their customer base. Customer churn, the rate at which customers leave the service provider and switch to competitors, is a recurring issue in the telecom sector.

Impact of Churn on Telcos

Customer churn has significant adverse effects on Telcos:

I. **Revenue Loss:** Losing customers means losing their recurring monthly revenue, impacting on the company's financial performance.
II. **Acquisition Costs:** Acquiring new customers is more expensive than retaining existing ones. Churn requires additional marketing and promotional efforts to attract new customers, further straining resources.
III. **Customer Lifetime Value (CLV)**: Churn directly affects the CLV of customers. Retaining customers for a longer period increases their overall value to the company.
IV. **Brand Reputation**: High churn rates can negatively impact the telco's reputation, leading potential customers to perceive it as offering subpar services.
V. **Market Share**: Churn leads to a decrease in market share, giving an advantage to competitors.

To address the churn problem, telecom companies collect vast amounts of customer data, including demographics, usage patterns, billing information, customer service interactions, contract details, and more. Analyzing this data can provide valuable insights into customer behavior, preferences, and pain points. This Project aims in building models for the Telecom Companies which they can use to predict Customers who are more likely to churn.

2. The Client

The client of the telco churn dataset refers to the entity or organization that owns and operates the telecommunications network and services, and who has provided the dataset for analysis. In this context, the client can be a specific telecommunication company, which may vary depending on the source of the dataset and the purpose of the analysis.

Their goal is to gain actionable insights into customer churn behavior to improve their business strategies, enhance customer retention, and maintain a competitive position in the telecommunications industry. The collaboration between the client and the data analysis team plays a crucial role in driving meaningful outcomes from the dataset analysis.

3. The Data

The dataset for Telco churn analysis is a collection of structured data containing information about customers and their interactions with a telecommunication company. It contains Customer Demographic information, usage patterns, service subscriptions, contract details, and tenure, among other relevant factors. It serves as the foundation for conducting in-depth analysis and developing predictive models to understand customer churn behavior and identify factors influencing churn.

We have 5 Data Sources as explained below:-

i. **Customer_churn_demographics.csv**:- This has the Demographic information about Customer like their Age, Gender, Marital Status, Number of Dependents etc.
ii. **Customer_churn_location.csv**: This has the location details for the Customer viz. Country, State, City, Latitude & Longitude among other details.
iii. **Customer_churn_population.csv**:- This has the Zip code along with their Population.
iv. **Customer_churn_services.csv**:- This has details about the Services availed by the customer like Streaming Services, multiple phone lines etc.
v. **Customer_churn_status.csv**:- This has the Churn Status of the customer along with the Satisfaction score, churn reason, category among other details.

3.1 Data Wrangling

Data wrangling, also known as data preprocessing or data cleaning, is a crucial step in telco churn analysis. It involves preparing the dataset for analysis by handling missing data, dealing with outliers, transforming variables, and ensuring data consistency. We did the below steps as part of Data Wrangling step of our Analysis:

i. **Data Inspection**: Explore the dataset's structure, inspect the first few rows, and check for the number of rows and columns to get a sense of the data's content and overall quality.
ii. **Handling Missing Values**: We found that the Churn Dataset has many missing values in the Churn Label field. But on analysis, we found that these are empty because for customers who are still Active and hence no action was needed for them. There were no missing values in any of the other datasets.
iii. **Data Type Conversion**: Ensured that the data types of each column are appropriate for analysis. All the Values were found to be in required formats and nothing had to be done for further processing.
iv. **Dealing with Outliers**: There were very few outliers in the Datasets and we will explore on this in the EDA Stage.

v. **Removing Redundant Columns**: We found a redundant column like ‘Count’ which was having the same value for all records and we removed it before merging the datasets into one for further analysis.

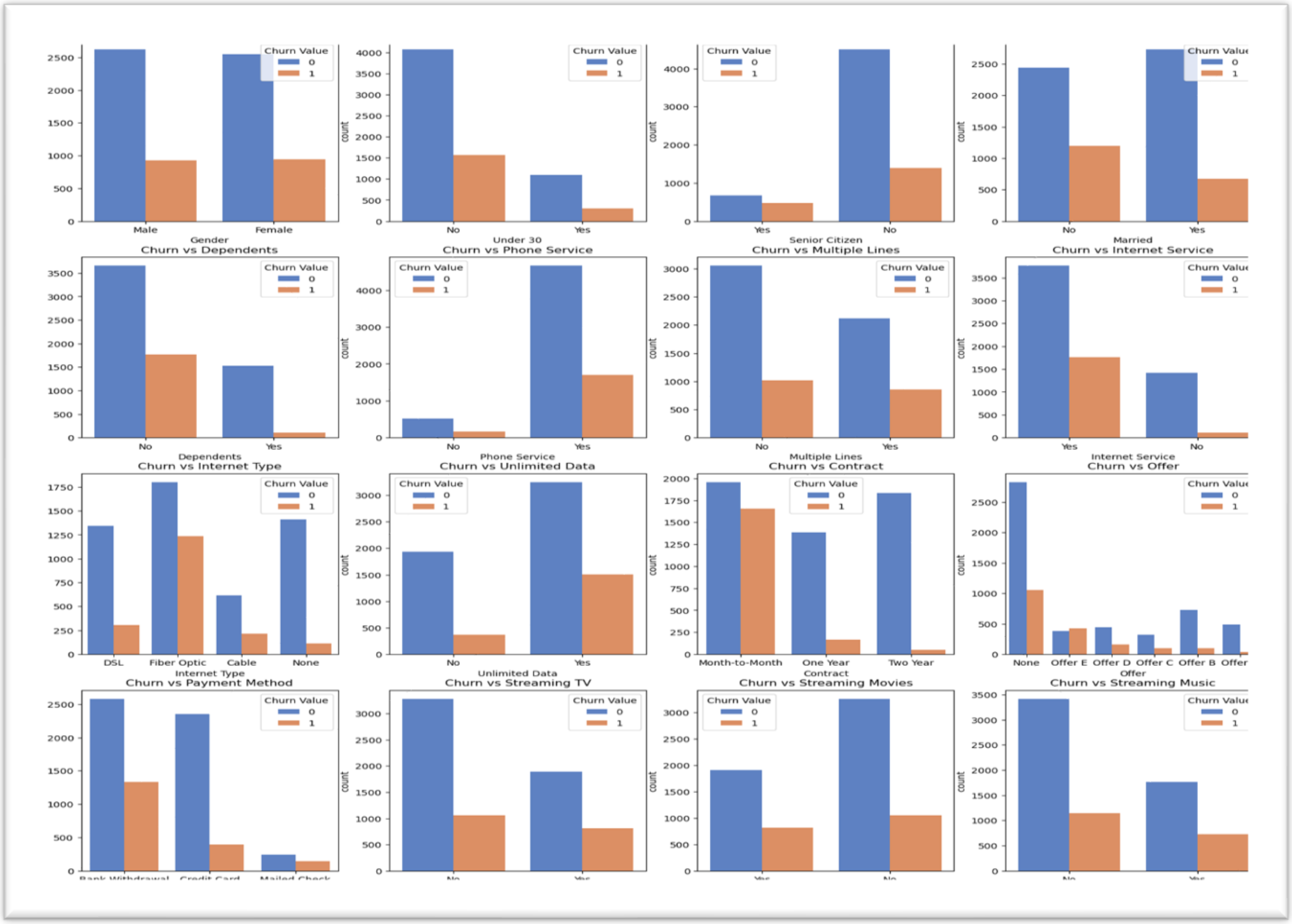
vi. **Data Merge**: We merged the Customer Dataset along with other dataset to come up with 1 Dataset which we will use for EDA and further stages.

4. Exploratory Data Analysis

We began our data exploration by distinguishing between numerical and non-numerical features in the Customer Dataset. The initial focus was on exploring the categorical variables, followed by an investigation of the numerical variables in the dataset.

4.1 Categorical Variables

We generated count plots for all the categorical variables, displaying the counts of churned and active members for each variable.



We identified the below trends: -

1. There is no relation between Gender and Churn Rate. Male and Female have almost equal Churn Rate.
2. People over 30 have a Higher Churn Rate than under it.
3. However, Senior Citizens have a lower Churn Rate.
4. Unmarried People or people with no dependents have a higher Churn Rate than Married ones or someone having dependents.
5. Customers with Phone Service have more tendency to churn than without it.
6. There is no appreciable correlation in churn rates and people with multiple or non-multiple lines.
7. People with Internet Service have churned more than without it.
8. Customers with their Internet Type as Fiber Optic has churned the most.
9. Customers availing Unlimited Data have churned the most than without them.
10. Customers who have no contract have churned the most.
11. Customers who have not availed any Offers have churned the maximum.
12. Customers who opted for "Bank Withdrawal" have a high Churn Rate.
13. There is no appreciable difference in Churn Rates between people opting and not opting media services.

