CHURN ANALYSIS

What does churn stand for?

Churn in a business setting refers to losing an acquired, potentially profitable customer. The definition of churn can vary by industry (in healthcare, dead people are considered churn while in finance, people with inactive cards are called churned).

Why do businesses want to prevent churn?

Acquiring a new customer is always more expensive than retaining an existing one. Hence, not letting them churn is the key to a sustained revenue stream.

What metrics do we optimize on while predicting churn?

F1-score and Recall are good ones, but you can also look at PR curves

ABOUT THE DATASET:

Kaggle telco churn dataset is a sample dataset from IBM, containing 21 attributes of approximately 7,043 telecommunication customers. In this Assessment, you are required to work with a modified version of this dataset (the dataset can be found at the URL provided below). Modify the dataset by removing the following attributes: 'MonthlyCharges', 'OnlineSecurity', 'StreamingTV', 'InternetService' and 'Partner'.

The dataset can be found at: https://www.kaggle.com/blastchar/telco-customer-churn

The data set includes information about:

- Customers who left within the last month – the column is called Churn.
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges.
- Demographic info about customers – gender, age range, and if they have partners and dependents.

PROBLEM STATEMENT:

I will try to explore the data and try to answer some questions (some questions may or may not be answered due to some constrants) like:

- What's the % of Churn Customers and customers that keep in with the active services?
- Understanding the gender and age range columns.
- Understanding the tenure of customers based on their contract type.

- Take a look at predictor variable (Churn) and understand its interaction with other important variables.
- Churn vs Tenure.
- Are there any patterns in Churn Customers based on ‘Contract Type’?
- Are there any patterns in Churn Customers based on ‘Seniority’?
- Are there any patterns in Churn Customers based on ‘Total Charge’?
- What's the most profitable service types?
- Which features and services are most profitable?

EXPLORATORY DATA ANALYSIS:

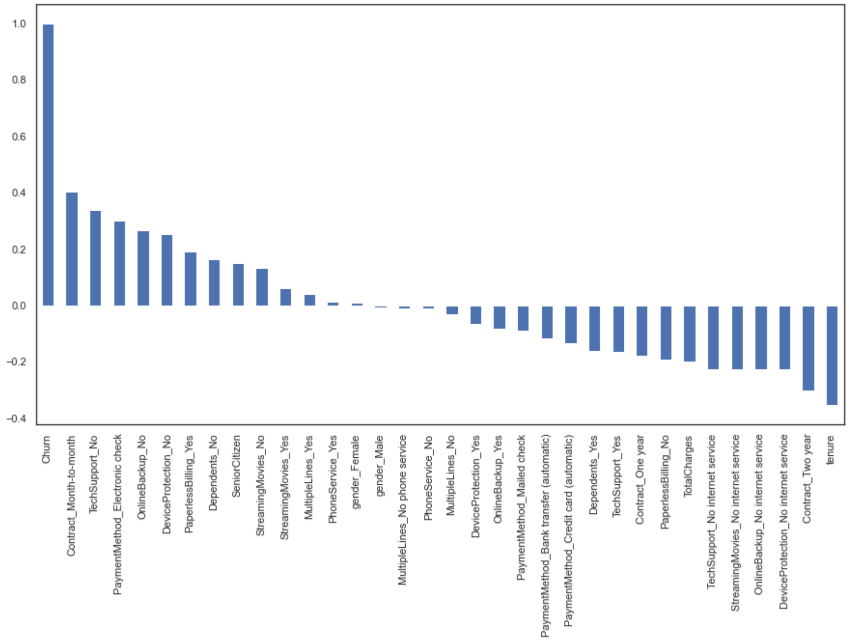

Fig 1. Correlation of "Churn" with other variables.

