PREDICTIVE ANALYSIS

Report

APRIL 7, 2022

Contents

Executive Summary

StirCom is a communications company that has been working in the streets of UK cities for the past few years to link people to their network. They are now eager to enter the 5G production and have initiated a marketing campaign to establish this. They called the customers to be a part of the new contract. We used machine learning to predict if the customer is going to signup for the contract or not.

Three different machine learning models are used for the predictive analysis. We used a logistic regression model, decision tree, and neural networks for the predictive analysis. The logistic regression model can achieve an accuracy score of 0.699, the neural network can achieve 0.732 and the Tree model can achieve an 0.747 accuracy score. In accuracy score, the Tree model has the highest score, and logistic regression has the least score among all three models. Logistic regression has an F1 score of 0.697, neural networks have 0.732, and the tree model has 0.747. Logistic regression has a precision score of 0.702, the neural network has a precision score of 0.732, and the tree model has a precision score of 0.747. Logistic regression has a recall score of 0.699, the neural network has a recall score of 0.731, and the tree model has a recall score of 0.747. Logistic regression has an area under the curve score of 0.765, the neural network has the area under the curve score of 0.804, and the tree model has the area under the curve score of 0.767.

Introduction

In this study, we are going to explore the Stircom data to predict the customer's new contract. We will use five stages cross-industry standard process for data mining (CRISP-DM) to answer the question. We will start with business understanding, data understanding, and defining problems. Once we have a proper problem in mind, we will explore the dataset and try to make it ready for analysis by preprocessing. After that, we will implement the supervised machine learning models to perform the predictive analysis. We will discuss in detail the performance of all the models and compare them with each other. To evaluate the performance of the model, we will discuss performance metrics like accuracy score, F1 score, area under the curve, recall, and precision. We will select the best-performing model and give suggestions for further studies.

Business understanding

StirCom is a telecommunications firm that has been digging up the streets of UK cities for the past few years to link people to their network. They are now eager to enter the 5G industry and have launched a marketing effort to support this. Customers who already have a landline can sign up for a mobile plan. Unfortunately, the call was not returned. Because the costs of running this campaign are substantial, they'd like to focus on customers. to prevent wasting calls more efficiently This is where we are going to enter the picture.

The main task of this analysis is to perform the predictive analysis. We are going to predict if the customer responded positively to the new contract call. Since we are going to predict if the customer is getting a new contract or not? It is a binary classification problem. In supervised machine learning, we have a lot of great machine learning models for predictive analytics. Some of the well-known models that can handle classification problems are logistic regression, decision tree, random forest, gradient-boosted tree, multilayer perceptron, one-vs-rest, and Naive Bayes. We will surely use supervised machine learning to answer this question. All the relevant questions will be referred to as independent variables (also known as predictors variable) and the variable containing information about the signing of the new contract will be used as the dependent variable (also known as the predicted variable). We might face problems like insufficient data, irrelevant data, poor quality of data, and imbalanced class data for the class variable. We will discuss the data in detail and try to extract the best insights from the data.

Data understanding

The data set contains twenty variables and 50662 rows of data. Each column refers to the specific question that the surveyor asked the customer, and each row gives all the responses from one client to the questions. The following table shows the name, description, and type of the variable.

Variable	Description	Type
INDEPENDENT VARIABLES		
ID	unique identifier for this record	Numeric
Town	hometown for customer	Categorial
Country	country for customer's home address	Categorial
Age	customer's age	Numeric
Job	customer's job	Categorial
Married	customer's marital status	Categorial
Education	customer's highest educational qualification level obtained	Categorial
Appears	has the customer failed to pay a recent bill?	Categorial
Current balance	the current amount in the customer's landline account in pounds	Numeric
Housing	is the customer a homeowner?	Categorial
Has TV package	has the customer got an additional TV and data package on their landline	Categorial
Last contact	type of communication used for the previous call to the customer	Numeric
Conn tr	connection type grouping ID	Numeric
Last contact this campaign month	last contact month of the year	Categorial
Last contact this campaign day	last contact day of the month	Numeric
This campaign	number of contacts performed during this campaign and for this client	Numeric
Days since last contact previous campaign	number of days that passed by after the client was last contacted from a previous campaign	Numeric
Contacted during the previous campaign	number of contacts performed before this campaign and for this client	Numeric

Outcome previous campaign	the outcome of the previous marketing campaign	Categorial
TARGET		
New contract this campaign	*has the client taken out a new contract?*	Categorial

There are 11 categorical variables and 9 numeric variables. Variable ID and town contain a lot of distinct categories which will lead to slow computation and less meaningful to the model. That's why I will drop these two variables.

