# Telco Customer Churn Rate Analysis


Completed by: Huzaifa Abid
Completed for:

STUDENT NAME HUZAIFA ABID
Student ID: 3520262470413
**Email**: huzaifa.abid50@gmail.com

## Contents

# BACKGROUND AND PROBLEM

## 1.1 BACKGROUND

The telco industry is focused on customer retention and attracting new customers. We assist telco companies in minimizing contract terminations (churn rate) and increasing customer acquisition. Predicting churn provides insights into customer retention efforts, helping companies understand why customers leave. By comparing offerings with competitors, companies can ensure competitiveness and meet customer preferences. Churn analysis guides the company in providing incentives to maintain customer loyalty, considering the higher cost of acquiring new customers. Our churn prediction analysis enables telco companies to proactively address retention challenges, enhance products, and foster long-term customer satisfaction and loyalty.

## 1.2 DATA SOURCE

The telco company has a dataset available on Kaggle, sourced from the IBM sample set collection. The dataset comprises information about 7,043 customers in California who utilize home and internet services. Our objective is to assist the company in predicting customer behaviour to enhance retention and develop targeted customer retention programs.

1. The dataset includes the following information:
2. Demographic details: Gender, age, and marital status of customers.
3. Customer account information: Length of time as a customer, paperless billing preference, payment method, monthly charges, and total charges.
4. Services subscribed to by customers: Phone service, multiple lines, internet service type, online security, online backup, device protection, and tech support.
5. Customer churn: Indication of customers who terminated their services within the last month.

By analysing this comprehensive dataset, we aim to provide insights that enable the telco company to implement targeted strategies for customer retention and develop initiatives to meet customer needs effectively. To ensure a more objective and consistent approach to candidate evaluation

### 1.3 RESEARCH OBJECTIVE

- Which is the most important factor that contributes to the high retention rate?
- Which analytics model can accurately predict a customer's churn rate?
- What are the advantages and disadvantages of using different analytical models?
- How could the telco company use our analysis to develop focused retention programs?

### 1.4 JUSTIFICATION OF THE RESEARCH

Churn analysis is crucial for understanding customer attrition and improving products/services. We employ various supervised learning models to analyse customer churn behaviour and provide actionable insights. Our research aims to reduce churn by targeting specific customer demographics, account information, usage behaviour, and subscribed services. This helps the telco company enhance customer retention and optimize its operations.

## 2. DATA SUMMARY AND EXPLORATOR ANALYSIS

### 2.1 DATA INTRODUCTION

After reading the data using Pandas in Python, we observed that the raw data set had no missing values. The majority of features, including gender, phone service, and payment method, were categorical. However, Monthly Charges and Total Charges were numerical data. Here are the summary of the data:

```
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
```

```
customerID          7043
gender                 2
SeniorCitizen          2
Partner                2
Dependents             2
tenure                73
PhoneService           2
MultipleLines          3
InternetService        3
OnlineSecurity         3
OnlineBackup           3
DeviceProtection       3
TechSupport            3
StreamingTV            3
StreamingMovies        3
Contract               3
PaperlessBilling       2
PaymentMethod          4
MonthlyCharges      1585
TotalCharges        6531
Churn                  2
dtype: int64
```

