Machine Learning and Big Data

AN EXECUTIVE REPORT ON PREDICTIVE MACHINE LEARNING MODEL FOR HYPER BIG BANK

Ayomikun Adebawojo

ICMA Centre, Henley Business School
University of Reading

Table of Contents

1. Introduction

This executive report presents the findings and recommendations from the project on predictive modelling for customer classification in marketing the Hyper Climate Change Fund. The study aims to provide useful information on the use of machine learning methods to forecast client categories and direct marketing initiatives.

The Hyper Climate Change Fund, established by the Hyper Big Bank's New Product Development Team, focuses on funding start-up enterprises developing climate-change-responsible solutions. In order to market the fund efficiently, it is critical to evaluate consumers' interest levels and categorise them accordingly. For a large client base, the team intends to apply machine learning models to anticipate customer segments based on attributes.

The major purpose of this project is to assess the feasibility of constructing a predictive machine learning model and how accurate it is at categorising clients. A collection of de-identified consumer data will be used to train and test the model. The dataset has 12 features labelled [FEAT_0, FEAT_1, FEAT_2, ... , FEAT_11] as well as related customer categories labelled [0, 1, 2, 3].

We will look at different strategies and techniques used to develop the predictive model in this study. The method includes stages for data collection and preprocessing, feature engineering and selection, model evaluation and selection, and validation. These techniques ensure that the model is accurate, durable, and can be generalised.

To assess the model's accuracy, we will employ performance evaluation metrics such as accuracy, precision, recall, and F1-score. A sample dataset from the study will be used for the evaluation, allowing us to assess how well the model performs for recognised consumer categories.

The report will include citations to pertinent research papers, methods, and frameworks to give readers a complete knowledge of the strategies employed. The report's References section will include citations to these sources.

The following sections address the methodology, findings, analysis, model suggestions, business implications, and recommendations for further research. It is expected that useful insights into the use of predictive modelling for customer classification and its potential impact on marketing the Hyper Climate Change Fund would be accessible by the end of this research.

2. Methodology

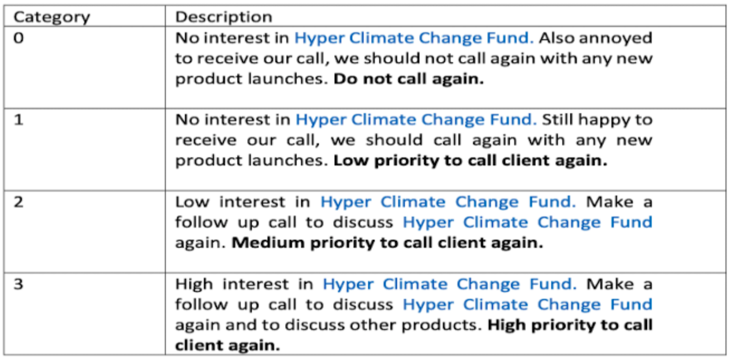
A reliable method was used to predict customer interest in the Hyper Climate Change Fund and divide clients into different categories. The process includes many primary phases, including data preparation, model selection, model training and evaluation, feature selection, and model performance analysis. Each of these techniques is explained briefly in the sections below.

2.1 Data Preprocessing

The first stage involved importing the provided dataset in the form of an Excel (CSV) file that comprised 2000 rows and 13 columns of client information. The dataset was anonymised to protect the customers' privacy. The goal was to find a viable model for predicting client groups based on the customer features provided.

Data visualisation was utilised to help understand the dataset. This included examining the data type, identifying missing numbers and outliers, and understanding how the data was dispersed among the below categories.

Figure 1: Customer Categories



Category	Description
0	No interest in Hyper Climate Change Fund. Also annoyed to receive our call, we should not call again with any new product launches. **Do not call again.**
1	No interest in Hyper Climate Change Fund. Still happy to receive our call, we should call again with any new product launches. **Low priority to call client again.**
2	Low interest in Hyper Climate Change Fund. Make a follow up call to discuss Hyper Climate Change Fund again. **Medium priority to call client again.**
3	High interest in Hyper Climate Change Fund. Make a follow up call to discuss Hyper Climate Change Fund again and to discuss other products. **High priority to call client again.**

