

**AN6003 Analytics Strategy**

**Group Report**

Group A Team 7

Instructor: Professor Chew Chee Hua, Neumann

|  |  |
| --- | --- |
| **Name** | **Matriculation Number** |
| Cheng Jie | G2000455H |
| Li Ziye | G2000873G |
| Lyu Yunfan | G2000805D |
| Ma Yuqing | G2000305E |
| Wang Lyu | G2003525C |

[**Executive Summary** 5](#_Toc52992267)

[**1. Business Problem and Business Outcome** 6](#_Toc52992268)

[**2. Analytics Problem** 7](#_Toc52992269)

[**3. Analytics Performance Measurement Metrics and Targets** 7](#_Toc52992270)

[3.1 Customer Churn Prediction 7](#_Toc52992271)

[3.1.1 Performance Measurement Metrics 7](#_Toc52992272)

[3.1.2 Performance Target 8](#_Toc52992273)

[3.2 Customer Segmentation 8](#_Toc52992274)

[3.2.1 Performance Measurement Metrics 8](#_Toc52992275)

[3.2.2 Performance Target 8](#_Toc52992276)

[**4. Analytics Methodology** 9](#_Toc52992277)

[Overview of Solution 9](#_Toc52992278)

[4.1 Dataset Relevance 9](#_Toc52992279)

[4.2 Data Exploration and Visualization 10](#_Toc52992280)

[4.2.1 Dataset for Churn Modeling 10](#_Toc52992281)

[4.2.2 Dataset for Customer Segmentation 13](#_Toc52992282)

[4.3 Model Building - Customer Churn Prediction 14](#_Toc52992283)

[4.3.1 Data Preprocessing 14](#_Toc52992284)

[4.3.2 Model Fitting (5 models) 15](#_Toc52992285)

[4.3.2.1 Logistic Regression 15](#_Toc52992286)

[4.3.2.2 Decision Tree 15](#_Toc52992287)

[4.3.2.3 Random Forest 16](#_Toc52992288)

[4.3.2.4 XGBoost 16](#_Toc52992289)

[4.3.2.5 Neural Network 16](#_Toc52992290)

[4.3.3 Significant features identification 17](#_Toc52992291)

[4.4 Model Building - Customer Segmentation 17](#_Toc52992292)

[4.4.1 Clustering Algorithm: Partitioning Around Medoids (PAM) 17](#_Toc52992293)

[4.4.2 Similarity Measurement 18](#_Toc52992294)

[4.4.3 Decide Number of Optimal Clusters, K 19](#_Toc52992295)

[4.4.4 Visualization on high-dimensional dataset 20](#_Toc52992296)

[**5. Result and Performance Evaluation** 20](#_Toc52992297)

[5.1 Customer Churn Prediction 20](#_Toc52992298)

[5.2 Customer Segmentation 21](#_Toc52992299)

[5.2.1 Customer Portrait 21](#_Toc52992300)

[5.2.2 Customer Portrait Performance Evaluation 22](#_Toc52992301)

[**6. Analytics strategy** 22](#_Toc52992302)

[**7. Future Actions** 23](#_Toc52992303)

[**8. Conclusion** 24](#_Toc52992304)

[**Appendices** 25](#_Toc52992305)

[Appendix 1: Customer Churn Prediction Dataset 25](#_Toc52992306)

[Appendix 1.1.1: Skewness of all numeric data 25](#_Toc52992307)

[Appendix 1.1.2: Kurtosis of all numeric data 25](#_Toc52992308)

[Appendix 1.2: Heat map of data correlation 26](#_Toc52992309)

[Appendix 1.3.1: Simulated distribution curve of numerical variables 26](#_Toc52992310)

[Appendix 1.3.2: Distribution of numerical variables given different exited situation 27](#_Toc52992311)

[Appendix 1.4.1: Distribution of categorical variables 27](#_Toc52992312)

[Appendix 1.4.2: Distribution of categorical variables given different exited situation 27](#_Toc52992313)

[Appendix 1.5.1: Pairplot 28](#_Toc52992314)

[Appendix 2: Customer Segmentation Dataset 30](#_Toc52992315)

[Appendix 2.1.1: Skewness of all numeric data 30](#_Toc52992316)

[Appendix 2.1.2: Kurtosis of all numeric data 30](#_Toc52992317)

[Appendix 2.2: Heat map of data correlation 31](#_Toc52992318)

[Appendix 2.3.1: Simulated distribution curve of numerical variables 31](#_Toc52992319)

[Appendix 2.3.2: Distribution of numerical variables given different response to campaign 32](#_Toc52992320)

[Appendix 2.4.1: Distribution of categorical variables 34](#_Toc52992321)

[Appendix 2.4.2: Success probability of categorical variables 34](#_Toc52992322)

[Appendix 2.5.1: Pairplot 35](#_Toc52992323)

[Appendix 3: Statistical summary for prediction results 37](#_Toc52992324)

[Appendix 3.1: Logistical Regression 37](#_Toc52992325)

[Appendix 3.2: Decision Tree 38](#_Toc52992326)

[Appendix 3.3: Random Forest 39](#_Toc52992327)

[Appendix 3.4: XGBoost 39](#_Toc52992328)

[Appendix 3.5: Neural Network 40](#_Toc52992329)

[Appendix 3.6: Feature Importance 40](#_Toc52992330)

[Appendix 4: Statistical summary for clustering results 41](#_Toc52992331)

[**References** 43](#_Toc52992332)

# **Executive Summary**

Considering the higher costs to acquire new customers compared with retaining existing customers, and the fact that long-term customers tend to produce more profits, churn management can help White Rock optimize customer retention and thus increase profitability.

