Churn Prediction Analysis

Milestone: Model Performance Evaluation and Interpretation

Group 54

Amoolya Bagalkoti

Akshita Singh

bagalkoti.a@northeastern.edu

singh.akshi@northeastern.edu

Effort Contributed by Student 1: 50%

Effort Contributed by Student 2: 50%

Signature of Student 1

Signature of Student 2

Submission Date: 03/24/2023

Table of content

Problem Setting

The aim of churn prediction analysis is to identify the customers who are most likely to discontinue using a company's goods or services or cancel their subscription. Customer churning is one of the significant issues faced by many industries such as banking, ecommerce, telecommunications, and subscription-based services. In general, the purpose of churn prediction research is to spot at-risk clients early on, allowing a firm to take steps to keep them before they cancel their subscription or stop utilizing the firm's goods or services. This can be a difficult challenge since it frequently needs a lot of data, and it can be challenging to figure out which features of the data are most important to the problem.

Problem Definition

Telecom companies face a high rate of customer churn. These companies use data from past customer behavior and interactions to predict which customers are likely to cancel their subscriptions in the future. The customer churn prediction is often addressed as a binary classification with the objective of predicting whether a customer would churn or not based on a set of input attribute values. These input attributes could include behavioral data, transactional data, and/or demographic information like age and gender. This can assist businesses in taking proactive steps to keep loyal consumers and reduce revenue loss because of customer churn and come up with efficient customer retention strategy.

Data Source

The dataset is taken from Kaggle, an open-sourced public datasets platform.
www.kaggle.com/datasets/blastchar/telco-customer-churn

Data Description

The dataset comprises 7043 records. There are 21 attributes in total and 1 target attribute "Churn", which indicates where the customer is likely to stay or leave. Following are the attributes: -

Column	Attribute	Description
1	Customer ID	To identify each customer
2	Gender	Gender of the customer (male/female)
3	Senior Citizen	If the customer is a senior citizen or not (1/0)
4	Partner	If the customer has a partner or not (Yes/No)
5	Dependents	If the customer has dependents or not (Yes/No)
6	tenure	Number of months the customer has stayed with the company
7	PhoneService	If the customer has a phone service or not (Yes/No)
8	MultipleLines	If the customer has multiple lines or not (Yes/No/No phone service)
9	InternetService	Customer's internet service provider (DSL/Fiber optic/No)
10	OnlineSecurity	If the customer has online security or not (Yes/ No/No internet service)
11	OnlineBackup	If the customer has online backup or not (Yes/ No/No internet service)
12	DeviceProtection	If the customer has device protection or not (Yes/ No/No internet service)
13	TechSupport	If the customer has Tech support or not (Yes/ No/No internet service)

14	StreamingTV	
15	StreamingMovies	
16	Contract	Type of contract (month-to-month/ one year/ Two year)
17	PaperlessBilling	If they have signed up for paperless billing option (yes/no)
18	PaymentMethod	Mode of payment by the customer (electronic, bank transfer, mailed check, credit card)
19	MonthlyCharges	Monthly bill rate
20	TotalCharges	Total charges of the plan

Data Collection

In this project, our goal is to estimate customer churn by analyzing a telecom dataset that was obtained from Kaggle. The dataset includes a variety of aspects, including usage patterns, customer service statistics, and demographics about the customers. We aim to create a model that can reliably predict if a client is likely to churn or not using exploratory data analysis, data preparation, and machine learning approaches. The conclusions and predictions drawn from this analysis can aid telecom firms in making decisions on client loyalty and retention.

Data Cleaning

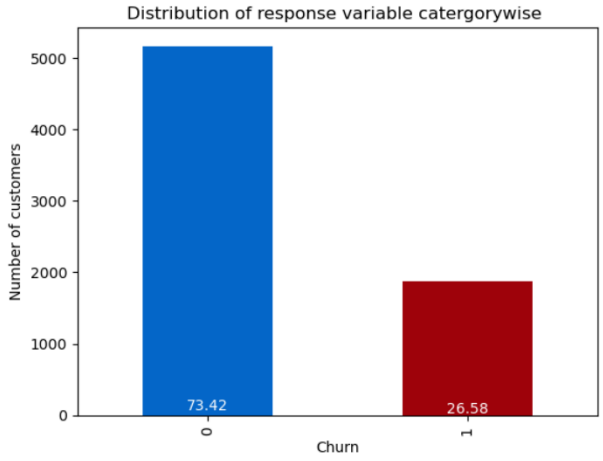
The attribute "CustomerID" was a unique identifier for each customer and hence was removed as it was an unnecessary variable for the analysis.
The number of null values in the dataset in each column was calculated, there was only one column "TotalCharges" that had null values. The records with nulls in "TotalCharges" had "tenure" as zero years despite having values in the "MonthlyCharges" column which is quite
contradictory, and since the nulls made up to only 0.15% of missing values, the respective rows were dropped.

