Name: Mandeep Singh Gill

Reg No. 12100568

Assignment: Predicting Customer Churn in a Telecommunications company

1. Introduction: Telco Customer Churn Prediction is a critical task for telecommunications companies aiming to retain customers and improve their service offerings. Churn refers to the phenomenon where customers discontinue their services with the company, often switching to competitors. Predicting churn allows telecom companies to take proactive measures to retain customers by identifying those at risk of churning and offering targeted incentives or improvements in service.

2. Dataset Description: The dataset used for Telco Customer Churn Prediction typically contains various features related to customer demographics, services subscribed to, contract details, and usage patterns. These features include customer ID, gender, age, tenure, contract type, internet service type, monthly charges, and churn status (whether the customer churned or not).

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupp
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	

```
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 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)
memory usage: 1.1+ MB
```

```
customerID          object
gender              object
SeniorCitizen        int64
Partner             object
Dependents          object
tenure               int64
PhoneService        object
MultipleLines       object
InternetService     object
OnlineSecurity      object
OnlineBackup        object
DeviceProtection    object
TechSupport         object
StreamingTV         object
StreamingMovies     object
Contract            object
PaperlessBilling    object
PaymentMethod       object
MonthlyCharges     float64
TotalCharges        object
Churn               object
dtype: object
```

	SeniorCitizen	tenure	MonthlyCharges	TotalCharges	Churn	gender_Female	gender_Male	Partner_No	Partner_Yes	Dependents_No	...	StreamingMovies_Ye
0	0	1	29.85	29.85	0	1	0	0	1	1	...	
1	0	34	56.95	1889.50	0	0	1	1	0	1	...	
2	0	2	53.85	108.15	1	0	1	1	0	1	...	
3	0	45	42.30	1840.75	0	0	1	1	0	1	...	
4	0	2	70.70	151.65	1	1	0	1	0	1	...	

5 rows × 46 columns

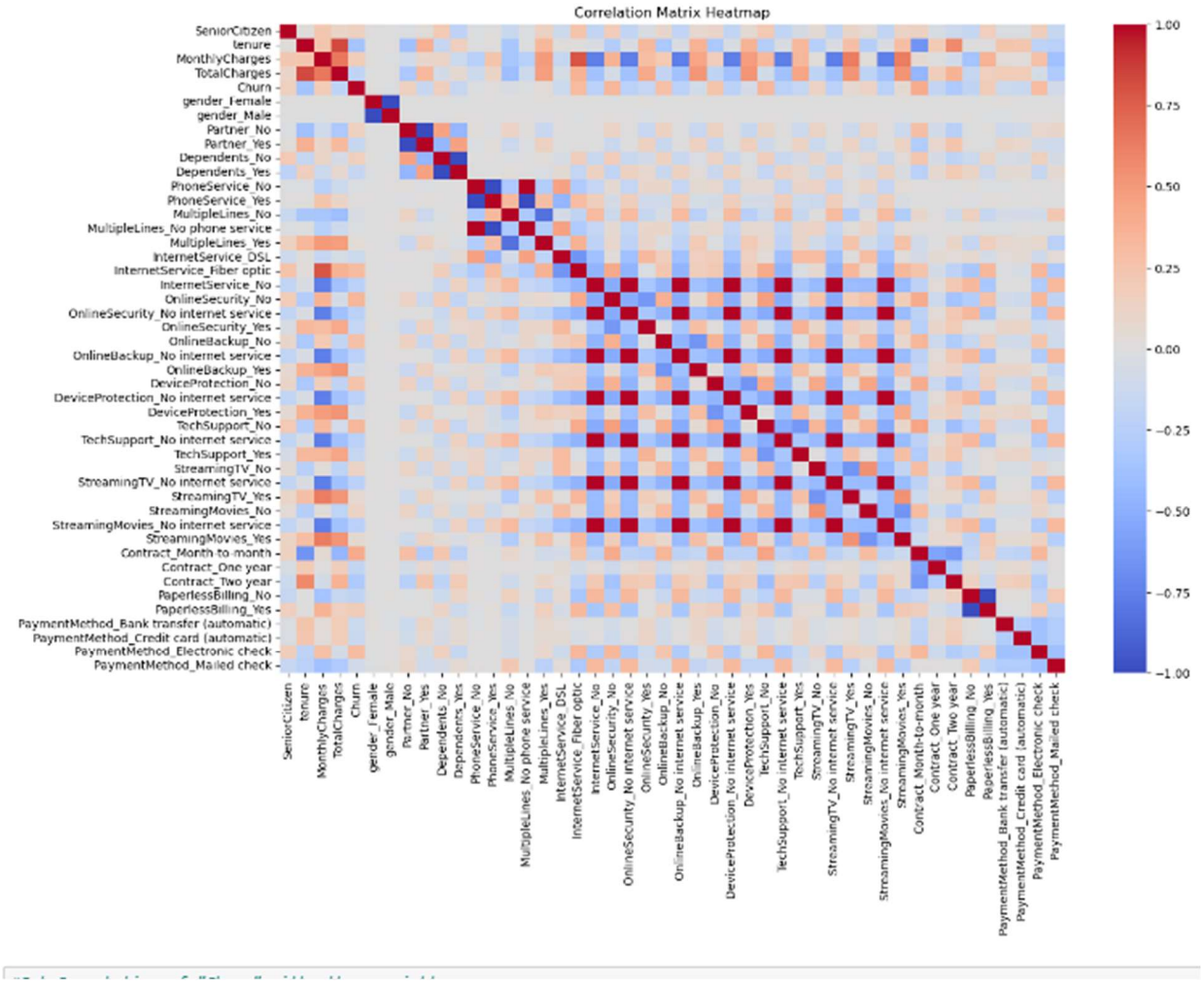
Converting categorical to numerical and creating dummies.