The purpose of this project is to solve the problem of reducing customer churn with a low cost by improving prediction accuracy of the churn model and identifying high net worth customers. Specifically, this project will provide White Rock with (1) an accurate classification model to identify churning customers and corresponding strategies to retain them, (2) customer segmentation to identify customers worth retaining and corresponding strategies for different segments.

With this in mind, the first part of this project focuses on customer churn prediction models: Logistics Regression, Decision Tree, Random Forest, XGBoost and Neural Network to identify churn customers based on their basic information and transaction data. Our finding showed that XGBoost is a desirable classifier on customer churn prediction: it shows a highest accuracy, a high recall and an optimal AUC score. For significant features identification, we calculate feature importance scores on different models, focusing on Age, NumofProducts, IsActiveMember, Gender, Germany (they contribute 95% to the model) and put forward some useful suggestions correspondingly.

The second part of our project strives to provide White Rock with reasonable customer segmentation and customer portrait to help increase revenue and reduce costs when retaining churn customers. We apply the K-Medoids algorithm to partition the dataset into 5 clusters and conclude each cluster with a specific pattern which is helpful in recognizing the high-value customer group we target with. The computed silhouette coefficient shows clusters are almost well apart from each other and are clearly distinguished. And we created customer portrait for these five cohorts, preparing for subsequent customer retention.

Overall, this project provides comprehensive analytical solutions to achieve churn management with low costs and high efficiencies for White Rock. With the customer churn prediction model, the marketing team can closely monitor the activity level of the customer and take actions accordingly in advance to reduce the probability of customer churn. At the same time, the customer segmentation model helps to distinguish between high value customers, high growing value customers, and low value customers, so that tailor made marketing strategies can be applied to each group, which will be more cost effective. If put into operation, this project will help White Rock mine the value of internal data from operational transactions by its automated and intelligent technology.

# **1. Business Problem and Business Outcome**

As for companies like White Rock, when the number of customers in the business peaks, the cost of finding and acquiring new customers becomes increasingly costly compared with retaining existing customers; meanwhile, long-term customers tend to produce more profits. In this case, Verbeke et al. (2011) asserts that customer retention increases profitability. According to Nie et al. (2011), a bank can increase its profits by up to 85 % by improving the retention rate by up to 5 %. Many competitive organizations have realized that a key strategy to survive is to retain existing customers through churn management.

In this case, the concept of churn management is to identify those customers who intend to close their account and stop trading in White Rock; especially those tend to move their accounts to competing service providers. Once these customers are identified, they can be targeted with proactive marketing campaigns for retention efforts. And deliberate churn is the problem that most churn management solutions try to tackle down. This type of churn occurs when a customer decides to move his/her account to a competing company. Reasons that can lead to deliberate churn include technology-based reasons, when a customer discovers that a competitor is offering the latest products that their current supplier is unable to provide. Economic reasons include finding better-priced products from competing companies. Other reasons for deliberate churn include quality factors, such as poor coverage or possibly a bad call center experience, etc.

Another important fact is that churn management efforts should not focus on the entire customer base, because (1) attempting to retain customers that have no intention of churning will waste resource, (2) not all customers are worth retaining.

For (1), our models can accurately identify churn customers, making it possible to target the individual to prevent customer churning, and to identify and improve upon areas where customer services is lacking.

For (2), our team includes customer segmentation in our models. By conducting proper customer segmentation and taking corresponding strategy, we can improve the overall efficiency and therefore reduce operating costs. For White Rock, a company who highly relies on customer relationship, it will help the company understand the cost to acquire, serve and retain clients. With proper segmentation, it can also help identify indicators of customer churn and actions to take to reduce it for different segments, which can also help increase revenue and reduce costs.

In summary, our solution to the business problem of customer retention is to reduce customer churn with the lowest cost by improving prediction accuracy of the churn model and identifying high net worth customers.

# **2. Analytics Problem**

In this project, the analytics problems need to be resolved are customer churn prediction and customer segmentation so that we can know to what extent the high-value customers have churned and which customers can we target for retention when they decide to leave.

For customer churn prediction, the challenge is how to precisely predict whether a customer will leave (close account) based on his or her personal information like gender, age, income, account balance, etc. , which means we need to improve recall while ensuring a high prediction accuracy. Otherwise, a low true positive rate will lead to some customer churn that could have been avoided and a low true negative rate can cause unnecessary cost of retention on wrong targets. Although we are provided with advanced matching learning models such as decision tree, XGBoost, neural network, etc. , prediction result is not always satisfactory due to improper setting of parameters, undesirable variables selection, imbalanced dataset, etc., which we should pay close attention to. Another challenge is significant features identification, i.e. identify which variables play most significant roles in the customer churn prediction model. Based on this, we can put forward suggestions to reduce customer churn combining with other factors in real business context.

For customer segmentation, the challenge is how we can precisely cluster different group of customers so that their relating characteristics, like age, job, default experience and etc., can be well grouped as well. With the rise of AI and machine learning technology, unsupervised learning like clustering can provide us with the appropriate solution. However, you may find that by adding all relevant fields to the cluster may not show satisfactory results, instead, a clutter will appear. Thus, a good way to avoid this is to do feature engineering by implementing data exploration and visualization. Another challenge that needs to be focused on is to decide the number of clusters in order to achieve both statistical significance and practical significance. And then we can easily tell from clustered group of customers about their key characteristics and achieving the optimal clustering effect at the same time.

