Predicting Telecom Churn:

A Machine Learning Approach to Retain Customers

Abstract

Churn is a one of the biggest problems in the telecom industry. Research has shown that the average monthly churn rate among the top 4 wireless carriers in the US is 1.9% - 2%.

Customer attrition is one of the biggest expenditures of any organization. Customer churn otherwise known as customer attrition or customer turnover is the percentage of customers that stopped using your company's product or service within a specified timeframe. For instance, if you began the year with 500 customers but later ended with 480 customers, the percentage of customers that left would be 4%. If we could figure out why a customer leaves and when they leave with reasonable accuracy, it would immensely help the organization to strategize their retention initiatives manifold.

In this project, we aim to find the likelihood of a customer leaving the organization, the key indicators of churn as well as the retention strategies that can be implemented to avert this problem

Introduction

Telecom Churn Prediction is a machine learning project that aims to predict customer churn in the telecom industry. Churn prediction is important for telecom companies because it helps them to retain their customers and reduce customer acquisition costs. The figure below depicts the description of the dataset importation and screenshot of same.

Problem statement

Telecommunication companies (telecom) have a significant problem with customer churn, which is the loss of customers who stop using their services. To tackle this problem, telecom need to identify the customers who are likely to churn and take proactive measures to retain them. Machine learning models can help telecom in predicting the customers who are most likely to churn, based on various factors such as customer demographics, usage patterns, and payment history.

Data Cleaning

The dataset used in this project contains 21 columns and 7,043 rows. The first step in the data cleaning process was to remove any duplicate rows in the dataset. After removing duplicates, there were 21 columns and 7,043 rows left in the dataset.

```
2]: teleco_churn=pd.read_csv('Telco-Customer-Churn.csv')
```

```
3]: teleco_churn.head()
```

3]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSup
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	

5 rows × 21 columns

The next step was to check for missing values. There were missing values in the columns 'Total Charges' and 'Churn'. The missing values in 'Total Charges' were removed since it constituted less than 0.15% of the dataset. Below represents the action.

```
10]: #Removing missing values
     teleco_churn.dropna(inplace = True)
```

```
11]: # Remove customer IDs from the data set
     df2 = teleco_churn.iloc[:,1:]
```

```
12]: teleco_churn.head()
```

12]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	Tec
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	

5 rows × 21 columns

Data Analysis

After cleaning the dataset, the next step was to perform exploratory data analysis (EDA). EDA is important because it helps to understand the distribution of the data and identify any trends or patterns in the data.

The first step in EDA was to analyze the distribution of the target variable 'Churn'. The distribution of 'Churn' was found to be imbalanced, with 73.4% of customers not churning and only 26.6% of customers churning. This imbalance in the target variable can lead to biased results.

We also check the distribution of the gender; it was revealed the distribution is also equal as male stood at 50.5% and 49.5%.

There are only 16% of the customers who are senior citizens. Thus, most of the customers in the dataset are younger people.

About 50% of the customers have a partner, while only 30% of the total customers have dependents.

Interestingly, among the customers who have a partner, only about half of them also have a dependent, while other half do not have any dependents.

Additionally, as expected, among the customers who do not have any partner, a majority (90%) of them do not have any dependents.

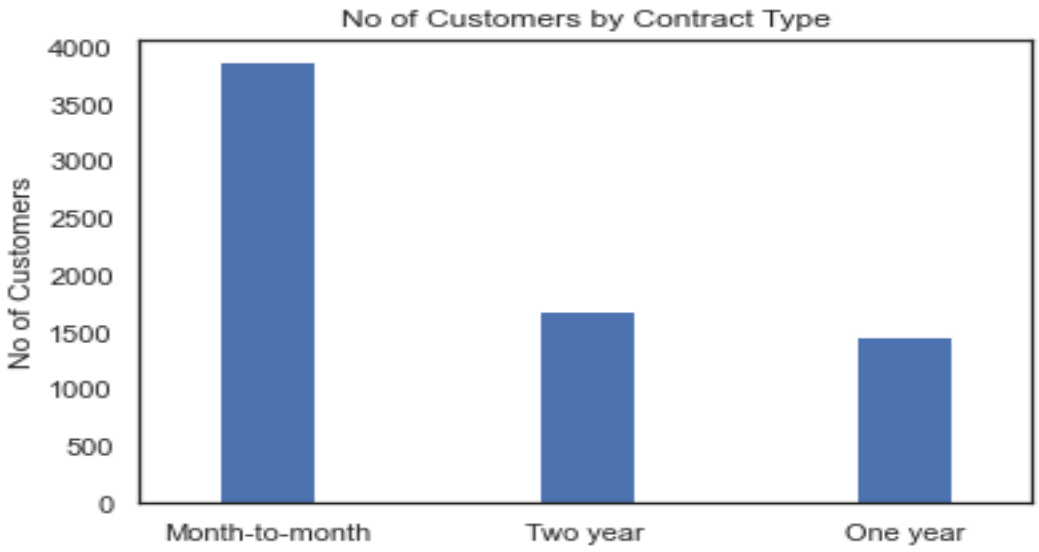
After looking at the below histogram we can see that a lot of customers have been with the telecom company for just a month, while quite a many are there for about 72 months. This could be potentially because different customers have different contracts. Thus, based on the contract they are into it could be more/less easy for the customers to stay/leave the telecom company.