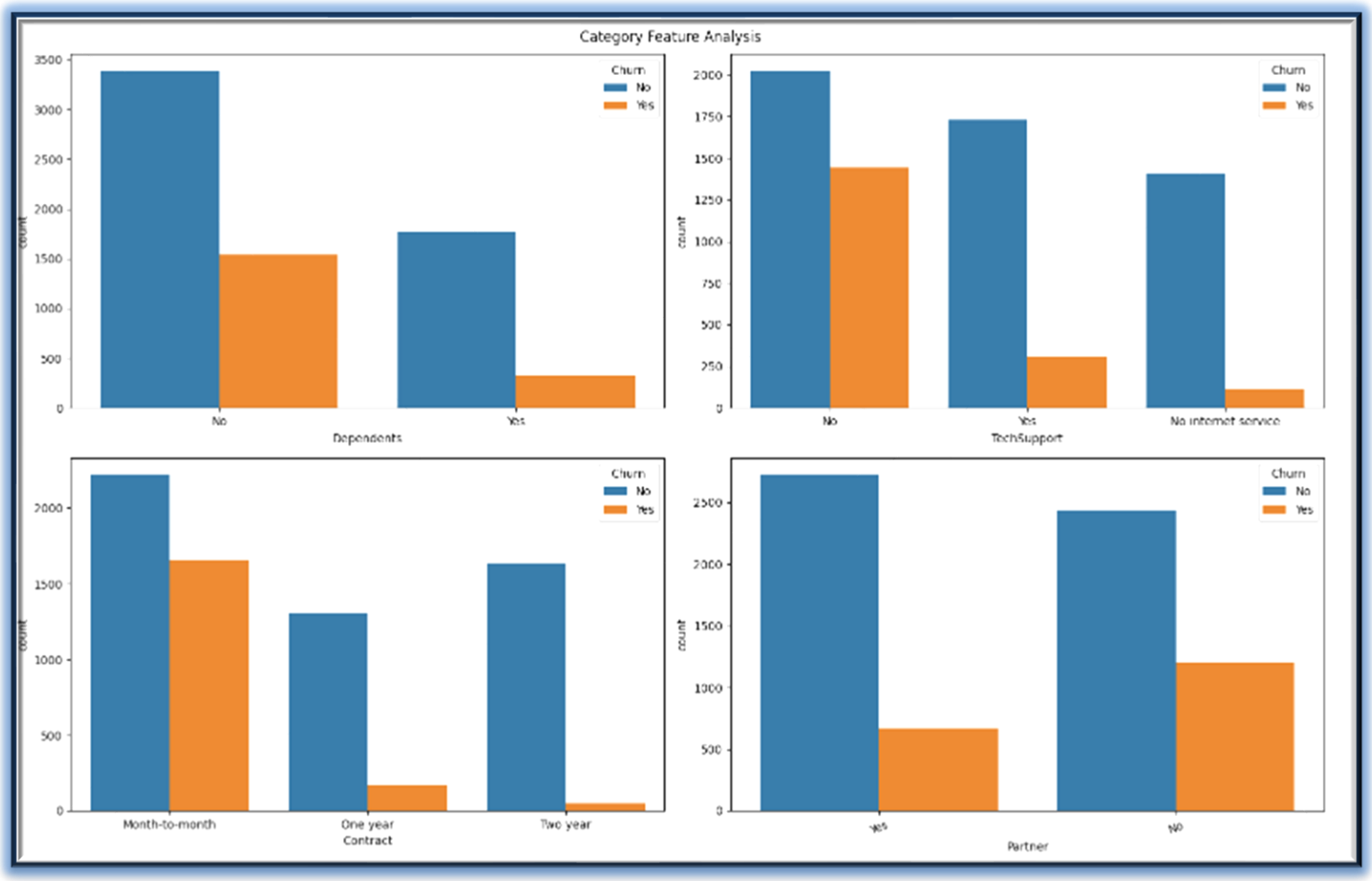
```
df.describe()
```

| | SeniorCitizen | tenure | MonthlyCharges |
|---|---|---|---|
| **count** | 7043.000000 | 7043.000000 | 7043.000000 |
| **mean** | 0.162147 | 32.371149 | 64.761692 |
| **std** | 0.368612 | 24.559481 | 30.090047 |
| **min** | 0.000000 | 0.000000 | 18.250000 |
| **25%** | 0.000000 | 9.000000 | 35.500000 |
| **50%** | 0.000000 | 29.000000 | 70.350000 |
| **75%** | 0.000000 | 55.000000 | 89.850000 |
| **max** | 1.000000 | 72.000000 | 118.750000 |

As shown above, the data set contains 7043 observations and 21 columns. Apparently, there are no null values on the data set; however, we observe that the column Total Charges was wrongly detected as an object. This column represents the total amount charged to the customer and it is, therefore, a numeric variable. For further analysis, we need to transform this column into a numeric data type. To do so, we can use the pd.to_numeric function. By default, this function raises an exception when it sees non-numeric data; however, we can use the argument errors='coerce' to skip those cases and replace them with a NaN.

