Customer Churn Prediction

Author: Nguyen Ngoc Tue Minh

2023

1. Table of Contents

2. Table of Figures

3. Abstract

This report presents a detailed analysis of customer churn in a Californian telecommunications company, using advanced data analytics and machine learning techniques on a dataset from quarter 2 in 2022. It focuses on identifying key factors contributing to customer churn through extensive data exploration and visualization. The study evaluates various classification models, including Logistic Regression, KNN, Decision Tree, Gaussian Naive Bayes, and Random Forest, to predict customer churn effectively. The analysis results in strategic recommendations for customer retention, emphasizing the effectiveness of the Random Forest model. This work highlights the critical role of machine learning in understanding customer behavior and enhancing business decision-making in the telecommunications industry.

4. Introduction

This report consists of the detail analysis of the findings of a dataset related to customer churn in a telecommunications company in California during quarter 2 in 2022 through the various techniques of Data Analytics and Machine Learning. The dataset can be found in this link. The details are further elaborated and visually represented in the form of graphs and charts as a part of data visualization process. The dataset which is used for the analysis is the sales detail of the telecommunication company, based on several continuous and categorial variables. The expected result of this machine learning assignment is to predict whether or not a customer will leave the company by developing suitable classification predictive models and also give some recommendations on what improvements the company can make to retain customers.

5. Data overview

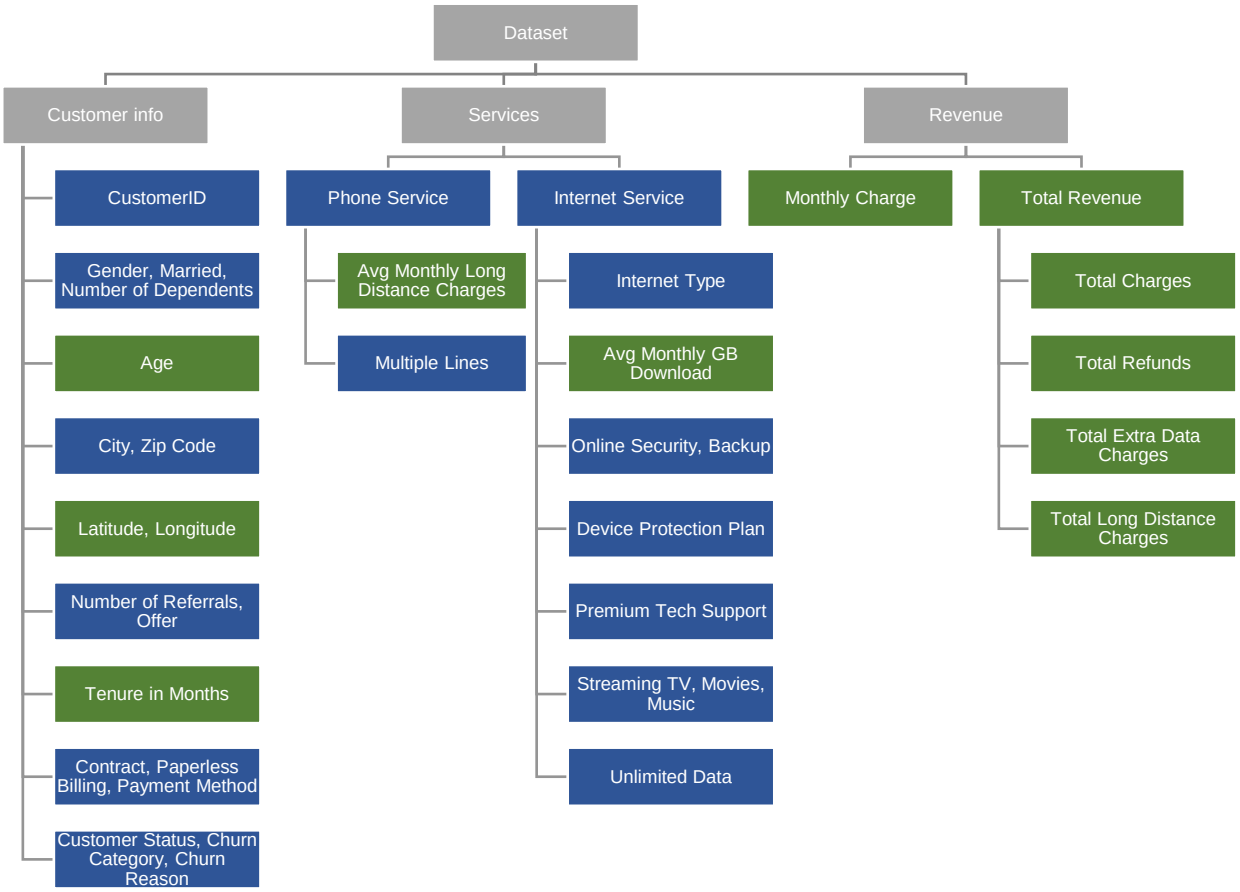

Figure 1

The selected data comprises 38 features which include 12 continuous and 26 categorical variables. Each row in the dataset represents for one customer including details about their information (ID, Gender, Married, Age, Contract, Status, ect), the services they are using (phone service and/or internet services) and the revenue generated from them. There are a total of 7,043 customers in the dataset. The figure above illustrates the dataset structure with blue boxes represent for categorical variables and green boxes represent for continuous variables.

6. Research idea and method

6.1. Research idea

Customer retention is a difficult problem for any company working in the service industry since customers are the most important factors to generate income. What are the factors contribute to customer churn and how to predict whether or not a customer will leave the company? Solving the customer retention problem is the main research motivation for this project.

The first purpose of this assignment is to investigate the factors that contribute to customer churn, such as demographics, subscription services, tenure, ect. Exploring the relationships between different variables, understanding the distribution of key features, and identifying patterns. Creating visualizations to better understand the patterns and trends in the data as a part of data visualization.

The second purpose of this assignment, as a part of the machine learning exercise, is to develop several different classification models to predict customer status. From that, trends and recommendations can be identified to bring the improvements to increase the sales of the company.

Two main research questions are: What are the factors that contribute to customer churn? and How to predict whether or not a customer will leave the company? To answer these two questions, the research team came up with 30 detailed questions, the questions list and answers for them can be found in the Appendix section.

6.2. Method

The method that was selected for this study is as follows. Data explorations is performed to identify the relationships and patterns between features, focuses on features that can impact the customer status (churned or stayed). After that the research team perform data preprocessing before training the models.

For the data exploration:

- Understand the distribution and unique values of each categorical variables using count plots.
- Understand the distribution and value range of each continuous variables using histplots, boxplots and mapbox (for geography-related features).
- Visualize the relationships between Customer Status and all other categorical and continuous features.
- Visualize the relationships between Total Revenue and Monthly Charge and all other categorical and continuous features.
- Visualize any other relationships needed to answer the research questions list.

For developing the predictive models:

- There are certain preprocessing steps required for data in order to implement it into our required models, which includes removing the columns that are not useful for prediction, converting categorical variables into numeric, splitting the dataset into training and testing, data cleaning checking for missing values, data standardization, checking for outliers and balancing the data.
- Develop 5 classification models which are Logistic Regression, K Nearest Neighbours, Decision Tree, Gaussian Naïve Bayes and Random Forest.

During the preprocessing steps, there are some notable points as below.

Remove columns that are not useful: Dependent columns are remove. There are many dependent columns in the dataset, for example, the Online Security Backup and Premium Tech Support among others are dependent on the Internet Service Columns. Those who have not chosen the internet package will not have any information in these columns and therefore have been removed. 454 rows of newly joined customers are also removed, as these customer data do not provide enough information for this study.

Convert categorical variables to numeric variables:

- In the final version of the notebook, nine categorical features are converted into numeric variables: “Married”, “Phone Service”, “Internet Service”, “Customer Status”, “Gender", "Offer", "Contract", "Paperless Billing", "Payment Method".
- We also tried another version in which we converted two more categorical features and grouping before converting for three features: “Number of Dependents” to 0 (no dependents) and 1 (have dependents), “Number of Referrals” to 0 (no referrals) and 1 (have referrals), “Offer” to 0 (no offers) and 1 (have offers), “Contract” to 0 (monthly contract) and 1 (yearly contract), “Payment Method” to 0 (Bank Withdrawal) and 1 (other methods). However, the model performance in this second version is lower than the first one. Therefore, we finalized the first version. The notebook for the second version can be found here.

Data standardization: Min Max Scaling is chosen instead of Standard Scaling for this dataset. There are several reasons for this:

- After viewing the distribution of all features (categorical and continuous) there is no feature follows normal distribution (Gaussian distribution), so the Standard Scaling which is useful for algorithm that assumes the data is normal distributed (Linear Regression, Logistics Regression, ect) is not beneficial here.
- Min Max Scaler translates each individual features into a same range 0 to 1, which is highly beneficial for distance-based algorithms like K Nearest Neighbor, which is also used in this project.
- Min Max Scaler is sensitive to outliers but this dataset does not have many outliers in a total of 7043 entries: 200 rows on “Total Refunds”, 197 rows on “Total Extra Data Charges”, 30 rows on “Total Long Distance Charges” and 2 rows on “Total Revenue”. However, “Total Refunds” and “Total Extra Data Charges” are not important features and they only have minor contributions to “Total Revenue” which is explained in later part.

