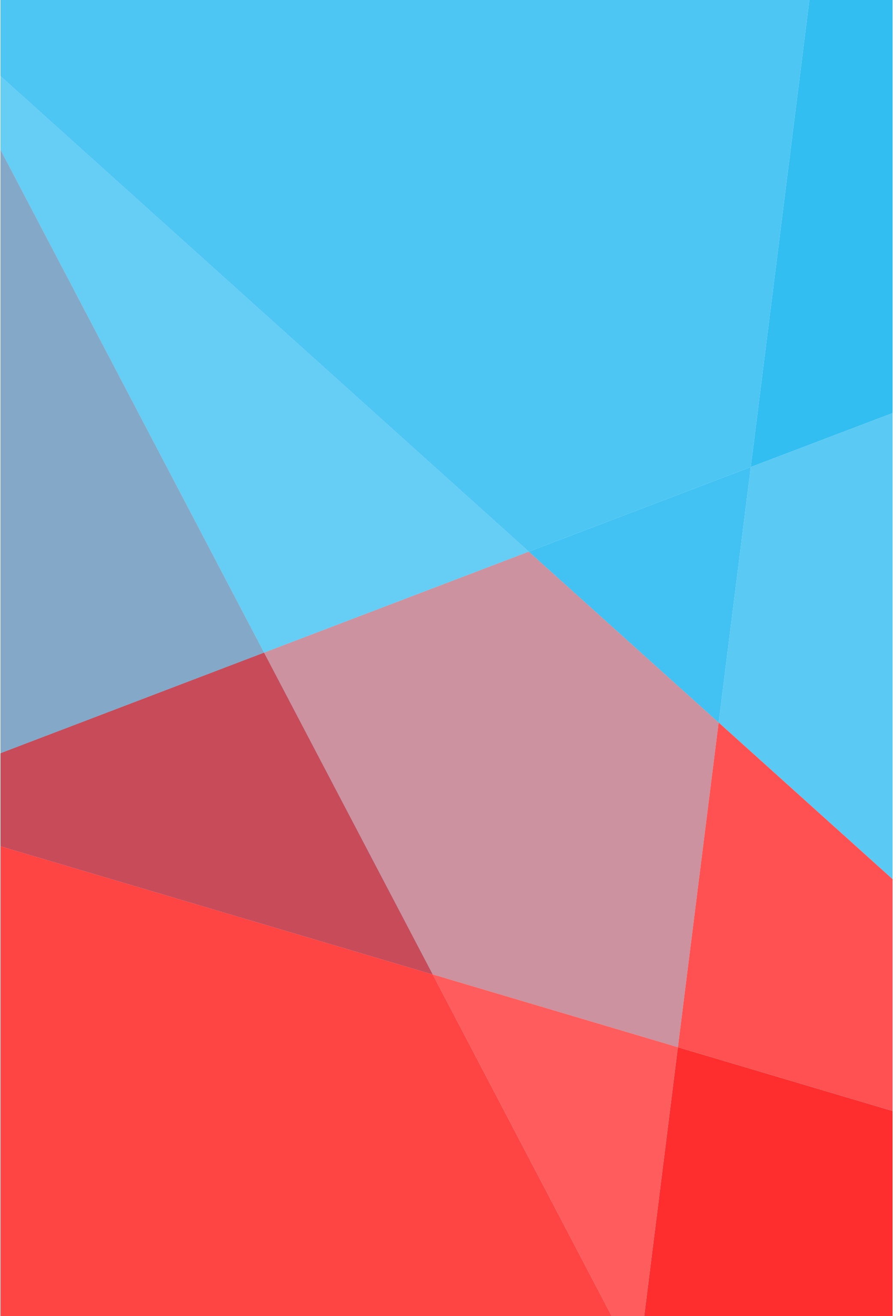


TELCO CUSTOMER CHURN PREDICTION

BY RAKA R.A PRASETYO

BACKGROUND

Churn quantifies the number of customers who have left a brand by cancelling their subscription or stopping paying for the services. This is bad news for any business as it costs five times as much to attract a new customer as it does to keep an existing one. A high customer churn rate will hit a company's finances hard.

https://www.invespcro.com/blog/customer-acquisition-retention/

OBJECTIVE

- Utilize machine learning for predicting customer churn
- Compare and choose the best machine learning model for this case
- Model Evaluation using Receiver Operating Characteristic

DATA INTRODUCTION

Source : https://www.kaggle.com/yeanzc/telco-customer-churn-ibm-dataset

Data related to a customer who left

churn_label, churn_scores, churn_value, churn reason

Data related to a customer subscribed service

phone_service, multiple line, internet_service, online_security, online_backup, device protection, tech_support, streaming_tv, streaming_movies

Data related to a customer demographic info

country, state, city, zip_code, lat_long, latitude, longitude, gender, senior_citizen, partner, dependents

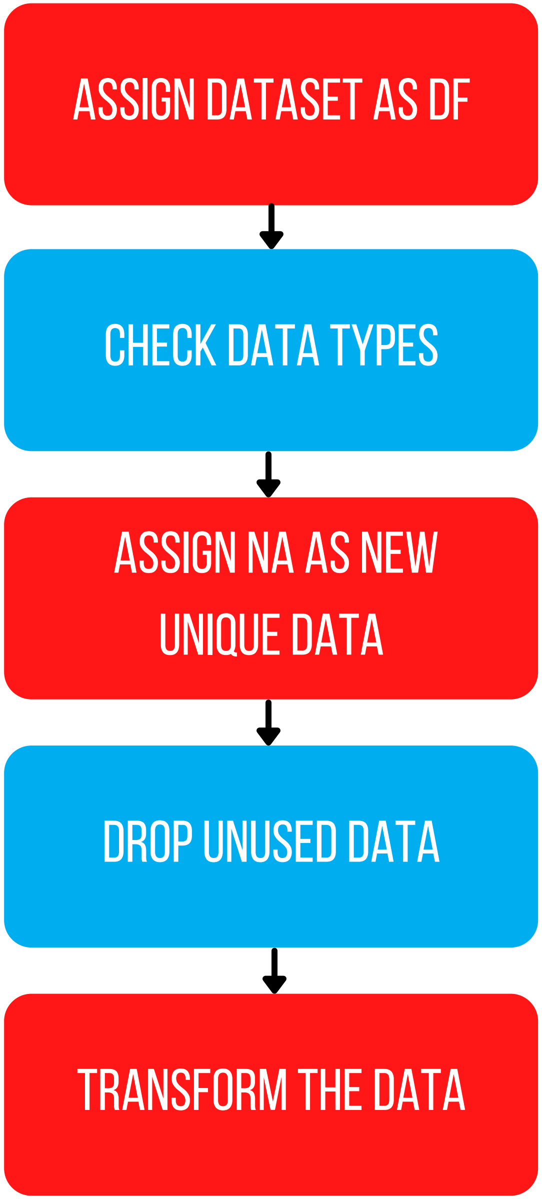
Data related to a customer account information

tenure_month, contract, paperless_billing, payment_method, monthly_charges, total_charges

Data related to IBM company data

customer_id, count, cltv

DATA CLEANSING



ASSIGN DATASET AS DF

CHECK DATA TYPES

ASSIGN NA AS NEW UNIQUE DATA

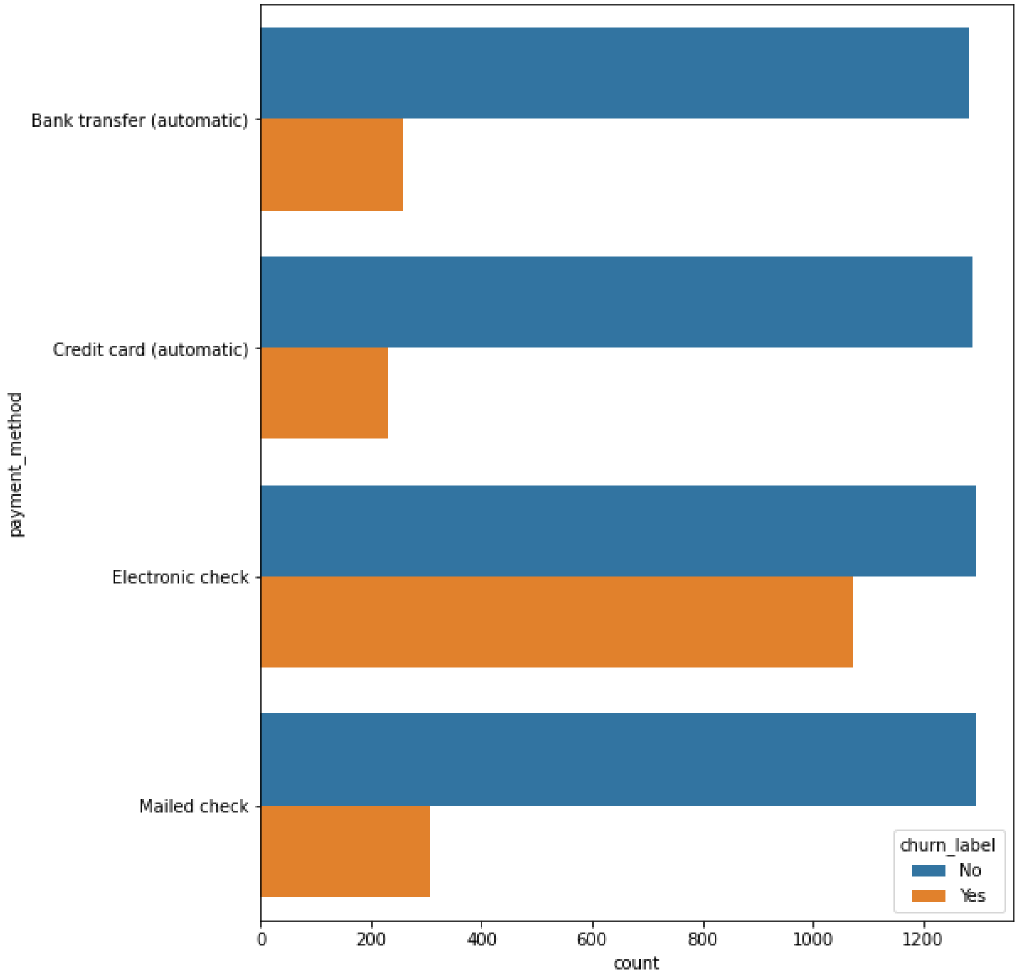
DROP UNUSED DATA

TRANSFORM THE DATA

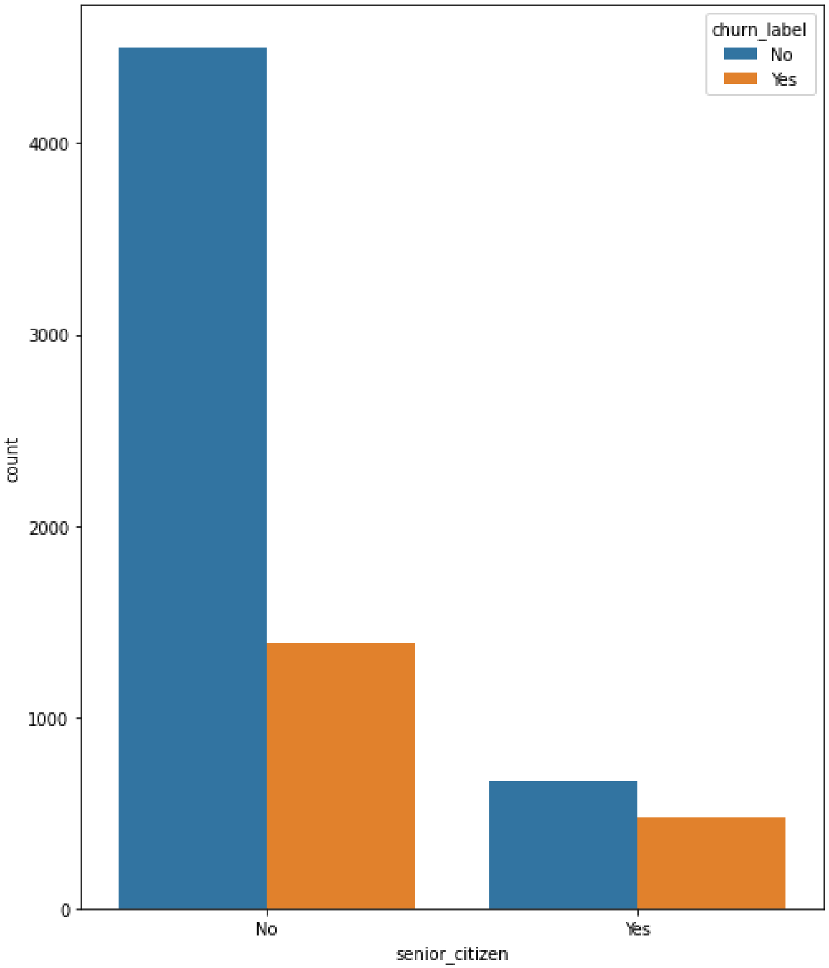
EXPLORATORY DATA ANALYSIS

CATEGORICAL VARIABLE

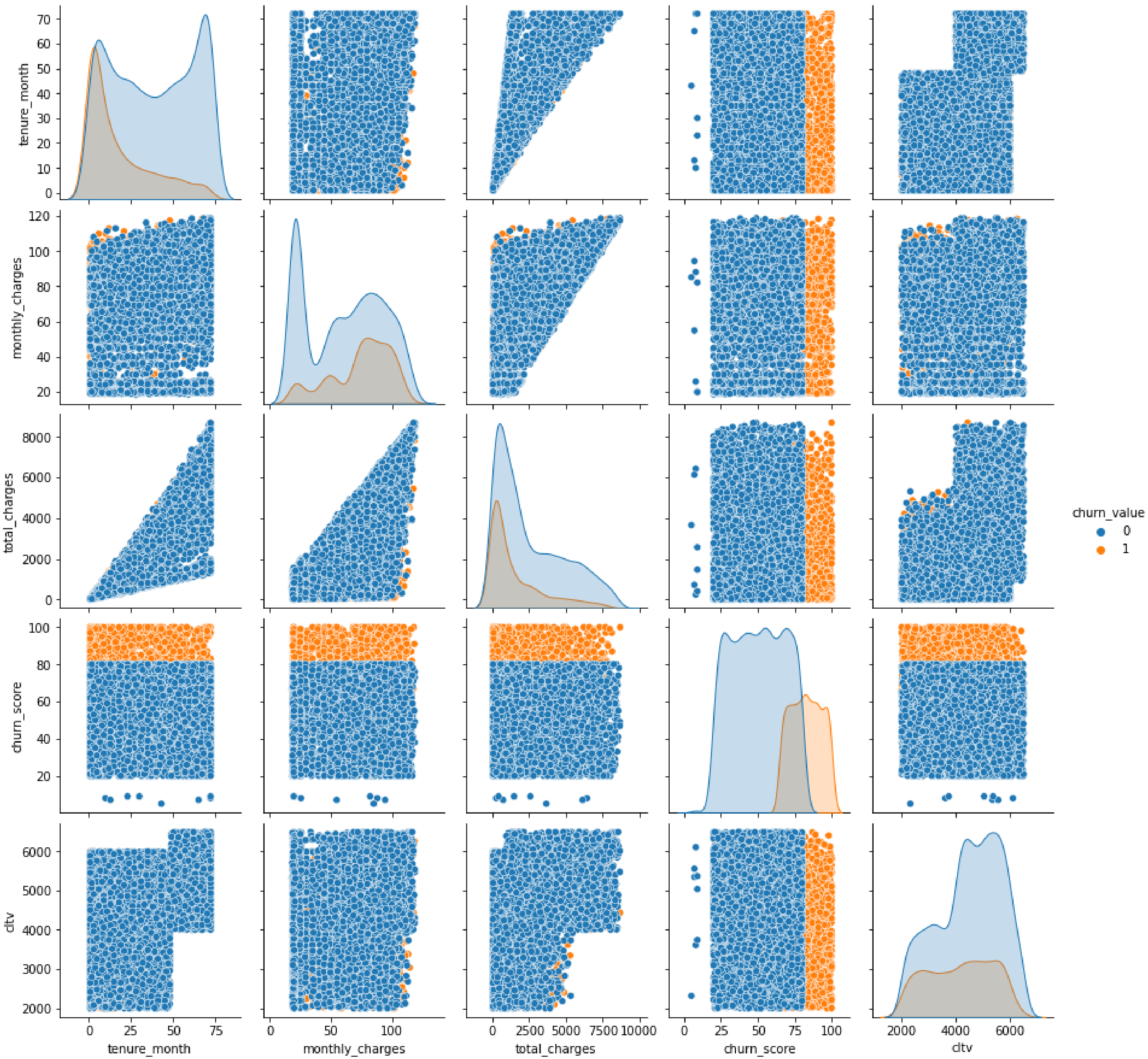
PAYMENT METHOD



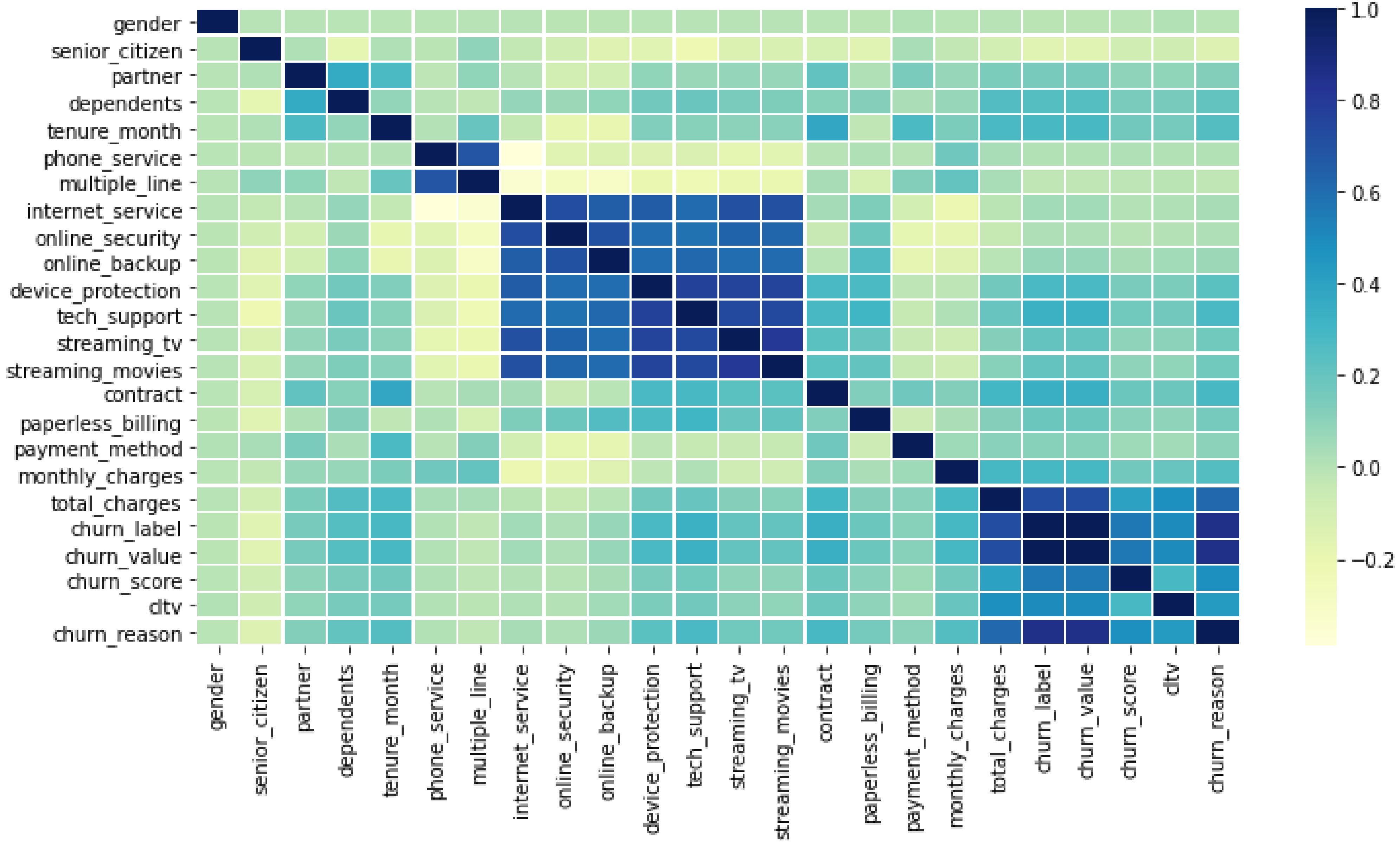
SENIOR CITIZEN



NUMERIC COLUMN



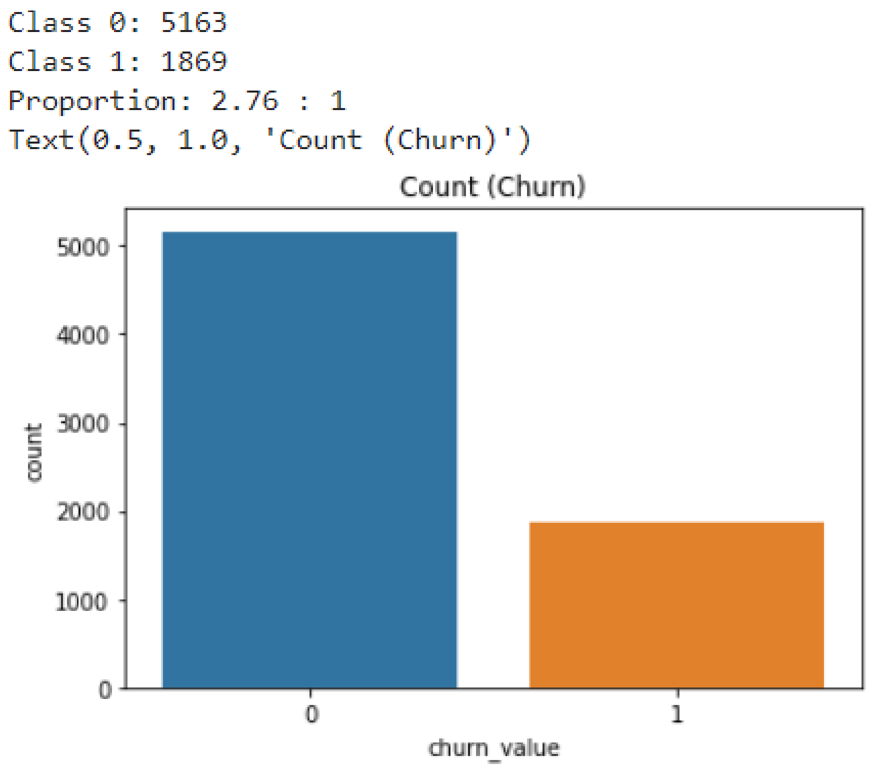
CORRELATION HEATMAP



RANDOM UNDERSAMPLING

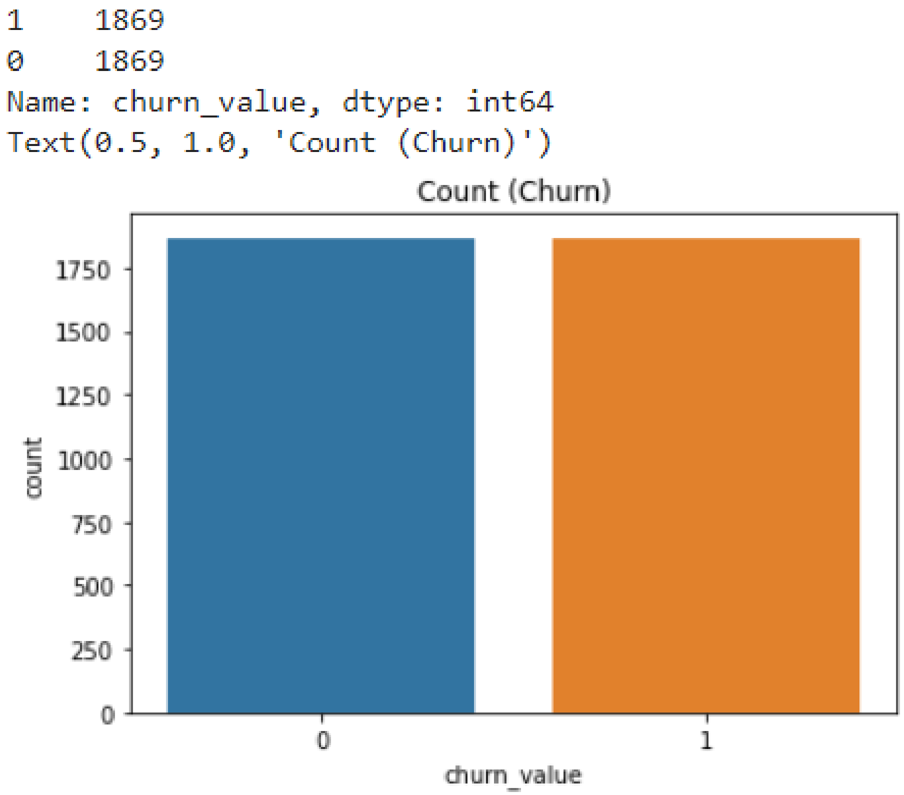
Random Undersampling involves randomly selecting examples from the majority class to delete from the training dataset.

BEFORE



```
Class 0: 5163
Class 1: 1869
Proportion: 2.76 : 1
Text(0.5, 1.0, 'Count (Churn)')
```

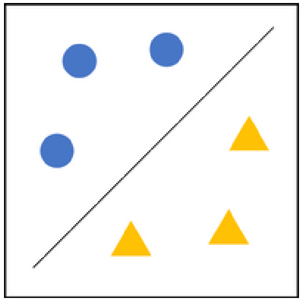
AFTER



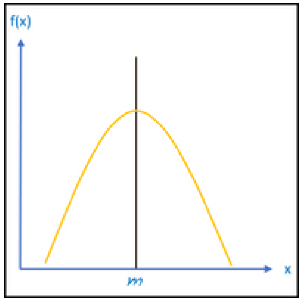
```
1    1869
0    1869
Name: churn_value, dtype: int64
Text(0.5, 1.0, 'Count (Churn)')
```