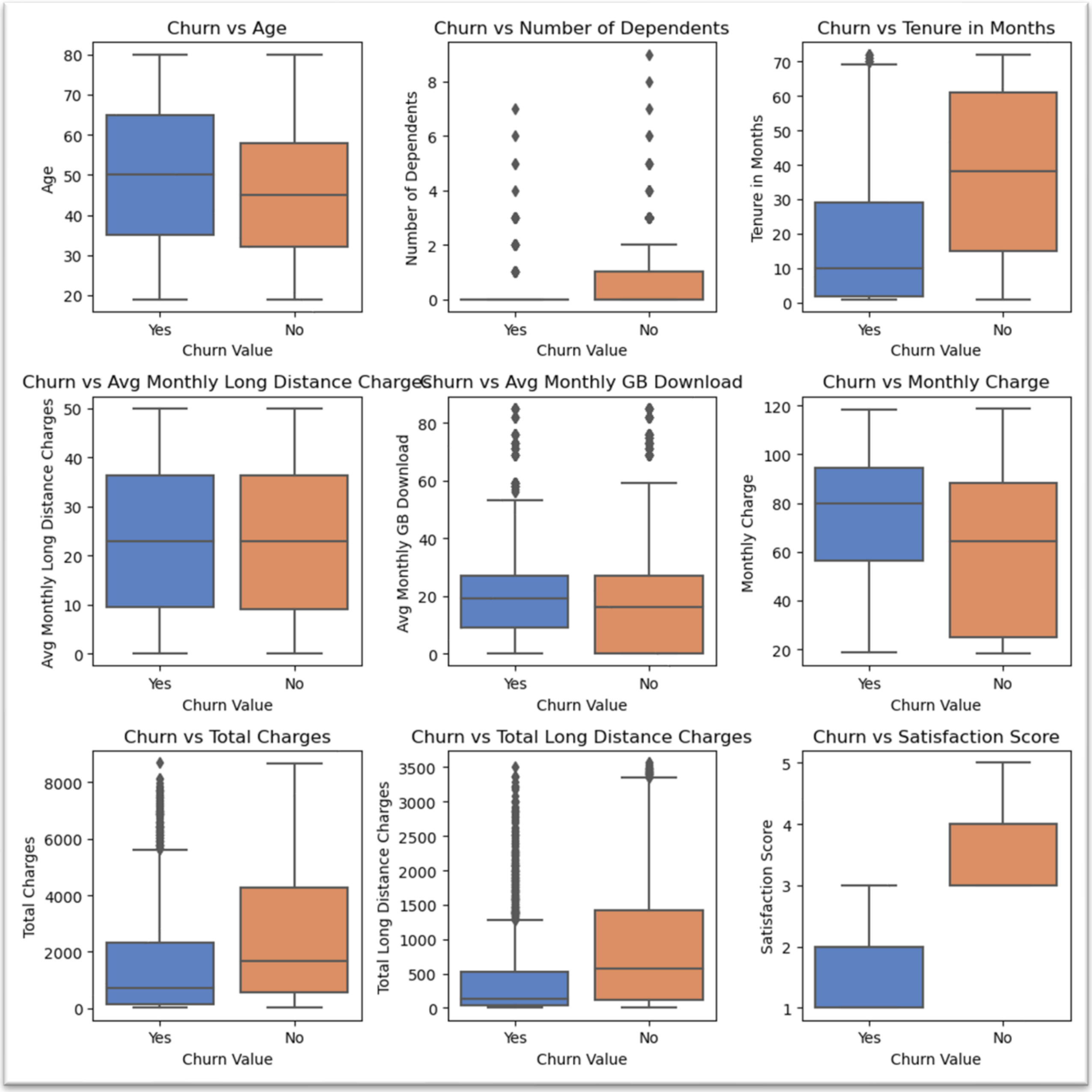
Subsequently, we determined the degree of dependency between the categorical variables and our target variable, which is the Churn Value. It emerged that the top 5 variables which have most impact on Churn Value are as follows: -

i. Contract: People with No or Monthly contract has churned the most
ii. Internet Type: People with Fiber Optics has a High Churn Rate
iii. Dependents: People with Dependents have churned the most
iv. Offers: Customers who was given no offers has left
v. Payment Method: Customers availing Automatic Bank Payments have churned the most

4.2 Non-Categorical Variables

Next, we analyzed the Numerical Features of the Dataset like Age, Number of Dependents, Tenure, Monthly and Total Costs, Services cost etc.

We plot them into a Boxplot to find the relation they have with churning.



From the above Plots, we identified the below pattern: -

i. There is not much difference in mean age of Customers who have churned and who haven't.

ii. Customers who are comparatively new have churned more.

iii. Customers with Zero Dependents or too many Dependents have a high churn rate.

iv. There is no appreciable relation between Churning and Average Long-Distance Calls and Data Download or extra Data download.

v. Customers with More Monthly charges have churned more.

vi. Customers who have a very High Total charges and Long-Distance Charges have also churned more.

Upon analyzing both the categorical and non-categorical features, a clear pattern emerged. We observed that customers who avail themselves of multiple services such as Internet, Phone lines, and Streaming services, those without any contracts, and individuals with higher monthly bills have a higher tendency to churn.

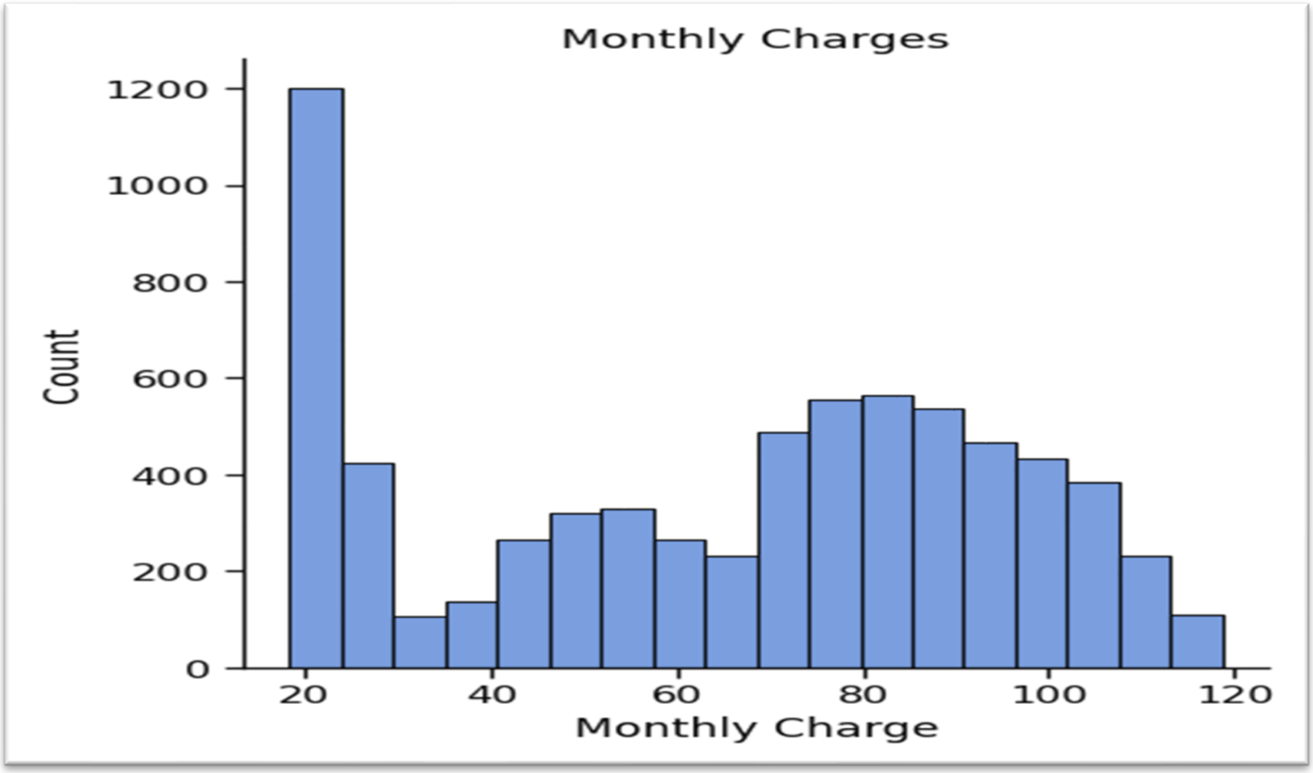
4.3 Null Hypothesis

We formulate a Null Hypothesis to test the absence of a relationship between Monthly Charges and Churn Rate.

H$_{\textbf{null}}$: Customers who Churned have no relation with High Monthly Charges

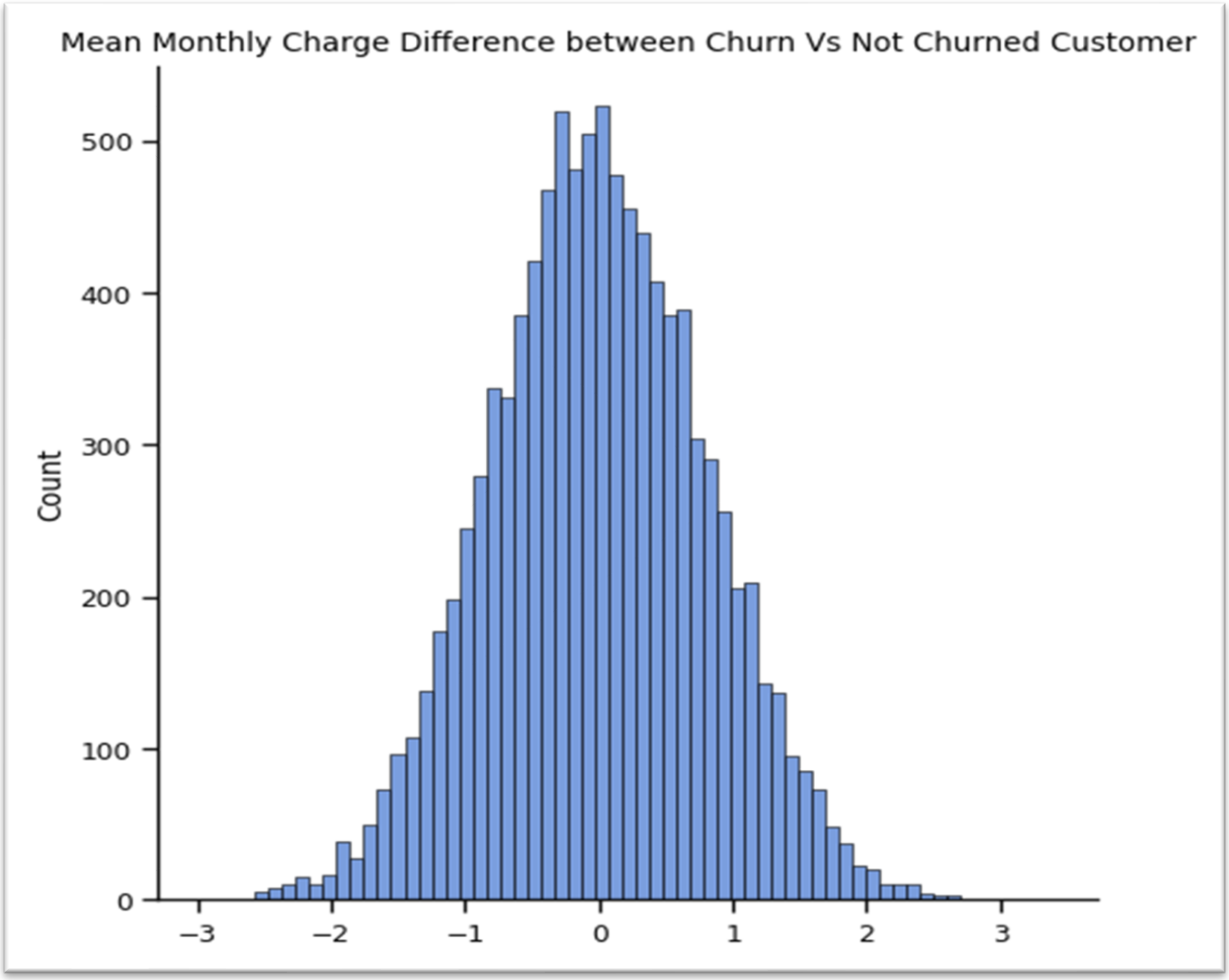
H$_{\textbf{Alternative}}$: Customers with High Monthly Charge is more like to churn

We picked a **significance level** of 0.05.