# **3. Analytics Performance Measurement Metrics and Targets**

In order to evaluate the success and effectiveness of the analytics models, we used the following performance measurement metrics and set the goals that the model should achieve.

## 3.1 Customer Churn Prediction

### 3.1.1 Performance Measurement Metrics

For customer churn prediction, we are going to focus on both recall and accuracy, so we consider indicators below as model performance measurement metrics.

1. showing the overall prediction accuracy of the model.
2. showing rate of precisely predicting customer churn
3. , the measure of the ability of a classifier, the higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

### 3.1.2 Performance Target

The performance target is to precisely conduct customer churn prediction. Specifically, we aim to achieve both accuracy and recall above 70% and AUC above 0.800. Besides, we need to identify reasonable significant features, based on which we can put forward suggestions for retention.

## 3.2 Customer Segmentation

### 3.2.1 Performance Measurement Metrics

- Before clustering

Before evaluating clustering performance, it is important to ensure that the dataset we use has a tendency to cluster and does not contain uniformly distributed points. If the dataset does not show in clustering trends, the clusters identified by any state-of-the-art clustering algorithm may be irrelevant. Thus, it is important to define the uneven distribution of points in the dataset. To solve this, a combination of statistical method called the Hopkins statistic (Lawson and Jurs 1990) and visual method (i.e. data exploration and visualization) can be used.

- After clustering

Once clustering is done, we need to check clustering quality as well. Ideally, clustering quality can be characterized by minimal intra cluster distance and maximal inter cluster distance. A small minimal intra cluster distance and a large maximal inter cluster distance is what we want. To check for this, we use Intrinsic Measure like Silhouette Coefficient which does not require ground truth labels.

### 3.2.2 Performance Target

The performance target is to obtain a clear and separated group of clusters with understandable and representable features so that we can well conclude each sub-group of our customers and identify the high-value customers we look for.

# **4. Analytics Methodology**

## Overview of Solution

First of all, to ensure both customer churn prediction model and customer segmentation model are compatible with our dataset, a data cleaning and exploration process will be implemented to identify the possible flaws of our data and model design.

Next, for customer churn prediction model building, we use 5 machine learning models, logistic regression, decision tree, random forest, XGBoost, and Neural Network, to conduct classification on dataset and then compare performance of these models to determine which the best one is.

As for significant features identification, we use feature importance, a technique that assigns a score to input features based on how useful they are at predicting a target variable, for logistic regression, decision tree, random forest and XGBoost, to identify which features are the most important and least important to the model when making a prediction.

In order to do customer segmentation, we use K-Medoids or Partitioning Around Medoids (PAM) as our clustering algorithm with Gower distance measuring the similarity between different cluster groups. By selecting 7 relative attributes from the provided Bank Customers Survey dataset and evaluating the clustering quality with Silhouette Coefficient, we eventually clustered the total customers into 5 different sub-groups where we summarized the characteristics of each group and developed different marketing and retention plans for our interested high-value customers.

## 4.1 Dataset Relevance

Based on publicly available information on kaggle, we use two separate datasets for prediction and clustering.

For Customer churn prediction, we use a real bank’s churn modeling dataset that contains customer personal information (credit score, gender, age, estimated salary) and bank related transaction information (tenure, balance, number of bank products, has a credit card with the bank). Using this bank dataset to build a churn prediction model for White Rock is reasonable because the attributes that indicate whether a client of bank or White Rock will churn are similar, like current balance, client income, the number of products purchased by clients, etc. And we suppose the churn customer of White Rock will have similar user portrait as in the bank case. So we can use the transaction data and personal information to build a classification model and find churn indicators for White Rock.

For customer segmentation, we use telemarketing sample data of a real bank which is derived from UCI machine learning repository. As an investment company, the most important data that can help us to judge the value of customers would be their basic information, credit information, loan information and balance information. And since the amount of original data is too large, we randomly select 24999 observations with 7 fields (age, job, marital status, education, credit default, average yearly balance, housing loan) revealing the necessary customer information for clustering. When read in the dataset, we strictly check that each character field will be recognized as a categorical column to guarantee the proper calculation for Gower distance between any two clusters. Also, we confirm that the dataset is complete enough with zero missing values so that clustering result can be more appropriate.

## 4.2 Data Exploration and Visualization

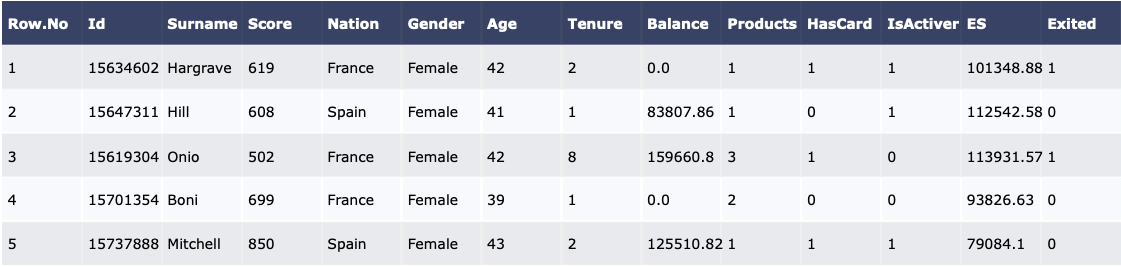
Deep understanding and data cleaning is important for making good predictions. So firstly we will explore and clean the data, and then we will make visualization to deepen our understanding of the data. After visualization we will decide if we need creating new variables or not. Finally we will use Machine Learning Algorithms for make predictions and achieve customer segmentation.

