Hyper Crypto
A Proposal for Hyper Big Bank

A M Mukhtar

University of Reading

Machine Learning and Big Data

Table of Contents

Abstract

This report looks in detail at the making of a predictive model which can be used to predict which customers of Hyper Bank are likely to invest in Hyper Crypto. The report investigates how and why we came to choose a Decision Tree Model over the others. We look at further details on how to enhance the process of identifying potential clients. We also show the features which have the most correlation of determining whether one would be potentially interested in Hyper Crypto along with the features that are redundant and play little to no role in the selection. Furthermore, this report highlights ways to improve the process of identifying the potential clients and the data required.

Keywords: Cryptocurrencies, Hyper Bank, Machine Learning, Decision Trees

Introduction

The predictive machine learning model will make it easier for Hyper Bank to determine which of its clients is likely to invest in Hyper Crypto. This model would be beneficial in helping reduce time, efforts, and costs. It would prevent cold calling all the customers and narrowing down the clients which have a higher likely hood of buying.

Building a predictive machine learning model

Predictive Modelling involves analyzing data to predict a future event or outcome. They are classified into two main categories: unsupervised and supervised models. We compare both to have insight on why we chose the particular model that we did.

- Unsupervised Models: Unlike supervised models, these allow for complex tasks to be carried out. The algorithms analyse and cluster the unlabelled dataset free from human intervention, hence the name unsupervised.
- Supervised Models: These make use of labelled data to train the algorithm to classify the data and make accurate decisions. These models can also be divided into two separate problems: regression and classification.

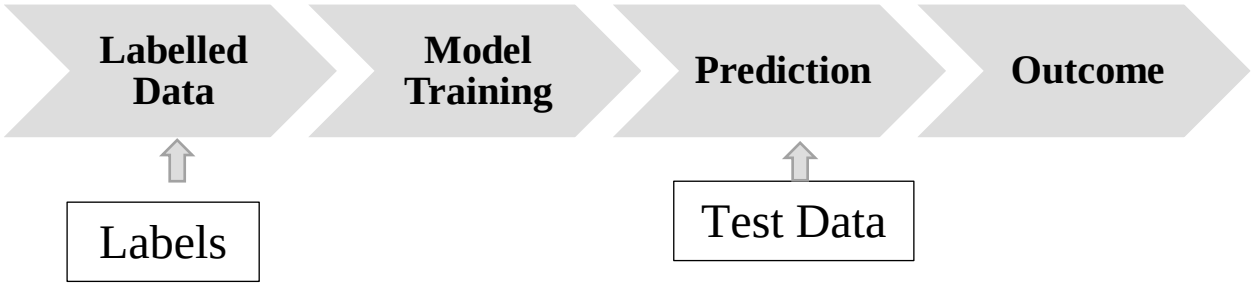

Figure 1 Flow chart of learning model

The flow chart above depicts the operation of the learning procedure. The same procedure is used to train our model. The dataset was split into three: the training dataset, test dataset and validation dataset. The input features of the training dataset was determined and run with the different models to find one that yields the highest accuracy.

Building a predictive machine learning model requires the use of labelled data with each record belonging to a certain class. The labelled data in this case being the features.

The predictions can be either classification or regression problems.

- Regression models allow finding the relationships between the values of one or more independent variables with one dependent variable (Balanchine, 2018). The relationship would then allow one to determine the output for another input. E.g., Linear regression, Logistic regression.
- Classification models on the other hand are identical to regression with the exception of the dependent variables being discrete values as opposed to continuous. E.g. decision trees, and random forests.

The data of the bank's customers are labelled data which belong to a classification problem. As the dataset contains several numerical columns and 1 categorical column it makes it possible to build a predictive machine learning model. The numerical variables are those that describe quantifiable characteristics as numbers (ABS, 2020) while the categorical have a finite number of outcomes such as binary. The class column is the categorical variable here as it contains 2 values of 0 and 1 making it a case of binary classification.

Recommended Machine Learning Model

As this model falls under supervised learning, we recommend making use of any of the supervised learning models. Supervised learning models are typically used to classify data and make predictions. The process of identifying these customers is a binary problem with labelled data available, making it possible to use any supervised model. The presence of the labelled data is important in the choice of the model.

