REPORT on CUSTOMER CHURN PREDICTION

Laiba Tanveer - Data Science Trainee

INTRODUCTION

In today's highly competitive business landscape, customer retention has become a critical factor for the success and growth of companies across various industries. Customer churn, or the rate at which customers discontinue their relationship with a company, poses a significant challenge for businesses. This report focuses on customer churn prediction, aiming to report a model that accurately identifies customers who are likely to churn.

This report aims to provide a comprehensive summary of a project, covering key findings, exploratory data analysis (EDA) insights, and an evaluation of the model's performance. The project focused on getting insights on how to retain customers and getting insights on what are the factors influencing churning.

This report provides a concise overview of the steps involved in conducting churn prediction analysis.

Steps:

- Pre-Processing
- Exploratory Data Analysis (EDA)
- Model

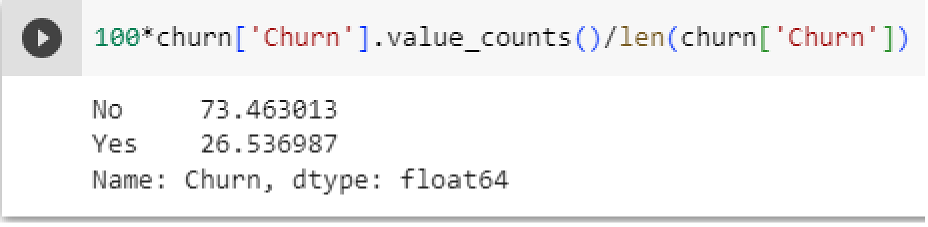
Each step is essential in accurately predicting customer churn and facilitating informed decision-making. The report delves into the methodologies employed, key findings from the EDA, and performance evaluation of the predictive model. By following this systematic approach, businesses can gain valuable insights into customer churn dynamics, leading to enhanced customer retention strategies.

EDA Insights

EDA played a crucial role in understanding the project's dataset and identifying patterns, trends, and relationships. The EDA phase provided valuable insights, including:

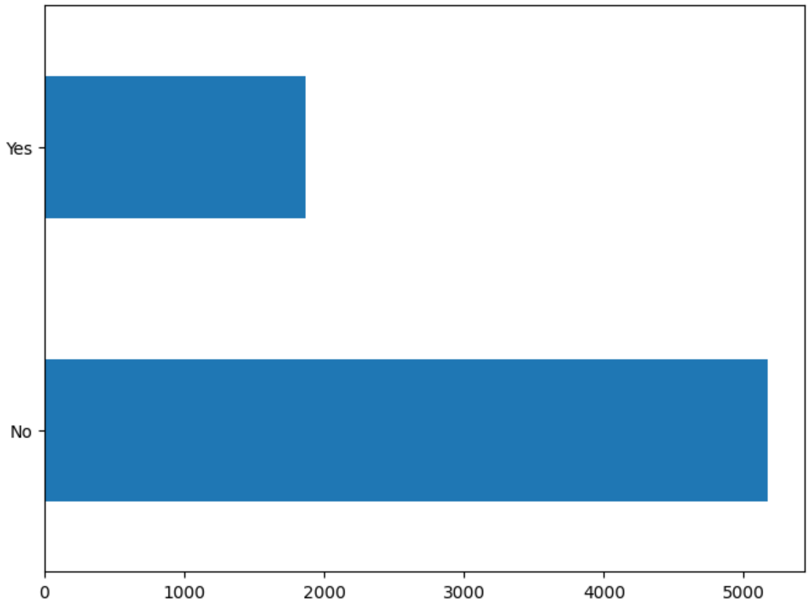
Data Imbalanced

The exploratory data analysis (EDA) revealed a significant class imbalance in the churn variable. The majority class, labelled as "No," represented approximately 73.46% of the dataset, while the minority class, labelled as "Yes," accounted for only 26.54%. This class imbalance poses challenges for accurate churn prediction and may result in biassed model performance.



```
100*churn['Churn'].value_counts()/len(churn['Churn'])
```