### 4.2.1 Dataset for Churn Modeling

This data set contains details of a bank's customers and the target variable is a binary variable reflecting the fact whether the customer left the bank (closed his account) or he continues to be a customer.

Summary Information about the variables and their types in the data:

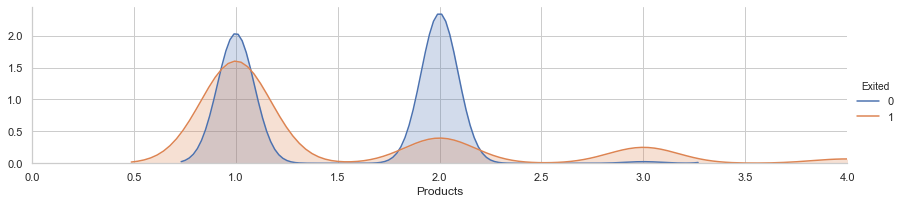
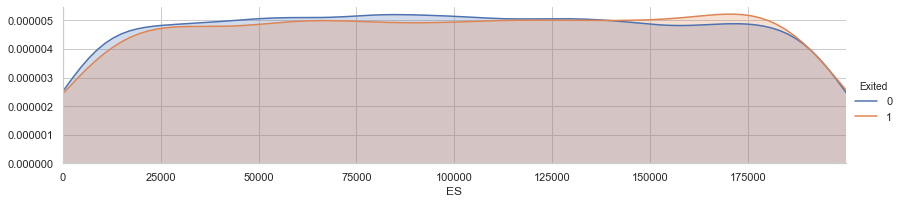
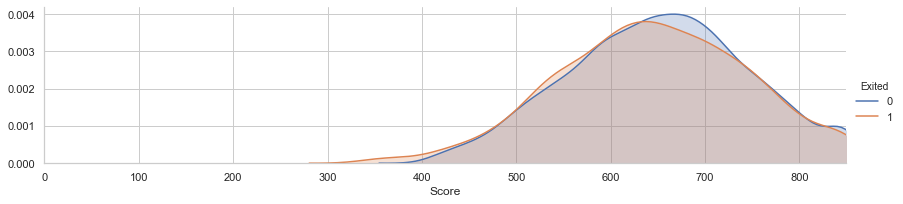
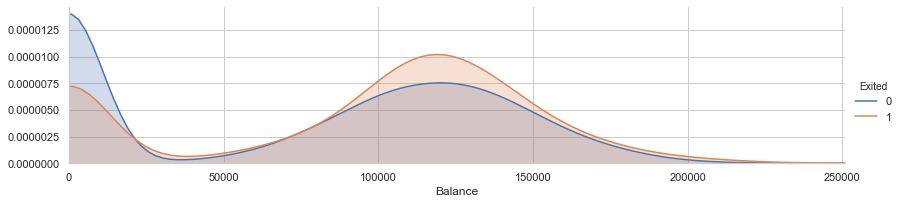
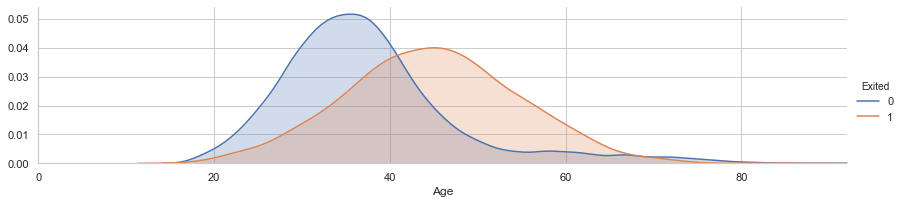
1. Surname: The surname of the customer
2. CreditScore: The credit score of the customer
3. Geography: The country of the customer (Germany[/France/Spain](https://colab.research.google.com/drive/1pecweHuBGf9bMhJ5zGGZIYpsgTK69MUj))
4. Gender: The gender of the customer (Female/Male)
5. Age: The age of the customer
6. Tenure: The customer's number of years in the in the bank
7. Balance: The customer's account balance
8. NumOfProducts: The number of bank products that the customer uses
9. HasCrCard: Does the customer has a card? (0=No,1=Yes)
10. IsActiveMember: Does the customer has an active membership (0=No,1=Yes)
11. EstimatedSalary: The estimated salary of the customer
12. Exited: Churned or not? (0=No,1=Yes)



