A logo for college computing

Description automatically generated

**Predicting Customer Churn and Enhancing Retention Strategies Through Machine Learning**

| *Swan Saung Lwin* |  |
| --- | --- |
| *2025282* |  |
| *Strategic Thinking* |  |
| *CA 2 – Capstone Report* |  |
| *Sunday, 18th May 2025 23:59* |  |
| *Friday, 16th May 2025* |  |



# Abstract

Customer Churn is a critical challenge faced by businesses across industries, especially in the digital market. Many companies struggle to predict customer churn accurately and have difficulties in carrying out effective retention strategies. Key challenges include ineffective traditional methods, lack of insights into impact of different services, generalized retention strategies, and the need to have cost-effective retention strategies. This project aims to predict customer churn using machine learning and identify the impact of key services offered by a telecommunication company.

The report contains a total of 2265 words from Section 1 to 9 excluding captions.

Both the code and report are host at <https://github.com/CCT-Dublin/ca1-capstone-project-proposal-swan-cct>

# Contents

[Table of Figures 4](#_heading=h.vnldta75exct)

[1. Strategic Overview of the Business Problem 6](#_heading=h.cersum6jhy0z)

[2. Project Plan 6](#_heading=h.cxop9cih540k)

[2.1. Project Scope & Prioritization 6](#_heading=h.l73codjoxyq3)

[2.2. Project Timeline 6](#_heading=h.4zkf6qiterjf)

[2.3. Models & Evaluation 7](#_heading=h.lkxp8gdtft1s)

[2.3.1. Models & Methods 7](#_heading=h.9imxo497e2li)

[2.3.2. Evaluation 7](#_heading=h.6nxc1u9h2yx)

[2.4. Challenges 7](#_heading=h.udcn17dkte2)

[3. Business Understanding 8](#_heading=h.ywi1t5rxoj6l)

[4. Data Understanding 8](#_heading=h.nv1z55kfvu4y)

[4.1. About Datasource 8](#_heading=h.t2t2y7xzq421)

[4.2. Exploratory Data Analysis 8](#_heading=h.x4g40vuohtei)

[4.2.1. Demographic 8](#_heading=h.z8vjo4qaqk46)

[4.2.2. Location 10](#_heading=h.sv8ncfc8yu9z)

[4.2.3. Population 11](#_heading=h.u7sj0ke3jii3)

[4.2.4. Services 12](#_heading=h.7ywgexw6x0zm)

[4.2.5. Churn Status 16](#_heading=h.87mmdponi472)

[5. Data Preparation 19](#_heading=h.n6di2y49gg7i)

[5.1. Merging Datasets 19](#_heading=h.x2yf97m8mrxr)

[5.2. Label Encoding 19](#_heading=h.73h9mjrujgrk)

[5.3. Customer Segmentation 19](#_heading=h.ad67tx1nvk06)

[5.3. Hypothesis Testing 22](#_heading=h.pyrkfsau5w9)

[5.3.1. Does longer contract length reduce churn? 22](#_heading=h.nrbs8ecwuea7)

[5.3.2. Does more service usage affect churn? 22](#_heading=h.hxw6w81fia8q)

[5.4. Feature Selection 23](#_heading=h.3aush0g0on2o)

[5.5. Splitting the data 24](#_heading=h.o4l2wca36vqp)

[5.6. Feature Scaling 24](#_heading=h.rix898y52saz)

[5.7. Dimensionality Reduction 24](#_heading=h.3srhbg2ljjt9)

[5.8. Class Balancing 25](#_heading=h.1s2bfdjwzb2d)

[6. Machine learning implementation 25](#_heading=h.kr58l442mce4)

[6.1. Model Selection 25](#_heading=h.r829qiyyzb1e)

[6.2. Cross Validation 26](#_heading=h.n6v9nxe5iimv)

[6.3. Hyperparameter Tuning 26](#_heading=h.b7hao0cto1cy)

[6.3.1. Logistic Regression 26](#_heading=h.8lpoohe68igx)

[6.3.2. Random Forest 26](#_heading=h.r290kbkr32hy)

[6.3.3. SVM 27](#_heading=h.nkuhzfit6upy)

[6.4. Evaluation 27](#_heading=h.f4bh41i50nek)

[7. Findings 28](#_heading=h.gk4e84wqmwki)

[8. Conclusions 28](#_heading=h.et5fmooej3uh)

[9. Any Future Recommendations 28](#_heading=h.x3gt7dihj1l4)

[References 29](#_heading=h.5akkuwhp012l)

# Table of Figures

| **Figure** | **Caption** | **Page** |
| --- | --- | --- |
| 1 | Project scope & task prioritization | 6 |
| 2 | CRISP-DM process diagram (Wikipedia Contributors, 2019) | 6 |
| 3 | High level project timeline for Customer Churn Prediction | 7 |
| 4 | First 5 rows of demographic dataset | 8 |
| 5 | Demographic dataset info | 9 |
| 6 | Demographic dataset statistics | 9 |
| 7 | Number of Dependents distribution | 9 |
| 8 | Gender count comparison | 10 |
| 9 | First 5 rows of location dataset | 10 |
| 10 | Location dataset info | 10 |
| 11 | Top 10 cities by customer population | 11 |
| 12 | First 5 rows of population dataset | 11 |
| 13 | Population dataset info | 11 |
| 14 | Population dataset statistics | 12 |
| 15 | Population distribution | 12 |
| 16 | Population box plot | 12 |
| 17 | First 5 rows of services dataset | 12 |
| 18 | Services dataset info | 13 |
| 19 | Services dataset statistics | 13 |
| 20 | Services dataset boxplots | 14 |
| 21 | Services usage pie charts | 15 |
| 22 | Customers by Contract Type | 15 |
| 23 | Customers by Internet Type | 16 |
| 24 | Customers by Payment Method | 16 |
| 25 | First 5 rows of churn status dataset | 16 |
| 26 | Churn status dataset info | 17 |
| 27 | Churn status dataset statistics | 17 |
| 28 | Customer churn status | 17 |
| 29 | Customer lifetime value distribution | 17 |
| 30 | CLTV vs Churn Score | 18 |
| 31 | Churn Category | 18 |
| 32 | Churn Reasons | 18 |
| 33 | Merged churn dataset | 19 |
| 34 | Elbow method for customer segmentation | 20 |
| 35 | Service usage by customer segment | 21 |
| 36 | Correlation between services and customer satisfaction | 22 |
| 37 | Hypothesis testing on if longer contract length reduce churn | 22 |
| 38 | Hypothesis testing on if more service usage affect churn | 23 |
| 39 | Unwanted columns with reasons | 23 |
| 40 | Feature Importance (Information Gain) | 23 |
| 41 | Feature Importance (Fisher Score) | 23 |
| 42 | Dataset after feature scaling | 24 |
| 43 | Principal component analysis | 24 |
| 44 | First 5 rows of X\_train after PCA | 24 |
| 45 | Churn Class Comparison | 25 |
| 46 | Training data observations after oversampling | 25 |
| 47 | Cross-validation results | 26 |
| 48 | Logistic regression hyperparameters | 26 |
| 49 | Random forest hyperparameters | 27 |
| 50 | SVM hyperparameters | 27 |
| 51 | Best Hyperparameters | 27 |
| 52 | Models testing performance | 27 |
| 53 | Models confusion matrix | 28 |

# 1. Strategic Overview of the Business Problem

In recent years, digital businesses have experienced rapid growth, leading to fierce competition among companies. As a result, sustaining a loyal customer base becomes more important than ever as acquiring a new customer can cost 5 to 25 times more than keeping an existing customer (Gallo, 2014). Dewan et al. (2020) found that effective engagement strategies, such as personalized communication and tailored product offerings, significantly enhance customer satisfaction and retention rates. Understanding how different services impact customer satisfaction can help businesses focus their efforts on high-impact areas, thereby improving overall retention.

# 2. Project Plan

## 2.1. Project Scope & Prioritization

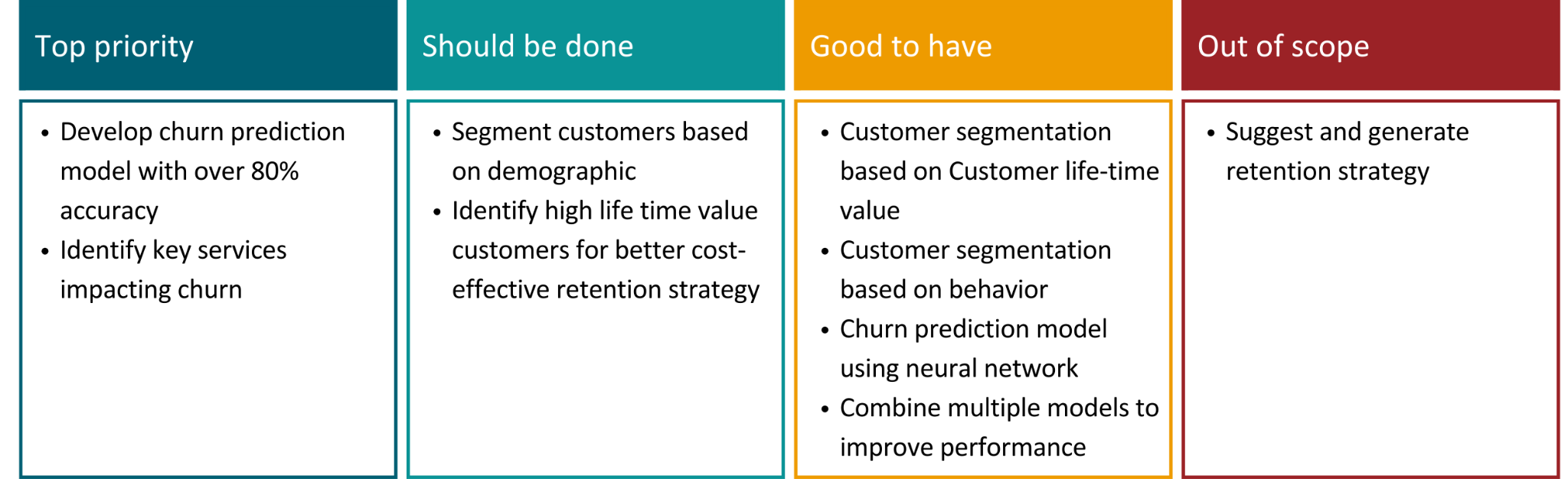
This project focuses on predicting customer churn using machine learning and identifying the impact of key services offered by a telecommunication company. 

Figure 1. Project scope & task prioritization

## 2.2. Project Timeline

This project follows the CRISP-DM methodology, which provides a structured approach to data-mining.