With the data data relating to finance, and the features have nonlinear relation to the target value, an easily interpretable model would be the best option. Based on mentioned reasons, tree-based models such as Decision trees, Random Forest, or Gradient boosted trees are all options that can be used. As random forest and gradient boosted trees are ensemble methods, they are complex while a single decision tree is easily explainable. The decision tree is recommended due to its simplicity in nature, making it easy to interpret, understand and visualize. The decision trees are also fast when compared to other classification algorithms.

However, if the dataset is set to increase in size, a random forest would be more ideal as it does not rely on a single tree and it won't lead to overfitting.

Data for less than 12 features

To narrow down the number of features required and simplify the model we observe the features with the highest correlation to buying Hyper Crypto.

The following are the categories of the features:

- Feat 8: Hyper Equity
- Feat 9: Hyper Tech
- Feat 10: Hyper Bonds
- Feat 11: Hyper Gold

	Feat0	Feat1	Feat2	Feat4	Feat5	Feat6	Feat7	Feat8	Feat9	Feat10	Feat11	PotentialBuyerHyperCrypto
Feat0	1.000000	-0.039633	0.023859	-0.008923	-0.049049	-0.033220	-0.012455	0.009814	-0.007110	-0.028196	-0.006155	-0.005293
Feat1	-0.039633	1.000000	0.030277	0.040114	-0.013127	0.012844	0.036842	0.037912	0.057861	-0.007072	-0.005147	-0.022530
Feat2	0.023859	0.030277	1.000000	-0.008627	-0.066468	0.000760	0.040986	0.000473	0.055493	0.031602	-0.029780	0.089447
Feat4	-0.008923	0.040114	-0.008627	1.000000	-0.001045	-0.020325	0.016779	0.004448	-0.003903	0.008604	0.026306	0.000959
Feat5	-0.049049	-0.013127	-0.066468	-0.001045	1.000000	0.004424	-0.002935	0.023419	-0.001186	0.015454	-0.000754	0.176224
Feat6	-0.033220	0.012844	0.000760	-0.020325	0.004424	1.000000	-0.022646	-0.000322	0.019584	0.055929	-0.010669	-0.007358
Feat7	-0.012455	0.036842	0.040986	0.016779	-0.002935	-0.022646	1.000000	0.074500	0.010640	0.013763	0.037763	0.005495
Feat8	0.009814	0.037912	0.000473	0.004448	0.023419	-0.000322	0.074500	1.000000	-0.007284	-0.009235	-0.051229	0.134106
Feat9	-0.007110	0.057861	0.055493	-0.003903	-0.001186	0.019584	0.010640	-0.007284	1.000000	0.006097	-0.019686	0.039892
Feat10	-0.028196	-0.007072	0.031602	0.008604	0.015454	0.055929	0.013763	-0.009235	0.006097	1.000000	-0.021014	-0.049023
Feat11	-0.006155	-0.005147	-0.029780	0.026306	-0.000754	-0.010669	0.037763	-0.051229	-0.019686	-0.021014	1.000000	-0.355211
HyperCrypto	-0.005293	-0.022530	0.089447	0.000959	0.176224	-0.007358	0.005495	0.134106	0.039892	-0.049023	-0.355211	1.000000

Figure 2 Heatmap Correlation

From the heatmap above we observe a high correlation of Feat 5, 8 and 11 with the target column. Hyper Gold holders denominated by feat 11 has the highest negative correlation indicating their buyers are least likely to buy hyper crypto. This finding is in line with our understanding of the investors, risk-averse investors such as gold investors tend to distance themselves from cryptocurrencies due to their inherent volatility. They are typically of the older generation which prefers their asset to be in a safe investment with little fluctuation.

Feat 8 and 11 would also be recommended due to the high positive relation. Feat 8 customers invest in the equity market which is a high-risk investment product. The Hyper Crypto would be suitable for customers in this market as it is also a high-risk high reward investment fund making it within their area of interest.

Collecting data from these 3 features would give us a clear understanding of which customer would be interested in the product and would help the model predict accurately.