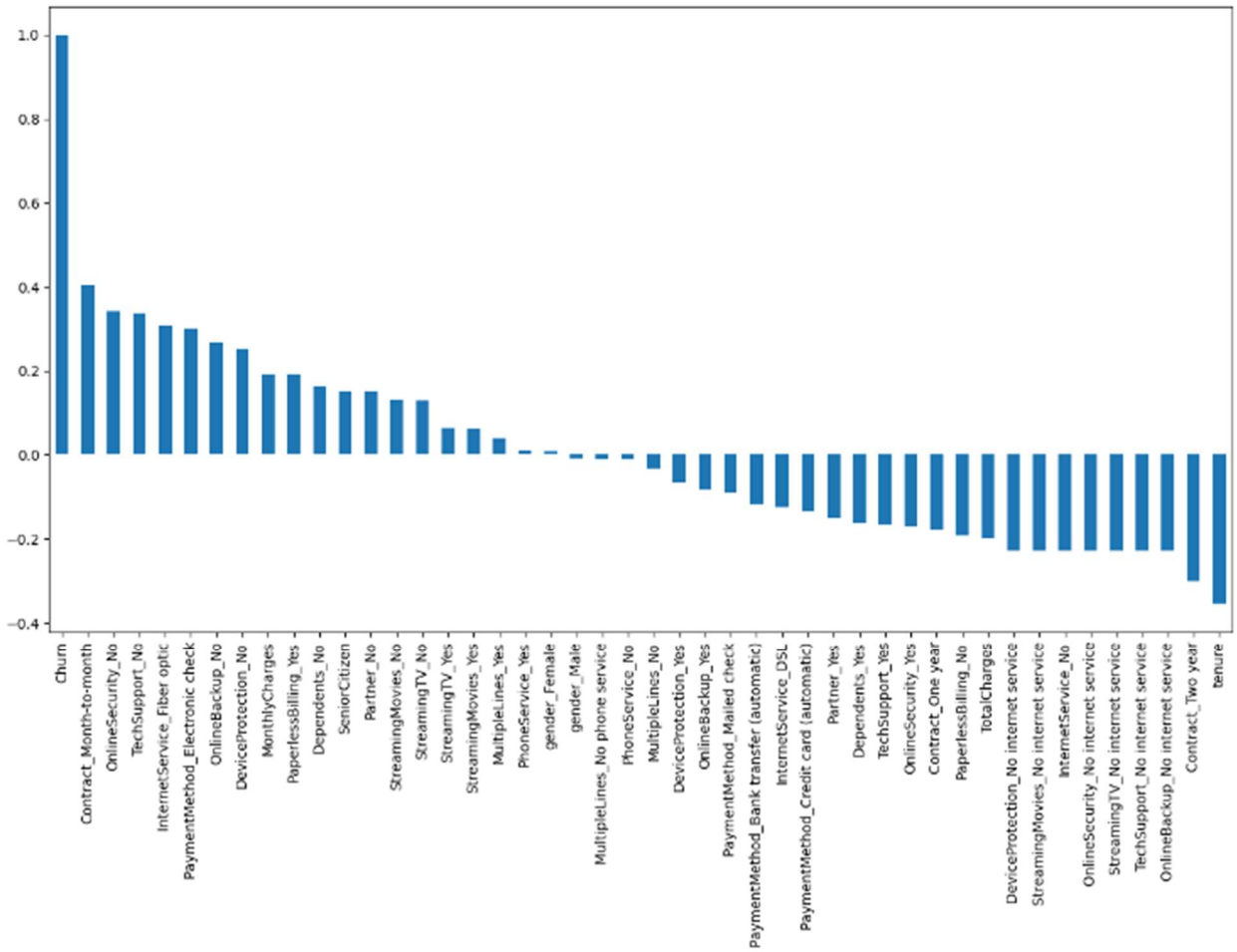
3. Data Preprocessing: Before building predictive models, the dataset undergoes preprocessing steps to handle missing values, encode categorical variables, and scale numerical features if necessary. Missing values may be imputed using mean, median, or mode values. Categorical variables are often encoded using techniques like one-hot encoding or label encoding to convert them into a format

