IBM Supervised Machine Learning: Classification Assignment

By

Ahmed Shahriar Sakib

Introduction:

The objective of this project is to perform exploratory analysis on **IBM Customer Churn Data** with hypothesis testing and build classification machine learning models to predict customer churn.

Dataset:

Source

- Kaggle Dataset URL: https://www.kaggle.com/blastchar/telco-customer-churn
- GitHub Dataset URL: https://github.com/IBM/telco-customer-churn-on-icp4d/tree/master/data

Descriptive Analysis

Dataset dimension: 7043 rows × 21 columns

Data Types:

Data Type	Count
object	18
int64	2
float64	1

Feature Name	Description	Data Type
customerID	Contains customer ID	categorical
gender	whether the customer female or male	categorical
SeniorCitizen	Whether the customer is a senior citizen or not (1, 0)	numeric, int
Partner	Whether the customer has a partner or not (Yes, No)	categorical
Dependents	Whether the customer has dependents or not (Yes, No)	categorical
tenure	Number of months the customer has stayed with the company	numeric, int
PhoneService	Whether the customer has a phone service or not (Yes, No)	categorical

MultipleLines	Whether the customer has multiple lines r not (Yes, No, No phone service)	categorical
InternetService	Customer's internet service provider (DSL, Fiber optic, No)	categorical
OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)	categorical
OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)	categorical
DeviceProtection	Whether the customer has device protection or not (Yes, No, No internet service)	categorical
TechSupport	Whether the customer has tech support or not (Yes, No, No internet service)	categorical
streamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)	categorical
streamingMovies	Whether the customer has streaming movies or not (Yes, No, No internet service)	categorical
Contract	The contract term of the customer (Month-to-month, One year, Two year)	categorical
PaperlessBilling	Whether the customer has paperless billing or not (Yes, No)	categorical
PaymentMethod	The customer's payment method (Electronic check, Mailed check, Bank transfer, Credit card)	categorical
MonthlyCharges	The amount charged to the customer monthly	numeric, int
TotalCharges	The total amount charged to the customer	object
Churn	Whether the customer churned or not (Yes or No)	categorical

Table: Feature Summary

Data Exploration Plan

First, I will check for data type mismatch, missing values, binning if applicable and clean the data accordingly. Then I will perform different kinds of hypothesis test such as normality test, variable dependency test, multicollinearity test based on the data types. After that I will explore the dataset using different types of visualization methods, encode both categorical and numerical feature and normalize the dataset and prepare dataset for ML modeling. Finally, I will apply classic ML algorithms as well as some popular gradient boosting algorithms to predict customer churn rate.

Data Cleaning and Feature Engineering

Data Cleaning

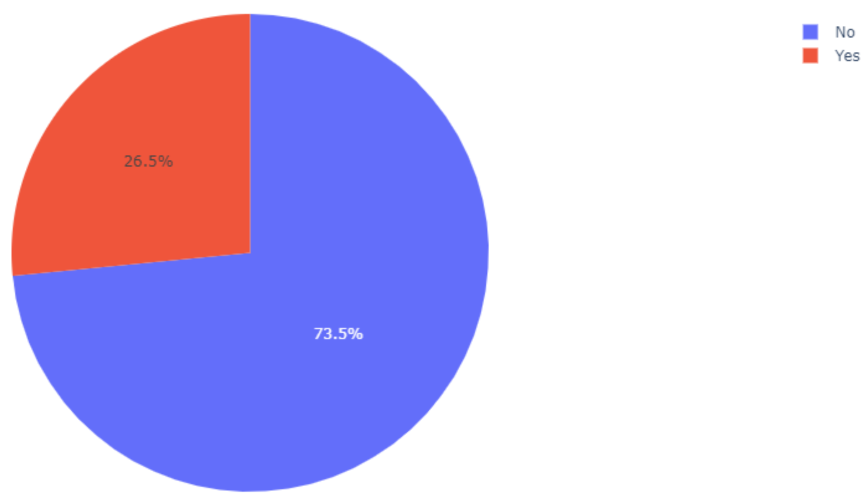
- Deleted unnecessary column ("**Id**")
- Checked data types and assign appropriate type ("**TotalCharges**" to numeric)
- Checked for missing values and imputation ("**TotalCharges**")

Feature Engineering

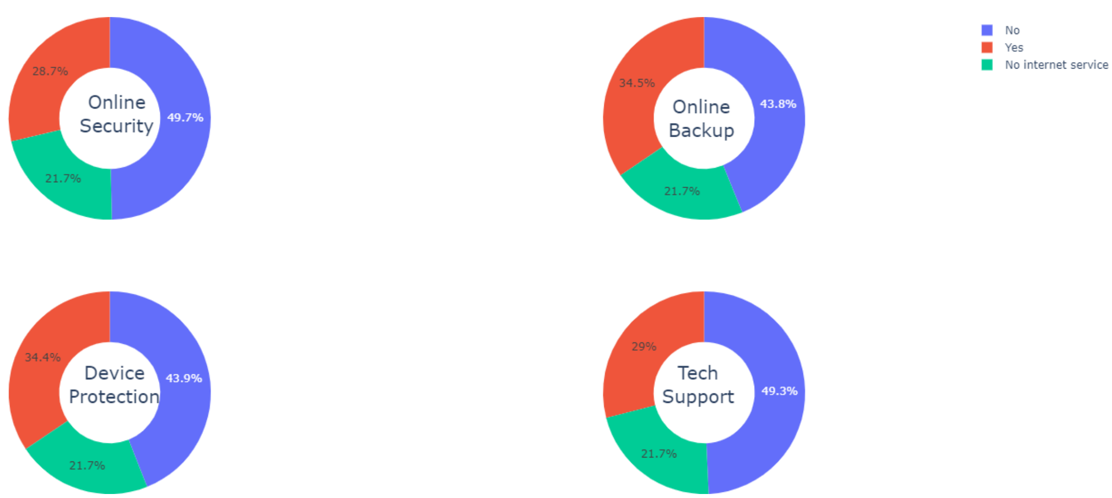
- Binning numerical features (**Tenure**, **MonthlyCharges** and **TotalCharges**)
- Correlation Test
- One hot, ordinal and label Encoding
- Check feature importance with gradient boosting models