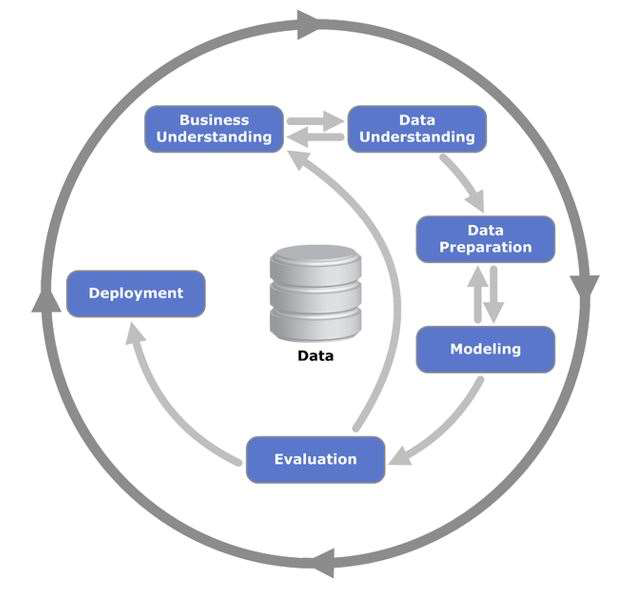


Figure 2. CRISP-DM process diagram (Wikipedia Contributors, 2019)

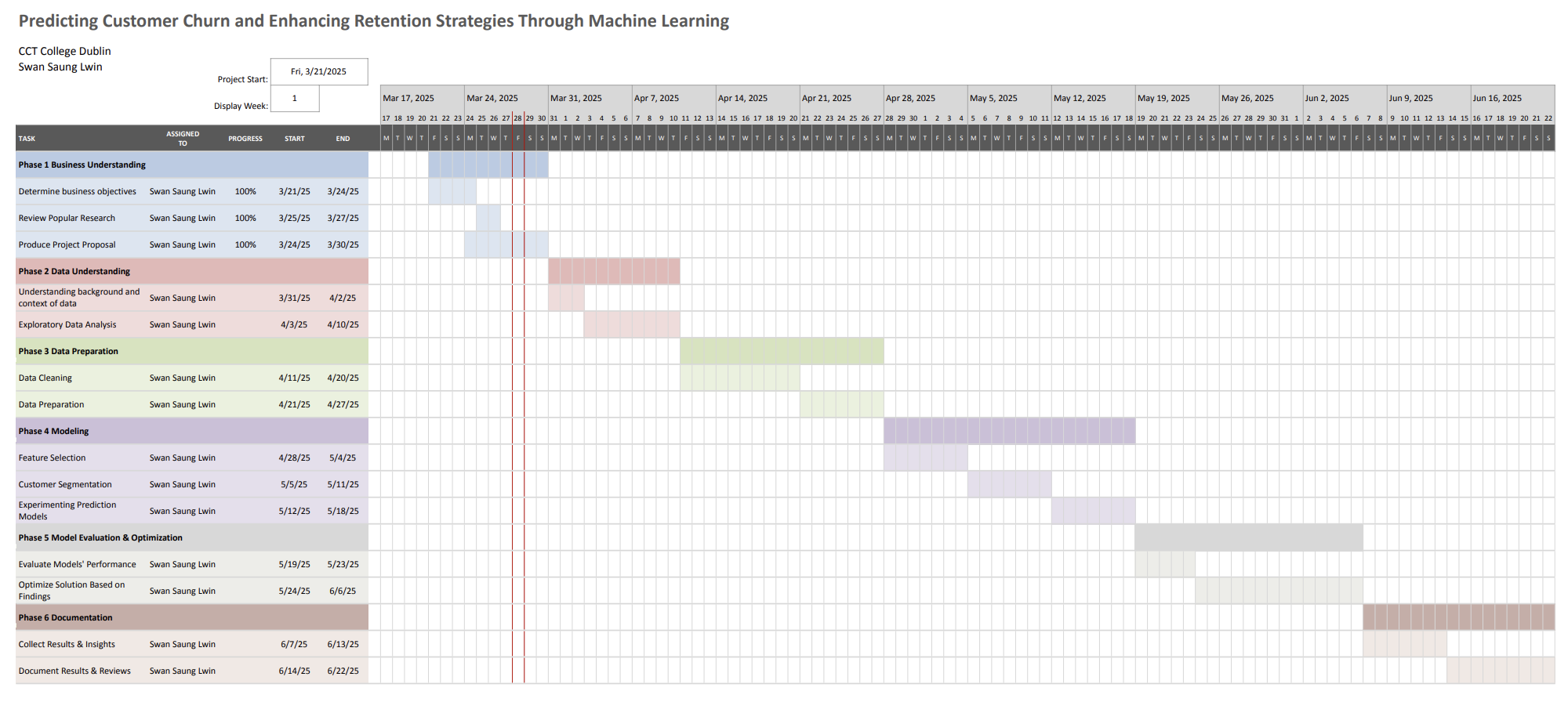
I have divided the project into 6 phases, according to the CRISP-DM methodology. However, the last deployment stage is replaced with documentation as the project will not be deployed.

Figure 3. High level project timeline for Customer Churn Prediction

## 2.3. Models & Evaluation

### 2.3.1. Models & Methods

For predicting churn, I will be using supervised learning models. Models include

* Logistic Regression
* K-Nearest Neighbours
* Random Forest
* XGBoost
* Gaussian Naive Bayes
* Support Vector Machine

For customer segmentation, I will be using K-means unsupervised learning model for demographic segmentation.

To identify most influential factors and services impacting churn, I will be using methods such as

* Correlation
* Information Gain
* Fisher Score

### 2.3.2. Evaluation

* **Supervised models:** Accuracy, Precision, Recall, F1-Score, Confusion Matrix
* **Customer segmentation:** Elbow method, Distortion score

## 2.4. Challenges

* Have to perform EDA on 5 datasets
* Churn was under-represented, having to resample data
* Needed to remove redundant and unimportant data via feature selection
* Ensuring results were understandable for business users

# 3. Business Understanding

Analyzing customer information in the telecommunication industry is similar to other service providing industries, in metrics such as billing information, service usage, demographics and customer information. For this project, I will focus on analyzing data and building a churn prediction model for a telecom company which provides home phone and Internet services to customers in California.

The key objectives are

* Develop an accurate churn prediction model
* Implement customer segmentation
* Identify the factors that influence customer churn and customer lifetime value
* Perform hypothesis testing to answer intriguing questions

# 4. Data Understanding

## 4.1. About Datasource

The Telco customer churn data contains information about a fictional telco company that provided home phone and Internet services to 7043 customers in California in Q3 (IBM, 2019). Following datasets are used in this project.

* Demographic
* Location
* Population
* Services
* Churn Status

## 4.2. Exploratory Data Analysis

### 4.2.1. Demographic

Each customer is uniquely identified by Customer ID. Under 30, and Dependents columns are already well represented by their numerical counterparts, and the Count column is unnecessary so I removed them.

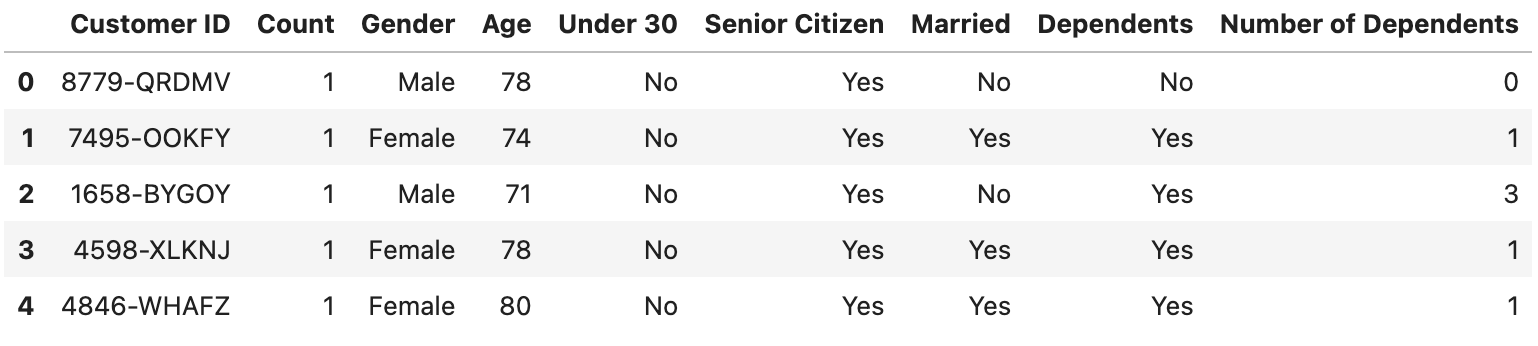


Figure 4. First 5 rows of demographic dataset

The dataset doesn’t seem to have either wrong data types, duplicated values or missing values. It has exactly 7043 observations as described in the data source. However, some of the status columns are represented by Yes/No. I will replace them with 1/0 as the models cannot understand text. By looking at the statistics, Age doesn't seem to have any outliers, while the Number of Dependents column seems to be sparse, have outliers and possibly right skewed.

|  |  |
| --- | --- |
| Figure 5. Demographic dataset info | Figure 6. Demographic dataset statistics |

I have plotted the distribution for the Number of Dependents as shown in Figure 7 and found my assumption is correct.

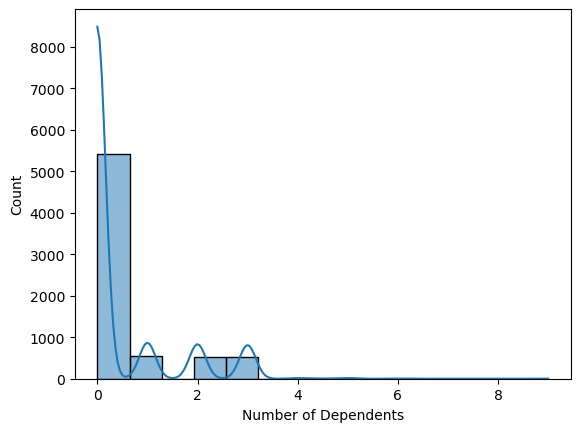


Figure 7. Number of Dependents distribution

To ensure the model doesn’t get influenced by these outlier values, I checked the value counts of each value and discretized the column by replacing the outliers with the upper bound value, which is 3. Male, Female ratio in the dataset is also roughly the same with 50.5% and 49.5% respectively.

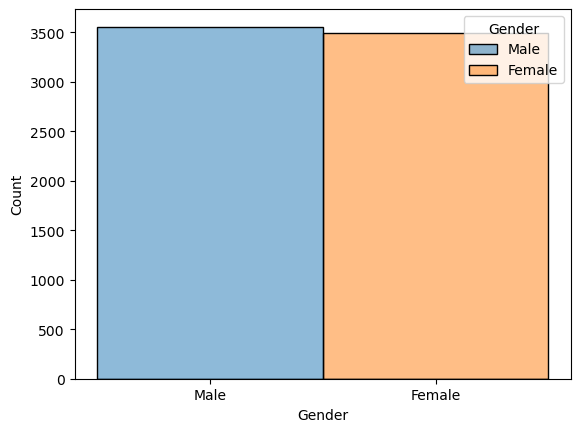


Figure 8. Gender count comparison

### 4.2.2. Location

Country and State seem to contain only one value so I confirm it by checking unique values and then remove them. I removed Location ID, Count, and Lat Long columns as they are not important in prediction. There are neither null values nor duplicate values in the dataset.

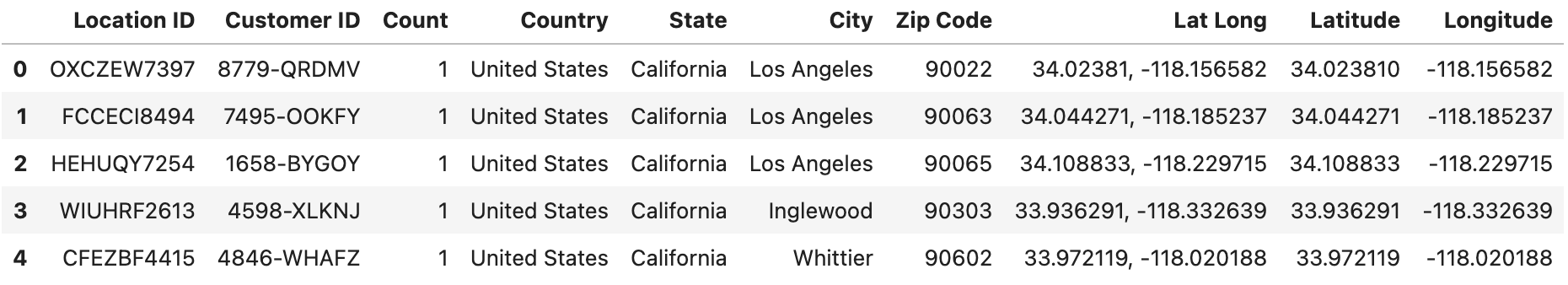


Figure 9. First 5 rows of location dataset

The dataset mostly has discrete values. As I learned about top cities by customer population, I found that Los Angeles and San Diego are the top cities with almost 300 customers each. They have almost triple the population than the rest of the top 10 cities.

