Data Science

Capstone Report - Spring 2022

Deep Dive in Customer Lifecycle

Gong Liancheng
Bai Xue
Zhang Tianyu

supervised by
Guo Li and Gong Xinyi

Preface

The customer lifecycle is always a thorny problem for banks or customer-based enterprises. This project is inspired and supported by HSBC. We are interested in this topic and want to discover how much we can contribute to such real cases as a "data scientist". The main target audience is the wealth advisors and data analysis team. We start from personal and transaction data into some business insights that not only provide new ideas to the data analysis team but also bring more recommendations to wealth advisors

Acknowledgements

This thesis would not come into being if we had not received help from our supervisor, professors, classmates, and friends. They offered us so much encouragement. But before we express our sincerest gratitude to them, we would thank our family first. During the COVID-19 pandemic, everyone has gone through difficult times. It was our family that always be here giving the support to us, without which we could have already given up.
Our professor Li Guo offered us so much help that we feel so grateful to her. She gave us advice on data processing and modeling. Every week's meeting pushed and encouraged us to move forward. And our supervisor Xinyi Gong offered us a lot of help as well. Without her help with data understanding and guidance on business interpretation, we could never finish this project.
Moreover, we would like to extend our gratitude to our friends Christopher Bang, Tooru Oikawa, Haiji Kiyose, and Zhiyang Wang. They were here for us during the hardest time, giving us support and encouragement.

Abstract

In this capstone project, we offered a number of strategies to help HSBC segment customers, predict inactivity, and make recommendations accordingly. These issues are challenging to handle because of their tight ties to corporate and consumer diversity. We propose three phases to the problem: 1) We will calculate customer values based on their personal information; 2) We will then create a customer Lifecycle and divide customers into three stages: developing, honeymoon, and retention. 3) Based on the behaviors and preferences of different consumer phases or segments, we will recommend customized products and services to them. We are able to create a high-performing machine learning prediction model, segment customers appropriately, and a business-relevant recommendation method.

Keywords

Capstone; Data science; NYU Shanghai; Lifecycle; HSBC; Value Segment; Machine Learning; RFM k-means; XGBoost

Contents

1 Introduction

HSBC (The Hong Kong and Shanghai Banking Corporation Limited) is a world-renowned bank serving a wide variety of customers. It provides a well-rounded chain of products and services, including mortgages, credit cards, saving accounts, investments, and insurance. Its key component is to grow business by attracting and keeping customers. To achieve this, it is essential to understand the customer lifecycle and personalize services for customers in their lifecycle. Customer lifecycle describes the stages that customers go through before, during, or after their transactions and hopefully they establish a loyal relationship with the bank.

The project will personalize the service to high-value customers for HSBC through the analysis of the customer's value and life stage. We will have three phases in this project. In the first phase, we will calculate customer values according to their personal information, purchase power, and consumption habit. Then we design a customer lifecycle and assign customers to three different stages we designed: the Developing Stage, the Honeymoon Stage, and Retention Stage in the second phase. In the final phase, we will recommend personalized products and services to different customer stages or segments according to their stage habits and preferences.

2 Related Work

2.1 Customer Life stages

Reach, acquire, develop, retain, and advocate are the five stages of the customer lifecycle [1]. Customers must be informed of the bank's goods during the reaching stage. It is regarded as the most crucial and costly stage, as advertisements must be placed in the most appropriate location for target clients [2]. Segmentation is a technique for grouping customers and identifying potential purchasers based on previous data [3]. In the second stage, customers are considering and inquiring about the products. As a result, prediction can be utilized to tailor service and adjust things to each unique customer [4]. The acquisition is completed successfully in the third stage. The next step is to keep existing clients. To achieve the goal, several strategies are utilized, including estimating the churn rate [5]. Customers in the last stage are not only loyal but also advocate for and extend the client base.

Following the identification of customers in their proper stage, the next critical step is to recommend appropriate items to various customers in order to enhance customer loyalty. Knott, Hayes, and Neslin describe how to create a recommendation model [6]. The procedure is broken

down into four parts in the paper: picking variables, selecting a model, evaluating the model, and scoring consumers. The procedure can be used as a guideline in our project.

2.2 Logistic Regression

Logistic regression is useful for expressing and evaluating ideas concerning categorical outcome variable relationships [7]. Considering our first goal is to put customers into different segments, which is an iconic categorical outcome. Also, Logistic Regression is a simple model that we can start with, diving into more complex inner-variables relationships with the following other models.

2.3 Decision Tree

A Decision Tree is a divide-and-conquer classification strategy that may be used to mine features and extract patterns from big databases [8]. The classes in a Decision Tree are mutually exclusive, and the result is a mapping between attribute values and classes [9]. The interpretability of the built model is one advantage of decision tree modelling over other pattern recognition algorithms [8].

2.4 Random Forest

Random Forest is an ensemble method that uses bootstrap aggregation to create numerous Decision Trees [10]. It also includes a feature selection method that allows it to handle multiple input parameters while maintaining all dimensions [10].

2.5 XGBoost

XGBoost is a scalable tree boosting machine learning technique that is commonly used in data mining [11]. XGBoost uses out-of-core processing to allow data scientists to handle hundreds of millions of records in a time and storage efficient manner [11].

2.6 RFM method

Customer lifetime value (CLV) is commonly used to determine which customers are valuable and loyal, as well as which methods to employ. Customers' lifetime value is frequently estimated using the RFM approach. The letters R and F stand for recency, or the time since the last transaction, and frequency, or the number of purchases made in a certain period. M stands for

monetary, which refers to the amount of money spent over time [12]. Customers with similar RFM values are more likely to be at the same stage. However, according to Liu and Shih's study, it is more useful among more committed consumers, which in our instance may be HSBC Premier and Jade customers [12]. Association rules for identifying the difference in customer personal information can be used alongside CLV. This is to better decide important patterns for clustering customers so that more tailored strategies can be made for them. After clustering customers with RFM method, features from their characteristics profile will be looked at to provide better service to different customers [13].

2.7 K-means

The K-means algorithm(KMA) is the most basic and widely used among iterative and hill-climbing clustering algorithms. This approach may result in a suboptimal partition [14]. Because stochastic optimization algorithms are good at preventing convergence to a locally optimal solution, they could be used to find a globally optimal solution. KMA will be used to group customers with similar CLV together.

2.8 Summary

Reading all the literature, we will first apply models to predict the possibility of attrition, and then use it as a feature to cluster similar customers into the same segment.

3 Solution

3.1 Data Analysis and Visualization

The raw data contains 13 txt worksheets from the HSBC system, describing three dimensions of the customers and related product information: Customer personal information, Product information, and Credit card transaction information.

The raw data columns were separated by "|" and rows were separated by the line break. So we first process all worksheet data to DataFrame for future use.

We have 300039 customers' information in total. The first problem with this dataset is that it is unbalanced on the feature of "Attrition". Among 300039 customers, there are only 20% attrition customers. We have much more Nonattrition customers, which will bring errors and inaccuracy in the prediction model. But we decided to keep all the data even though it is unbalanced since

the real-world distribution rule is like this.

	Amount	Percentage
Attrition	60744	20.25%
Non Attrition	239295	79.75%

Table 1: Raw data classification

To clean the wrong and missing data in the "Age" feature, we put all people with the age higher than 100 in the group that has ages between 70 and 100, and all people with negative ages in groups 0-18. From Figure 1 below we tell that the age of most of the customers is in the range of 26 to 70. Age is positively correlated with the inactive rate.

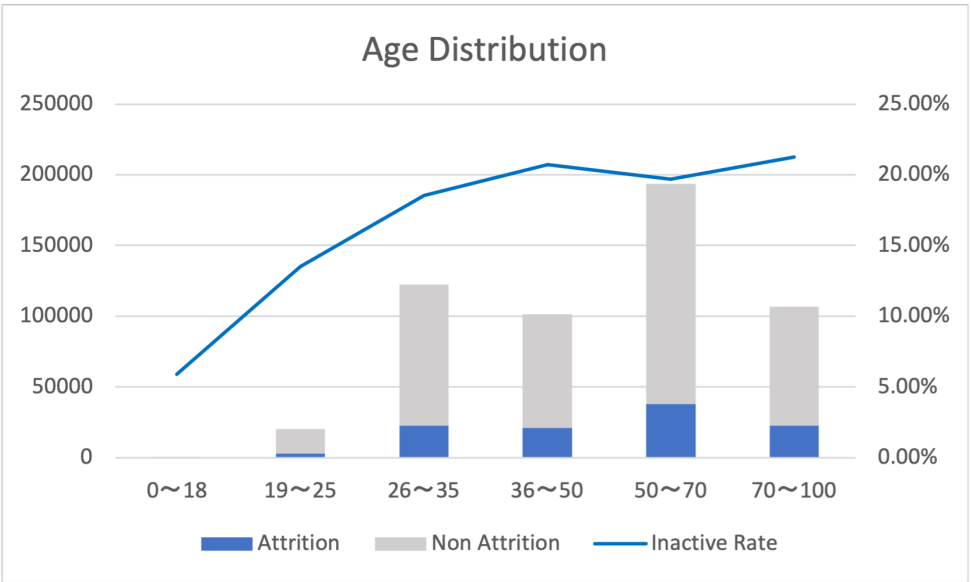

Figure 1: Age Distribution

Another significant feature we choose is "Years after first contact". Learning from the RFM model, we realized the significance of the recency in the customer value. Here we decide to use the year after first contact to represent this feature. The nearest year is 2018, so we use 2018 minus the years of each customer's first contact with the bank to get the recency. From Figure 2 we can see that most of the customers have a lifetime between 0 and 2 years.