Balancing the dataset: Over sampling is chosen instead of under sampling for this dataset. There are several reasons for this:

- This is not a large dataset (only 7043 records) which takes too much time or resources to run full dataset. We want to utilize all the data for developing the models and not want to decrease the size and loss any information.
- Over sampling can increase the chance of overfitting, however, the good models’ performance in testing set shown in later parts proved that over sampling in this case does not cause overfitting problem.

6.3. Model evaluation principle

The company wants to correctly predict which customers will churn so that they can take timely actions to retain those customers. An ideal predictive model will have high Recall and Precision of Churned but in reality a model cannot have both because there is usually a Precision-Recall trade-off.

Low Recall of Churned means there are more customers who were predicted to stay but churned, this will incur the cost of losing existing customers. This cost is very high, because customers often stay with the same provider for years in the Telecommunication Industry (Lifetime Value). Losing a customer now means not only losing his revenue for this month or this year, but also losing his revenue for several years in the future. Meanwhile, low Precision of Churned class means there are more customers who were predicted to churn but stayed. The company would waste money (special offers, discounts, etc.) to retain more

customers who were predicted to churn but stayed anyway. Compared with the cost of losing customers, this is of course less costly.

Moreover, retaining customers would be more profitable for a company than finding new customers. This can be up to five times. Profits could also go up between 25% and 95% by doing the same. This strategy goes beyond just financials. This goes on to state that loyal customers would be likely to refer services to others as well as try new offers. The current customers would be ready to forgive too and repurchase products. The simple way to keep customers is by staying in touch with them and reminding them of events that are coming up shortly.(Landis, 2022)

Telecommunication communication companies worldwide are aware of the fact that they need to be very careful with how they treat their existing customers as they realize that they will spend more on acquiring new customers than keeping old ones. (AL-Shatnwai, A.M. and Faris, M., 2020)

Most firms that are in their growing stage try to find a balance between retaining old customers and acquiring new customers. They need to be reminded that retaining old customers would work out to be cheaper than acquiring new ones. This is very visible, especially in the SaaS sector. (Kumar, S., 2022)

Most retained customers have a trust factor when it comes down to a product that they have been using for some time and therefore it is not necessary for a firm to remind them about the features of the product. It also goes on to show that as they get familiar with the product over some time, they will not require training as they already have enough knowledge about the product. On the other hand, new customers have difficulty accepting new products and are skeptical of just the advertisement alone. This goes to show that acquiring new customers is more expensive than retaining old ones. (You, Y. and Joshi, A.M., 2020)

Based on the above facts and reasons, the company should focus more on retaining existing customers than acquiring new one. The cost of losing existing customers is very costly compared to cost of retaining wrong customers, therefore, ***the best performance model should have high Recall of Churned class but still maintain an acceptable Precision of Churned class.***

7. Analysis result

7.1. Descriptive analysis

This part focuses on describing some notable features and presenting some important insights that relevant with customer status. The visualization results of other features and less important relationships can be found in the notebook.

7.1.1. Understand categorical variables

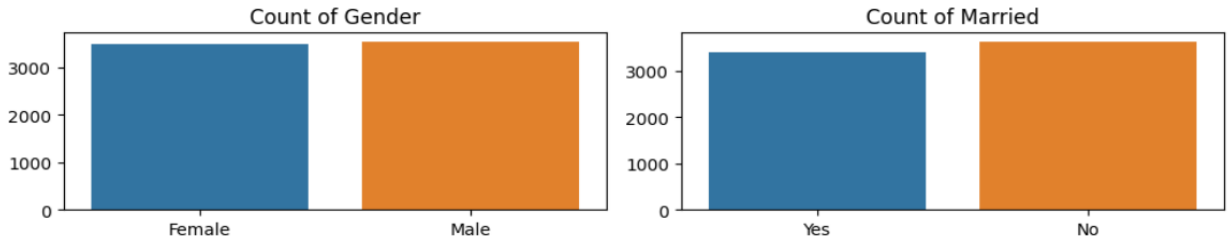

Figure 2

The first category we have is the count of males & females who are using the services & the marital status of the people. Male & females are almost in an equal proportion whereas, the count of single people is slightly higher than the married people.

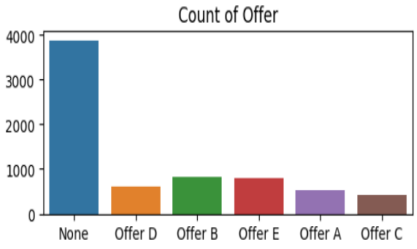

Figure 5

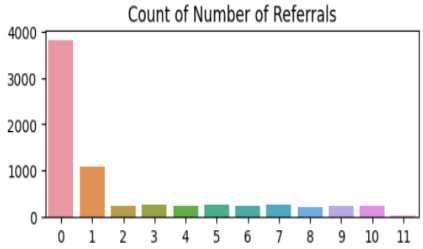

Figure 4

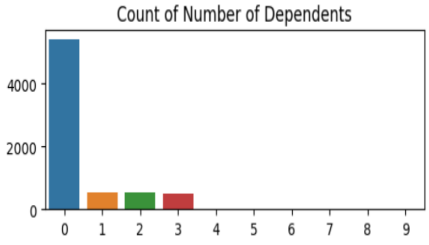

Figure 3

Other categories are Offer, Number of Referrals and Number of Dependents. According to our data visualization, the customers who don't use any offers are far higher than the customers who avail any offer. Customers without any referrals are also higher than the customers who started using the services due to any referrals. Lastly, the count of people who don't have any dependent is nearly 6000, on the other hand people who have any dependents are not even nearly 1000.

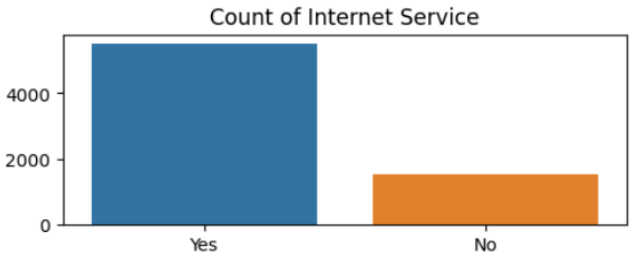

Figure 6

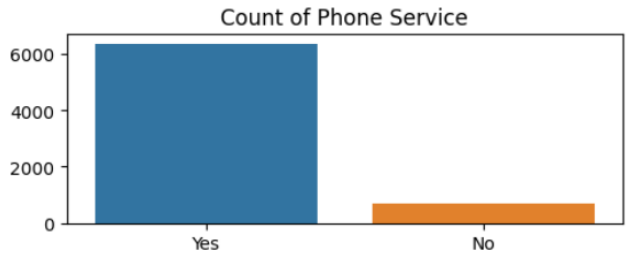

Figure 7

The number of customers who are using internet services and phone services are higher than the ones who are not using these services.

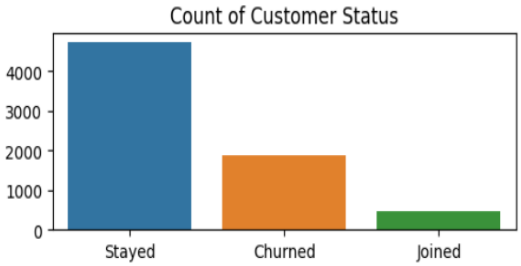

Figure 8

Another variable is customer status which shows us how many customers stayed, churned and joined. The highest proportion is for the customers who stayed, churned are on the second number and the new customers have the lowest proportion.