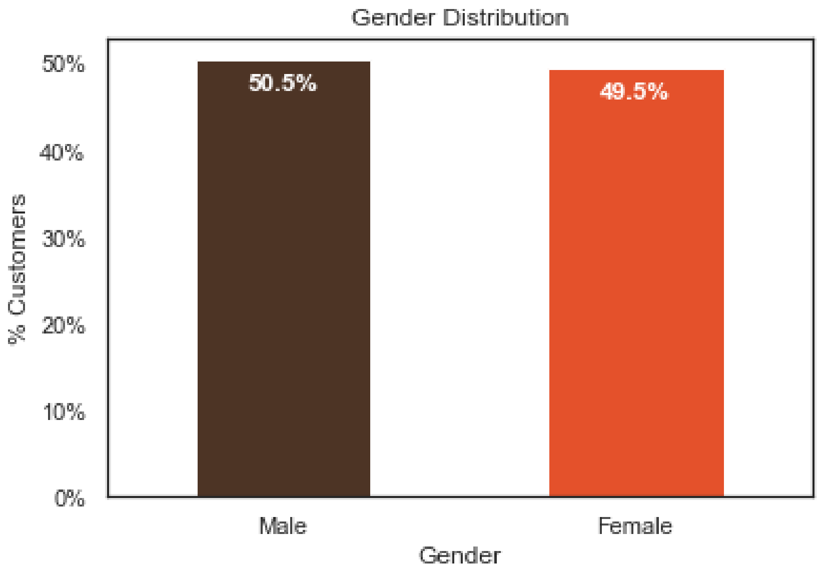

Fig 2. Gender distribution.

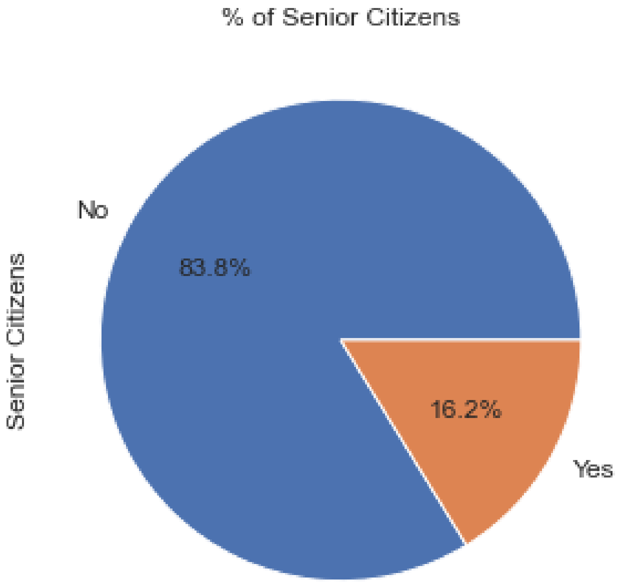

Fig 3. Percentage of Senior Citizens.

