Customer Churn Model

-Bhargavi Jadhav

Problem Statement and Objective:

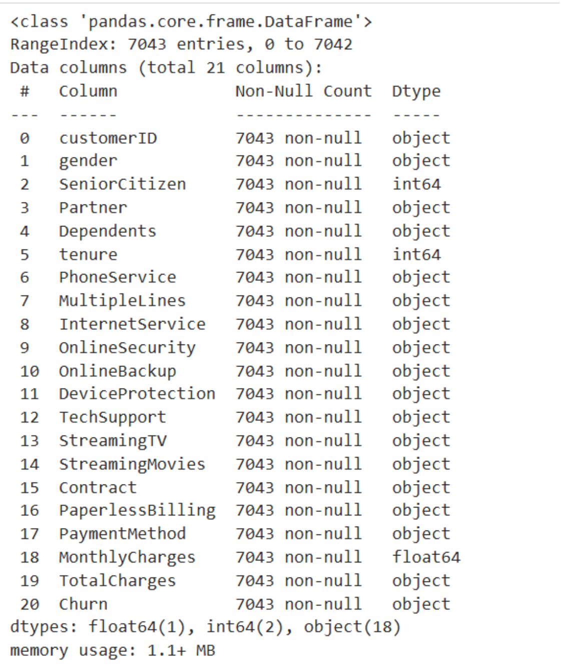
Customer churn is an important factor that affects the revenue, market share, and operational cost of a company. This makes it important for a company to understand patterns in customer churn in order to work on customer retention policies. This project aims to develop a predictive model to identify customers who are likely to churn, enabling the company to proactively engage and retain the customers who are likely to churn through possible interventions.

Scope of the project:

- Understand the key drivers of churn.
- Build an ML model using historical data to predict the likelihood of churn.
- Provide actionable insights to reduce churn rate

Approach:

1. Data loading and basic understanding:
   Understanding the information presented in the file, the columns, the size of the data, the features, and the data types.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
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
```

Insights:

- The missing values in the TotalCharges column are indicated by a space
- Total charges should be a float
- Customer ID can be dropped
- Senior citizen is a categorical column

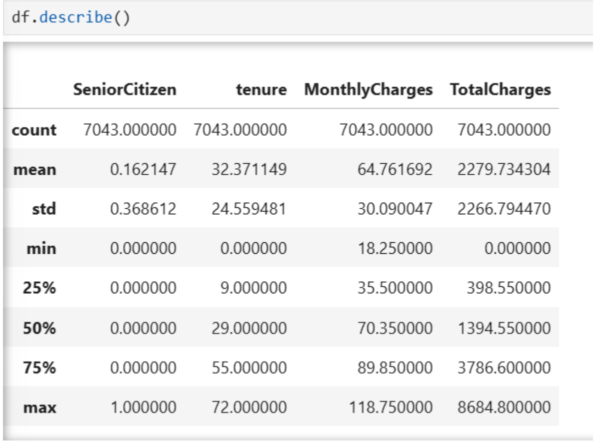
2. Data Preprocessing
   - Replacing the missing values in the TotalCharges column with 0.0
   - Changing the datatype of Total Charges to float
   - Dropping the customerID column
   - Dropping the churn column from the object columns and adding the senior citizens column

3. EDA (Exploratory Data Analysis):
   a) Understanding the distribution of the target variable

