TELCO CUSTOMER CHURN PREDICTION

Kalum De Alwis - 265

INTRODUCTION

The telecom industry has traditionally experienced an average churn rate of 30% to 35%, while the cost per acquisition (CPA) typically ranges from five to 10 times higher than the cost of retaining an existing customer, according to the 2016 sector review Big Data Applications in the Telecommunications Industry.

PROBLEM

This Project tries to predict churn of customer in a telco environment using machine learning

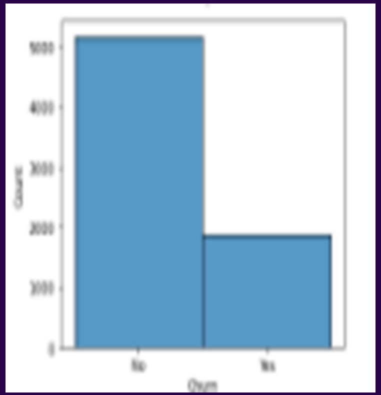
Data set

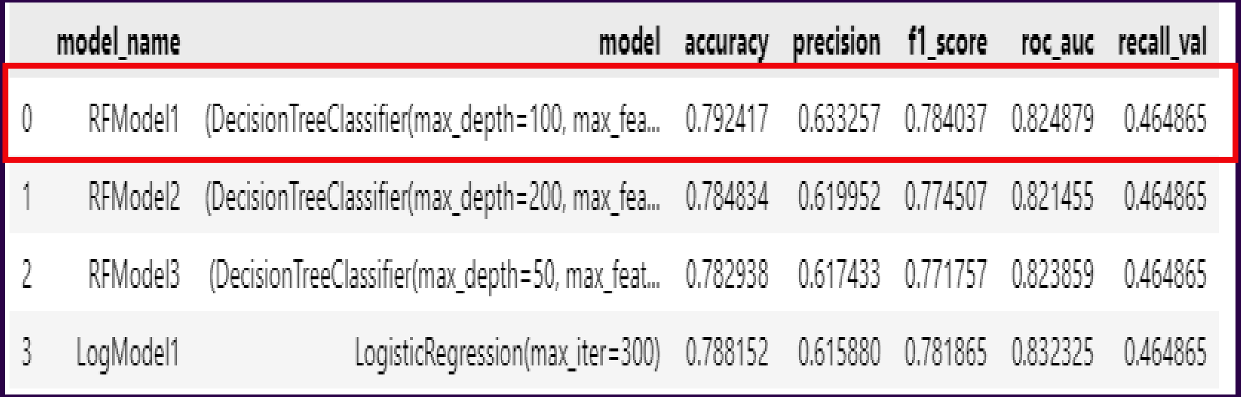
Data set is WA_Fn-UseC_-Telco-Customer-Churn.csv which is sourced from https://www.kaggle.com/datasets/palashfendarkar/wa-fnusec-telcocustomerchurn

Algorithms in consideration

Since this is a binary classification problem Random forest and logistic regression were targeted as possible ML algorithms.

MODEL





	model_name	model	accuracy	precision	f1_score	roc_auc	recall_val
0	RFModel1	(DecisionTreeClassifier(max_depth=100, max_fea...	0.792417	0.633257	0.784037	0.824879	0.464865
1	RFModel2	(DecisionTreeClassifier(max_depth=200, max_fea...	0.784834	0.619952	0.774507	0.821455	0.464865
2	RFModel3	(DecisionTreeClassifier(max_depth=50, max_feat...	0.782938	0.617433	0.771757	0.823859	0.464865
3	LogModel1	LogisticRegression(max_iter=300)	0.788152	0.615880	0.781865	0.832325	0.464865

Pre processing

Data set consisted of 21 columns and 7043 rows. Data set had only 3 numerical features and 17 categorical features, it did not contain null values

Model building

Target variable which is 'Churn' had 1869 (26%) samples with 'Yes" values and 5174 with 'no' values (figure 3). Therefore, no special sampling was not done to the test data set considering the nature of telco churn

Model

Random forest and Logistic regression algorithms were used.