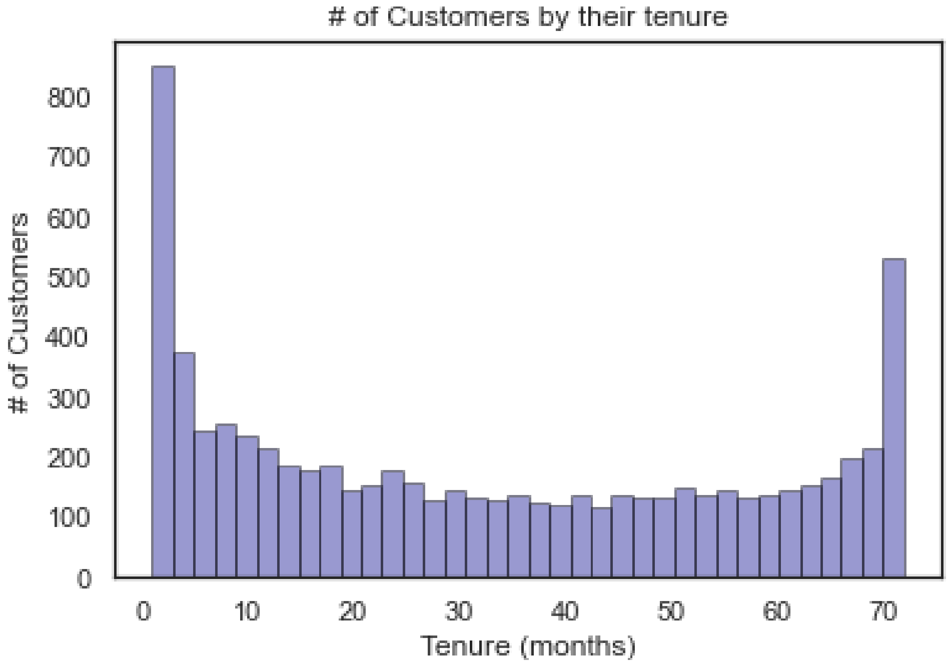

Fig 4. Number of customers by their tenure.

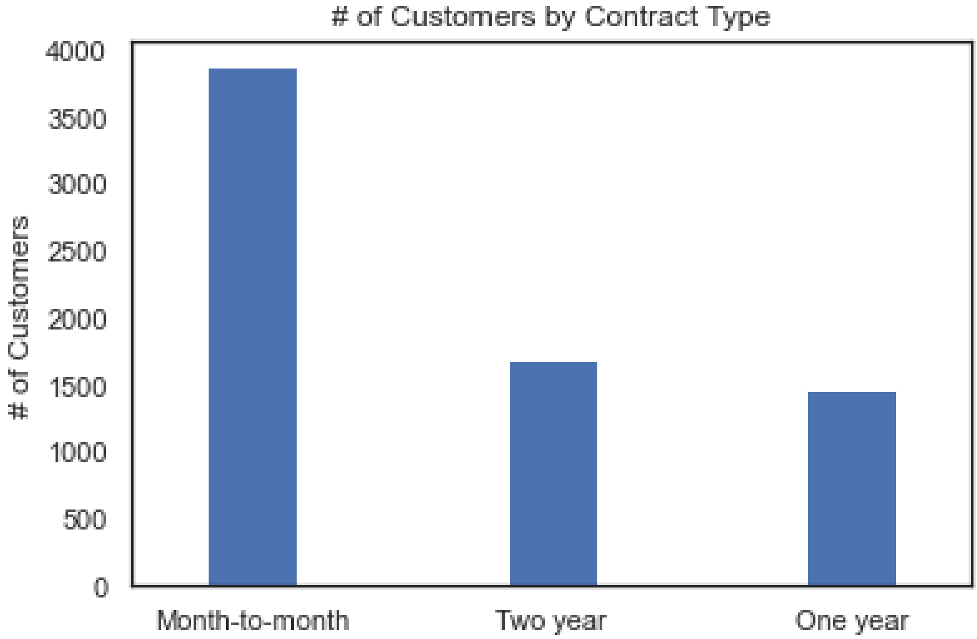

Fig 5. Number of customers by Contract Type.