The following table shows the feature statistics of all the remaining variables.

Name	**Mean**	**Median**	**Dispersion**	**Min.**	**Max.**	**Missing**
age	40.98	39	0.27	18	95	0 (0%)
current_balance	13.9912	4.72	2.2084	-80.19	984.17	0 (0%)
conn_tr	3.01	3	0.47	1	5	0 (0%)
last_contact_this_campaign_day	15.77	16	0.53	1	31	0 (0%)
this_campaign	2.71	2	1.12	1	63	0 (0%)
days_since_last_contact_previous_campaign	42.64	-1	2.40	-1	871	0 (0%)
contacted_during_previous_campaign	0.64	0	3.68	0	275	0 (0%)
country		UK	0.00114			0 (0%)
job		management	2.13			0 (0%)
married		married	0.918			0 (0%)
education		secondary	1.12			0 (0%)
arrears		no	0.0867			0 (0%)
housing		yes	0.69			0 (0%)
has_tv_package		no	0.43			0 (0%)
last_contact		cellular	0.805			0 (0%)
last_contact_this_campaign_month		may	2.08			0 (0%)
outcome_previous_campaign		unknown	0.697			0 (0%)
new_contract_this_campaign		no	0.494			0 (0%)

Data preparation

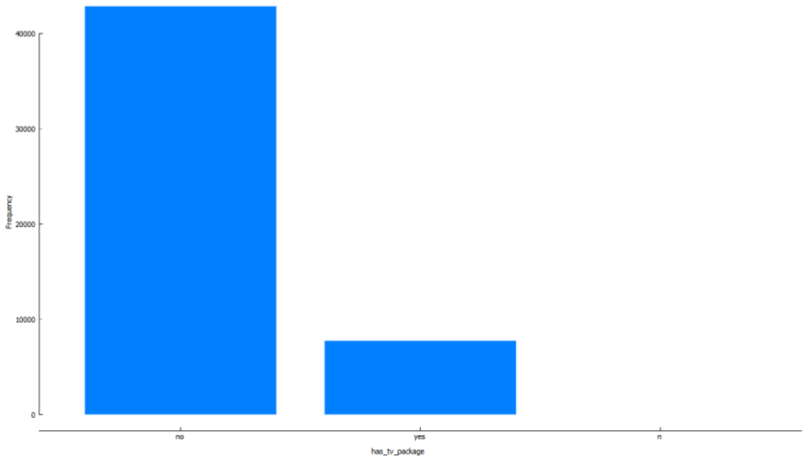
Some variable has wrong entries in them. To deal with these spelling problems, I used the edit domain feature in orange to merge or drop them according to their nature. I faced issues in the following variables.

- Has tv package
- Last contact
- Last contact this campaign

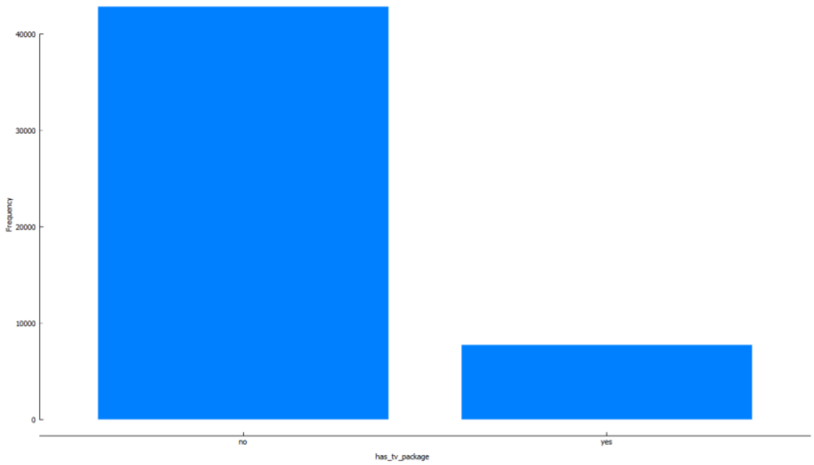
Has tv package

In has tv package variable we should have only two categories yes and no. But I had three different labels, yes-no, and n. Here I am supposing that n is typing mistake and it should be no instead of n. That's why I merge both.