Source: Project brief

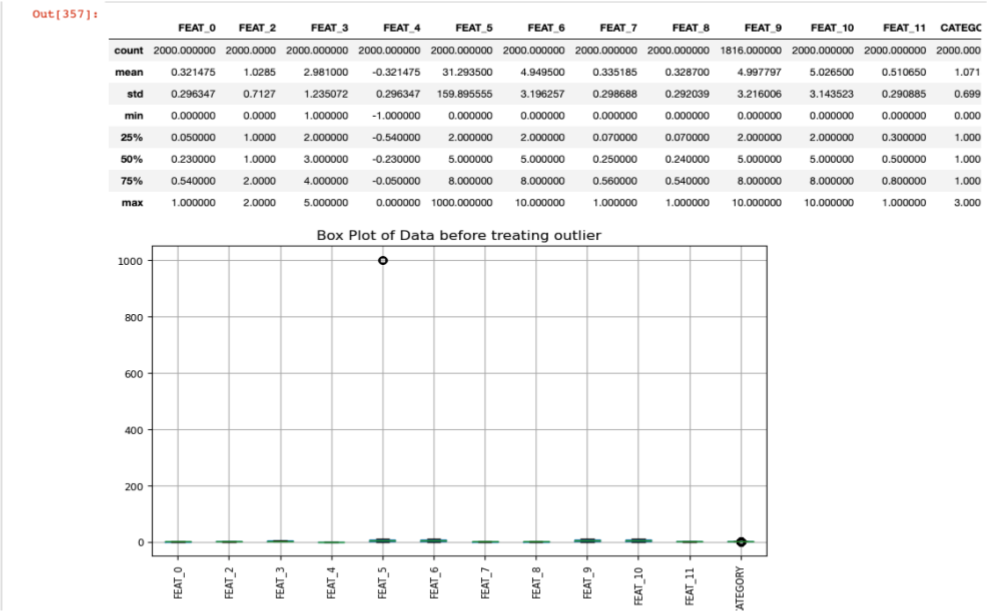
A range of techniques, including summary statistics, box plots, and heat maps, were employed to identify outliers in the dataset. In this work, a box plot was used to identify an outlier in FEAT_5. As a result, the outlier is associated with 53 rows and was removed.

Furthermore, it was determined that FEAT_9 contained 184 missing values, accounting for 9.2% of the dataset. Given that the data was numerical and there were few to no outliers, missing values were handled by substituting the mean value of 5.0.

FEAT_1 was also a categorical feature that needed to be converted into a numerical representation. Using label encoding, the category values (JOB_A through JOB_D) in FEAT_1 were assigned integer values (0 to 3).

After the data preprocessing phases were completed, the dataset included 1947 rows, reflecting the revisions.