The next step in EDA was to analyze the distribution of the independent variables. The following observations were made:

The majority of customers were on a month-to-month contract.

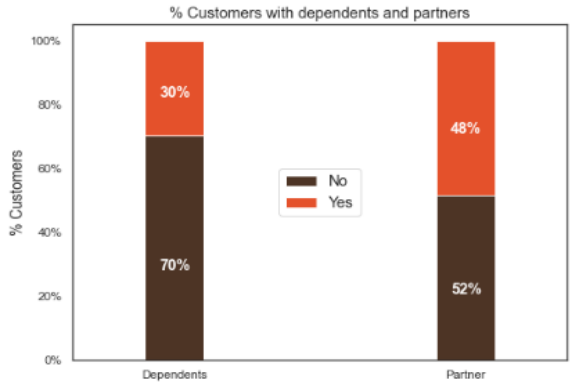
```
ax = teleco_churn['Contract'].value_counts().plot(kind = 'bar',rot = 0, width = 0.3)
ax.set_ylabel('No of Customers')
ax.set_title('No of Customers by Contract Type')

plt.show()
```



The majority of customers did not have a partner or dependents.

```
In [18]:
df2 = pd.melt(teleco_churn, id_vars=['customerID'], value_vars=['Dependents','Partner'])
df3 = df2.groupby(['variable','value']).count().unstack()
df3 = df3*100/len(teleco_churn)
colors = ['#4D3425','#E4512B']
ax = df3.loc[:,'customerID'].plot.bar(stacked=True, color=colors,figsize=(8,6), rot = 0, width = 0.2)
ax.yaxis.set_major_formatter(mtick.PercentFormatter())
ax.set_ylabel('% Customers',size = 14)
ax.set_xlabel('')
ax.set_title('% Customers with dependents and partners',size = 14)
ax.legend(loc = 'center',prop={'size':14})
for p in ax.patches:
    width, height = p.get_width(), p.get_height()
    x, y = p.get_xy()
    ax.annotate('{:.0f}%'.format(height), (p.get_x()+.25*width, p.get_y()+.4*height), color = 'white', weight = 'bold', size
```