suitable for machine learning algorithms.

```
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges        0
Churn               0
dtype: int64
```

4. Exploratory Data Analysis (EDA): EDA involves analyzing and visualizing the dataset to gain insights into the relationships between different variables and their impact on churn. EDA techniques such as summary statistics, histograms, box plots, correlation analysis, and visualization tools like scatter plots and heatmaps help identify patterns and trends in the data.

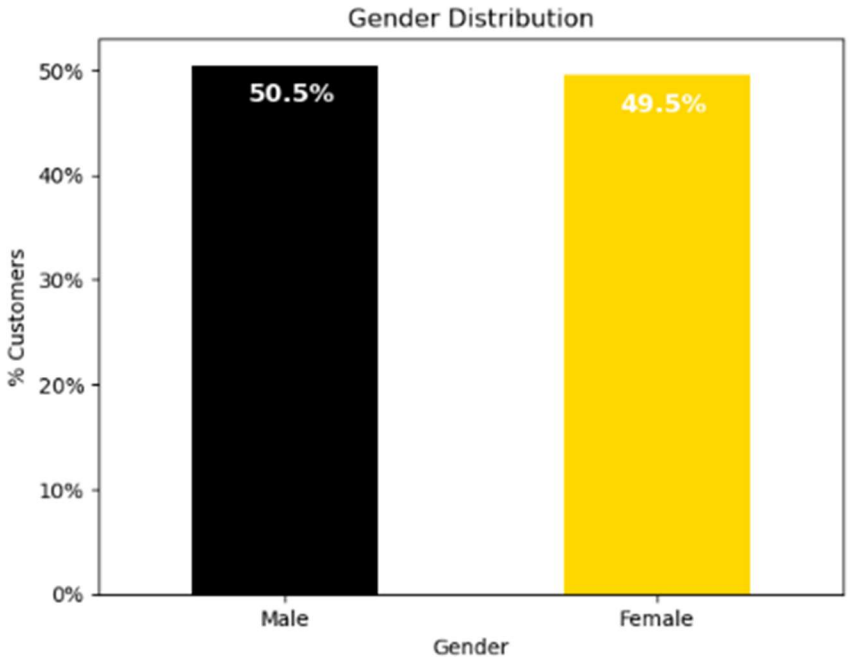




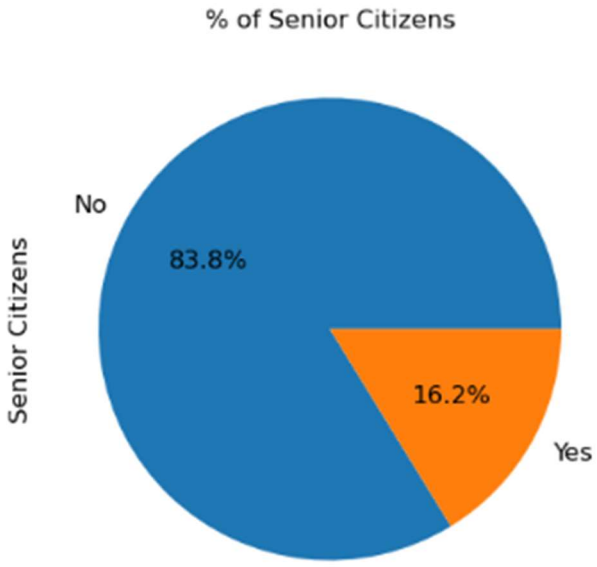
Month-to-month contracts, absence of online security, and lack of tech support appear to be positively correlated with churn. Conversely, longer tenure and two-year contracts show a negative correlation with churn.