First 5 Rows of Data



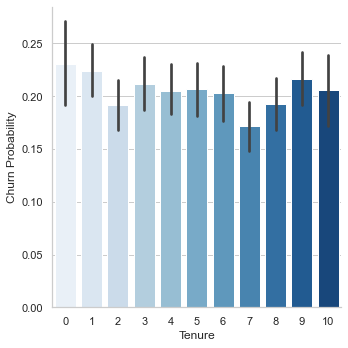
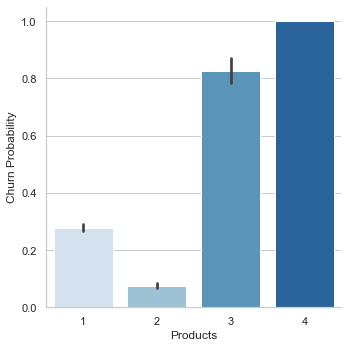
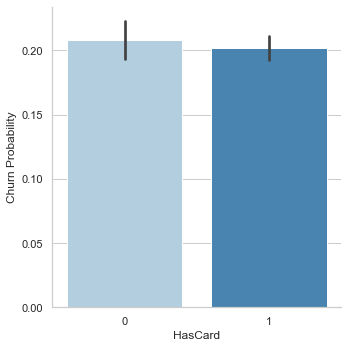
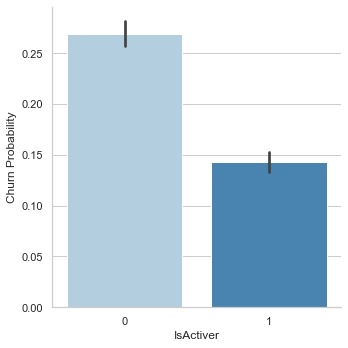
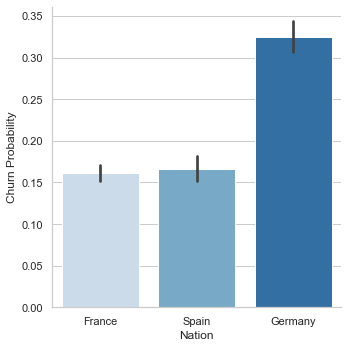
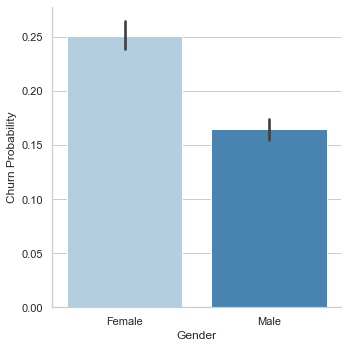
Descriptive statistics



Distribution of numerical variables given different exited situation

We can conclude from the Chart that:

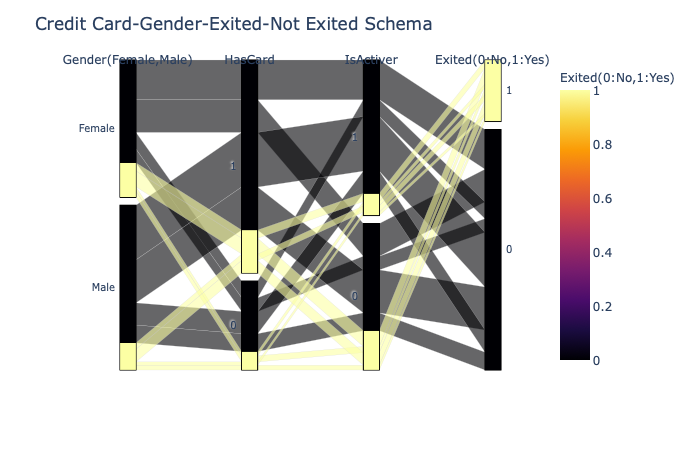
* Older customers tend to exit compared to young customer.
* When customers don’t have outstanding balance, they are less likely to exit.



Churn probability of categorical variables

We can conclude from the Chart that:

* Female has more tendencies to exit than male.
* When the account is active, customer tend not to exit.
* Customers who have many products tend not to exit.
* Germany, compared to others, tends not to exit.



Credit Card-Gender-Exited-Not Exited Schema Gender-Geography-Exited-Not Exited Schema

We can conclude from the schema that

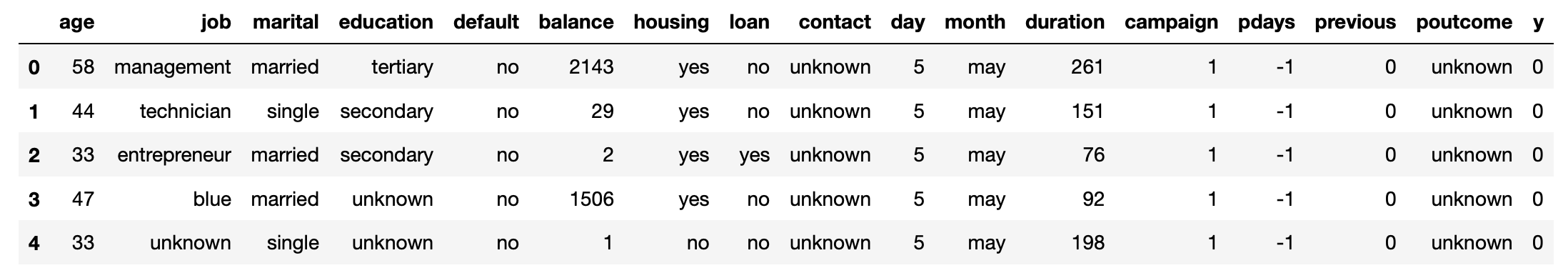
* Around 20% of people exited.
* Females exited proportionally and numerically more (1139 Female/898 Male).
* Germans proportionally and numerically exited most (448 Female - 366 Male).
* French female numerically exited more (460).
* Frenches continued with a bank numerically and proportionally most (5014-810 = 4196).
* Females and inactive members more prone to exit.
* Credit card users are also a bit more prone to exit than non-credit card users.