Hyper parameter change

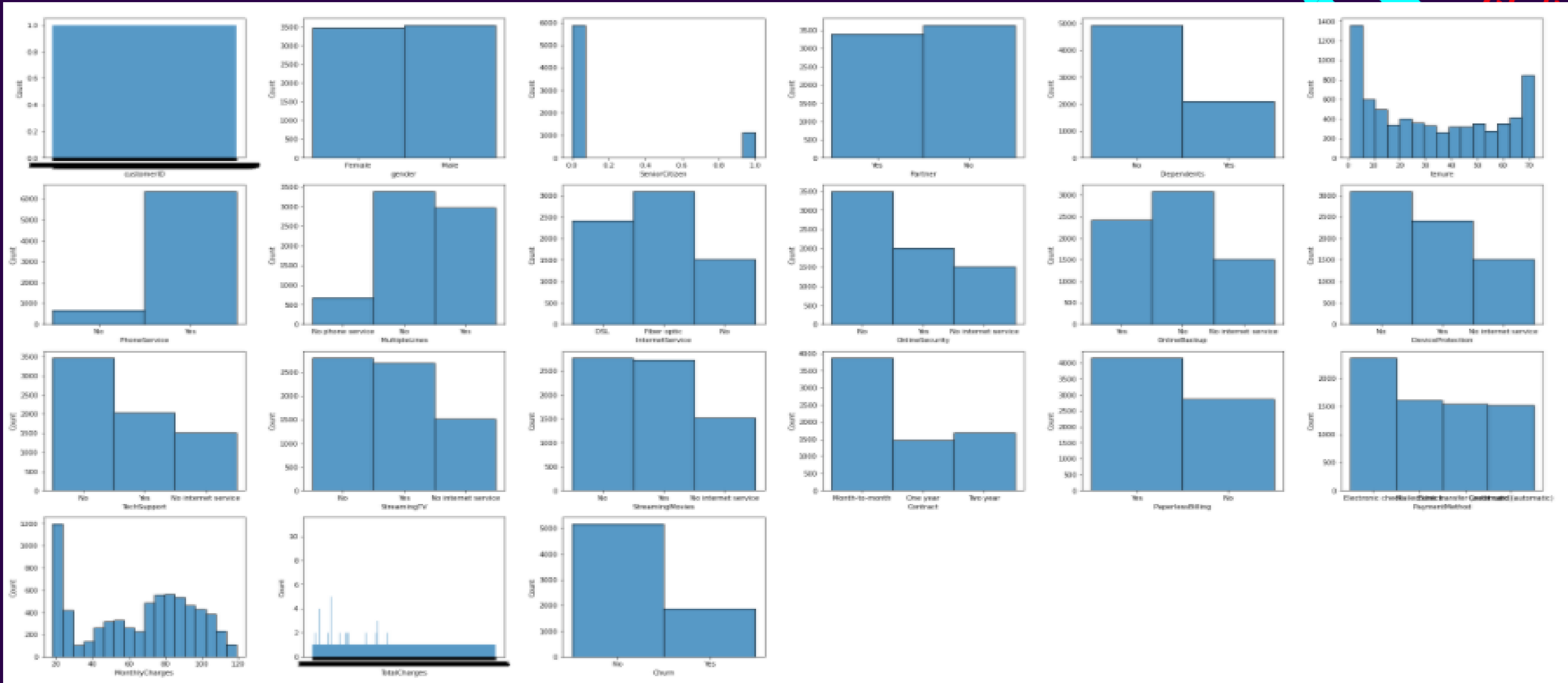
Logistic regression failed to converge with default hyper [parameters and had to set "max_iter=300" to solve the issue.

Deployment

Model was saved using joblib and functions were created to load the model and provide churn prediction for input data sets by user.