Notably, services such as online security, streaming TV, online backup, and tech support, when provided independently of internet connection, exhibit a negative relationship with churn.

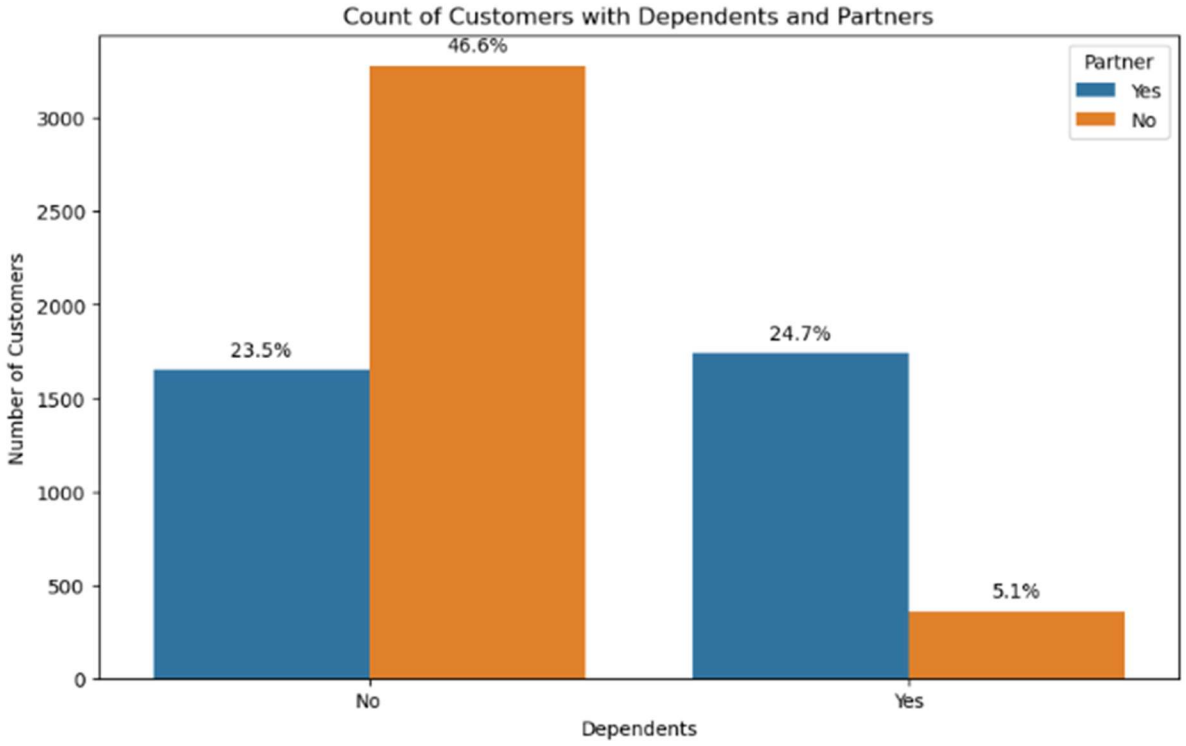
Below, we will investigate these correlations further before proceeding to model building and identifying key variables.



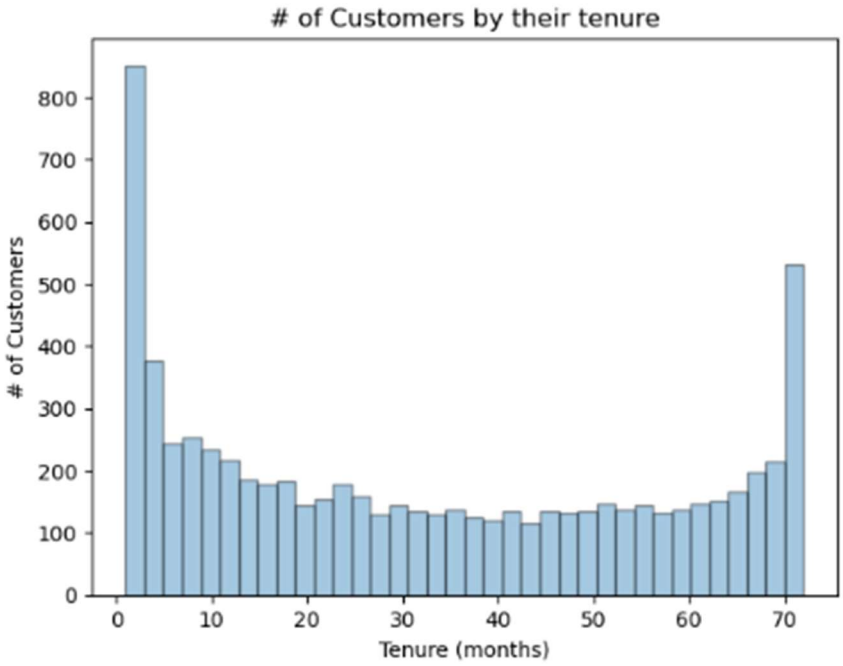
About half of the customers in our data set are male while the other half are female