The lowest features to maintain 70% accuracy

A single feature would be enough to ensure 70% accuracy. By using feat 11, we obtain an accuracy of 0.88 (88%). This shows that by simply owning HyperGold, the customers are 88% unlikely to buy Hyper Crypto. Other features have been tested but none give better results.

When testing the correlation with Pearson Correlation coefficient, it measures the correlation between two datasets, in this case the correlation between the features and Buyers of Hyper Crypto. It is observed that Feat11 had the highest negative correlation.

These results show how Feat 11 has the highest correlation on both positive and negative spectrum. The results ate also consistent when tested with the Spearman’s Correlation, Phik Correlation, Kendall’s Tau, and Cramer’s V.

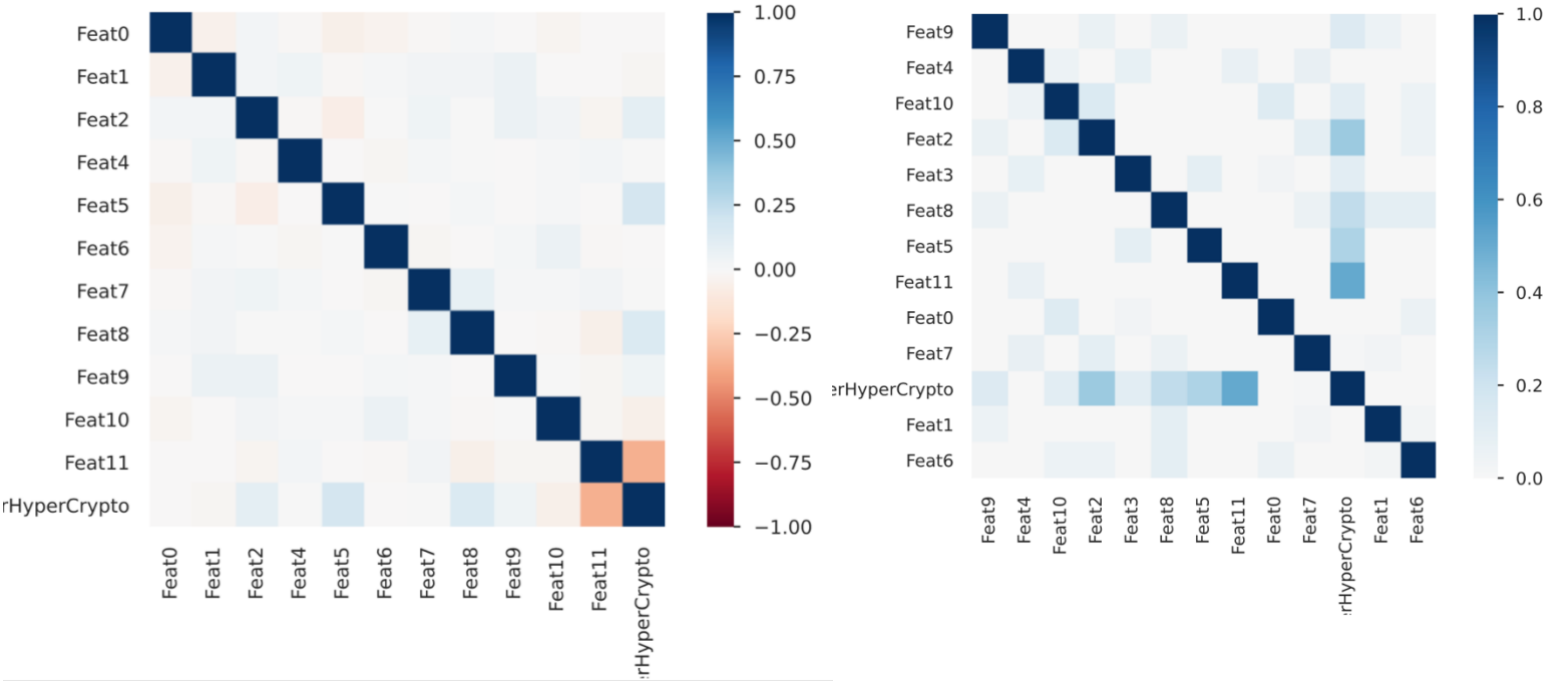

Figure 3 Correlation Map with Pearson Correlation on the left and with Phik (φk) Correlation on the right