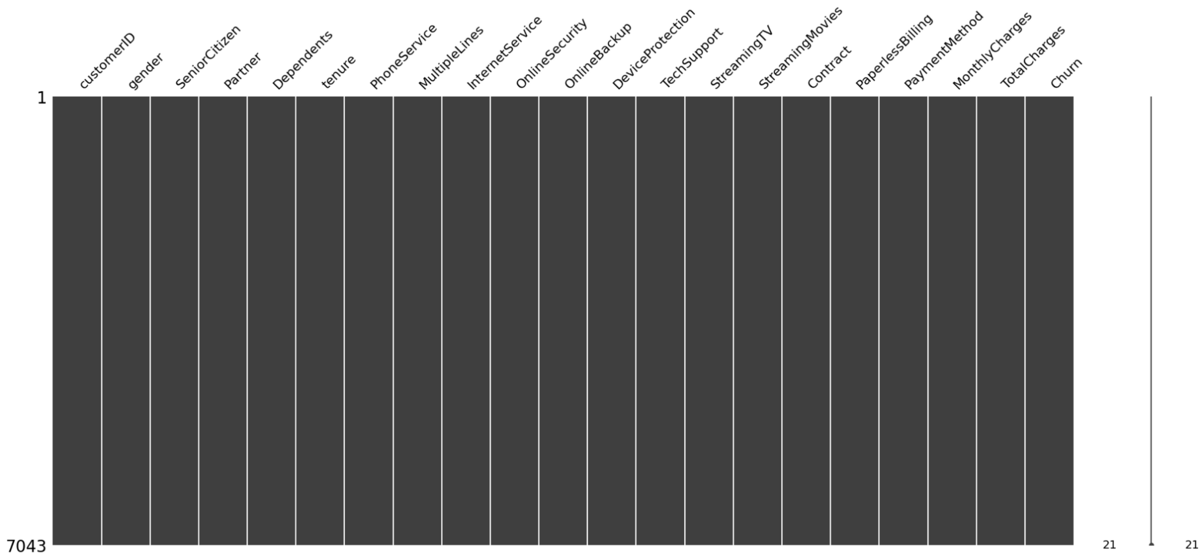
```
No     73.463013
Yes    26.536987
Name: Churn, dtype: float64
```



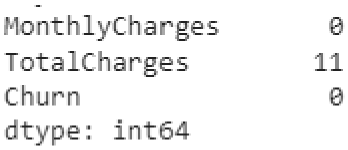
We will be handling this situation in model by methods using undersampling the majority class and advanced sampling techniques like SMOTE (Synthetic Minority Over-sampling Technique). It can help balance the class distribution and enhance the model's ability to predict both churn and non-churn instances accurately.

Missing Values

During the EDA phase, it was observed that the dataset used for churn prediction analysis exhibited no missing values.



But when we check data types of features. We see something suspicious. “TotalCharges” has an object type when we change it to numerics. We unravel 11 missing values.

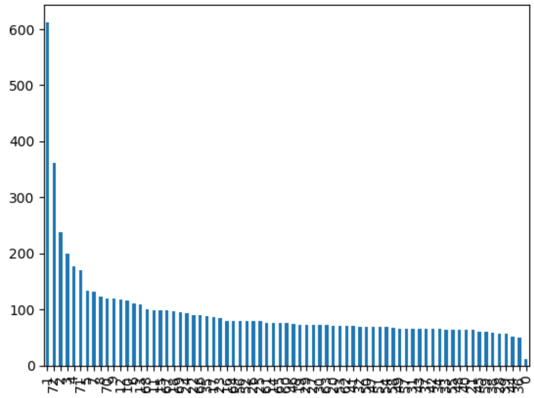


```
MonthlyCharges      0
TotalCharges       11
Churn               0
dtype: int64
```

The values are then replaced by mean.

Unnecessary Skewness of Tenure Feature

The tenure feature graph exhibit skewness:

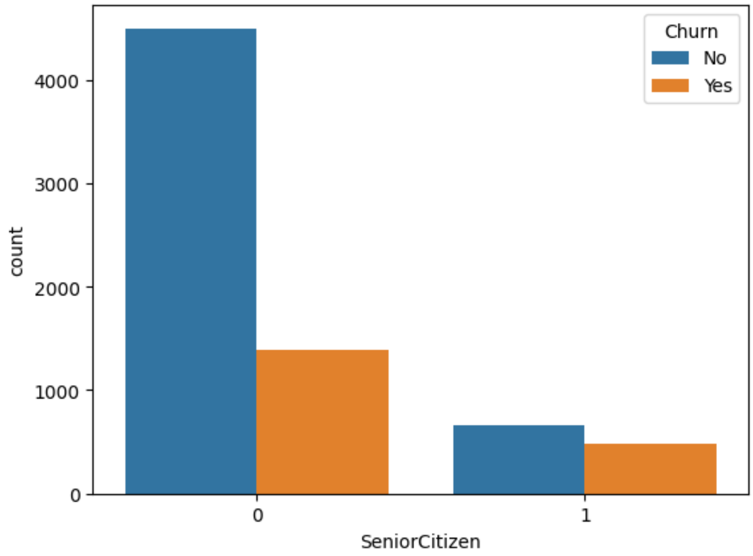


But the skewness doesn't mean anything here. As we see, the most people churned are at 0 months. So, we eliminate the people who churned at 0 months.Its elimination also helps in less bias.