Figure 2: Visualisation before treating outlier and missing value

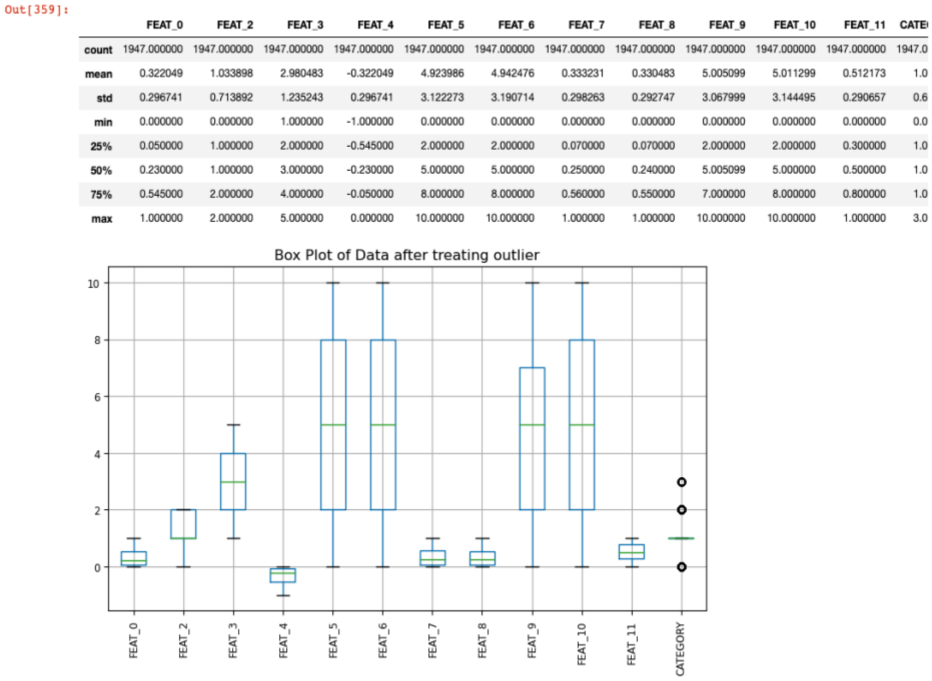


Out[357]:

	FEAT_0	FEAT_2	FEAT_3	FEAT_4	FEAT_5	FEAT_6	FEAT_7	FEAT_8	FEAT_9	FEAT_10	FEAT_11	CATEGC
count	2000.000000	2000.0000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	1816.000000	2000.000000	2000.000000	2000.000
mean	0.321475	1.0285	2.981000	-0.321475	31.293500	4.949500	0.335185	0.328700	4.997797	5.026500	0.510650	1.071
std	0.296347	0.7127	1.235072	0.296347	159.895555	3.196257	0.298688	0.292039	3.216006	3.143523	0.290885	0.699
min	0.000000	0.0000	1.000000	-1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
25%	0.050000	1.0000	2.000000	-0.540000	2.000000	2.000000	0.070000	0.070000	2.000000	2.000000	0.300000	1.000
50%	0.230000	1.0000	3.000000	-0.230000	5.000000	5.000000	0.250000	0.240000	5.000000	5.000000	0.500000	1.000
75%	0.540000	2.0000	4.000000	-0.050000	8.000000	8.000000	0.560000	0.540000	8.000000	8.000000	0.800000	1.000
max	1.000000	2.0000	5.000000	0.000000	1000.000000	10.000000	1.000000	1.000000	10.000000	10.000000	1.000000	3.000

Source: Jupyter Notebook

Figure 3: Visualisation after treating outlier and missing value



Out[359]:

	FEAT_0	FEAT_2	FEAT_3	FEAT_4	FEAT_5	FEAT_6	FEAT_7	FEAT_8	FEAT_9	FEAT_10	FEAT_11	CATE
count	1947.000000	1947.000000	1947.000000	1947.000000	1947.000000	1947.000000	1947.000000	1947.000000	1947.000000	1947.000000	1947.000000	1947.0
mean	0.322049	1.033898	2.980483	-0.322049	4.923986	4.942476	0.333231	0.330483	5.005099	5.011299	0.512173	1.0
std	0.296741	0.713892	1.235243	0.296741	3.122273	3.190714	0.298263	0.292747	3.067999	3.144495	0.290657	0.6
min	0.000000	0.000000	1.000000	-1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
25%	0.050000	1.000000	2.000000	-0.545000	2.000000	2.000000	0.070000	0.070000	2.000000	2.000000	0.300000	1.0
50%	0.230000	1.000000	3.000000	-0.230000	5.000000	5.000000	0.250000	0.240000	5.005099	5.000000	0.500000	1.0
75%	0.545000	2.000000	4.000000	-0.050000	8.000000	8.000000	0.560000	0.550000	7.000000	8.000000	0.800000	1.0
max	1.000000	2.000000	5.000000	0.000000	10.000000	10.000000	1.000000	1.000000	10.000000	10.000000	1.000000	3.0

Source: Jupyter Notebook

2.2 Model Training and Evaluation

The performance of the predicted machine learning model was evaluated using the train-test split method, as is typical for classification or regression problems (Brownlee, 2020). The dataset was divided into training and testing subgroups.

The performance of the machine learning model was assessed using the testing dataset, which contained 389 rows (20% of the original data). The training dataset, which contains 1558 rows (or 80% of the data), was used to train the model.

3. Recommendation of Predictive Machine learning model and accuracy

For predictive modelling, six machine learning classification models have been utilized and assessed using the accuracy score of the test data set. An accuracy score is a metric used to evaluate the performance of a classification model. It displays the proportion of correctly predicted cases out of all the examples in the dataset.