Data Set Statistics

The dataset contains 1500 rows and 13 features. One feature is categorical in nature while others are numerical. The categorical feature is found in Feat3 which contains 3 values A. B and D. When analysing the dataset, the following statistics are obtained:

Dataset Statistics	
Number of Variables	13
Number of observations	1500
Missing cells	57
Missing cells %	0.3%
Duplicate rows	0

Table 1 Correlation Map with Pearson Correlation on the left and with Phik (φk) Correlation on the right

Missing Data

From the data obtained we found 57 missing data which accounts for only 0.3% of the data, we decided to remove this data. The visualization below shows the normal distribution of Feat 5, allowing us to fill in the missing data using the mean value. The mean value is chosen as the data is numerical with little to no outliers.

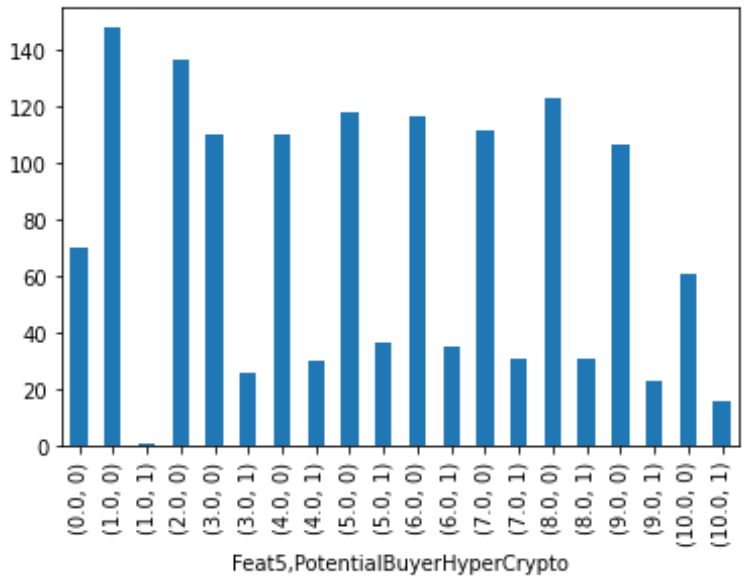

Figure 4 Distribution of Feat 5

Outliers

To detect outliers, we made use of Boxplot, a chart that visualizes how data is distributed using quartiles. Any data points present outside of the third quartile are considered as outliers.