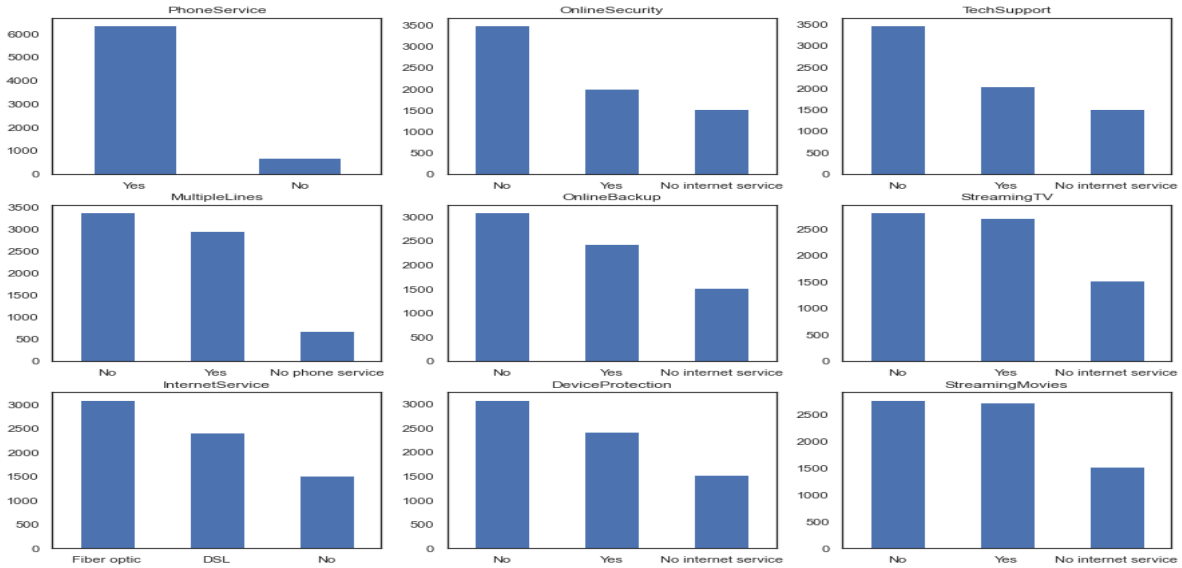
Partner and dependent status - About 50% of the customers have a partner, while only 30% of the total customers have dependents

1. **Distribution of various services used by customers**

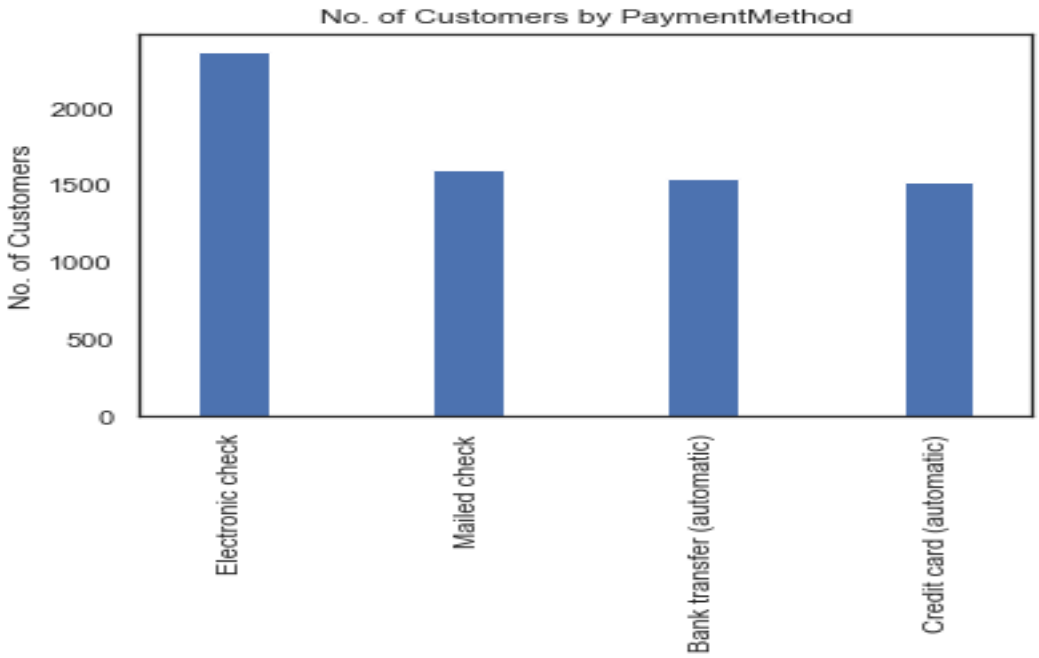
The majority of customers did not have multiple lines, online security services, Tech support services, Online Backup services, device protection services and stream TV services

The majority of customers had internet service and Phone services.