Figure 4: Accuracy of machine learning models

	Accuracy Score	**Error Score**
Logistic Regression	77.18%	22.82%
Logistic Regression Polynomial (2 Order)	84.10%	15.90%
Logistic Regression Polynomial (3 Order)	86.67%	13.33%
KNN	72.82%	27.18%
Decision Tree	70.51%	29.49%
Random Forest	82.82%	17.18%

Source: Compiled from Jupyter Notebook

Considering the accuracy scores for the various models supplied above, we can interpret the findings as follows:

3.1 Logistic Regression:

Logistic regression is extensively used in predictive modelling, particularly for binary classification problems. It is a supervised learning technique that predicts the chance of a given instance falling into a specific class.

The logistic regression model was 77.18% accurate. This suggests that the model predicted the class labels accurately for roughly 77.18% of the occurrences in the dataset.

3.2 Logistic Regression Polynomial (Mapping order 2):

When the features and the target variable have a non-linear connection, logistic regression with polynomial features is highly useful. By incorporating higher-order interactions between the features, the model can capture more complex patterns and improve prediction accuracy. Overfitting should be avoided wherever possible, especially when dealing with tiny or noisy datasets. However, regularisation techniques including varying the degree of polynomial features can help to reduce overfitting and improve generalisation.

The logistic regression model with second-order polynomial features enhanced accuracy to 84.10%. The model was able to capture higher-order interactions between variables by including polynomial features, resulting in improved prediction performance than the usual logistic regression model.

3.3 Logistic Regression Polynomial (Mapping order 3):

The polynomial model is an effective tool for predictive modelling in machine learning. Polynomial features are added to classic logistic regression, allowing it to include non-linear relationships between predictors and the target variable.

The logistic regression model with third-order polynomial features raised the accuracy to 86.67% which is the highest of all the models. The model improved its predicted performance by integrating more polynomial characteristics and identifying more complicated interactions between variables. This can mean that the relationship that exists between the dependent and independent variables is non-linear.

3.4 K-nearest Neighbours (KNN):

K Nearest Neighbours (KNN) is a popular supervised learning technique for predictive modelling applications. This non-parametric technique can be used to address classification and regression problems. The KNN algorithm makes predictions based on how similar the input data points are to their closest feature space neighbours.

Before settling on a class label, the KNN model determines the separation between the input data points and their neighbours by aggregating the votes of the nearest neighbours. The number of neighbours considered for the forecast is determined by the value of k. It is critical to choose the correct value for k because a low value may result in overfitting and a high value may result in bias.

It is advised to carry out data pretreatment operations such as feature scaling and, if necessary, addressing missing values as we did before using the KNN method. Additionally, to enhance the model's performance and lessen computational complexity, dimensionality reduction and feature selection approaches can be used. The model's accuracy in this case was roughly 72.82%.

3.5 Decision Tree:

Using a Decision Tree model, we can capture complicated relationships and make predictions based on a set of features. A decision tree is a popular and basic machine-learning

Missing data are handled automatically, and both categorical and numerical characteristics are supported. To minimise overfitting, hyper-parameters such as maximum depth and a minimum number of data per leaf must be fine-tuned. If decision trees are not appropriately regularised, they tend to overfit the data. Decision trees are hierarchical models that make predictions based on a series of if-then scenarios. In this case, the model's accuracy was approximately 70.51%.

3.6 Random Forest:

Random Forest is a powerful machine learning method for predictive modelling. To generate predictions, this combination of learning techniques combines many decision trees.

Random Forest's strengths include the ability to handle both category and numerical features, to handle complex correlations and nonlinearities in data, and to provide feature priority rankings. It is often used for a wide range of predictive modelling tasks, including classification and regression, and it can be an effective tool for developing trustworthy and accurate predictive models. The random forest model had an accuracy rate of 82.82%.

The logistic regression model with polynomial features of the third order (86.67%) had the maximum accuracy when compared to other models. This is closely followed by the logistic regression model with polynomial features of the second order (84.10%) and the random forest model with an accuracy of 82.82%. These methods outperform the decision tree model, KNN, and basic logistic regression model in terms of predictive performance.