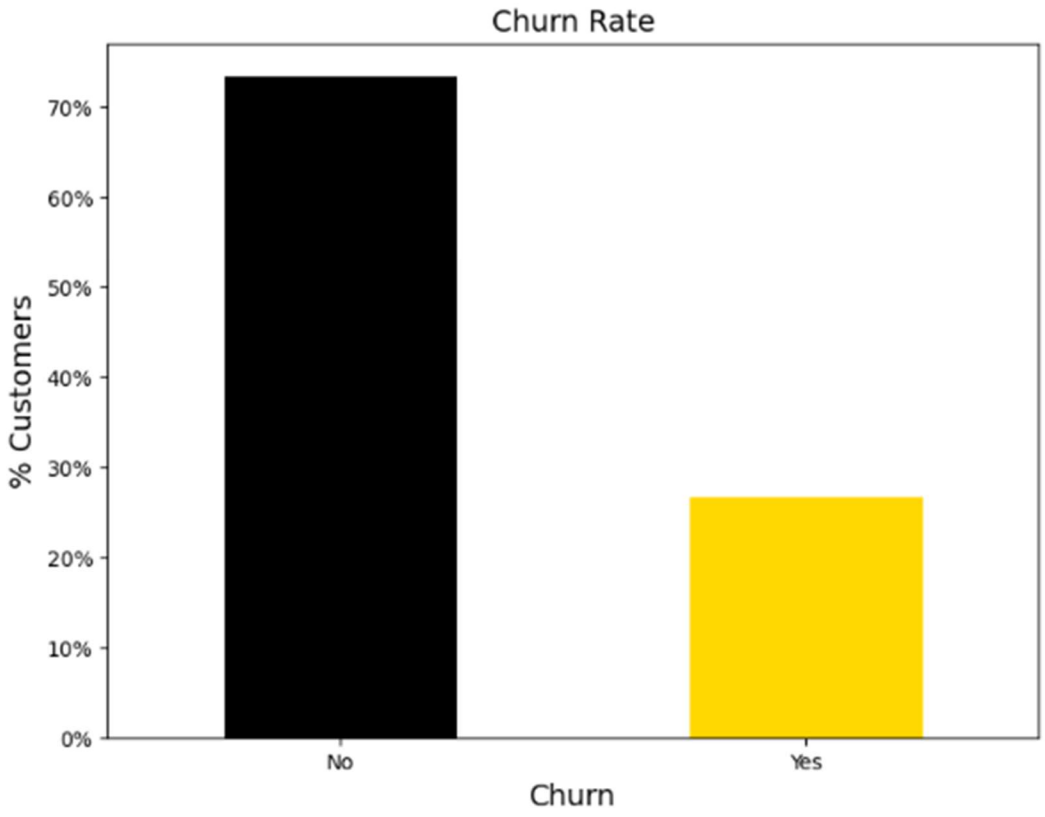
Only 16% of the customers are senior citizens. Thus, most of our customers in the data are younger people.



