Machine Learning and Big Data in Finance (Part 1)

May 2023

Executive Summary Report

Hyper Climate Change Fund

Machine Learning classification model for Newly Developed Fund

Adedayo Adewole 30810670

Table of Contents

Introduction

Artificial Intelligence and Machine Learning are poised to play a large role in marketing optimization. Amongst other brilliant uses, Machine Learning features build models and classify customers based on certain behaviours, so marketing teams can form accurate insights and buyer personas. The goal of this report is to show the value of Machine Learning and form customer personas from the database of existing customers of Hyper Big Bank as well as be able to classify new customers.

Is it possible to build a predictive Machine Learning model?

A predictive Machine Learning Model is a type of Artificial Intelligence (AI) system that is designed to automatically analyze data, learn from it, and then make predictions or decisions based on that knowledge. These predictions can be used to optimize processes, develop new products or in this case, make better decisions as to which prospective client to schedule follow up calls with. The main goal of predictive Machine Learning models is to make accurate predictions about future outcomes based on historical data. By adopting this, we can harness the value of any amount of data (even Big data) we get from future prospective customers to ascertain to a certain level of confidence if we will need our Sales team to follow up for effective conversion or not. So yes, we can build a predictive Machine Learning model for this purpose.

There are two main types of predictive Machine Learning models: classification models and regression models.

- Classification model is trained to assign new data points to specific categories based on the analysis of previous data points.
- Regression models, on the other hand, estimate the relationship between input data and output data to predict continuous values.

Predictive Machine Learning models can be further categorized into two types of learning: unsupervised and supervised learning.

- Unsupervised learning involves analyzing data without any prior knowledge of the output variable.
- In contrast, supervised learning involves using labelled data to train the model to make predictions on new, unseen data.

The benefits of predictive Machine Learning models include increased efficiency, better accuracy, and cost savings. This would prevent cold calling all prospective customers of Hyper Climate Change Fund and narrowing down to only those with a greater chance of subscribing to this fund. These models automate

processes that were once done manually, reducing errors, and enabling businesses to make quick, data-driven decisions. They can also identify patterns and insights in data that would be difficult to detect manually. Some other benefits include:

- They can help businesses make better decisions based on data-driven insights.
- They can improve efficiency and accuracy by automating tasks that are otherwise tedious or prone to human error.
- They can enhance customer satisfaction and loyalty by providing personalized recommendations or services.
- They can uncover hidden opportunities or risks by detecting patterns or trends that are not obvious to human eyes.

Predictive Machine Learning models can also be used in areas of credit card fraud detection, customer segmentation, demand forecasting, and image recognition. Banks already use predictive Machine Learning models to detect patterns and fraudulent transactions, while Retailers are using them to forecast demand and optimize supply chains.

- In finance, predictive models can be used for credit scoring, portfolio optimization, fraud detection.
- In healthcare, predictive models can be used for diagnosis, prognosis, treatment recommendation.
- In marketing, predictive models can be used for customer segmentation, churn prediction, cross-selling.
- In e-commerce, predictive models can be used for product recommendation, price optimization, inventory management.
- In education, predictive models can be used for student performance prediction, curriculum design, dropout prevention.

For this project, we used a supervised machine learning approach since we have labelled data indicating the customer's interest level in the Hyper Climate Change Fund- Category. Specifically, we used a classification model to predict which category a customer belongs to as the Target variable (Category 0, 1, 2 and 3) is a categorical feature.

Data Overview and Pre-processing

We started by exploring the dataset to understand the distribution of each feature and the relationships between features and the target class. The dataset provided consists of 2,000 rows and 13 columns, with each row representing a customer and each column representing a feature. It contained numerical variables

and categorical variables. Among the categorical variables are 'CATEGORY' which had four categories and is our Target variable and 'FEAT_2' which we label encoded. The numerical variables are those that describe quantifiable characteristics as numbers while the categorical have a finite number of outcomes such as binary. (ABS, 2020)

Out of all the 13 features, only 'FEAT_9' had missing values- a total of 184 missing values. Also, only 'FEAT_5' had outliers amongst all features with Positive skewness as the Mean- 31.29 was greater than the Median- 5. Our Box plot chart reveals this with this feature having an outlier value 1,000.