There are various benefits to using third-order polynomial features in logistic regression in some circumstances. We can understand the relationship between features and the goal variable since the results are interpretable. It can represent non-linear interactions by introducing polynomial characteristics, giving it more flexibility than linear decision boundaries. Complex models, such as neural networks or random forests, are computationally inefficient and less easy than logistic regression. Also, regularisation techniques assist it in managing more complicated relationships without overfitting. The model gives feature importance through coefficients to aid in feature selection. With small datasets, logistic regression performs well and captures non-linear correlations quickly.

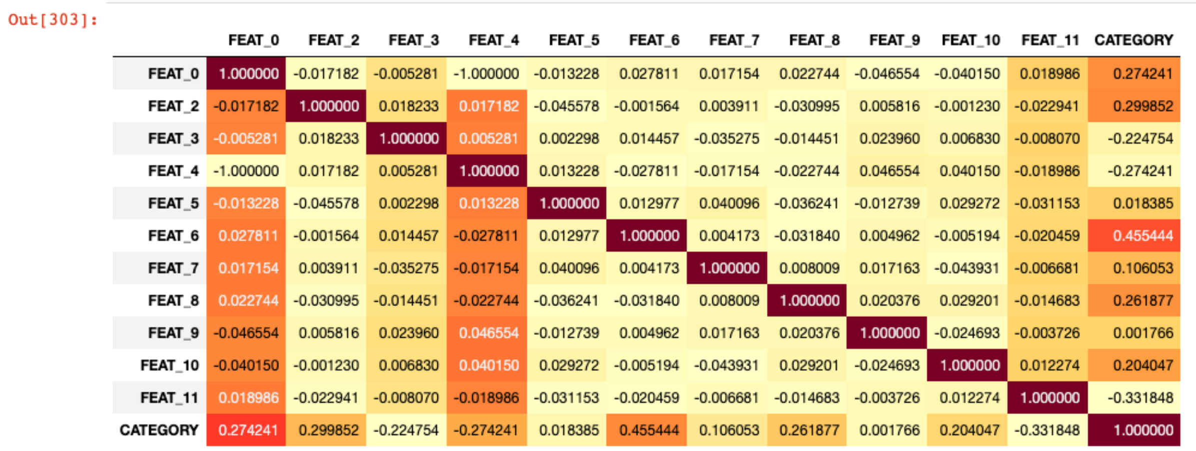
Therefore, based on the accuracy score and the above-stated advantages of the model, the logistic regression model with polynomial features of the third order is recommended to be used for classification.

4. Feature importance

4.1 Features recommendation

It is critical to understand the potential for the prediction of various variables in a machine-learning model. By examining the connection between features and the target variable, we may identify the critical elements that have a major impact on the predictions. In the context of our effort, the Hyper Climate Change Fund, we conducted a feature importance study and obtained some intriguing results.

Figure 5: Heatmap Correlation



Out[303]:

	FEAT_0	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	-0.017182	-0.005281	-1.000000	-0.013228	0.027811	0.017154	0.022744	-0.046554	-0.040150	0.018986	0.274241
FEAT_2	-0.017182	1.000000	0.018233	0.017182	-0.045578	-0.001564	0.003911	-0.030995	0.005816	-0.001230	-0.022941	0.299852
FEAT_3	-0.005281	0.018233	1.000000	0.005281	0.002298	0.014457	-0.035275	-0.014451	0.023960	0.006830	-0.008070	-0.224754
FEAT_4	-1.000000	0.017182	0.005281	1.000000	0.013228	-0.027811	-0.017154	-0.022744	0.046554	0.040150	-0.018986	-0.274241
FEAT_5	-0.013228	-0.045578	0.002298	0.013228	1.000000	0.012977	0.040096	-0.036241	-0.012739	0.029272	-0.031153	0.018385
FEAT_6	0.027811	-0.001564	0.014457	-0.027811	0.012977	1.000000	0.004173	-0.031840	0.004962	-0.005194	-0.020459	0.455444
FEAT_7	0.017154	0.003911	-0.035275	-0.017154	0.040096	0.004173	1.000000	0.008009	0.017163	-0.043931	-0.006681	0.106053
FEAT_8	0.022744	-0.030995	-0.014451	-0.022744	-0.036241	-0.031840	0.008009	1.000000	0.020376	0.029201	-0.014683	0.261877
FEAT_9	-0.046554	0.005816	0.023960	0.046554	-0.012739	0.004962	0.017163	0.020376	1.000000	-0.024693	-0.003726	0.001766
FEAT_10	-0.040150	-0.001230	0.006830	0.040150	0.029272	-0.005194	-0.043931	0.029201	-0.024693	1.000000	0.012274	0.204047
FEAT_11	0.018986	-0.022941	-0.008070	-0.018986	-0.031153	-0.020459	-0.006681	-0.014683	-0.003726	0.012274	1.000000	-0.331848
CATEGORY	0.274241	0.299852	-0.224754	-0.274241	0.018385	0.455444	0.106053	0.261877	0.001766	0.204047	-0.331848	1.000000