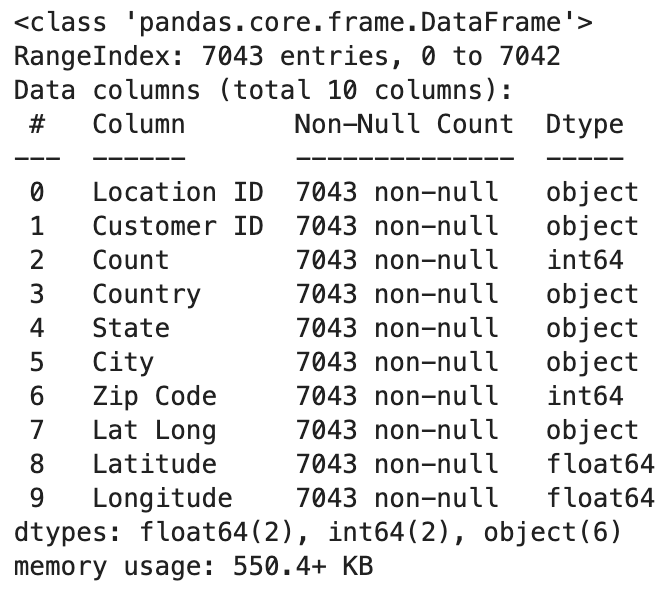


Figure 10. Location dataset info

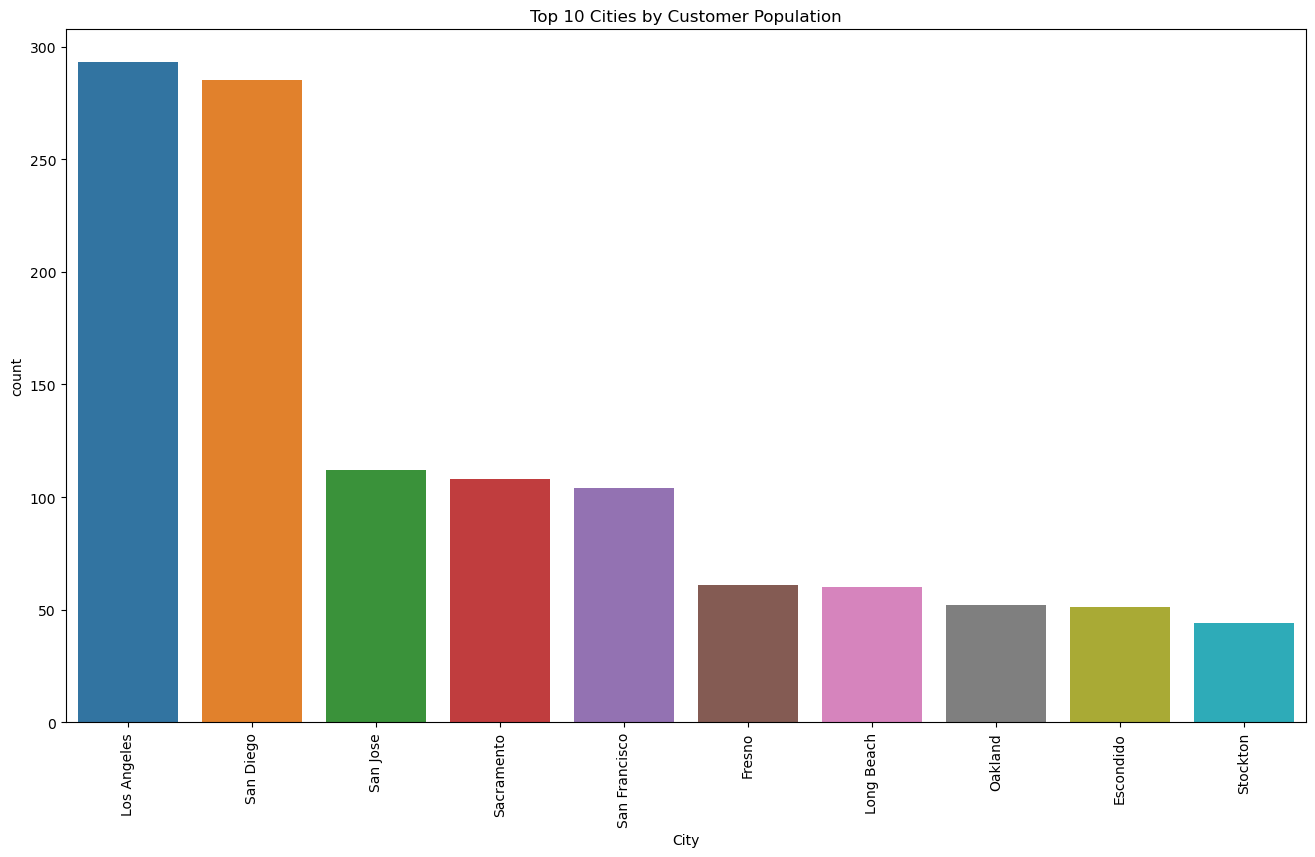


Figure 11. Top 10 cities by customer population

### 4.2.3. Population

There are no null values nor duplicate values in the dataset. After checking info about the dataset, I correct the population column by removing comma separators. I also removed the ID column as it is not needed.

|  |  |
| --- | --- |
| Figure 12. First 5 rows of population dataset | Figure 13. Population dataset info |

After checking the descriptive statistics, Population seems to have very high variance with a high number of outliers.

|  |  |
| --- | --- |
| Figure 14. Population dataset statistics | Figure 15. Population distribution |

I confirmed my assumption by creating a box plot and discovered that there’s a high number of outliers in the upper end.

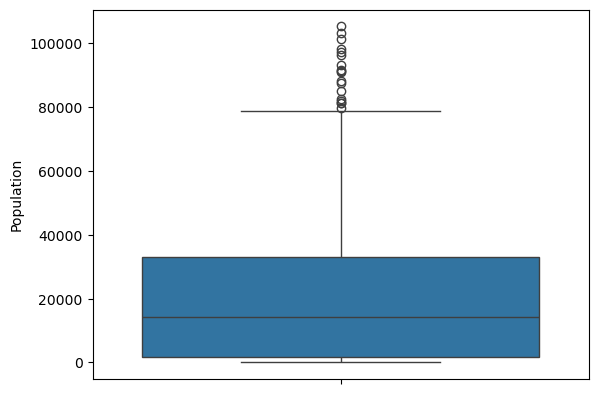


Figure 16. Population box plot

### 4.2.4. Services

Services dataset contains information about service usage and bill information of each customer.

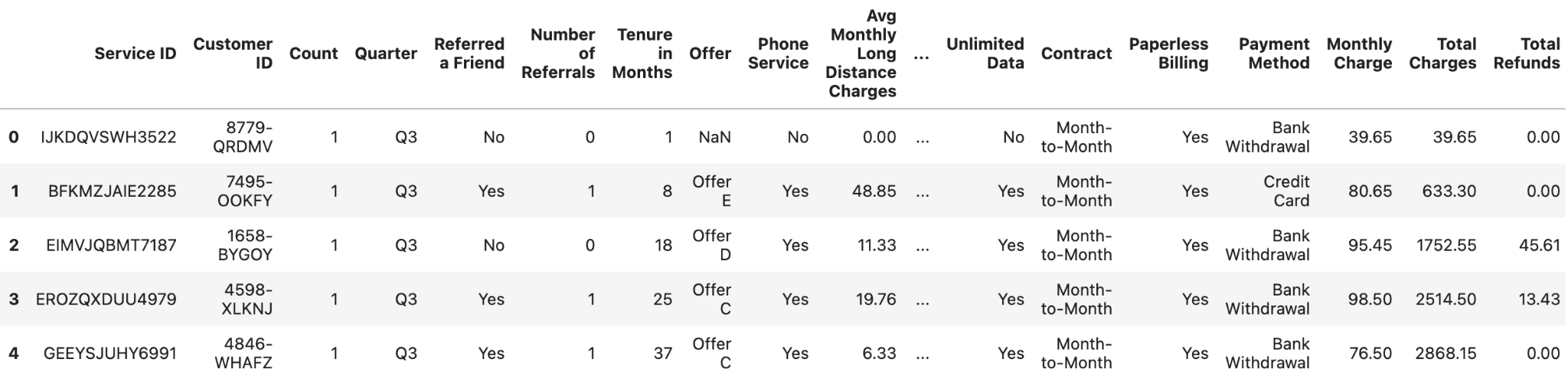


Figure 17. First 5 rows of services dataset

There are 31 columns in this dataset. There are no missing values except Offer and Internet Type. For Internet Type, the missing value is present when a customer is not using Internet Service. Therefore all the missing values are replaced with “None”. Offer column represents the offer a customer has taken. It is dropped as it has 5 unique values with about 5000 missing values. Other redundant columns are also dropped. Yes/No status values are also transformed using the same function applied in the Location dataset.

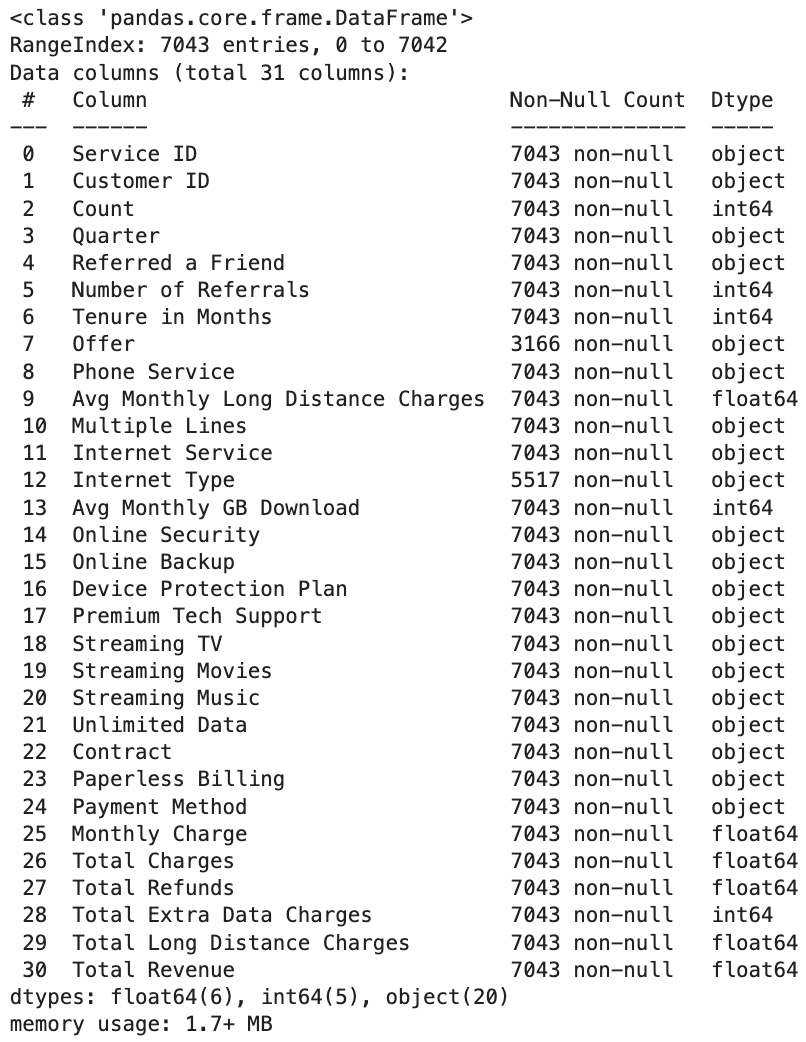


Figure 18. Services dataset info

By looking at the statistics, columns such as Number of Referrals, Avg Monthly GB Download, Total Refunds, Total Extra Data Charges and Total Long Distance Charges seem to contain a high number of outlier values.