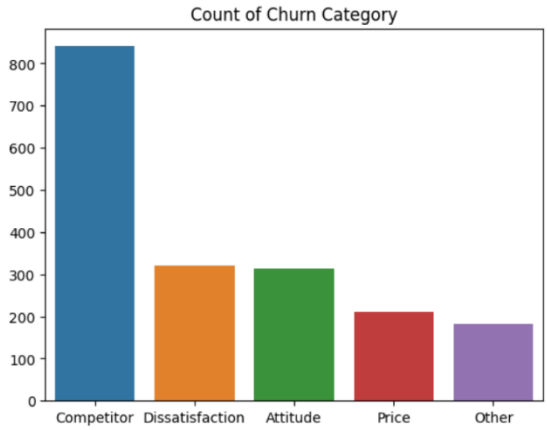

Figure 9

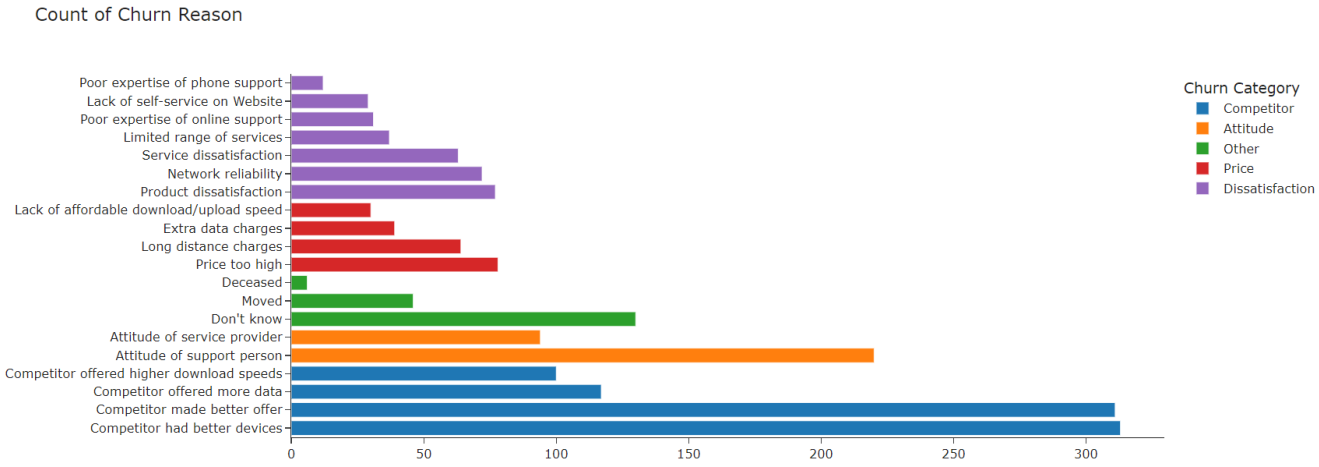

Figure 10

Above mentioned are the major variables which shows us what are the reasons why we lost the customers. Competitors stood at first, dissatisfaction and attitude of the service providers are on second and third position respectively. The price and other factors come at last. Our customers are shifting to competitors service, so the main reasons are more data offered by competitor, they have better offers and better devices. Secondly, our attitude towards the customers in terms of service providing & support also the major factor of churn.

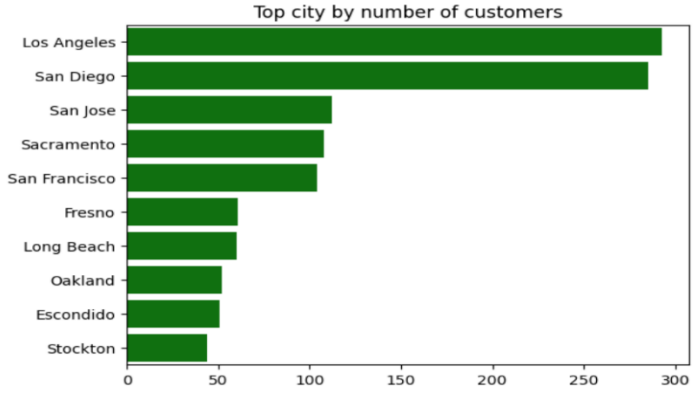

Figure 11

The above bar chart represents the location of the customers, Los Angeles and San Diego have the highest customers. San Jose, Sacramento and San Francisco have a very slight difference after the top 2 locations. Whereas Stockton is the city with the least customers.

7.1.2. Understand continuous variables

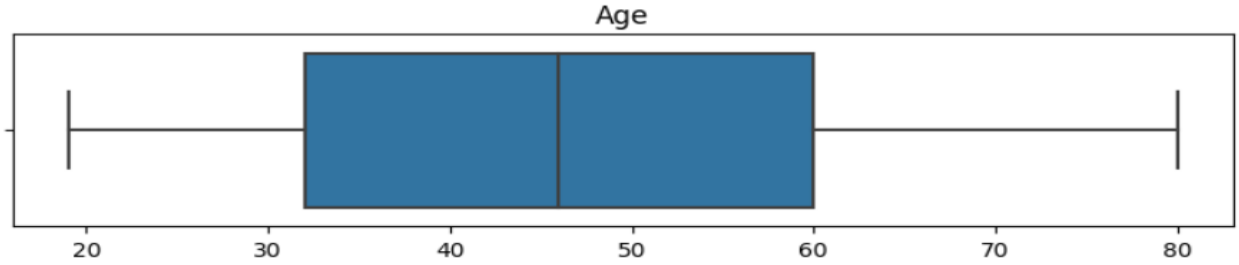

Figure 12

The customer's age who are using our services is in the range of teenagers to 80's. Almost 25% age of the customers are mid 30's, 50% people are in the range of late 40's , 75% people are up to 60 years old & the maximum age of our customers are 80 years old.

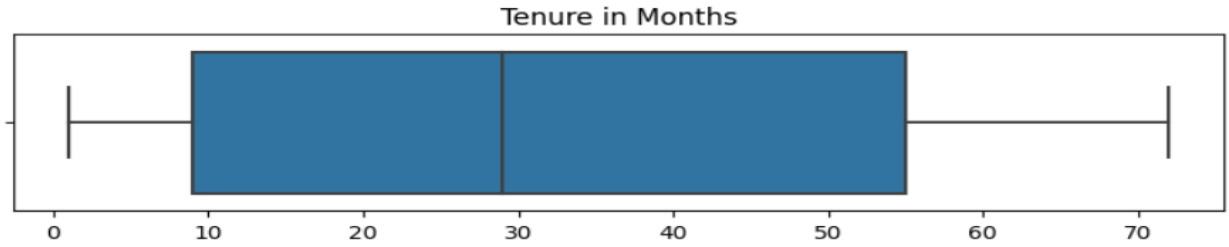

Figure 13

A contract's duration also varies from customer to customer & the offer they are using. There are some customers who don't have any contract & use the services on a monthly pre-paid basis. Whereas almost 50% of the customers have less than 29 months contract, 75% of the customer's contract tenure is approximately less than 55 months & the maximum contract duration is 72 months.

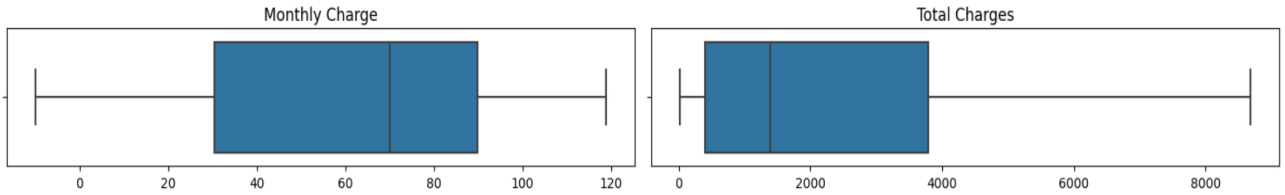

Figure 14

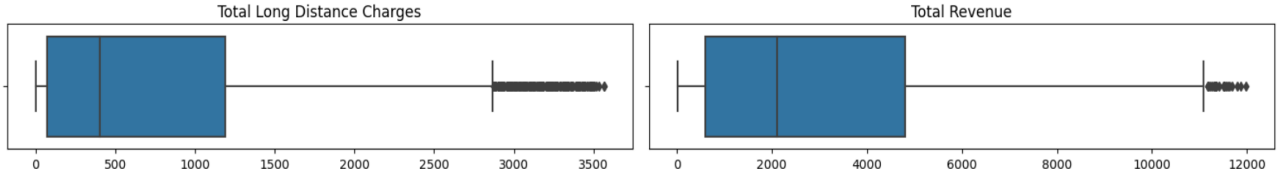

Figure 15

Total revenue is determined by subtracting total refunds and adding total extra data charges and total long-distance charges to the total charges. Additionally, total revenue mainly comes from total charges and total long-distance charges, while total refunds and total extra data charges contribute relatively small to the overall revenue.

7.1.3. Key findings

Below mentioned are some key findings from our data visualization & some suggestions based on that.