Data Exploration

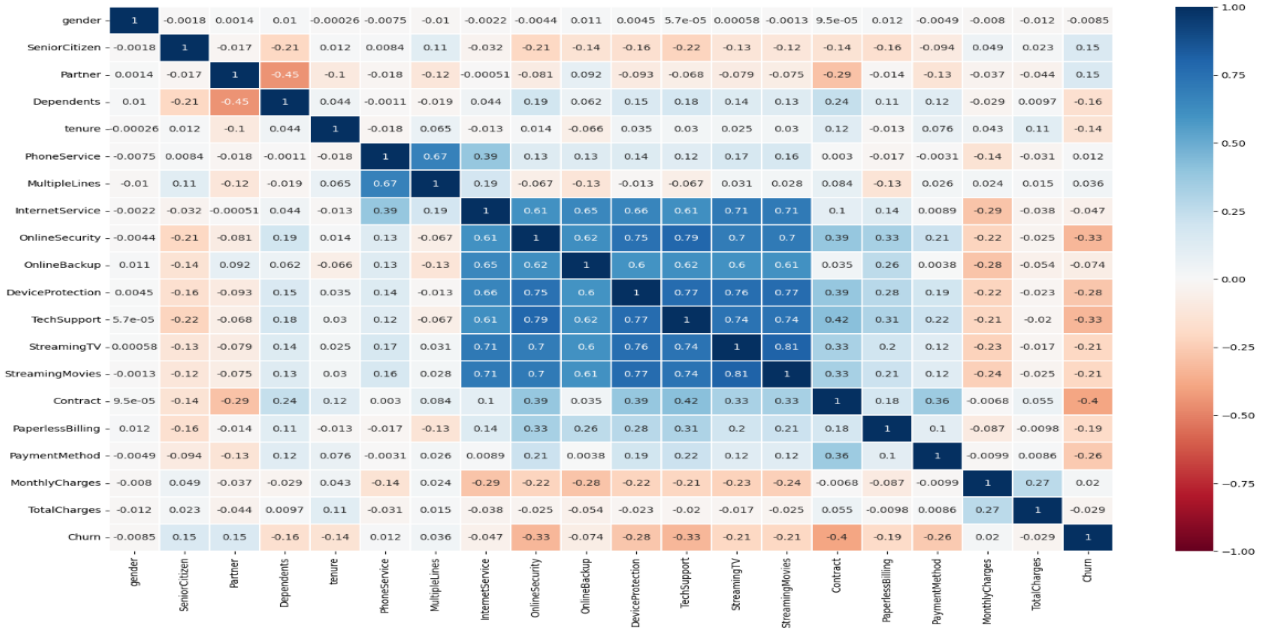
The column "TotalCharges" was originally identified as object type. It was converted into a numeric datatype before checking for and removing the null values.
The shape of the data set after cleaning the data of null values and dropping the unique identifier column is (7032, 20). The dataset now consists of 3 numeric variables, 6 binary categorical variables and 10 nominal variables and 1 response variable "Churn" which is converted into a numeric data type to hold the values 1(churned) and 0(loyal).

Data Visualization

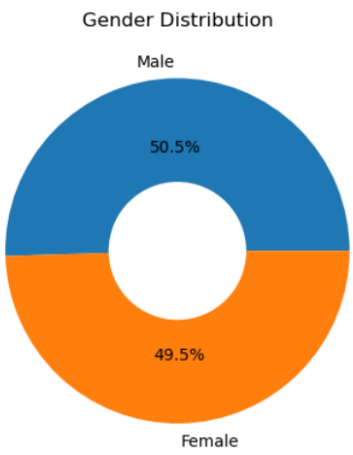
We start by finding the number of churn and no churn cases in the dataset. This gives us an idea of how skewed or imbalanced the dataset is. By plotting the bar chart, we can see that the data is skewed as the ratio of classes is 73:27.

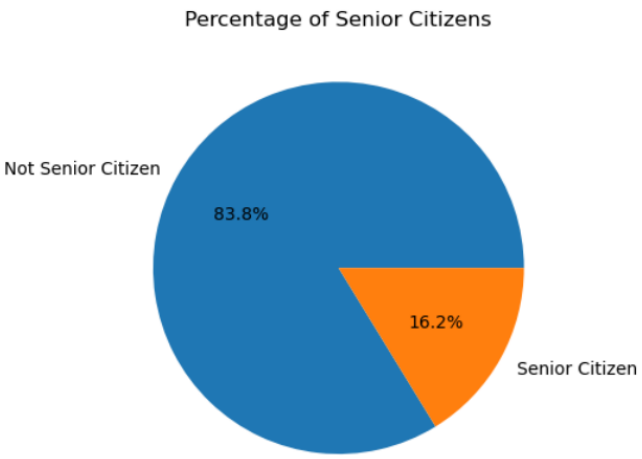


The variables that are related to the response variable by performing an initial correlation analysis among the variables can be found using a heatmap. As per the heatmap Gender, Phone service, Multiple lines, Internet Service, Online Backup, Monthly Charges and total charges seem to have no correlation with churning. But we cannot still conclude that these variables are not associated with the response variable. Hence, we need to further check the trends between individual variables with churn to get and definite answer.