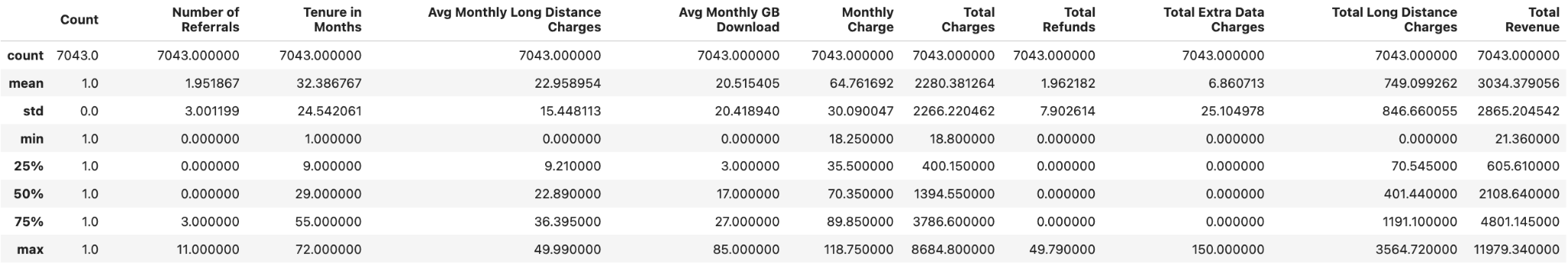


Figure 19. Services dataset statistics

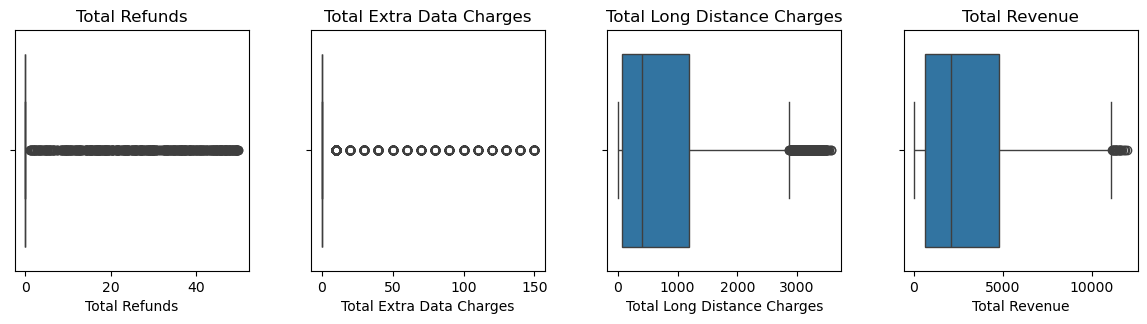
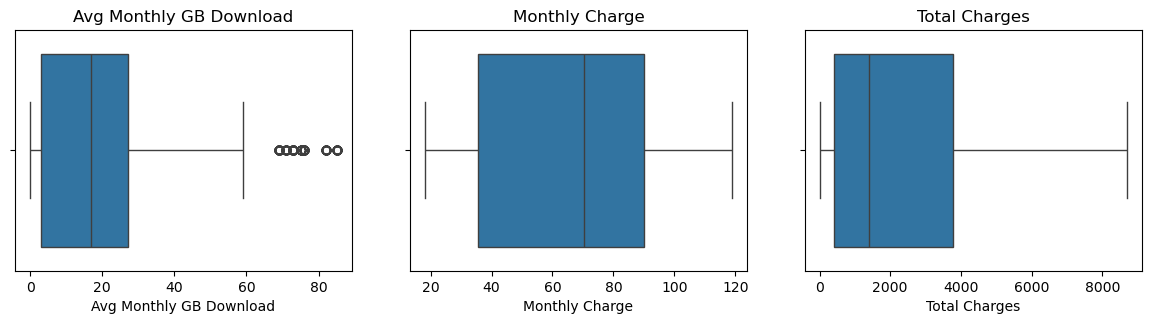
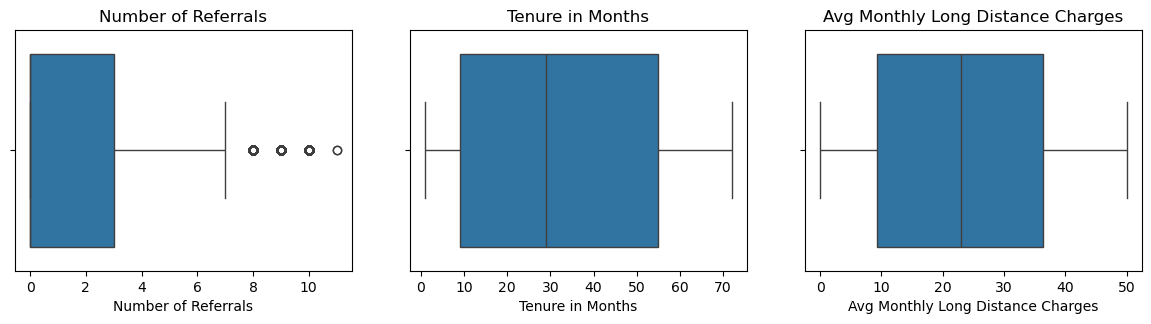


Figure 20. Services dataset boxplots

Service usage percentages are depicted as pie charts as shown in Figure 21. Internet Types, Contract Types, and Payment Methods are also compared using bar charts in Figure 22, 23, and 24.

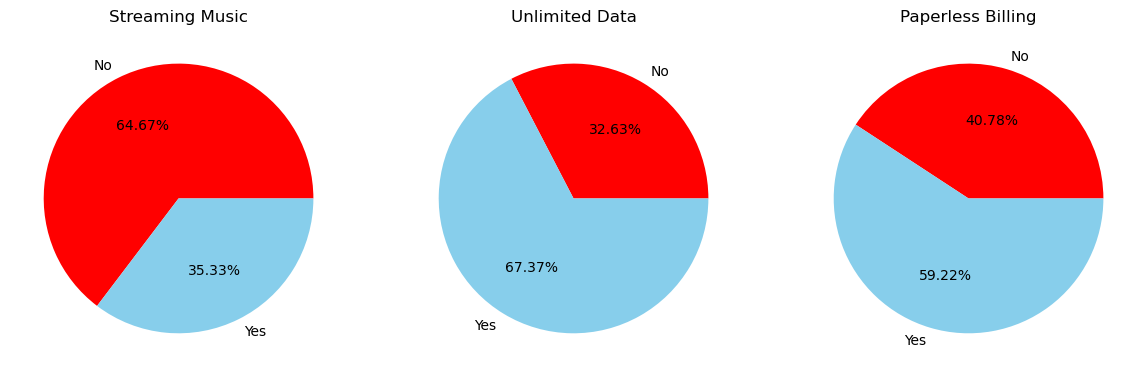
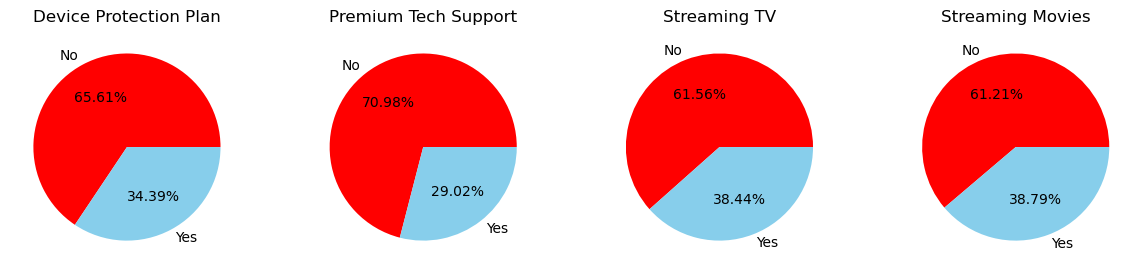
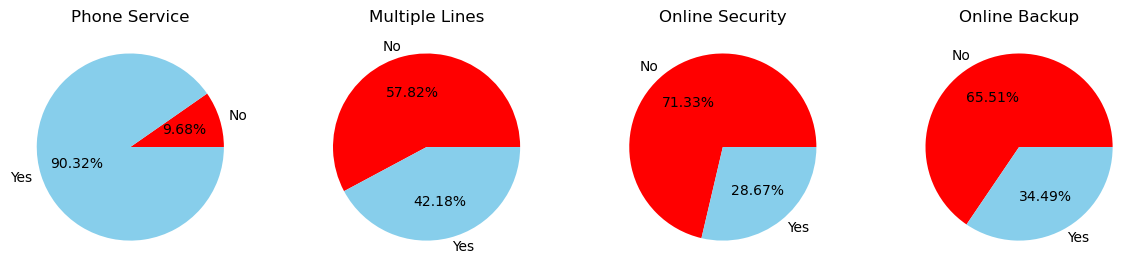


Figure 21. Services usage pie charts

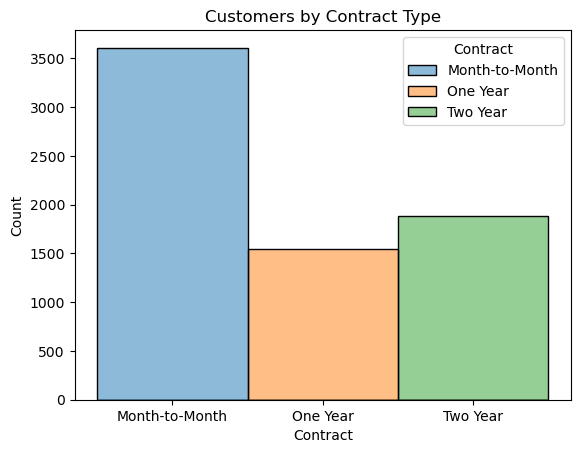


Figure 22. Customers by Contract Type

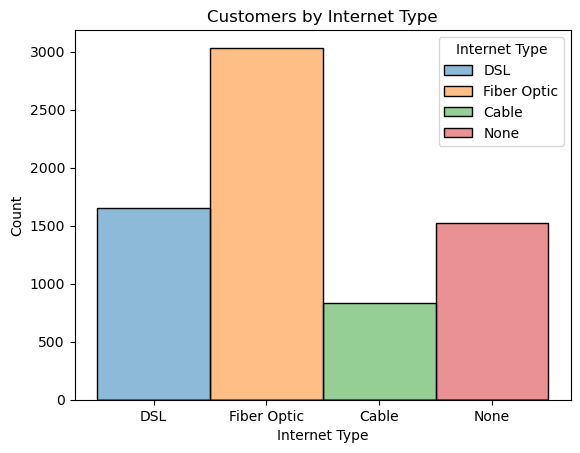


Figure 23. Customers by Internet Type

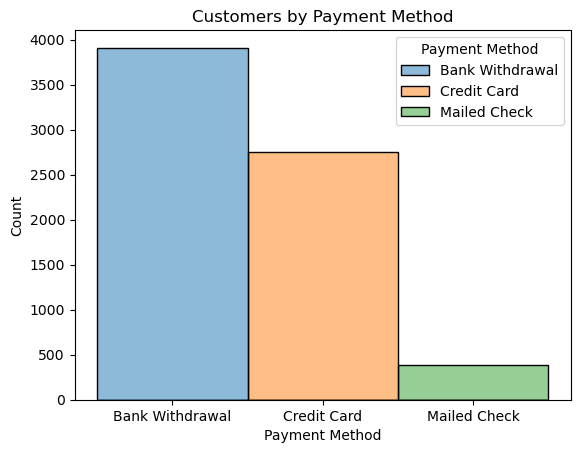


Figure 24. Customers by Payment Method

### 4.2.5. Churn Status

Churn status dataset contains various customer information. Redundant columns such as Status ID, Count, and Quarter are dropped.