### 4.2.2 Dataset for Customer Segmentation

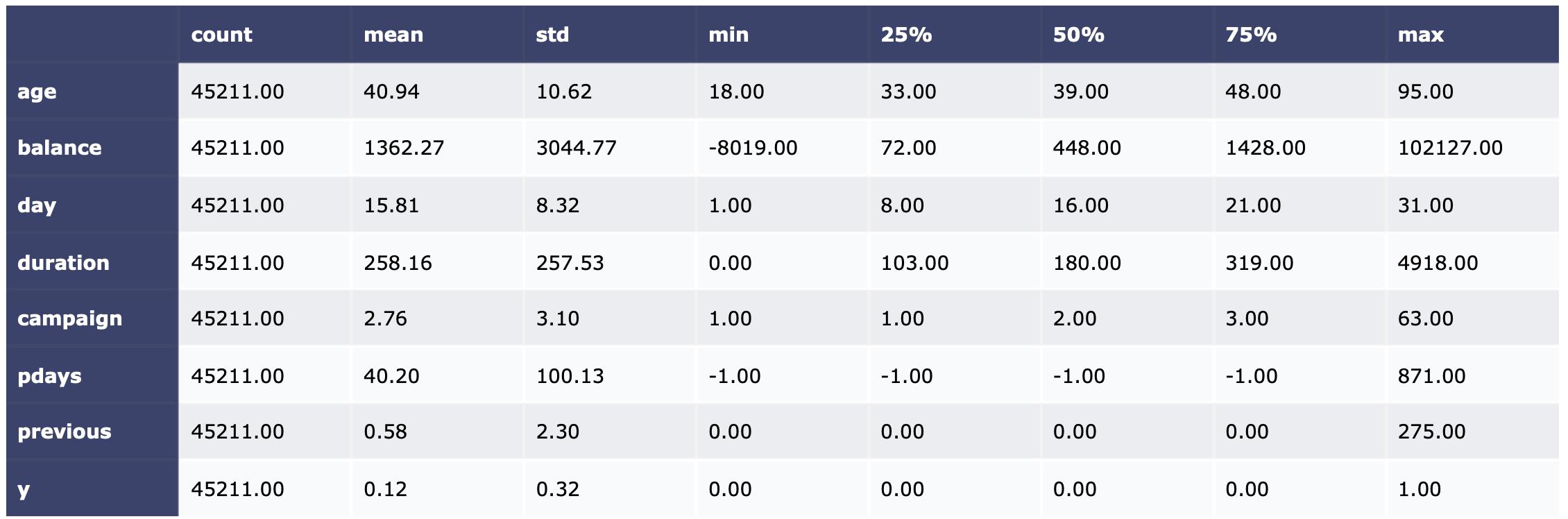
This dataset contains data from a marketing campaign that aimed to capture time deposits. With this data, it is possible to use classification techniques and profile customers. The data of the clients approached were divided into 17 attributes, being 8 numeric classes and 9 of factors. The campaign reached 45211 customers and at the end of the campaign 5286 customers made the deposit.

Summary Information about the variables and their types in the data:

1. Age: The age of the customer
2. Job: Type of job
3. Marital
4. Education
5. Default: Has credit in default? (Yes/No)
6. Balance: Average yearly balance
7. Housing: Has housing loan? (Yes/No)
8. Loan: Has personal loan? (Yes/No)
9. Contact: Contact communication type (unknown/telephone/cellular)
10. Day: Last contact day of the month
11. Month: Last contact month of year
12. Duration: last contact duration, in seconds
13. Campaign: number of contacts performed during this campaign and for this client
14. Pdays: number of days that passed by after the client was last contacted from a previous campaign
15. Previous: number of contacts performed before this campaign and for this client
16. Poutcome: outcome of the previous marketing campaign (unknown/other/failure/success)
17. Y: Has the client subscribed a term deposit? (0=No,1=Yes)