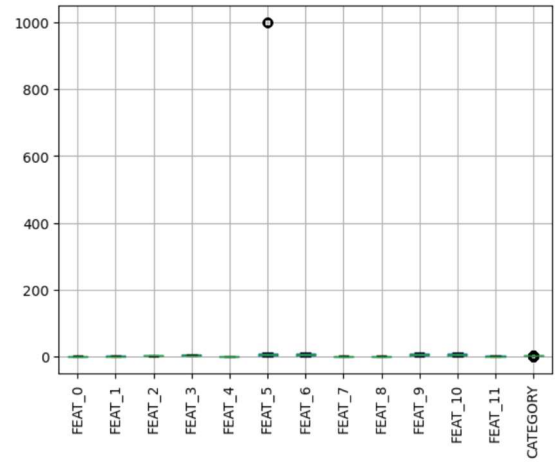

Figure 1- Box Plot before removing Outliers (Jupyter Notebook)

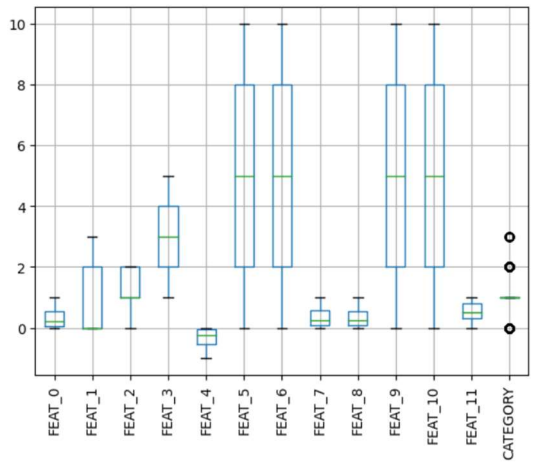

Figure 2- Box Plot after removing Outliers (Jupyter Notebook)

According to edX (2023), "when dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis". We have assumed that the missing values are all at random as we have not been told otherwise. Also, we have an adequate observation pool of 2000 customers, and the missing values amount to only 9.2% of that so remaining observations were assumed enough for our purpose. In dealing with the Outliers, we identified them using the Interquartile range method and were removed from the dataset.

We found that some features have a significant correlation with the target class, while others have weak or no correlation. 'FEAT_6' had the strongest correlation with our target feature- 'CATEGORY'. It was also observed that 'FEAT_4' was perfectly negatively correlated with 'FEAT_0'. This reveals that the variables act in totally opposite directions, as one goes up or increases, the other goes down or decreases.

	FEAT_0	FEAT_1	FEAT_2	FEAT_3	FEAT_4	FEAT_5	FEAT_6	FEAT_7	FEAT_8	FEAT_9	FEAT_10	FEAT_11	CATEGORY
FEAT_0	1.000000	0.017022	-0.012107	-0.009415	-1.000000	-0.078655	0.030349	0.018928	0.014502	-0.047970	-0.037504	0.022824	0.262390
FEAT_1	0.017022	1.000000	0.049550	-0.014508	-0.017022	-0.023025	-0.003848	-0.015095	-0.037842	-0.027081	-0.016835	-0.006033	-0.063525
FEAT_2	-0.012107	0.049550	1.000000	0.015242	0.012107	0.012261	0.005363	0.014656	-0.028487	0.008603	-0.005181	-0.019425	0.309642
FEAT_3	-0.009415	-0.014508	0.015242	1.000000	0.009415	-0.011187	0.006190	-0.035589	-0.025130	0.027652	-0.006040	0.002848	-0.244237
FEAT_4	-1.000000	-0.017022	0.012107	0.009415	1.000000	0.078655	-0.030349	-0.018928	-0.014502	0.047970	0.037504	-0.022824	-0.262390
FEAT_5	-0.078655	-0.023025	0.012261	-0.011187	0.078655	1.000000	-0.016044	0.026817	0.044584	0.033333	-0.004951	0.022340	-0.011611
FEAT_6	0.030349	-0.003848	0.005363	0.006190	-0.030349	-0.016044	1.000000	0.004231	-0.033385	-0.000636	-0.000518	-0.042559	0.468551
FEAT_7	0.018928	-0.015095	0.014656	-0.035589	-0.018928	0.026817	0.004231	1.000000	0.002936	0.018883	-0.045512	-0.006105	0.105105
FEAT_8	0.014502	-0.037842	-0.028487	-0.025130	-0.014502	0.044584	-0.033385	0.002936	1.000000	0.022781	0.028810	-0.015947	0.249564
FEAT_9	-0.047970	-0.027081	0.008603	0.027652	0.047970	0.033333	-0.000636	0.018883	0.022781	1.000000	-0.023743	-0.005698	-0.000890
FEAT_10	-0.037504	-0.016835	-0.005181	-0.006040	0.037504	-0.004951	-0.000518	-0.045512	0.028810	-0.023743	1.000000	0.019101	0.205532
FEAT_11	0.022824	-0.006033	-0.019425	0.002848	-0.022824	0.022340	-0.042559	-0.006105	-0.015947	-0.005698	0.019101	1.000000	-0.334387
CATEGORY	0.262390	-0.063525	0.309642	-0.244237	-0.262390	-0.011611	0.468551	0.105105	0.249564	-0.000890	0.205532	-0.334387	1.000000