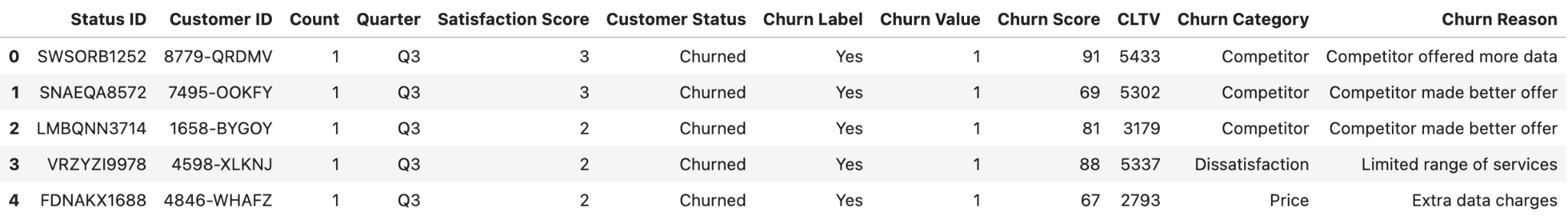


Figure 25. First 5 rows of churn status dataset

There are no missing values in the dataset except Churn Reason and Churn Category columns. Upon investigation, the values are missing when a customer is not churn. Most columns seem to be normally or uniformly distributed on looking at the statistics.

|  |  |
| --- | --- |
| Figure 26. Churn status dataset info | Figure 27. Churn status dataset statistics |

As shown in Figure 29, churn occurs in all ranges of CLTV but has a higher percentage in lower lifetime value. Also customers become churn when the churn score is 60 and above.

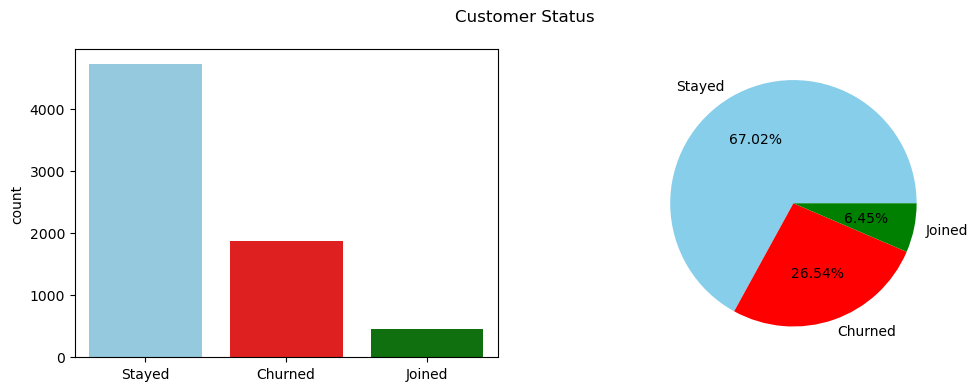


Figure 28. Customer churn status

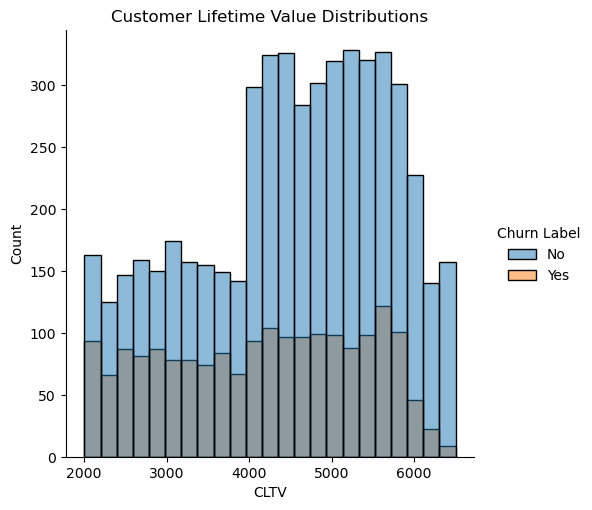


Figure 29. Customer lifetime value distribution

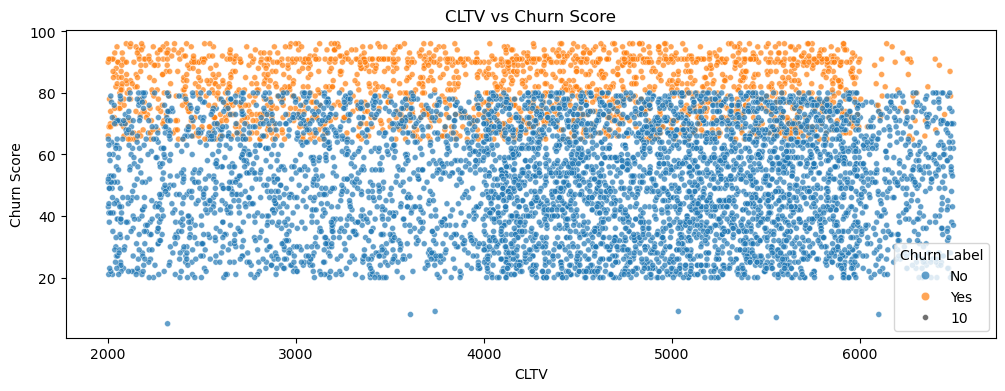


Figure 30. CLTV vs Churn Score

Top churn reasons and churn categories are identified as shown in Figure 31 and 32.

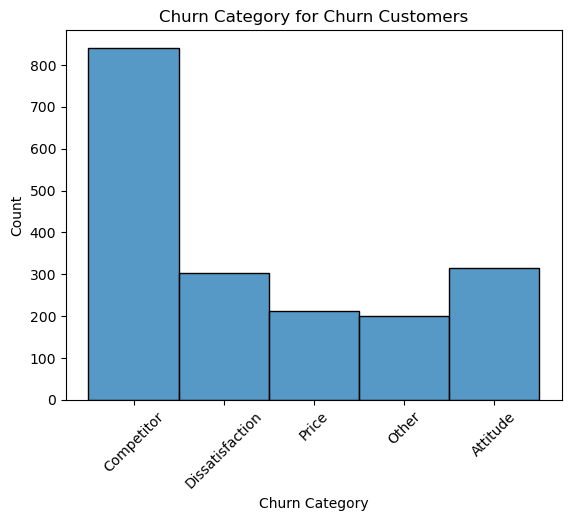


Figure 31. Churn Category

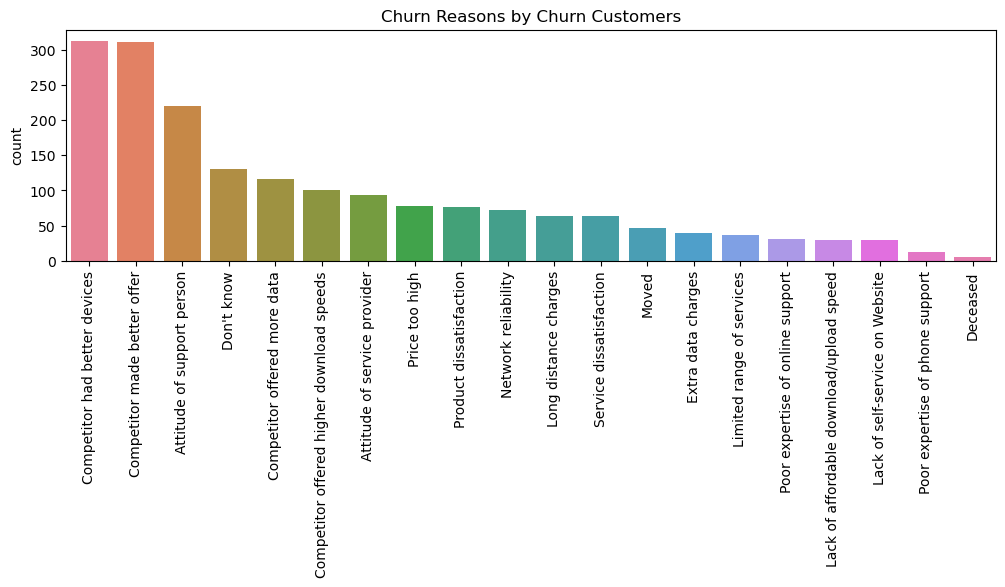


Figure 32. Churn Reasons

# 5. Data Preparation

## 5.1. Merging Datasets

The datasets are merged on the common unique key, Customer ID.

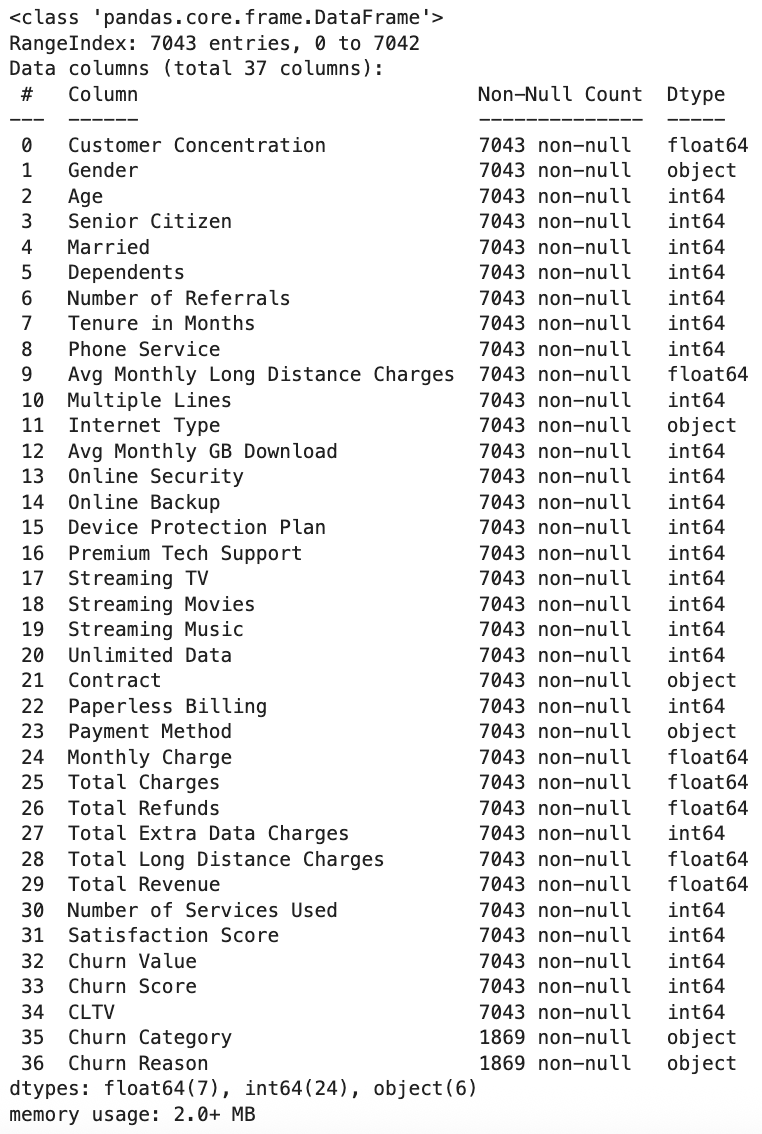


Figure 33. Merged churn dataset

## 5.2. Label Encoding

Object type features are encoded using the label encoder instead of other methods as they have low cardinality and don't create too much of an order.

## 5.3. Customer Segmentation

Customers are segmented based on demographics using K-Means clustering. Optimal k-value is chosen using the elbow method. For each k value, WCSS is calculated and identifies the point where the reduction in WCSS starts to diminish, forming an elbow-like shape. (Sankalana, 2023). Customers are then clustered according to optimal k-value, 4.