Before preprocessing,



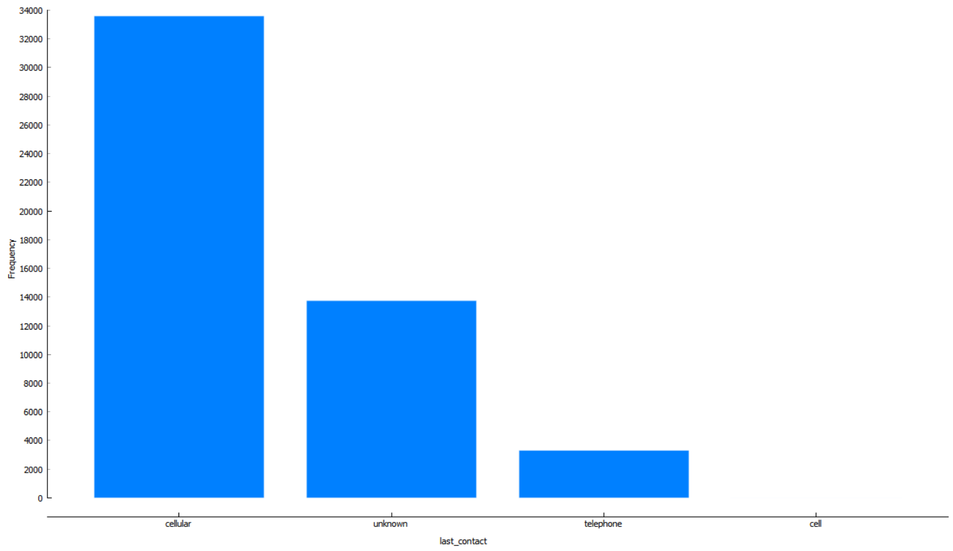
After preprocessing,



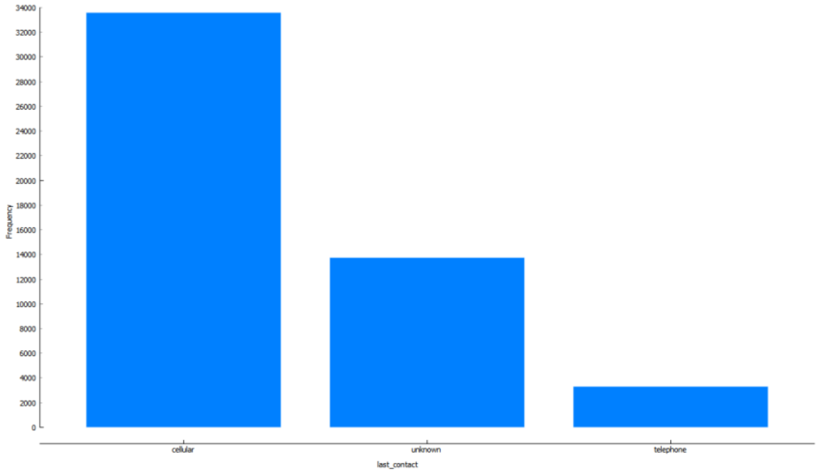
Last contact

Similarly in last_contact variable, we should have three categories like cellular, unknown, and telephone. But we have another category named cel. Which I merged with the cellular.