Key finding 1: Offers contributing more/less to Revenue

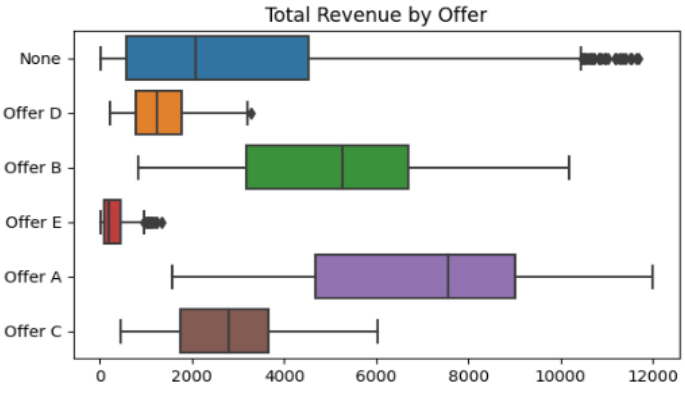

Figure 17

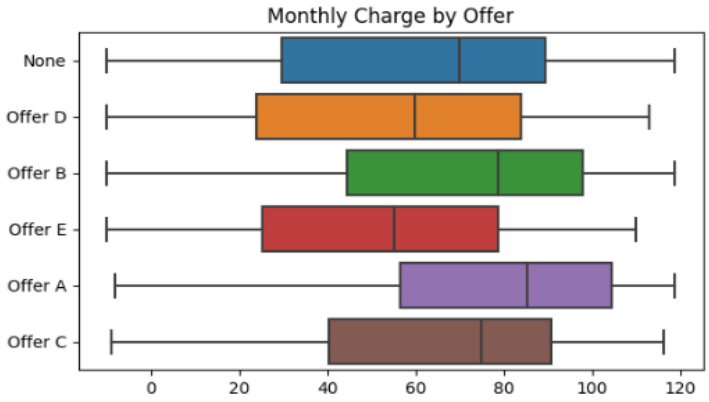

Figure 16

Figure 18

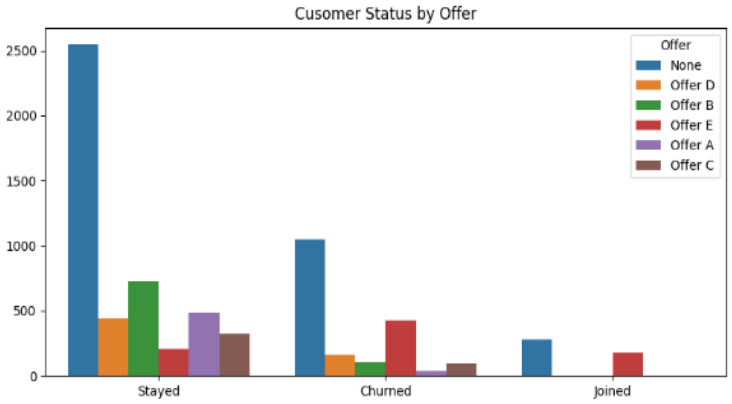

Figure 19

While considering Offer E in terms of total revenue, monthly charges, offer tenure & churned status by offer, it's visible that Offer-E is the least performing & least revenue generating offer. Offer E also has the least tenure as compared with other offers & also the churned proportion is higher for Offer-E among other offers. Company can consider making adjustments to Offer E based on customer feedback and performance analysis. Alternatively, if it continues to underperform, discontinuing it may be a viable option. The company should discontinue offering Option E due to its low total revenue and low monthly charge, especially if it leads to more customer churn. Instead, focusing on promoting Offer A and B. This decision might assume that these alternatives (A and B) are more profitable or have a better customer retention rate.

Key finding 2: Assessment of Revenue Streams

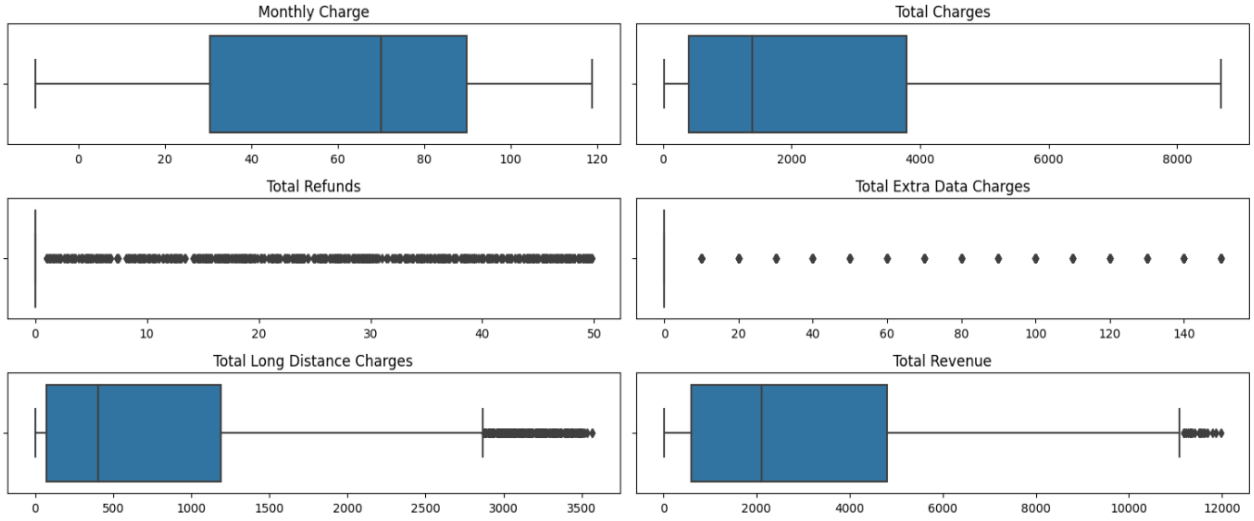

Figure 20

Total revenue is based on total charges & total charges are mainly contributed by the long-distance charges. Total refunds & extra data charges are very low. Promoting the Long-Distance service can significantly contribute to the total revenue, implementing more promotional programs can be a strategic move. Offer bundles that include Long Distance Service along with other services, encouraging customers to opt for comprehensive packages. Company can also consider to implement referral programs that reward existing customers for referring others to use the Long Distance Services.

Key finding 3: Long-term and High-value customers leaving

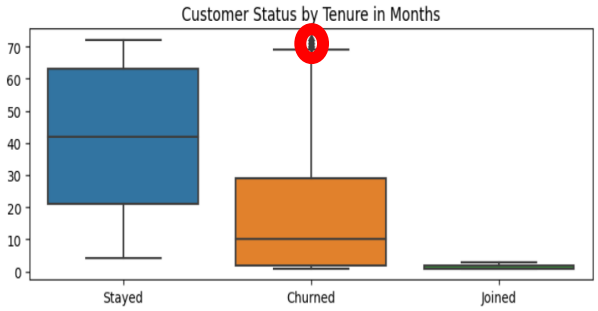

Figure 21

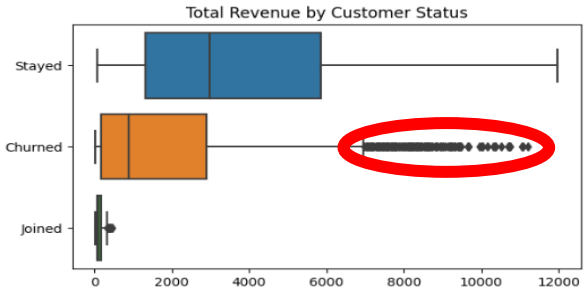

Figure 22

	All customers	Long-term customers	High-value customers
Total Revenue Mean	3,034	8,470	8403
Monthly Charge Mean	64	97	102

Figure 23

Churn Reason	Long-term customers	High-value customers
Competitor	39%	52%
Dissatisfaction	22%	15%
Price	22%	11%
Attitude	17%	16%
Other	0%	6%

Figure 24

The company is losing its revenue because of churned customers, let's have a look the factors which underpins their decision & impact on the overall revenue of the company. When consider customer status by their tenure and total revenue, there are some long-term customers (more than 70 months) and high-value customers (more than 7000) who left the company. These are important customers to the company because their average Total Revenue and Monthly Charge are both higher than the whole population, so investigating their churn reasons is necessary.

There are several factors which contributes to the decision discontinuing of the offer with our company and the major factors are competitors (39% and 52%), price (22% and 11%), dissatisfaction (22% and 15%) and attitude of the company (11% and 16%). The company should conduct a thorough analysis of competitors' services, pricing models, and customer satisfaction levels. Identify areas where competitors excel and areas where the company can differentiate itself as a part of their competitive strategy. Gather feedback from existing and churned high-value customers to understand their specific concerns and reasons for leaving. Train employees to provide excellent customer service. Satisfied and well-informed employees are more likely to create positive experiences for customers, contributing to customer retention.

Key finding 4: Revenue from Married/Single people

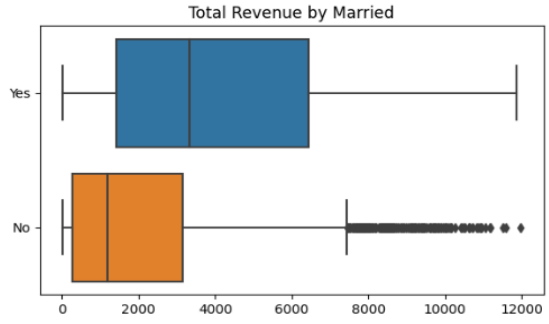

Figure 26

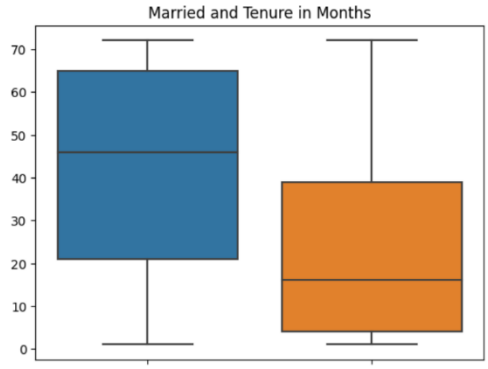

Figure 25

The company is generating significantly more revenue from married people & the tenure of the contract is also more as compared to the single people. Designing targeted promotion programs for married customers is a strategic approach that can enhance customer loyalty and maximize revenue. Create bundled service packages that cater to the needs of families. This could include combining internet, TV, and phone services at a discounted rate for married customers. Develop loyalty programs that recognize and reward long-term engagement. Provide special perks, discounts, or exclusive access to events for married customers who have been with the company for an extended period.