MODELLING

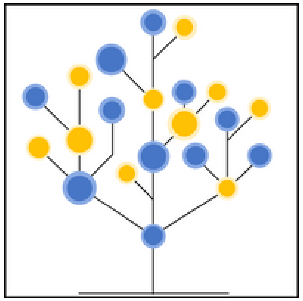
Support Vector Machine



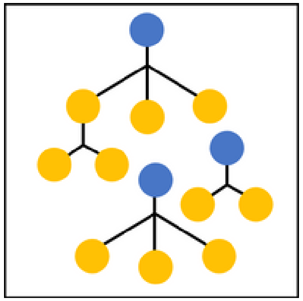
Gaussian Naive Bayes



Decision Tree Classifier



Random Forest Classifier



Train Set

80 % Total Data
X = 2990, 19
y = 2990, 1

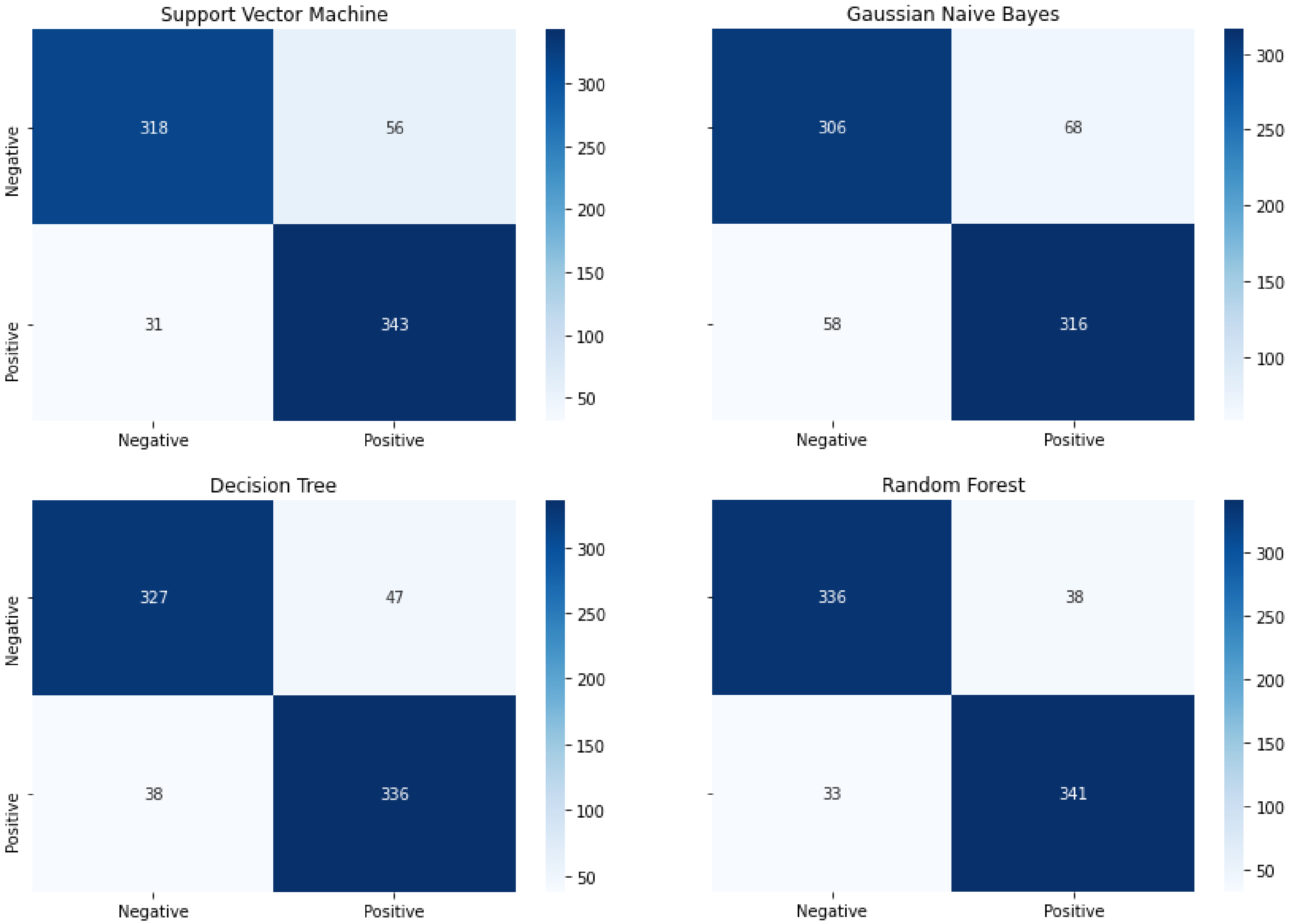
Test Set

20 % Total Data
X = 748, 19
y = 748, 1

MODEL COMPARISON

MODEL	ACCURACY	RECALL	PRECISION	F1 SCORE
Support Vector Machine	0.883	0.917	0.859	0.887
Gaussian Naive Bayes	0.831	0.844	0.822	0.833
Decision Tree Classifier	0.886	0.898	0.877	0.887
Random Forest Classifier	0.905	0.911	0.899	0.905