Source: Jupyter Notebook

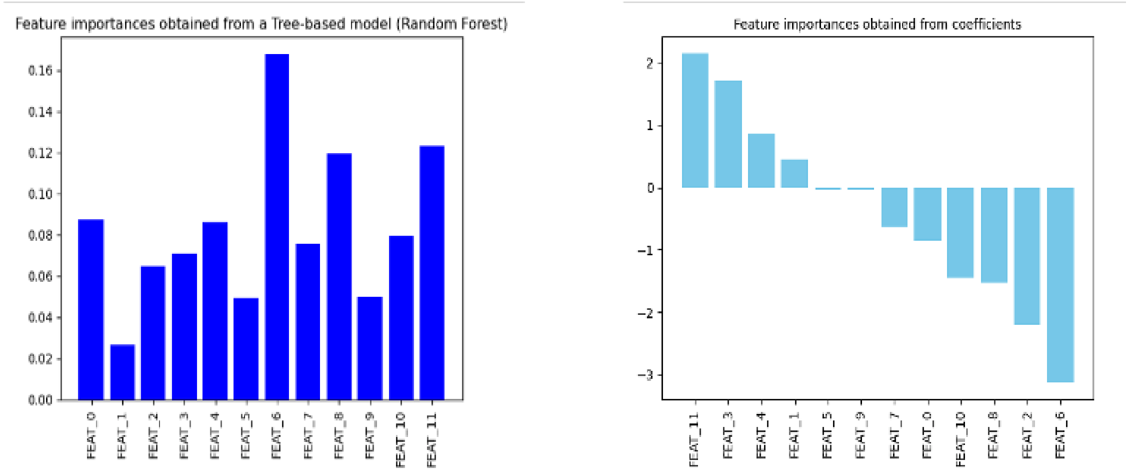
According to the heatmap analysis, the target variable (CATEGORY) and most attributes had a strong connection. FEAT_1, FEAT_5, and FEAT_9, three other features, had significantly lower relationships with the target. As a result, these traits may have less of an impact on a customer's decision to invest in the Hyper Climate Change Fund.

According to the analysis, customers with FEAT_11 had the strongest negative relationship with the target variable. This implies that purchasers with this specific characteristic are less likely to be interested in the fund offering. Furthermore, FEAT_3 and FEAT_4 found negative associations, lending credence to the notion that these features may indicate a lack of interest in the product.

A limited number of features, on the other hand, demonstrated significant positive associations with the target variable. Some of these traits include FEAT_0, FEAT_2, FEAT_6, FEAT_7, FEAT_8, and FEAT_10. According to the positive correlation, customers who possess these characteristics are more likely to be interested in the Hyper Climate Change Fund. As a result, acquiring information about these characteristics would provide insightful information and assist the model in producing credible predictions.

Other approaches to feature selection, such as the analysis of feature relevance derived from coefficients and a tree-based model like random forest, have also proven the value of these traits. These models also highlighted the importance of FEAT_0, FEAT_2, FEAT_6, FEAT_7, FEAT_8, and FEAT_10 in projecting client interest.

Figure 6: Feature Importance Chart



Source: Jupyter Notebook

4.2 Minimum features of 75% accuracy

Using the findings of our Jupyter predictive model, we calculated the variance of each feature in the dataset. The variance values for each feature are as follows:

- FEAT_0: 16.874
- FEAT_1: 9.309
- FEAT_2: 9.083
- FEAT_3: 8.812
- FEAT_4: 8.600
- FEAT_5: 8.354
- FEAT_6: 8.244
- FEAT_7: 7.917
- FEAT_8: 7.908
- FEAT_9: 7.553
- FEAT_10: 7.347
- FEAT_11: 0.000

Feature selection is critical in machine learning for developing an accurate and efficient predictive model. It aids in identifying the most important features that significantly contribute to the model's performance while decreasing noise and computational complexity.