Since Data is Not Normal, we did a Parametric Test.

We computed the Mean Monthly Charge Difference between Churn Vs Not Churned Customer.



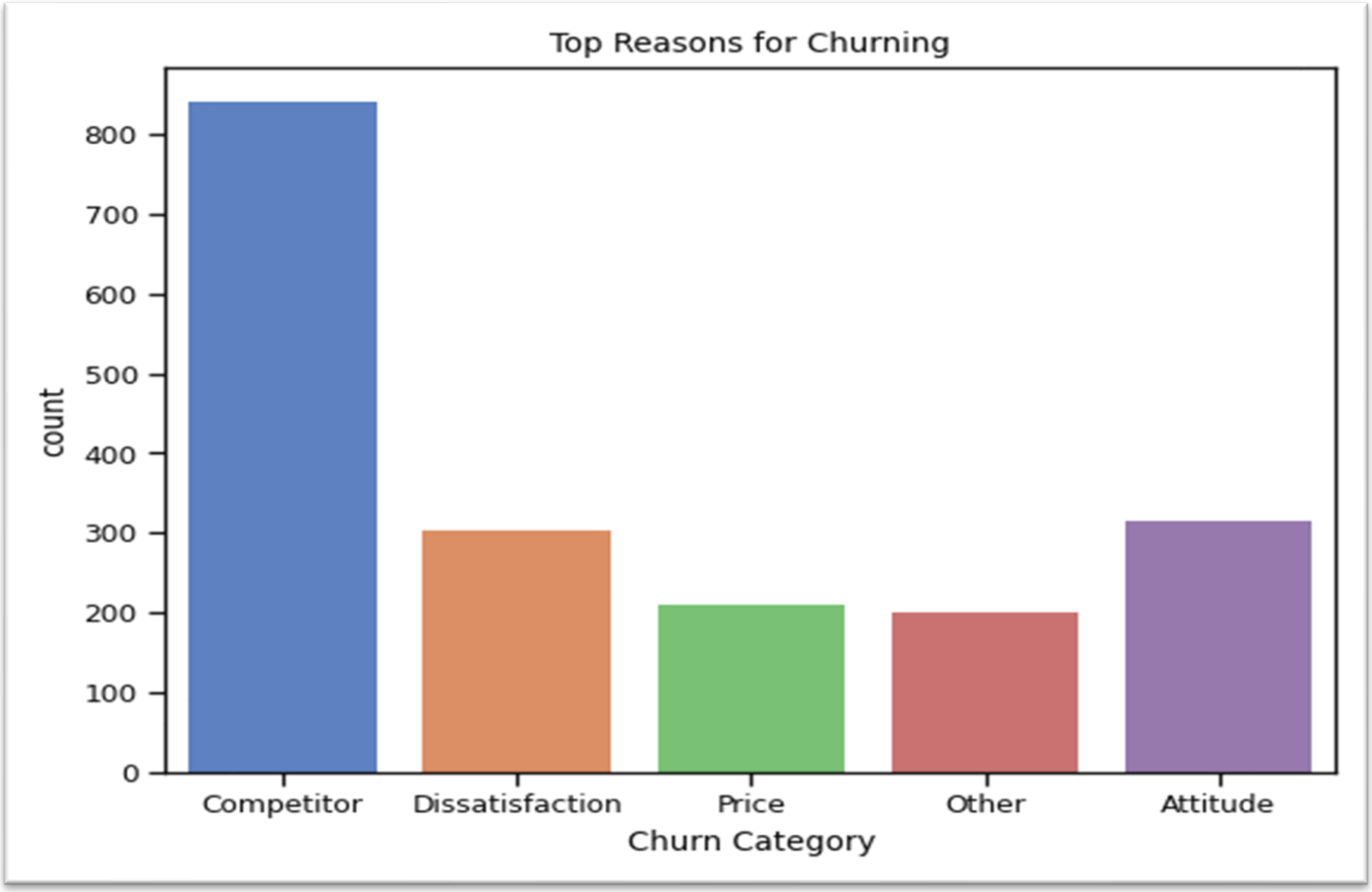
We computed the p-value and it came as 0.0

So it doesn't matter which significance level we pick; our observed data was statistically significant, and we rejected the Null.

Therefore, we concluded that Monthly Charges **<u>DO</u>** have an impact in the Churn Rate.

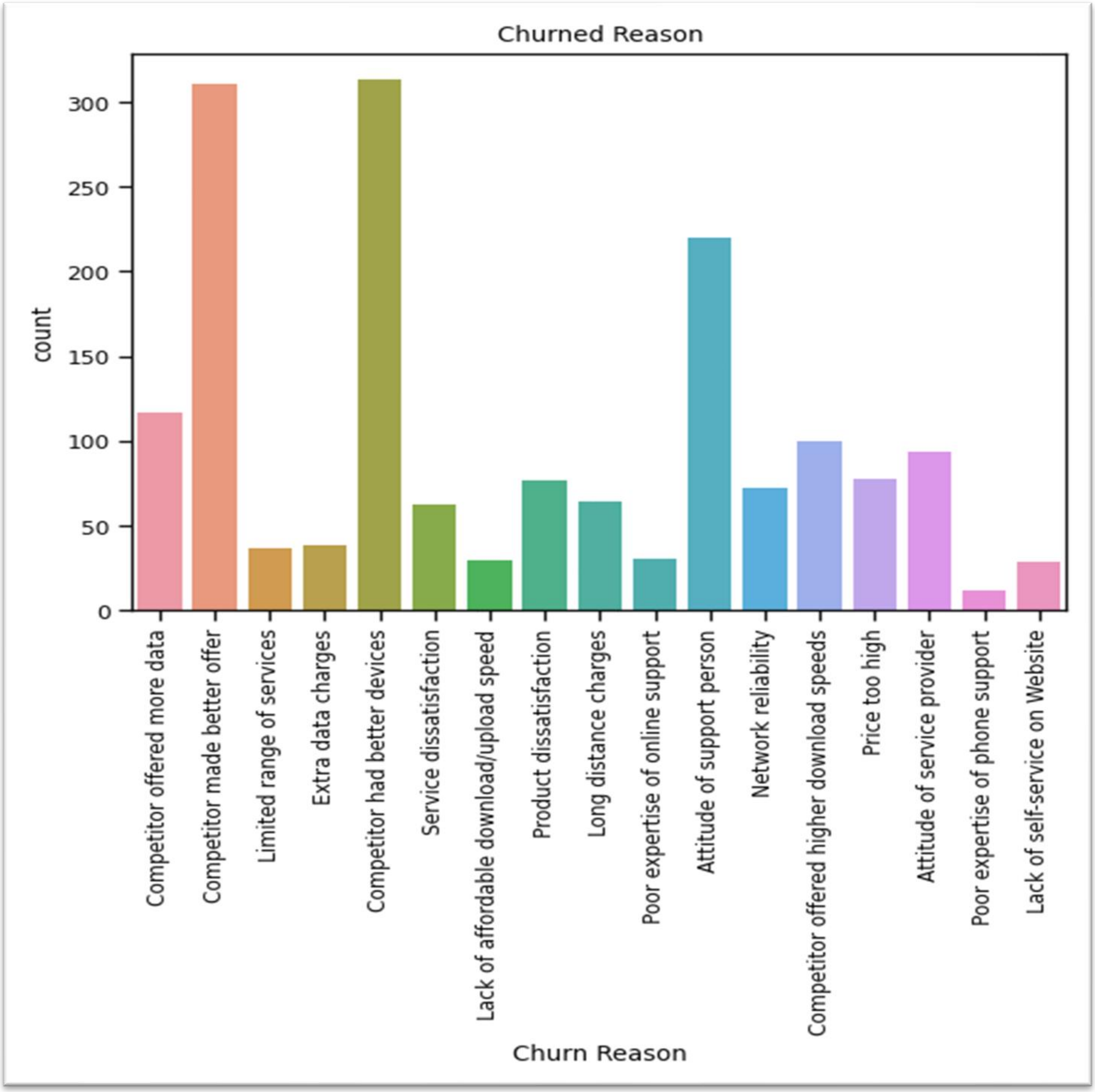
4.4 Churn Analysis

We diverted our attention to the Churn Dataset to find the Top Reasons for Churning.



Competitor, Dissatisfaction and Attitude of Support Staff are Top 3 Reasons for churning.

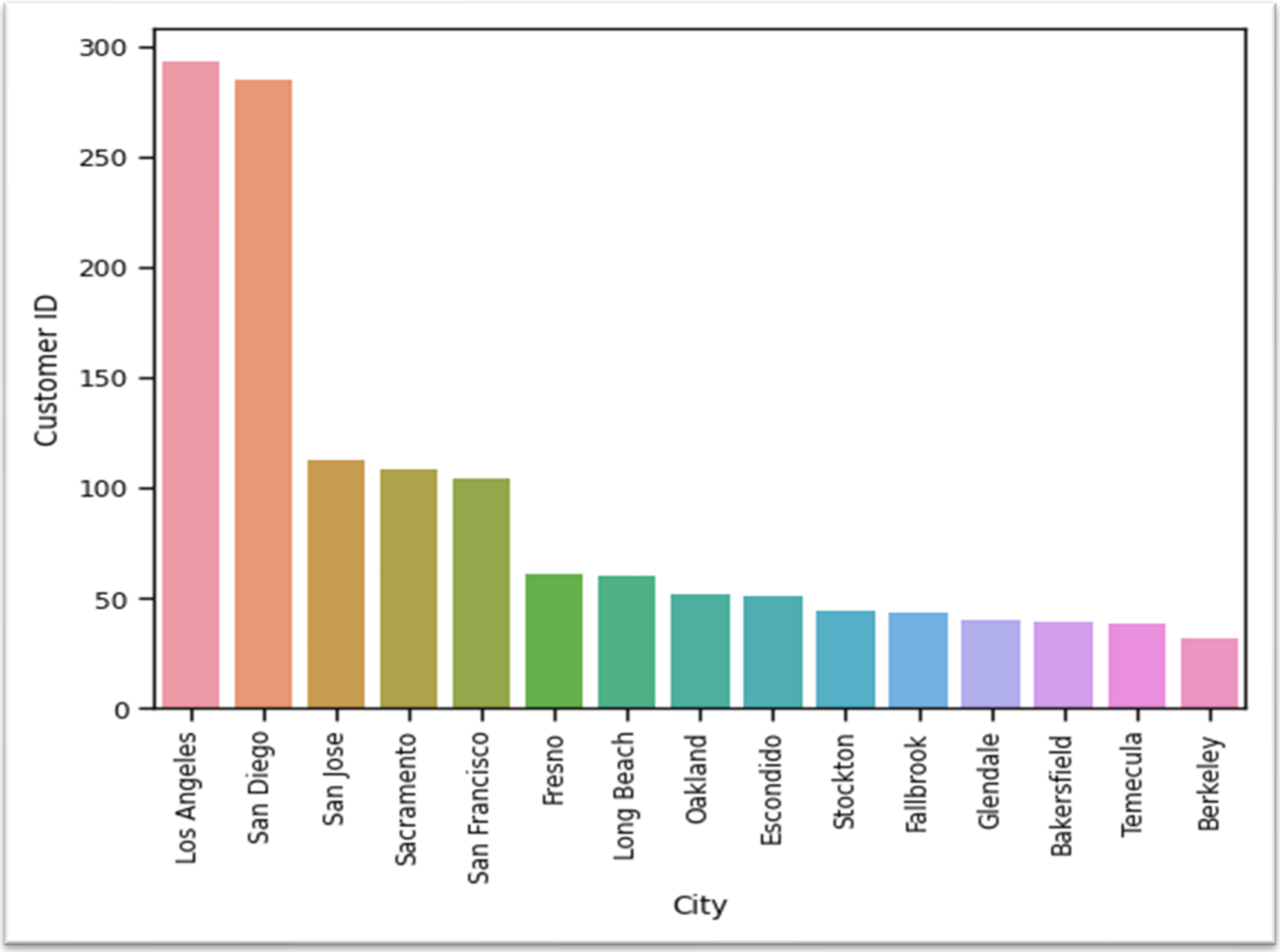
Drilling further by removing some Values like “Moved”, “Deceased”, “No Answer” we arrived at the below reasons for Customer Churning:-