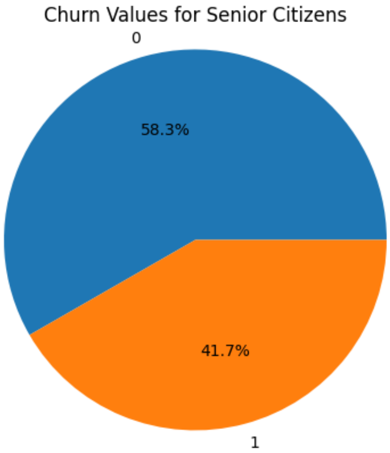
Bivariate Analysis

Bivariate analysis shows relationship or correlation between other features to “Churn” to gain insights into their mutual interactions.
Following are some insights:

SeniorCitizen



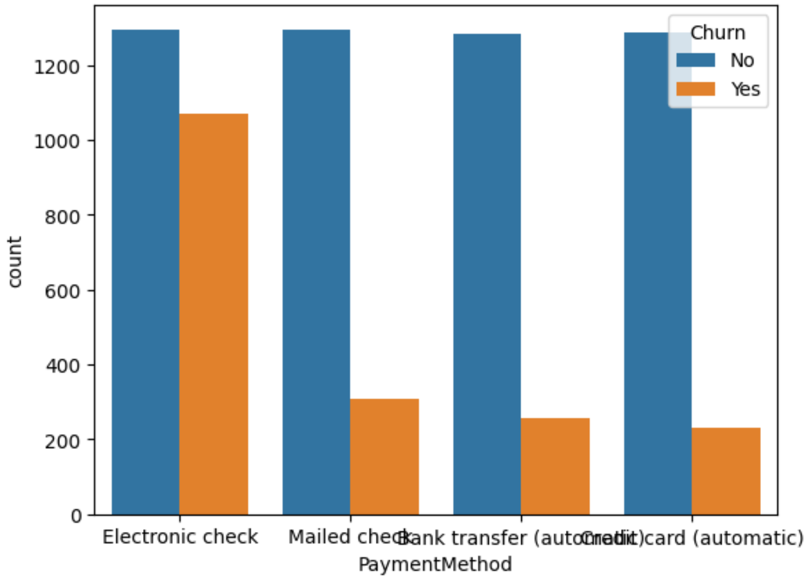
The ones who are senior citizens are more likely to churn out. But the difference is not that much.



For these 41%, we need to offer solutions to retain the senior citizens' customers.

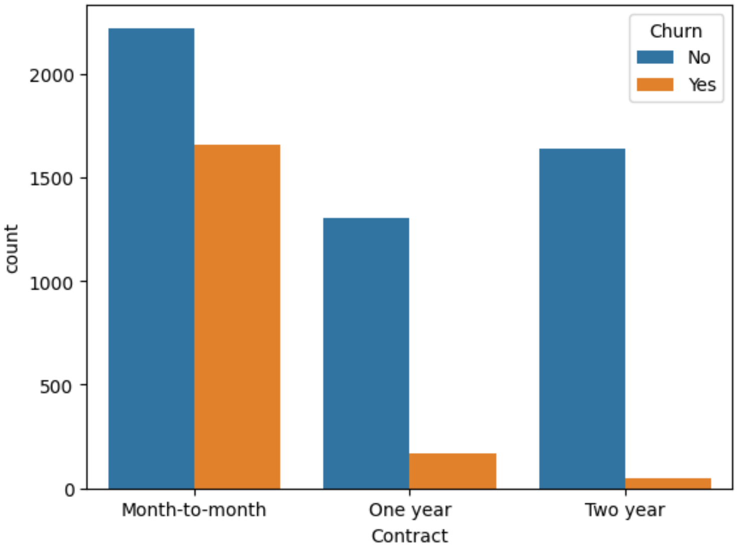
PaymentMethod

Here, we see the automatic payment, bank transfer and mailed check have a great non churning ratio. This insight is valuable as it tells us to move customers from Electronic check to other mediums.