Key finding 5: Contract's tenure & revenue generated from each type of contract

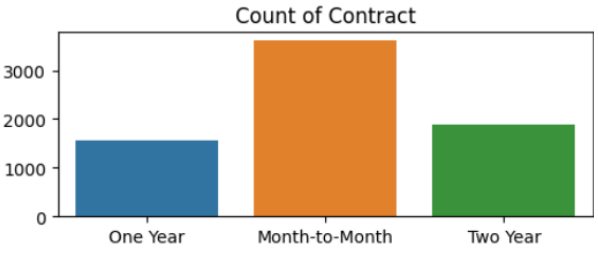

Figure 28

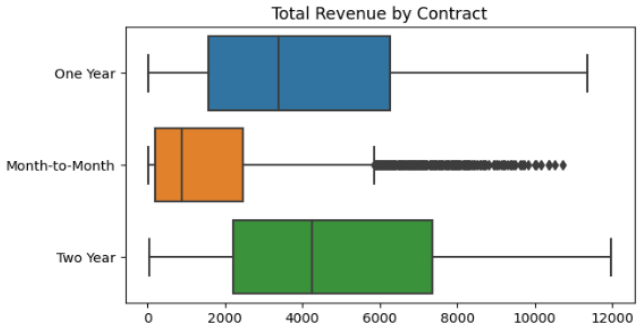

Figure 27

According to our data visualization, the graphs show that the largest number of contracts are customers who use monthly basis services, instead of any long-term commitment such as one or two-year contracts. On the other hand, while looking at the total revenue from monthly contracts is less among all types of contracts. The company should establish open communication channels to gather feedback from the customers who use monthly services & use this feedback to continuously refine and improve the services and promotions offered to meet their evolving needs. Offer attractive discounts or incentives for customers who choose to subscribe annually rather than monthly. Highlight the cost savings and added value of the yearly plan. Allow customers to customize their annual plans based on their needs. This flexibility can make the annual subscription more appealing by catering to individual preferences.

7.2. Predictive analysis

In this section, five predictive classification models are chosen to predict Customer Status: Logistic Regression, K Nearest Neighbors, Decision Tree, Gaussian Naive Bayes, Random Forest. The same training and testing set, including the same balanced and resampled data set, are used for the training process of all five model types. Each section below includes parameter tuning, model configuration and model evaluation.

7.2.1. Logistics Regression

Algorithm brief overview: Logistic Regression is a statistical method used for binary classification problems, where the outcome variable is categorical and has two or more classes. In our data we are predicting the number of customers who left & stayed, therefore Logistics Regression is suitable for this case.

Data preprocessing: There are four version of data are used in training this model: unscaled and unbalanced, scaled and unbalanced, unscaled and balanced, and scaled and unbalanced.

Model evaluation: Below are the performance from four version of data:

No	Data version	Accuracy Score	Balanced Accuracy Score	Precision of Churned	Recall of Churned	F1
1	Unscaled and unbalanced data	0.819423	0.752537	0.720257	0.597333	0.653061
2	Scaled and unbalanced data	0.846737	0.807765	0.736986	0.717333	0.727027
3	Unscaled and balanced data	0.741275	0.748520	0.531481	0.765333	0.627322
4	Scaled and balanced data	0.814871	0.828061	0.627680	0.858667	0.725225

Table 1

Consider all models above, **scaled and balanced data** gives the best performance. The final equation is: Log(Y/(1-Y)) = 0.0914 - 0.7514 * Age - 1.4383 * Married + 4.8782 * Number of Dependents + 5.6443 * Number of Referrals + 2.5284 * Tenure in Months + 0.5265 * Phone Service - 2.9587 * Monthly Charge - 0.8094 * Population - 0.9386 * Offer A +0.2914 * Offer B + 0.4950 * Offer C + 0.8884 * Offer D - 0.6984 * Offer E + 1.4927 * Contract One Year + 2.9210 * Contract Two Year - 0.3256 * Paperless Billing + 0.4601 * Payment Method Credit Card - 0.4226 * Payment Method Mailed Check

- Among 1318 customers in the testing set, there are 322 customers were correctly predicted to churn (True Positive), 752 customers were correctly predicted to stay (True Negative), 53 customers were predicted to stay but actually churned (False Negative), 191 customers were predicted to churn but actually stayed (False Positive).
- Recall of churned is 86% which means 86% of customers who actually churned were predicted to churn. This is a good performance.

- Precision of churned is 63% which means 63% of customers who were predicted to churn actually churn. This is not as good as other versions but if consider the trade-off with Recall, this performance is acceptable.
- The harmonic mean of precision and recall is 73% (F1 score) which is high.
- Balanced accuracy score is 83% which means the model can correctly predict 83% of the balanced data. This is a highest performance compared with other versions.

7.2.2. K Nearest Neighbors

Algorithm brief overview: K Nearest Neighbors (KNN) Classifier identifies the "K" nearest data points to a new data point and assigns the most common class among these neighbors to the new data point. KNN is a simple, intuitive and powerful technique for classification problem and it is suitable for this dataset and useful to predict Customer Status in this case.

Data preprocessing: Because KNN classifier bases on the nearest neighbors to make decisions, scaling data in this model type is very important to ensure all features are on the same scale. For this model, the same training process is performed for two versions of data: 1/ scaled and unbalanced data and 2/ scaled and balanced data.

Parameter tuning and model configuration: In each version of data, different combinations of parameters are tested using 3 nested for-loop commands until the best-performing model is identified:

- For "n_neighbors": from 1 to 10
- For "metrics": "minkowski","manhattan","euclidean","canberra"
- For "weights": "uniform","distance"

Model evaluation: Below are the top 5 performing models from each version of data

With scaled and unbalanced data:

No	Combination	Accuracy Score	Balanced Accuracy Score	Precision of Churned	Recall of Churned	F1
14	n=2, m=canberra, w=uniform	0.760243	0.779444	0.552773	0.824000	0.661670
10	n=2, m=manhattan, w=uniform	0.754173	0.764762	0.547135	0.789333	0.646288
62	n=8, m=canberra, w=uniform	0.837633	0.819071	0.691211	0.776000	0.731156
46	n=6, m=canberra, w=uniform	0.824734	0.810057	0.664384	0.776000	0.715867
30	n=4, m=canberra, w=uniform	0.798938	0.791227	0.617021	0.773333	0.686391

Table 2

With scaled and balanced data:

No	Combination	Accuracy Score	Balanced Accuracy Score	Precision of Churned	Recall of Churned	F1
62	n=8, m=canberra, w=uniform	0.757967	0.794719	0.546358	0.880000	0.674157
78	n=10, m=canberra, w=uniform	0.757967	0.794719	0.546358	0.880000	0.674157
46	n=6, m=canberra, w=uniform	0.743551	0.782235	0.529984	0.872000	0.659274
74	n=10, m=manhattan, w=uniform	0.761002	0.793627	0.550676	0.869333	0.674250

58	n=8, m=manhattan, w=uniform	0.760243	0.793097	0.549747	0.869333	0.673554

Table 3

Consider all models above, with **scaled and balanced data** the combination of **n=8, m=canberra, w=uniform** gives the best performance:

- Among 1318 customers in the testing set, there are 330 customers were correctly predicted to churn (True Positive), 669 customers were correctly predicted to stay (True Negative), 45 customers were predicted to stay but actually churned (False Negative), 274 customers were predicted to churn but actually stayed (False Positive).
- Recall of churned is 88% which means 88% of customers who actually churned were predicted to churn. This is a good performance.
- Precision of churned is 55% which means 55% of customers who were predicted to churn actually churn. This is not a good performance but acceptable when compared with other versions.
- The harmonic mean of precision and recall is 67% (F1 score)
- Balanced accuracy score is 79% which means the model can correctly predict 79% of the balanced data. This is a pretty good performance.

The fact that the top-performing models above predominantly use the metric of Canberra is understandable because this distance metric is sensitive to small changes of low-valued features (Eskandar, 2021). The original dataset has many features with different magnitudes (for example, the range of “Number of Dependent” is 0 to 11, the range of “Total Revenue” is 0 to 12000). The data is then scaled using the Min Max Scaler, which translates all features into the same scale of 0-1. Therefore, some wide-range continuous variables like “Total Revenue” will vary widely, which can be the key reason why the Canberra distance metric performs better than other metrics.

7.2.3. Decision Tree

Algorithm brief overview: A decision tree classifier is a popular machine learning algorithm used for classification tasks. It works by recursively partitioning the dataset into subsets based on the most significant attribute at each step. The goal is to create a tree structure where the leaves represent the class labels.

Data preprocessing: Although scaling data does not impact the performance of Decision Tree, in this model, we still use the same training process for four versions of data: 1/ unscaled and unbalanced data, 2/ scaled and unbalanced data, 3/ unscaled and balanced data and 4/ scaled and balanced data.