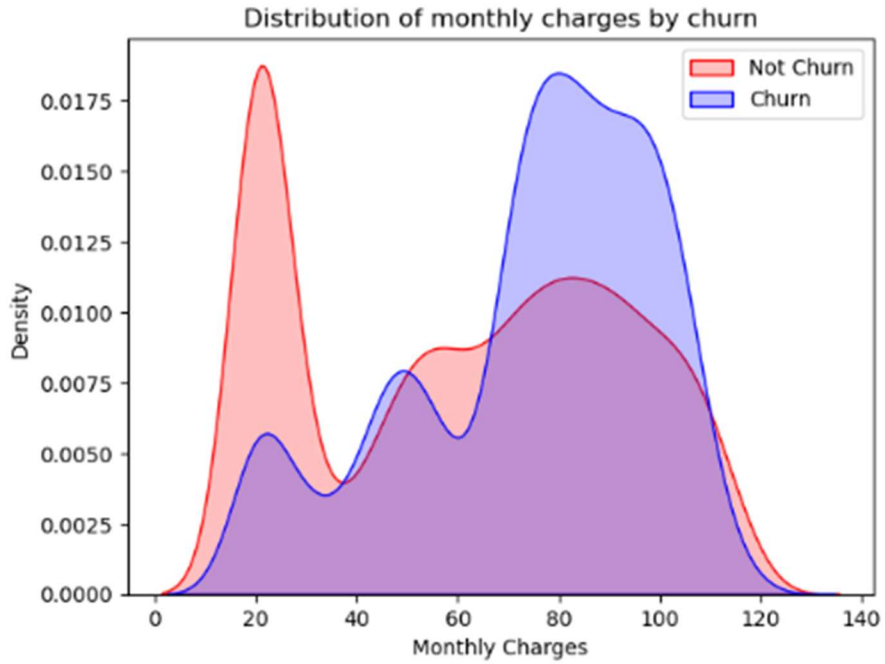
About 50% of the customers have a partner, while only 30% of the total customers have dependents.



After looking at the above 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.



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 skeweness could lead to a lot of false negatives. We will see in the modelling section on how to avoid skewness in the data.



Churn by Monthly Charges: Higher % of customers churn when the monthly charges are high.

5.Model Building: After going through the above EDA we will develop some predictive models and compare them.¶

We will develop Logistic Regression, Random Forest.

A} Logistic regression

```
# We will use the data frame where we had created dummy variables
y = df_dummies['Churn'].values
X = df_dummies.drop(columns = ['Churn'])

# Scaling all the variables to a range of 0 to 1
from sklearn.preprocessing import MinMaxScaler
features = X.columns.values
scaler = MinMaxScaler(feature_range = (0,1))
scaler.fit(X)
X = pd.DataFrame(scaler.transform(X))
X.columns = features
```

```
# Create Train & Test Data
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=101)
# Running logistic regression model
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
result = model.fit(X_train, y_train)
```

```
from sklearn import metrics
prediction_test = model.predict(X_test)
# Print the prediction accuracy
print (metrics.accuracy_score(y_test, prediction_test))
```

```
0.8075829383886256
```

B} Random Forest

```
from sklearn.ensemble import RandomForestClassifier
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=101)
model_rf = RandomForestClassifier(n_estimators=1000 , oob_score = True, n_jobs = -1,
                                  random_state =50, max_features = 100,
                                  max_leaf_nodes = 30)
model_rf.fit(X_train, y_train)

# Make predictions
prediction_test = model_rf.predict(X_test)
print (metrics.accuracy_score(y_test, prediction_test))
```

```
0.8081023454157783
```

C} SVC

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=99)
from sklearn.svm import SVC

model.svm = SVC(kernel='linear')
model.svm.fit(X_train,y_train)
preds = model.svm.predict(X_test)
metrics.accuracy_score(y_test, preds)
```