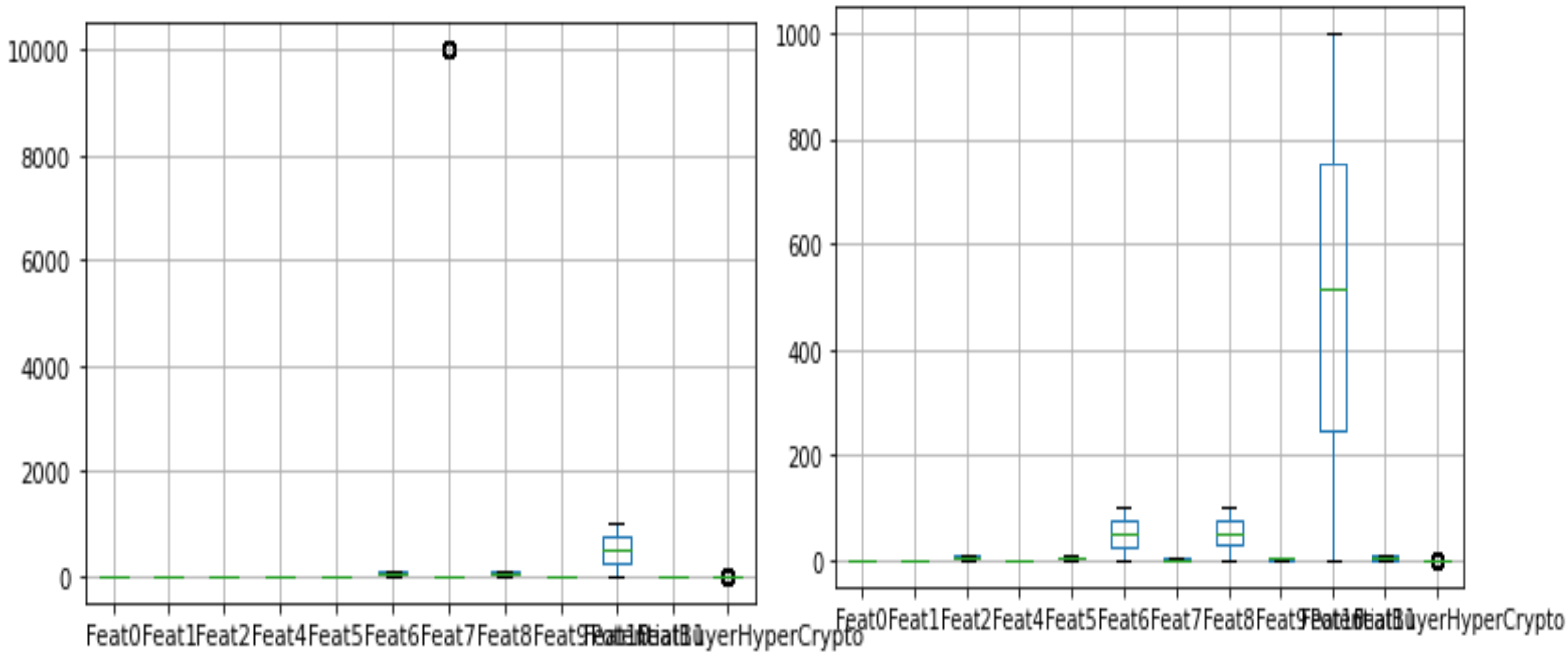

Figure 5 Quartiles showing outliers

We remove all rows where the value of feat 7 was greater than the mean value of feat7 thereby eliminating the outliers. After which we are left with 1460 rows in the dataset.

Categorical Feature

Most machine learning models only process numerical variables, thus to make use of the categorical data in feat 3 it is mapped to numerical.

- A is mapped to 0
- B is mapped to 1
- D is mapped to 2

After which the data is scaled as some columns contained data within range of 0-1 while others had range of 0-1000. The table displays the different methods used and their descriptions.

	DESCRIPTION
BOX PLOT	To display outliers in the dataset.
SIMPLE SCALAR	To ensure all the data are within the same ranges.
DF FILLNA	Used to fill missing values with specific values (Mean of the feature)
ONE HOT ENCODING	Used to encode the categorical data from Feat 3 converting it to numerical.

Table 2 showing the description

Train-Test Split

The train test split technique is typically used in classification or regression problems. It was used to evaluate the performance of the predictive machine learning model by splitting the data into two subsets.

The Train Dataset is used to fit the machine learning model, while the Test Dataset is used to evaluate the fit machine learning model. (Brownlee, 2020) The Train data compromised of 1168 rows which is 80% of the data while the test data contained the rest of the 292 rows (20% of the original data).

Chosen Model

After running trials on several models, we found the decision tree models to be the most efficient. The decision trees fall under supervised learning algorithms. The decision trees are versatile and can be used to solve any classification or regression problems. Below we outline the many advantages of using decision trees.

- Require minimal effort and are easier to understand.
- Fast and efficient compared to KNN and others
- Requires minimal data preparation and is useful in data exploration.
- Can capture non linear relationships and can handle any type of data (Numerical, Categorical and Boolean).

Decision tress also come with some disadvantages. As the dataset increases in size a decision tree would be less efficient. The decision tree would grow a lot of nodes thereby resulting in overfitting. The complexity of the trees also increases as the input increases and also increasing the training time. In the event of larger data sets we would recommend the use of the next best alternative, the Random Forest. The random forest handles all the limitations of the single decision trees. Random forest makes use of various decision trees to arrive at a conclusion. They are however slower and can be biased when dealing with categorical variables. The decision trees are chosen due to the combination of their accuracy, AUC and F1 score.