Key Findings and Insights

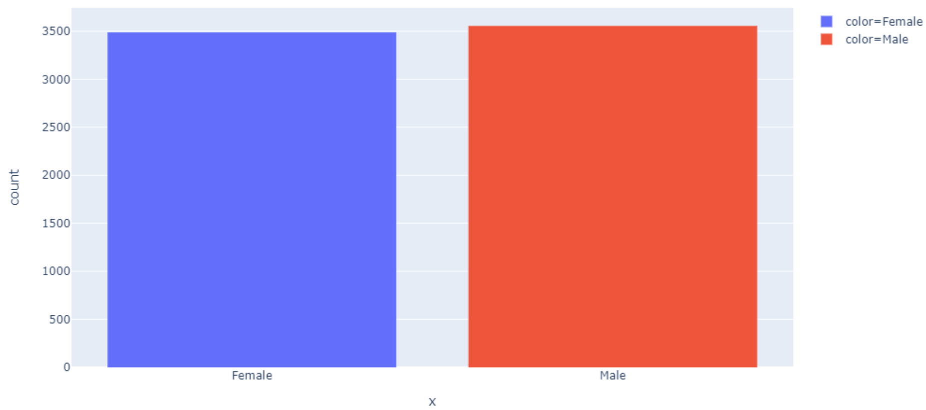
1. Imbalanced Dataset (target - churn)



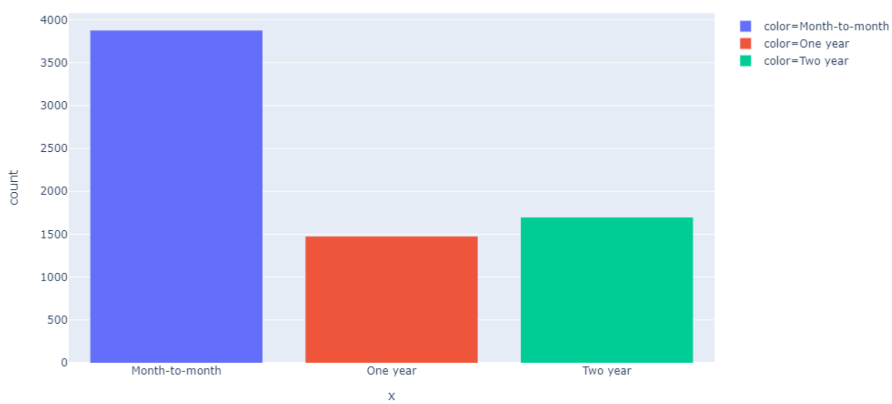
2. Proportion of online utility usage in similar among customers



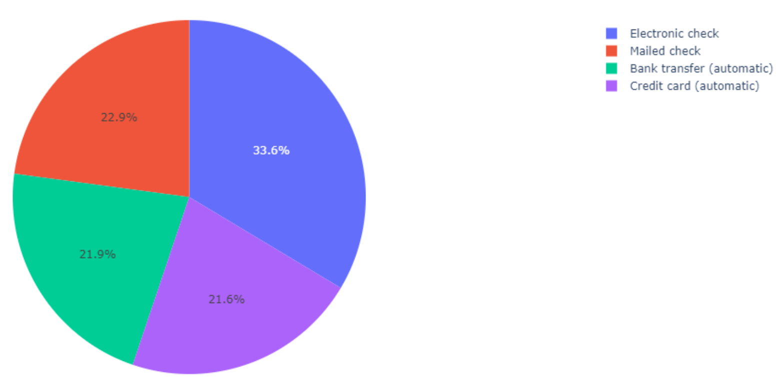
3. Almost 1:1 Gender ratio, relatively balanced



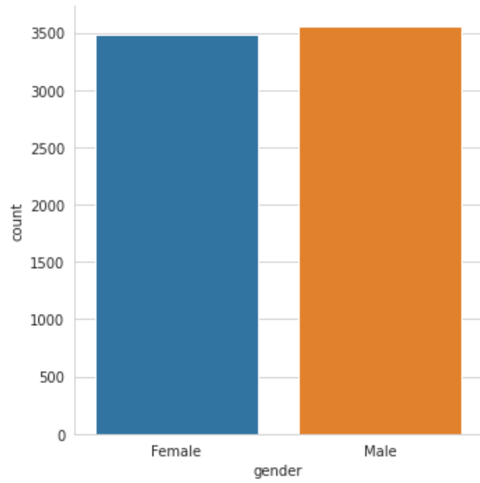
4. Month-to-month contract is higher than others



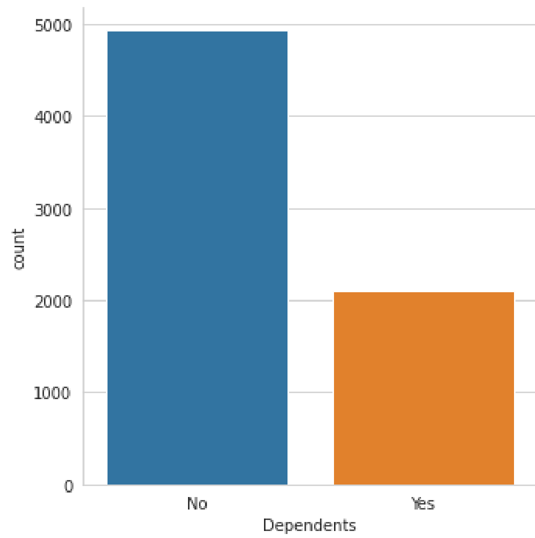
5. Most of the customers use E-check



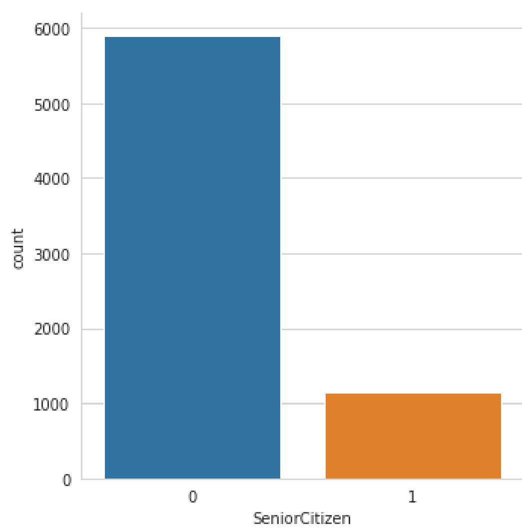
6. Approximately 50/50 gender ratio



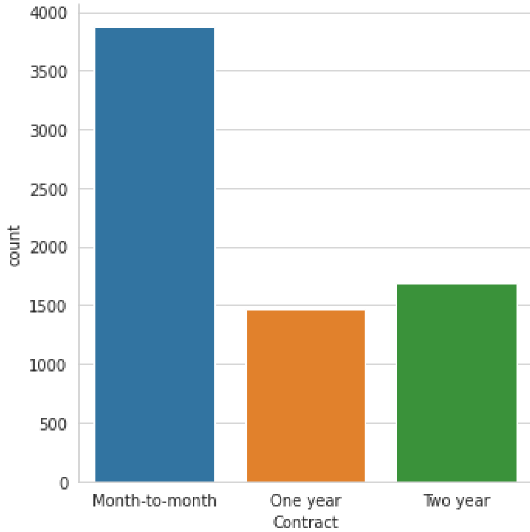
7. Users who have non-dependents are approximately two times more than users having dependents