```
0.820184790334044
```

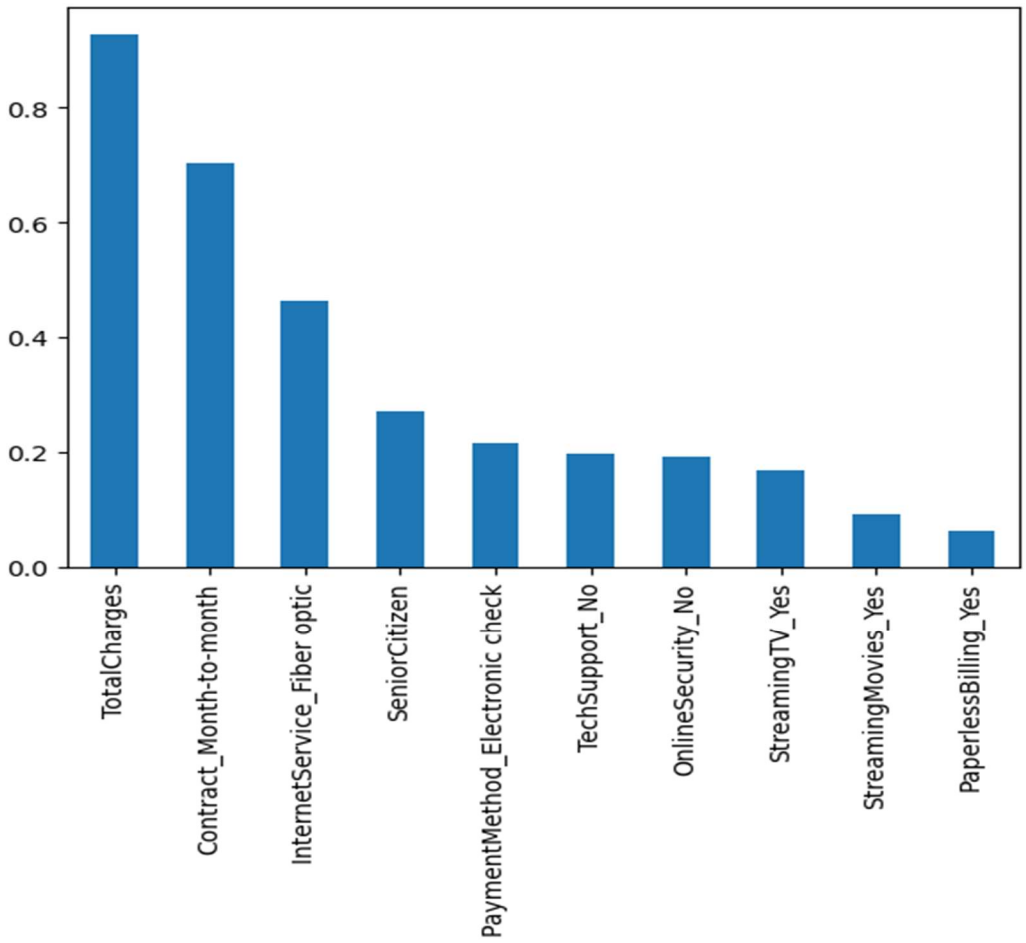
6. Insights:

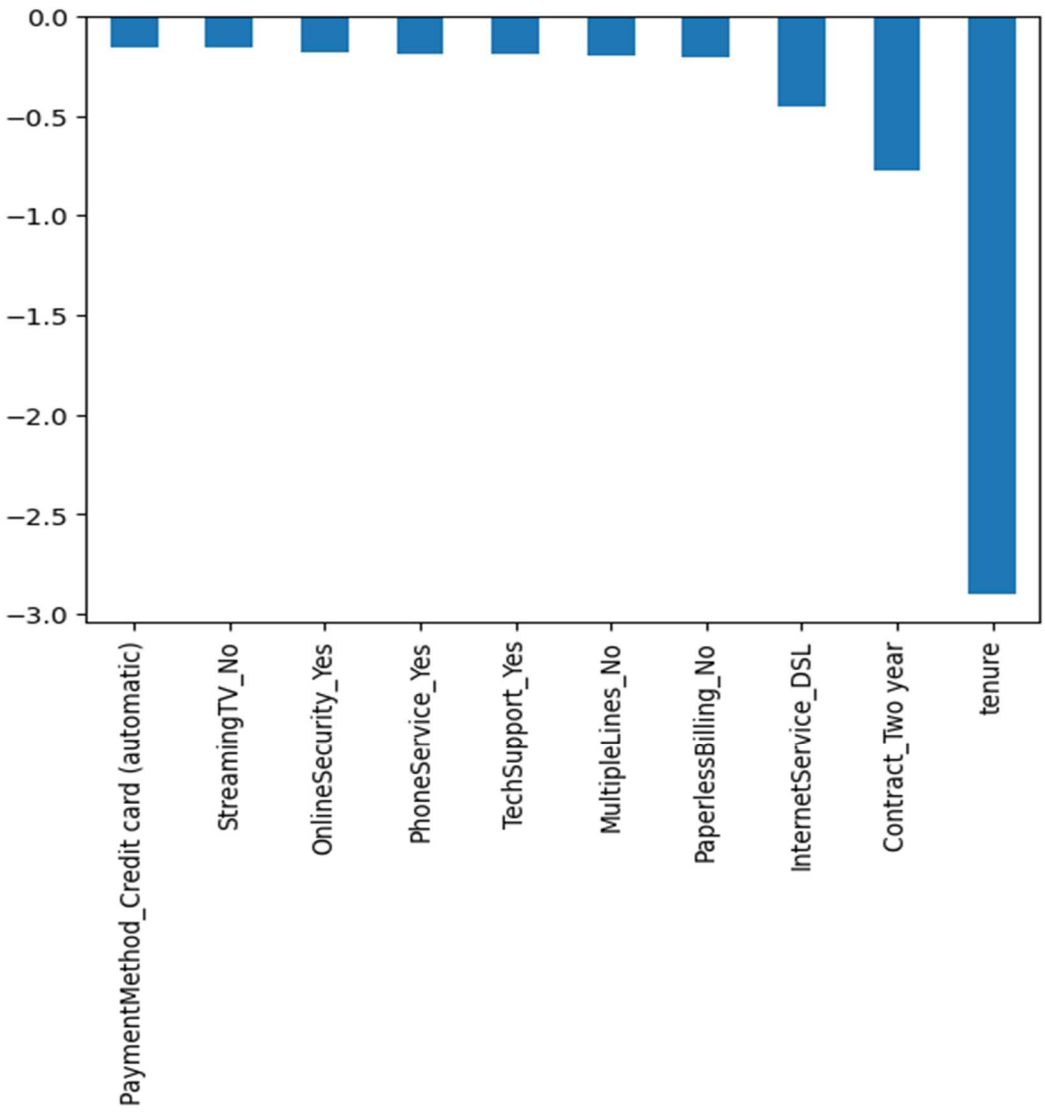
we uncovered several noteworthy features that exhibit significant relationships with our predicted variable:

1. **Contract Length and Tenure:** Our analysis revealed that customers with two-month contracts exhibit a reduced likelihood of churning. Furthermore, tenure, particularly in conjunction with two-month contracts, demonstrates the most pronounced negative correlation with churn, as indicated by logistic regression predictions.
2. **DSL Internet Service:** Customers utilizing DSL internet service also display a diminished probability of churning. This suggests that DSL subscribers are more likely to remain with the telecom provider compared to those with other internet service types.
3. **Total Charges and Monthly Contracts:** Interestingly, higher total charges, monthly contracts, and the availability of fiber optic internet services are associated with elevated churn rates. Despite the superior speed of fiber optic