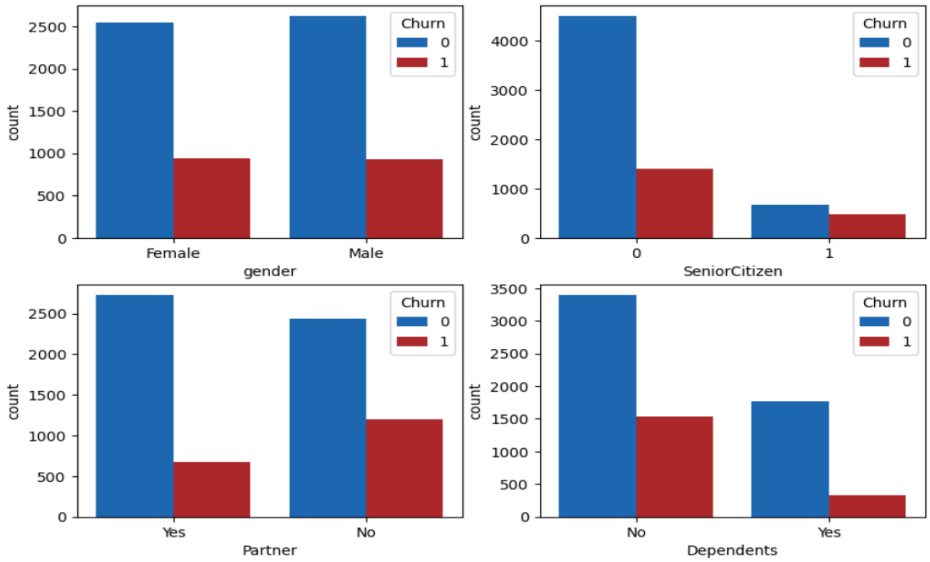


We then perform a few univariate analyses to understand the patterns in the data better and form some assumptions. In order to find any intriguing trends, we will first examine the distribution of the various variables. Let's first examine the demographics of the clients, including their gender, age range, partner status, and dependency status.

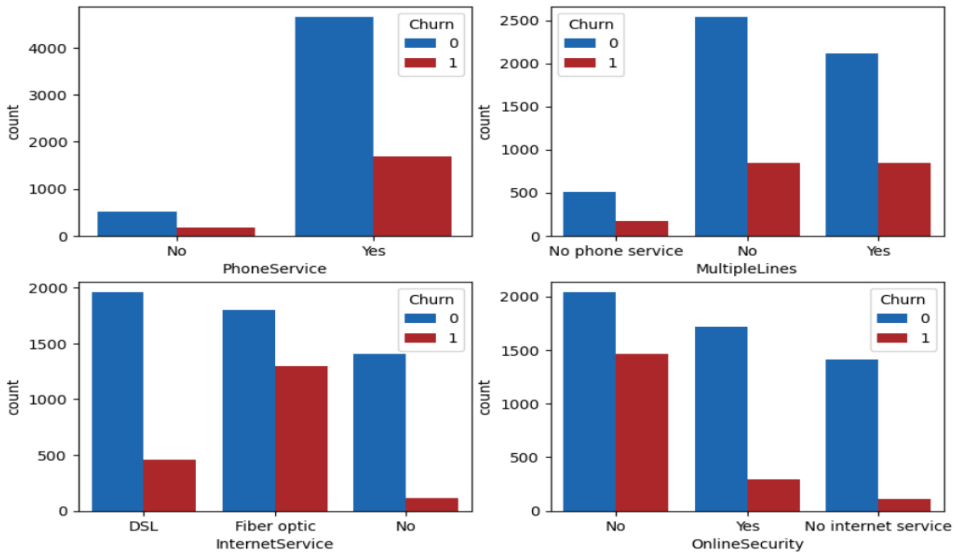




Of the clients in our data collection, roughly half are men and the other half are women. Only 16% of the clients are beyond the age of sixty. The statistics shows that young people make up the majority of customer base.

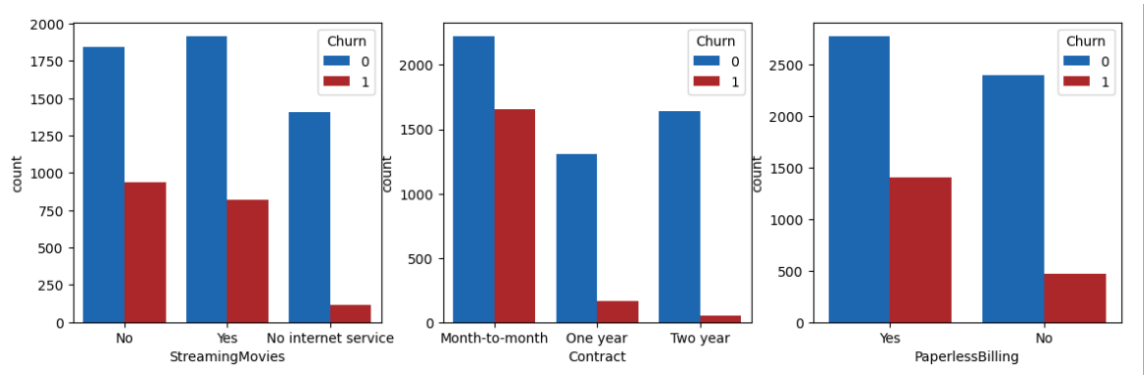


From the above plots we get a stronger answer that gender has no effect on the churn rate as seen in the heatmap. Further, we see that the rest of the attributes do have some impact on the response variable. Even if the observed percentage of seniors is quite low, the majority of them churn. Customers who do not have partners are more likely to churn compared to the ones that have a partner. Customers who don't have dependents have a higher churn rate.