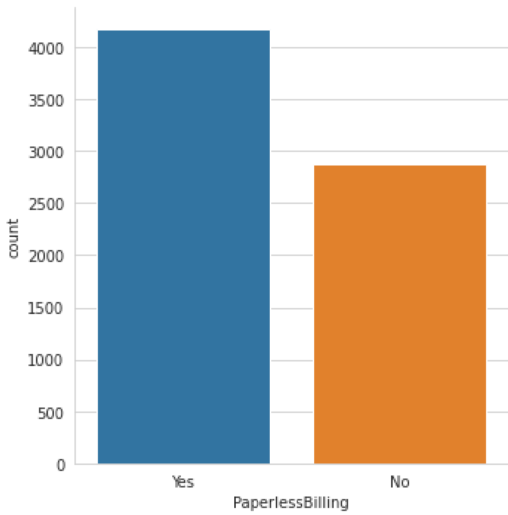
8. Most of the users are not Senior Citizen



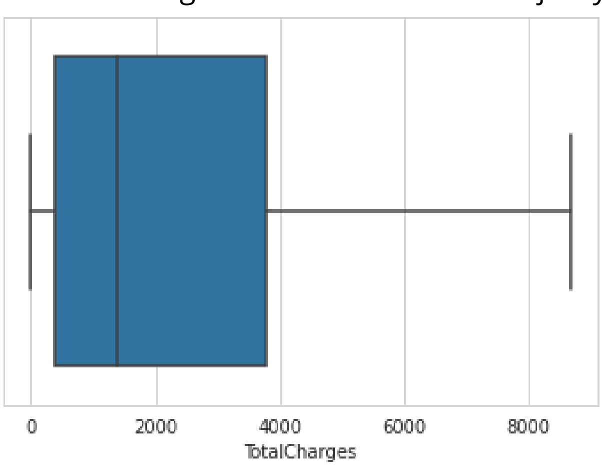
9. Most of the users prefer Month-to-month contract



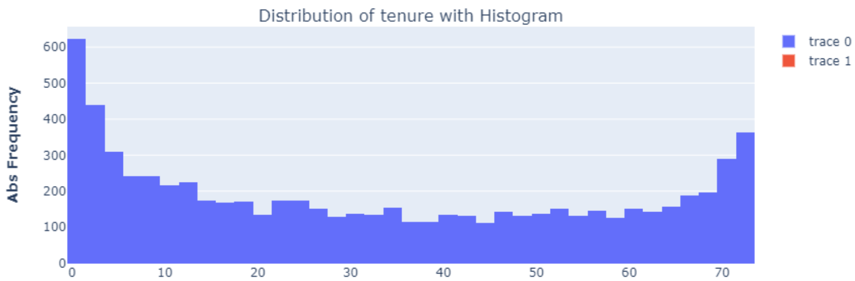
10. Most of the users prefer paperless billing