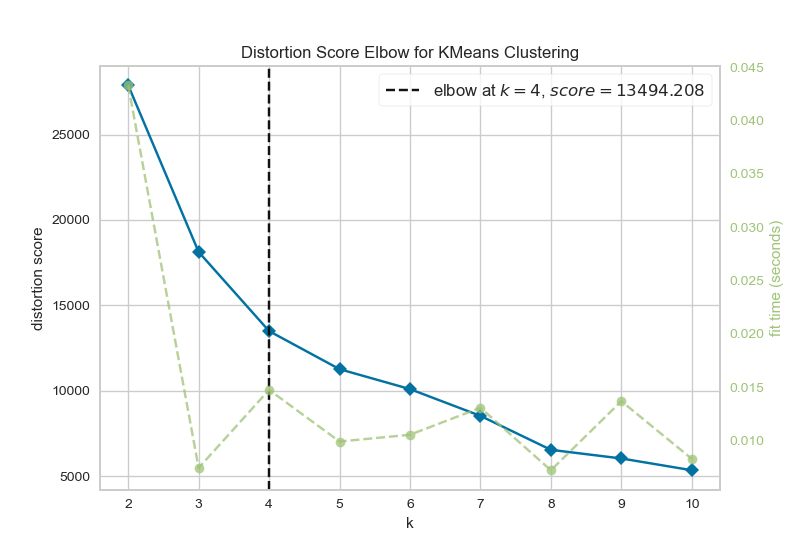
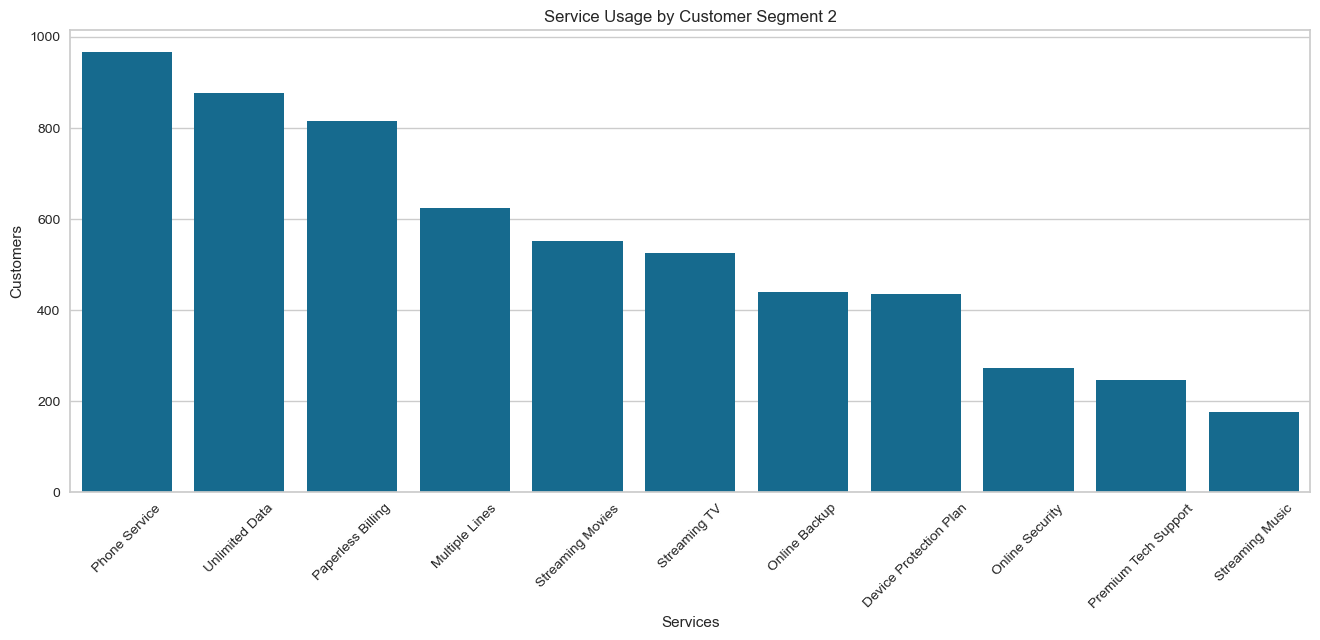
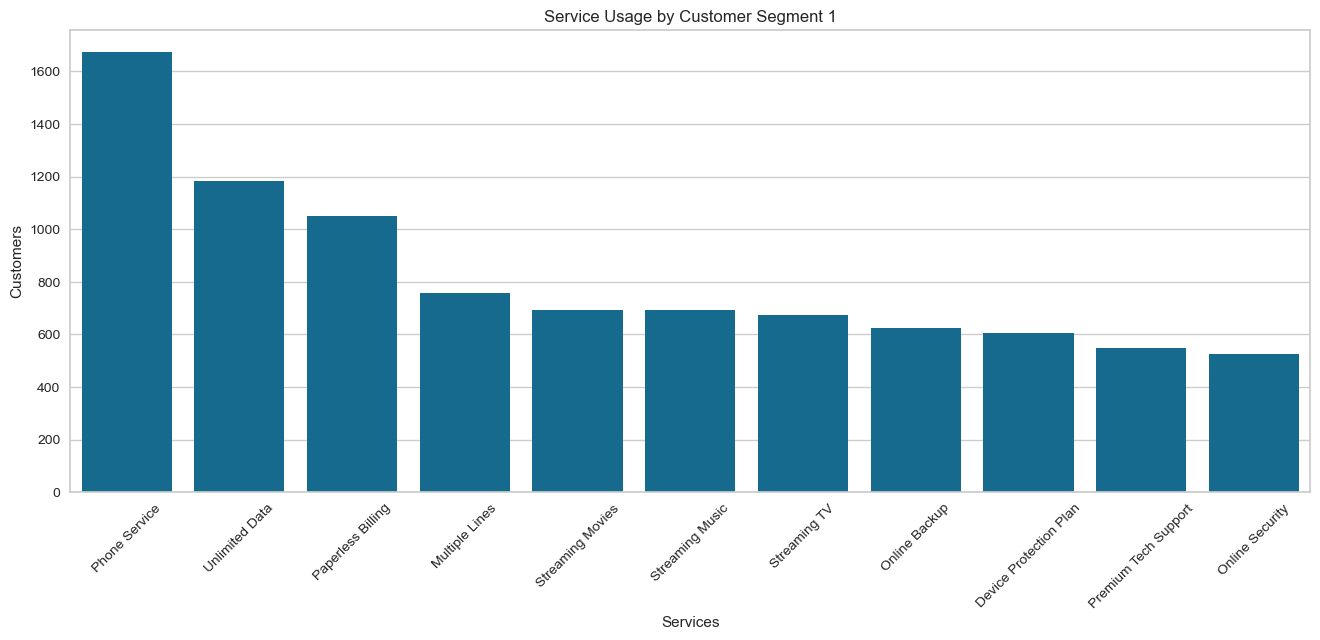
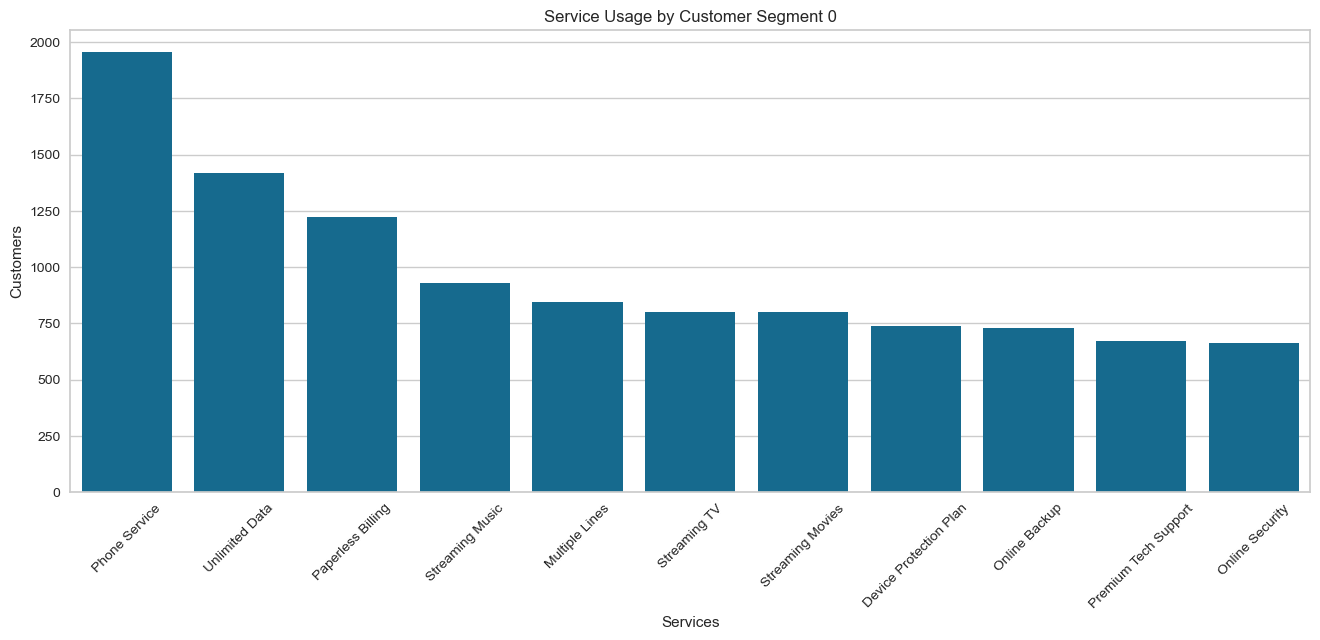


Figure 34. Elbow method for customer segmentation

Insights relating service usage and customer satisfaction for each customer segment are drawn as shown in Figure 35 and 36.



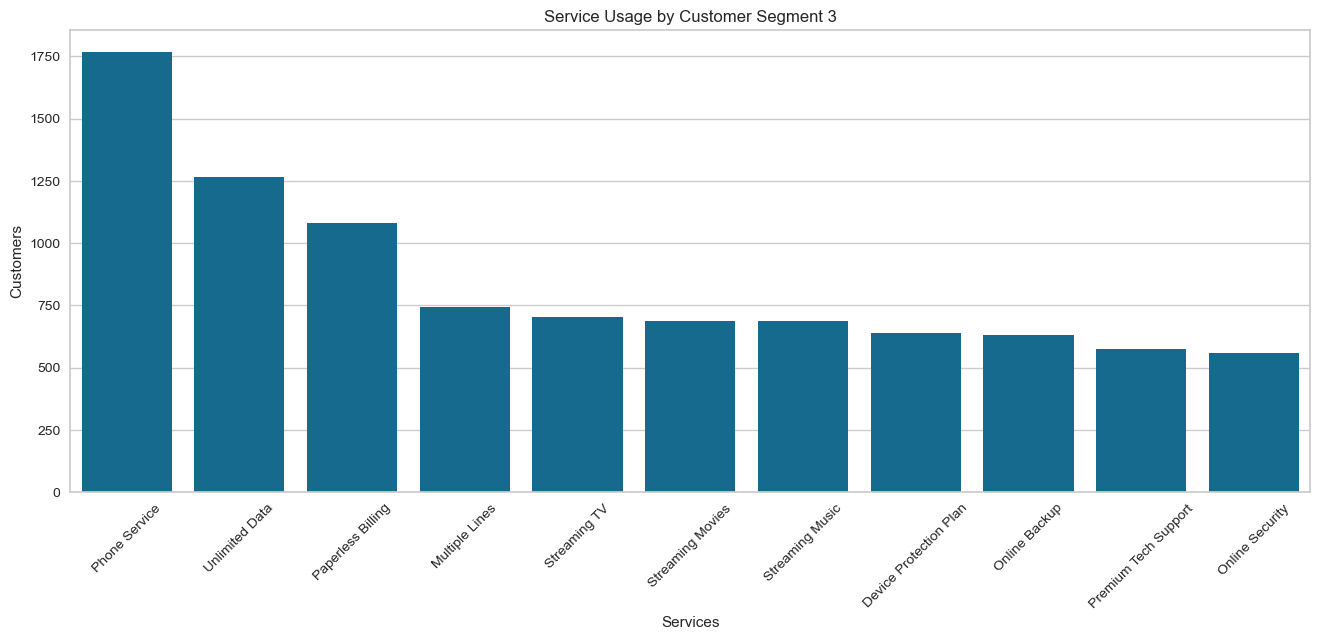


Figure 35. Service usage by customer segment

|  |  |
| --- | --- |
|  |  |

Figure 36. Correlation between services and customer satisfaction

## 5.3. Hypothesis Testing

### 5.3.1. Does longer contract length reduce churn?

I will carry out hypothesis testing by pair-wised comparison using proportion z-test on different contract lengths to see if longer contracts have less percentage of churn customers. Proportion z-test is used instead of normal z-test as the populations are different as shown in Figure 22. (GeeksforGeeks, 2022)