Parameter tuning and model configuration: In each version of data, different combinations of parameters are tested using 2 nested for-loop commands until the best-performing model is identified:

- For “max_depth”: from 1 to 20
- For “criterion”: "gini", "entropy", "log_loss"

Model evaluation: Below are the top 5 performing models from each version of data

With unscaled and unbalanced data:

No	Combination	Accuracy Score	Balanced Accuracy Score	Precision of Churned	Recall of Churned	F1
13	i=5, k=entropy	0.844461	0.815009	0.717949	0.746667	0.732026
14	i=5, k=log_loss	0.844461	0.815009	0.717949	0.746667	0.732026
16	i=6, k=entropy	0.843703	0.813676	0.717224	0.744000	0.730366

17	i=6, k=log_loss	0.843703	0.813676	0.717224	0.744000	0.730366
10	i=4, k=entropy	0.817147	0.795118	0.658019	0.744000	0.698373

Table 4

With scaled and unbalanced data:

No	Combination	Accuracy Score	Balanced Accuracy Score	Precision of Churned	Recall of Churned	F1
13	i=5, k=entropy	0.844461	0.815009	0.717949	0.746667	0.732026
14	i=5, k=log_loss	0.844461	0.815009	0.717949	0.746667	0.732026
16	i=6, k=entropy	0.843703	0.813676	0.717224	0.744000	0.730366
17	i=6, k=log_loss	0.843703	0.813676	0.717224	0.744000	0.730366
10	i=4, k=entropy	0.817147	0.795118	0.658019	0.744000	0.698373

Table 5

With unscaled and balanced data:

No	Combination	Accuracy Score	Balanced Accuracy Score	Precision of Churned	Recall of Churned	F1
4	i=2, k=entropy	0.553111	0.675652	0.385439	0.960000	0.550038
5	i=2, k=log_loss	0.553111	0.675652	0.385439	0.960000	0.550038
7	i=3, k=entropy	0.730653	0.779647	0.515385	0.893333	0.653659
8	i=3, k=log_loss	0.730653	0.779647	0.515385	0.893333	0.653659
21	i=8, k=gini	0.832322	0.839453	0.657787	0.856000	0.743917

Table 6

With scaled and balanced data:

No	Combination	Accuracy Score	Balanced Accuracy Score	Precision of Churned	Recall of Churned	F1
4	i=2, k=entropy	0.553111	0.675652	0.385439	0.960000	0.550038
5	i=2, k=log_loss	0.553111	0.675652	0.385439	0.960000	0.550038
7	i=3, k=entropy	0.730653	0.779647	0.515385	0.893333	0.653659
8	i=3, k=log_loss	0.730653	0.779647	0.515385	0.893333	0.653659
21	i=8, k=gini	0.831563	0.838923	0.656442	0.856000	0.743056

Table 7

Consider all models above, with **unscaled and balanced data** the combination of **max_depth=3, criterion=entropy** gives the best performance. Although scaling data does not impact the performance of Decision Tree, the unscaled version is chosen over the scaled one to prioritize the interpretation of the tree. The performance and visualization of the tree is as below:

- Among 1318 customers in the testing set, there are 335 customers were correctly predicted to churn (True Positive), 628 customers were correctly predicted to stay (True Negative), 40 customers were predicted to stay but actually churned (False Negative), 315 customers were predicted to churn but actually stayed (False Positive).
- Recall of churned is 89% which means 89% of customers who actually churned were predicted to churn. This is a good performance.
- Precision of churned is 52% which means 52% of customers who were predicted to churn actually churn. This is not a good level of precision however compared with other very low precision (39%) and consider the trade-off, this 52% is acceptable.
- The harmonic mean of precision and recall is 65% (F1 score)
- Balanced accuracy score is 85% which means the model can correctly predict 85% of the balanced data. This is a good performance.

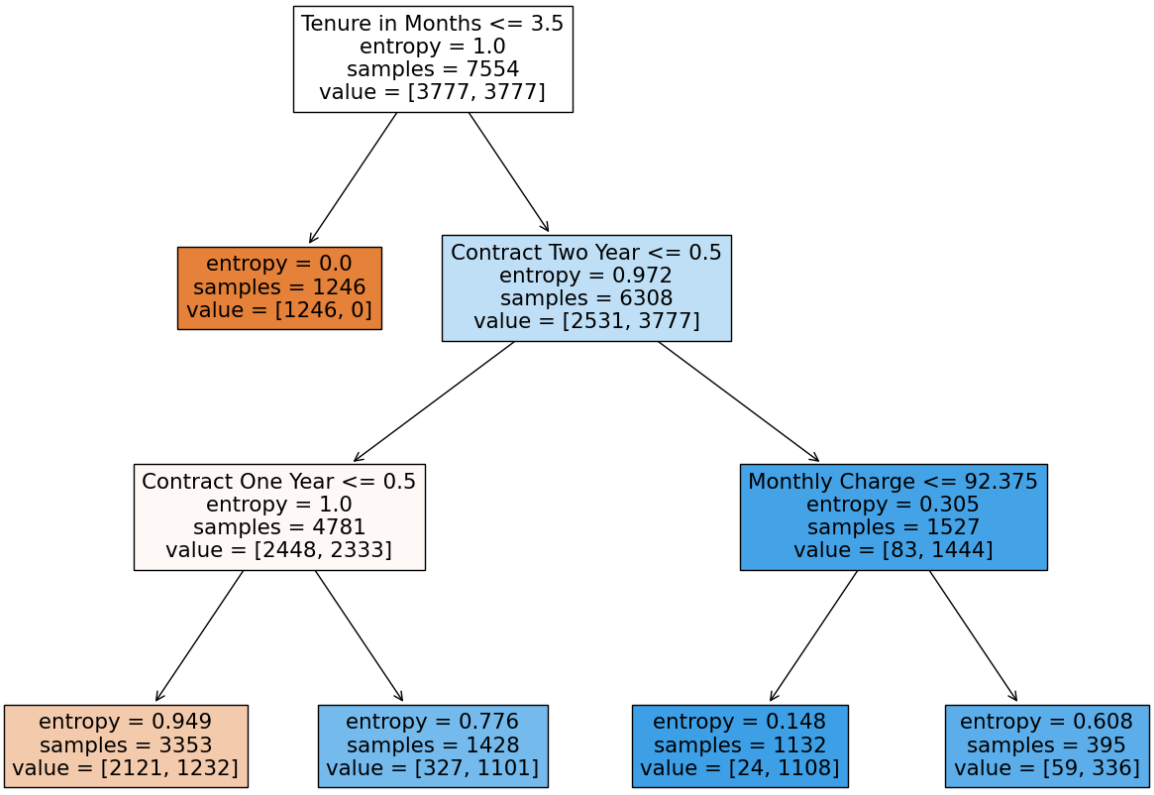

Figure 29

7.2.4. Gaussian Naive Bayes

Algorithm brief overview: Gaussian Naive Bayes (GNB) is a probabilistic classifier based on Bayes' theorem, with the assumption that features are independent and normally distributed. It's efficient and effective for datasets with continuous features, making it suitable for predicting Customer Status.

Data preprocessing: Scaling does not impact the performance of GNB, for this model, the same training process is performed for three versions of data: 1/ unscaled and unbalanced data, 2/ unscaled and over-sampling data, and 3/ unscaled and under-sampling data.

In this model we have processed the data in three ways 1/ Unbalanced data, 2/ Oversampling data, and 3/ Under sampling data

Model evaluation: Below are the performance from each version of data

No	Data version	Accuracy Score	Balanced Accuracy Score	Precision of Churned	Recall of Churned	F1
1	Unscaled and unbalanced data	0.711684	0.740691	0.500000	0.810000	0.610000
2	Unscaled and over-sampled data	0.695751	0.739997	0.480000	0.840000	0.610000
3	Unscaled and under-sampled data	0.694234	0.737331	0.480000	0.840000	0.610000

Table 8

Consider all models above, **unscaled and over-sampled data** gives the best performance. Although over-sampled and under-sampled version have similar performance, the over-sampled version is chosen over the under-sampled for consistency with other model types.

- Among 1318 customers in the testing set, there are 316 customers were correctly predicted to churn (True Positive), 601 customers were correctly predicted to stay (True Negative), 59 customers were predicted to stay but actually churned (False Negative), 342 customers were predicted to churn but actually stayed (False Positive).
- Recall of churned is 84% which means 84% of customers who actually churned were predicted to churn. This is a good performance.
- Precision of churned is 48% which means 48% of customers who were predicted to churn actually churn. This is not a good performance but acceptable when consider the precision-recall trade-off.
- Balanced accuracy score is 74% which means the model can correctly predict 74% of the balanced data.

7.2.5. Random Forest

Algorithm brief overview: Random Forest Classifier is an ensemble learning method that operates by combining many decision trees at training time and outputting the class that is the mode of the classes of the individual trees. This approach combines the simplicity of decision trees with flexibility, resulting in a robust model against overfitting. Random Forest is well-suited for this dataset, providing a reliable method to predict Customer Status.

Data preprocessing: Although scaling data does not impact the performance of Random Forest, in this model, we still use the same training process for four versions of data: 1/ unscaled and unbalanced data , 2/ scaled and unbalanced data, 3/ unscaled and balanced data and 4/ scaled and balanced data.