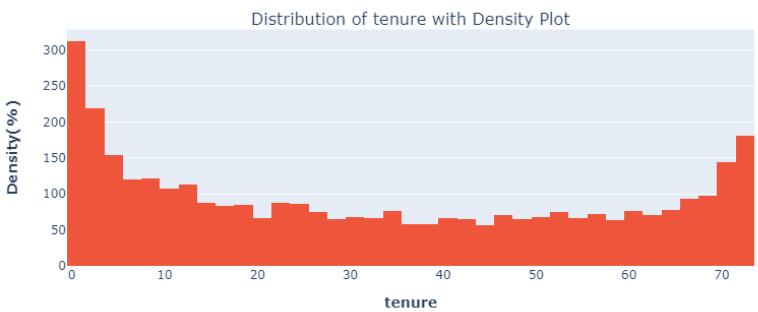


11. The total charges fall under 4000 for majority of the users

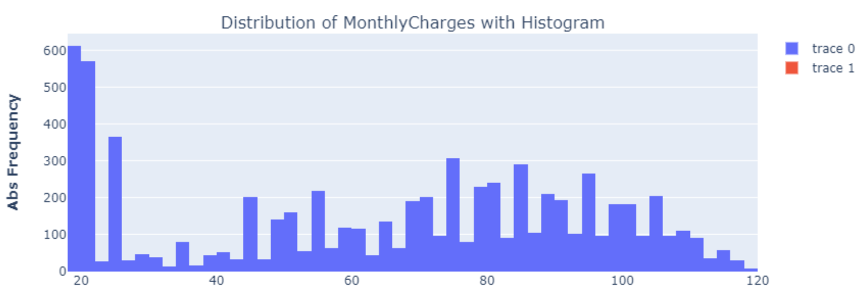


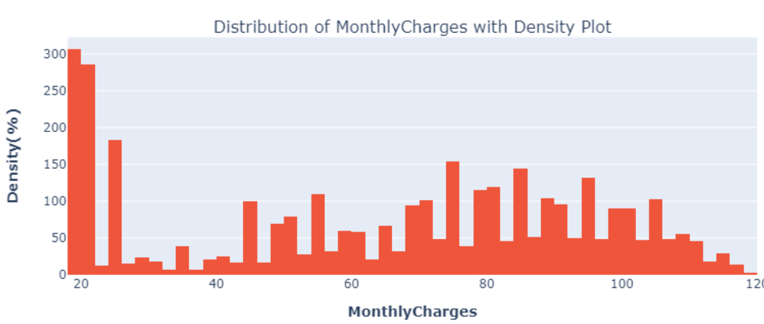
12. Tenure is U-shaped distributed



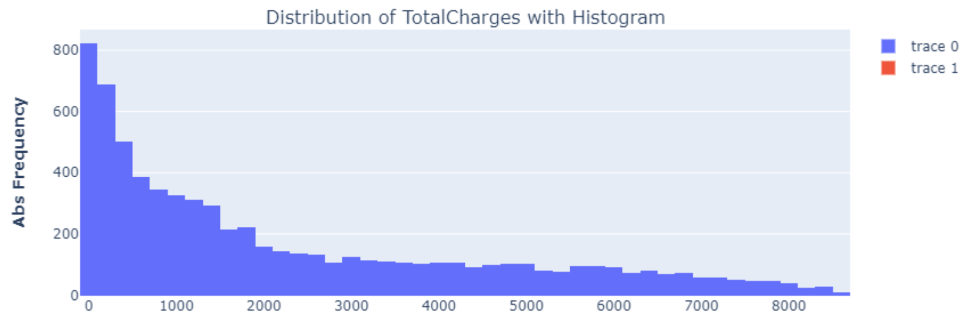


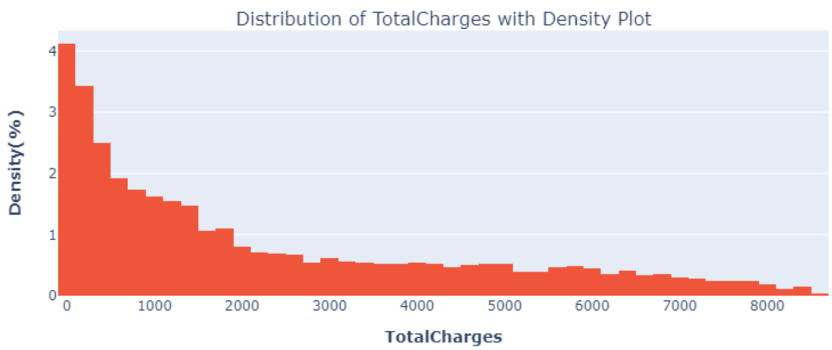
13. Monthly Charges is heavily skewed



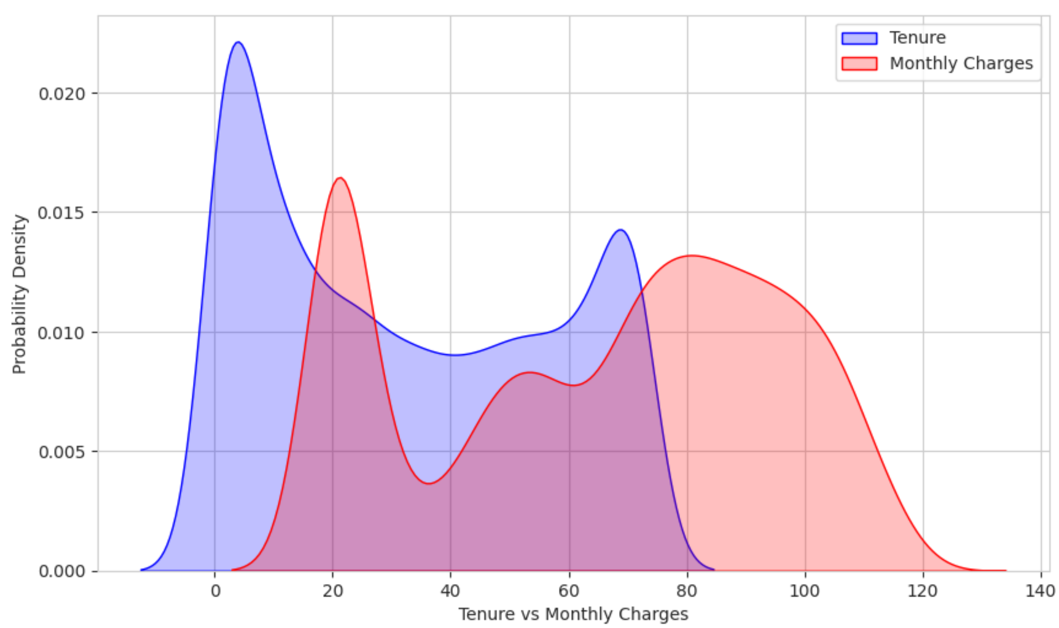


14. Total Charges is reversed J-shaped distributed

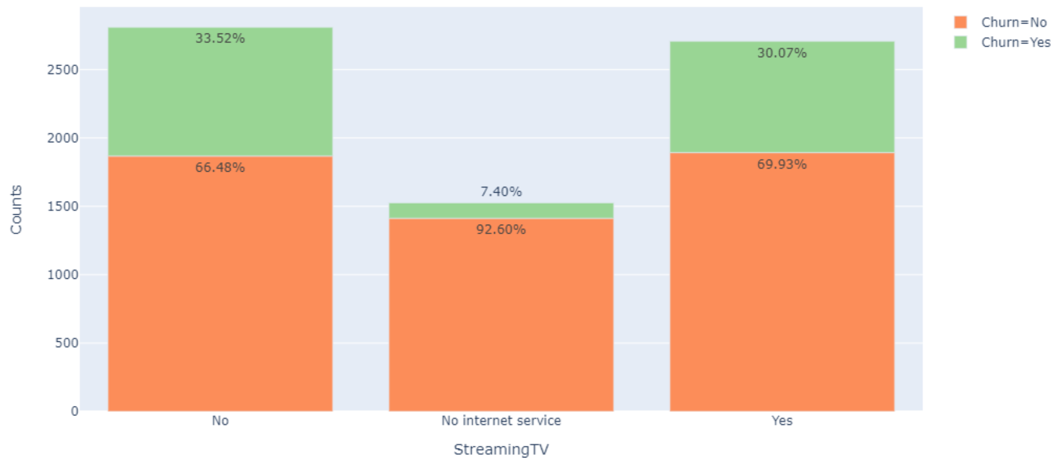




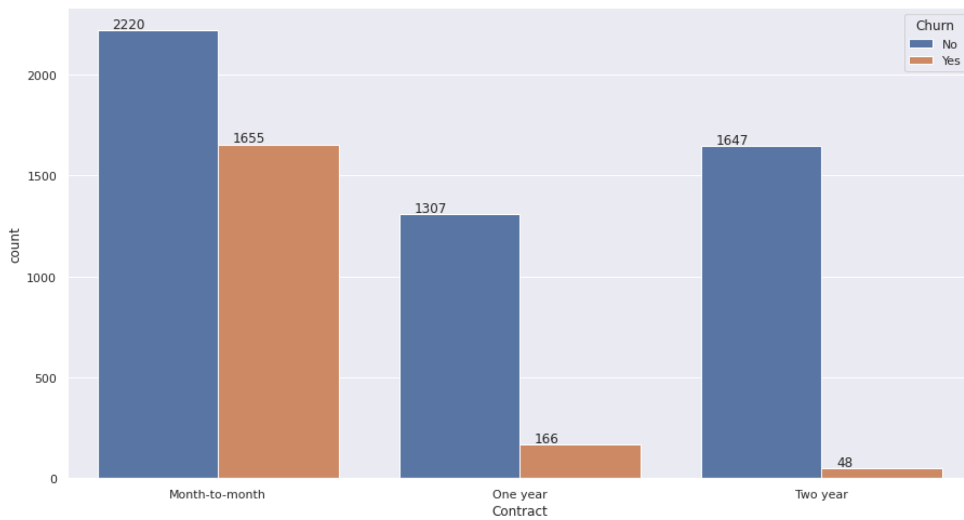
15. Both are are not normally distributed, skewed. Tenure has a Bi-modal distribution. Most users stayed for less than 20 months, Monthly Charges for most people is nearly 20 unit