First 5 Rows of Data



Descriptive statistics

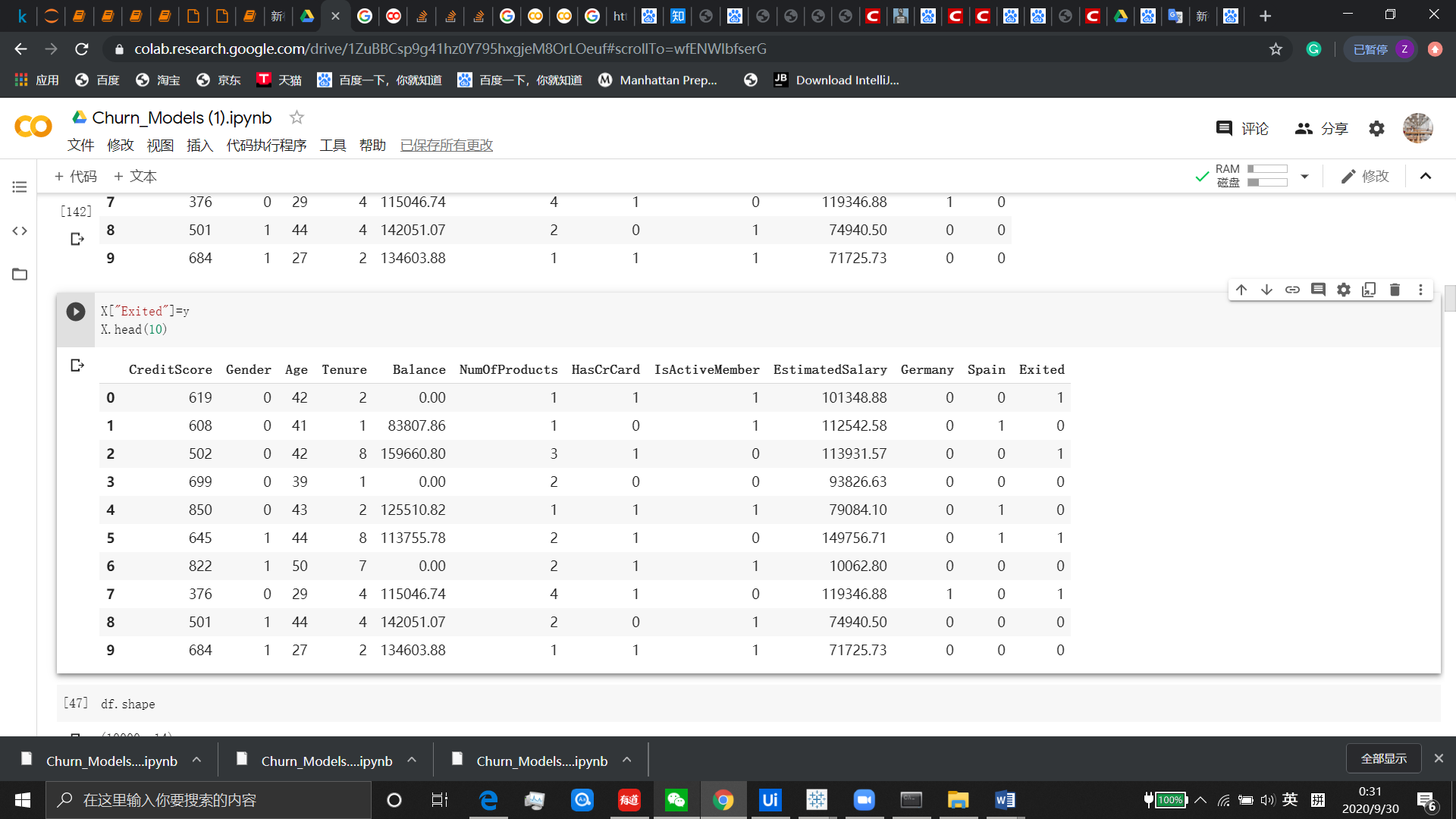
## 4.3 Model Building - Customer Churn Prediction

### 4.3.1 Data Preprocessing

Before build machine learning models, we firstly process dataset as follows:

1. Find missing value and outliers and then delete them or replace them with average of the other values
2. Remove irrelevant variables (RowNumber, CustomerID, Surname)
3. One-hot encoding on categorical variable (Gender and Geography)

By now, we get a processed dataset below (10000 rows x 11 columns):



And then:

1. Select dummy variable (Exited) as y, and others as X
2. Split the data frame on a train and test basis with 0.7 split ratio
3. Scale X of trainset and testset to around 0 to 1 with MinMaxScaler().

The main idea behind normalization/standardization is that variables that are measured at different scales do not contribute equally to the model fitting & model learned function and might end up creating a bias. Thus, to deal with this potential problem feature-wise normalization such as MinMax Scaling is usually used prior to model fitting.

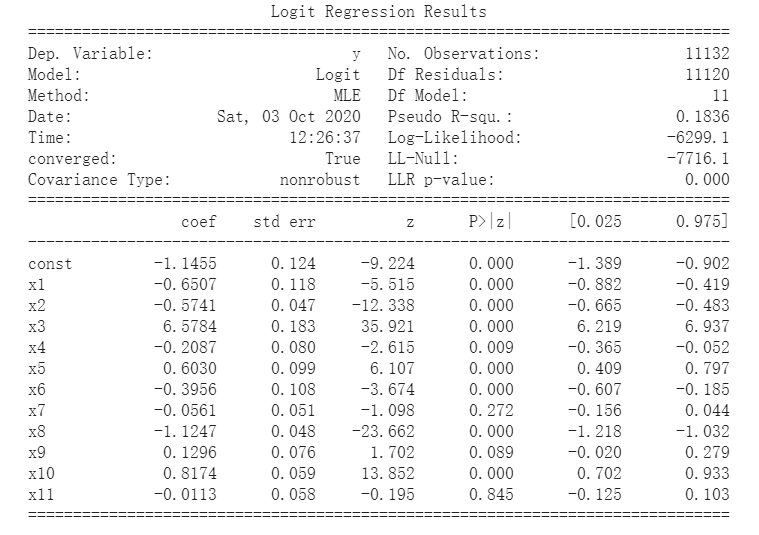
1. Over-sampling on trainset.

In our original data, the ratio of the number of y=0 to y=1 is around 4:1 indicating that the dataset is imbalance, which is confirmed in the subsequent model fitting. Besides, the dataset is not so big (tens of thousands of records or less), and to avoid removing necessary sample data, we use SMOTE algorithm to conduct over-sampling instead of under-sampling on trainset (SMOTE works by creating synthetic samples from the minor class instead of creating copies. The algorithm selects two or more similar instances (using a distance measure) and perturbing an instance one attribute at a time by a random amount within the difference to the neighboring instances). As for resampled ratios, we firstly follow the default, 1:1, and we will make adjustment on it if the classification result is not satisfactory.