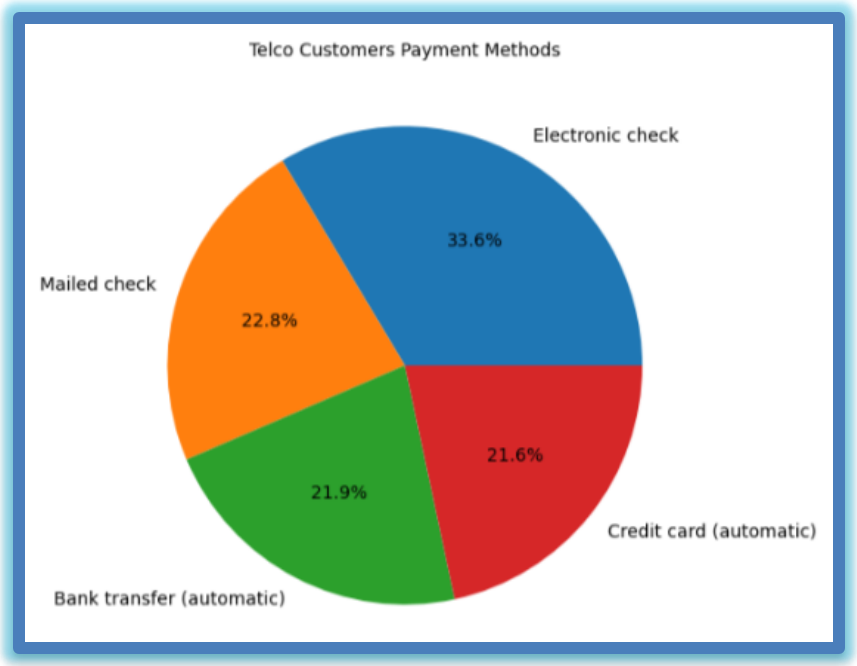
```
df['TotalCharges']=pd.to_numeric(df.TotalCharges, errors='coerce')
```

## 2.3 EXPLORATORY DATA ANLYSIS



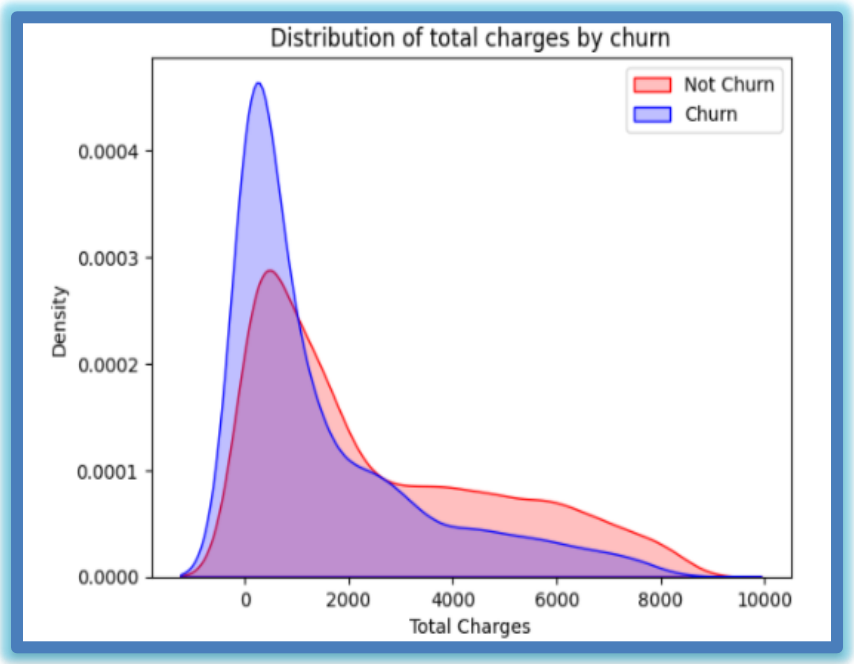

As shown Above:

1. Dependent, the customer who doesn't have Dependents will be more likely to churn for the Telco service.
2. Internet Service, it looks like most people are using Fiber internet and the customers who subscribe to Fiber internet are more likely to Churn.
3. Customers who have month to month subscription are more likely to churn
4. Partner, the customer who doesn't have a partner will be more likely to churn for the Telco service.