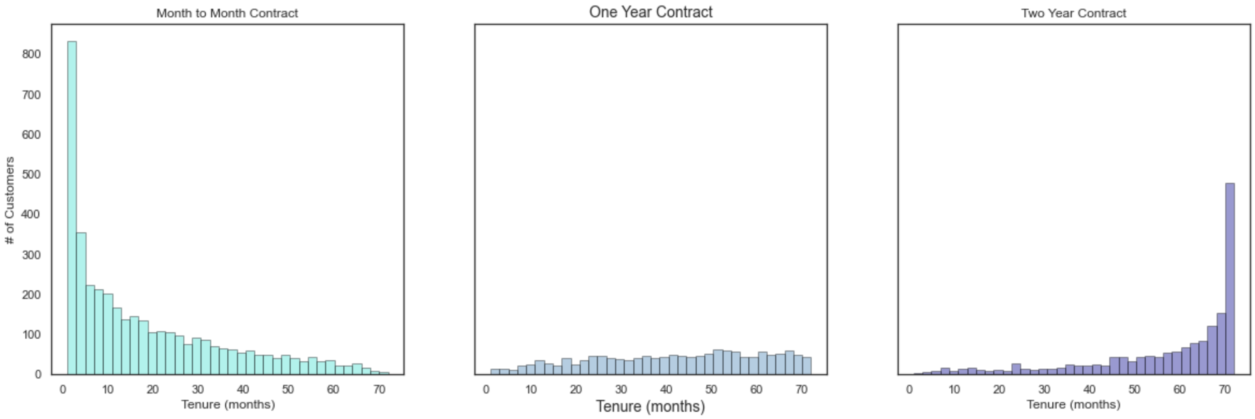

Fig 6. Tenure of customers based on their contract type.

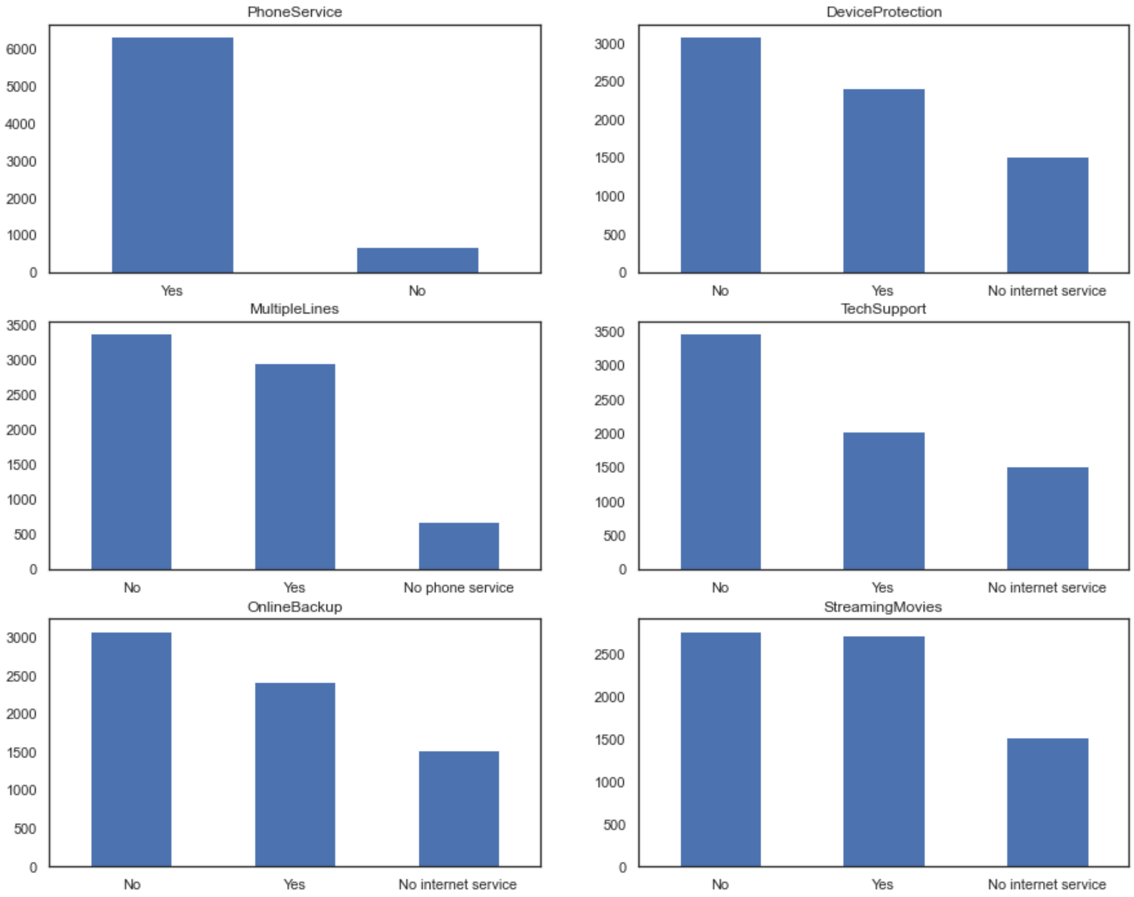

Fig 7.