Evidently, the main driver behind customer churn is the availability of better offers and devices from competitors. Therefore, our Telecom Client should focus on enhancing their offerings in terms of competitive deals and devices to retain customers effectively. Additionally, there is a need for significant improvement in the attitude of the support staff, as it plays a crucial role in customer retention. By prioritizing these areas, the Telecom Client can improve overall customer satisfaction and reduce churn rates.

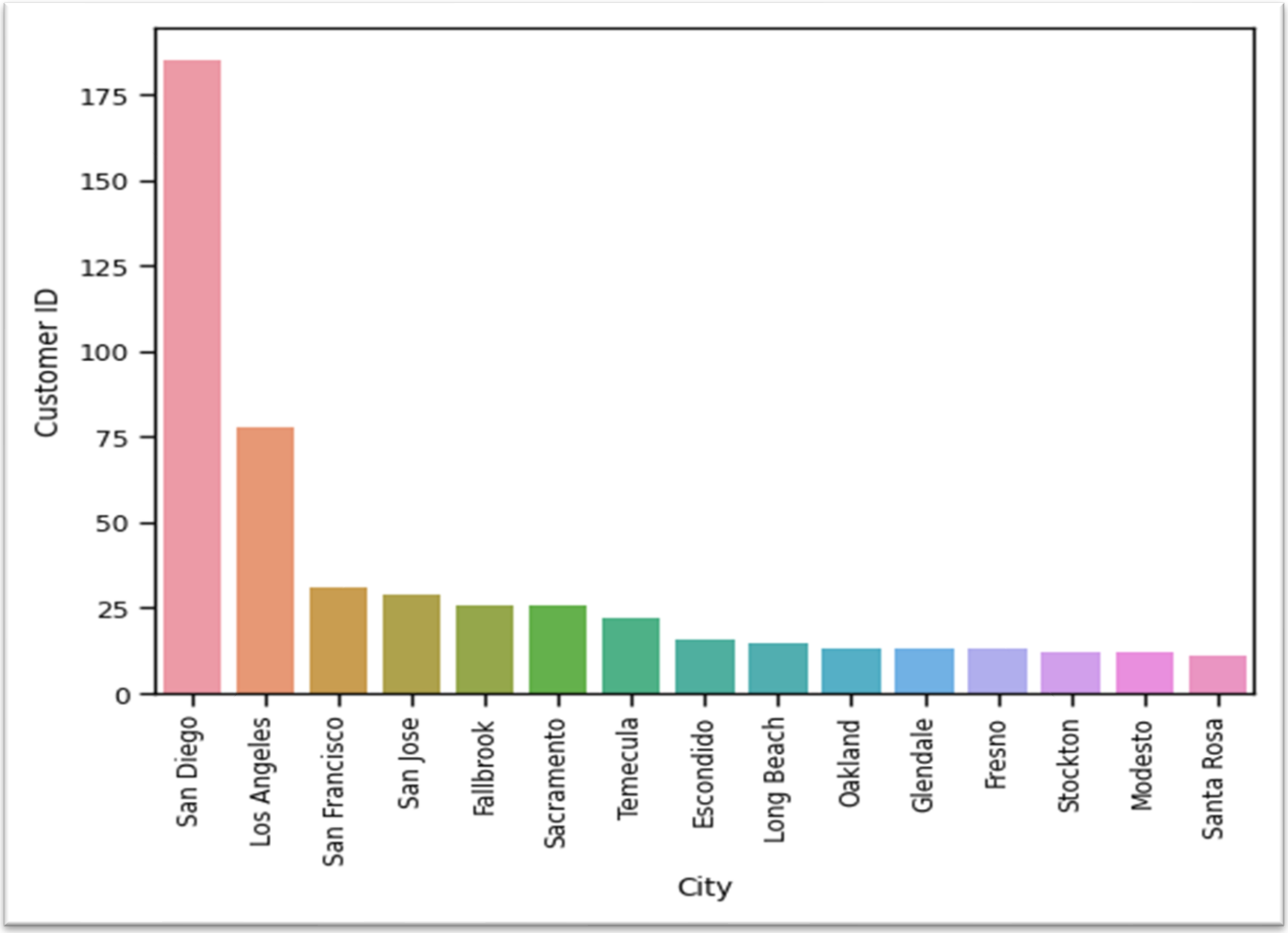
4.5 Effect of Location

During our investigation, we explored whether the location had any correlation with the churn rate. All our Customers were from California, United States. Our analysis revealed that certain specific cities displayed a significantly higher churn rate compared to others. Identifying these areas with a high churn rate allows us to develop a targeted plan to address the concerns of those regions, aiming to prevent churn and improve overall customer retention.



Los Angeles & San Diego has the largest Customer base. However, this is not surprising as the Cities are big with huge population.

We have now examined the churn rate across different locations. By analyzing the churn rate variations among various regions, we gain valuable insights that can help us develop location-specific strategies to address customer concerns and reduce churn effectively.

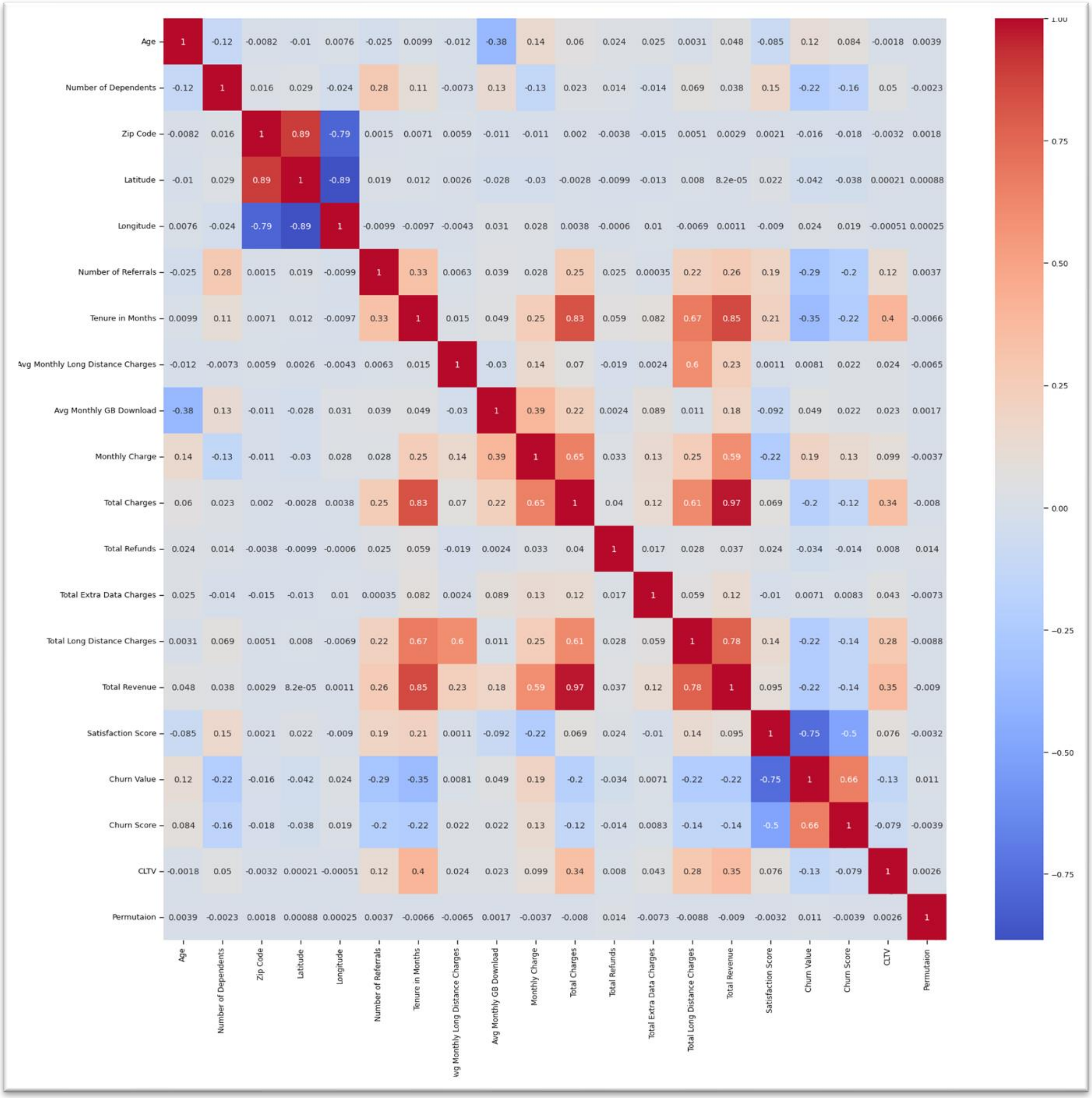


The Churn Rate is proportional to the Population of the Cities and hence we conclude that location has no relation with Churn Rate.