```
Churn
No      5174
Yes     1869
Name: count, dtype: int64
```

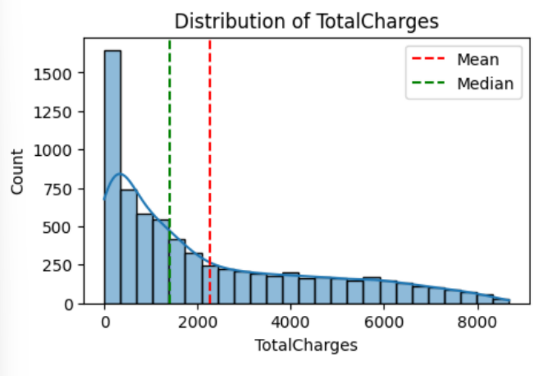
As one category dominates over the other, there is an imbalance in the dataset

b) Understanding the distribution of numerical categories, further through histograms as well.

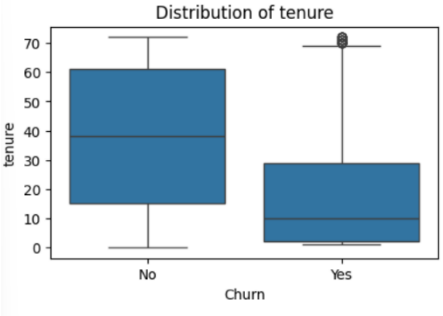