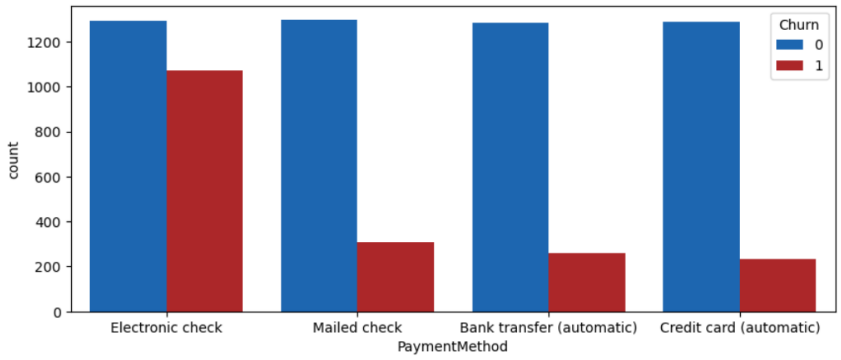


The absence of online security makes most customers churn. Many customers choose fiber optic service, and a significant rate of customer churn exists within this group. This may reveal a problem with the fiber optic service that had many of its customers unhappy; additional investigation could lead to the discovery of a better and more suitable alternative.

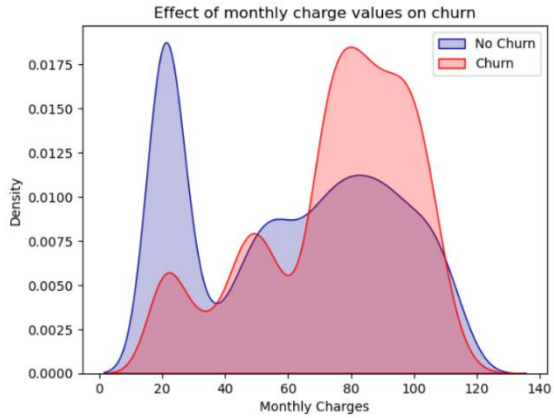
Customers who choose DSL service are more abundant and had a lower rate of turnover than those who chose fiber optic service.

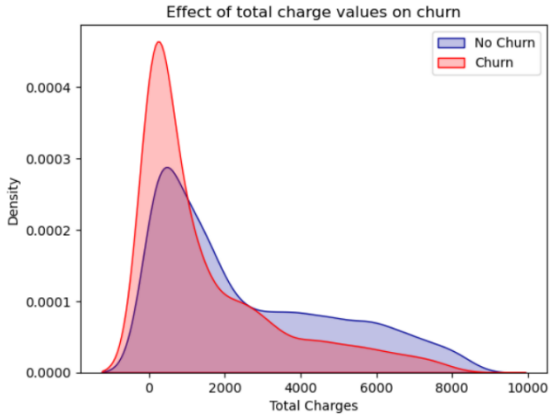


Paperless Billing seems like one of the reasons why customers are most likely to churn. Approximately 75% of customers with a month-to-month contract have chosen to leave. On the contrary, customers with one-year and two-year contracts have very low chances of discontinuing. We have also seen that people who have no internet streaming are likely to continue the service.

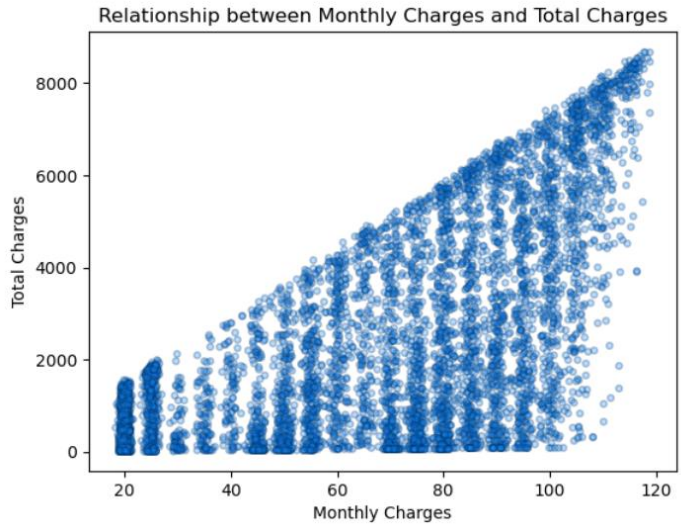


Most of the customers who left used electronic checks as their primary payment method, whereas those who used credit-card automatic transfers, bank automatic transfers, and mailed checks were less likely to switch.