Figure 3- Correlation chart (Jupyter Notebook)

Machine Learning modelling

In using our supervised machine learning models, we also established earlier that classification models were adopted. Supervised machine learning models are usually used to classify data and make predictions as the process of identifying these customers is a classification problem with labelled data available, making it possible to use any supervised model. It is important to note that the presence of labelled data is important in the choice of Machine learning model. That said, we used Decision Tree, Random Forest, K Nearest Neighbour and Logistic Regression in our modelling process.

First off, we defined our Target variable ('CATEGORY'), splitting it as the Dependent variable after which we go ahead to split our data into the training and test set. We split the data into a training set and a testing set using stratified sampling, with each category represented equally in both sets. This approach helps to prevent bias in the model and ensures that it is robust across different categories of customers. According to Quang Hung Nguyen et al (2021), Investigation on the performance of models showed that the predictive capability of ML models was greatly affected by the training/testing ratios, where the 70/30 one presented the best performance of the models. Hence, we adopted the 70/30 split.

Decision Tree: It builds a tree-like structure where each node represents a test on a feature, each branch represents an outcome of the test, and each leaf node represents a class label. The decision tree is mostly common owing to its simplicity in nature, making it easy to interpret, understand and visualize. The decision trees are also fast in comparison to other classification models. The model returned an 100% training accuracy but 71.89% test accuracy.

Random Forest: It is based on the idea of combining multiple decision trees and using their majority vote or average prediction as the final output. This model is always ideal when data set is expected to increase in size, as in our case as we expect both Hyper Big Bank existing customers and new customers. It is ideal because it does not rely on a single tree and will not lead to overfitting. It is a type of ensemble learning method that aims to reduce the variance and overfitting of individual trees by introducing randomness in the data and features selection. The model also returned an 100% training accuracy but 80.57% test accuracy.

K Nearest Neighbour: It is based on the idea of finding the most similar data points to a given query point and using their labels or values to make a prediction. Using K as equal to 5, and running 200 train/test splits, this model returned an 82.43% training accuracy with 71.38% test accuracy.

Logistic Regression: This model uses the logistic function to transform the linear combination of the independent variables into a probability value between 0 and 1. The model returned a 78.38% training accuracy and 78.11% test accuracy.

MODEL COMPARISON	Accuracy Score	Error Score
Random Forest	0.81	0.19
Logistic Regression	0.78	0.22
Decision Tree	0.72	0.28
K Nearest Neighbour	0.71	0.29

Figure 4- Model comparison

In evaluating the performance of our model on the testing data, we recommend using a Random Forest Classifier. The reason for this recommendation is that Random Forest has proven to be effective in classification tasks, it is robust to outliers and missing data, it can handle high-dimensional data, and it provides feature importance, which can help us understand the impact of different features on the classification results. Furthermore, it can handle imbalanced classes, which is important in this case, where we have a class (category 0) that we want to predict with high accuracy. Most importantly, the above Model comparison table tells that the Random Forest Classifier model is the best by its Accuracy score- 81%, hence why it is recommended for use to the Global Head of Sales.

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

Feature Importance on model

One of the best advantages of using a Random Forest Classifier is that it can capture complex non-linear relationships and interactions among features and more importantly provide feature importance scores and out-of-bag error estimates. Our feature selection process involved analyzing the importance of each feature using a Random Forest algorithm and selecting the top nine features with the highest importance scores- FEAT_0, FEAT_2, FEAT_3, FEAT_4, FEAT_6, FEAT_7, FEAT_8, FEAT_10, FEAT_11. These features were chosen based on their ability to accurately classify customers into the four categories and their potential usefulness in predicting customer behaviour and preferences. Hence if we are to start collecting Data for less than the current twelve features, then these nine features above are ones we would recommend.