Almost equal number of the customers stream movies

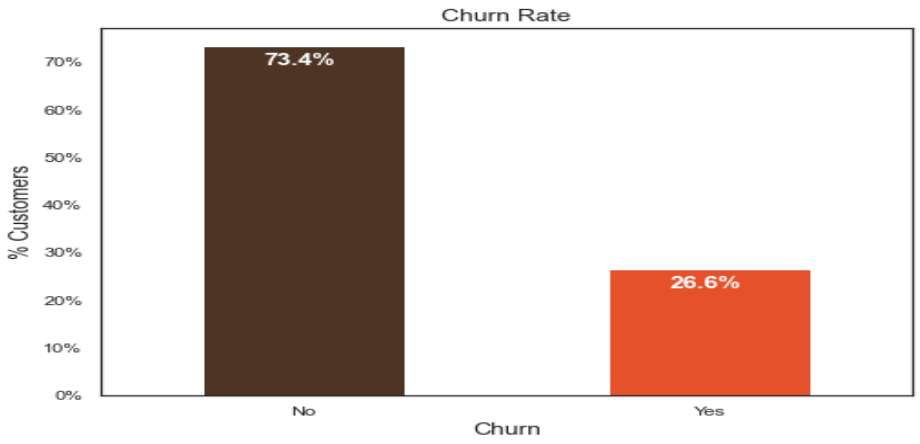


The majority of customers had a monthly payment method/plan.



Predictor variable (Churn) and its interaction with other variables

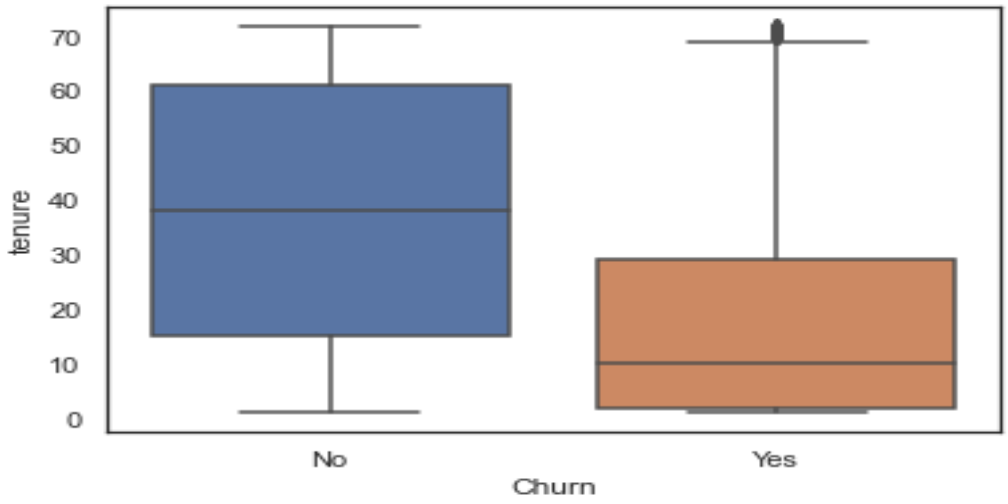
1. **Churn rate**



In our data, 74% of the customers do not churn. Clearly the data is skewed as we would expect a large majority of the customers to not churn. This is important to keep in mind for our modelling as skewedness could lead to a lot of false negatives. We will see in the modelling section on how to avoid skewness in the data.

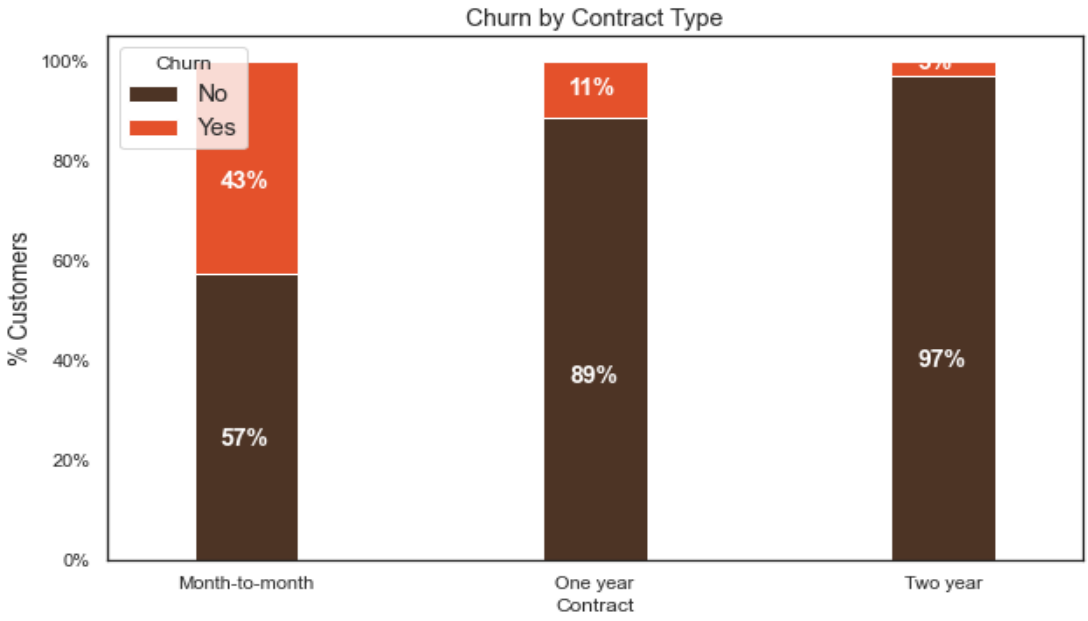
2. **Explore the churn rate by tenure, seniority, contract type, monthly charges and total charges to see how it varies by these variables**