4.6 Correlation Analysis

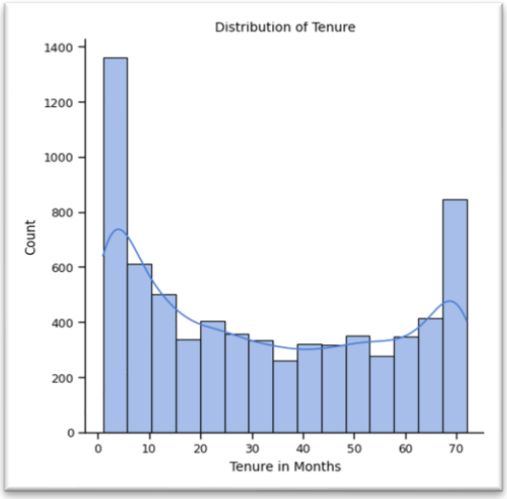
Exploring Variable's Relationship with Churning:

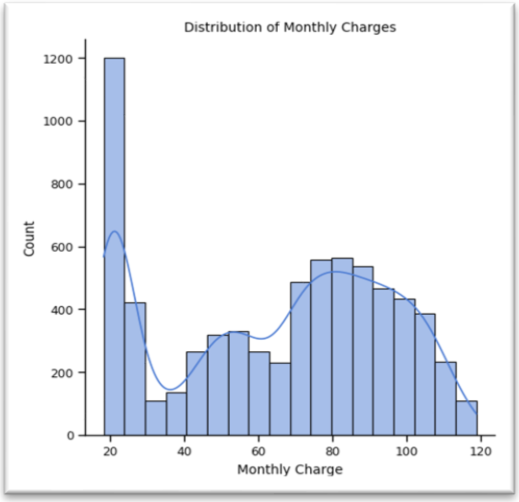
Having established the relationship between the relevant features and our target variable, the Churn Rate, we further visualized the trends by generating a heatmap. This heatmap provides a clear and intuitive representation of the correlations between different features and the Churn Rate, helping us gain deeper insights into the factors influencing customer churn.

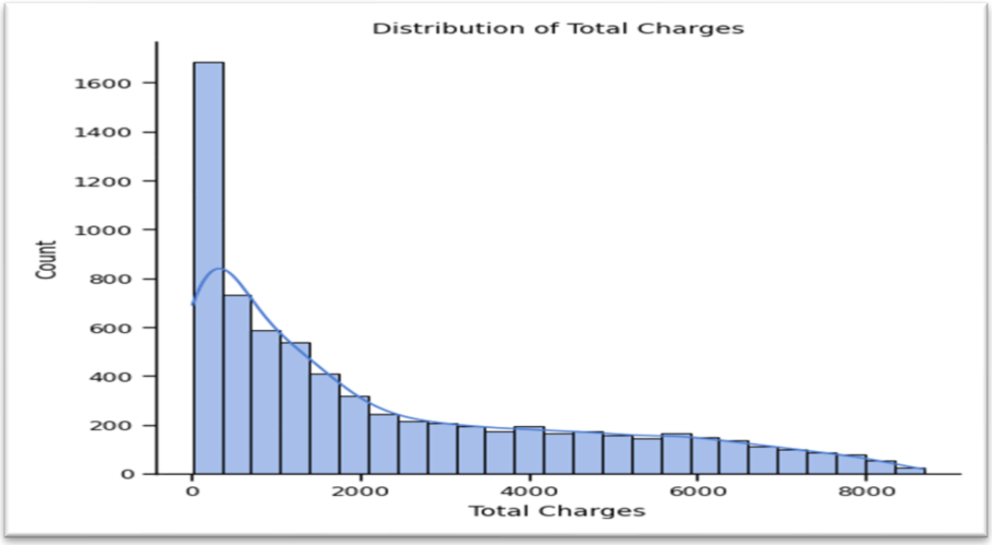


	Age	Number of Dependents	Zip Code	Latitude	Longitude	Number of Referrals	Tenure in Months	vg Monthly Long Distance Charges	Avg Monthly GB Download	Monthly Charge	Total Charges	Total Refunds	Total Extra Data Charges	Total Long Distance Charges	Total Revenue	Satisfaction Score	Churn Value	Churn Score	CLTV	Permutaion
Age	1	-0.12	-0.0082	-0.01	0.0076	-0.025	0.0099	-0.012	-0.38	0.14	0.06	0.024	0.025	0.0031	0.048	-0.085	0.12	0.084	-0.0018	0.0039
Number of Dependents	-0.12	1	0.016	0.029	-0.024	0.28	0.11	-0.0073	0.13	-0.13	0.023	0.014	-0.014	0.069	0.038	0.15	-0.22	-0.16	0.05	-0.0023
Zip Code	-0.0082	0.016	1	0.89	-0.79	0.0015	0.0071	0.0059	-0.011	-0.011	0.002	-0.0038	-0.015	0.0051	0.0029	0.0021	-0.016	-0.018	-0.0032	0.0018
Latitude	-0.01	0.029	0.89	1	-0.89	0.019	0.012	0.0026	-0.028	-0.03	-0.0028	-0.0099	-0.013	0.008	8.2e-05	0.022	-0.042	-0.038	0.00021	0.00088
Longitude	0.0076	-0.024	-0.79	-0.89	1	-0.0099	-0.0097	-0.0043	0.031	0.028	0.0038	-0.0006	0.01	-0.0069	0.0011	-0.009	0.024	0.019	-0.00051	0.00025
Number of Referrals	-0.025	0.28	0.0015	0.019	-0.0099	1	0.33	0.0063	0.039	0.028	0.25	0.025	0.00035	0.22	0.26	0.19	-0.29	-0.2	0.12	0.0037
Tenure in Months	0.0099	0.11	0.0071	0.012	-0.0097	0.33	1	0.015	0.049	0.25	0.83	0.059	0.082	0.67	0.85	0.21	-0.35	-0.22	0.4	-0.0066
Avg Monthly Long Distance Charges	-0.012	-0.0073	0.0059	0.0026	-0.0043	0.0063	0.015	1	-0.03	0.14	0.07	-0.019	0.0024	0.6	0.23	0.0011	0.0081	0.022	0.024	-0.0065
Avg Monthly GB Download	-0.38	0.13	-0.011	-0.028	0.031	0.039	0.049	-0.03	1	0.39	0.22	0.0024	0.089	0.011	0.18	-0.092	0.049	0.022	0.023	0.0017
Monthly Charge	0.14	-0.13	-0.011	-0.03	0.028	0.028	0.25	0.14	0.39	1	0.65	0.033	0.13	0.25	0.59	-0.22	0.19	0.13	0.099	-0.0037
Total Charges	0.06	0.023	0.002	-0.0028	0.0038	0.25	0.83	0.07	0.22	0.65	1	0.04	0.12	0.61	0.97	0.069	-0.2	-0.12	0.34	-0.008
Total Refunds	0.024	0.014	-0.0038	-0.0099	-0.0006	0.025	0.059	-0.019	0.0024	0.033	0.04	1	0.017	0.028	0.037	0.024	-0.034	-0.014	0.008	0.014
Total Extra Data Charges	0.025	-0.014	-0.015	-0.013	0.01	0.00035	0.082	0.0024	0.089	0.13	0.12	0.017	1	0.059	0.12	-0.01	0.0071	0.0083	0.043	-0.0073
Total Long Distance Charges	0.0031	0.069	0.0051	0.008	-0.0069	0.22	0.67	0.6	0.011	0.25	0.61	0.028	0.059	1	0.78	0.14	-0.22	-0.14	0.28	-0.0088
Total Revenue	0.048	0.038	0.0029	8.2e-05	0.0011	0.26	0.85	0.23	0.18	0.59	0.97	0.037	0.12	0.78	1	0.095	-0.22	-0.14	0.35	-0.009
Satisfaction Score	-0.085	0.15	0.0021	0.022	-0.009	0.19	0.21	0.0011	-0.092	-0.22	0.069	0.024	-0.01	0.14	0.095	1	-0.75	-0.5	0.076	-0.0032
Churn Value	0.12	-0.22	-0.016	-0.042	0.024	-0.29	-0.35	0.0081	0.049	0.19	-0.2	-0.034	0.0071	-0.22	-0.22	-0.75	1	0.66	-0.13	0.011
Churn Score	0.084	-0.16	-0.018	-0.038	0.019	-0.2	-0.22	0.022	0.022	0.13	-0.12	-0.014	0.0083	-0.14	-0.14	-0.5	0.66	1	-0.079	-0.0039
CLTV	-0.0018	0.05	-0.0032	0.00021	-0.00051	0.12	0.4	0.024	0.023	0.099	0.34	0.008	0.043	0.28	0.35	0.076	-0.13	-0.079	1	0.0026
Permutaion	0.0039	-0.0023	0.0018	0.00088	0.00025	0.0037	-0.0066	-0.0065	0.0017	-0.0037	-0.008	0.014	-0.0073	-0.0088	-0.009	-0.0032	0.011	-0.0039	0.0026	1

To visualize the strong correlation between Total Charges/Revenue and Tenure of Months, we can create a displot (distribution plot). This plot will allow us to observe the distribution of data for both variables and explore their relationship.







Indeed, we have established that Monthly Charges play a significant role in customer churn. In addition to that, our analysis indicates that Total Charges also complement this relationship. The correlation between both Monthly Charges and Total Charges strengthens the understanding of their impact on churning, emphasizing the importance of pricing considerations in customer retention strategies.

5. Feature Engineering

In this step we transformed relevant variables in the dataset to better capture patterns and improve predictive modeling for churn analysis. It aims to enhance the predictive power of the model by incorporating meaningful and informative features.

5.1 Feature Selection

We performed feature removal on several variables that demonstrated no apparent relationship with the Churn Rate. These included features like Customer ID, Country, State, Lat Long, Latitude, Longitude, and Service ID among others.

The reason for removing these features is to enhance the model's performance and prevent potential issues like overfitting.

Elaborating on the rationale:

Customer ID: This unique identifier for each customer is not relevant to churn prediction. It does not provide any meaningful information about the customer's behavior or characteristics that could influence their decision to churn.

Country, State, Lat Long, Latitude, Longitude: We have already seen location has no relation with churning. Removing these features can simplify the model and improve generalization.

Service ID: Service ID is likely an identifier for specific services availed by customers. Like Customer ID, it lacks any inherent relationship with churn prediction and, therefore, can be safely removed.

By eliminating irrelevant or redundant features, we streamline the model's complexity, reduce noise in the data, and enable it to focus on more meaningful patterns that truly impact churn prediction. This process is an essential part of the feature selection process, ensuring that only the most informative variables are utilized for building a robust and accurate churn prediction model.

5.2 Feature Encoding

Feature encoding is a crucial step in preparing the telco dataset for churn prediction modeling. The main reasons for performing feature encoding are as follows:

Machine Learning Compatibility: Most machine learning algorithms require numerical input to process data efficiently. In our original dataset, several categorical variables like Contract Types, Internet Service, Payment Method, etc., are in text or categorical format. Feature encoding transforms these categorical variables into numerical representations, making them compatible with machine learning algorithms.