ANNEXURES

DATA ANALYSIS -1



DATA ANALYSIS -1

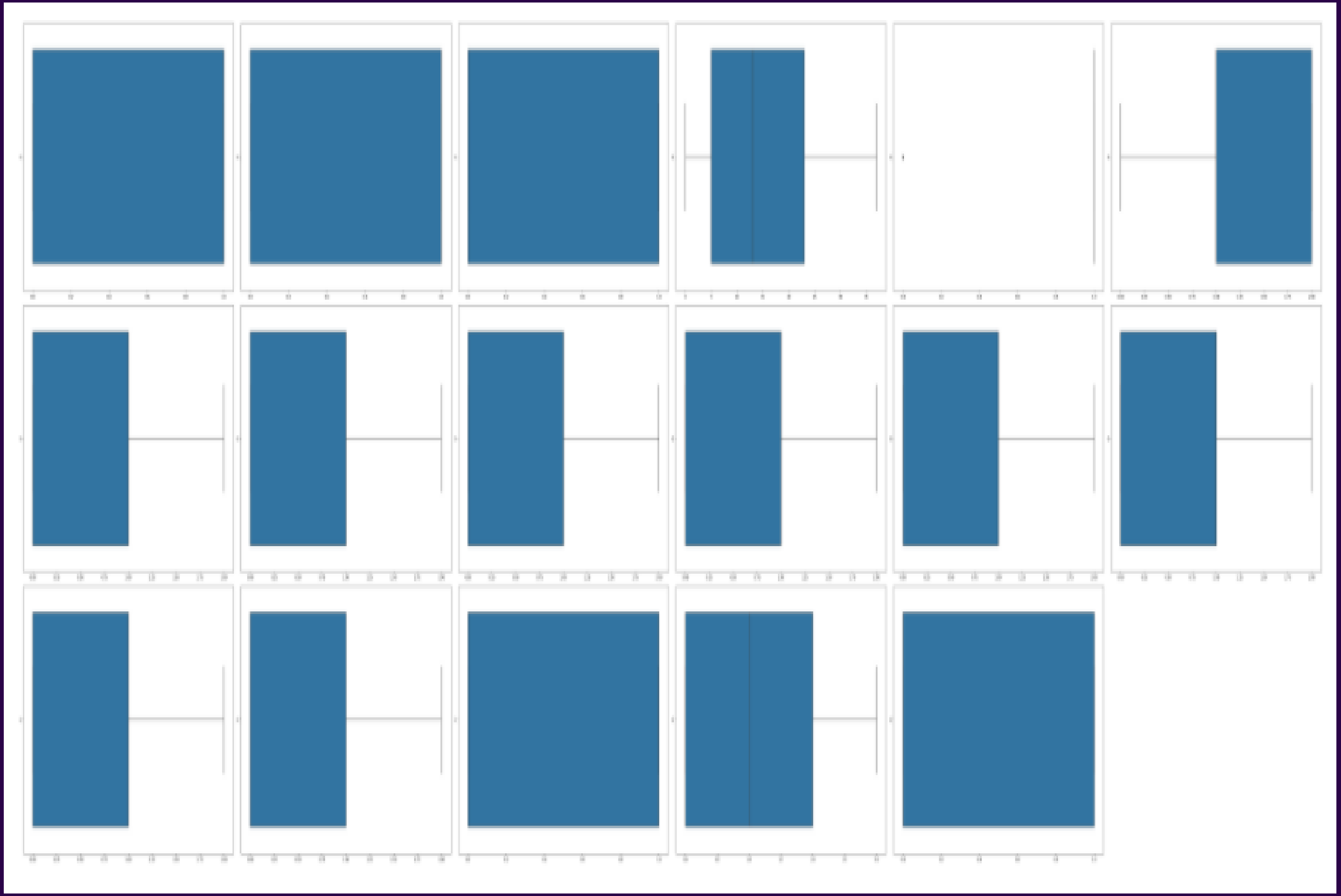


Figure 2 outliers

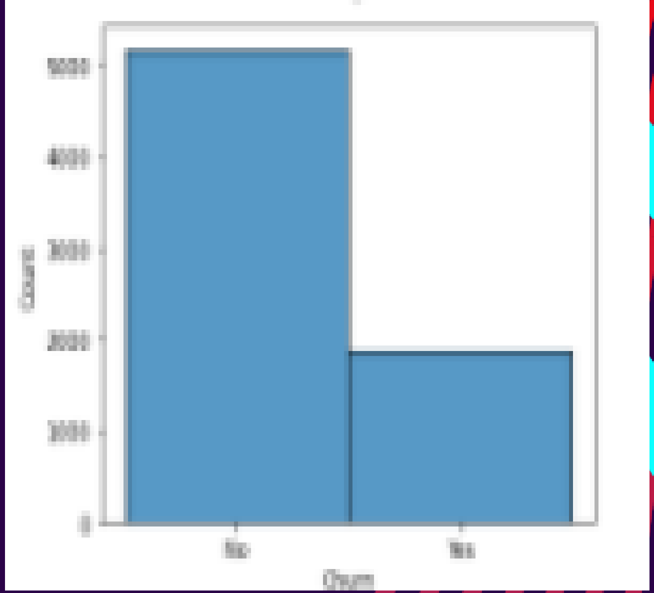
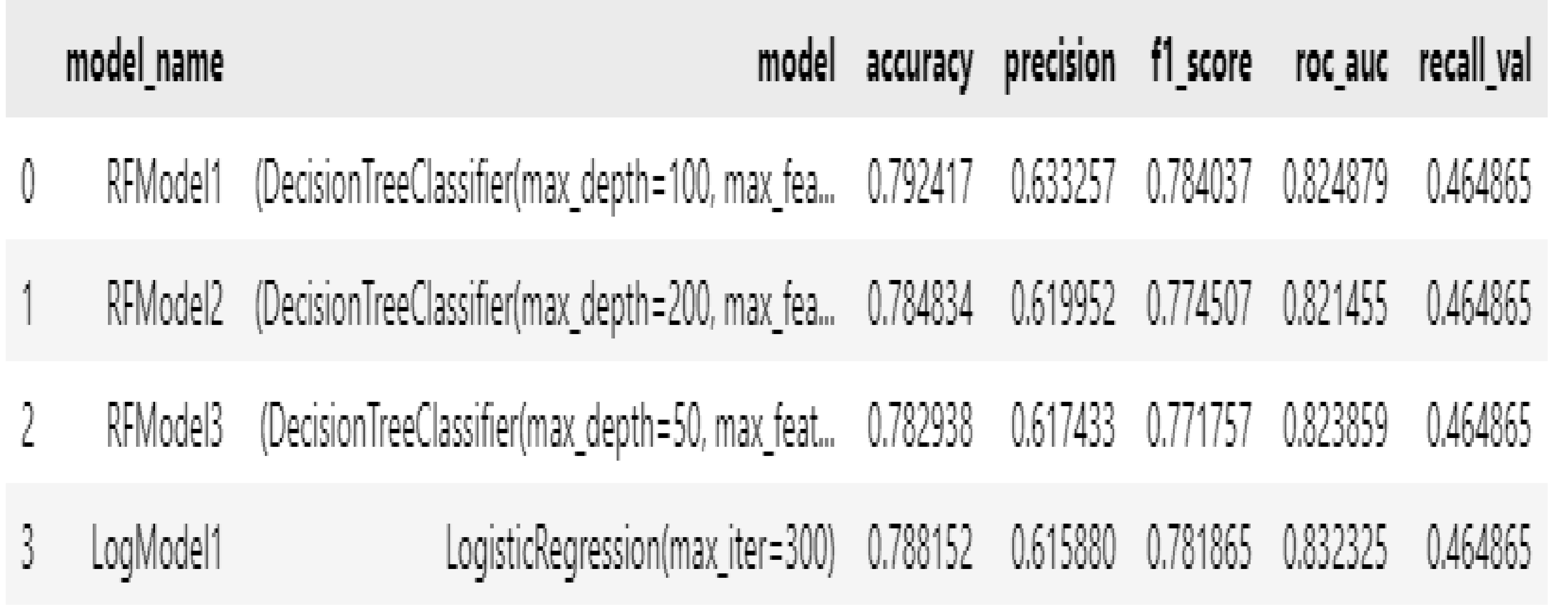


Figure 3 Target Variable Bias

MODEL EVALUATION



	model_name	model	accuracy	precision	f1_score	roc_auc	recall_val
0	RFModel1	(DecisionTreeClassifier(max_depth=100, max_fea...	0.792417	0.633257	0.784037	0.824879	0.464865
1	RFModel2	(DecisionTreeClassifier(max_depth=200, max_fea...	0.784834	0.619952	0.774507	0.821455	0.464865
2	RFModel3	(DecisionTreeClassifier(max_depth=50, max_feat...	0.782938	0.617433	0.771757	0.823859	0.464865
3	LogModel1	LogisticRegression(max_iter=300)	0.788152	0.615880	0.781865	0.832325	0.464865

Random forest algorithm-based model RFModel1 gave much better F1, recall, precision scores

THANK YOU

Kalum De Alwis

265

ML Batch 2