We can consider the features with the highest variation to estimate the minimum features required to reach a 75% accuracy level. The variation or variability of data points within a feature is measured by variance. Generally, features with higher variances have more significant variations between their values, indicating potential discriminative power for classification.

Based on the stated variance values, we can prioritise the features with larger variances because they may have a greater selective capacity in our model. FEAT_0 (16.874), FEAT_1 (9.309), and FEAT_2 (9.083) have significantly greater variances. More variation in the values of these variables indicates that they may include valuable classification-related information.

Previous research such as (Guyon, I., & Elisseeff, A. 2003) and (Chandrashekar, G., & Sahin, F. 2014) has also stressed the importance of feature selection in machine learning models. Techniques for feature selection help to improve model performance by improving interpretability, lowering overfitting, and increasing model performance. We aim to collect relevant data that will help us accomplish our target accuracy level by taking larger variation variables into account.

Finally, we propose using FEAT_0 to FEAT_7 as the minimum features to achieve a 75% accuracy requirement based on variance values. These factors have bigger variances, which may indicate that they are better at predicting distinct sorts of consumers. Instead of employing all 12 features, choosing fewer but more essential characteristics can help deal with data complexity and avoid overfitting because the filtered features have no significant impact on the accuracy score. This would also help Hyper Bank save money by lowering staff and data collection costs.

5. Limitations on the study and potential areas of future research

While working on the model prediction project for customer classification, the following limitations and potential research topics emerged:

5.1 Ability to Interpret and Explain: The findings from the selected machine learning models might be quite accurate, but their capacity to be understood and explained may be limited. Understanding the factors affecting the model's predictions is necessary to developing confidence and ensuring regulatory compliance. Using explainable AI techniques, such as feature importance analysis or model-agnostic approaches, would increase the model's transparency and interpretability (Angwin et al., 2016).

5.2 Continuous Model Monitoring: As a result of changes in consumer behaviour, market dynamics, or the introduction of new items, the predictive model's performance might decrease over time. To keep the model accurate and current, it must be tested for effectiveness on a regular basis and updated with new data. To ensure long-term success, future research should focus on developing systems for continuous model monitoring and adaptation.

5.3 Ethical Concerns: Using predictive modelling to categorise clients raises concerns about privacy, data security, and potential algorithmic flaws. Addressing these issues of ethics and ensuring that relevant laws and policies are followed are critical (Bellamy et al., 2019). Future study should look on techniques for evaluating and decreasing distortion in the model's predictions in order to ensure fairness and avoid discriminatory outcomes.

5.4 Feature Engineering: The current study classified clients using a specified set of features. However, the analysis could have overlooked some important details. The model's prediction capability could be improved by looking at additional data sources or including domain-specific information (Caruana et al., 2015). Future research could consider the inclusion of cutting-edge approaches such as natural language processing on data from user comments (text mining).

5.5 Limited Dataset: Only 2,000 of the 3,000 customer records in the study's dataset were made available for research. The dataset's size and scope may not effectively reflect the variety of customer preferences and traits, which may affect the applicability of the model's predictions are (Caruana et al., 2015). A larger and more diverse dataset may be valuable for future research to improve model performance and resilience.

6. References

Angwin, J., Larson, J., Mattu, S. and Kirchner, L., 2016. *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks*. ProPublica.

Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K. and Varshney, K.R., 2019. 'AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias'. *IBM Journal of Research and Development*, 63(4/5), pp.1-15.

Brownlee, J., 2020. Train-Test Split for Evaluating Machine Learning Algorithms. Machine Learning Mastery. Available at: https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/ (Accessed: 09 May 2023)

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. and Elhadad, N., 2015. *Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission.* Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1721-1730.

Chandrashekar, G. and Sahin, F., 2014. *A Survey on Feature Selection Methods*. Computers & Electrical Engineering, 40(1), pp.16-28.

Guyon, I. and Elisseeff, A., 2003. 'An Introduction to Variable and Feature Selection'. *Journal of Machine Learning Research*, 3, pp.1157-1182.