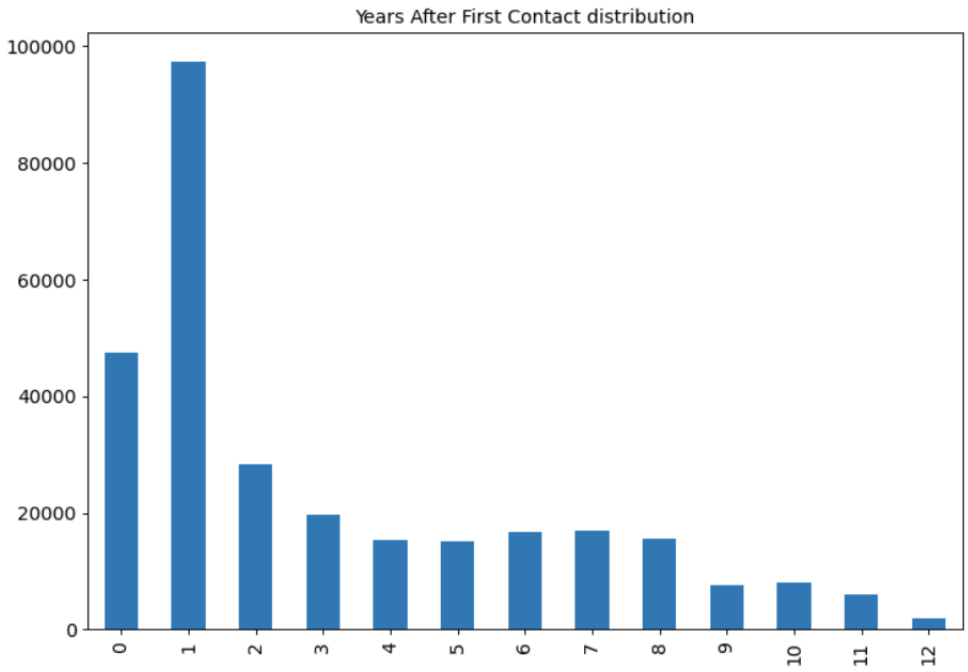

Figure 2: Years After First Contact Distribution

3.2 Merging Data

Considering that directly merging all tables will bring redundancy and difficulty in model running, we decided to select only part of the data to merge for future analysis.

For the Customer Information, we selected age, PBK_Ind, Gender, YearsAfterFirstContact, InterCorpACIndicator, Cust_Segment, and Number_Children. We made Gender and Cust_Segment into a binary variable, specifically, we wanted to predict the probability of attrition, so we made "Cust_Segment" into two variables: "Attrition" and "Not Attrition".

Among 300039 customers, only 28636, less than 10% of the customer purchased products. This group of customers is too niche and it would generate biased results and errors using a small sample size. But we still wanted to focus on the customers with higher value, which have more transaction activities in the bank. So finally we decided to merge the customer info table and Credit Card table, with selected features.

3.3 Machine Learning Classification Models

3.3.1 Data Preparation

From the previous data merging part, we have 300,039 data in total. However, we focus on the customers with credit card records. We now filtered into 83578 in total for the binary classification

of whether the customer is attrition or not. After the data preprocessing part, we now have 44 variables, including age, YearsAfterFirstContact, InstallmentTransCounts, etc.

Our object, y dataset, is AttritionOrNot where 1 represents Attrition and 0 represents not Attrition. 43 variables except AttritionOrNot form our x dataset.

The dataset was split into 80% training dataset (66862) and 20% testing dataset (16716). The training dataset helps train our models while the testing dataset is used to test the accuracy of our models.

We also standardize 43 features by removing the mean and scaling to unit variance. In this case, our features are balanced and satisfy the normal distribution.

3.3.2 Four models

After data preparation, we start to fit the processed data into four classification models, including the Logistic Regression, Decision Tree, Random Forest, and XGBoost. As stated in the Related Work section, our models are first selected because they satisfy the classification task and return the probability of Customer Attrition. Since combining more models can reduce the bias, we select the four models with the highest accuracy. We haven't tuned many parameters when the overall accuracy is high enough.

The main criteria for our model are the accuracy, i.e. the percentage of true prediction. Besides, we don't do oversampling or undersampling to keep the imbalanced data in real cases. In this case, we didn't do the oversampling or undersampling. Instead, we use the confusion matrix to check whether our model does predict the Attrition of customers. We also use ROC Curve and AUC to evaluate the performance of the four models.

In addition to the evaluation, feature importance helps us in future feature selection and threshold settings. All evaluations and the comparison of the four models will be discussed in the Result section.

3.3.3 Probability analysis

Each of the four models will predict individual customers whether attrition or not. At the same time, the model will return the attrition probability where a higher probability represents the customer more likely to become attrition. In other words, the probability approaching 0 means that the customer is least likely to become attrition. We will have four probabilities for each customer since we have four classification models. The four probabilities are "LR_Prob",

"DT_Prob", "RF_Prob", and "XGB_Prob" in Figure 3. In this case, we calculated the average of four probabilities to minimize our bias shown as "average_probability" in Figure 3.

In addition to the continuous value of probability, we categorized those values into integers 1 to 10 shown as "average_category" in Figure 3. When the probability falls into the range from 0 to 0.1, then it is categorized into 1. When the probability falls into the range from 0.1 to 0.2, then it is categorized into 2. So on and so forth. In this case, the average of four probabilities will be categorized into 1 to 10. This predicted category will help to verify the rationality of this classification in the result section when analyzing the trend. Besides, to look at the trend in more detail, we also cut probability into 40 pieces shown as "detailed_category" in Figure 3. Based on the "average_category", we use k-means to cluster into four customer value segments shown as "level" in Figure 3. Those high-value customers are our target and will be assigned stages in the next RFM Clustering section.

Customer_id	LR_Prob	DT_Prob	RF_Prob	XGB_Prob	average_probability	average_category	detailed_category	level
500999800895910	0.003083	0.010319	0.00	0.009772	0.005793	1.0	1.0	high value
500999800779783	0.199561	0.270270	0.68	0.291658	0.360372	4.0	15.0	relatively low value
500999800757376	0.284355	0.105395	0.04	0.168539	0.149572	2.0	6.0	relatively high value
500999800756005	0.685026	0.250000	0.75	0.645293	0.582580	6.0	24.0	low value

Figure 3: distribution of the predicted categories

3.4 RFM Value

After predicting the attrition rate of customers, we picked out the high-value customers with lower attrition probabilities and further clustered them into different stages. We used the weighted RFM method and chose the weights according to Liu and Shih's paper [12]. While R is usually the period since the last purchase and the data we have is the transaction history from 2017/07 to 2018/06, the last purchase date may not be able to distinguish customers. Therefore, we replaced it with years after the first contact. We used the number of spending from credit cards as F and the total amount of spending from credit cards as M. According to the feature importance of XGBoost, these three features are the most significant ones, which shows that this method could be valid in addressing the segmentation problem.

The R, F, and M values are normalized separately that x' = (max - x) / (max-min) where x is the original values, max and min are the largest and smallest R, F, and M values among all customers. RFM values are then multiplied by relative weights 0.731, 0.188, and 0.081 using the

values from Liu and Shih's paper [12].

4 Results

4.1 Classification Model Evaluations

To evaluate the four classification models, we use several criteria including confusion matrix, testing accuracy, and ROC Curve (AUC).

4.1.1 Confusion Matrix

	True	**False**
Positive	15919	54
Negative	701	42

Table 2: Confusion matrix for Logistic Regression

	True	**False**
Positive	15930	43
Negative	606	137

Table 3: Confusion matrix for Decision Tree

	True	**False**
Positive	15890	83
Negative	535	208

Table 4: Confusion matrix for Random Forest

According to Table 2 to Table 4, we can find the data is imbalanced and there are more positives than negatives. We keep that imbalance in the real cases. Besides, the predicted results between Decision Tree and Random Forest are the same. The detailed accuracy of the four models will be discussed in the next section.

	True	False
Positive	15912	61
Negative	558	185

Table 5: Confusion matrix for XGBoost

4.1.2 F1-score (accuracy)

	non-attrition	attrition	macro avg	weighted avg	accuracy
Logistic Regression	0.98	0.12	0.55	0.94	0.96
Decision Tree	0.98	0.31	0.65	0.95	0.96
Random Forest	0.98	0.31	0.65	0.95	0.96
XGBoost	0.98	0.35	0.67	0.95	0.96

Table 6: F1-scores for the four models

	non-attrition	attrition	recall	precision	auc	accuracy
Logistic Regression	0.98	0.12	0.954	0.934	0.526	0.95
Decision Tree	0.98	0.31	0.961	0.954	0.590	0.96
Random Forest	0.98	0.31	0.963	0.956	0.637	0.96
XGBoost	0.98	0.35	0.962	0.956	0.622	0.96

Table 7: F1-scores for the four models

Table 5 shows the F1 score for the four models. To better evaluate model performance, F1-score is a helpful metric that considers both precision and recall. We have F1 in the following formula where TP is True Positive, FP is False Positive, FN is False Negative: $F1 = 2\frac{Precision*Recall}{Precision+Recall} = \frac{TP}{TP+\frac{1}{2}(FP+FN)}$

The non-attrition columns and attrition columns are the scores for each class. However, instead of having multiple F1-scores, we mainly use the average score to describe overall performance. The macro average is the most straightforward one in that it takes the arithmetic mean of all the per-class F1 scores. The weighted average is calculated by the weighted mean of all the per-class F1 scores considering the number of each class. Accuracy computes the proportion of correctly classified observations out of all observations, which is the indicator we focus mostly on.

Within the expectation, the Decision Tree has the same result as Random Forest. The accuracy is all 0.96, which already shows good model performance. However, the score of attrition is low for Logistic Regression, which leads to a lower macro average score and weighted average score. It means that Logistic Regression has worse performance than the other three models when predicting attrition.

4.1.3 ROC/AUC

Figure 4 is the ROC Curves for the four models.

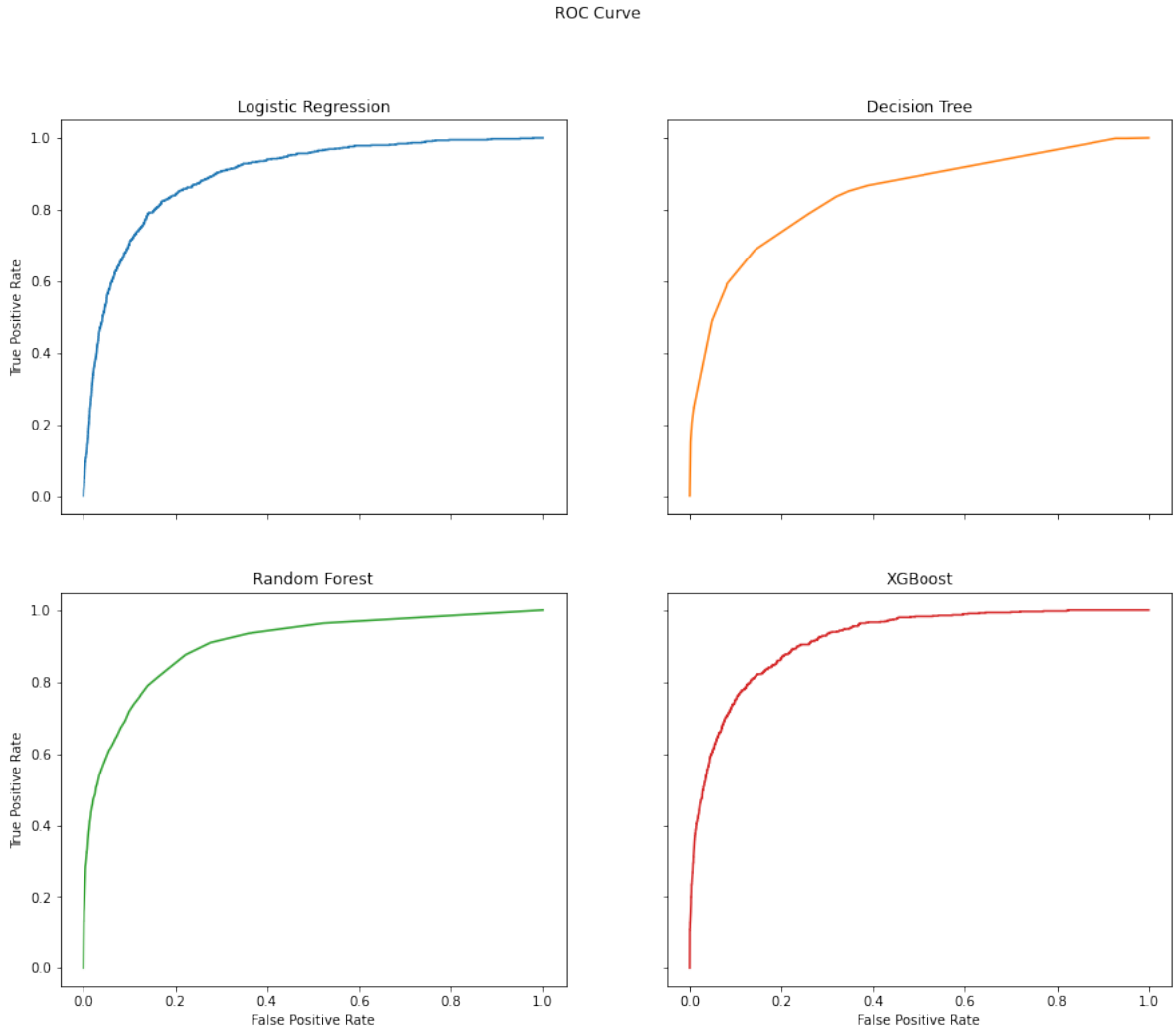

Figure 4: ROC Curves for four classification models

Receiver operating characteristic curve (ROC Curve) has a False Positive Rate as the x-axis and a True Positive Rate as the y-axis. It is a performance measurement for classification problems at various threshold settings. Area Under the Curve (AUC) represents the measure of separability and model performance. From the Figure, Decision Tree has a slightly smaller AUC than others. Overall speaking, all models have good performance.

4.1.4 Feature Importance

Feature importance refers to a class of techniques for assigning scores to input features to a predictive model that indicates the relative importance of each feature when making a prediction.

Instead of telling absolute important and non-important features, we analyze the relative importance under comparison. Those relatively important features are crucial in our future analysis. The following bar charts show the feature importance of the four models.

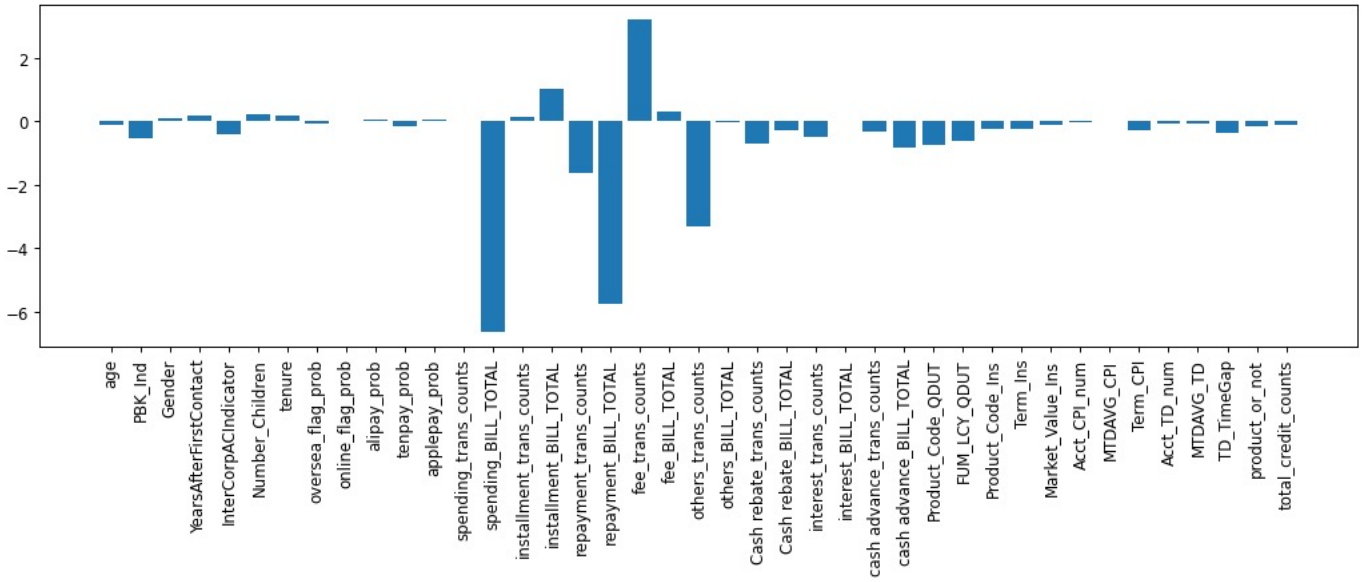

Figure 5: feature importance of logistic regression

Like Figure 5, the parameters in logistic regression are positive and negative. The positive scores represent the feature importance that predicts attrition (label 1) and the negative scores represent the feature importance that predicts non-attrition (label 0). From the graph, we can find spending_Bill_Total, repayment_Bill_Total, and others_trans_counts are obviously important to predict non-attrition. On the other hand, fee_trans_counts in logistic regression is important to predict attrition.

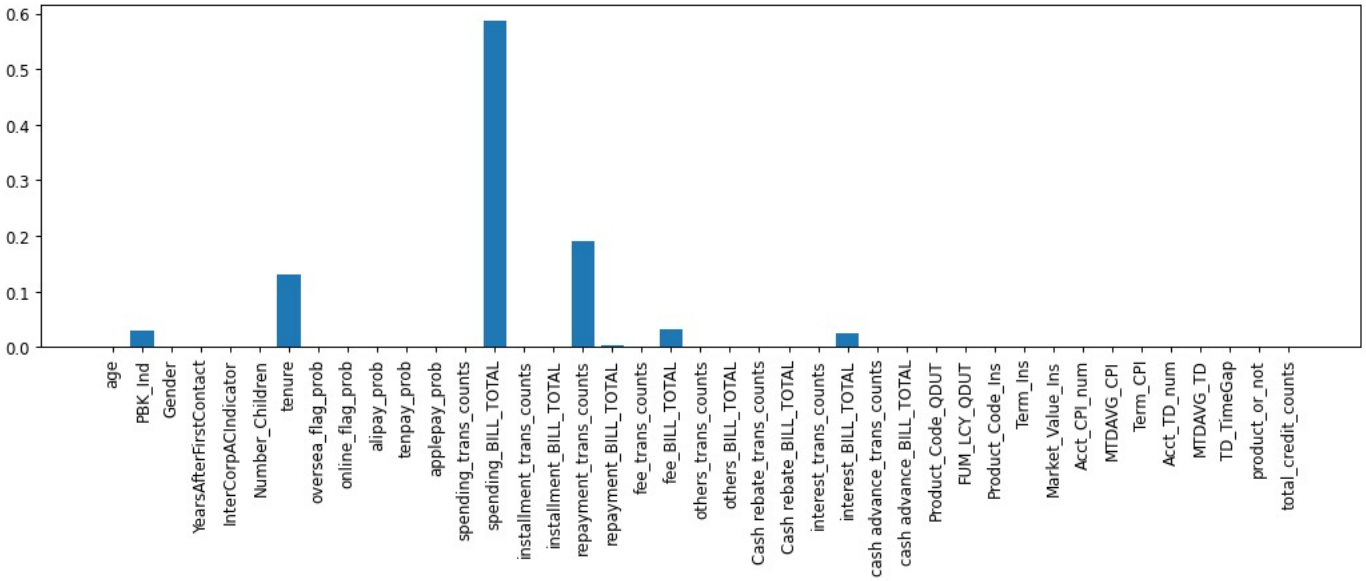

Figure 6: feature importance of decision tree

In Figure 6, we can find that spending_BILL_TOTAL is the most important feature in the decision tree model to predict the attrition of customers. The tenure (similar to YearsAfter-FirstContact) and repayment_trans_counts are relatively important to predict the attrition of customers. The PBK_Ind, fee_BILL_TOTAL, and interest_BILL_TOTAL are parameters relevant to predicting the attrition of customers.

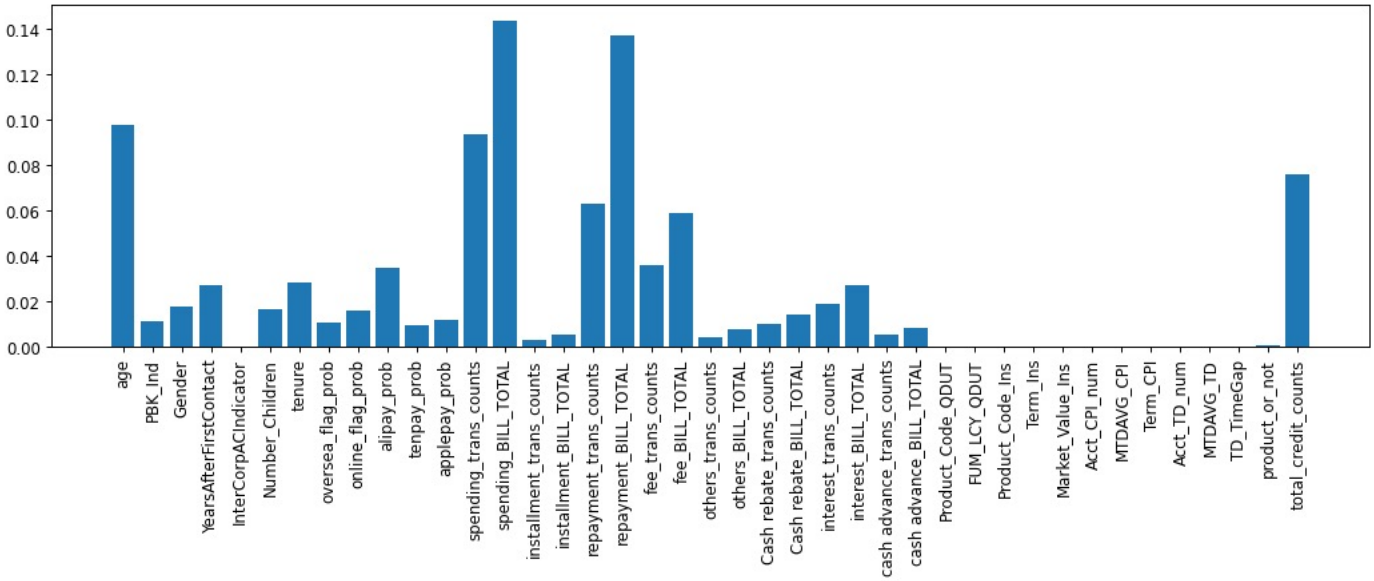

Figure 7: feature importance of Random Forest

In Figure 7, we can find that spending_BILL_TOTAL and repayment_BILL_TOTAL are the

most important features in the Random Forest model to predict the attrition of customers. The age, spending_trans_counts, repayment_trans_counts, fee_BILL_TOTAL, and total_credit_counts are relatively important to predict the attrition of customers by the Random Forest model.

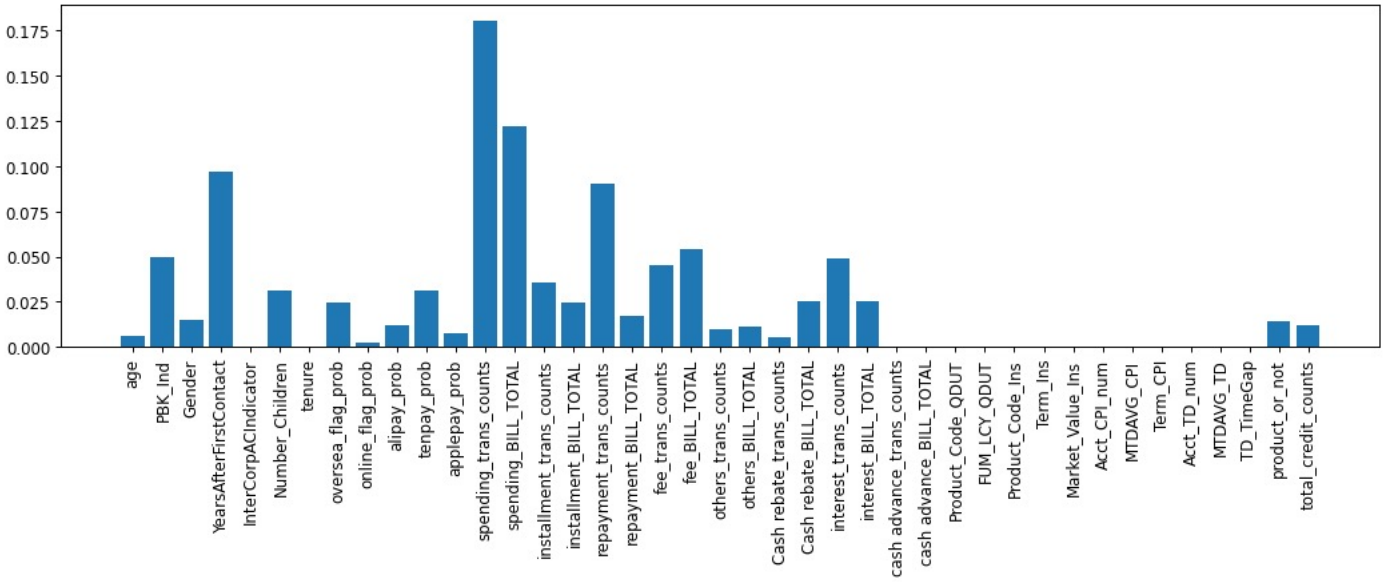

Figure 8: feature importance of XGBoost

In Figure 8, we can find that spending_trans_counts is the most important feature in the XGBoost model to predict the attrition of customers. The YearAfterFirstContact, spending_BILL_TOTAL, and repayment_trans_counts are relatively important to predict the attrition of customers by the XGBoost model. The other parameters are relevant to predicting the attrition of customers.

From the four feature importance graphs, we conclude relatively important features for future analysis while other features without obvious feature importance can not be directly regarded as non-important. Verified by those models, spending_trans_counts, spending_BILL_TOTAL, YearAfterFirstContact, and repayment_trans_counts are parameters that are the most important to predicting the attrition of customers. In the later analysis, we will more focus on them.

4.2 Probability Analysis

As explained in the previous Solution section, we already have the predicted probability and category for 83578 credit card customers.

4.2.1 Predicted Category Analysis

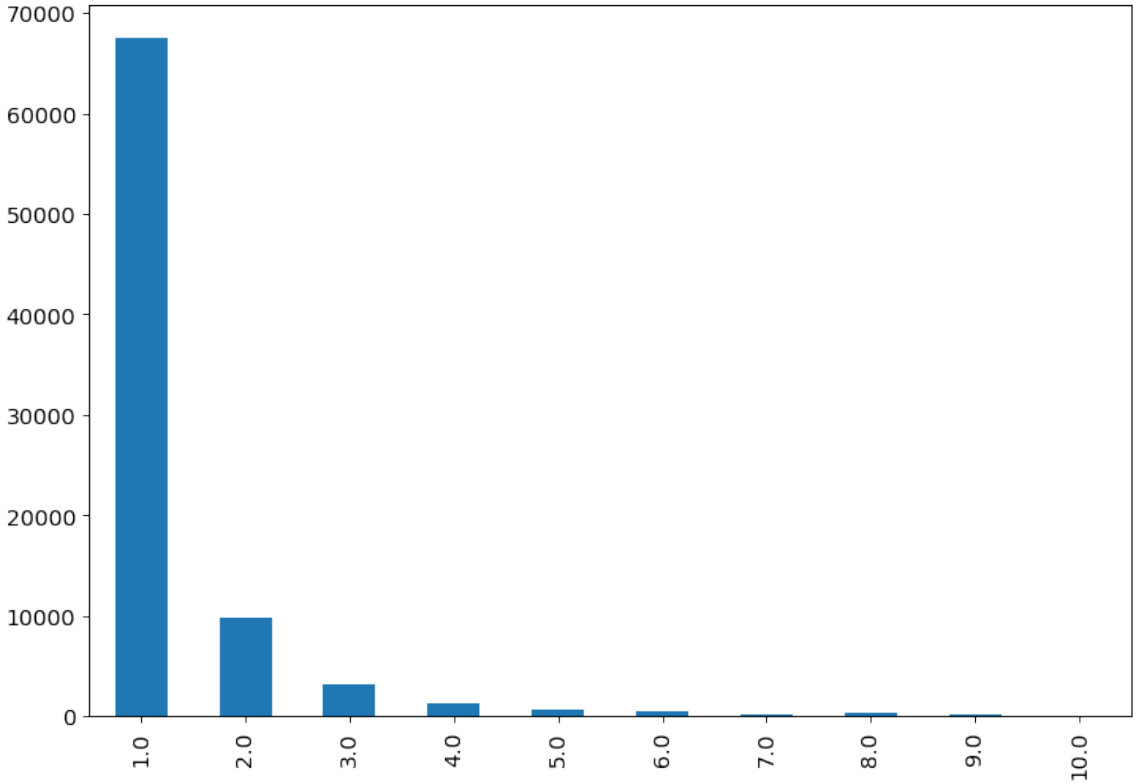

Figure 9: distribution of the predicted categories

The above figure is the distribution of ten categories, set from the fixed 10% average probability explained in 3.4.3 Probability analysis. Though the distribution is imbalanced that Category 1 accounts for nearly 80% while the rest of the categories are fewer. This category helps us to verify the reasonable attrition probability and divide customers by values in the next Customer Value Segments section.

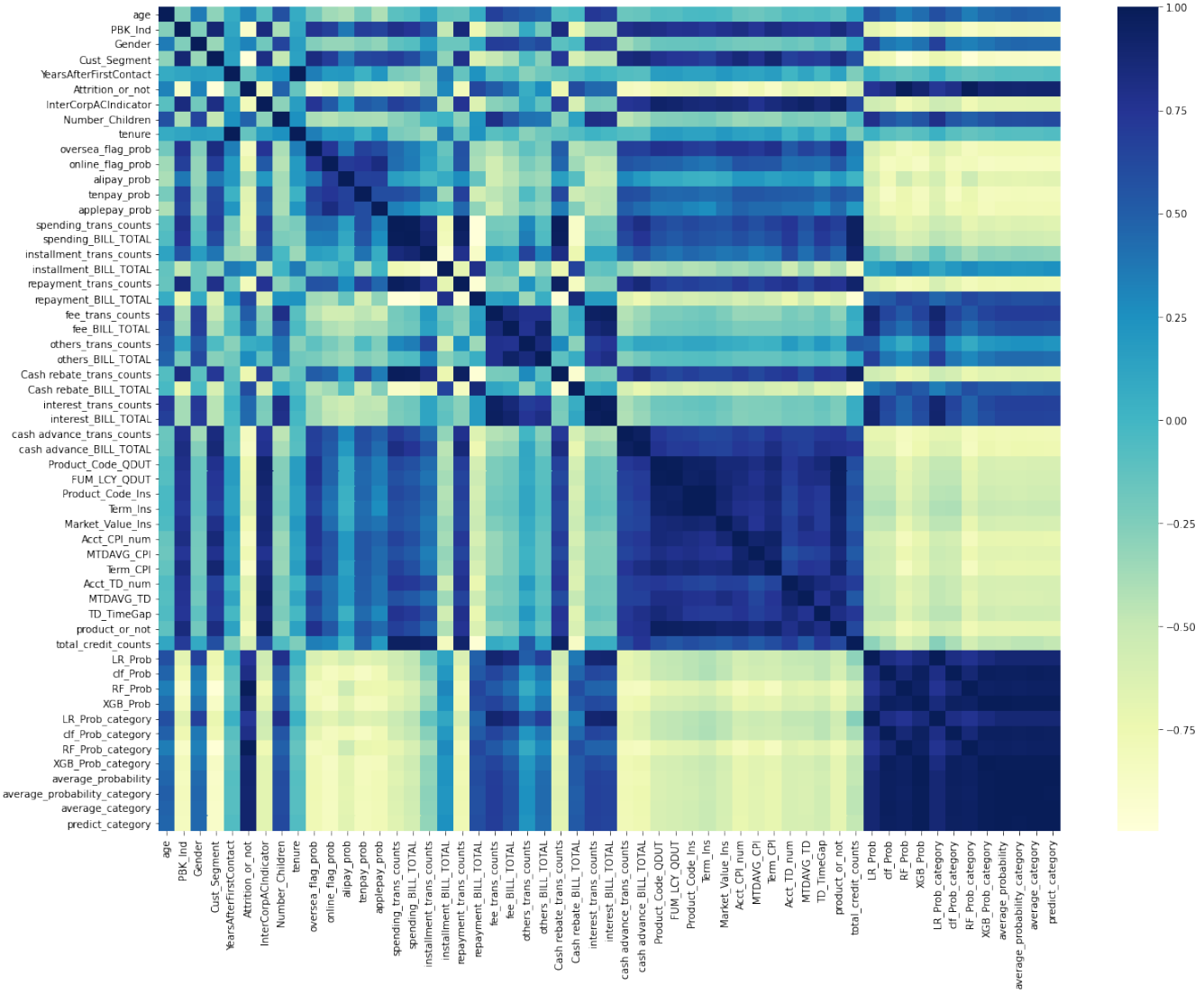

Figure 10: the correlation matrix

After being grouped by the ten categories and taking the mean value of each category, we have the correlation matrix in Figure 10. We pay close attention to the bottom part that the correlation between the predicted value and other features. We could find that almost all of the features are related. We then want to find how specific features are related to the predicted categories and verify them.

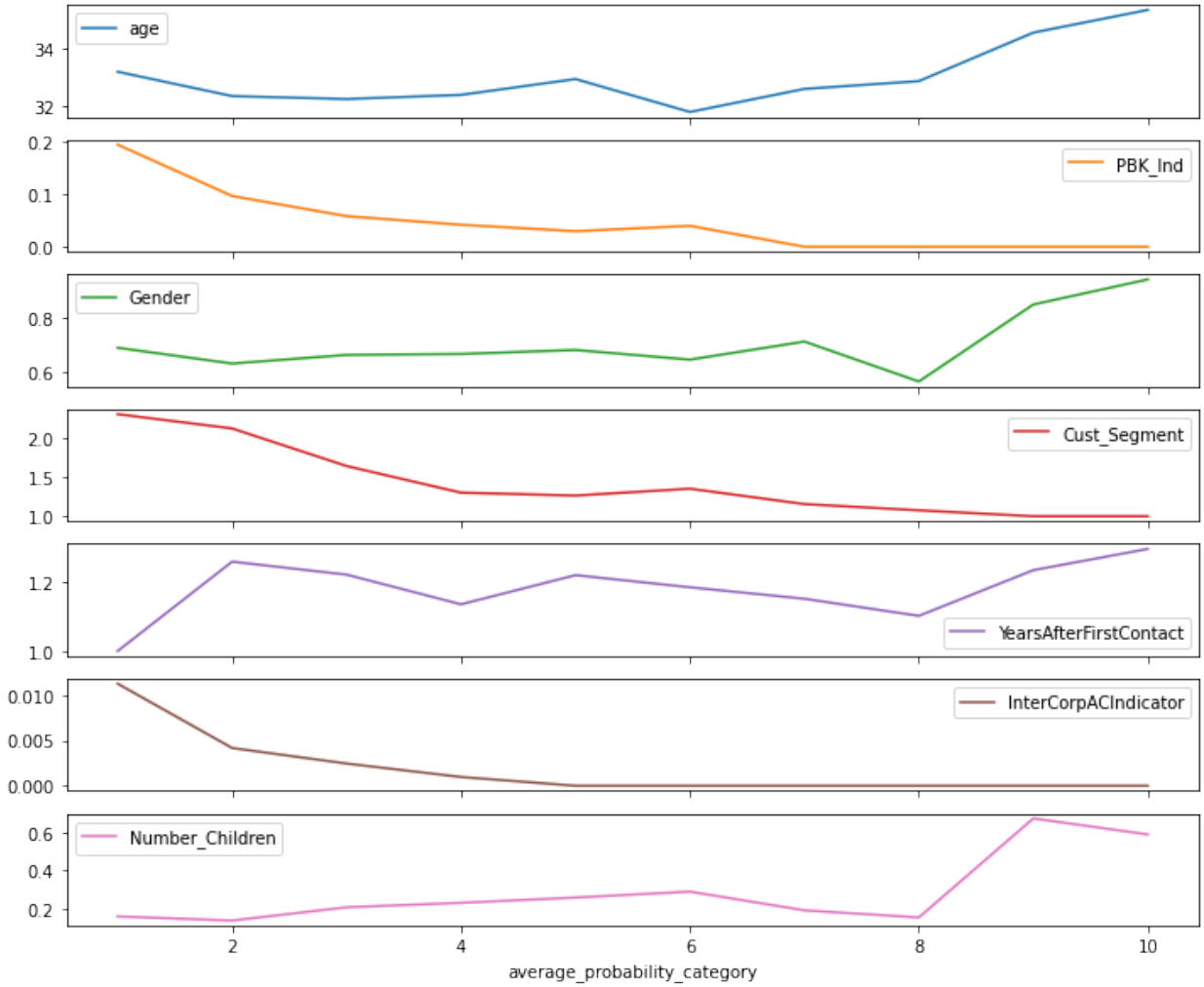

Figure 11: the trend of personal information changes with predicted categories

As shown in figure 11, the trend of Cust_Segment is decreasing, which shows that the 1-10 category prediction satisfies the original HSBC classification in five labels. The graph of PBK_Ind (phone banking) and InterCorpACIndicator (overseas account) also decrease with higher attrition category, which means that customers who have phone banking or overseas accounts are most likely to be categorized into category 1. The graph of age has an increasing trend that indicates its positive correlation with attrition. Gender and Number_Children's increasing trend needs more extra analysis that we will skip in this paper. When coming to YearsAfterFirstContact, we have the following scatter and line plot:

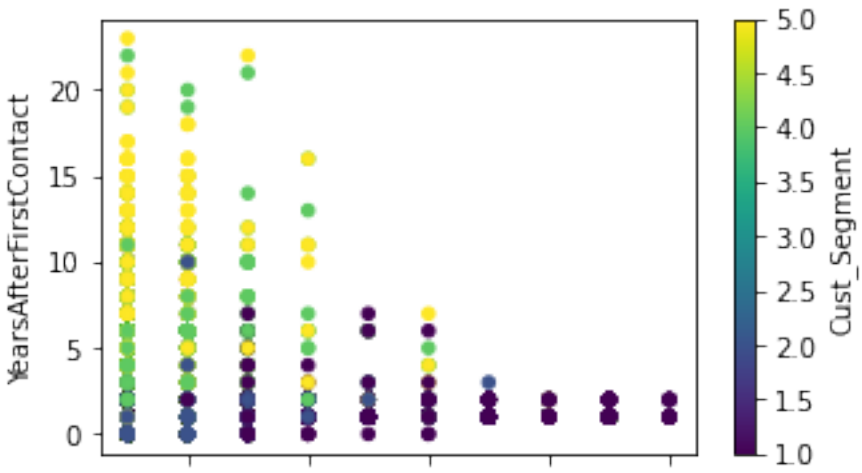

Figure 12: scatter distribution of predicted categories and Cust_Segment on YearsAfterFirstContact

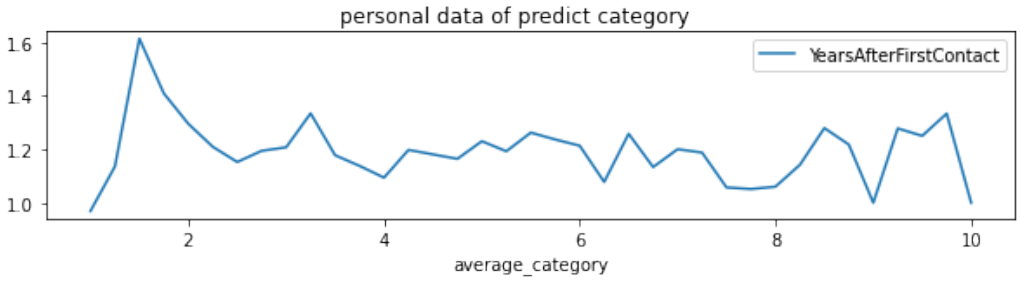

Figure 13: the mean of YearsAfterFirstContact according to average_category

From the above figures 12 and 13, the mean of YearsAfterFirstContact is high in a smaller category. Besides, if the YearsAfterFirstContact is greater than 10, customers are most likely to be in the first four categories and labeled as Premier & Jade and Advance from HSBC. It is a good indicator when analyzing customer segments.

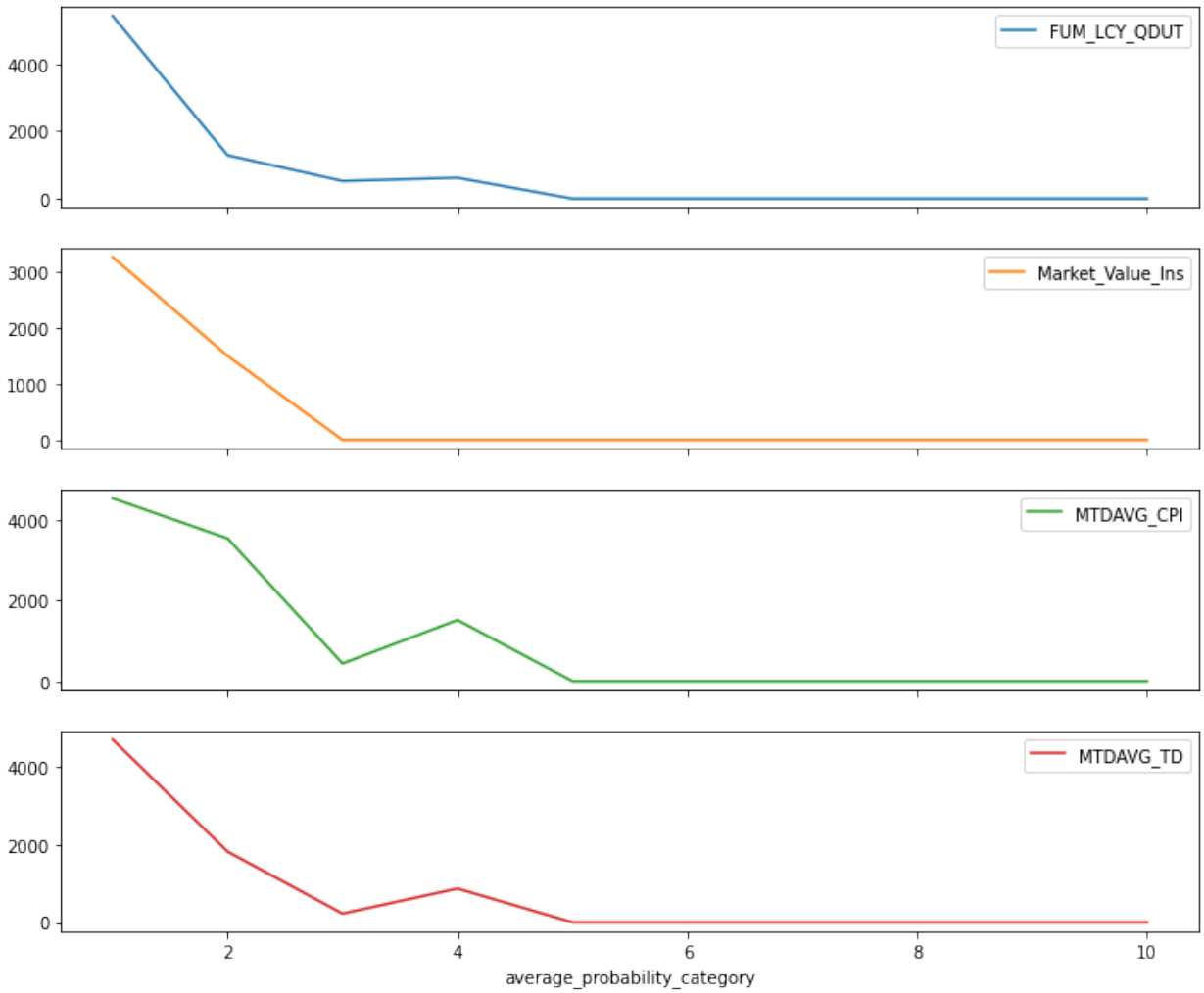

Figure 14: the trend of product information changes with predicted categories

The four features in the above Figure 14 represent four types of products: Mutual Fund (FUM_LCY_QDUT); Insurance (Market_Value_Ins); CPI (MTDAVG_CPI); TD (MTDAVG_TD). The decreasing trends of all four line graphs show that customers in lower attrition categories are more likely to buy any of the products. In short, the product is also a good indicator of customer value segments.

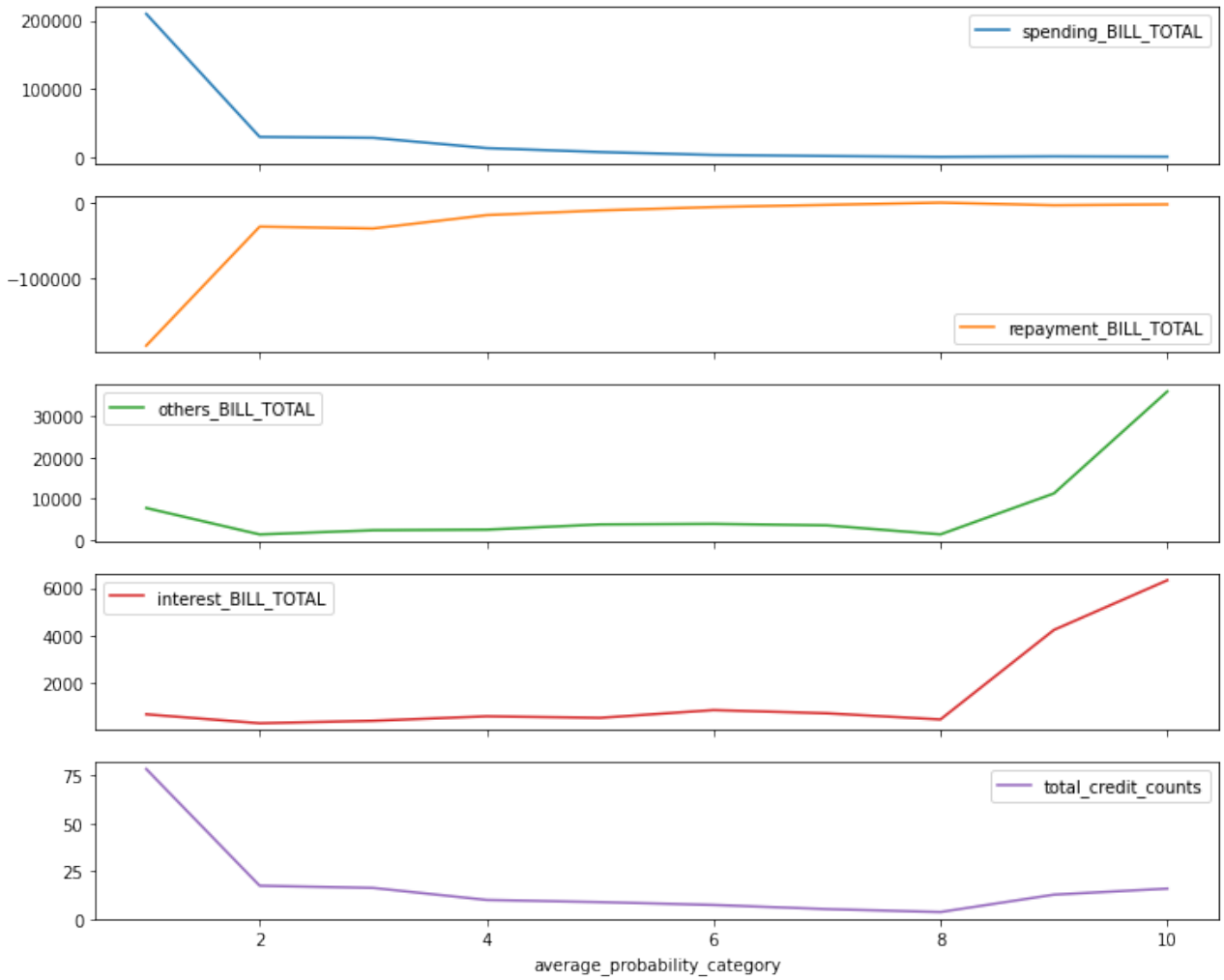

Figure 15: the trend of credit card information changes with predicted categories

Recalling from data preprocessing sections that the credit card records are divided into spending, repayment, interest, and other bill types.

There is an obvious trend for each feature in the above Figure 15. Customers in the low attrition category have high spending bills, high repayment bills, and more total transaction records; while customers in the high attrition category have high-interest bills and other bills.

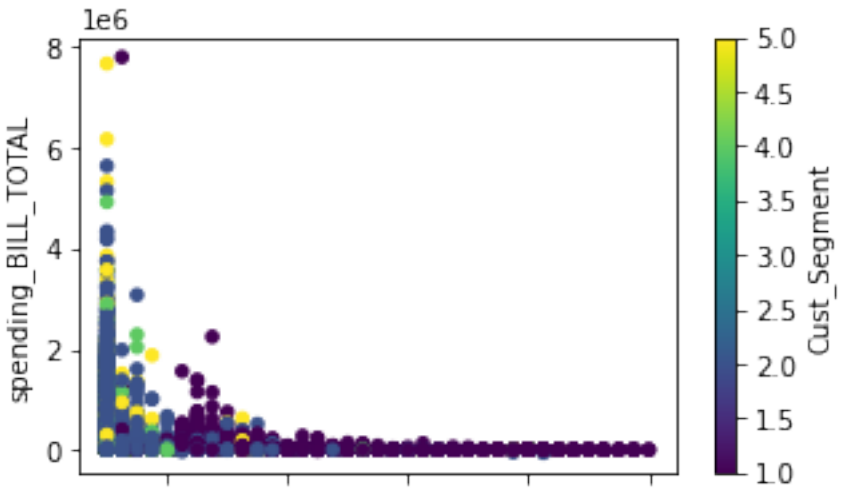

Figure 16: distribution of customer value segments

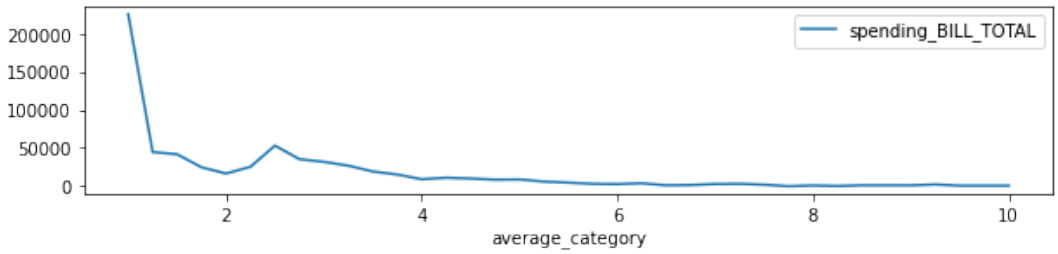

Figure 17: distribution of customer value segments

Since we want to focus on high-value customers, spending is an important factor. The above scatter plot and line graph show that those with predicted category 1 have higher spending bills. If the spending bill is$>$ 3,000,000, the customer will most likely be in a lower attrition category.

4.2.2 Customer Value Segments

From the trends and analysis in the previous Predicted category analysis, we then use k-means to better categorize customers according to their average probability. The customers' new segments are called high-value customers with center 0.0156, relatively high-value customers with center 0.139, relatively low-value customers with center 0.331, and low-value customers with center 0.684.

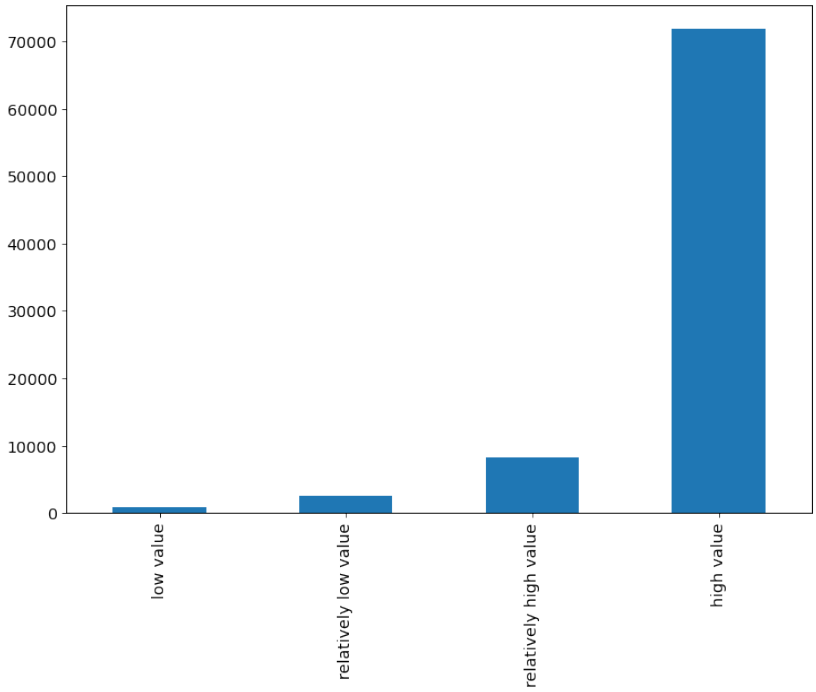

Figure 18: distribution of customer value segments

The above figure 18 shows the four customer value segments from k-means based on the average probability. We have more high-value customers than other customer segments. Those high-value customers are our target and will be assigned different stages in the next section.

4.3 RFM Clusters

4.3.1 Data Clustering

We used k-means to cluster customers with similar weighted RFM into six stages and ordered them by ranking their CLV, where customers are more loyal in higher stages. CLV is calculated so that CLV = wR * CR + wF * CF + wM * CM (wR = 0.731, wR =0.188, wR =0.081) where CR, CF, and CM are the RFM values of cluster centres. The number of stages was determined using the analysis of variance. We were able to reject the null hypothesis with a p-value less than 0.05 and conclude that all three values, years after the first contact, the number of spending, and the amount of spending are significantly different among these six stages. Table 6 shows the number of customers and the mean RFM and CLV in each stage.

Clusters	No. of customers	Recency	Frequency	Monetary	CLV
1	19610	0	20.7	83561	0.00335
2	51317	1	41.14	209391	0.03823
3	6771	2.2	88.87	278411	0.08127
4	2455	1.16	481.16	424691	0.08450
5	1471	6.13	140.11	167823	0.20960
6	548	11	145.83	252847	0.36635

Table 8: Six clusters by K-means clustering

4.3.2 Cluster Analysis

To verify the correctness of our clustering, we looked at the mean amount of balance, investment products and insurance of the customers in each cluster. The clear increasing trends in figure 19 show that customers with higher lifetime value (clusters 4 and 5) are more likely to have deposits and buy products, which indicates that these customers indeed have high values for the bank. We were then able to use k-nearest neighbour to decide which cluster a customer should be put in based on their RFM values with an accuracy of 0.98.

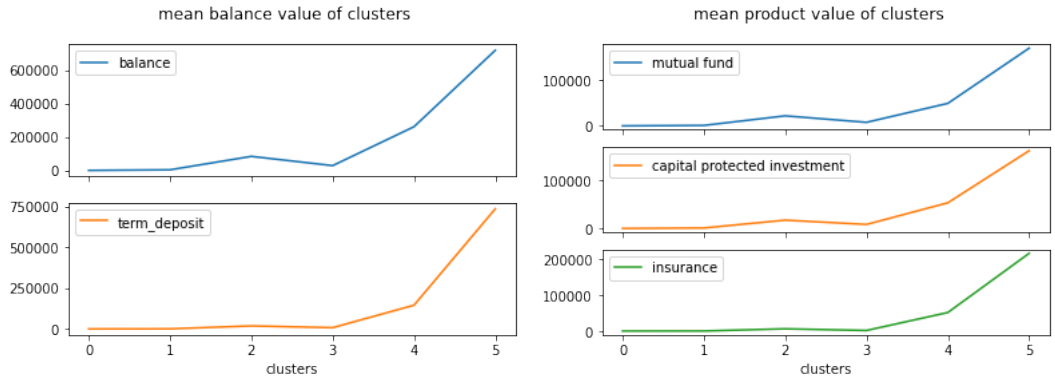

Figure 19: mean value of clusters

Figure 20 shows the box plots of RFM. The x-axis is the six clusters and the y-axis is the RFM values.

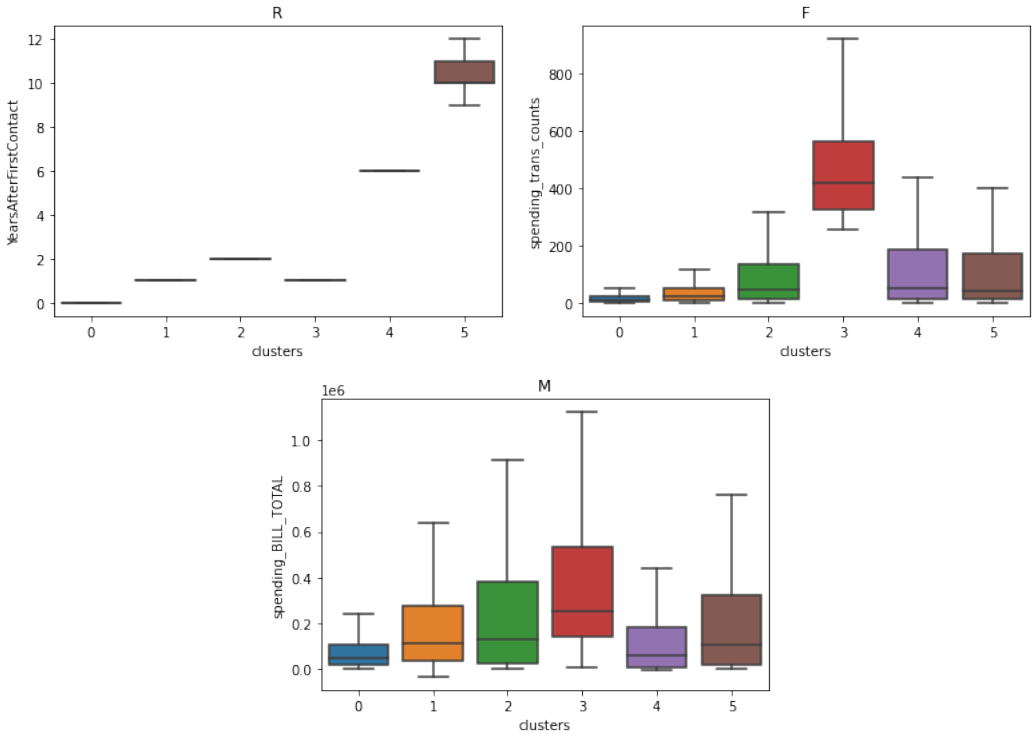

Figure 20: box plots of RFM

As discussed above, customers from clusters 4 and 5 tend to buy more products. Therefore, in general, we may conclude that the years after contact indicates customers longer than 2 years are more likely to become loyal customers. For the six clusters, we divided them into three stages, with clusters 0 and 1 being the developing stage, clusters 2 and 3 being the honeymoon stage, and clusters 4 and 5 being the retention stage. During the developing stage, customers are rather new to the bank and more insights will thus be needed to find out their purchase preferences. For cluster 1 customers, in particular, they have been with the bank for some time and are spending rather a large amount of money. Therefore, they are important to develop so that they will use credit cards more frequently and reach the frequency and amount of customers in the honeymoon stage. For the honeymoon stage, it is important to keep the purchase ability of customers. These customers have more spending behaviours and are very valuable to the bank. However, they are not buying enough products and insurance. Because they are rather similar to customers in the retention stage except for the number of years they are with the bank, it is reasonable to assume that they will be interested in products and can stay with the bank for a long time. For the retention stage, it is necessary to provide customers with more personalized services and promote newly developed products to them.

4.4 Representative customers or recommendation

To further understand the difference among clusters within the same stage and develop specific strategies to recommend products, we chose some features from customers' personal information to study their preferences. Two features of interest are the age and the number of children. We can see from figure 21 that the mean age of customers has an increasing trend with clusters 4 and 5 being older. For the number of children, we have clusters 0 and 4 smaller than 1, 3, and 5 smaller than 2.

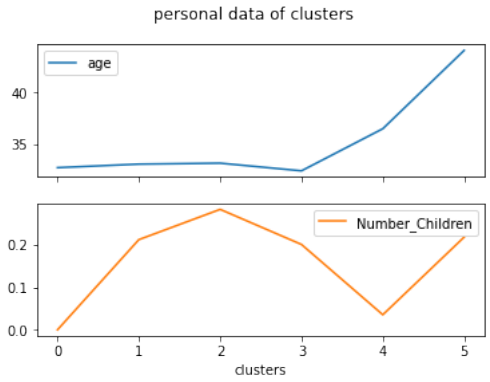

Figure 21: mean personal info of clusters

Combining their personal information and lifetime values, we then came up with some ideas for recommendations. For example, in the honeymoon stage, customers in clusters 2 and 3 are presenting different characteristics. Since cluster 2 customers are the ones likely to have more children, children-related products can be recommended to them. Cluster 4 customers, on the other hand, are the least likely to have children among customers in the retention stage. Therefore, more risky investment products may suit them better.

5 Discussion

For the data preprocessing part, there are lots of missing values during which we didn't use the features with over 100,000 missing values. If we replace these values with the mean and mode, we may also add some valuable features.

Our recommendation is more business-focused, we want to collect more data on products so that we can make a better recommendation.

Also, we want to create a web for HSBC, entering customer ID at the searching box, the page will return his/her profile, product information, credit card transaction, and our recommenda-

tions. Because of the time limitation, we just put a demo design here, if we got the chance, we want to finish the website in the future.

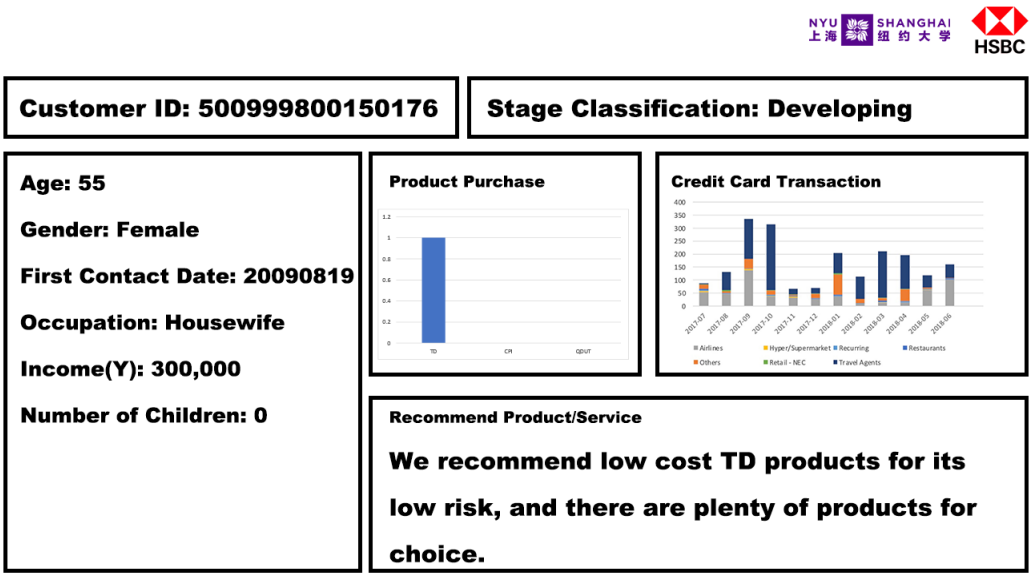

Figure 22: Customer Personal Page DEMO

6 Conclusion

In conclusion, we hope to assist HSBC in conducting customer segment predictions and recommendations for this project in order to maximize customers' lifetime value. Our approach is mainly composed of three steps. To begin, we computed consumer values based on personal data, purchasing power, and consumption habits. We also anticipated that a consumer might become attrition. As for our attrition prediction, we reached 96% accuracy. Then we created a client lifecycle and divided customers into three stages: developing, honeymoon, and retention. We designed these three stages according to RFM k-means. Finally, we made individualized product and service recommendations to distinct consumer phases or segments based on their habits and preferences.

For future work, while we have looked into some of the personal information of customers, more features such as whether they own overseas or online banking can also be useful in deciding the future behaviours of customers.

Our methods and considerations go beyond model realism. Our solutions are tailored to real-world business problems, and we foresee deployable techniques in HSBC's operations. We also consider the most profitable purchasing conversion strategy for HSBC. If embraced and imple-

mented, we are confident that our solutions will help to improve the current banking situation.

References

[1] S. Amaresan, "Everything you need to know about customer lifecycle management," Jan 2022. [Online]. Available: https://blog.hubspot.com/service/customer-lifecycle-management#:~:text=The%20customer%20lifecycle%20refers%20to,conversion%2C%20retention%2C%20and%20loyalty

[2] R. Izquierdo, "5 stages of the customer life cycle (updated 2022)," Jan 2021. [Online]. Available: https://www.fool.com/the-blueprint/customer-life-cycle/

[3] I. Smeureanu, G. Ruxanda, and L. M. Badea, "Customer segmentation in private banking sector using machine learning techniques," *Journal of Business Economics and Management*, vol. 14, no. 5, p. 923–939, 2013.

[4] A. Kazemi and M. Babaei, "Modelling customer attraction prediction in customer relation management using decision tree: A data mining approach," vol. 4, pp. 37–45, 06 2011.

[5] S. F., "Machine-learning techniques for customer retention: A comparative study," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 2, 2018.

[6] A. Knott, A. Hayes, and S. A. Neslin, "Next-product-to-buy models for cross-selling applications," *Journal of Interactive Marketing*, vol. 16, no. 3, p. 59–75, 2002.

[7] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The Journal of Educational Research*, vol. 96, no. 1, pp. 3–14, 2002. [Online]. Available: https://doi.org/10.1080/00220670209598786

[8] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *Journal of Chemometrics*, vol. 18, no. 6, pp. 275–285, 2004. [Online]. Available: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.873

[9] J. R. QUINLAN, "Learning decision tree classifiers," *ACM Computing Surveys*, vol. 28, no. 1, pp. 71–72, 1996. [Online]. Available: https://dl.acm.org/doi/10.1145/234313.234346

[10] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation," *Biomedical Signal Processing and Control*, vol. 52, pp. 456–462, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809417300204

[11] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[12] D.-R. Liu and Y.-Y. Shih, "Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences," *Journal of Systems and Software*, vol. 77, no. 2, p. 181–191, 2005.

[13] N.-C. Hsieh, "An integrated data mining and behavioral scoring model for analyzing bank customers," *Expert Systems with Applications*, vol. 27, no. 4, p. 623–633, 2004.

[14] K. Krishna and M. Narasimha Murty, "Genetic k-means algorithm," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 29, no. 3, pp. 433–439, 1999.