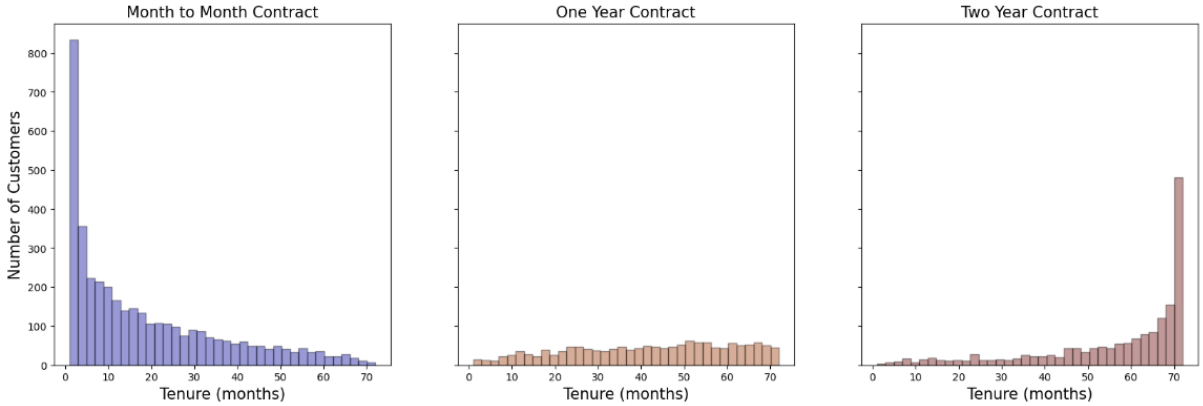




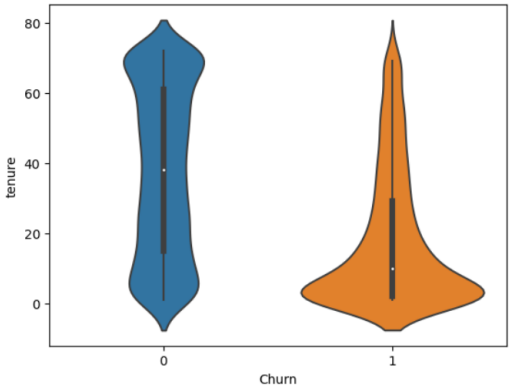
We see a higher churn rate with an increase in monthly charges and the opposite case in total charges.



As we saw in the heatmap that total charges and monthly charges seemed to have some positive correlation between them. The above scatterplot also confirms that as monthly bill increases the total charge for a customer increases.



Many monthly contracts only last a few weeks to a few months, although two-year contracts often continue for over 70 months. This demonstrates that clients who sign longer contracts are more devoted to the business and have a propensity to stick with it.

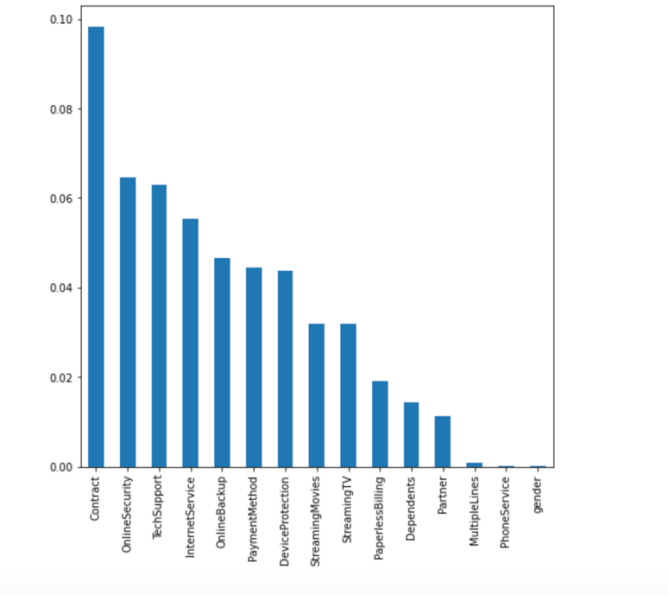


Customers who have lesser tenure are more likely to churn than the customers who are on the higher scale of tenure.

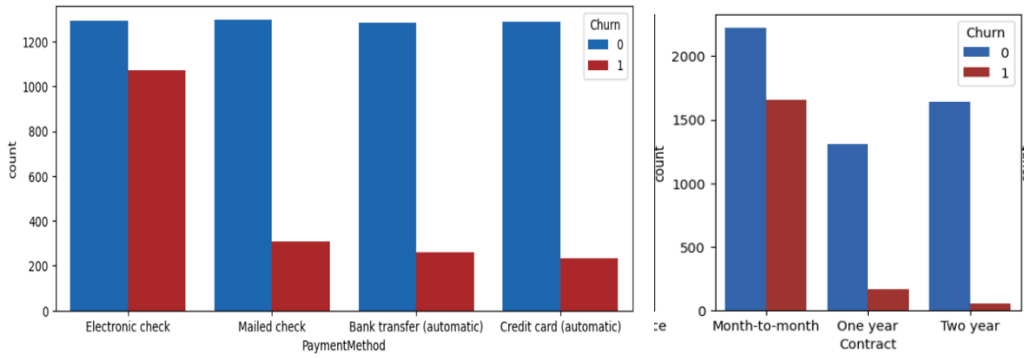
Data Preprocessing:

The dataset has 16 categorical variables. Out of which 6 are binary categorical variables and 10 are nominal categorical variables. The remaining i.e., monthly charges, total charges and tenure are of numeric type.

The mutual_info_score() function is made use to find the degree of dependency of every column with the target variable column. The variables with higher mutual information scores have higher impact on predicting the response variable and the columns with lower values of mutual information score will have less impact the response variable.



It is observable from the bar chart of each categorical variable with their respective mutual information scores that the variables gender, MultipleLines and PhoneService have values very close to zero and these columns can be dropped.



Furthermore, for the variables “Contract” and “PaymentMethod” it is evidently clear from the bar graphs that the churn rate is much higher for month-on-month contract and Electronic check payment method respectively. Thus, the value of Electronic check in PaymentMethod can be encoded as 1 and remaining categories can be considered 0. Similarly for the Contract column, the category month-on-month can be encoded as 1 and the remaining categories as 0.

The binary categorical variables (Partner, Dependents, PaperlessBilling, PhoneService) are encoded using the LabelEncoder() from sklearn.preprocessing. The remaining nominal categorical variables are encoded with One-Hot encoding technique by converting the columns into dummy columns with the use of get_dummies().