services, customers may be more inclined to churn when opting for this option. This paradox warrants further investigation to uncover the underlying reasons behind this trend.

In conclusion, our analysis highlights the importance of contract length, internet service type, and total charges in influencing customer churn behavior. While certain factors mitigate churn risk, others appear to exacerbate it, underscoring the need for deeper exploration and understanding of customer preferences and behaviors in the telecom industry.

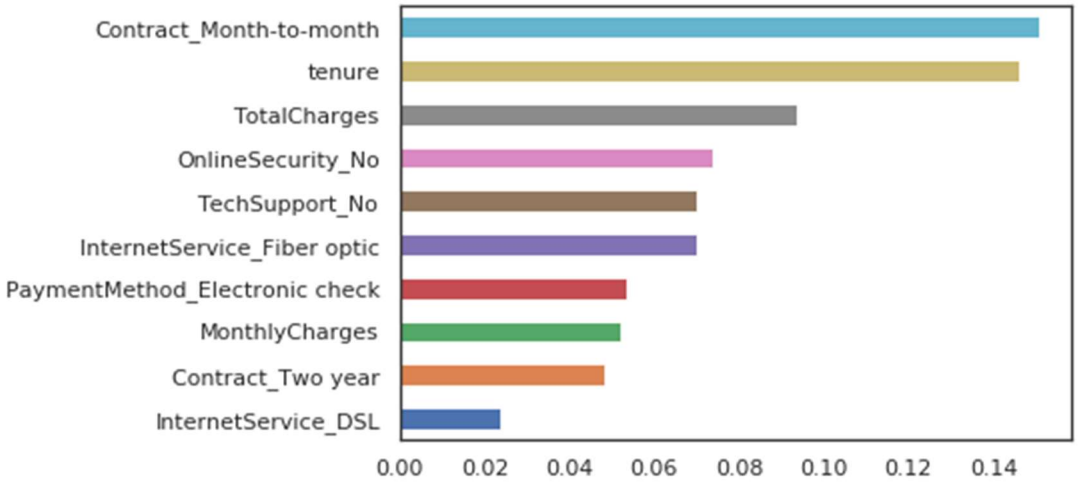




For Random Forest,

- monthly contract, tenure and total charges are the most important predictor variables to predict churn.

- The results from random forest are very similar to that of the logistic regression and in line to what we had expected from our EDA



7. Evaluation: