Telco Customer Churn

SHUCHITA MISHRA

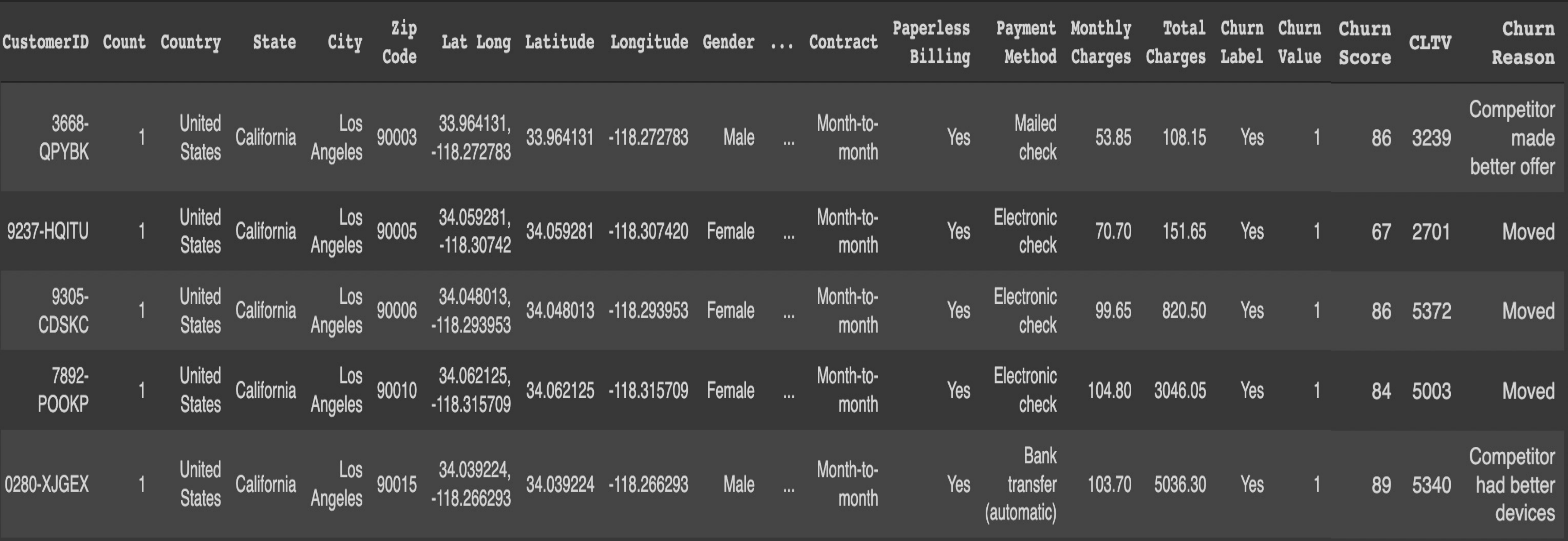
001020146

Did you know that attracting a new customer costs five times as much as keeping an existing one?

- The telecommunications business has an annual churn rate of 15-25 percent in this highly competitive market.
- Corporations and businesses can forecast which customers are likely to leave ahead of time and focus on customer retention efforts.
- As a result,
  - *preserve their market position,*
  - *grow and thrive*
  - *lower the cost of initiation*
  - *larger the profit*

DATA OVERVIEW

- Each row represents a customer, each column contains customer's attributes described on the column Metadata.
- 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*

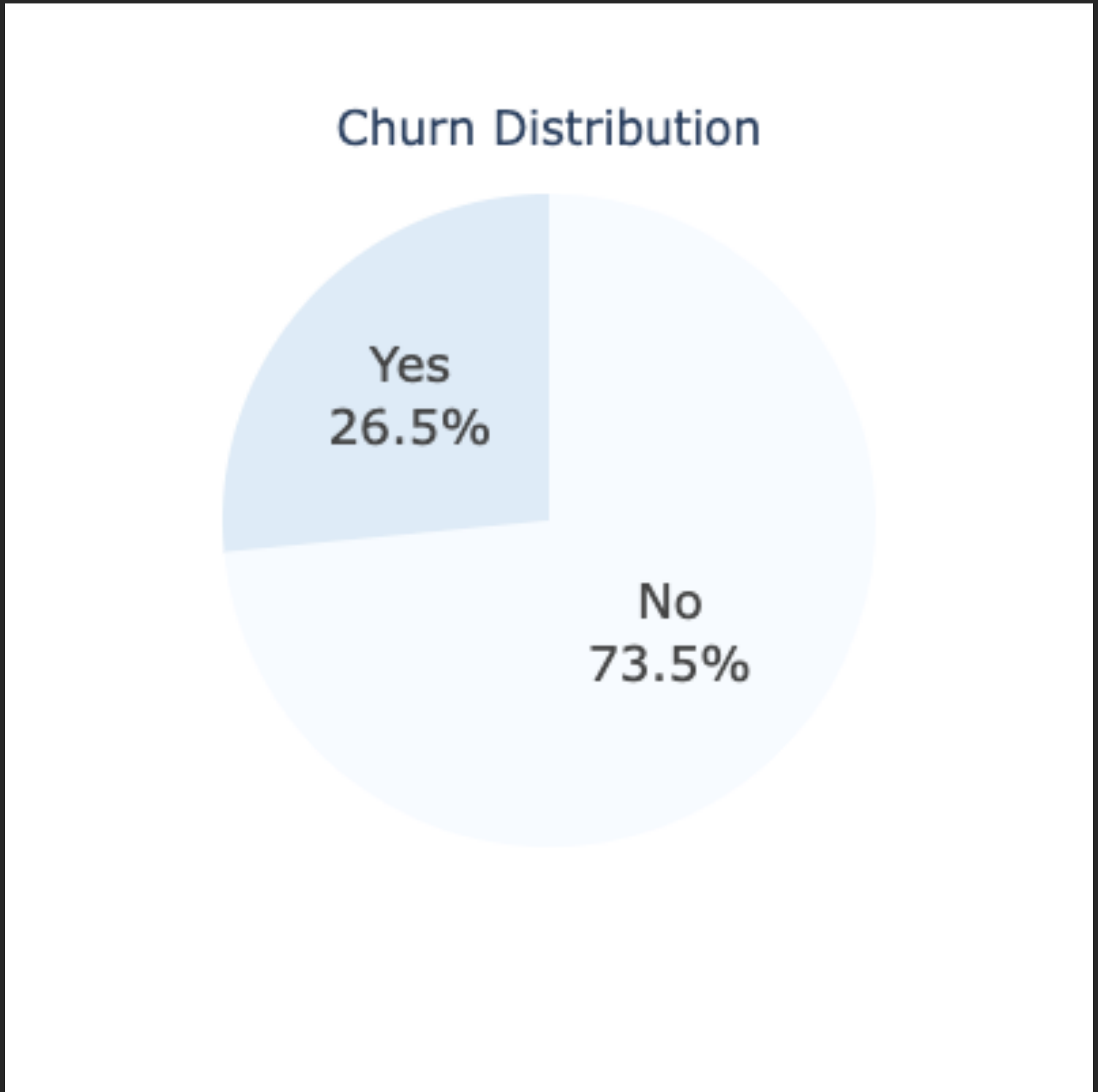


CustomerID	Count	Country	State	City	Zip Code	Lat Long	Latitude	Longitude	Gender	...	Contract	Paperless Billing	Payment Method	Monthly Charges	Total Charges	Churn Label	Churn Value	Churn Score	CLTV	Churn Reason
3668-QPYBK	1	United States	California	Los Angeles	90003	33.964131, -118.272783	33.964131	-118.272783	Male	...	Month-to-month	Yes	Mailed check	53.85	108.15	Yes	1	86	3239	Competitor made better offer
9237-HQITU	1	United States	California	Los Angeles	90005	34.059281, -118.30742	34.059281	-118.307420	Female	...	Month-to-month	Yes	Electronic check	70.70	151.65	Yes	1	67	2701	Moved
9305-CDSKC	1	United States	California	Los Angeles	90006	34.048013, -118.293953	34.048013	-118.293953	Female	...	Month-to-month	Yes	Electronic check	99.65	820.50	Yes	1	86	5372	Moved
7892-POOKP	1	United States	California	Los Angeles	90010	34.062125, -118.315709	34.062125	-118.315709	Female	...	Month-to-month	Yes	Electronic check	104.80	3046.05	Yes	1	84	5003	Moved
0280-XJGEX	1	United States	California	Los Angeles	90015	34.039224, -118.266293	34.039224	-118.266293	Male	...	Month-to-month	Yes	Bank transfer (automatic)	103.70	5036.30	Yes	1	89	5340	Competitor had better devices

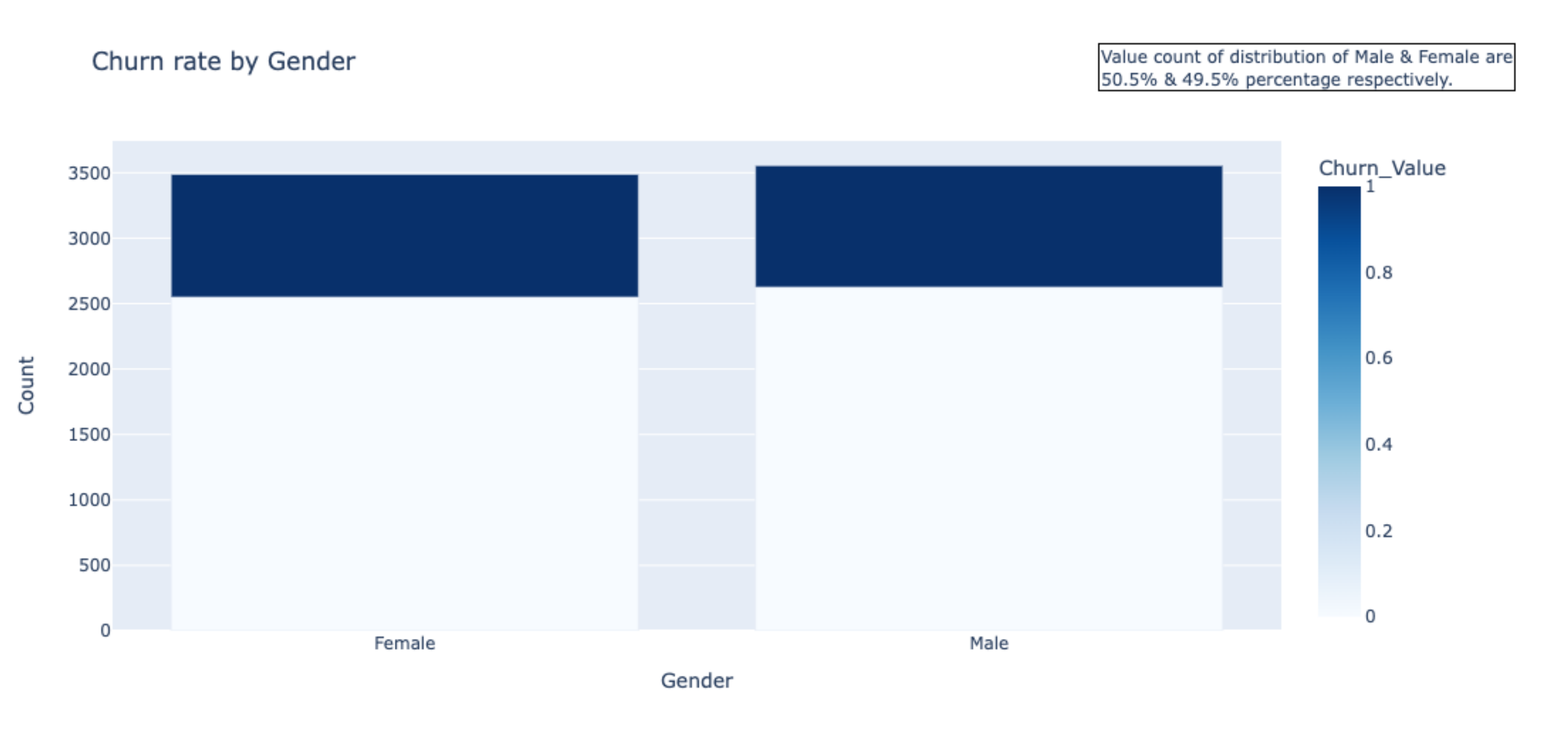
EXPLORATORY DATA ANALYSIS

LET'S EXPLORE THE DATA AND TRY TO ANSWER SOME QUESTIONS

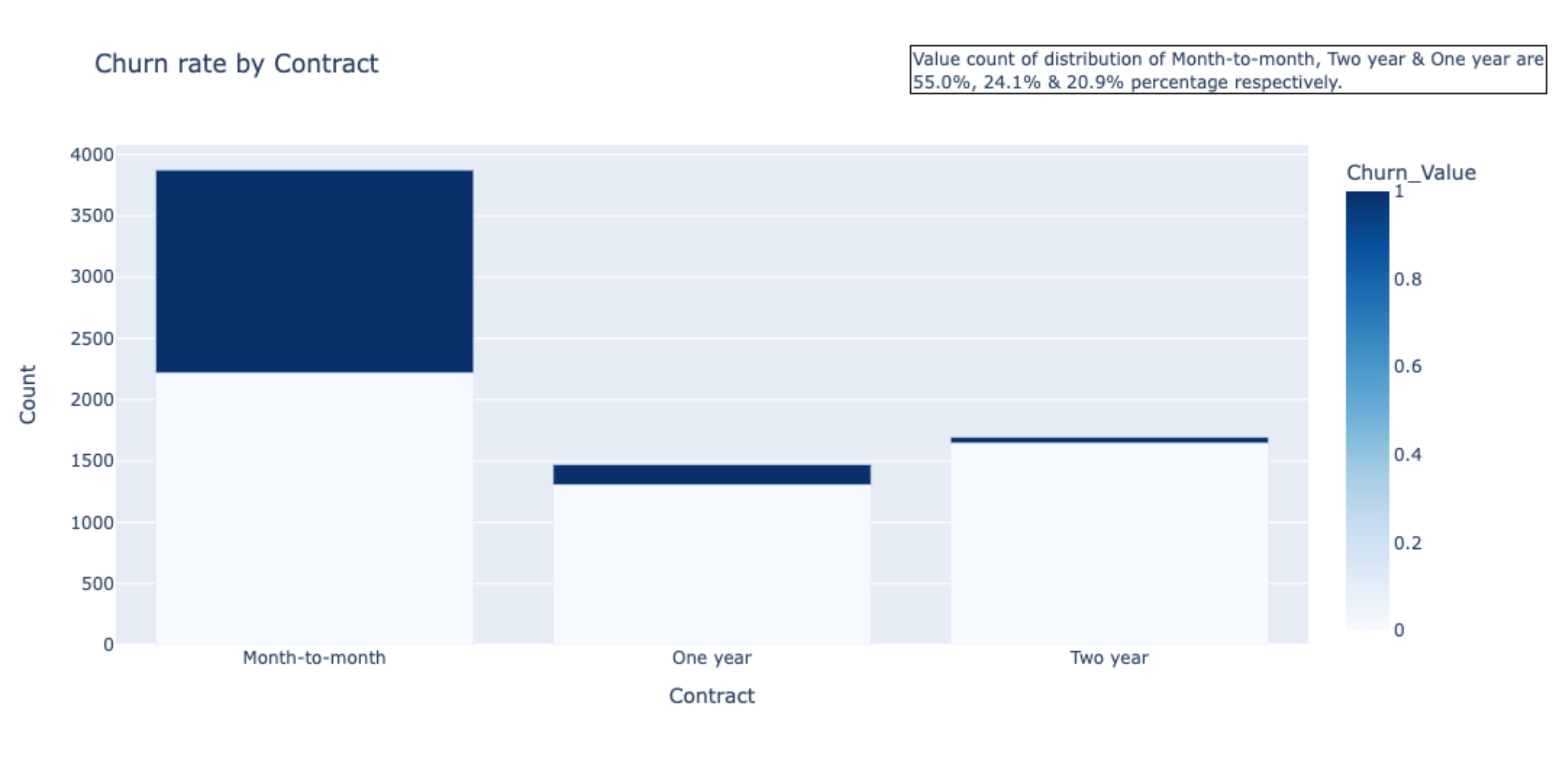
26.5 % OF CUSTOMERS SWITCHED TO ANOTHER FIRM.



BOTH GENDERS BEHAVED IN SIMILAR FASHION WHEN IT COMES TO MIGRATING TO ANOTHER SERVICE PROVIDER/FIRM



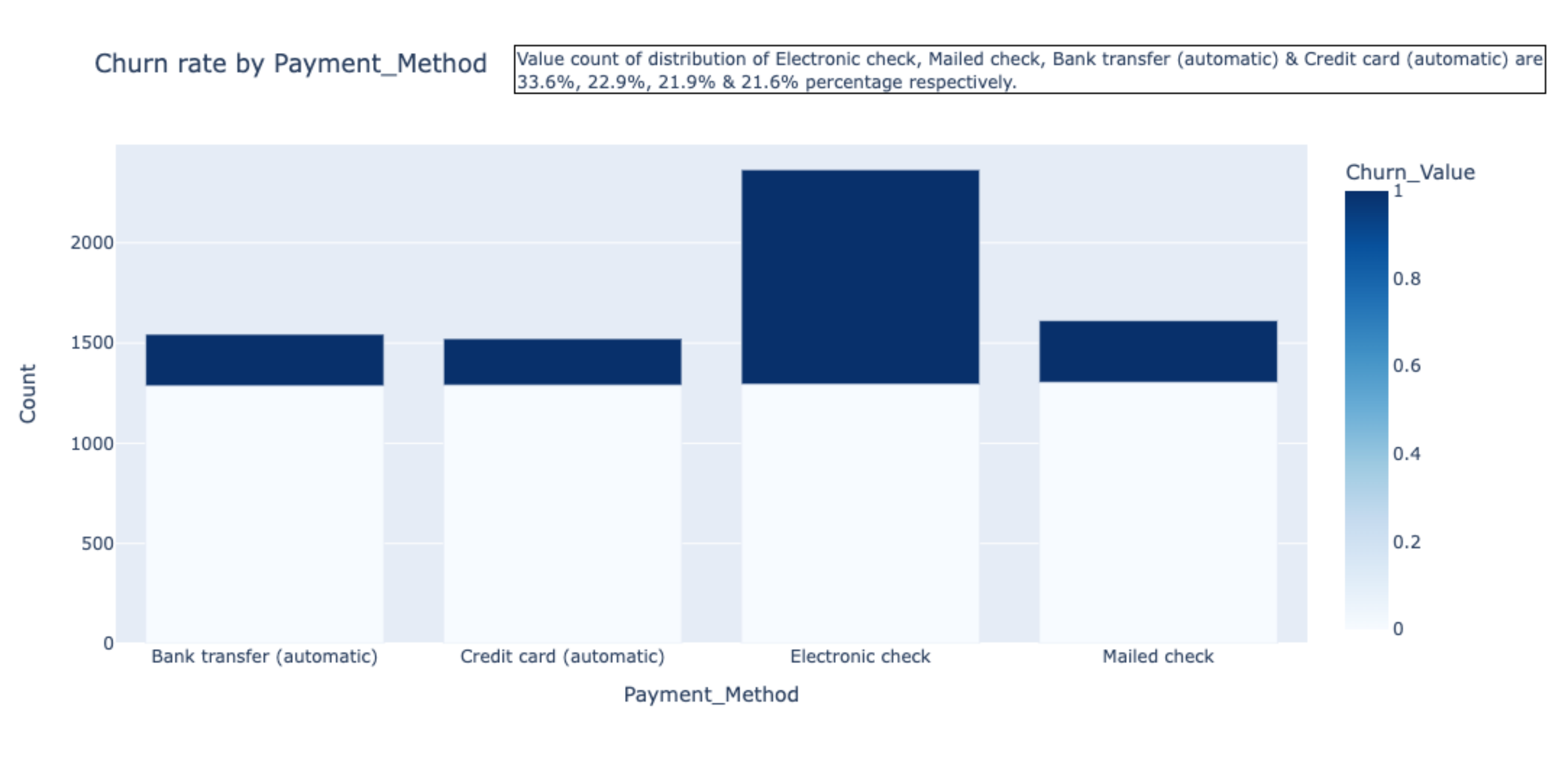
ABOUT 75% OF CUSTOMER WITH MONTH-TO-MONTH CONTRACT OPTED TO MOVE OUT AS COMPARED TO 13% OF CUSTOMERS WITH ONE YEAR CONTRACT AND 3% WITH TWO YEAR CONTRACT


Churn rate by Contract

Value count of distribution of Month-to-month, Two year & One year are 55.0%, 24.1% & 20.9% percentage respectively.

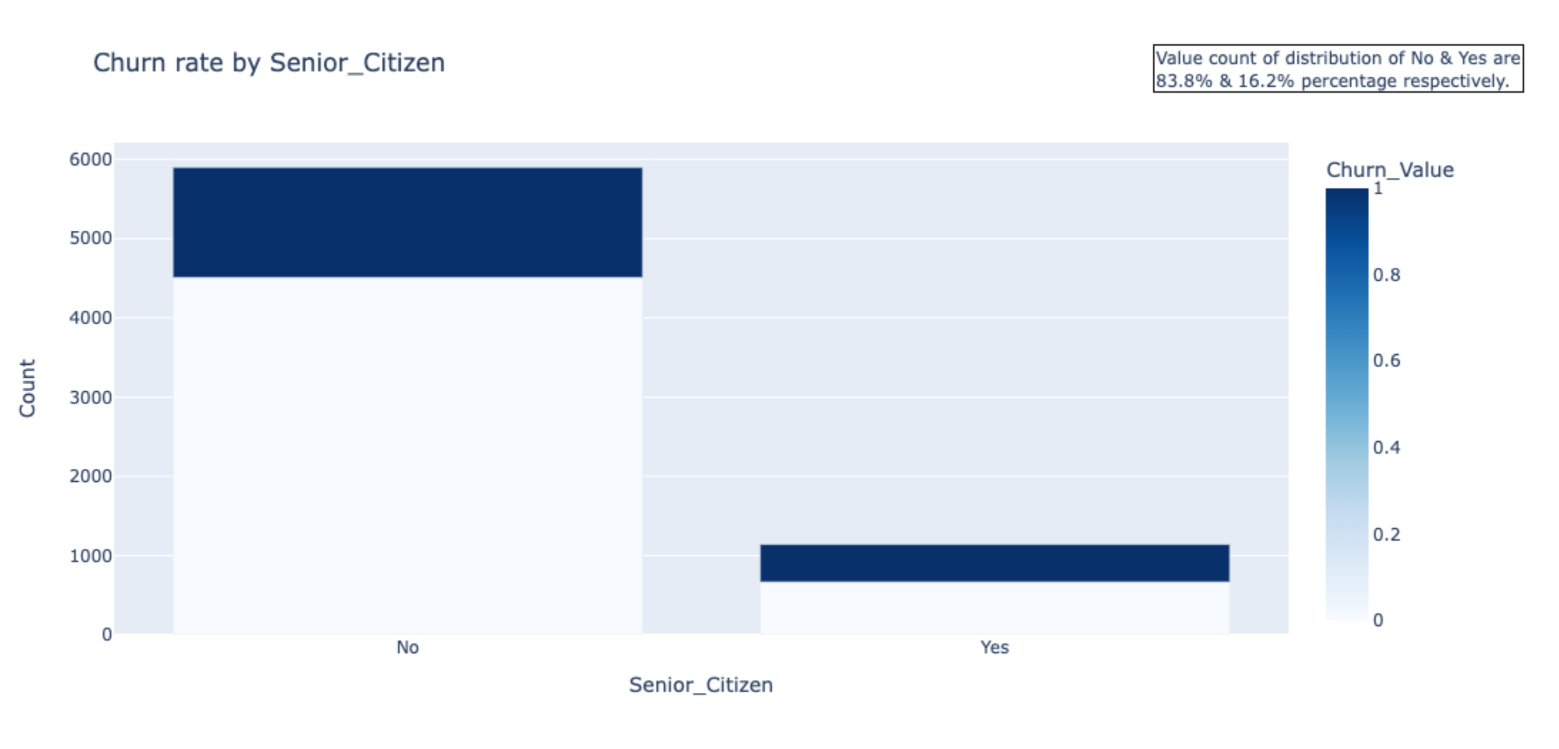
MAJOR CUSTOMERS WHO MOVED OUT HAD AN ELECTRONIC CHECK AS PAYMENT METHOD ON FILE

CUSTOMERS WHO OPTED FOR CREDIT-CARD AUTOMATIC TRANSFER OR BANK AUTOMATIC TRANSFER AND MAILED CHECK AS PAYMENT METHOD WERE LESS LIKELY TO MOVE OUT

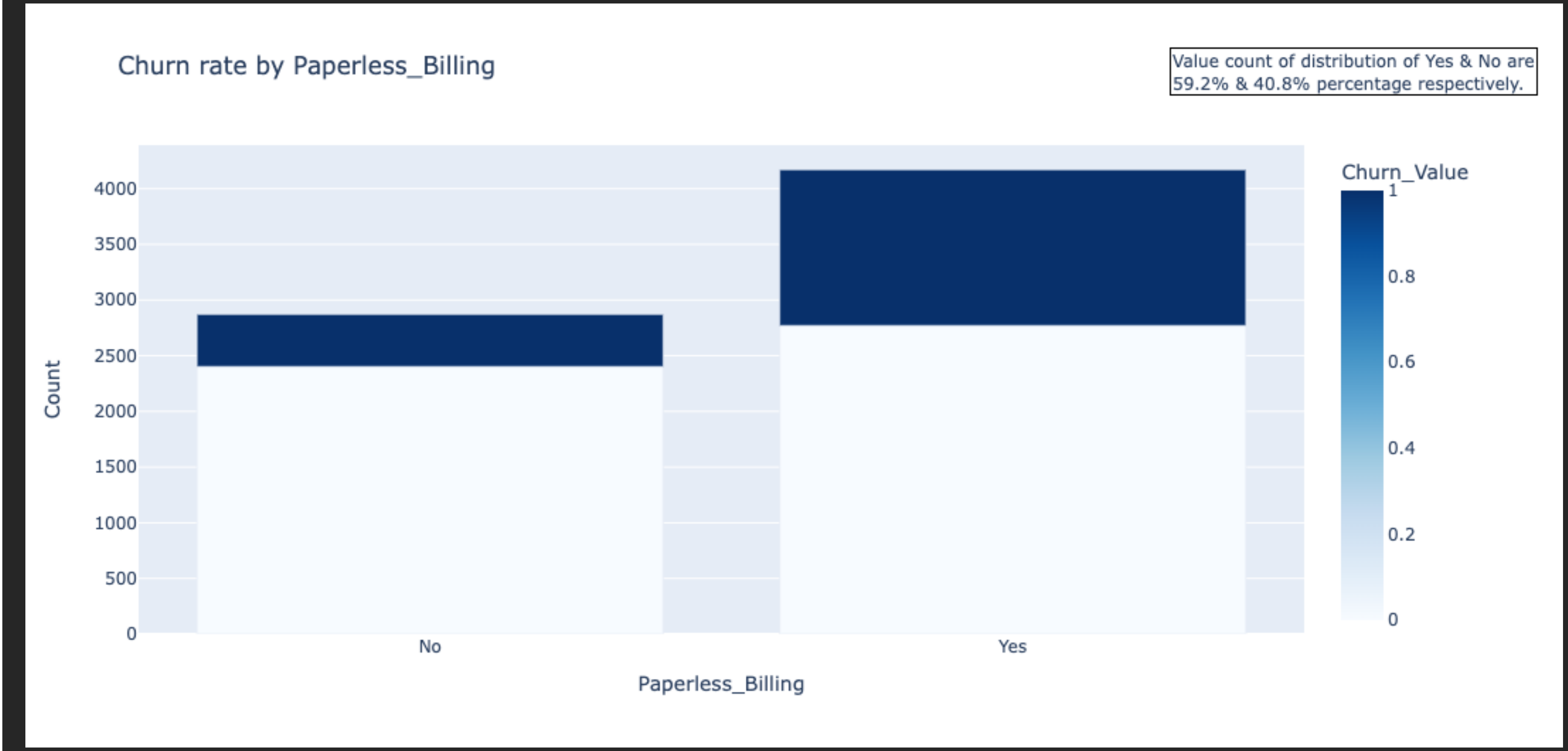


Value count of distribution of Electronic check, Mailed check, Bank transfer (automatic) & Credit card (automatic) are 33.6%, 22.9%, 21.9% & 21.6% percentage respectively.

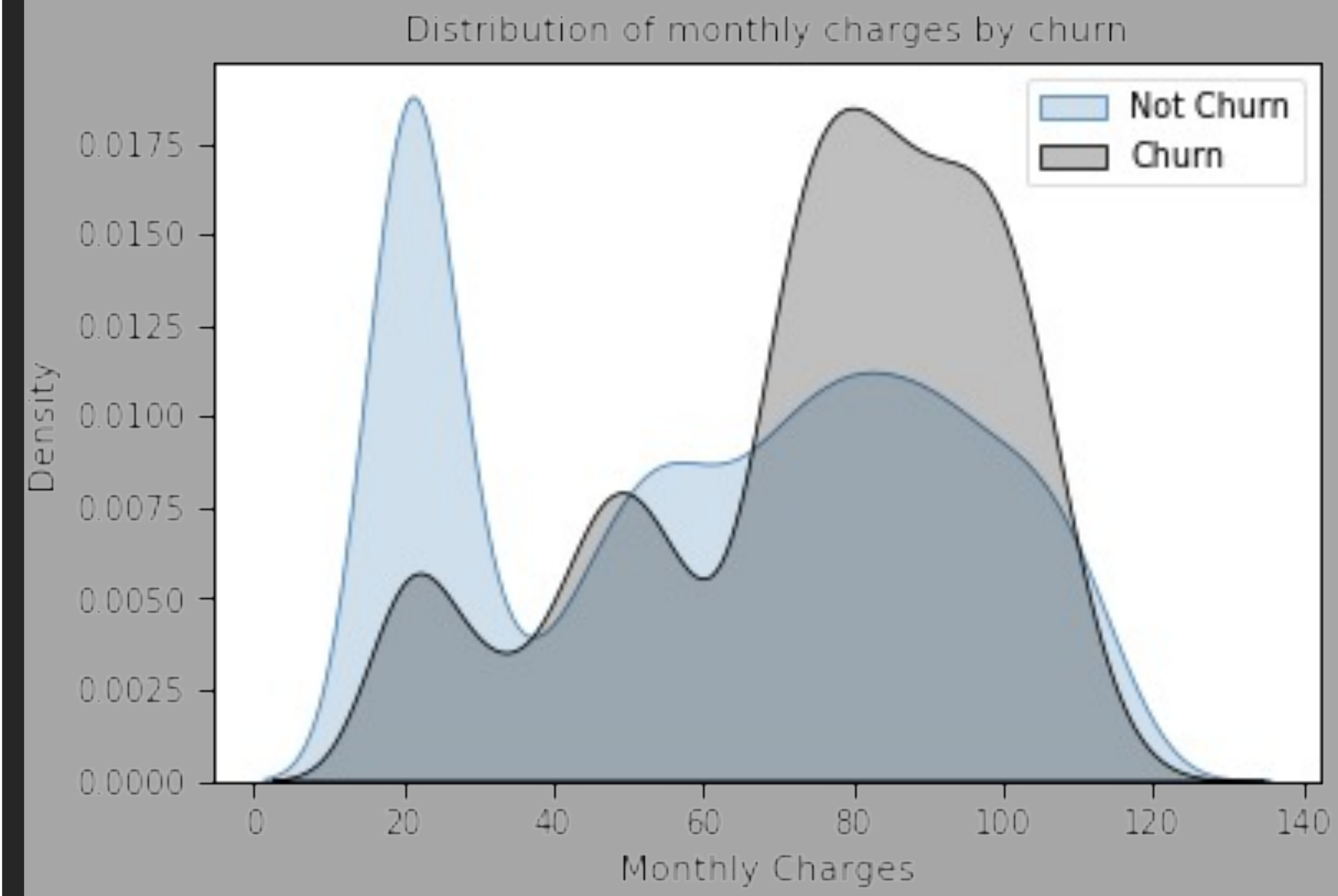
MOST OF THE SENIOR CITIZENS CHURN



CUSTOMERS WITH PAPERLESS BILLING ARE MOST LIKELY TO CHURN



CUSTOMERS WITH HIGHER MONTHLY CHARGES ARE ALSO MORE LIKELY TO CHURN



DATA PRE-PROCESSING AND CLEANING

Standard scalar to scale numerical columns down to the same range

Splitting the data into train and test sets

Manually categorizing the data in 0,1 form

One hot encoding the total charges column

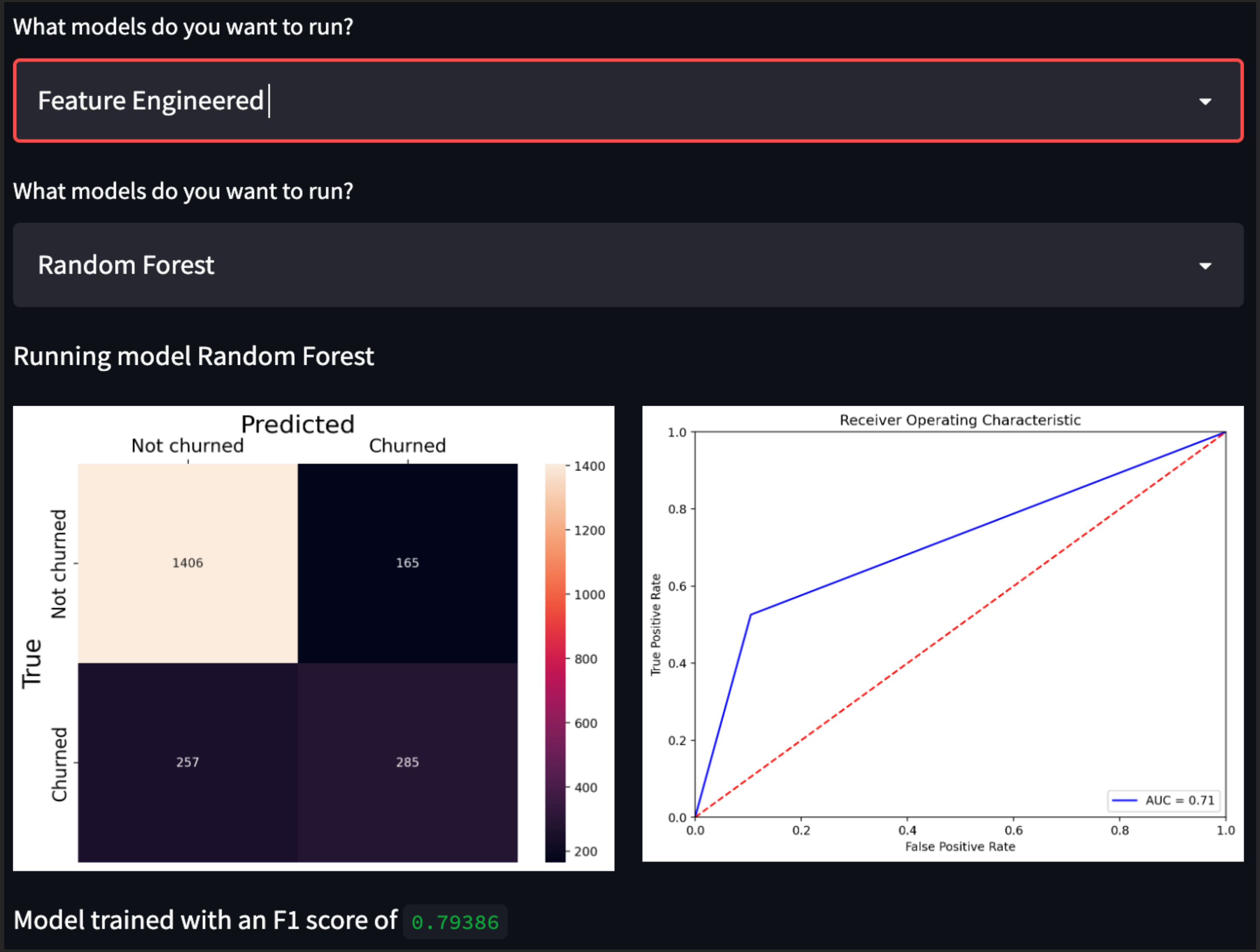
Label encoding

Dropping the redundant columns such as country, state, count, latitude, longitude

ML MODEL EVALUATIONS AND PREDICTING

NOW THAT OUR DATA IS PROCESSED AND CLEANED, LET'S START PREDICTING THE CHURN STATUS

RANDOM FOREST CLASSIFIER GIVES BEST PREDICTION ON RAW UNSCALED DATA WITH F1 SCORE OF 79%



What models do you want to run?

Feature Engineered

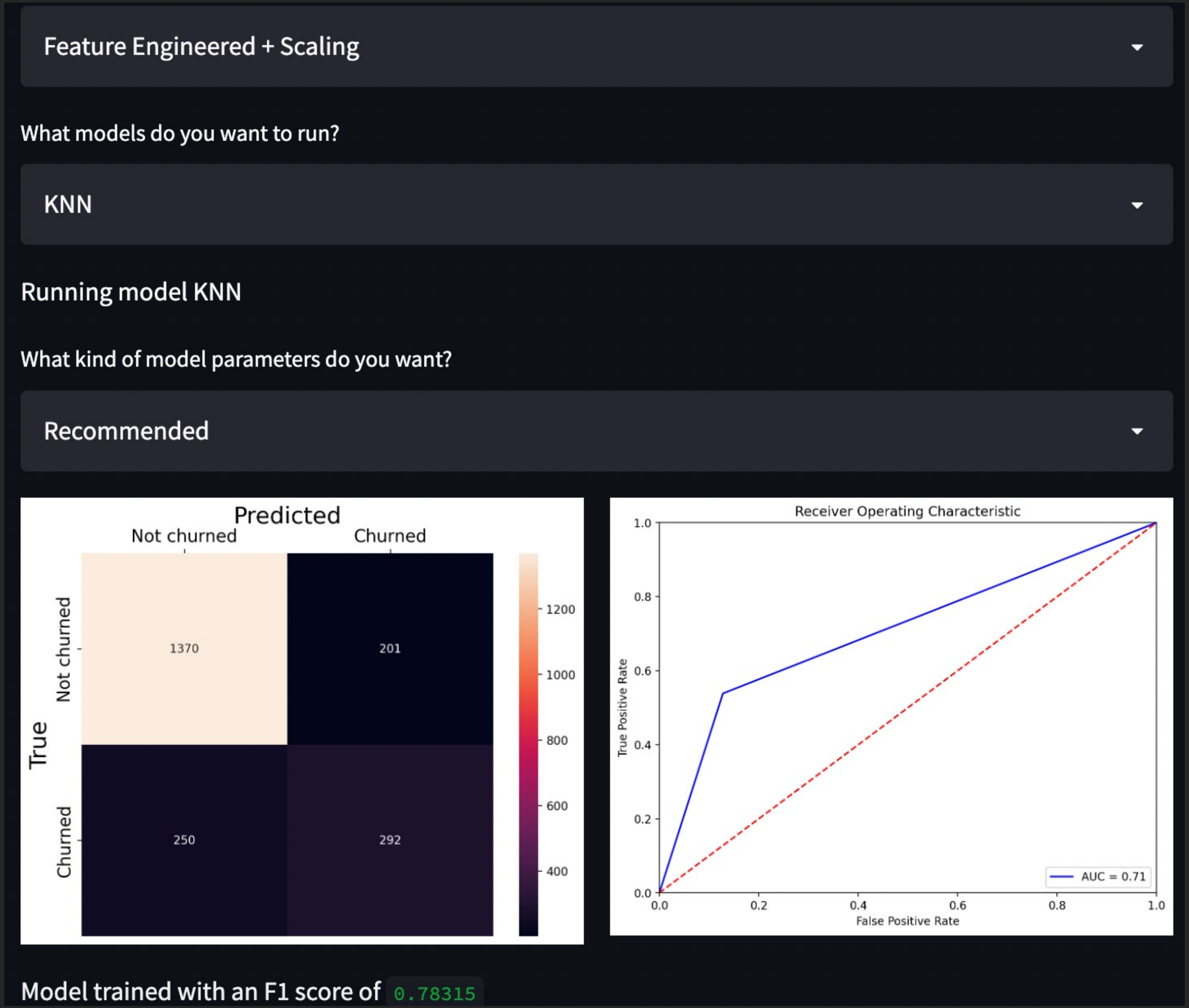
What models do you want to run?

Random Forest

Running model Random Forest

Model trained with an F1 score of `0.79386`

KNN CLASSIFIER GIVES A 78% F1 SCORE WITH SCALED DATA



Feature Engineered + Scaling

What models do you want to run?

KNN

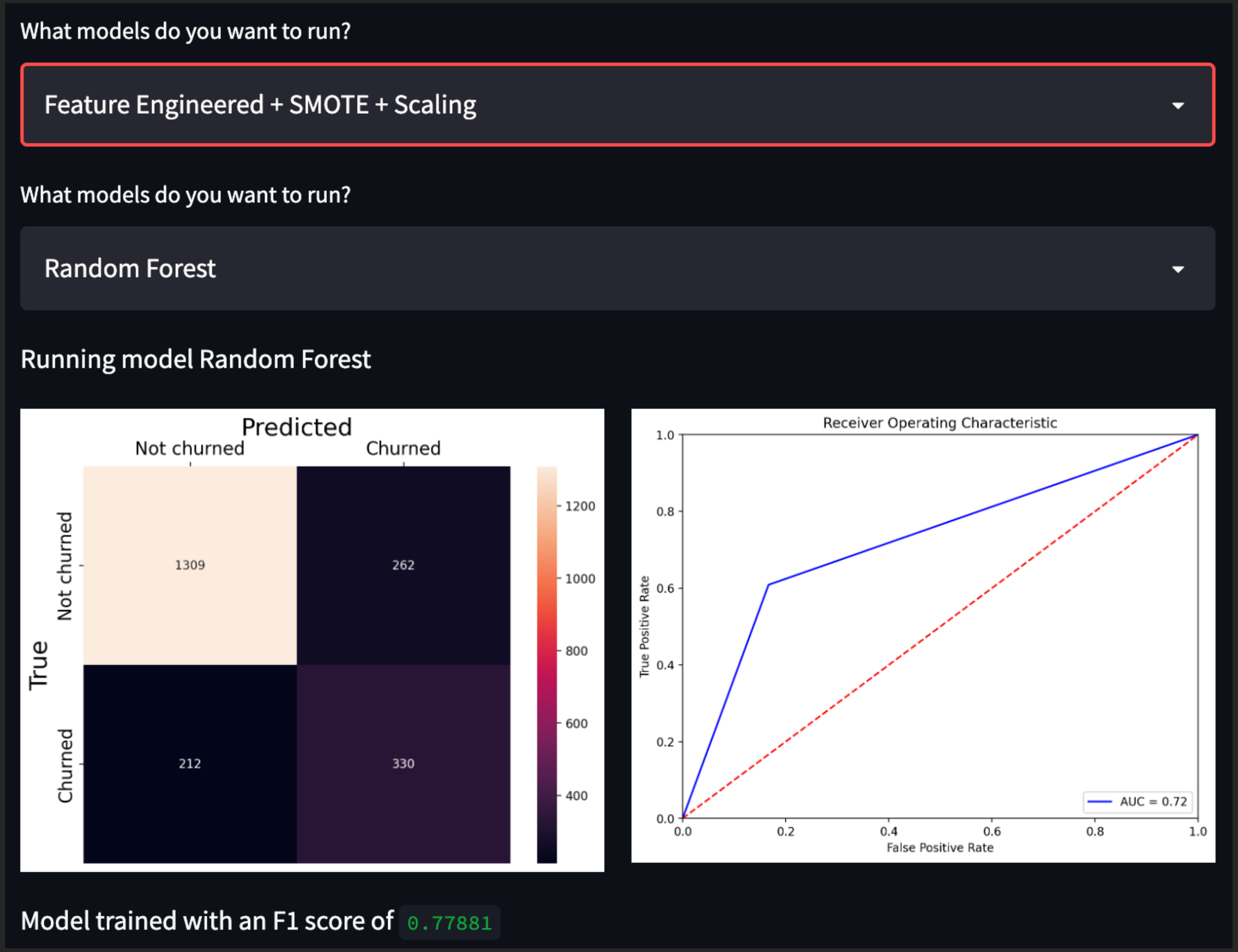
Running model KNN

What kind of model parameters do you want?

Recommended

Model trained with an F1 score of `0.78315`

RANDOM FOREST CLASSIFIER GIVES 77% F1 SCORE WITH BALANCED, SCALED DATA



What models do you want to run?

Feature Engineered + SMOTE + Scaling

What models do you want to run?

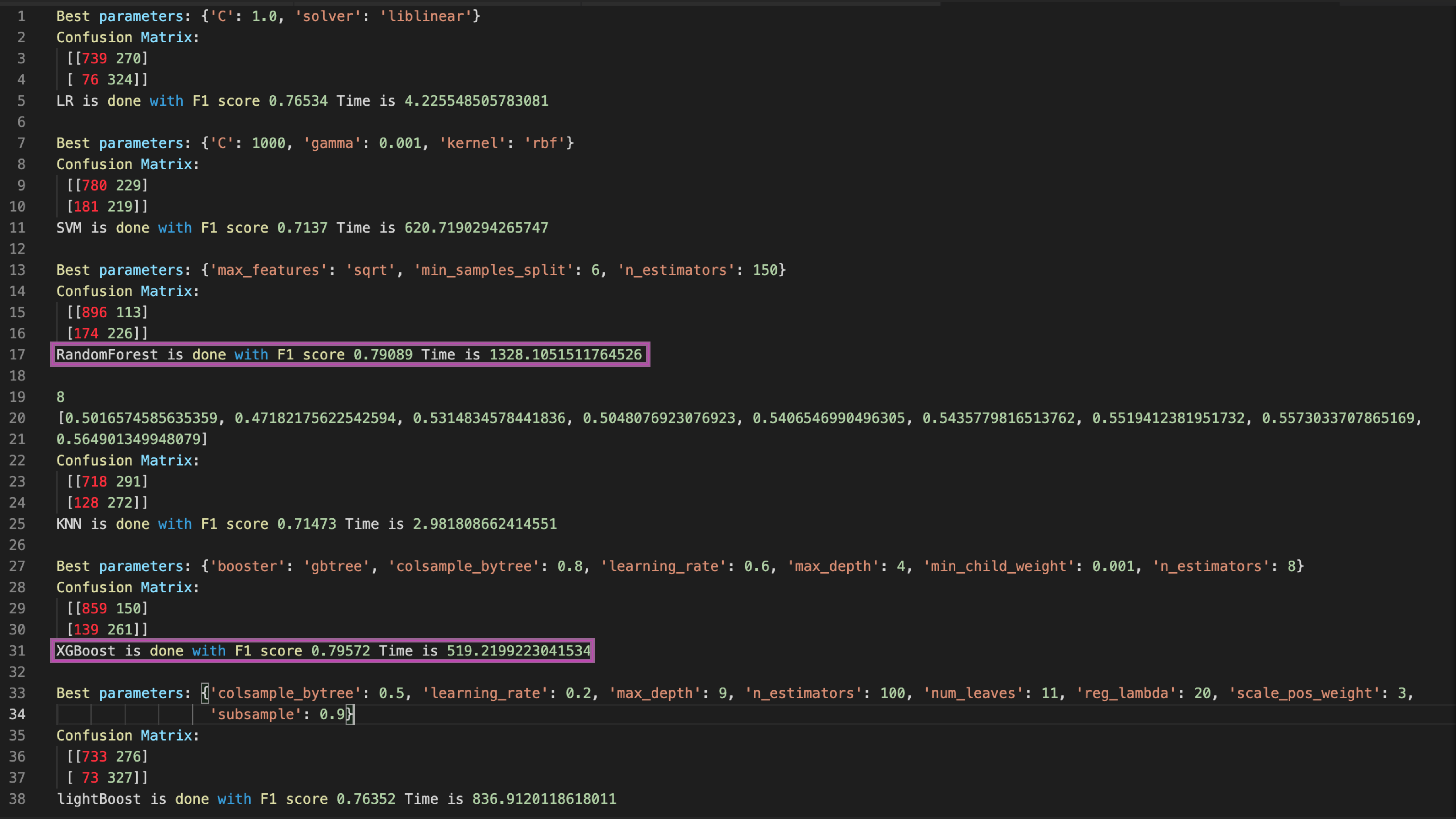
Random Forest

Running model Random Forest

Model trained with an F1 score of `0.77881`

COMPARATIVE EVALUATIONS OF ALL ML MODELS USED

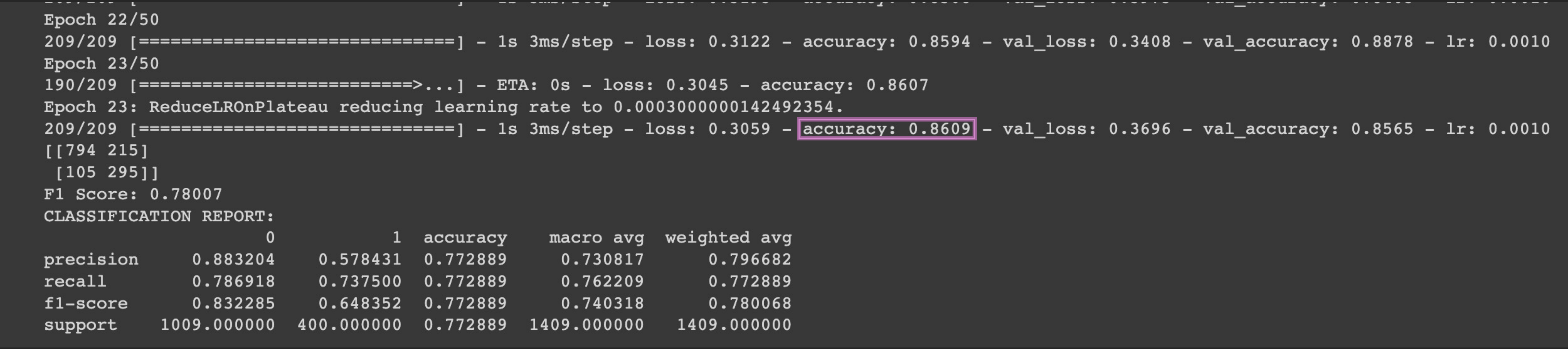
Logistic regression, SVM Classifier, Random Forest, KNN, XGBoost Classifier, LightGBM Classifier tuned with best/recommended parameters using cross-validation



```
Best parameters: {'C': 1.0, 'solver': 'liblinear'}
Confusion Matrix:
 [[739 270]
 [ 76 324]]
LR is done with F1 score 0.76534 Time is 4.225548505783081

Best parameters: {'C': 1000, 'gamma': 0.001, 'kernel': 'rbf'}
Confusion Matrix:
 [[780 229]
 [181 219]]
SVM is done with F1 score 0.7137 Time is 620.7190294265747

Best parameters: {'max_features': 'sqrt', 'min_samples_split': 6, 'n_estimators': 150}
Confusion Matrix:
 [[896 113]
 [174 226]]
RandomForest is done with F1 score 0.79089 Time is 1328.1051511764526

8
[0.5016574585635359, 0.47182175622542594, 0.5314834578441836, 0.5048076923076923, 0.5406546990496305, 0.5435779816513762, 0.5519412381951732, 0.5573033707865169,
0.564901349948079]
Confusion Matrix:
 [[718 291]
 [128 272]]
KNN is done with F1 score 0.71473 Time is 2.981808662414551

Best parameters: {'booster': 'gbtree', 'colsample_bytree': 0.8, 'learning_rate': 0.6, 'max_depth': 4, 'min_child_weight': 0.001, 'n_estimators': 8}
Confusion Matrix:
 [[859 150]
 [139 261]]
XGBoost is done with F1 score 0.79572 Time is 519.2199223041534

Best parameters: {'colsample_bytree': 0.5, 'learning_rate': 0.2, 'max_depth': 9, 'n_estimators': 100, 'num_leaves': 11, 'reg_lambda': 20, 'scale_pos_weight': 3,
                  'subsample': 0.9}
Confusion Matrix:
 [[733 276]
 [ 73 327]]
lightBoost is done with F1 score 0.76352 Time is 836.9120118618011
```

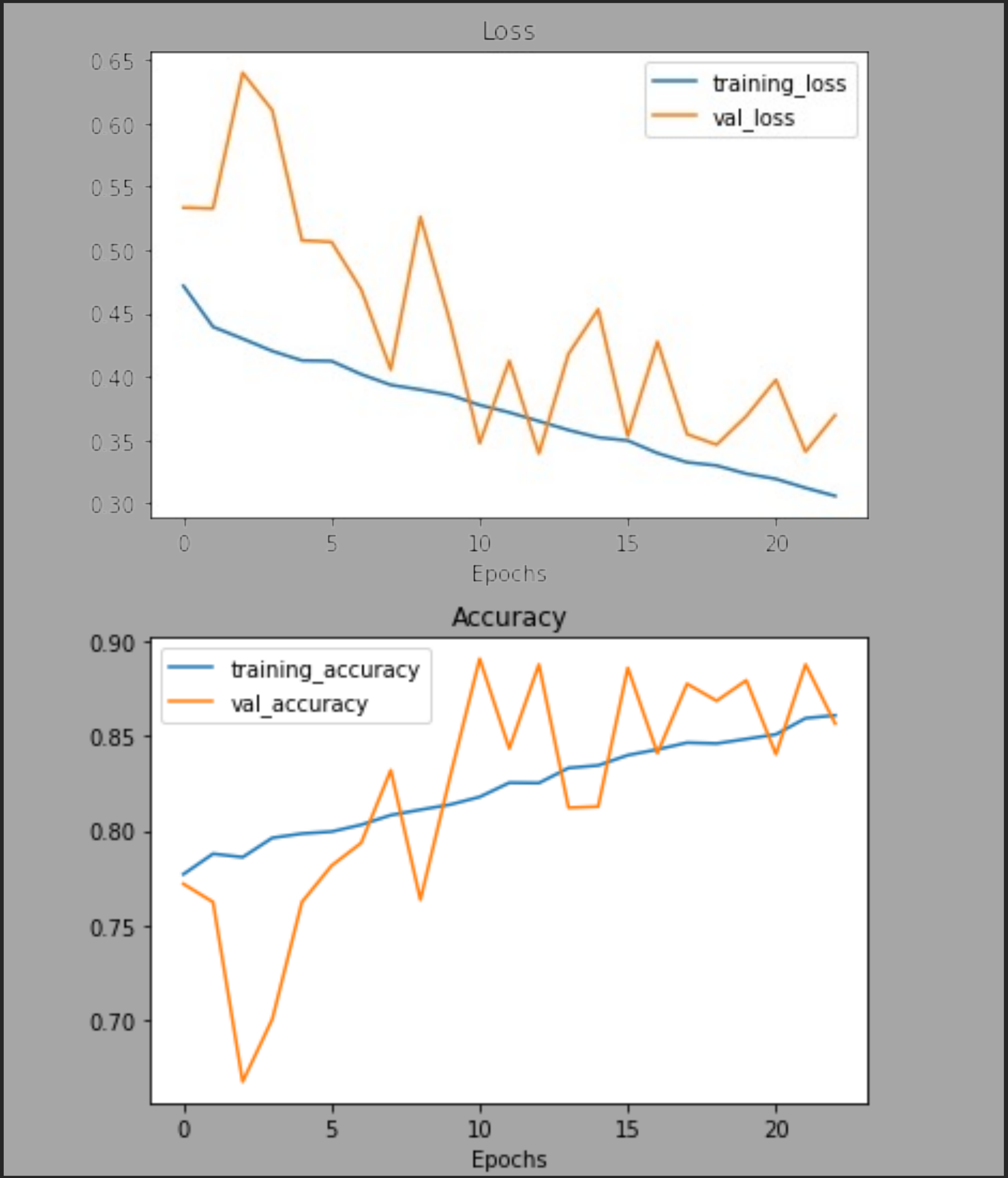
NEURAL NETWORKS GIVES THE BEST ACCURACY SCORE OF 86%

Last few optimal epochs and classification report

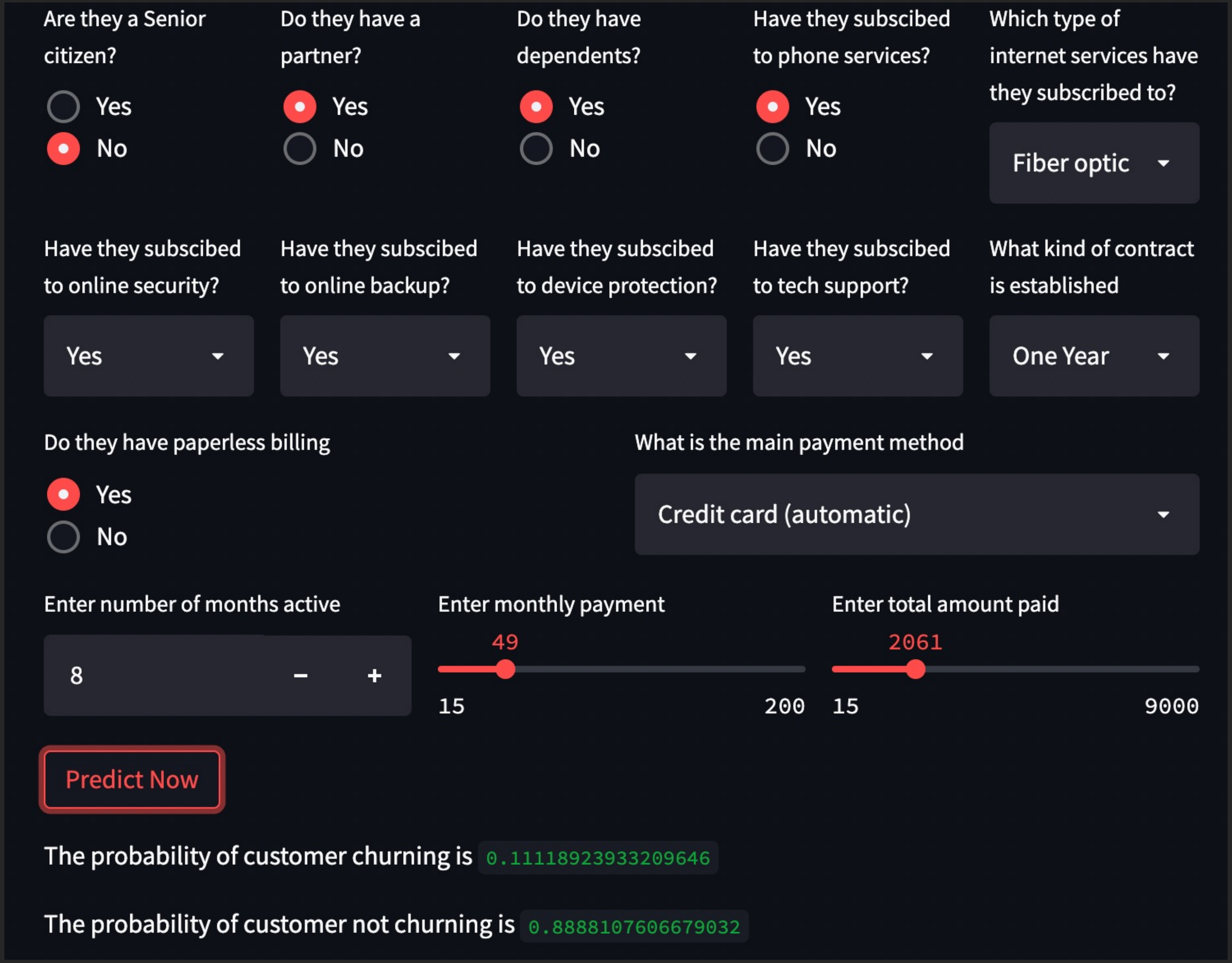


```
Epoch 22/50
209/209 [==============================] - 1s 3ms/step - loss: 0.3122 - accuracy: 0.8594 - val_loss: 0.3408 - val_accuracy: 0.8878 - lr: 0.0010
Epoch 23/50
190/209 [==========================>...] - ETA: 0s - loss: 0.3045 - accuracy: 0.8607
Epoch 23: ReduceLROnPlateau reducing learning rate to 0.0003000000142492354.
209/209 [==============================] - 1s 3ms/step - loss: 0.3059 - accuracy: 0.8609 - val_loss: 0.3696 - val_accuracy: 0.8565 - lr: 0.0010
[[794 215]
 [105 295]]
F1 Score: 0.78007
CLASSIFICATION REPORT:
                     0           1  accuracy    macro avg  weighted avg
precision     0.883204    0.578431  0.772889     0.730817      0.796682
recall        0.786918    0.737500  0.772889     0.762209      0.772889
f1-score      0.832285    0.648352  0.772889     0.740318      0.780068
support    1009.000000  400.000000  0.772889  1409.000000   1409.000000
```

LOSS CURVES FOR TRAINING AND VALIDATION METRICS



DEPLOYMENT USING STREAMLIT



Are they a Senior citizen?
○ Yes
● No

Do they have a partner?
● Yes
○ No

Do they have dependents?
● Yes
○ No

Have they subscibed to phone services?
● Yes
○ No

Which type of internet services have they subscribed to?
Fiber optic

Have they subscibed to online security?
Yes

Have they subscibed to online backup?
Yes

Have they subscibed to device protection?
Yes

Have they subscibed to tech support?
Yes

What kind of contract is established
One Year

Do they have paperless billing
● Yes
○ No

What is the main payment method
Credit card (automatic)

Enter number of months active
8

Enter monthly payment
49

Enter total amount paid
2061

The probability of customer churning is `0.11118923933209646`

The probability of customer not churning is `0.8888107606679032`

CONCLUSION

- The best way to avoid customer churn is to identify customers who are at risk of churning and working to improve their satisfaction.
- The most **important features** that helped this models are **”Tenure” which had the biggest effect and then “TechSupport” and “TotalCharges”**
- Based on my project and results, Random Forests and Neural Network models predict the probability of “high risk” customers very effectively.
- I decided to use **ROC AUC** as the evaluation metric
  - ➢ *suitable to classification problems*
  - ➢ ***robust to imbalance** of the target classes compared to accuracy*
- The **confusion matrix** was used to check if I am avoiding both **type I error and type II errors.**

THANK YOU