Before moving on to the model selection and model building, the data in the numeric attributes must be standardized to create consistency and reduce bias due to magnitude. Along with that we also need to deal with class imbalance issues.

Exploration of Data mining Models:

Support Vector Machines (SVM):
SVM works by sorting observations into different groups. It tries to find the best line or curve to draw between the groups in a way that gives the largest possible gap between them. This helps to make sure that the datapoints are sorted correctly, and new data can be easily sorted in the future.
This model can be an effective choice in predicting the churn because we are dealing with large dataset having many features. It can also handle noise and outliers in the data.

Logistic Regression:
Logistics regression is a prediction algorithm for classification of categorical data based on the concept of probability. The output of a linear regression model is converted by the logistic regression algorithm into a probability score between 0 and 1 using a sigmoid function. This probability score indicates the possibility of an event occurring given a set of independent variables.
Advantages of using logistic regression is that it typically shows good performance for binary classification and it can show robustness towards the noise in the data. On the downside, logistic regression model might get biased towards the majority class with imbalanced data. It also assumes a linear relation between the predictors and the response variable.

Random Forest:
Random Forest combines multiple decision trees to produce a more robust model that is less susceptible to overfitting. The algorithm builds each decision tree by randomly selecting a subset of features and samples from the data, and then combines the predictions of all trees to make a final prediction. To optimize performance, the number of trees in the forest, as well as other hyperparameters such as the maximum depth of each tree, can be tuned.
A Random Forest classifier might not perform well if the data is high dimensional.

KNN (K-Nearest Neighbors):
The KNN algorithm works by locating the k nearest data points in the training set to a given query data point in the feature space, and then using these k data points to predict the class of the query point. A majority vote among the k neighbors determines the predicted class. The value of k is a hyperparameter that must be specified before training the model.
KNN is a non-parametric algorithm, which means that no assumptions are made about the underlying data distribution. It is susceptible to the noise in the data and might show bias towards the majority class.

Decision Tree:
A decision tree is a supervised learning algorithm that can be used for classification as well as regression. It is a non-parametric method for creating a tree-structured model. Each internal node of the tree represents a decision based on a feature's value, whereas each leaf node represents a class label or a regression problem value. The algorithm's goal is to build a tree that predicts the target variable as accurately as possible by deciding which features to split on at each level. Decision trees can capture non-linear relationships between features and response variables. Unlike some other models, decision trees make no assumptions about the underlying distribution of the data. But on the other hand, there is a risk of bias and overfitting of the data.

Above are a few data mining models which can be utilized to predict if a customer will churn or not. To finalize a model, all the training models must be implemented, and performance measure scores have to be

considered. The model/models that will show best accuracy and minimum error can be selected as final model for the churn prediction analysis.

Implementing Machine Learning Models:

After performing EDA, finding correlation and mutual information we understood the most important features and dropped 7 columns out of 21. We had 14 predictors for model building.

The following models were implemented, and we obtained the results as follows:

1. Logistic Regression:

   By performing logistic regression on the training data, we got an accuracy of 0.79. The model's precision for class 0 (non-churn) was 0.90, which indicates that 90% of all predictions for this class were accurate. According to the recall for class 0, which is 0.83, 83% of the actual data that belonged to this category were correctly classified by the model.

   The model's precision for class 1 (churn) was 0.51; which indicates that only 51% of the predictions for class 1 were accurate. Recall for class 1 is 0.65, which indicates that out of all the actual observations, 65% of them were properly classified as being in class 1. Class 1's F1-score is 0.57.

```
Classification report of Logistic Regression with imbalance data :
              precision    recall  f1-score   support

           0       0.90      0.83      0.86      1666
           1       0.51      0.65      0.57       444

    accuracy                           0.79      2110
   macro avg       0.70      0.74      0.72      2110
weighted avg       0.82      0.79      0.80      2110
```

2. Random forest classifier:

   The accuracy of the random forest classifier for the churn prediction analysis is 0.79. With a precision of 0.90 and recall of 0.83 for non-churn customers vs a precision of 0.51 and recall of 0.65 for churn customers. The F1-score is 0.86 for non-churning clients and 0.57 for churning ones. Again, the classifier performs better at predicting non-churn consumers.

```
------------------------------------------------------------
Classification report for Random forest classifier with imbalanced data :
              precision    recall  f1-score   support

           0       0.90      0.83      0.86      1666
           1       0.51      0.65      0.57       444

    accuracy                           0.79      2110
   macro avg       0.70      0.74      0.72      2110
weighted avg       0.82      0.79      0.80      2110

------------------------------------------------------------
```

In both cases, the model appears to be performing relatively well for class 0, but not so well for class 1. We need to enhance the classifier's accuracy for churn prediction. Therefore, we need to employ oversampling

or under sampling methods to correct the data's imbalance. Hence, we are using the SMOTE technique to deal with the imbalance.

```
The number of classes before fit Counter({0: 3622, 1: 1300})
The number of classes after fit Counter({0: 3622, 1: 3622})
```

Now, both the classes are balanced and have 3622 data points in them. Again, we deployed the above-mentioned models, and we obtained the results as follows:

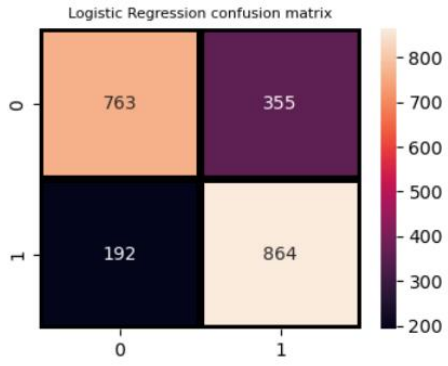
1. Logistic Regression:

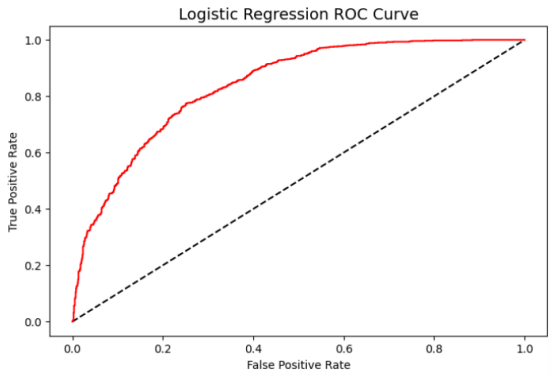
   The logistic regression model achieved an accuracy of 75% on predicting churn using balanced data. It had a precision of 68% for non-churn and 82% for churn customers, and a recall of 80% for non-churn and 71% for churn customers. The f1-score for non-churn was 0.74 and for churn it was 0.76.

```
Accuracy: 0.7483900643974241
Classification report of Logistic Regression with balance data :
              precision    recall  f1-score   support

           0       0.68      0.80      0.74       955
           1       0.82      0.71      0.76      1219

    accuracy                           0.75      2174
   macro avg       0.75      0.75      0.75      2174
weighted avg       0.76      0.75      0.75      2174
```





2. Support Vector Classifier:

   Churn prediction models utilize multiple features, such as customer demographics, historical behavior, and usage patterns, SVMs can handle high-dimensional feature spaces efficiently and can identify relevant features that contribute to churn prediction, leading to accurate and interpretable results.

   Compared to the previous logistic regression model, the SVM model has a lower overall accuracy of 63% (vs. 75% for the logistic regression model). It also has lower f1-scores for both non-churn (0.65 vs. 0.74) and churn (0.60 vs. 0.76) customers.

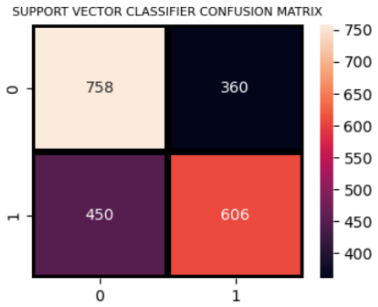
   In terms of precision and recall, the SVM model has similar precision for both non-churn (0.63 vs. 0.68) and churn (0.63 vs. 0.82) customers compared to the logistic regression model. However, it has lower recall for both non-churn (0.68 vs. 0.80) and churn (0.57 vs. 0.71) customers.

```
------------------------------------------------------------
Classification report for SVC with balanced data :
              precision    recall  f1-score   support

           0       0.63      0.68      0.65      1118
           1       0.63      0.57      0.60      1056

    accuracy                           0.63      2174
   macro avg       0.63      0.63      0.63      2174
weighted avg       0.63      0.63      0.63      2174

------------------------------------------------------------
```



3. Decision Tree:

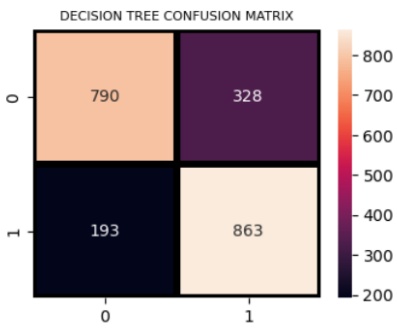
Decision trees can handle both categorical and numerical features, making them versatile for churn prediction tasks that may involve a mix of different types of data.

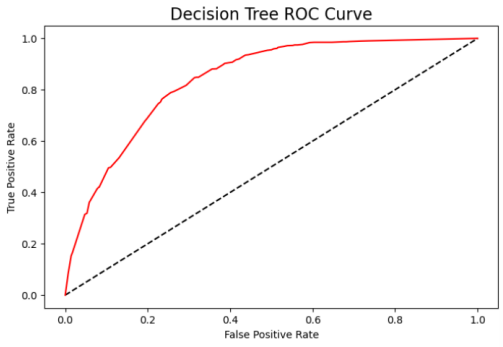
The Decision Tree model has a higher accuracy (76%) than the logistic regression (75%) and SVM models (63%), and similar f1-scores for both non-churn and churn customers compared to the logistic regression model. The model has higher precision for non-churn customers (0.80) and similar precision for churn customers (0.72) compared to both logistic regression and SVM models. The recall for non-churn is lower than logistic regression but higher than the SVM model.
Decision Tree model is performing better than both logistic regression and SVM models for predicting churn on this dataset.

```
Accuracy score : 0.7603495860165593
Classification report for Decision Tree with balanced data :
              precision    recall  f1-score   support

           0       0.80      0.71      0.75      1118
           1       0.72      0.82      0.77      1056

    accuracy                           0.76      2174
   macro avg       0.76      0.76      0.76      2174
weighted avg       0.77      0.76      0.76      2174
```





4. Random Forest Classifier:

The Random Forest classifier has a similar accuracy and f1-scores compared to the Decision Tree model. It has higher precision for churn customers but lower precision for non-churn customers compared to both Logistic Regression and Decision Tree models. The recall for both non-churn and

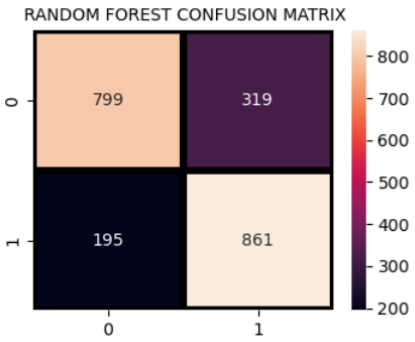
churn customers is lower than the Decision Tree model, but higher than the SVM model.