a. churn rate by tenure



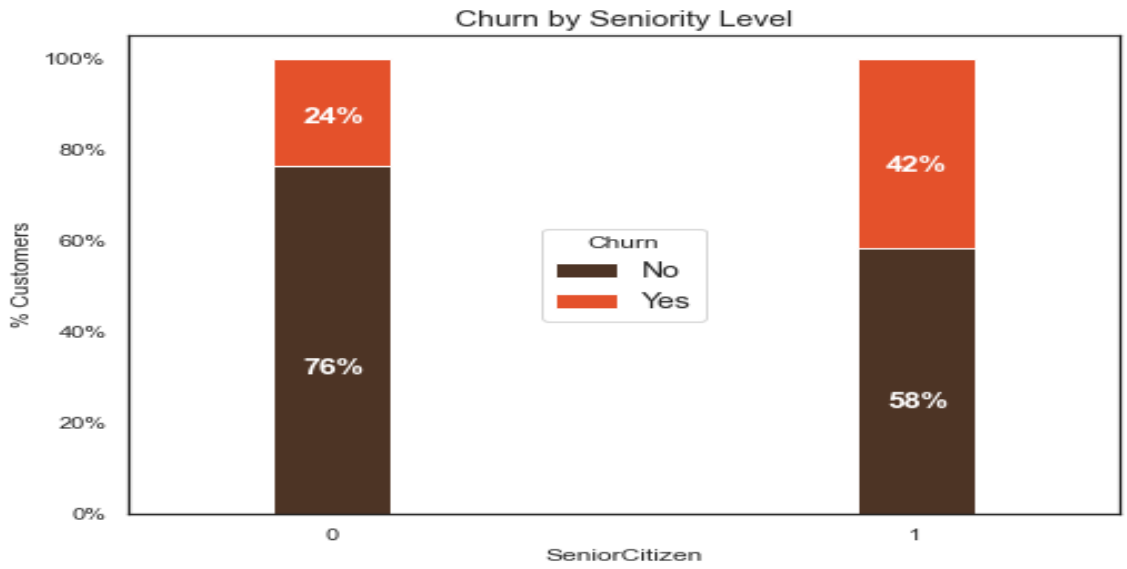
As we can see from the above plot, the customers who do not churn, they tend to stay for a longer tenure with the telecom company.