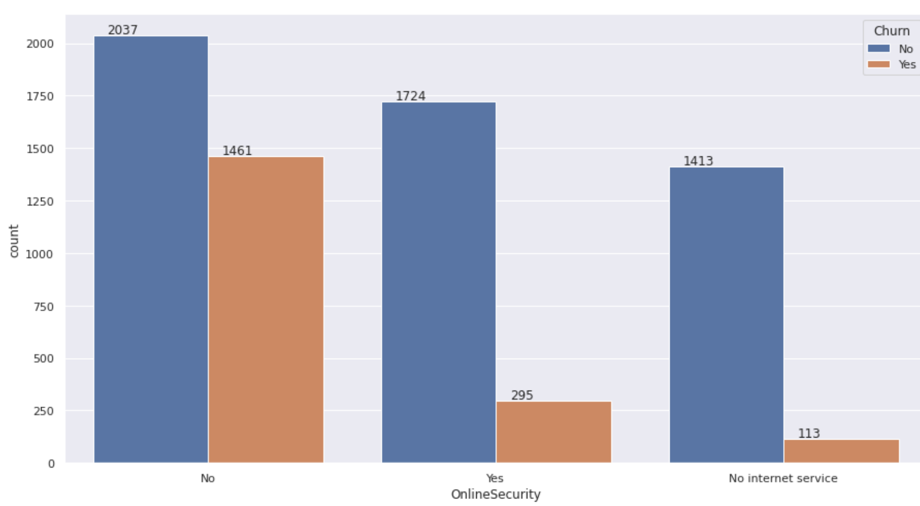
16. Similar ratio between streamer vs non-streamer in churned users



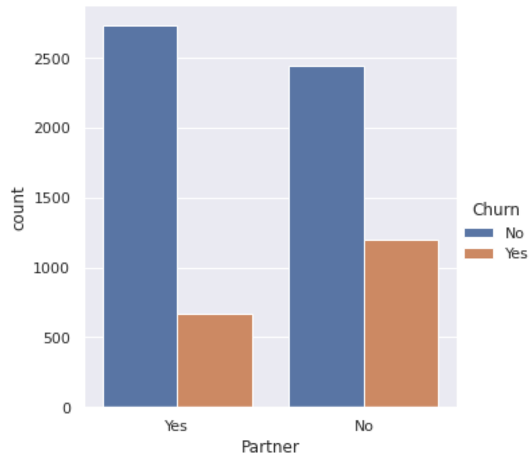
17. Most churned users has Month-to-month contract



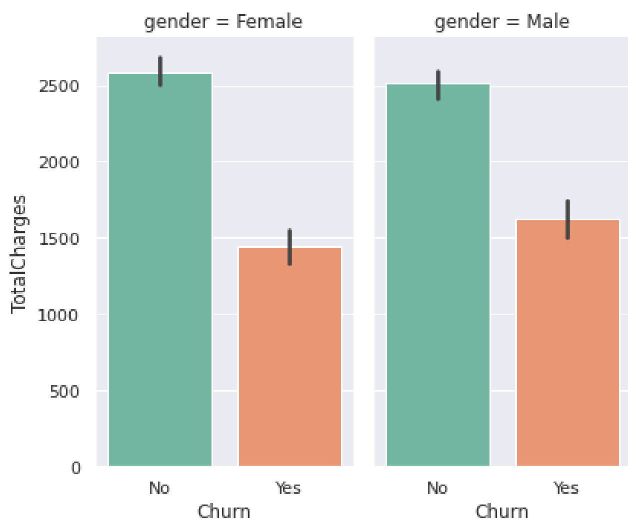
18. Most churned users didn't have online security



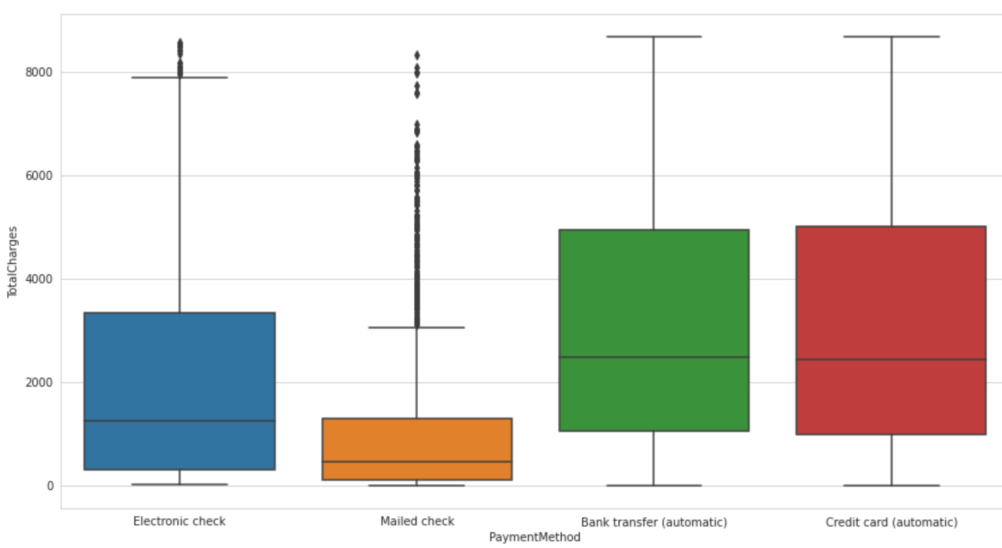
19. Most users who churned does not have a partner in contrast to the users who does



20. Gender is uncorrelated with churn rate



21. Total Charges for many users are in extreme level in Mailed Check payment method



Hypothesis Test

Normality test

Hypotheses

H0: the sample has a Gaussian distribution

H1: the sample does not have a Gaussian distribution

alpha (significance level) = 0.05

If p-value > alpha, then Null hypotheses gets rejected.