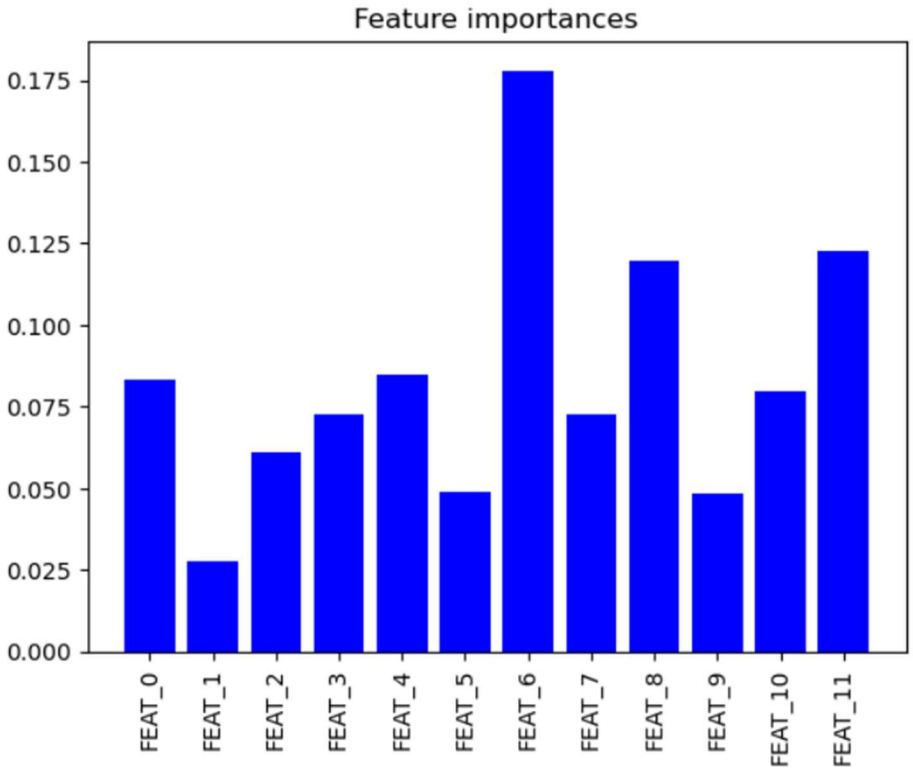

Figure 5- Feature importance for Random Forest

The bar chart plot above reveals that FEAT_6 is the most important of all features, which is also confirmed by the Feature importance chart of other models. Also, we recommend using at least eight of the twelve features provided in the dataset to maintain an accuracy of 75% or higher.

Further insights on Data

In terms of extra insights on collecting future data, we recommend:

- Collecting more data on customer demographics, such as age, income, and location, as these may impact a customer's interest in the Hyper Climate Change Fund.
- Collecting data on customer's investment history to determine whether they are a suitable candidate for the fund.
- Certain features such as occupation, employment status, and account balance will help improve the accuracy of the model.
- Collecting more data on customer behaviour, interests, and preferences, as well as conducting surveys and focus groups to gain insights into customer attitudes towards climate change and related investment opportunities.

All these are especially important in predicting which category a customer will fall into.

Regarding the concerns from our Global Head of Sales on correct classification of Category 0 customers, it may be challenging to achieve 100% classification accuracy on unknown data in Machine learning (which is not unusual) however our model- Random Forest gives the best result toward this also. We have 100% classification of Category 0 (and other categories) on the Training set, while we achieved approximately 70% classification success [73/ (73+32)] on the Test set for Category 0.