b. Churn by Contract Type



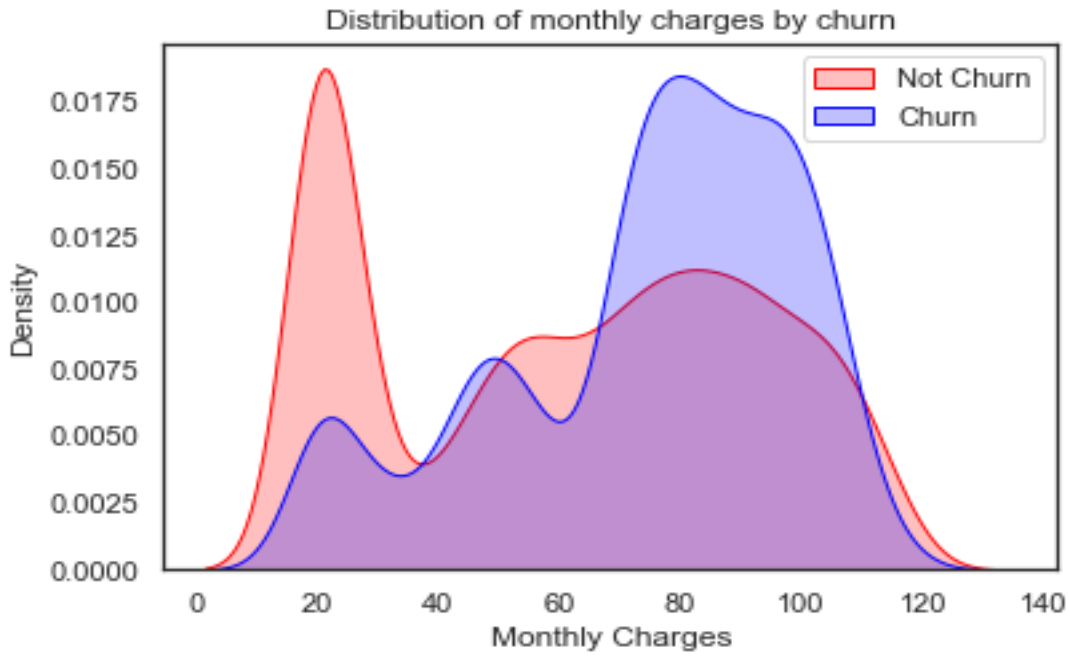
Similar to what we saw in the correlation plot, the customers who have a month-to-month contract have a very high churn rate

c. Churn by Seniority



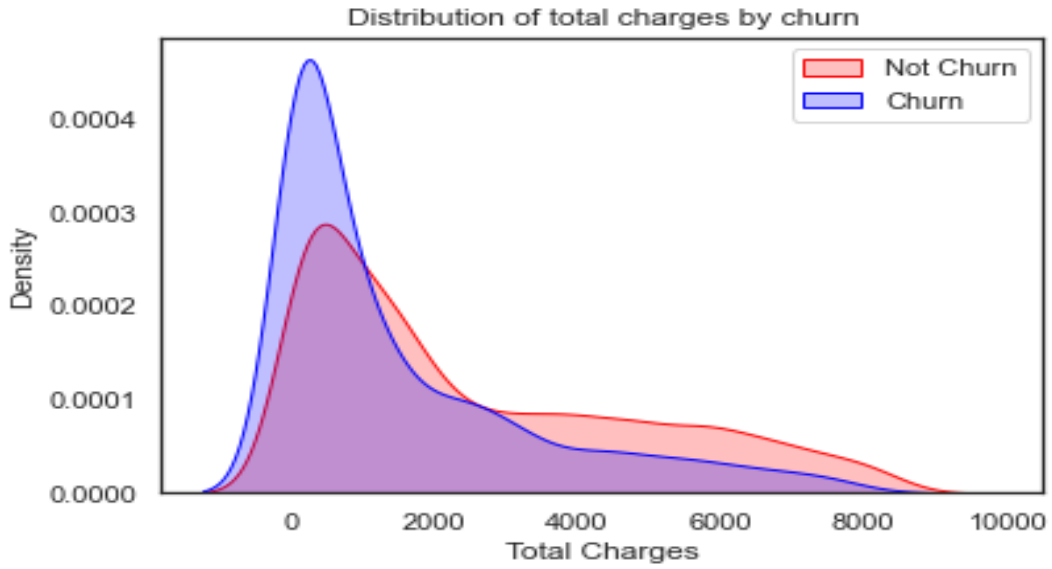
Senior Citizens have almost double the churn rate than younger population

d. Churn by Monthly Charges



Higher percentage of customers churn when the monthly charges are high

e. Churn by Total Charges



It seems that there is higher churn when the total charges are lower

Hypothesis Testing

Null: Customers who have a month-to-month contract, do not have tech support or online security services, and have a high monthly charge are more likely to churn.

Alternative: Customers who have a month-to-month contract, do not have tech support or online security services, and have a high monthly charge are more likely not to churn.

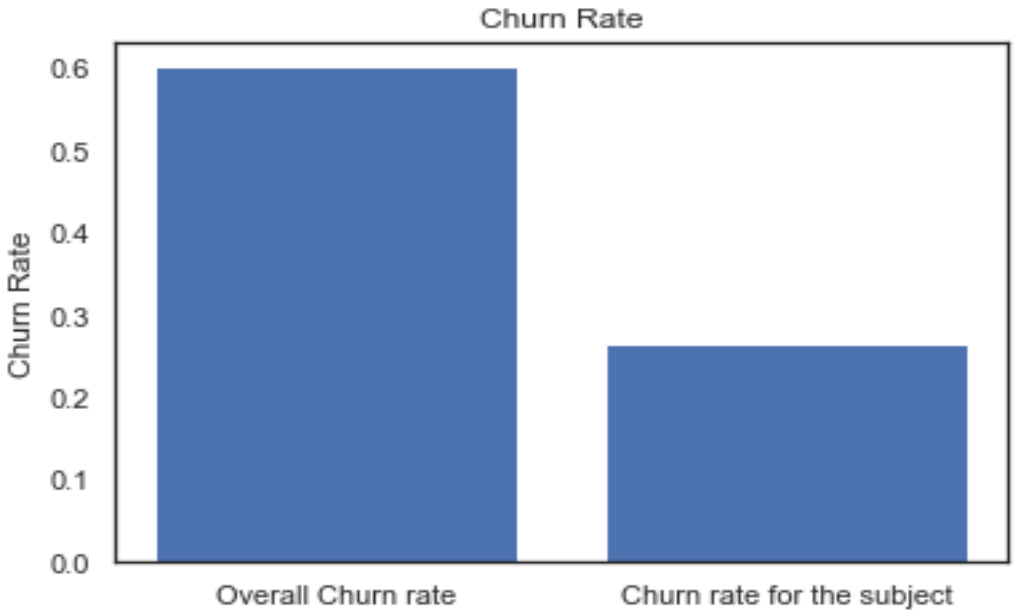
If the churn rate for this subset is significantly higher than the overall churn rate, then it indicates that customers in that subset are more likely to churn.

To test the above hypothesis, we used the following code to create a subset of the data based on the conditions mentioned in the hypothesis and calculate the churn rate for this subset.

```
# Create a subset of the data based on the hypothesis
subset = teleco_churn[(teleco_churn.Contract == "Month-to-month") & (teleco_churn.TechSupport == "No")
                      & (teleco_churn.OnlineSecurity == "No") & (teleco_churn.MonthlyCharges > 70)]

# Calculate the churn rate for the subset
churn_rate = subset.Churn.mean()

print("Churn rate for the subset:", churn_rate)
```



We accept the Null hypothesis since the churn rate for the subset is significantly highly than the overall churn rate

Machine learning Modeling

After performing EDA, the next step was to build a machine learning model to predict customer churn. The following models were used for this project; Logistic Regression, Random Forest, XG-Boost, Ada Boost, Support Vector Machine Classifiers.

The models were trained on the training data and evaluated on the test data using accuracy, Precision, Recall, and F1-score metrics. The Ada Boost model had the highest accuracy. Again, the Ada Boost, Support Vector Machine had highest Recall score.

Presentation and discussion of Results

Metrics / Model	Accuracy	Recall	Precision	F1-Score
Support Vector Machine	0.78	0.78	0.77	0.78
Random Forest	0.75	0.75	0.75	0.67
Logistic Regression	0.79	0.79	0.78	0.78
Ada Boost	0.80	0.80	0.79	0.79
XG-Boost	0.78	0.78	0.77	0.78

Table 1.0: presentation and discussion of results

Based on the Accuracy, Precision, Recall and F1-Score metrics, the best model for the classification problem appears to be Ada Boost. Although all the models had similar accuracy scores, Ada Boost had the highest Precision, Recall and F1-Score. Precision, Recall and F1-Score are important metrics to consider in classification problem because they take into account both false positive and false negative errors which can have different impact depending on specific problem.

While XG-Boost had a good accuracy score of 0.78, its Precision, Recall and F1-Score is slightly lower than those of Ada Boost. Therefore, based on the metrics evaluated Ada Boost appears to be the best model for this classification problem.

However, it is important to note that other factors such as model complexity, computerization, efficiency and interpretability such be also be considered when choosing a model for deployment.

Conclusion

In conclusion, this project aimed to predict customer churn in the telecom industry using a machine learning model. The dataset was cleaned by removing duplicates and imputing missing values. EDA was performed to analyze the distribution of the independent variables and the target variable. Five models were used to predict customer churn, and the Ada Boost Classifier model had the highest accuracy and F1-score.

This project can help telecom companies to predict customer churn and take proactive measures to retain their customers.