Parameter tuning and model configuration: In each version of data, different combinations of parameters are tested using 2 nested for-loop commands until the best-performing model is identified:

- For “max_depth”: from 1 to 20
- For “criterion”: "gini", "entropy", "log_loss"

Model evaluation: Below are the top 5 performing models from each version of data

With unscaled and unbalanced data:

No	Combination	Accuracy Score	Balanced Accuracy Score	Precision of Churned	Recall of Churned	F1
58	i=20, k=entropy	0.883156	0.82920	0.859935	0.704000	0.774194

59	i=20, k=log_loss	0.883156	0.82920	0.859935	0.704000	0.774194
49	i=17, k=entropy	0.880121	0.82467	0.855738	0.696000	0.767647
50	i=17, k=log_loss	0.880121	0.82467	0.855738	0.696000	0.767647
55	i=19, k=entropy	0.879363	0.82414	0.852941	0.696000	0.766520

Table 9

With scaled and unbalanced data:

No	Combination	Accuracy Score	Balanced Accuracy Score	Precision of Churned	Recall of Churned	F1
58	i=20, k=entropy	0.882398	0.82867	0.857143	0.704000	0.773060
59	i=20, k=log_loss	0.882398	0.82867	0.857143	0.704000	0.773060
55	i=19, k=entropy	0.880121	0.82467	0.855738	0.696000	0.767647
56	i=19, k=log_loss	0.880121	0.82467	0.855738	0.696000	0.767647
54	i=19, k=gini	0.881639	0.824928	0.863787	0.693333	0.769231

Table 10

With unscaled and balanced data:

No	Combination	Accuracy Score	Balanced Accuracy Score	Precision of Churned	Recall of Churned	F1
22	i=8, k=entropy	0.843703	0.846603	0.679406	0.853333	0.756501
23	i=8, k=log_loss	0.843703	0.846603	0.679406	0.853333	0.756501
19	i=7, k=entropy	0.831563	0.836513	0.658385	0.848000	0.741259
20	i=7, k=log_loss	0.831563	0.836513	0.658385	0.848000	0.741259
16	i=6, k=entropy	0.830046	0.83465	0.656315	0.845333	0.738928

Table 11

With scaled and balanced data:

No	Combination	Accuracy Score	Balanced Accuracy Score	Precision of Churned	Recall of Churned	F1
22	i=8, k=entropy	0.843703	0.846603	0.679406	0.853333	0.756501
23	i=8, k=log_loss	0.843703	0.846603	0.679406	0.853333	0.756501
19	i=7, k=entropy	0.831563	0.836513	0.658385	0.848000	0.741259
20	i=7, k=log_loss	0.831563	0.836513	0.658385	0.848000	0.741259
16	i=6, k=entropy	0.830046	0.834650	0.656315	0.845333	0.738928

Table 12

Consider all models above, with **unscaled and balanced data** the combination of **max_depth=8, criterion=entropy** gives the best performance. Although scaling data does not impact the performance of Random Forest, the unscaled version is chosen over the scaled one to prioritize the interpretation of individual trees.

- Among 1318 customers in the testing set, there are 320 customers were correctly predicted to churn (True Positive), 792 customers were correctly predicted to stay (True Negative), 55 customers were predicted to stay but actually churned (False Negative), 151 customers were predicted to churn but actually stayed (False Positive).
- Recall of churned is 85% which means 85% of customers who actually churned were predicted to churn. This is a good performance.
- Precision of churned is 68% which means 68% of customers who were predicted to churn actually churn. This is a good level of precision given the precision-recall trade-off.
- The harmonic mean of precision and recall is 76% (F1 score)
- Balanced accuracy score is 85% which means the model can correctly predict 85% of the balanced data. This is a good performance.

The unbalanced versions have larger max_depth (17, 19, 20) than the balanced version (6, 7, 8) because with unbalanced data deeper trees may requires to capture the minority class effectively. A higher max_depth allows the trees to learn better details when one class dominates the dataset. When the data is balanced, a shallow tree can be more effective because it does not need to go deep to learn the difference between classes. More important, a lower max_depth can reduce the chance of overfitting.

7.2.6. Comparison and discussion

This section compares and discusses the performance of five different model types: Logistic Regression, KNN, Decision Tree, Gaussian Naive Bayes, and Random Forest. The table below summarizes the top-performing models from all five different model types together with their data version, parameter combination and performance details.

Data version	Parameter combination	Model	Accuracy Score	Balanced Accuracy Score	Precision of Churned	Recall of Churned	F1	Ranking
Scaled and balanced data		Logistics	81.49%	82.81%	62.77%	85.87%	72.52%	2
Scaled and balanced data	n=8, m=canberra, w=uniform	KNN	75.80%	79.47%	54.64%	88.00%	67.42%	4
Unscaled and balanced data	i=3, k=entropy	Decision Tree	73.07%	77.96%	51.54%	89.33%	65.37%	3
Over sampling data		Gaussian Naïve Bayes	69.58%	74.00%	48.00%	84.00%	61.00%	5
Unscaled and balanced data	i=8, k=entropy	Random Forest	84.37%	84.66%	67.94%	85.33%	75.65%	1

Table 13

As mentioned in earlier part, the evaluation principle is identifying the model with high Recall of Churned class but still have an acceptable level of Precision of Churned class. Based on this predefined principle, the comparative analysis is as below.

Random Forest:

- The strongest accuracy and balanced accuracy score (84% and 85%).
- The highest harmonic mean of precision and recall of 76% (F1 score) proves that both precision and recall are high.
- The highest precision of 68% means there are fewer customers who were predicted to churn but stayed, hence the company does not have to waste the cost of retaining wrong customers.
- The use of multiple decision trees and averaging their predictions likely contributed to its robustness and generalization capabilities.
- Therefore, the ranking for this model is highest.

Logistic Regression:

- Similar performance to Random Forest, strong accuracy and balanced accuracy score (81% and 83%)
- Highest recall of 86% indicating its strength in correctly identifying churned customers, however, the F1 score of 73% only comes highest second
- Therefore, this model is ranked second, showing high effectiveness despite its simple nature compared to Random Forest.

Decision Tree:

- The highest recall of 89% indicates its effectiveness in identifying churned customers despite of its simple nature with just 3 simple splits
- However, the precision is relatively low at 52%, which means that nearly half of the customers are incorrectly predicted to churn
- The simplicity of a single tree might have limited its predictive power compared to the ensemble approach of Random Forest.
- Therefore, this model is ranked third, with performance metrics slightly lower than Logistic Regression but higher than KNN.

KNN:

- Similar to Decision Tree, KNN has the second highest recall of 88% showing its good performance in predicting churned customers.
- However, it has a low precision at just 55%, a bit better than Decision Tree but still low.
- This model is ranked fourth.

Gaussian Naive Bayes:

- High recall of churned at 84% but it has the lowest precision at just 48% which means that more than half of the customers are incorrectly predicted to churn
- It also has the lowest F1 score, accuracy score and balanced accuracy score among five models.
- Therefore, this model is ranked last.

There seems to be a relationship between the complexity and performance of the model, with more complex models like Random Forest outperforming simpler ones. However, Logistic Regression's performance indicates that simpler models can still be very effective, depending on the dataset.

Each model shows different trade-offs among the performance metrics. For instance, while Decision Tree has the highest recall, its precision is relatively lower than other models. This suggests the need to choose

a model based on the specific business objective – in this case, there is still a need to maintain an acceptable precision.

The performance differences between models also highlight the importance of data preprocessing and parameter tuning. Indeed, all the models perform better with balanced data than the unbalanced version. This emphasizes the impact of addressing data imbalance.

7.2.7. Feature selection and Cross validation

Beside the traditional way of training models with a single split, the project team performed cross validation without feature selection assesses the overall model performance, and performed cross validation with feature selection to optimize and simplify models by choosing the most important features.

Cross validation without figure selection: Below is the performance of five models using Cross validation without Feature selection. Random Forest still performs well in terms of accuracy (87.45%) and balanced accuracy (81.90%) compared with other models. However, the recall for the churned class dropped to 69.07%, which is lower compared to the traditional single split for Random Forest.

For all models, the overall performance (accuracy score and balanced accuracy score) are improved with cross validation compared with the traditional single split. However, the recall of churned class of all models also lower than those in the traditional single split in the Table 13 above.

This is understandable because cross validation improves the overall performance and not just focus on any class. Lower recall in the models also are observed with cross validation because of the imbalance data in each fold. This can be addressed by some advanced technique like the SMOTE but will not be covered in this assignment.

Data version	Parameter combination	Model	Accuracy Score	Balanced Accuracy Score	Precision of Churned	Recall of Churned	F1	Ranking
Scaled and unbalanced		Logistics	85.05%	81.24%	74.00%	72.00%	73.00%	4
Scaled and unbalanced	n=2, m=canberra, w=uniform	KNN	76.02%	77.96%	55.17%	82.45%	66.11%	2
Scaled and unbalanced	i=6, k=gini	Decision Tree	85.16%	82.14%	73.22%	75.17%	74.18%	3
Scaled and unbalanced		Gaussian NB	79.50%	79.84%	60.00%	81.00%	69.00%	1
Scaled and unbalanced	i=20, k=entropy	Random Forest	87.45%	81.90%	83.83%	69.07%	75.74%	5

Table 14

Cross validation with figure selection: Below is the performance of five models using Cross validation with Feature selection. For the parameters tuning used in feature selection, which ideally we should re-evaluate parameters for each significant subset of features, we used the best parameter from previous steps (cross validation without figure selection) due to the problem of resource and time consuming.