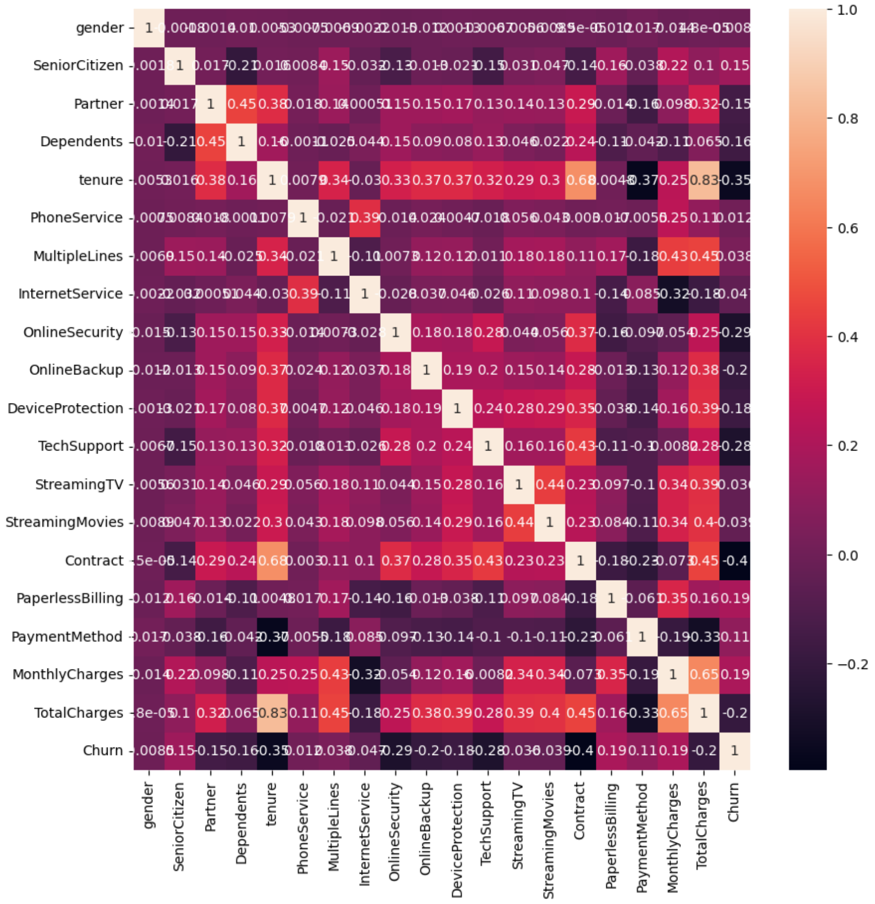
Contract

Here, Month-to-Month has the highest retaining customer. And Two year contracts have a great churning ratio. This insight is valuable as it tells us we need to move people to two year contracts or at least maintain a monthly contract.



HEATMAP

The heatmap visualisation indicates that the data exhibits weak correlation among the variables..



MODEL

Data Split and Training

20% Testing, 80% Training

During the data preparation phase, the dataset was split into two portions: 20% of the data was allocated for testing purposes, while the remaining 80% was used for training the predictive model. This split ratio ensures that a substantial portion of the data is dedicated to training the model, allowing it to learn patterns and make accurate predictions. The testing data, on the other hand, serves as an independent evaluation set to assess the model's performance and generalisation capabilities. By adhering to this 20:80 data split, the analysis can provide reliable insights into the model's effectiveness and its ability to predict churn accurately.

Model Architecture and Training

In this report, the RandomForestClassifier algorithm was employed as the chosen model for predicting customer churn. RandomForestClassifier is an ensemble learning technique that combines multiple decision trees to generate accurate predictions. It is widely used for classification tasks due to its ability to handle complex data, high-dimensional feature spaces, and capture non-linear relationships effectively.

The primary objective of this report is to apply the RandomForestClassifier algorithm to predict customer churn. By leveraging its strengths in handling diverse feature types and capturing intricate relationships, the model aims to deliver accurate churn predictions.Throughout this report, we will explore the model development process, including parameter tuning, feature importance analysis, and performance evaluation.

Performance Metrics

Accuracy - 81%

The model achieved an accuracy of 81%, indicating that it correctly classified 81% of instances in the dataset. This level of accuracy suggests that the model has a relatively high ability to predict customer churn accurately.