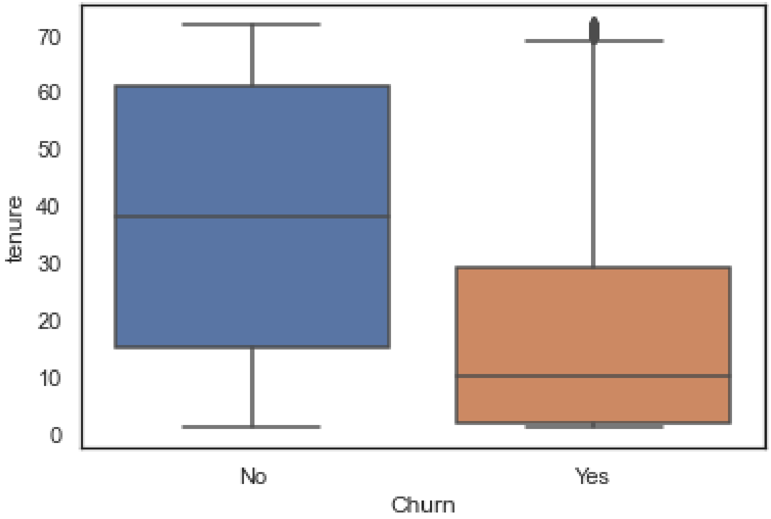

Fig 8. Churn vs Tenure.