1. D'Agostino's K^2 Test

Variable	P	Statistics	Decision
MonthlyCharges	0.0000	11419.5287	Sample does not look gaussian(reject Ho)
Tenure	0.0000	76258.5051	Sample does not look gaussian(reject Ho)

2. Anderson-Darling Test

Result of the test on the "TotalCharges" column

Significance Level (%)	Critical value	Decision
15.000	0.576	data does not look normal (reject H0)
10.000	0.656	data does not look normal (reject H0)
5.000	0.787	data does not look normal (reject H0)
2.000	0.917	data does not look normal (reject H0)
1.000	1.091	data does not look normal (reject H0)

Correlation Significance Test

1. Spearman rank-order correlation

Hypotheses

H0: the two samples do not have monotonic relationship

H1: there is a monotonic relationship between the samples

alpha (significance level) = 0.05

If p-value > alpha, then Null hypotheses gets rejected

Result

Variable Pair	Correlation	P	Decision
tenure, MonthlyCharges	0.276	1e-123	have monotonic relationship (reject H0)
TotalCharges, tenure	0.133	2e-29	have monotonic relationship (reject H0)
TotalCharges, MonthlyCharges	0.285	7e-132	have monotonic relationship (reject H0)

2. Kendall rank correlation coefficient

Hypotheses

H0: the two samples are not correlated

H1: Probably correlated

alpha (significance level) = 0.05

If p-value > alpha, then Null hypotheses gets rejected

Result

Variable Pair	Correlation	P	Decision
MonthlyCharges and TotalCharges-binned	-0.00861	0.470	uncorrelated (fail to reject H0)
TotalCharges, tenure-binned	-0.236	0.000	correlated (reject H0)
Tenure, MonthlyCharges-binned	-0.164	0.000	correlated (reject H0)

3. Mann-Whitney U Test

Hypotheses –

H0: population medians are equal.

H1: population medians are not equal.

alpha (significance level) = 0.05

If p-value > alpha, then Null hypotheses gets rejected

Result

Variable With Target : Churn	Correlation	P	Decision
tenure	48981984	0.470	Different distribution (reject H0)
TotalCharges	49603849	0.000	Different distribution (reject H0)
MonthlyCharges	49554833	0.000	Different distribution (reject H0)

4. Chi-Square

Hypotheses

H0: the two samples are not dependent

H1: Probably dependent

alpha (significance level) = 0.05

Test statistic in the context of the chi-squared distribution with the requisite number of degrees of freedom

In terms of a p-value and a chosen significance level (alpha):

- If p-value <= alpha: significant result, reject null hypothesis (H0), dependent.
- If p-value > alpha: not significant result, fail to reject null hypothesis (H0), independent.

Crosstab between "OnlineSecurity" and "PaymentMethod" columns

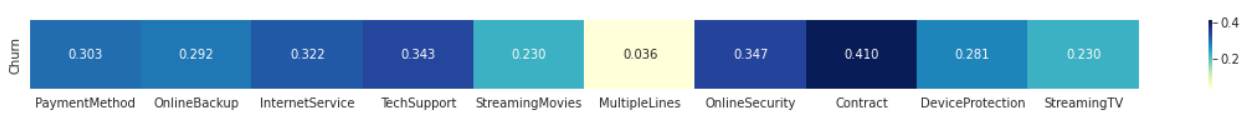
PaymentMethod	**Bank transfer (automatic)**	**Credit card (automatic)**	**Electronic check**	**Mailed check**
OnlineSecurity				
No	644	603	1734	517
No internet service	332	331	122	741
Yes	568	588	509	354

Result

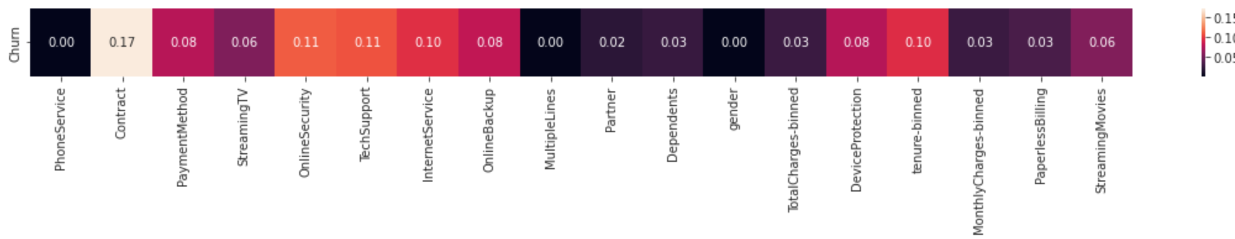
p-value : 7.427176940680162e-280, **degree of freedom**: 6

Test Type	Values		Decision
Test-statistic	critical = 12.592	Statistic= 1309.996	Dependent (reject H0)
p-value	significance=0.050	P value=0.000	Dependent (reject H0)