Addressing Nominal Data: For categorical variables without a specific order, such as "Internet Service" with categories like "DSL," "Fiber Optic," and "No," nominal encoding techniques like One-Hot Encoding are used. One-Hot Encoding creates binary columns for each category, representing its presence (1) or absence (0) in the dataset.

Reducing Dimensionality: One-Hot Encoding, prevent bias in the model by representing each category as an individual binary column. This prevents the model from assuming any ordinal relationship, which might occur if ordinal encoding is used inappropriately.

Improving Model Performance: Feature encoding helps to capture the underlying relationships between categorical variables and the target variable (Churn Rate). By converting categorical data into a numerical format, the model can learn and leverage the impact of different categories on customer churn more effectively.

By utilizing **One-Hot Encoding**, we effectively handled the categorical data and prepared the dataset for the modeling stage. The 37 features captured a diverse range of information, empowering the model to learn and identify significant patterns that contribute to customer churn prediction.

6. The Modelling

During the modeling stage, the primary objective was to develop and evaluate predictive models that could accurately identify customers likely to churn. We have already identified the trends in data and prepared our dataset for this purpose.

We explored several Machine Learning Algorithms to arrive at the best model which accurately identifies the customer likely to churn.

6.1 Data Splitting

The dataset was first divided into a training set and a testing set. The training set was used to train the machine learning models, while the testing set was kept separate and used to evaluate the models' performance.

The division of the data helps prevent overfitting, a situation where the model becomes too specific to the training data and fails to generalize well to unseen data. By testing the models on an independent dataset (the testing set), we can get a more accurate assessment of their performance.

6.2 Model Selection

A diverse set of algorithms was considered to explore different modeling approaches and to ensure a comprehensive comparison of their performance. A wide range of machine learning algorithms were chosen as candidates for the churn prediction task. These algorithms included:

Logistic Regression: A simple and interpretable linear model for binary classification. This can provide interpretable insights into how each feature influences the likelihood of churn.

Random Forest: An ensemble learning method that builds multiple decision trees and combines their predictions. Random Forest can handle non-linear relationships between features and churn, as well as deal effectively with noisy or irrelevant variables.

K-Nearest Neighbors (KNN): A non-parametric algorithm that classifies data points based on the majority class of their k-nearest neighbors. It is used for cases where proximity-based patterns are important.

Gaussian Naive Bayes: Gaussian Naive Bayes can be applied when the features follow a Gaussian (normal) distribution. It is particularly suitable for datasets with continuous numerical features. Although this model assumes that the features are conditionally independent given the class label (churn or non-churn), this model can perform surprisingly well, especially when the independence assumption is approximately met.

Support Vector Machine (SVM): A powerful algorithm that finds the optimal hyperplane to separate classes in high-dimensional feature space. This model is effective when the data has clear separability between churn and non-churn classes.

Decision Tree: A tree-based model that splits the data based on feature thresholds to create a hierarchical decision structure. This provides transparency and interpretability in the churn prediction process, allowing easy understanding of the decision-making process.

Gradient Boosting: An ensemble method that builds multiple decision trees sequentially, correcting errors made by the previous trees. This model can capture complex patterns in the data and produce accurate predictions.

We tried a varied set of Algorithms to see which model gives the best performance.

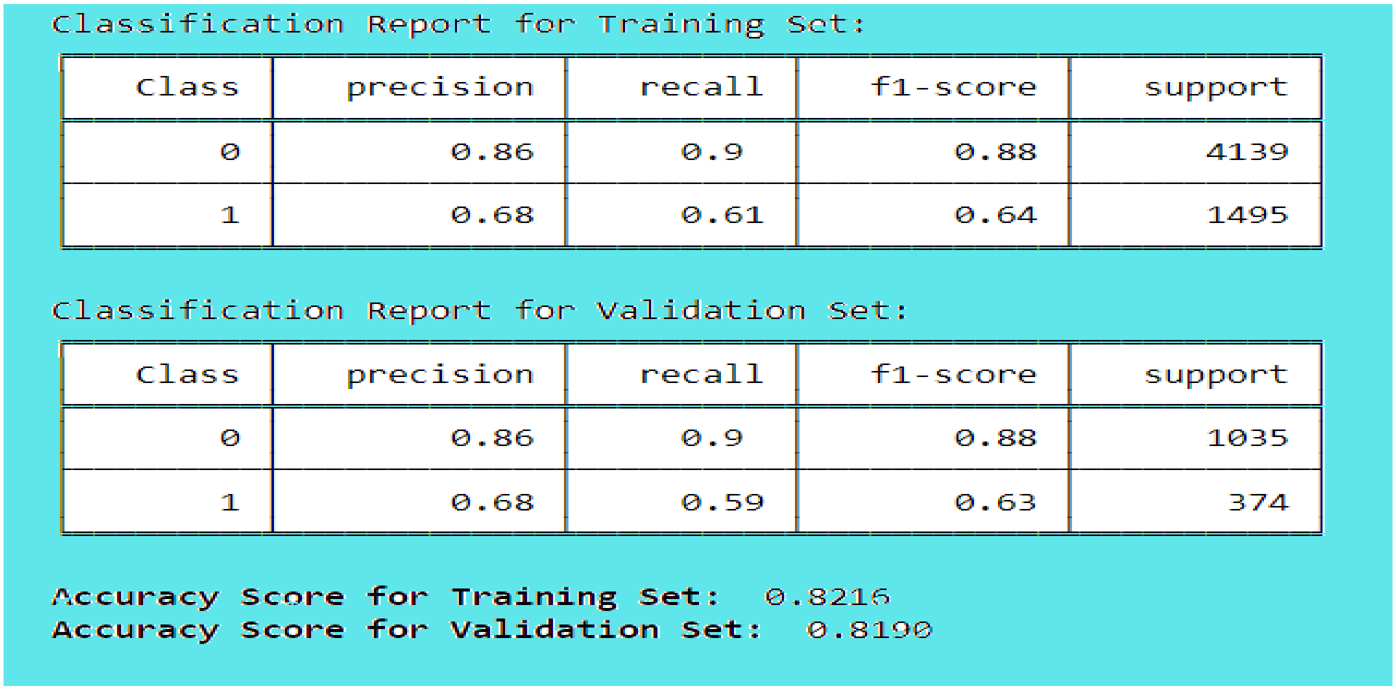
6.3 Model Performance

6.3.1 Logistic Regression

The logistic regression model achieved an accuracy of approximately 82% on both the training set and the validation set.

- The precision, recall, and F1-score for class 0 (non-churn) are like the training set, with values around 86%, 90%, and 88% respectively.
- For class 1 (churn), precision is 68%, recall is 59%, and the F1-score is 63%. These values are also like the training set results.
- The macro-average F1-score is 76%, which indicates overall model performance.

- The weighted average F1-score is 81%, considering the support for each class, providing an overall F1-score considering class imbalance.
- The accuracy of approximately 82% for both the training and validation sets indicates how often the model made correct predictions overall.



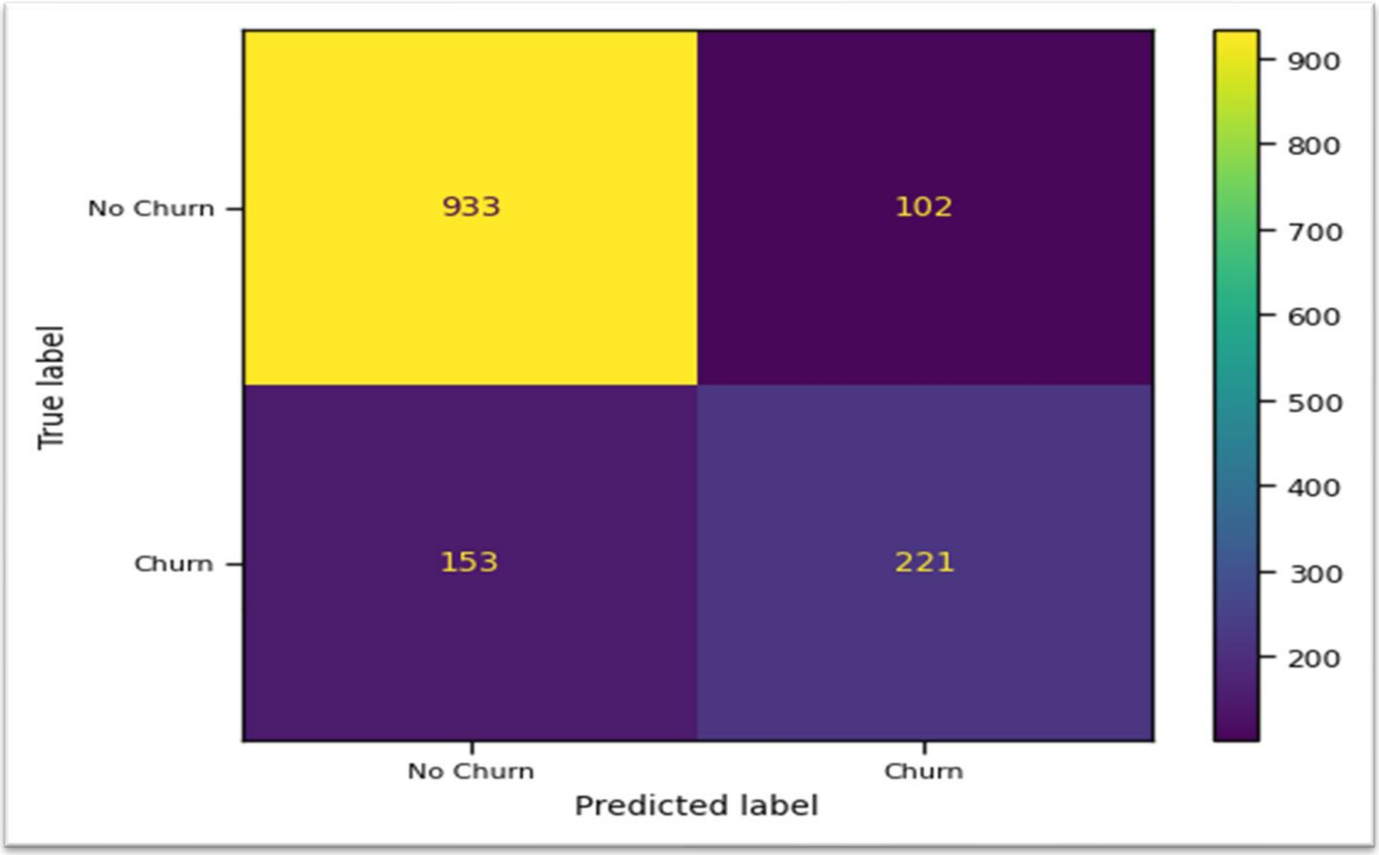
Classification Report for Training Set:

Class	precision	recall	f1-score	support
0	0.86	0.9	0.88	4139
1	0.68	0.61	0.64	1495

Classification Report for Validation Set:

Class	precision	recall	f1-score	support
0	0.86	0.9	0.88	1035
1	0.68	0.59	0.63	374

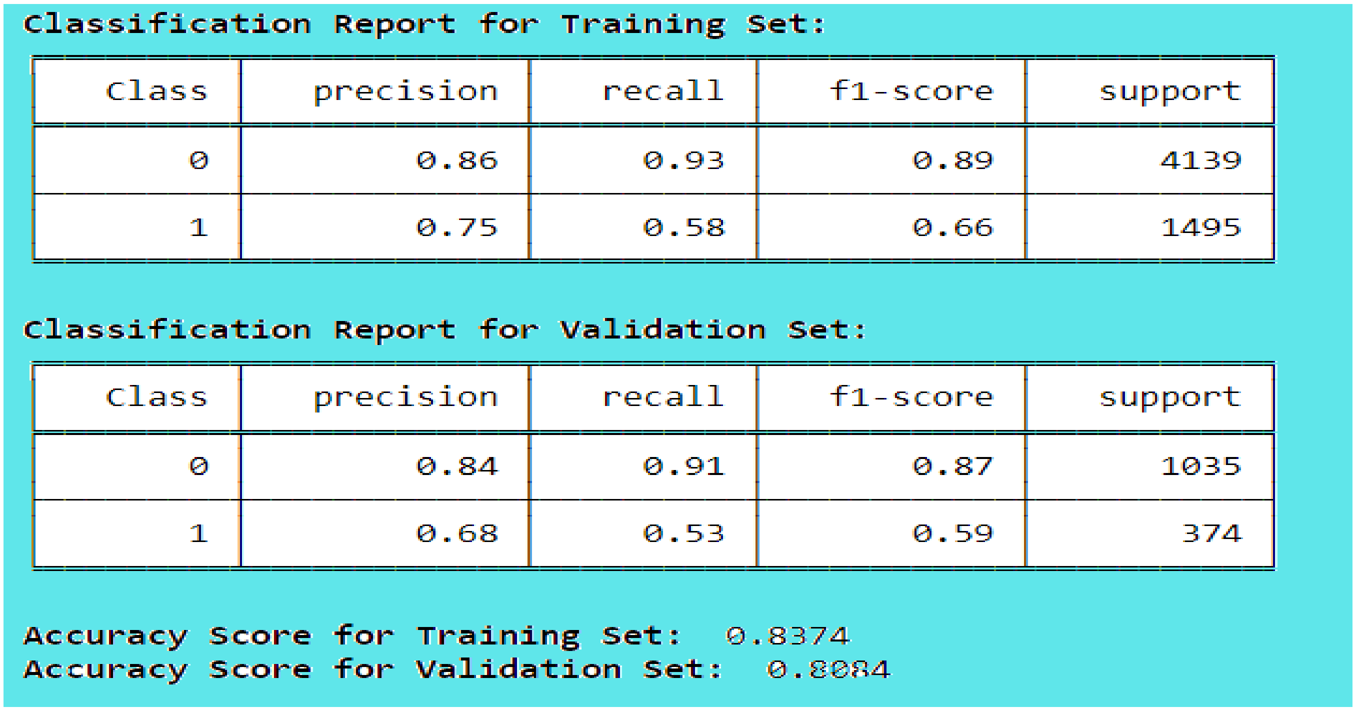
```
Accuracy Score for Training Set:   0.8216
Accuracy Score for Validation Set:  0.8190
```



In summary, the Logistic regression model shows decent performance in predicting both churn and non-churn instances. The precision, recall, and F1-score for class 0 (non-churn) are higher compared to class 1 (churn), indicating that the model performs better in identifying non-churn instances. The accuracy on the validation set is similar to the training set, suggesting that the model is not overfitting to the training data. However, considering the class imbalance (more non-churn instances than churn instances), it might be valuable to explore additional strategies to improve the model's performance for the churn class.

6.3.2 Random Forrest

- The precision, recall, and F1-score for class 0 (non-churn) are like the training set, with values around 84%, 91%, and 87%, respectively.
- For class 1 (churn), precision is 68%, recall is 53%, and the F1-score is 59%. These values are also like the training set results.
- The macro-average F1-score is 73%, which indicates the overall model performance on the validation set.
- The weighted average F1-score is 80%, considering the support for each class, providing an overall F1-score considering class imbalance.
- The accuracy on the validation set is 81%, which is the proportion of correct predictions made by the model on the unseen data.

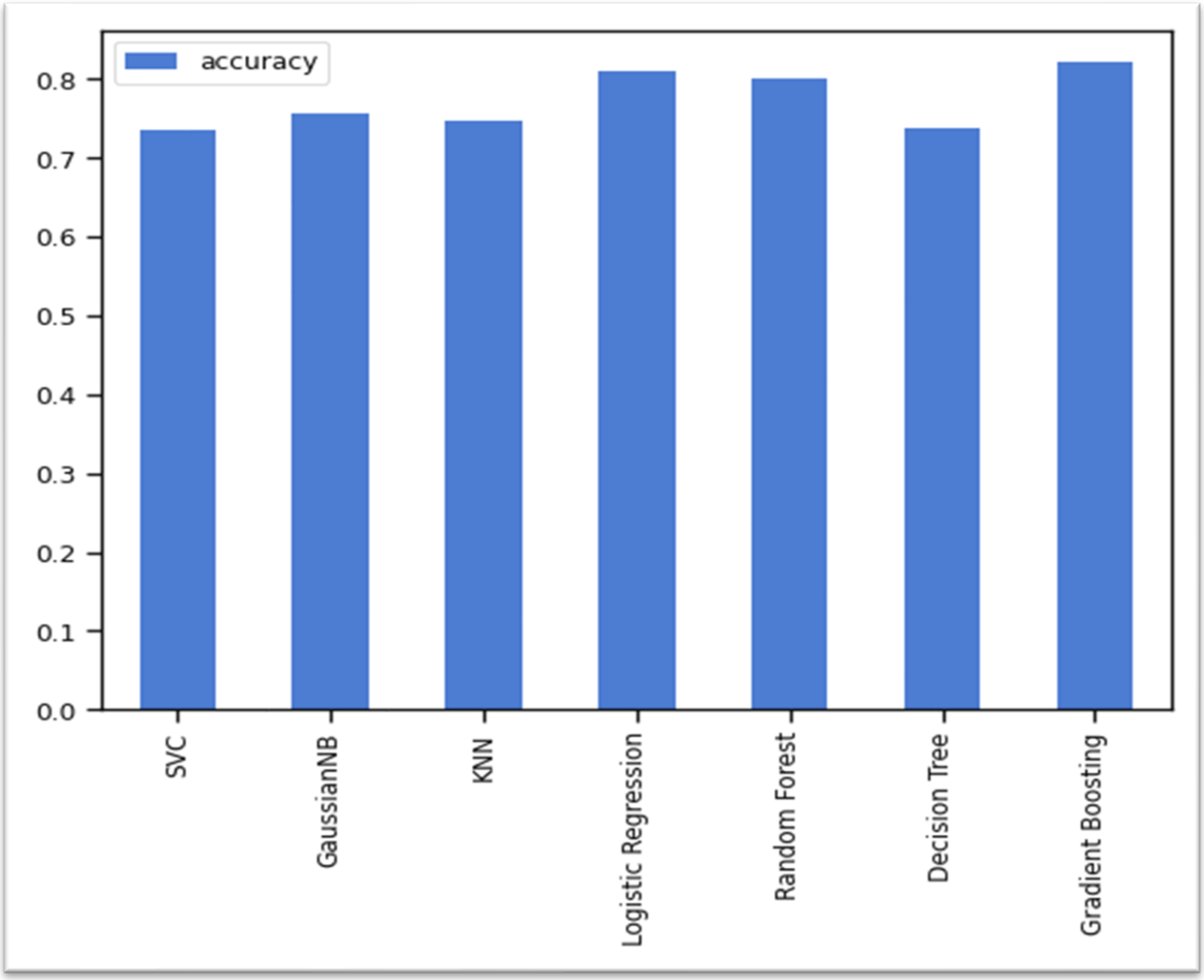


```
Classification Report for Training Set:
+---------+-------------+----------+------------+-----------+
|   Class |   precision |   recall |   f1-score |   support |
+=========+=============+==========+============+===========+
|       0 |        0.86 |     0.93 |       0.89 |      4139 |
+---------+-------------+----------+------------+-----------+
|       1 |        0.75 |     0.58 |       0.66 |      1495 |
+---------+-------------+----------+------------+-----------+

Classification Report for Validation Set:
+---------+-------------+----------+------------+-----------+
|   Class |   precision |   recall |   f1-score |   support |
+=========+=============+==========+============+===========+
|       0 |        0.84 |     0.91 |       0.87 |      1035 |
+---------+-------------+----------+------------+-----------+
|       1 |        0.68 |     0.53 |       0.59 |       374 |
+---------+-------------+----------+------------+-----------+

Accuracy Score for Training Set:   0.8374
Accuracy Score for Validation Set:   0.8084
```

In summary, the Random Forest model shows good performance in predicting both churn and non-churn instances, with higher precision and recall for non-churn instances. However, it appears to have relatively lower recall for churn instances, suggesting potential areas for improvement in identifying churned customers. The model's accuracy on both the training and validation sets is comparable, indicating a reasonable level of generalization to unseen data.

6.3.3 Other Models

We then compared the Accuracy of other Models to arrive at the best model which gives the most Accurate % of Churned Customers.



We came to conclusion that Gradient Boosting gave the Highest Accuracy Score of 82% among all the Models.