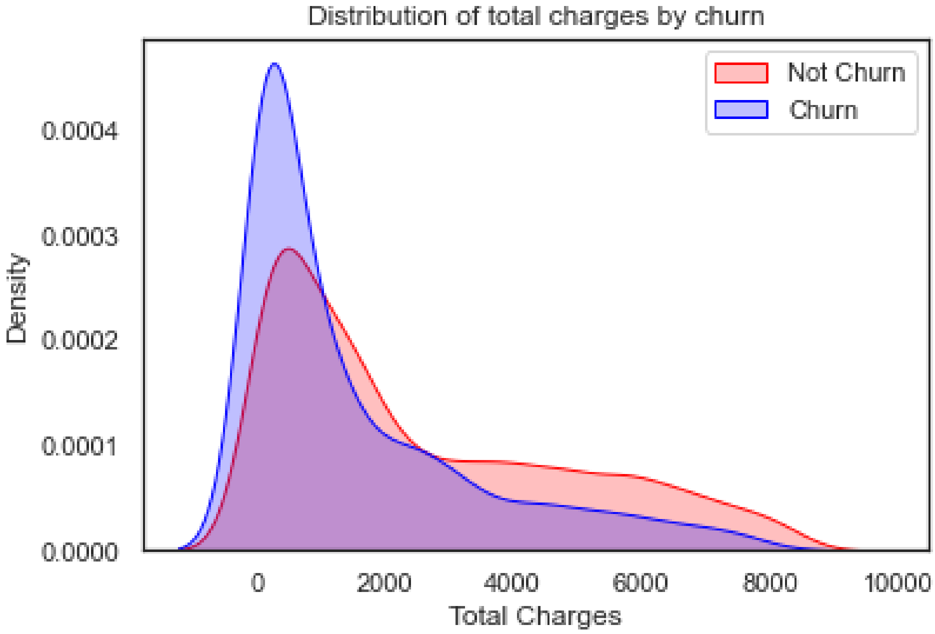

Fig 9. Churn by Total Charges

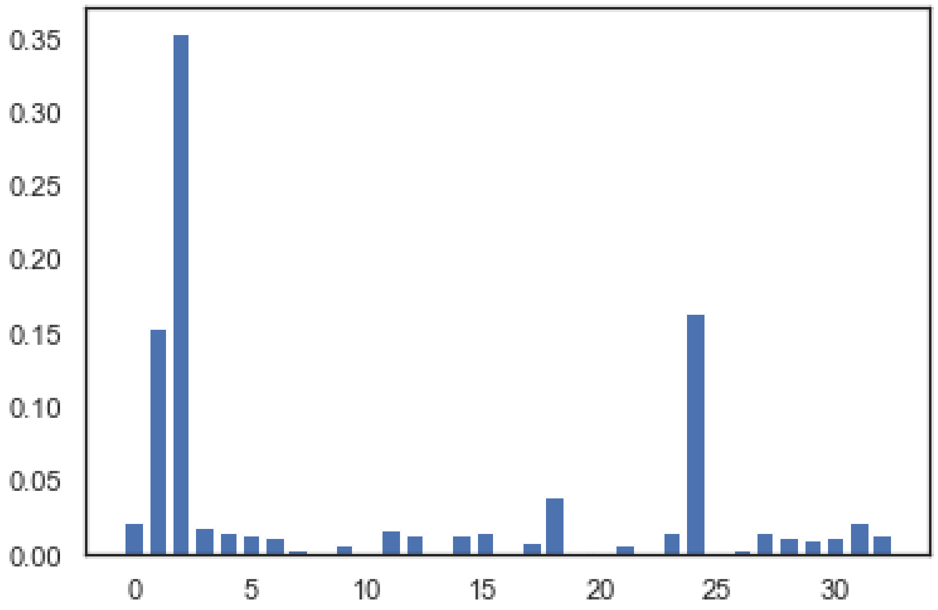

Fig 10. Feature Importance

About the EDA:

Fig 1: It shows the correlation between independent variables/features and dependent variables/features.

Fig 2: It shows the gender distribution. About half of the customers in our data set are male while the other half are female, that means the gender variable/feature is balanced and we do not need to treat for the unbalanced data.

Fig 3: It shows the Percentage of Senior Citizens. We can use this to observe this column/variable (independent variable).

Fig 4: It shows the Number of customers by their tenure. After looking at the 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.

Fig 5: It shows the Number of customers by Contract Type. As we can see from this graph most of the customers are in the month-to-month contract. While there are equal number of customers in the 1 year and 2-year contracts.

Fig 8: Similar to what we saw in the correlation plot, the customers who have a month-to-month contract have a very high churn rate.

Fig 9: It seems that there is higher churn when the total charges are lower.

Fig 10: It shows the importance of various independent features with the dependent feature (i.e., churn).

INTERPRETATION OF CHURN ANALYSIS:

Question. What was the percentage of time at which your analysis was able to correctly identify the churn? Can this be considered a satisfactory outcome? Explain why or why not.

Answer. The percentage of time at which our analysis was able to correctly identify the churn was 99.73%. Yes, this is a very satisfactory outcome because we can say that out of 100 times, our model/analysis was able to predict correct results 99 times.

Question. Describe the attributes of the customers who are churning and explain what is driving the churn.

Answer. Attributes of the customers who are churning can be determined using the 'feature_importances_' method used after training the decision tree. This method will give us the importance of each attribute and we can then conclude which are making the customers to churn. The attributes are: 'TotalCharges', 'tenure', and 'Contract_Month-to-month'.

Question. Describe the effects that your previous steps, model development and handling of missing values had on the outcome of your churn analysis and how the accuracy of your churn analysis could be improved.

Answer. Since, there was only one column/feature that had some missing values. That column was 'TotalCharges'. We got to know that this column was very important by using 'feature_importances_' method. We treated the missing values by dropping them because there were negligible number of NaNs and it was not affecting the result after dropping the missing values.

Things we can do to improve our model's accuracy or prediction:

- Implement StandardScaler in order to scale various attributes/features which will enhance our model's accuracy.
- We can optimize the hyperparameter of the decision tree to get better accuracy/results.
- We can use XGBoost Classifier or LightGBM Classifier, which may give better accuracy/results.