5. **Cramers V** Heatmap on Polytomous Features and Target: Churn



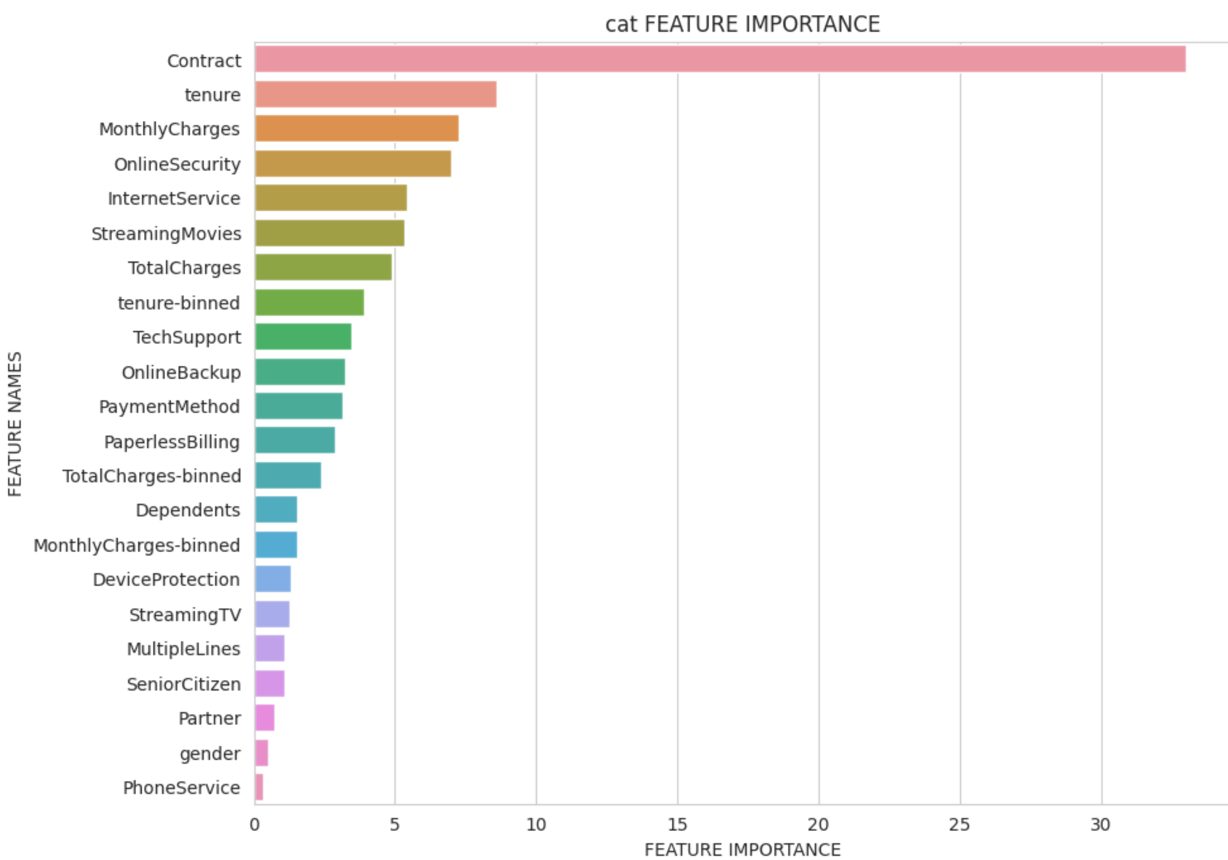
6. **Theil's U** test - Contract, OnlineSecurity, TechSupport, tenure-binned are moderately correlated with Churn



Modeling

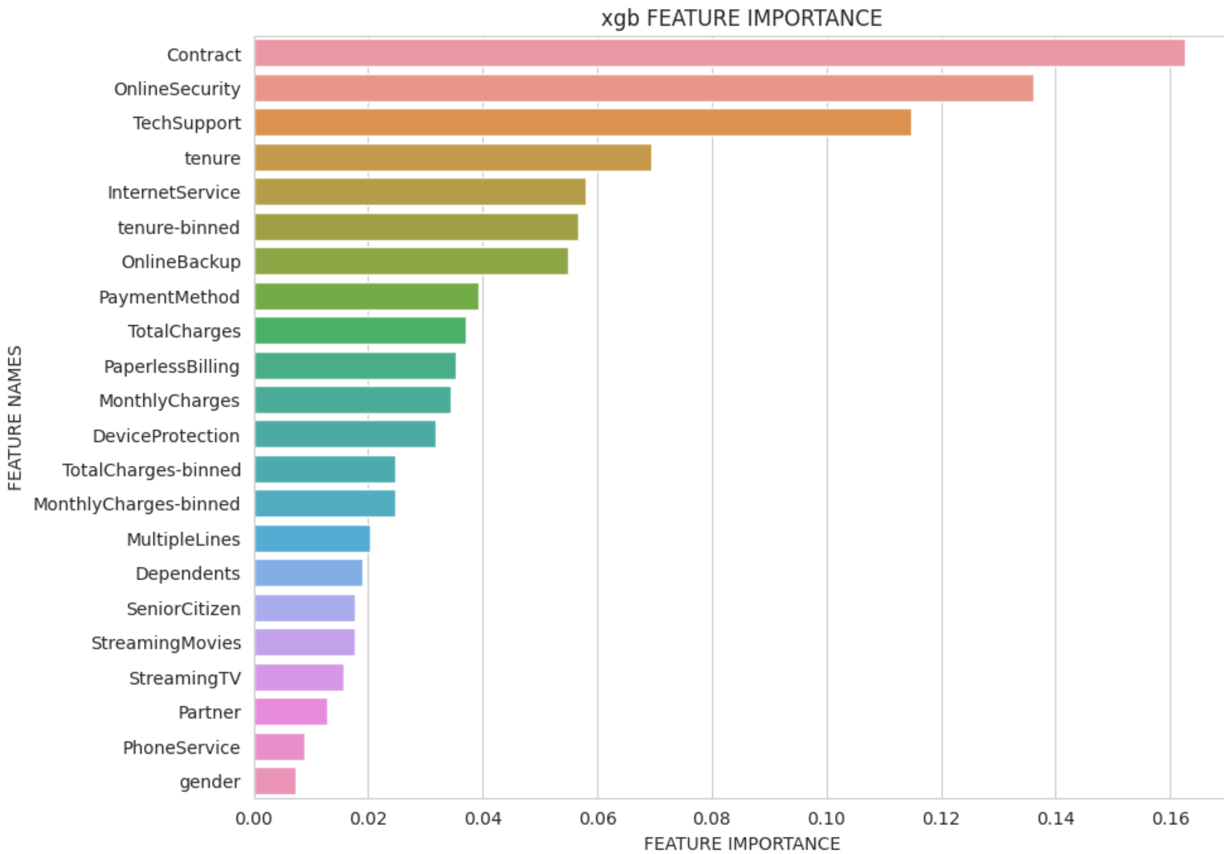
CatBoost

With 10-fold cross validation it reached an AUC of 0.8472. The hyperparameters were tuned using Optuna.