	Accuracy	F1 Score	AUC
Decision Tree	0.98	0.96	0.96
Random Forest	0.97	0.94	0.99
Gradient Boosted Trees	0.97	0.94	0.99
SVM	0.92	0.76	0.95
Neural Network	0.90	0.72	0.90
Logistic Regression	0.87	0.60	0.81

Table 3 comparing performance of different models

The table above compares the performance of the models. The accuracy measures how well the models are performing. It can be converted into percentage by multiplying by 100 and it is found using the formula:

$$accuracy = \frac{number\ of\ correctly\ predicted\ samples}{total\ number\ of\ samples}.$$

AUC stands for area under the curve, it's a statistical measure that we can use to evaluate the model predictions using a probabilistic framework. (Riva, 2021) The higher the AUC the better the model is at prediction of classes. The F1 score is used to compare the performance of two classifiers, it combines the precision and recall of the model and is defined as the harmonic mean of the model's precision and call. (Wood, 2021)

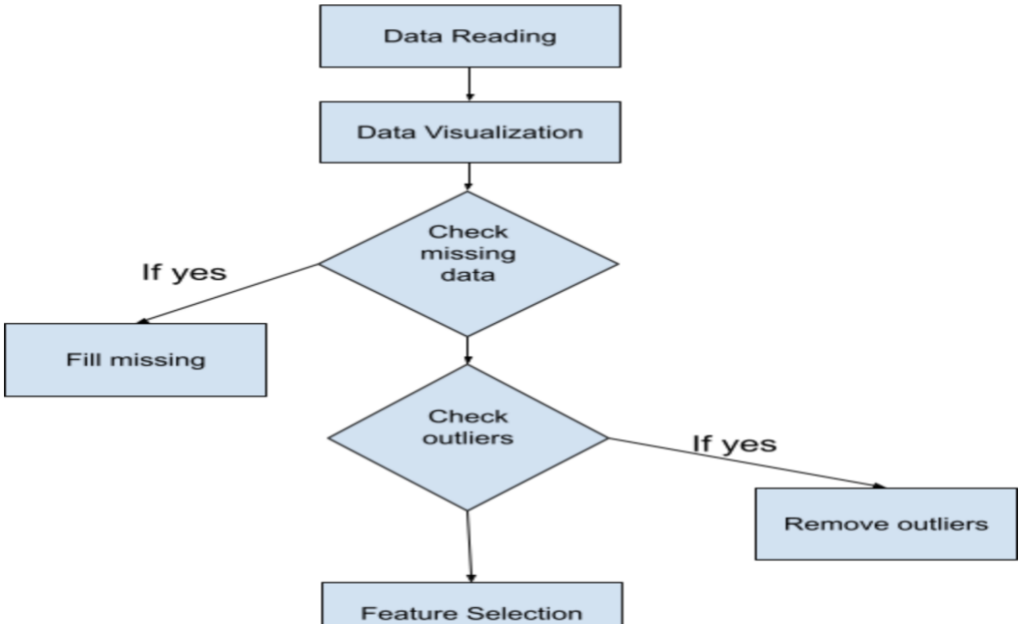

Figure 6 Tree explaining code

The Tree above explains the working principles of the Jupyter code.

- The first cell contains imports of different libraries and data read. Pandas, Numpy, Sklearn, and Matplotlib are used and these all are available in the standard installation.
- The second cell contains a correlation graph of all features.
- The third cell contains code for finding that Feat11 is very related to the target value. If the value in feat11 is equal to 6 or greater there are no customers who want to buy hyper crypto.
- In the fourth cell, the missing data is filled in Feat5 column by using the mean value. The visualization shows Feat5 has a nearly normal distribution so it can be filled with the mean value.
- Ouliers that were present in single-column Feat7 were removed. They were detected using boxplot.
- The categorical column (Feat 3) gets encoded.
- In the next cell, the data is scaled which was necessary because some columns contains data within the 0-1 range while others had 0-1000. I have used a simple scalar.

In the next cell, I have again plotted the correlation map.

In the next cell, I have selected some features. This is purely trial and test-based selection. This features performed well in prediction.