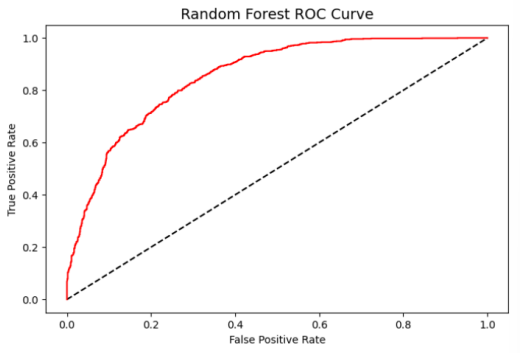
```
0.7635694572217111
------------------------------------------------------------
Classification report for Random forest classifier with balanced data :
              precision    recall  f1-score   support

           0       0.71      0.80      0.76       994
           1       0.82      0.73      0.77      1180

    accuracy                           0.76      2174
   macro avg       0.77      0.77      0.76      2174
weighted avg       0.77      0.76      0.76      2174

------------------------------------------------------------
```





5. K-NN Classifier:

Knn is a good model for predicting churn as it is simple and has the ability to capture local patterns in the data. It also can capture non-liner relationships between various features.

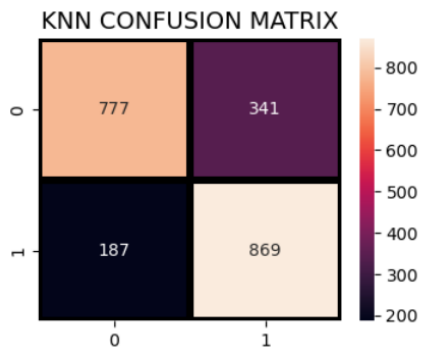
In terms of precision and recall, the KNN classifier has higher precision for non-churn customers (0.81) compared to all previous models, but lower precision for churn customers (0.72) compared to the Random Forest and Decision Tree models. The recall for non-churn customers (0.69) is lower than the recall of the Decision Tree model but higher than the recall of the Logistic Regression and Random Forest models. The recall for churn customers (0.82) is similar to the recall of the Decision Tree and Random Forest models. Overall, the KNN classifier is performing similarly to the Decision Tree and Random

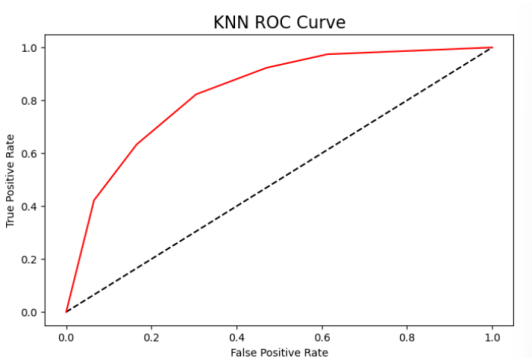
```
------------------------------------------------------------
Classification report for KNN classifier with balanced data :
              precision    recall  f1-score   support

           0       0.81      0.69      0.75      1118
           1       0.72      0.82      0.77      1056

    accuracy                           0.76      2174
   macro avg       0.76      0.76      0.76      2174
weighted avg       0.76      0.76      0.76      2174

------------------------------------------------------------
```





Performance evaluation:

Based on the performance metrics provided, the KNN classifier model seems to be the best one as it has the highest recall (0.82) and f1-score (0.77) among all models, indicating that it correctly identifies a higher percentage of churn customers and has a good balance of precision and recall. The Decision Tree and Random Forest models also perform well, with similar accuracy and f1-score to the KNN classifier, but slightly lower recall for churn customers.

	accuracy	recall	precision	f1
LogisticRegression	0.748390	0.818182	0.708778	0.759560
SVC	0.627415	0.573864	0.627329	0.599407
DecisionTree	0.760350	0.817235	0.724601	0.768135
RandomForest	0.763569	0.815341	0.729661	0.770125
KNN	0.757130	0.822917	0.718182	0.766990

Conclusion

In conclusion, our project on customer churn analysis has shown that KNN classifier model has the highest recall and f1-score among all the models we tested. It correctly identifies a higher percentage of churn customers and has a good balance of precision and recall. The Decision Tree and Random Forest models also perform well, with similar accuracy and f1-score to the KNN classifier. These models have slightly lower recall for churn customers compared to KNN. By identifying factors that drive customer churn, our analysis provides insights for developing targeted retention strategies. It enables companies to take proactive measures to prevent customer churn, which can lead to increased customer loyalty and revenue. Our project has provided a competitive advantage by understanding customer behavior better than competitors. Overall, our analysis of customer churn can help businesses make informed decisions to retain customers and increase their profitability.