Random Forest still performs well with accuracy (87.56%) and balanced accuracy (82.65%). With figure selection the overall performance of all model slightly increase compare with table 14 above. However, similar to table 14, the overall performance of all model is improved but the recall of churned class decrease compared with the traditional single split. The main reason is cross validation with feature selection help to optimize the overall performance and the imbalance in data still exist in each fold caused the poor

performance of recall of churned class. And again, this imbalance in each fold can be addressed by some advanced technique like the SMOTE but will not be covered in this assignment.

Data version	Parameter combination	Model	Accuracy Score	Balanced Accuracy Score	Precision of Churned	Recall of Churned	F1	Ranking
Scaled and unbalanced		Logistics	84.29%	81.81%	71.00%	76.00%	73.00%	4
Scaled and unbalanced	n=2, m=canberra, w=uniform	KNN	81.44%	81.91%	63.00%	83.00%	72.00%	1
Scaled and unbalanced	i=6, k=gini	Decision Tree	85.38%	82.79%	73.00%	77.00%	75.00%	3
Scaled and unbalanced		Gaussian NB	80.86%	81.37%	62.00%	83.00%	71.00%	2
Scaled and unbalanced	i=20, k=entropy	Random Forest	87.56%	82.65%	82.00%	71.00%	76.00%	5

Table 15

8. Conclusion

8.1. Final model

The comparative analysis of different models demonstrates that no single model totally outperforms others across all metrics. The choice of the best model depends on the specific business objectives, in this case, the model must have a high recall of churned but still have to maintain an acceptable precision of churned.

The Random Forest model consistently performed well in terms of overall performance in all three approaches (single split, cross validation without feature selection, cross validation with feature selection), indicating its robustness and effectiveness in this problem.

Compare all models in all three approaches, ***Random Forest is the best overall model*** (traditional split with unscaled and balanced data and the combination of max_depth=8, criterion=entropy) in this analysis. Still, the performance of logistic regression suggests that simpler models can be effective and should not be overlooked. The analysis also underscores the importance of appropriate data preprocessing and parameter combination testing in model performance.

8.2. Recommendations and impacts of the project

Main recommendations for the company

- **Recommendation 1:** The company can stop offer E which is a new offer (low Total Revenue) and only accept among new customers because it does not perform well (low Monthly Charge), and more people who churned when accepted this offer. They should promote Offer A and B instead.
- **Recommendation 2**: The company should have more promotion programs to promote Long Distance Service because Total revenue mainly comes from Total Charges and Total Long Distance Charge (Total Refunds and Total Extra Data Charges are small)
- **Recommendation 3:** The company should improve their competitiveness by researching about competitors' services and improving their own services because there are many long-term customers and high-value customers who churned, they have higher average Total Revenue and Monthly Charge than all customers, and most of them leaving because of Competitor, Dissatisfaction and Price.

- **Recommendation 4:** The company should have more promotion programs for Married customers because they tend to stay longer and single customers and they generate more revenue.
- **Recommendation 5:** The company should focus on converting monthly customers to yearly customers instead of acquiring new customers, and adjust business strategy to generate more revenue from monthly customers because yearly customers create more revenue and the company already have so many monthly customers.

Impact of the project

- The project's outcomes can guide strategic decisions, enhancing customer retention and improving overall business performance.
- By understanding and addressing the factors leading to churn, improve overall customer satisfaction and loyalty, contributing to long-term business success.
- Prevent revenue loss due to churn and optimize marketing spend by focusing on high-risk customers, thus improving cost efficiency.

Potential and future development of the project

- To further improve model performance, explore more advanced feature selection techniques or parameter tuning for Random Forest. Additionally, addressing the class imbalance issue in cross validation, as mentioned earlier, can also help improve recall.
- Extend the model to predict customer behaviours besides churn, such as the likelihood to purchase additional products or responsiveness to promotions.
- Continuously update the model with new data and perform feature analysis to identify emerging trends or changing customer behaviours over time.
- Integrate predictive models into Customer Relationship Management systems for real-time prediction and response, enhancing customer relationship strategies.

This project demonstrates the potential of machine learning in translating customer data into actionable insights, driving strategic decisions, and enhancing customer engagement. By continuously evolving the models and strategies based on data-driven insights, businesses can stay ahead in a competitive landscape and foster strong, lasting relationships with their customers.

9. Appendix

9.1. Research questions and answers

	Questions	Answers
I/	**Descriptive Questions**	
1	Gender proportion between customers who exited and who stayed	Female is approximately the same with Male in both group
2	Age average between the customers who exited and those who stayed	Stayed: 45, Churned: 49
3	Age average between the customers who exited and who stayed by gender	Both gender are the same. Stayed: 45, Churned: 49
4	Population density between the customers who exited and who stayed	Stayed: 21k, Churned: 24k
5	The relation between Tenure and Customer Status	Stayed: 41 months. Churned: 10 months
6	What is the most popular offer among all customers	None -> B -> E
7	What is the most popular offer among the customers who stayed	None -> B
8	What is the most popular offer among the customers who exited	None -> E
9	Find out the most popular offer among Married customers and not married customers	Married: B, Single: E. Same with Q7 Q8 because in those who churned more are single, and in those who stayed more are married
10	Find the relation between population and customer status	Statyed live in slighty lower density area than Churned. Maybe in high density area there are more competitors

	Questions	Answers
11	What would be the relation between population and referrals	There is no relationship, all Number of Referrals have similar Population
12	Which gender refers more customers	Both genders are similar in Number of Referrals
13	Tenure vs offer	Customers who took Offer A are the longest customers, descreasing in Tenure from A to B, C, D and E. Offer E takers tend to be new customers, that is why Offer E is popular among Churned customer who are also just 10 months tenures in average (refer to Q5)
14	Charges vs Tenure	For monthly charge vs tenure: no clear relationship. For total revenue vs tenure: increase if tenure increase (of course it is)
15	Customer status vs Avg Monthly GB Download vs Avg Monthly Long Distance Charges	Avg Monthly GB Download: same for Churned and Stayed and Joined ~ 20-21. Avg Monthly Long Distance Charges: same for all status ~ 26-27
16	Gender Vs Multiple Lines vs Dependents	Same for both gender
17	Contract - Short-term vs long-term - Customer retention rate	Churned: mostly Month-to-Month, very few one-year and two-year. Statyed: two-year > month-to-month > one-year
18	Which offer is offering unlimited data	Not clearly, most of customers in every offer use Unlimited Data
19	Top customer by Revenue	Top 100 highest Total Revenue customers
19.1	Gender	50 ~ 50 => same the whole dataset
19.2	Age	Mostly in 33~60 => same the whole dataset
19.3	Area	Mostly in 0-38k => same the whole dataset
19.4	Marital Status	Majority is Married, when ~50:50 for the whole dataset
19.5	Dependent	Most has no, then 1 2 3 => same the whole dataset
19.6	Tenure	Marjority is 70-71, when 9-55 for the whole dataset
19.7	Customer status	Stayed nearly 10 times Churned, when it is nearly 3 times for the whole dataset
20	What is the popular offer and customer status among the customers who received refunds?	Same with whole population. Offer B slightly higher than other offers.
II/	**Predictive Questions**	
1	What are the factors that improve customer status	Age, Married, Number of Dependents, Number of Referrals, Tenure in Months, Phone Service, Monthly Charge, Population, Offer, Contract, Payment method, Paperless billing
2	Predict the customer status using 5 different predictive models (Logistic, KNN, Decision Tree, Naives Bayes, Random Forest)	Done
3	What is the best model among the above 5 models to predict customer status	Random Forest

9.2. References

Bahri-Ammari, N. and Bilgihan, A., 2019. Customer retention to mobile telecommunication service providers: the roles of perceived justice and customer loyalty program. *International Journal of Mobile Communications*, *17*(1), pp.82-107.

Willys, N., 2018. Customer satisfaction, switching costs and customer loyalty: An empirical study on the mobile telecommunication service. *American Journal of Industrial and Business Management*, *8*(04), p.1022.

Landis, T., 2022. *Customer Retention Marketing vs. Customer Acquisition Marketing.* [Online] Available at: https://www.outboundengine.com/blog/customer-retention-marketing-vs-customer-acquisition-marketing/#:~:text=Acquiring%20a%20new%20customer%20can,customer%20is%205%2D20%25. [Accessed 10 12 2023].

AL-Shatnwai, A.M. and Faris, M., 2020. Predicting customer retention using XGBoost and balancing

methods. *International Journal of Advanced Computer Science and Applications*, *11*(7).

Kumar, S., 2022. Customer retention versus customer acquisition. *Forbes Business Council (https://www. forbes. com/sites/forbesbusinesscouncil/2022/12/12/customer-retention-versus-customeracquisition*.

You, Y. and Joshi, A.M., 2020. The impact of user-generated content and traditional media on customer acquisition and retention. *Journal of Advertising*, *49*(3), pp.213-233.