```
df.describe()
```

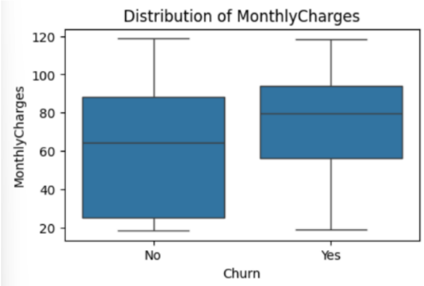
	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
count	7043.000000	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692	2279.734304
std	0.368612	24.559481	30.090047	2266.794470
min	0.000000	0.000000	18.250000	0.000000
25%	0.000000	9.000000	35.500000	398.550000
50%	0.000000	29.000000	70.350000	1394.550000
75%	0.000000	55.000000	89.850000	3786.600000
max	1.000000	72.000000	118.750000	8684.800000



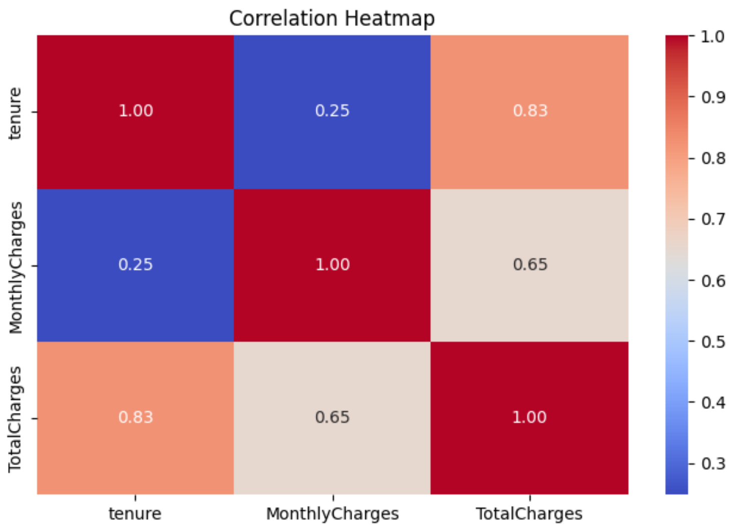
c) Understanding how the distribution of numerical parameters might affect Churn.



Customers with small tenures are more likely to churn than customers with long tenures. This may be because of short-term engagement.



Customers with higher monthly charges are more likely to churn than customers with lower monthly charges.

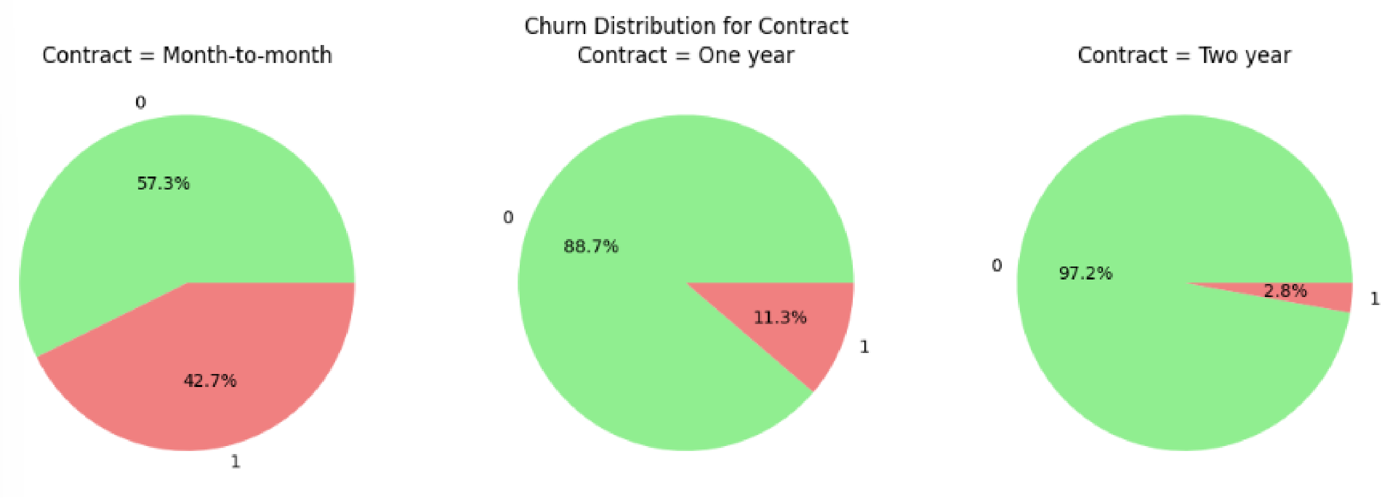


While tenure and monthly charges both affect the total charges, tenure has a higher correlation with the total charges paid by the customer as compared to the monthly charges.

d) Understanding the categorial features:

i) Count plots of all the categories are plotted to understand the distribution of each column.

ii) Understanding the churn distribution of each column with respect to all the categories in it. This helps to identify the key categorical drivers for churn.



The following are the insights from the same:

- **Contract: The customers with a month-to-month contract tend to churn more**, possibly because of short-term plans for customers who do not look forward to long-term engagement.
- **Payment mode: Customers with electronic check as a mode of payment churn more** probably because of the inconvenience of payment over automatic modes
- **Tech Support: Customers with no tech support churn more because** they might feel unsupported when an issue is raised
- **Customers with no device protection, online security, or backup online churn out more**
- **Fiber optic users churn out more,** which might be due to pricing sensitivity or higher expectations
- **Customers with no dependents churn out more as compared to those with dependents,** churning affects fewer people (individual plans vs family plans)

Interventions must be done to retain customers based on these parameters.

Business Recommendations:

- Incentivize long-term contracts with special discounts
- Simplify the process of automatic payments. They can even be incentivized through offers or discounts
- Improvement in tech support to make it more accessible. This can be done using various channels rather than relying on limited channels (example: chat, app, email, call)
- Provide incentives for the purchase of device protection and backups
- Improve fiber optic engagement through loyalty rewards or discounts
- Promote the purchase of family plans or plans in groups to increase dependents and reduce churn in individuals

e) Encoding the categorial columns using LabelEncoder and saving them using pickle files

4. Train-test data split:
Data is split into train data and test data, with test size=0.2

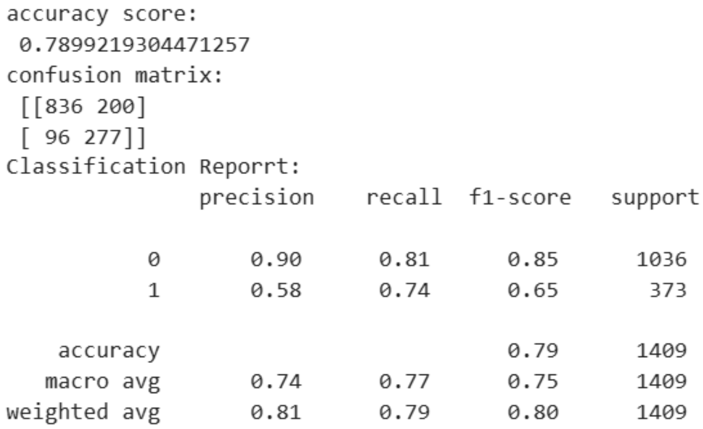
5. Model Training:
Multiple classification models were evaluated using cross-validation with default parameters. Based on their performance, a **Random Forest Classifier** with class_weight='balanced' was used to address class imbalance in the churn labels, ensuring the model gives appropriate importance to the minority class (churned customers) and avoids biased predictions skewed toward the majority class. **Hyperparameter tuning** was then performed using

Grid Search Cross-Validation to identify the optimal combination of parameters and improve the model's predictive performance.

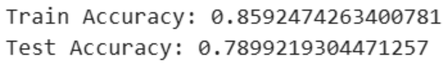
6. Predictions:
   The model was saved and later used for predictions using pickle files

Model Evaluation:

Accuracy score, Confusion matrix, and Classification Report:

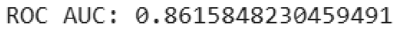