Before preprocessing,



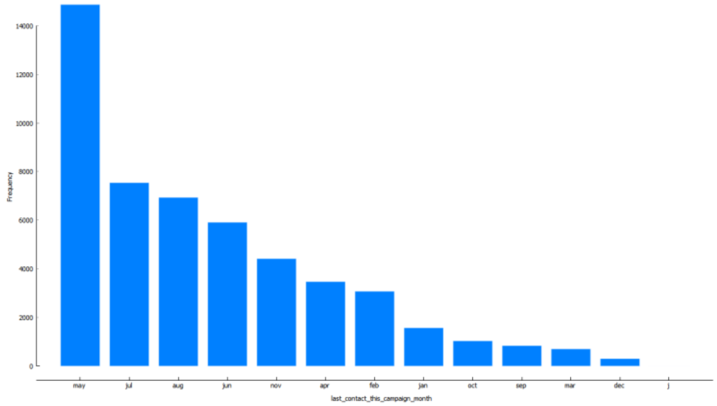
After preprocessing,



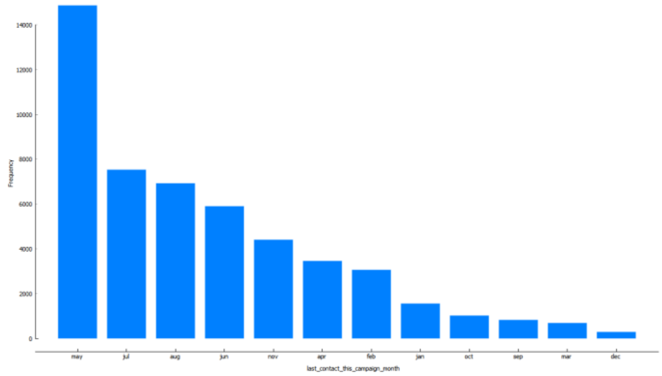
Last contact this campaign

Last_contact_this_camaign variable should have one category for each month. But we have one extra category named j only. It can be either January or July. Since we are not sure about its actual value, that's why I dropped this value.