In the next cell, training data is split into train test for validating the training of the model. Split is made on 80/20 rule with random_state of 42 which is the default.

The next cells contain different classifiers I have tested.

Decision Trees

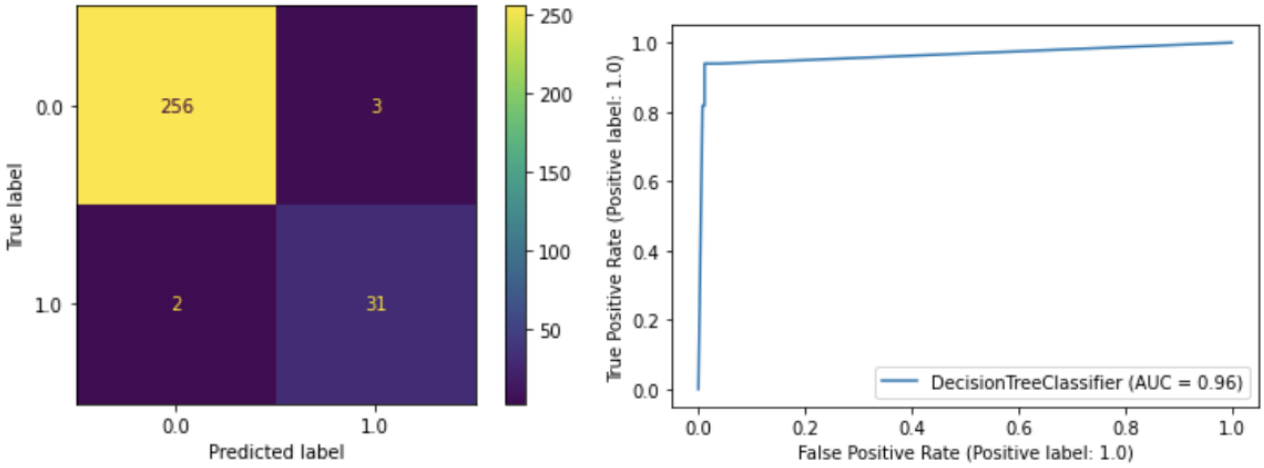

Figure 7 Decision Tree Matrix and AUC

Decision tree models is used from sklearn and has lowest False Positives (Only 3). Hence why it is chosen. The logistic regression had 14 false positives. The decision tree model provides the best results. It is an easily interpretable model and rule-based model.

Discussions

One of the major findings during the course of developing the machine learning model is that most holders of gold will not be interested in buying Hyper Crypto. We found that if the value in feat11 is equal to 6 or greater there are no customers who want to buy hyper crypto.

To enhance the results and obtain better predictions in the future, Hyper Bank could provide the age of the customers. Research has shown that younger audience is more likely to invest into the cryptocurrency market. Hyper Bank can also analyse which customers have made payments to crypto exchange platforms. Gender can also play an important role in determining whether one is to buy Hyper Crypto, we learn that women are more cautious investors than men. (WirexTeam, 2022)

Bibliography

ABS, 2020. *ABS.* [Online]
Available at: https://www.abs.gov.au/websitedbs/D3310114.nsf/home/statistical+language+-+what+are+variables

Balanchine, S., 2018. *Certixcloud.* [Online]
Available at: certixcloudservices.com
[Accessed 12 05 2022].

Brownlee, J., 2020. *Machine Learning Mastery.* [Online]
Available at: https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/
[Accessed 18 05 2022].

Last Name, F. M., Year. Article Title. *Journal Title*, pp. Pages From - To.

Last Name, F. M., Year. *Book Title.* City Name: Publisher Name.

Riva, M., 2021. *Baeldung.* [Online]
Available at: https://www.baeldung.com/cs/ml-loss-accuracy#accuracy

WirexTeam, 2022. *Wire.* [Online]
Available at: https://wirexapp.com/blog/post/is-the-gender-gap-closing-in-the-crypto-community-0569

Wood, T., 2021. *DeepAi.* [Online]
Available at: https://deepai.org/machine-learning-glossary-and-terms/f-score