```
accuracy score:
 0.7899219304471257
confusion matrix:
 [[836 200]
 [ 96 277]]
Classification Reporrt:
               precision    recall  f1-score   support

           0       0.90      0.81      0.85      1036
           1       0.58      0.74      0.65       373

    accuracy                           0.79      1409
   macro avg       0.74      0.77      0.75      1409
weighted avg       0.81      0.79      0.80      1409
```



```
Train Accuracy: 0.8592474263400781
Test Accuracy: 0.7899219304471257
```

ROC AUC:



```
ROC AUC: 0.8615848230459491
```

The model accurately predicted the class for approximately 79% of the samples in the test set.

True Negatives (TN): 836 instances were correctly predicted as class 0.

False Positives (FP): 200 instances were incorrectly predicted as class 1 (they were actually class 0). This is a Type I error.

False Negatives (FN): 96 instances were incorrectly predicted as class 0 (they were actually class 1). This is a Type II error.

True Positives (TP): 277 instances were correctly predicted as class 1.

Class 0:

Precision: 0.90 - When the model predicts class 0, it is correct 90% of the time.

Recall: 0.81 - The model identifies 81% of all actual class 0 instances.

F1-Score: 0.85 - A good balance between precision and recall for class 0.

Support: 1036 - There are 1036 instances of class 0 in the test set.

Class 1 (Minority class):

Precision: 0.58 - When the model predicts class 1, it is correct only 58% of the time. This is relatively low, indicating a significant number of false positives for this class.

Recall: 0.74 - The model identifies 74% of all actual class 1 instances. This is reasonably good, meaning it's not missing too many actual positives.

F1-Score: 0.65 - This F1-score is lower than that of class 0, indicating weaker performance in balancing precision and recall for the minority class.

Support: 373 - There are 373 instances of class 1 in the test set.

Macro Avg: This calculates the unweighted mean of the metrics for each class.

- Precision: 0.74
- Recall: 0.77
- F1-Score: 0.75

Weighted Avg: This calculates the mean of the metrics for each class, weighted by their support (number of true instances).

- Precision: 0.81
- Recall: 0.79
- F1-Score: 0.80

The **ROC AUC** value of approximately 0.86 suggests that the model has a good ability to distinguish between the positive and negative classes.

The model performs well overall, with an accuracy of **79%** and a good **ROC AUC of 0.86**. However, there's a noticeable class imbalance (1036 class 0 vs. 373 class 1), and the model's performance on the minority class (Class 1) is weaker compared to the majority class (Class 0).

Outcomes:

A data-driven churn prediction system that identifies high-risk customers with strong accuracy. This enables the telecom provider to take **proactive retention actions**, reducing revenue loss and improving customer lifetime value.