Before preprocessing,



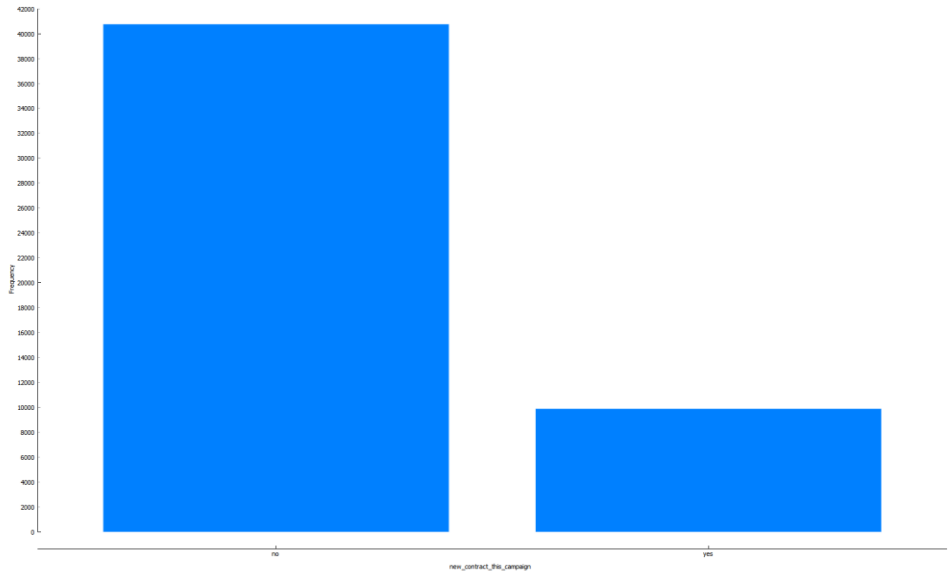
After preprocessing,



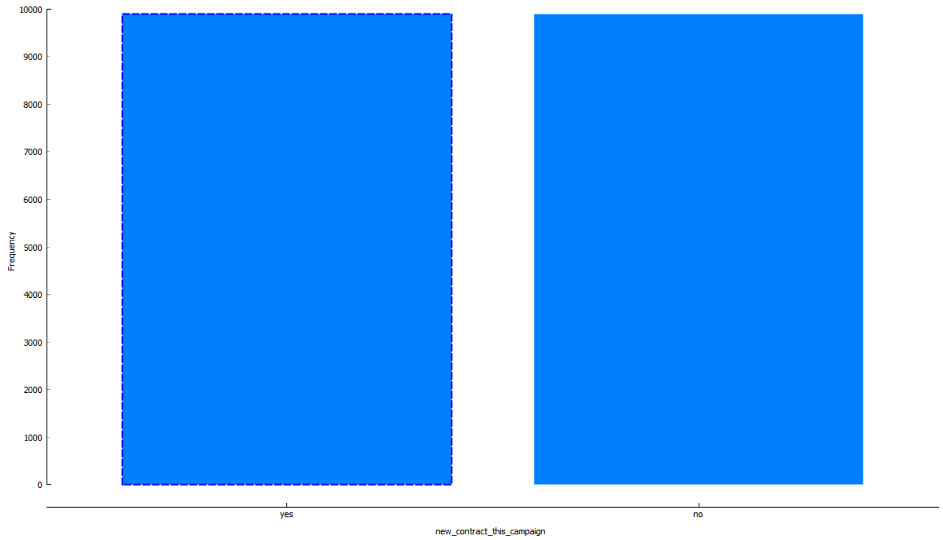
99% of the values in the country variable are the United Kingdom and less than 1 % of values contain France, Germany, and Portugal. That’s why there is no point in adding the country variable because it doesn’t provide enough useful information to the model that’s why I am skipping it for predictive analysis.

Dependent variable new contract this campaign is imbalanced. Follow classification analysis, we need to deal with the class balance. To balance the dataset, I used combination of select rows, data sampler and concatenate and create a balanced data set for our response variable.

Before preprocessing,



After preprocessing,



Modeling

We used three different models from supervised machine learning, these models are

- Decision tree
- Logistic regression
- Neural networks

Decision tree

The decision tree is used with the following settings,

The minimum number of instances in leave is set to 2.

Subsets smaller than 5 will not be split.

The limit for the maximum tree depth is 100.

Classification will be stopped when the majority reaches 95%.

Logistic regression

For the logistic regression, we used L2 regularization and C=1

Neural network

For the neural network, we used the following settings,

Neurons in the hidden layer will be 100.

We used rectified linear activation function as our activation function in the network

We used the Adam solver for computation purposes.

The maximum number of iterations is set to 200.

For regularization we used a = 0.0001

Performance metrics

To evaluate the performance of the model, the following metrics will be discussed.

Confusion matrix

A confusion matrix is a table that is frequently used to describe the capability of a classification model on a set of test data for which the actual values are acknowledged. The confusion matrix is quite simple to understand, but the related terminology can be puzzling.

Basic terms of confusion matrix are,

True positives (TP): These are cases in which we predicted yes, and the actual value is also yes.

True negatives (TN): When a predicted value is no, and the actual value is no.

False positives (FP): When predicted yes, and the actual value is no. (Also known as a "Type I error.")

False negatives (FN): When predicted no, and the actual value is yes. (Also known as a "Type II error.")

Accuracy score: The accuracy score tells us about the correctly classified entries.

Precision

The ratio of accurately predicted positive observations to total expected positive observations is known as precision.

$$Precision \ = \frac{TP}{TP + FP}$$

Recall

The ratio of accurately predicted positive observations to all observations in the actual class is known as recall.