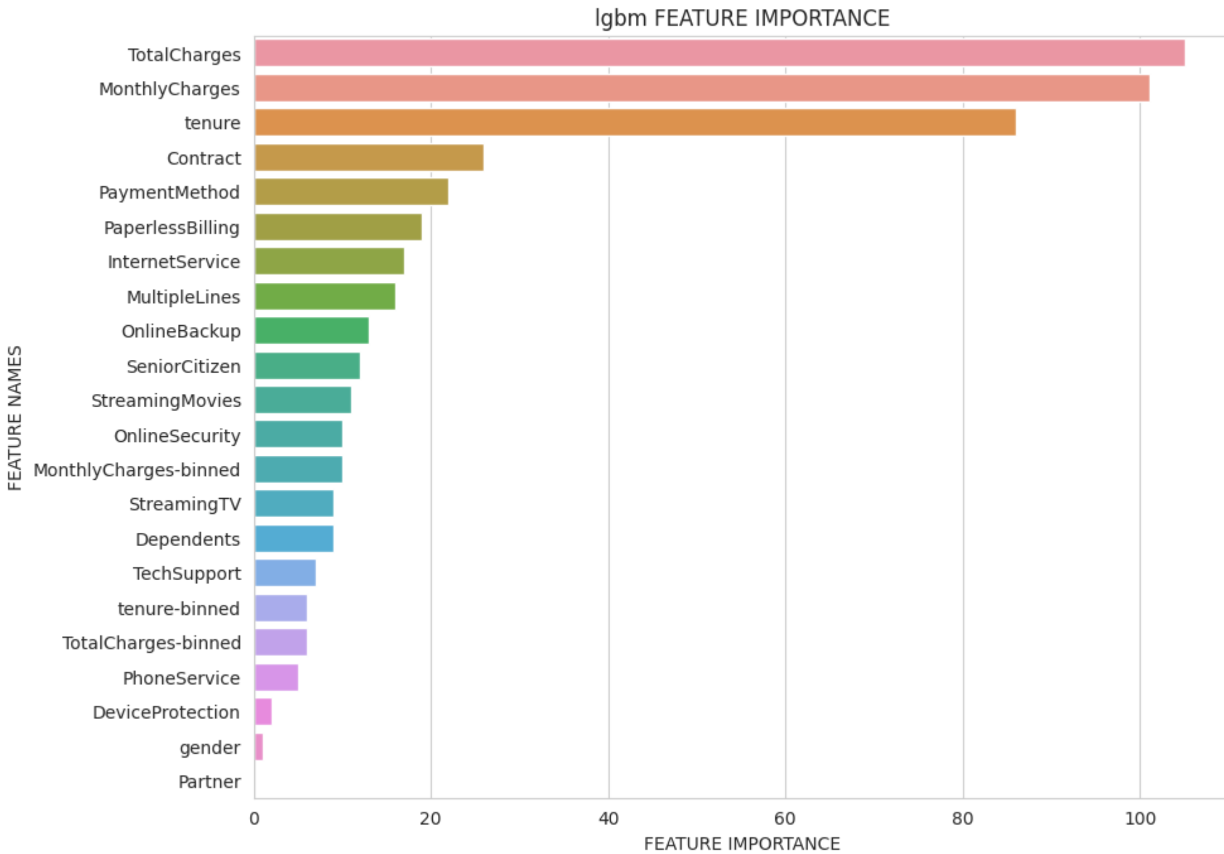
XGBoost

With 10-fold cross validation it reached an AUC of 0.8468. The hyperparameters were tuned using Optuna.



LightGBM

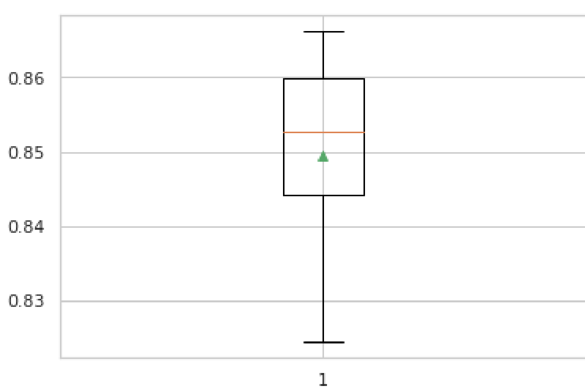
With 10-fold cross validation it reached an AUC of 0.8498. The hyperparameters were tuned using Optuna.



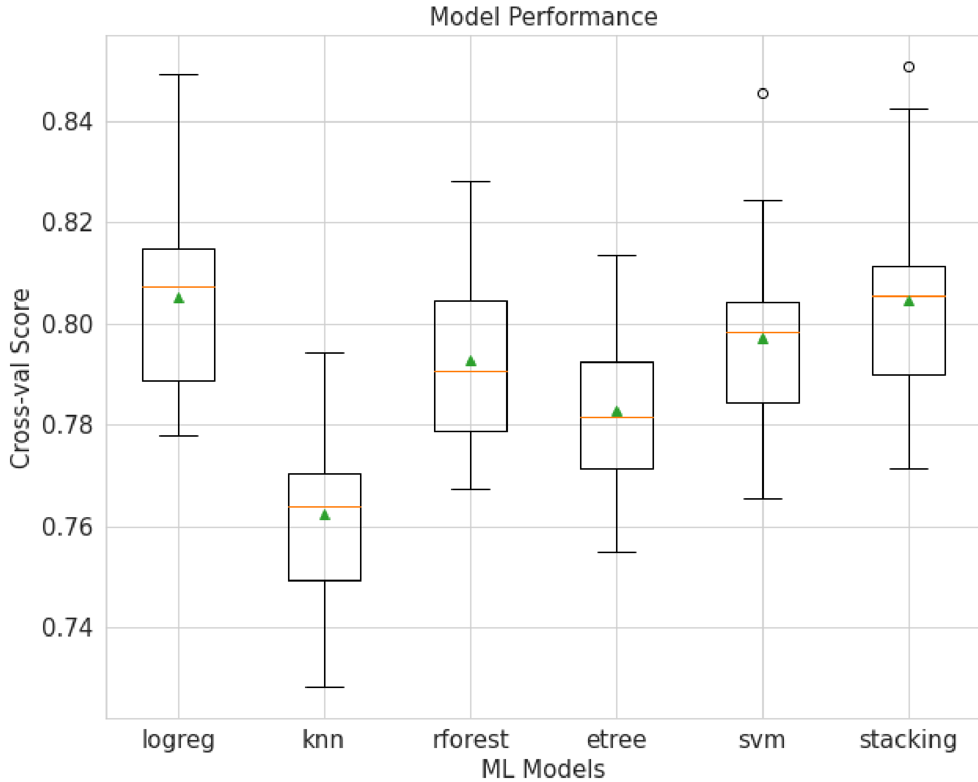
Stacking Ensemble

1. **Gradient Boosting Models**: With LightGBM, XGBoost and CatBoost as base models and Logistic regression as meta model, the single level stacking ensemble reached an AUC of ~0.8486.

5-fold Training AUC plot:



2. **Classic ML Models**: With KNN, Extra Tree, Random Forest and Logistic regression the stacking classifier reached an AUC of ~0.805.



Future Work

- Apply Neural Network
- Ensemble Modeling (Blending)
- Build an API

Conclusion

This dataset is small, it has less features too. To improve the Machine learning model and for a rigorous analysis more data is needed.