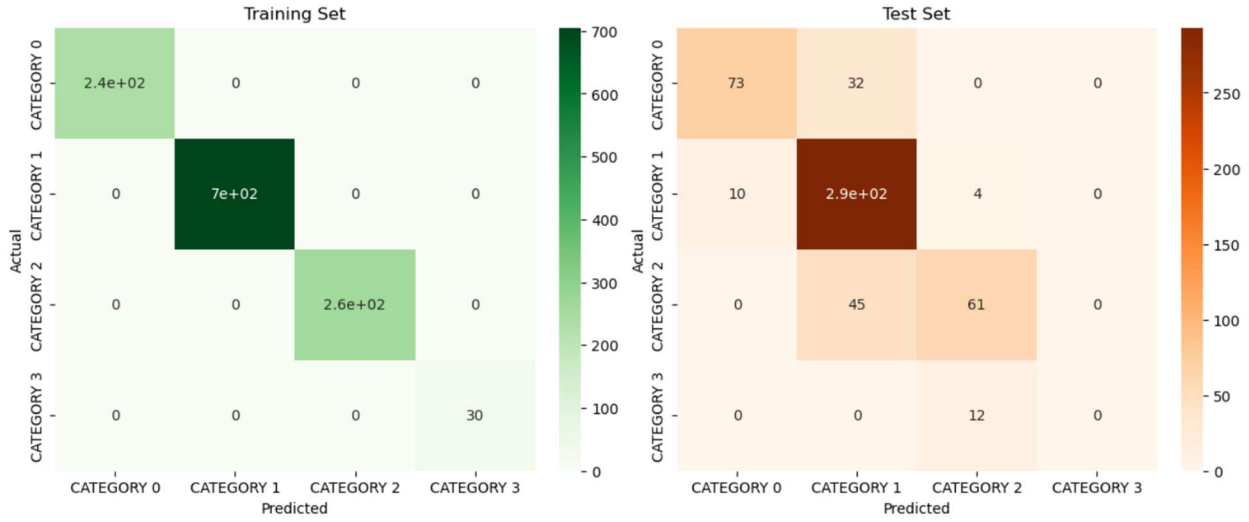

Figure 6- Confusion matrix for Random Forest

Conclusion

Overall, our analysis indicates that it is possible to build a predictive Machine Learning model to classify customers into the four categories of interest as we have done, and we believe that our model will help Hyper Big Bank and in particular, each sales-person or the sales team to effectively target potential customers and improve the success of its marketing initiatives on this new fund product- Hyper Climate Change Fund.

Reference

Alpaydin, E. (2020). Introduction to machine learning (3rd ed.). MIT Press.

Kelleher, J. D., Tierney, B., & Tierney, B. (2018). Data science an introduction. Chapman and Hall/CRC.

Witten, I. H., Frank, E., & Hall, M. A. (2016). Data mining: Practical machine learning tools and techniques. Morgan Kaufmann.

edX (2023). How to Deal with Missing Data- Imputation vs. Removing Data. Available at: https://www.mastersindatascience.org/learning/how-to-deal-with-missing-data/ (Accessed 28/05/2023)

ABS (2020). ABS- Variables [Online] Available at: Variables | Australian Bureau of Statistics (abs.gov.au) (Accessed 05/05/2023)

Alpaydin, E. (2010). Introduction to machine learning (2nd ed.). Cambridge, MA: MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). New York: Springer.

Anna Anisin (2022). Introduction To Machine Learning for Marketing. Forbes. https://www.forbes.com/sites/theyec/2022/11/08/introduction-to-machine-learning-for-marketing/ (Accessed 05/05/2023)

Marketing Evolution (2019). How to Use Predictive Analytics in Data-Driven Marketing. Available at: How to Use Predictive Analytics in Data-Driven Marketing (marketingevolution.com) (Accessed 05/05/2023)

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research, 15(1), 3133-3181.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems (1st ed.). Sebastopol, CA: O'Reilly Media.

Kelleher, J. D., & Tierney, B. (2018). Data science: An introduction (1st ed.). Boca Raton, FL: CRC Press.

Chen, P. Y., & Chen, Y. L. (2019). A deep learning approach to demand forecasting in e-commerce. IEEE Transactions on Industrial Informatics, 15(1), 562-570.

Chen, T., & Guestrin, C. (2016). XG Boost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

Hariharan, S., & Chan, S. (2018). Deep convolutional neural networks for image recognition. International Journal of Computer Applications, 180(26), 41-47.

Langseth, H., & Løland, A. (2019). Predictive machine learning for fraud detection: An empirical comparison of supervised and unsupervised approaches. Journal of Financial Crime, 26(1), 155-167.

Verlinden, J., Baesens, B., & Vanthienen, J. (2017). A comprehensive review of ensemble learning for credit scoring: Recent research and trends. Artificial Intelligence Review, 47(4), 485-513.

Grace Igbinosa (2023). Data Cleaning with Python: Tips and Best Practices. Available at: https://medium.com/@egigbinosa/data-cleaning-with-python-tips-and-best-practices-d31a26e111c (Accessed 09/05/2023)