As show above:

- the pie chart shows the percentage of usage of payment methods by the customers




- The above kdeplot shows customer total charges and their churn rate

## 2.3 FACTOR CONVERSION

A lot of our data columns' default values are binary: they are either " Yes" or "No" so I convert them into int values.
There are three columns of features that should be coded into factors:

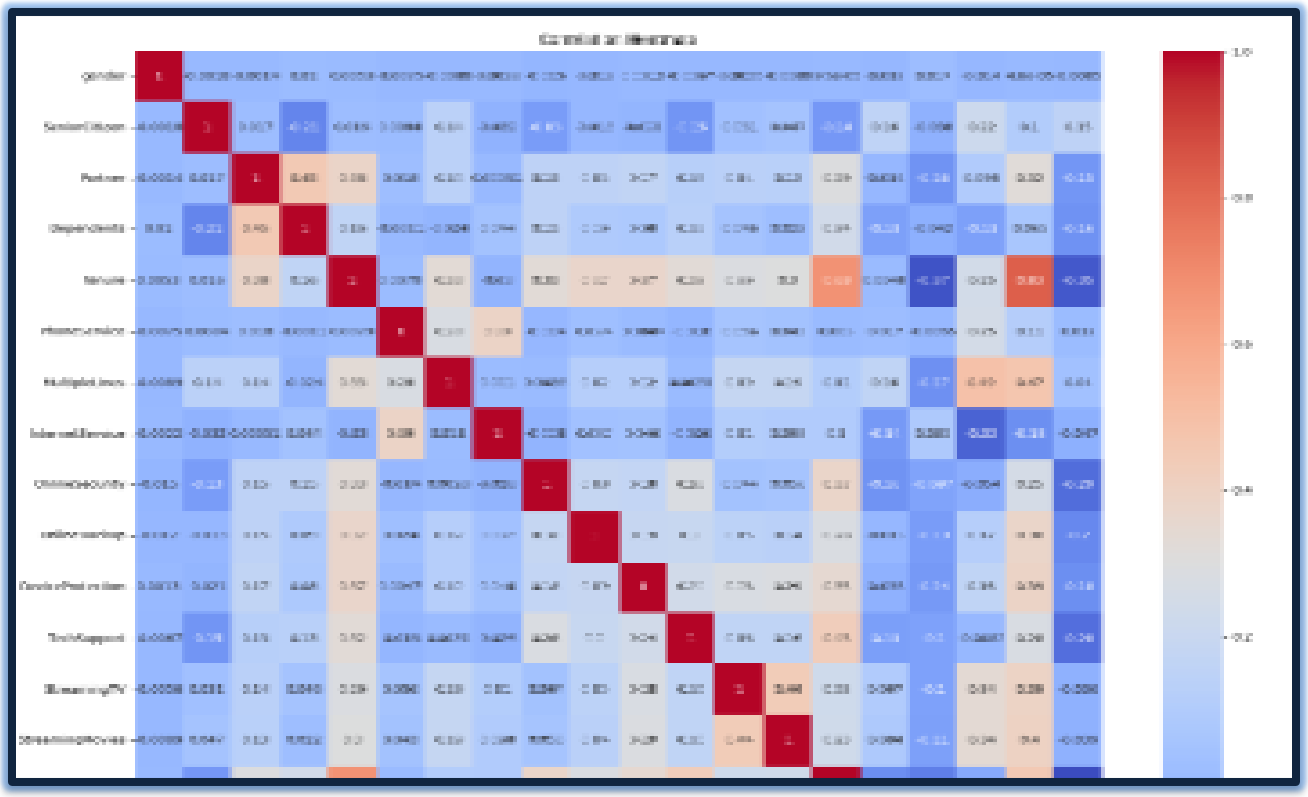
- Internet Service which includes DSL, Fiber optic, and no.
- The contract which includes month to month, one year, and two year.
- Payment Method which includes Bank transfer, Credit Card, Electronic Check, and Mailed check.