CONFUSION MATRIX



Support Vector Machine

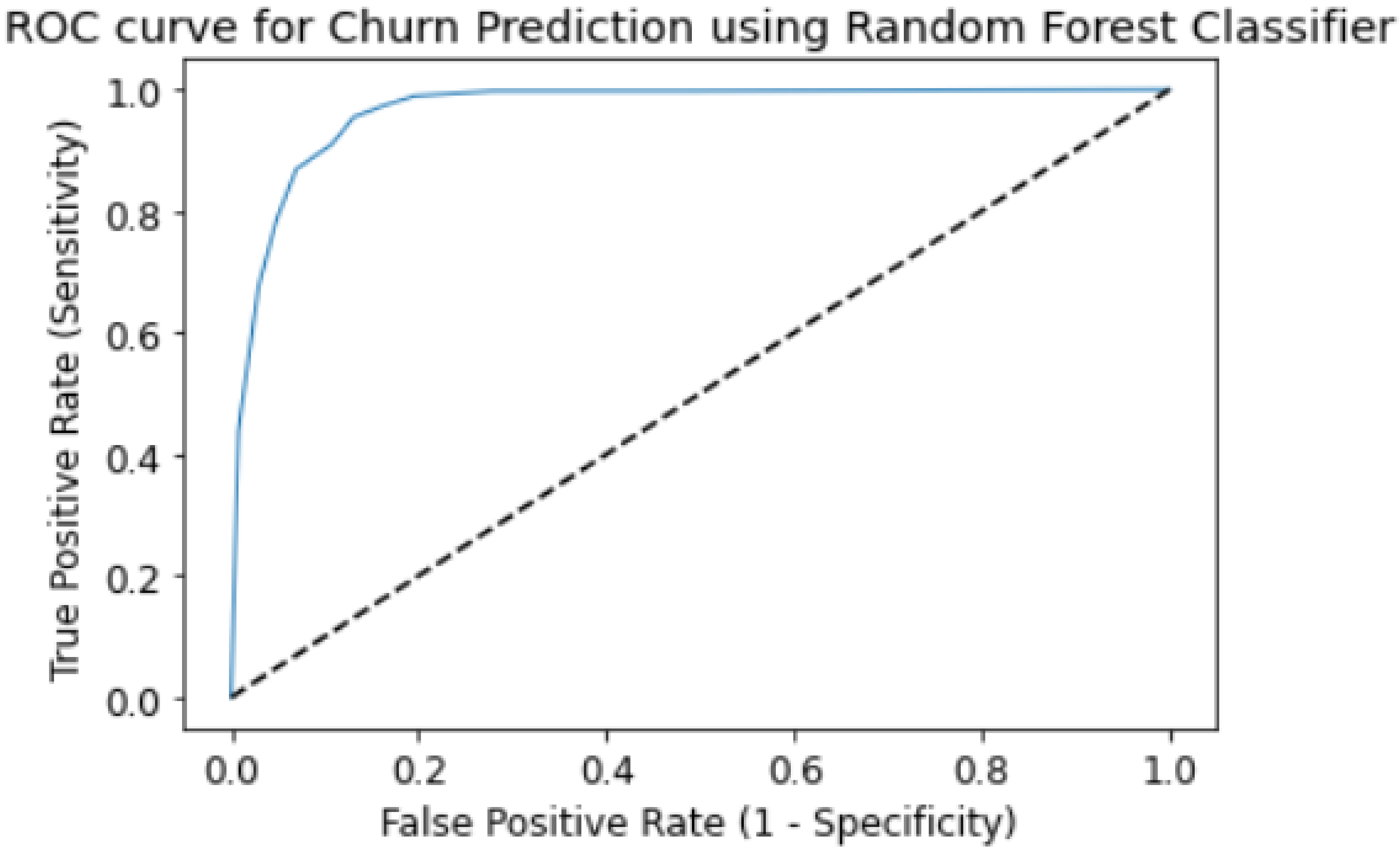
Gaussian Naive Bayes

Decision Tree

Random Forest

RECEIVER OPERATING CHARACTERISTIC

```
Cross validated : 0.9672
ROC Score : 0.9051
```



CONCLUSION

- Some column was dropped because they can cause the model to overfit.
- Random Undersampling was chosen to handle the imbalance data.
- Random Forest Classifier was chosen as it is the best model compared to the other model in this case.
- Model were validated using ROC AUC.

THANK YOU

https://www.linkedin.com/in/raka-raprast/

https://github.com/raka-raprast