Precision - 82%

The precision score of 82% indicates that the model accurately identifies 82% of the instances predicted as positive (churn) correctly. A high precision suggests a low false positive rate, meaning that the model is precise in its positive predictions

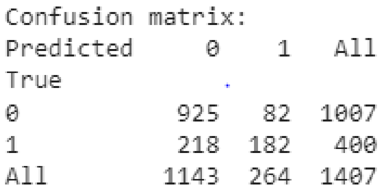
F1 Score - 81%

An F1 score of 81% indicates that the model achieves a good balance between precision and recall, resulting in accurate predictions for both churn and non-churn instances.

Recall - 81%

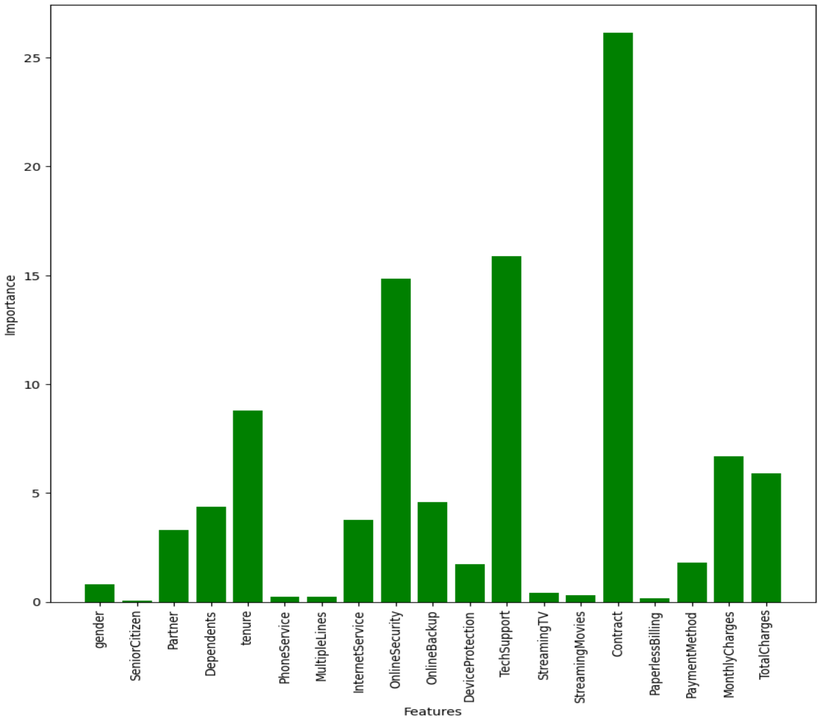
With a recall score of 81%, the model demonstrates its capability to effectively capture most of the churn instances present in the dataset. This implies that the model has a relatively low rate of missing customers who are likely to churn.

Confusion Matrix



```
Confusion matrix:
Predicted      0    1   All
True
0            925   82  1007
1            218  182   400
All         1143  264  1407
```

FEATURE IMPORTANCE



During the churn prediction analysis, the importance of features was evaluated to understand their contribution in predicting customer churn. The analysis revealed that the following features were significant in determining churn:

Contract

The contract type was identified as an important feature in predicting churn. Customers with certain contract types may exhibit different churn behaviors, making it a crucial factor in understanding and predicting churn likelihood.

TechSupport

The availability and quality of technical support emerged as another significant feature. Customers who receive satisfactory technical support may have a lower likelihood of churning, highlighting the importance of this service in customer retention efforts.

OnlineSecurity

The presence or absence of online security measures was found to impact churn prediction. Customers who perceive their online security needs are adequately met may be less likely to churn.

Tenure

The length of a customer's tenure with the company was identified as a relevant feature. Long-term customers may exhibit different churn patterns compared to new customers, making tenure an essential factor in understanding and predicting churn. These features provide valuable insights into the drivers of customer churn.

Conclusion

In this churn prediction analysis, the RandomForestClassifier model demonstrated satisfactory performance with an accuracy, precision, F1 score, and recall of 81%.

These metrics indicate that the model can predict customer churn with a reasonable level of accuracy and effectively identify customers at risk of churning.

Overall, the RandomForestClassifier model shows promise in accurately predicting customer churn and provides valuable insights for businesses to implement targeted retention strategies. However, continuous monitoring and refinement of the model are recommended to further improve its accuracy and optimise customer retention efforts.

By leveraging the strengths of the RandomForestClassifier model, businesses can make informed decisions, enhance customer satisfaction, and reduce churn rates, ultimately leading to improved business outcomes.