$$Recall \ = \frac{TP}{TP + FN}$$

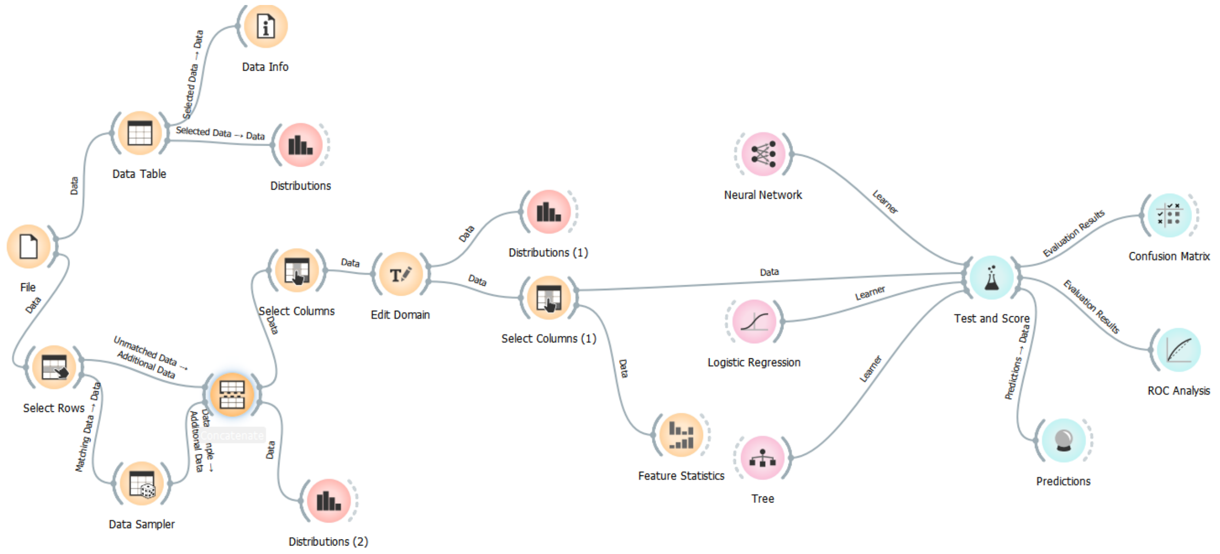
F1_score

The F1 score can be understood as a weighted average of the precision and recall values, where an

F1 score reaches its best value at 1 and worst value at 0.

Orange workflow

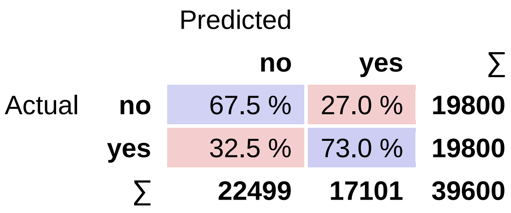
The following picture represents the orange workflow done for this report.



Evaluation

Logistic regression can predict 67.5% of the no values in the dependent variable correctly, and 27% of values are falsely classified. 73% values are correctly classified as yes and 32.5% of the values are falsely classified as yes. The following table represents the confusion matrix of the logistic regression model.

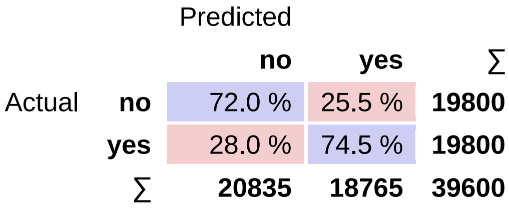
Confusion matrix for Logistic Regression (showing proportion of predicted)



		Predicted		
		no	yes	∑
Actual	no	67.5 %	27.0 %	19800
	yes	32.5 %	73.0 %	19800
	∑	22499	17101	39600

The neural network can predict 72% of the values in the dependent variable correctly classified as no, and 25.5% of the values are falsely classified as no. 74.5% of the values are correctly classified as yes and 28% of the yes values are falsely classified as yes. The following table represents the confusion matrix of the neural network model.

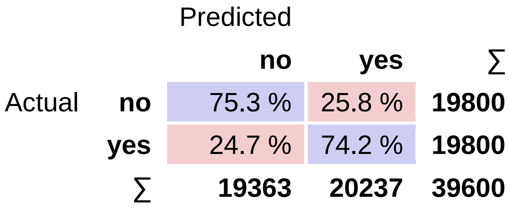
Confusion matrix for Neural Network (showing proportion of predicted)



		Predicted		
		no	**yes**	∑
Actual	**no**	72.0 %	25.5 %	**19800**
	yes	28.0 %	74.5 %	**19800**
	∑	**20835**	**18765**	**39600**

The tree model can predict 75.3% of the values in the dependent variable correctly classified as no, and 25.8% of values are falsely classified as yes. 74.2% of the values are correctly classified as yes and 24.7% of the values are falsely classified as yes. The following table represents the confusion matrix of the tree model.