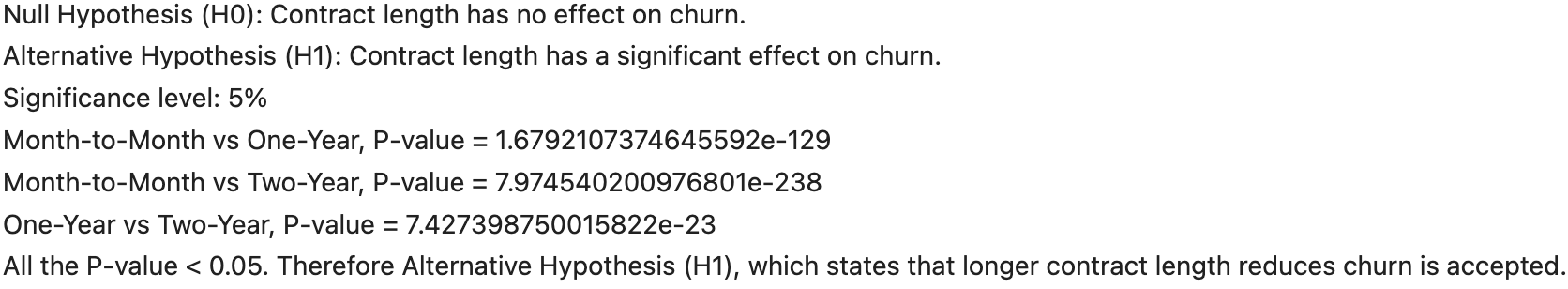


Figure 37. Hypothesis testing on if longer contract length reduce churn

### 5.3.2. Does more service usage affect churn?

T-test is carried out on the mean number of services used by both churn and non-churn customers to see if there is significant difference between the number of services used for churn and non-churn customers.

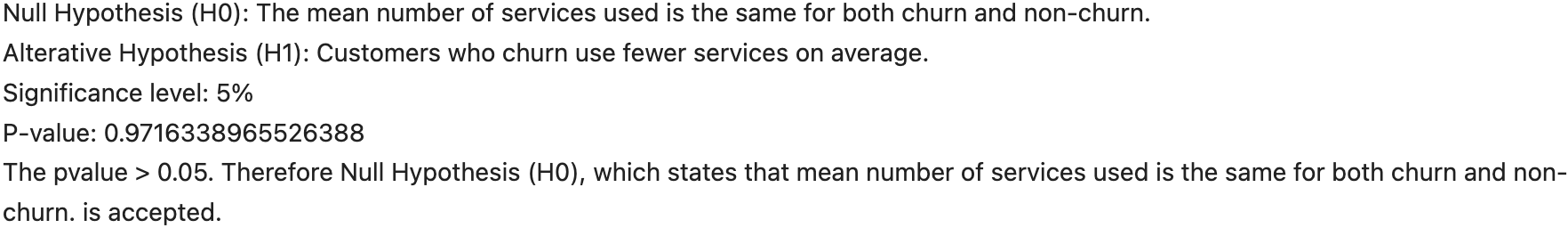


Figure 38. Hypothesis testing on if more service usage affect churn

## 5.4. Feature Selection

Some columns are identified and removed based on the reasons listed in Figure 39.

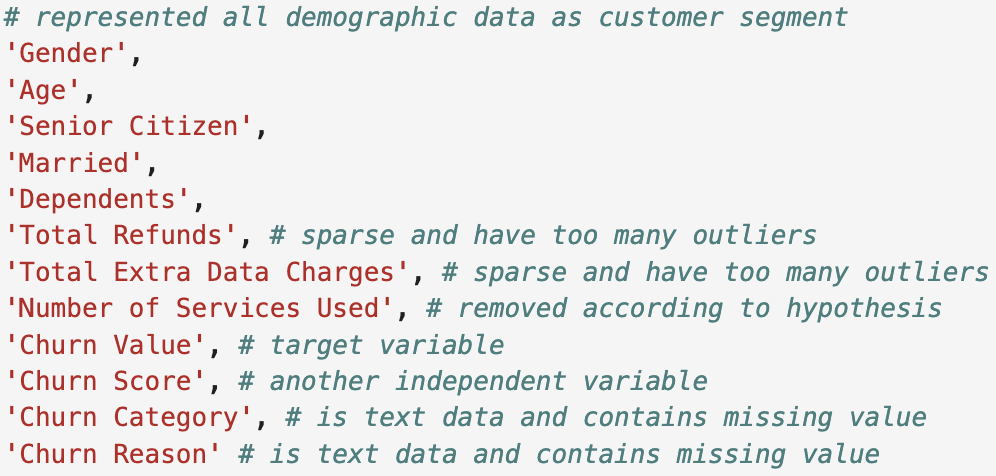


Figure 39. Unwanted columns with reasons

Feature selection is performed using Information Gain, and Fisher Score. Information gain calculates the reduction in entropy by evaluating each variable in the context of the target variable. (Aman, 2020). Fisher Score is a feature selection approach that ranks features based on their ability to differentiate various classes. A higher Fisher Score implies the characteristic is more discriminative and valuable for classification. (Rosidi, 2023).

|  |  |
| --- | --- |
| Figure 40. Feature Importance (Information Gain) | Figure 41. Feature Importance (Fisher Score) |

Features that are low on both metrics are identified and removed.

## 5.5. Splitting the data

Data is splitted into 70% training and 30% testing before all the preprocessing so that the testing data is completely unknown to the preprocessors, assuring no data leakage. (IBM, 2024)

## 5.6. Feature Scaling

I use MinMax scaler to scale all the values so that they all fall under the range of 0 to 1 and don't create bias based on the magnitude of the values.

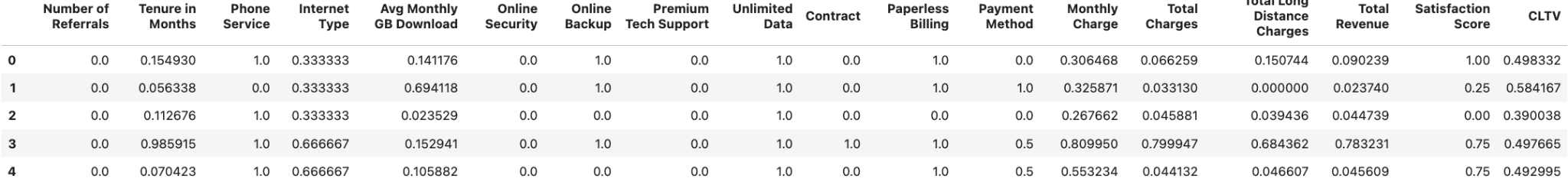


Figure 42. Dataset after feature scaling

## 5.7. Dimensionality Reduction

Principal Component Analysis (PCA) is used to reduce dimension from the dataset. Using PCA for dimensionality reduction involves zeroing out one or more of the smallest principal components, resulting in a lower-dimensional projection of the data that preserves the maximal data variance. For this dataset, I want to retain the 99% variance in the data.

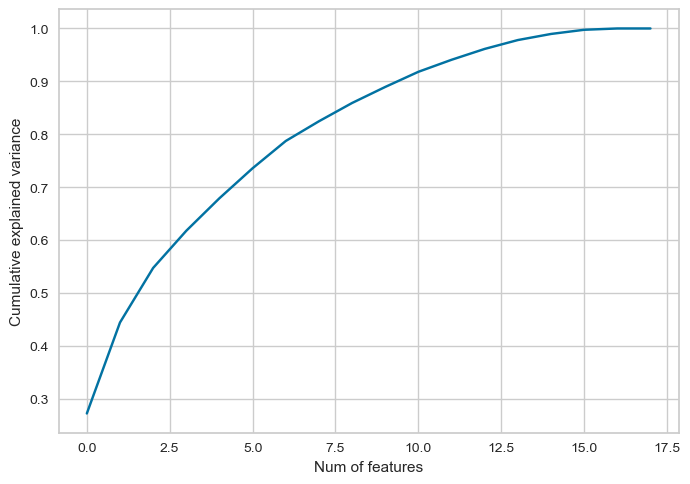


Figure 43. Principal component analysis

After plotting and calculating the cumulative sum of explained variance ratio, I got 16 components to achieve 99% variance.

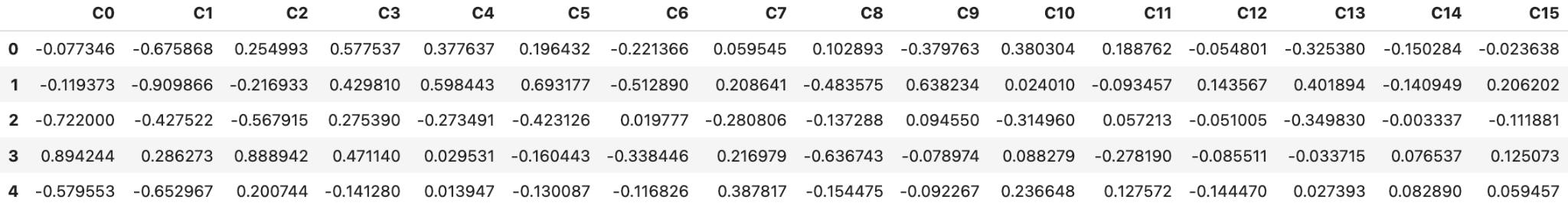


Figure 44. First 5 rows of X\_train after PCA

## 5.8. Class Balancing

Churn and Non-Churn classes are imbalanced and have about 73.4% and 26.6% respectively.

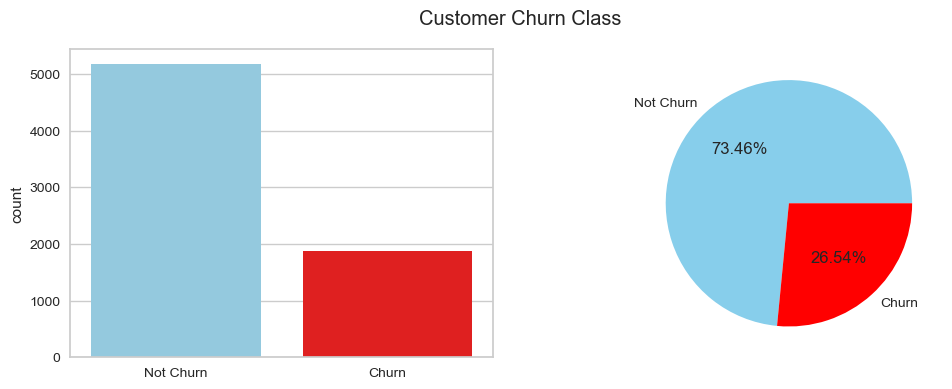


Figure 45. Churn Class Comparison

As the classes are imbalanced, the model will be biased towards the majority, resulting in low recall for the minority class. To solve this, an oversampling method called SMOTE will be used to generate synthetic data for minority classes to match the majority.

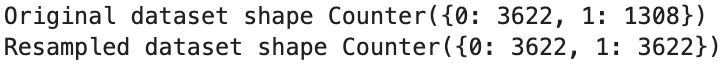


Figure 46. Training data observations after oversampling

# 6. Machine learning implementation

## 6.1. Model Selection

Models are considered based on the findings that there are less outliers, have moderately correlated variables, and being a binary classification problem.