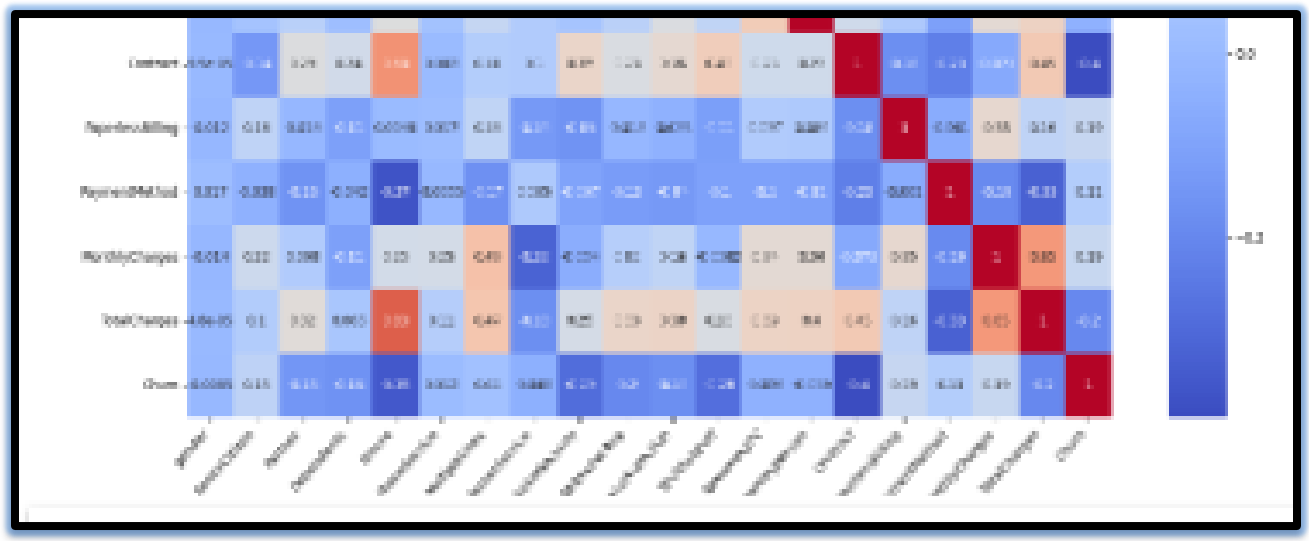
| | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | MultipleLines | InternetService | OnlineSecurity | OnlineBackup |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 0 | 0 | 0 | 34 | 1 | 0 | 0 | 2 | 0 |
| 2 | 1 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 2 | 2 |
| 3 | 1 | 0 | 0 | 0 | 45 | 0 | 0 | 0 | 2 | 0 |
| 4 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 1 | 0 | 0 |

| DeviceProtection | TechSupport | StreamingTV | StreamingMovies | Contract | PaperlessBilling | PaymentMethod | MonthlyCharges | TotalCharges | Churn |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 1 | 2 | 29.85 | 29.85 | 0 |
| 2 | 0 | 0 | 0 | 1 | 0 | 3 | 56.95 | 1889.50 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 3 | 53.85 | 108.15 | 1 |
| 2 | 2 | 0 | 0 | 1 | 0 | 0 | 42.30 | 1840.75 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 2 | 70.70 | 151.65 | 1 |

The project will start on March 17, 2023, and is expected to be completed by April 10, 2023. The project schedule will be divided into several phases, including planning, development, testing, and deployment.

## 2.4 CORELATIONAL HEAT MAP





- After converting all of the categorical data using Label Encoding and encoder, we ran a pair-wise correlation for all of the features
- From the heatmap, we could see that the features ‘Contract’ and ‘Tenure’ have a high correlation. It makes sense because these features measure the loyalty of the customer.

# 3. MODEL IMPLEMENTION AND KEY FINDINGS

## 3.1 MODEL INTRODUCTION

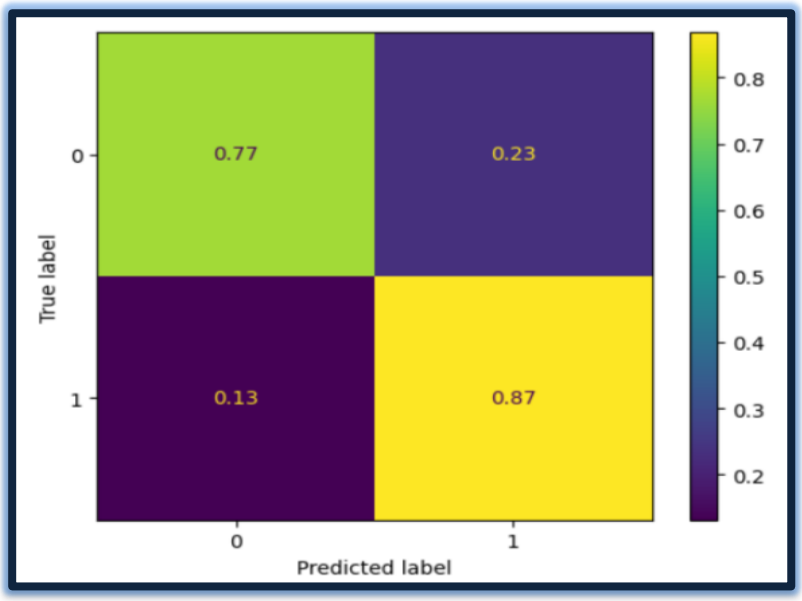
The random forest model consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction, and the class with the most votes becomes our model’s prediction.