Confusion matrix for Tree (showing proportion of predicted)

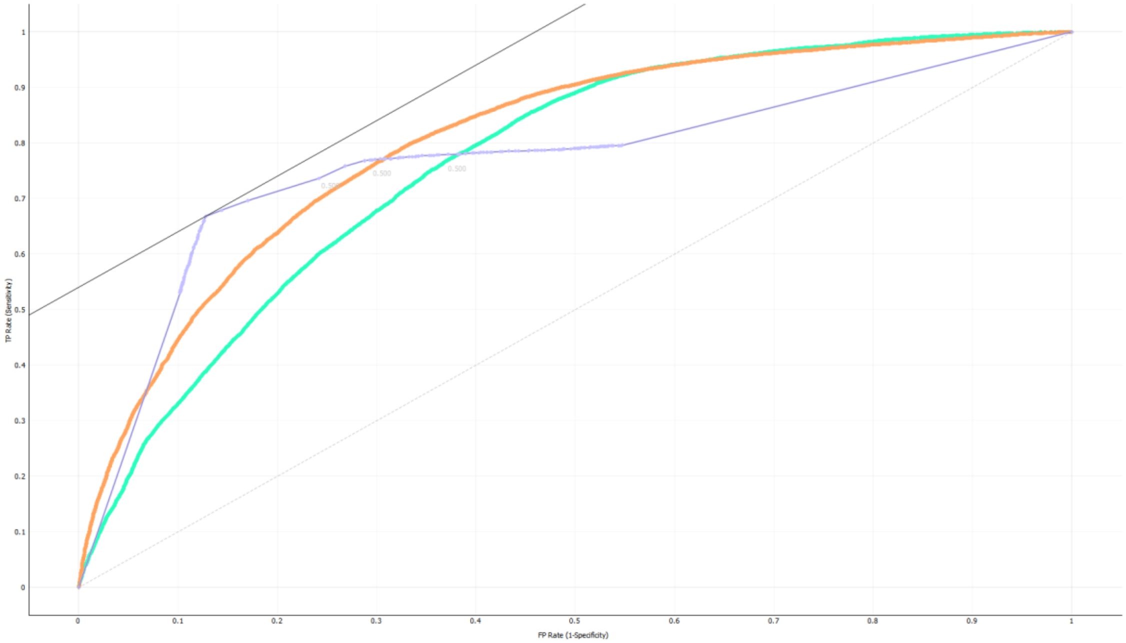


		Predicted		
		no	**yes**	∑
Actual	**no**	75.3 %	25.8 %	**19800**
	yes	24.7 %	74.2 %	**19800**
	∑	**19363**	**20237**	**39600**

For training and testing purposes we used a stratified shuffle split, 10 random samples with 80% data. The logistic regression model can achieve an accuracy score of 0.699, the neural network can achieve 0.732 and the Tree model can achieve an 0.747 accuracy score. In accuracy score, the Tree model has the highest score, and logistic regression has the least score among all three models. Logistic regression has an F1 score of 0.697, neural networks have 0.732, and the tree model has 0.747. Logistic regression has a precision score of 0.702, the neural network has a precision score of 0.732, and the tree model has a precision score of 0.747. Logistic regression has a recall score of 0.699, the neural network has a recall score of 0.731, and the tree model has a recall score of 0.747. Logistic regression has an area under the curve score of 0.765, the neural network has the area under the curve score of 0.804, and the tree model has the area under the curve score of 0.767.

Model	**AUC**	**CA**	**F1**	**Precision**	**Recall**
Tree	0.7669183935312724	0.7470959595959596	0.7470651574367734	0.7472163826787669	0.7470959595959596
Neural Network	0.8043875293337416	0.7317424242424242	0.7315590496779608	0.732377381032539	0.7317424242424242
Logistic Regression	0.7651637205387206	0.6986616161616162	0.6972552707905162	0.7024228904972428	0.6986616161616162

The following graph represents the area under the curve for all three models,



Accuracy score suggests that the tree model is one of the best models for classification of this data. The logistic regression model is the worst performing among all three models.

Suggestions:

We lose a lot of data in under sampling. We should try to gather more balanced data for the analysis. We can also implement different machine learning models to increase accuracy scores, we can implement random forest, k-nearest neighbor, support vector machines, naïve Bayes, and stochastic gradient descent algorithms that might help us to increase the model performance.