**Logistic regression:** a baseline model for binary classification as it provides interpretability.

**K-Nearest Neighbors:** chosen as the dataset doesn’t have many outliers to distort the distance measures.

**Random Forest:** good for classification tasks and robust to outliers.

**XGBoost:** often a top-performing model in churn prediction and can handle nonlinear patterns (Mouli et al., 2024)

**Gaussian Naive-Bayes:** fast classification model and works very well with independence features.

**Support Vector Machine:** can capture non-linear relationships via different kernels.

## 6.2. Cross Validation

The models are tested with 5-fold cross validation so that we get an accurate measure of performance.

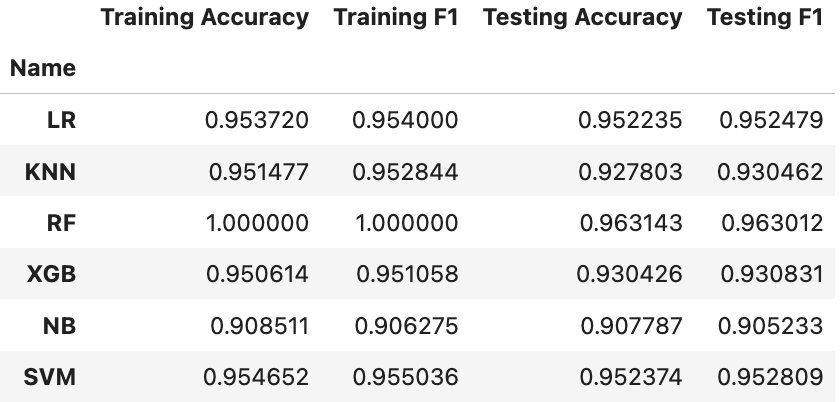


Figure 47. Cross-validation results

Linear Regression, Random Forest, and SVM come out on top for both accuracy and f1-score with Random Forest being a little bit overfitted. In the context of churn prediction, f1-score is the main metric for measure as the churn class will be the minority in real life too. Therefore it is important for models to not be biased towards the majority class, and f1-score will help identify that. I will perform hyperparameter tuning on these 3 models to get the best performance.

## 6.3. Hyperparameter Tuning

I will be using GridSearchCV to hyperparameter tune the 3 best models. GridSearchCV will fit multiple combinations of parameters on the models and validate with crossfold to find the best fit. F1-score will be used to measure the performance of the models.

### 6.3.1. Logistic Regression

* penalty (for regularization)
* C (smaller value for stronger regularization)
* solver (algorithms)

(GeeksforGeeks, 2024)

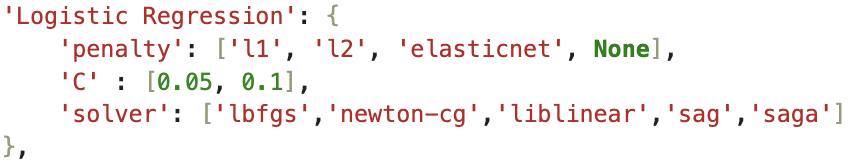


Figure 48. Logistic regression hyperparameters

### 6.3.2. Random Forest

* n\_estimators (number of trees)
* min\_samples\_split, min\_samples\_left (to prevent overfitting)

(Saxena, 2020)

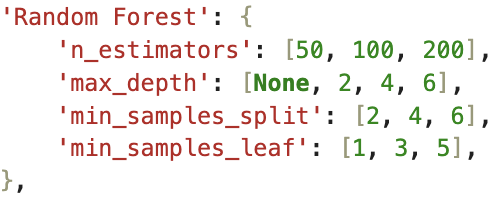


Figure 49. Random forest hyperparameters

### 6.3.3. SVM

* degree (complexity)
* C (regularization)
* gamma (sensitivity)
* kernel (draw boundaries between classes)

(Shrestha, 2024)

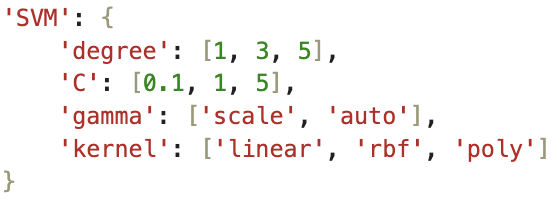


Figure 50. SVM hyperparameters

There is only a slight increase in F1-score after hyperparameter tuning. I will use these parameters to train the models and evaluate on test data.

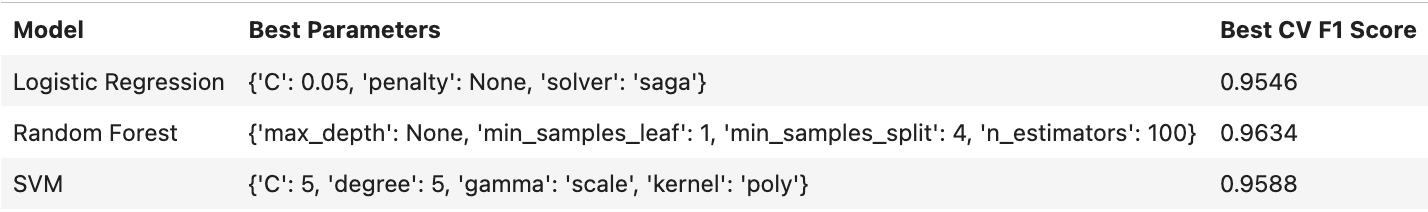


Figure 51. Best Hyperparameters

## 6.4. Evaluation

To choose the best model, I have considered based on the context that the business goal is to identify the most churn customer as possible, which means that recall is the measure to use. As shown in Figure 53, among all the models, only 36 churn customers are mis-predicted in Logistic Regression. Therefore, Logistic Regression is chosen as the best model to predict churn customers.

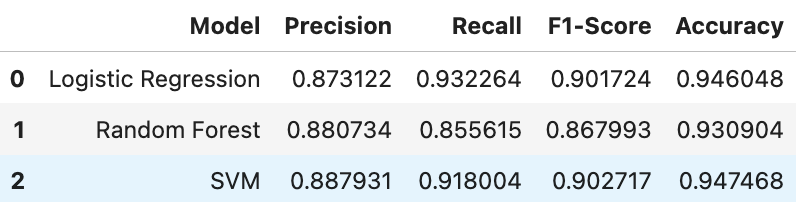


Figure 52. Models testing performance

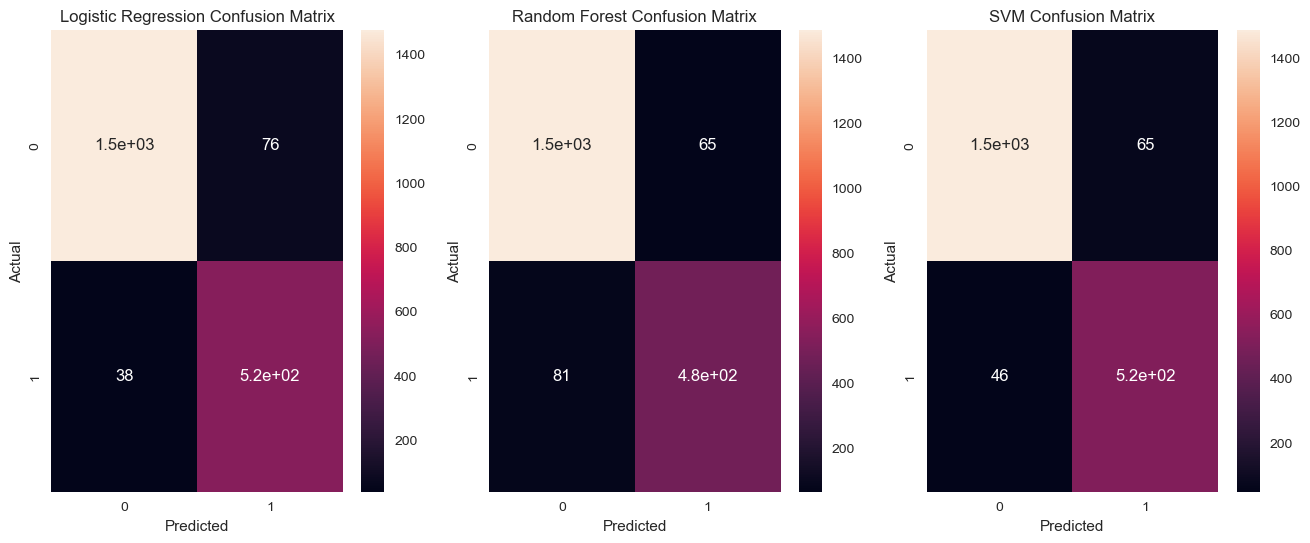


Figure 53. Models confusion matrix

# 7. Findings

Oversampling seems to be very effective as the model is not biased even though there is class imbalance in testing data. Dimensionality reduction using PCA and feature selection based on information gain and fisher score are very important in getting these results. It reduces a lot of training time, and optimizes performance as it saves the models from discovering meaningless patterns which makes the dataset more complex.

# 8. Conclusions

Multiple insights are discovered, including impact of different services on customer satisfaction, and churn. Customers are segmented into 4 groups and correlation of features to churn are identified for each segment. Hypotheses are tested and found that longer contracts reduce churn but more service usage doesn’t. Several models are experimented and hyperparameter tuned. Logistic regression is chosen as the best model based on having the best recall of 93%, which makes sure that churn customers are not missed out.

# 9. Any Future Recommendations

* Try more complex deep learning models, or time-series models to predict quarter by quarter
* Suggest to collect marketing spendings for each service so that effective retention strategies can be derived
* Use actual churn outcomes to continuously update and improve the model

# References

Gallo, A. (2014). *The Value of Keeping the Right Customers.* Harvard Business Review.

Dewan, S., Wu, D. J., & Trivedi, R. (2020). Customer engagement in fintech: A study of factors affecting retention. Journal of Banking and Finance, 118, 105888.

Wikipedia Contributors (2019). Cross-industry standard process for data mining. Available at: <https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining>.

IBM (2019). Telco customer churn. Available at: <https://community.ibm.com/community/user/blogs/steven-macko/2019/07/11/telco-customer-churn-1113>.

Sankalana, N. (2023). K-means Clustering: Choosing Optimal K, Process, and Evaluation Methods. Available at: <https://medium.com/@nirmalsankalana/k-means-clustering-choosing-optimal-k-process-and-evaluation-methods-2c69377a7ee4>.

GeeksforGeeks (2022). How to Perform a One Proportion ZTest in Python. [online] GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/how-to-perform-a-one-proportion-z-test-in-python>.

Aman (2020). Feature Selection in Machine Learning. Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/#h-supervised-techniques>.

Rosidi, N. (2023). Advanced Feature Selection Techniques for Machine Learning Models. KDnuggets. Available at: <https://www.kdnuggets.com/2023/06/advanced-feature-selection-techniques-machine-learning-models.html>.

IBM (2024). Data leakage machine learning. Available at: <https://www.ibm.com/think/topics/data-leakage-machine-learning>.

Mouli, K. C. et al. (2024) ‘An analysis on classification models for customer churn prediction’, Cogent Engineering, 11(1). doi: 10.1080/23311916.2024.2378877.

GeeksforGeeks. (2024). How to Optimize Logistic Regression Performance. Available at: <https://www.geeksforgeeks.org/how-to-optimize-logistic-regression-performance/>.

Saxena, S. (2020). Random Forest Hyperparameter Tuning in Python | Machine learning. Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2020/03/beginners-guide-random-forest-hyperparameter-tuning/>.

Shrestha, P. (2024). Tuning SVM Hyperparameters: Making Your Classifier Shine Like a Pro! Medium. Available at: <https://medium.com/@prayushshrestha89/tuning-svm-hyperparameters-making-your-classifier-shine-like-a-pro-8673639ddb16>.