In our case, the pro side includes (1) It generally provides high accuracy and balances the bias-variance trade-off well. (2) It can be used as feature importance visualization. (3) It is not influenced by outliers to a fair degree. (4) It can handle both linear and nonlinear relationships. And the cons are (1) It is much harder to interpret compared to previous models. (2) It will take a much longer time if the dataset is huge.

## 3.2 RANDOM FOREST AND EXPLORATION

As for comparing the accuracy of different models, we first split the dataset into 70% as training and 30% as testing. We then refit the models and we have summarized the result in the AIC/BIC table below.

## 3.3 CONFUSION MATRIX



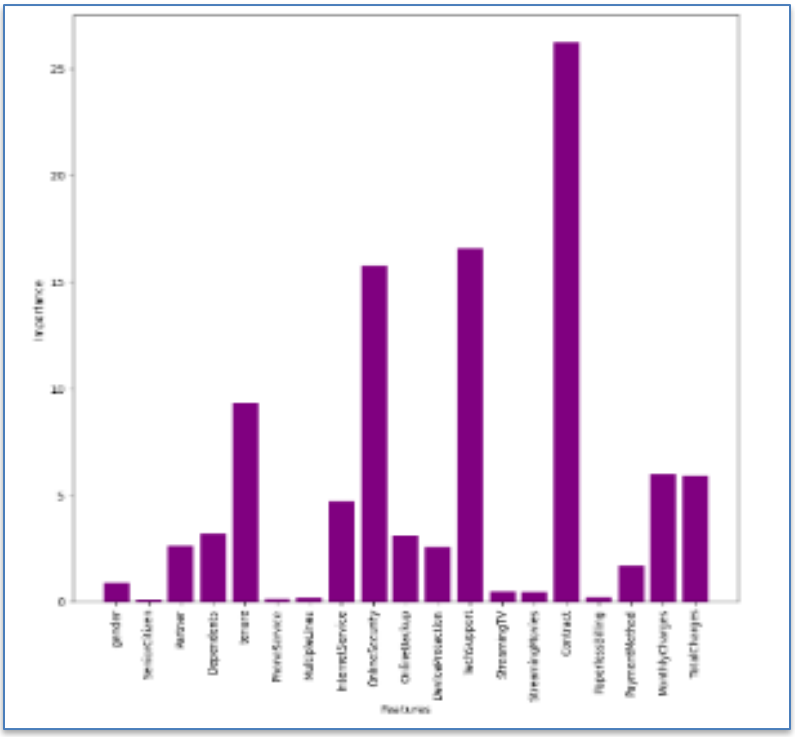
We can also use the aggregated confusion matrix to distil plenty of useful information.

### 3.4 Important information from Confusion Matrix

```
Accuracy: 0.8170377541142304
Precision: 0.8209487127429085
Recall: 0.8170377541142304
F1 Score: 0.8166599773929797
Accuracy: 81.70%
```

## 4. IMPORTANT FEATURES RELATED TO LABEL

### 4.1 IMPORTANT FEATURES



The above figure shows that the bars with more hight are more Important features related to the label Churn.

## 5.Conclusion

The importance of this type of research in the telecom market is to help companies make more profit. It has become known that predicting churn is one of the most important sources of income to telecom companies. Hence, this research aimed to build a system that predicts the churn of customers. These prediction models need to achieve high AUC values. To test and train the model, the sample data is divided into 80% for training and 20% for testing. We have applied feature engineering, effective feature transformation and selection approach to make the features ready for machine learning algorithms. In addition, we encountered another problem: the data was not balanced. Only about 5% of the entries represent customers' churn.