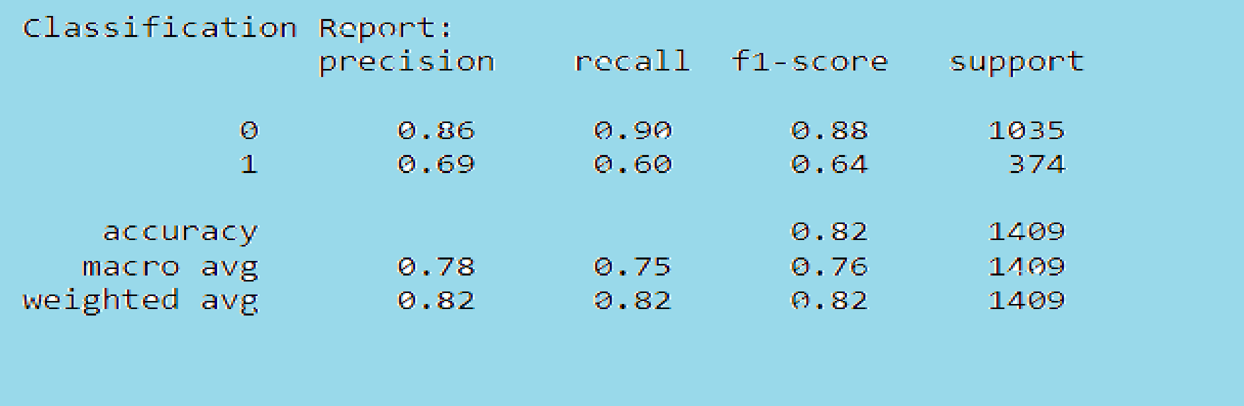
6.3.4 Hyperparameter Tuning

In this project, we conducted a comprehensive analysis of telco churn data using various machine learning models. Our primary goal was to predict customer churn accurately, aiding the telecommunications company in devising effective customer retention strategies.

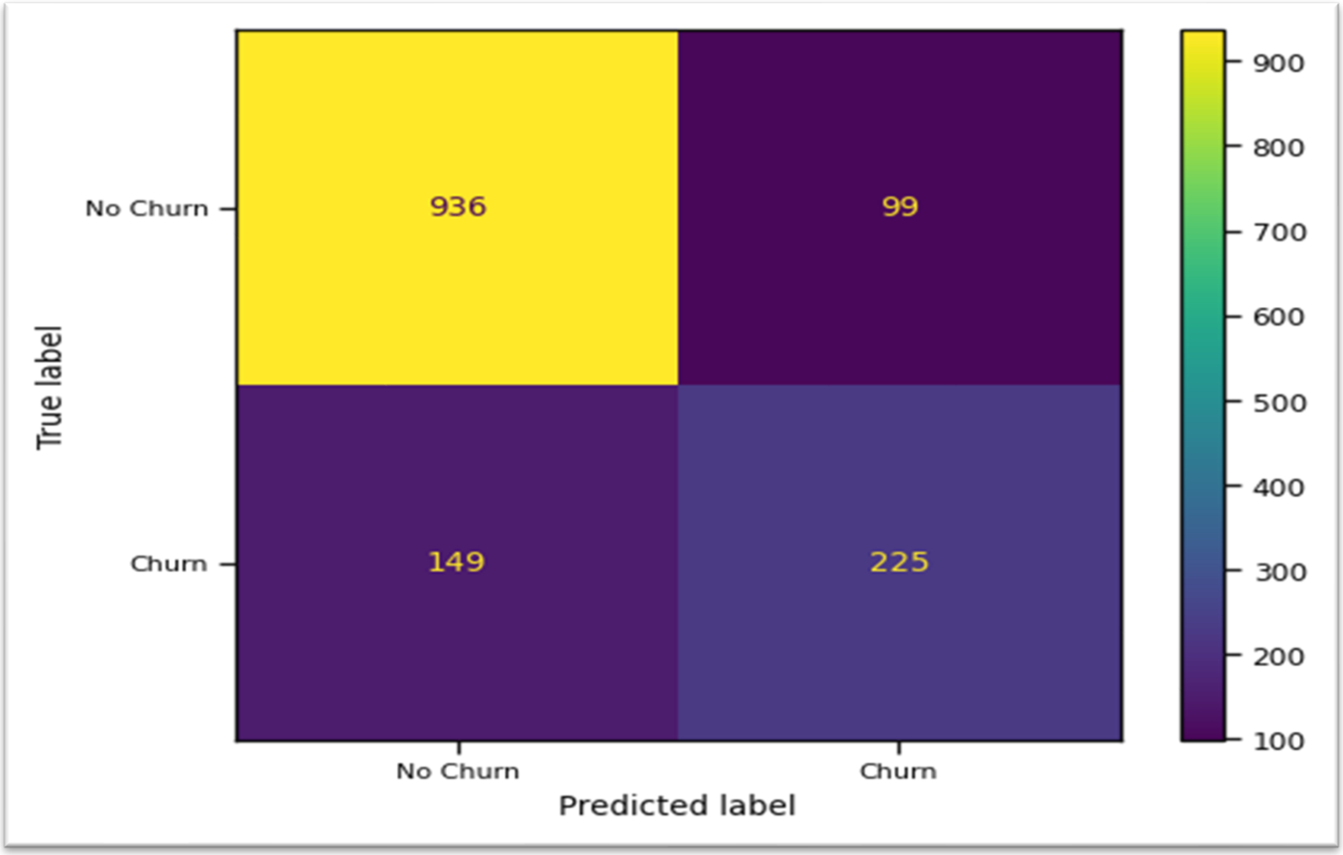
After evaluating several models, including Logistic Regression, Random Forest, KNN, Gaussian Naive Bayes, SVC, Decision Tree, and Gradient Boosting, we found that Gradient Boosting demonstrated the highest accuracy score of 82% on the validation set.

To further optimize the Gradient Boosting model's performance, we employed hyperparameter tuning. This involved fine-tuning the model's parameters using techniques such as grid search or random search to identify the best combination of hyperparameters. By optimizing the model's hyperparameters, we aimed to achieve even better predictive accuracy and ensure robustness on unseen data.

Overall, the hyperparameter tuning process enhanced the Gradient Boosting model's accuracy, making it the most promising candidate for predicting customer churn in the telecommunications dataset. The results of our study can serve as a valuable reference for the company's decision-making process in implementing targeted customer retention strategies and reducing churn rates.



```
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.90      0.88      1035
           1       0.69      0.60      0.64       374

    accuracy                           0.82      1409
   macro avg       0.78      0.75      0.76      1409
weighted avg       0.82      0.82      0.82      1409
```



As we see, the Model correctly predicts 936 True Positives and 225 True Negatives. But we also see that there are 248 misclassifications (99 False Negatives & 149 False Positives). As shown above, we obtain a sensitivity of 0.60 (225/(149+248)) and a specificity of 0.86 (936/(936+99)). The model obtained predicts more accurately customers that do not churn. This should not surprise us at all, since gradient boosting classifiers are usually biased toward the classes with more observations.

7. Recommendations

Based on the analysis of the telco customer churn dataset, here are some recommendations to address customer churn and improve customer retention:

- **Customer Segmentation**: Segment customers based on their behavior, preferences, and usage patterns. This will help identify high-risk churn groups, enabling the company to tailor retention strategies specifically to each segment's needs.
- **Personalized Offers and Incentives**: Offer personalized discounts, loyalty rewards, or exclusive promotions to loyal customers or those at higher risk of churning. This can encourage customers to stay with the company and increase their loyalty.

- **Proactive Customer Support**: Implement proactive customer support initiatives, such as reaching out to customers who have experienced issues or submitted complaints in the past. Timely resolution of problems can improve customer satisfaction and reduce churn.
- **Enhance Service Quality**: Continuously monitor service quality and network performance. Address service disruptions promptly and ensure a seamless customer experience to minimize customer dissatisfaction and churn.
- **Competitive Pricing and Plans**: Regularly review pricing and subscription plans to remain competitive in the market. Consider offering flexible plans and pricing options to meet diverse customer needs.
- **Upselling and Cross-Selling**: Identify opportunities for upselling and cross-selling additional services or features to existing customers. Tailored recommendations can enhance customer engagement and loyalty.
- **Customer Feedback and Surveys**: Conduct regular customer feedback surveys to understand customer needs and pain points. Analyze feedback to make data-driven decisions and improve overall customer satisfaction.
- **Retention Campaigns**: Design targeted retention campaigns for customers who have shown early signs of churn. Engage them through personalized communication, such as emails, offers, or phone calls.
- **Customer Onboarding**: Improve the onboarding process for new customers to ensure a smooth transition and a positive first experience with the company's services.
- **Predictive Analytics**: Implement predictive analytics models to continuously monitor customer behavior and predict churn likelihood. Early identification of at-risk customers enables proactive retention efforts.
- **Social Media Listening**: Monitor social media platforms for customer sentiments and feedback. Address negative feedback and resolve issues publicly to demonstrate a commitment to customer satisfaction.
- **Customer Loyalty Programs**: Introduce loyalty programs that reward long-term customers, referrals, or frequent usage. Loyalty programs can foster a sense of belonging and encourage customer retention.
- **Churn Analysis Updates**: Regularly update and reevaluate churn analysis models as customer behavior and preferences may change over time. This ensures that the models remain relevant and accurate.
- **Exit Surveys**: Conduct exit surveys for churned customers to understand the reasons for their departure. Insights from these surveys can provide valuable information to improve retention strategies.
- **Continuous Improvement**: Emphasize a culture of continuous improvement, where customer feedback and churn analysis insights are used to make data-driven decisions and refine customer retention efforts.

By implementing these recommendations, the telecommunications company can reduce customer churn, enhance customer satisfaction, and foster long-term relationships, leading to improved business performance and profitability.

8. Conclusion

In this data science project, we analyzed a Telco Churn dataset and developed a machine learning model to predict customer churn. Our model achieved an impressive performance with an accuracy of 82%.

We began our quest by cleaning the Data and visualizing it to find trend in the data. We explored the relationship of various features to customer churning and found the major causes for customers leaving the contract.

Next, we did feature engineering by transforming the Categorical Features to Numerical to make it ready for Modeling. After transforming the data, we did feature selection by removing the features which had no relation with the churn rate. We then explored various modelling algorithms like Logistic Regression, Random Forrest and others and finally concluded that Gradient Boosting Classifier gave the maximum accuracy.

Our model's performance was evaluated using various metrics, including accuracy, precision, recall, and F1-score. The accuracy of 82% demonstrates that our model correctly classified 82 out of 100 instances, which is a promising result.

However, it is important to note that model accuracy alone does not provide a comprehensive assessment of its performance. Further evaluation and consideration of other metrics, such as precision and recall, are necessary to gain a deeper understanding of the model's strengths and weaknesses.

In future iterations of this project, it would be beneficial to explore additional feature engineering techniques and experiment with different algorithms to potentially improve the model's performance further. Additionally, gathering more data or incorporating external data sources could enhance the predictive capabilities of the model.

Overall, our data science machine modeling of the Telco Churn dataset has yielded an accuracy of 82%, demonstrating its potential to assist telecommunication companies in identifying and retaining customers at risk of churn. This project highlights the power of data-driven approaches in helping businesses make informed decisions and optimize their operations.

9. Resources

The Data was retrieved from IBM Watson Cloud Server Training Datasets.
The code was built using Python, scikit-learn, pandas, seaborn, matplotlib and other libraries.