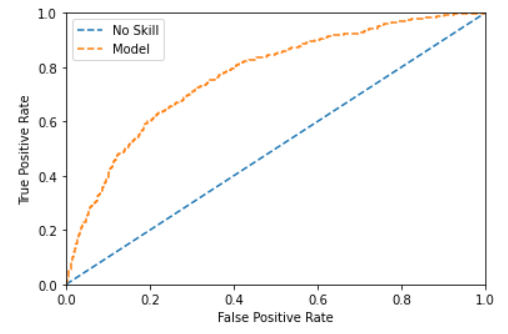
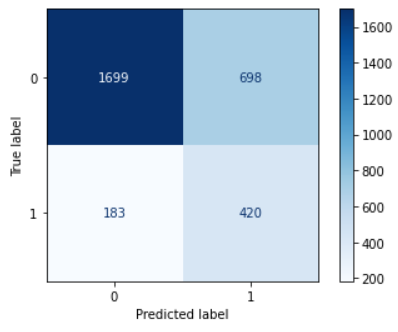
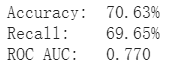
### 4.3.2 Model Fitting (5 models)

### 4.3.2.1 Logistic Regression

We firstly use the simplest model as a basis on comparison since it can process both numeric and categorical predictors, and also we can easily assess the significance level through P-value and feature importance through coefficients, which are shown below.

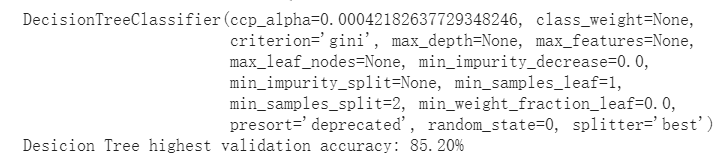


It’s notable that 3 variables(x7(HasCrCard), x9(EstimatedSalary), x11(Spain)) are not statistically significant at 5% level, so we remove these variables and fit logistic model again to improve its performance. The result (confusion matrix on test set, accuracy, recall, and AUC score) is as below.



### 4.3.2.2 Decision Tree

We then build the decision tree model. We first set the tree to the max with no pruning, setting cpp\_alpha=0. This creates a very large tree and the accuracy on trainset is very high (100%) but the accuracy on test set is low, only 66.3%. Apparently, this decision tree without pruning is overfitting, so to conduct post-pruning on the decision tree, we plot the relationship between ccp\_alpha and accuracy for training and testing sets as graph below. In the meanwhile, we get the highest accuracy score (85.20%) and corresponding ccp\_alpha (0.00422). Then we calculate feature importance to select significant features and run the model again.



With the ccp\_alpha (0.00422), we conduct post-pruning and reach a better result below.

|  |  |  |  |
| --- | --- | --- | --- |
| Indicators | Accuracy | Recall | AUC |
| Decision Tree | 76.53% | 74.96% | 0.831 |

### 4.3.2.3 Random Forest

We then use random forests model on the same dataset. Random forest grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest). It always performs well on classification. We adjust resample ratio and parameters in random forest model, like n\_estimators and max\_depth, and after feature selection, we obtain a satisfactory result below.

|  |  |  |  |
| --- | --- | --- | --- |
| Indicators | Accuracy | Recall | AUC |
| Random forest | 77.67% | 72.31% | 0.844 |

### 4.3.2.4 XGBoost

Next we turn to XGBoost, whose name stands for eXtreme Gradient Boosting, is a scalable and accurate implementation of gradient boosting machines and it has proven to push the limits of computing power for boosted trees algorithms as it was built and developed for the sole purpose of model performance and computational speed. Specifically, it was engineered to exploit every bit of memory and hardware resources for tree boosting algorithms. So we expect high on XGBoost. Adjusting parameters and resample ratio and after feature selection, we get a desirable result.

|  |  |  |  |
| --- | --- | --- | --- |
| Indicators | Accuracy | Recall | AUC |
| XGBoost | 80.40% | 72.14% | 0.859 |

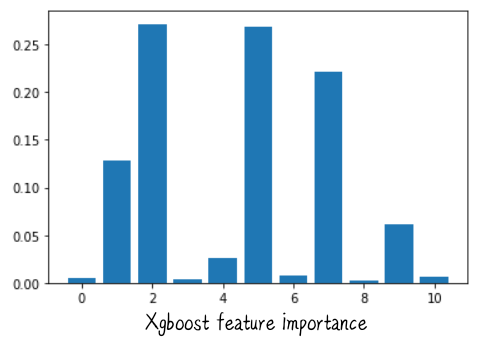
### 4.3.2.5 Neural Network

Lastly, we use artificial neural network for classification. ANN processes records one at a time, and learn by comparing their classification of the record (i.e., largely arbitrary) with the known actual classification of the record. The errors from the initial classification of the first record is fed back into the network, and used to modify the networks algorithm for further iterations. ANN always do well in classification. We set parameters in ANN (epoches=300, batch\_size=128 and dropout (0.2)) and get a good result below.

|  |  |  |  |
| --- | --- | --- | --- |
| Indicators | Accuracy | Recall | AUC |
| Neural Network | 76.45% | 71.84% | 0.838 |

### 4.3.3 Significant features identification

To identify which features is the most significant predictors, we conduct feature importance analysis on logistic regression, decision tree, random forest and XGBoost. Also we use it for feature selection to improve the efficiency and effectiveness of predictive models above.



## 4.4 Model Building - Customer Segmentation

4.4.1 Clustering Algorithm: Partitioning Around Medoids (PAM)

The K-Medoids algorithm or Partitioning Around Medoids (PAM) is a clustering approach related to K-Means clustering method for partitioning data into k groups. In K-Medoids clustering, each cluster is represented by one of the data points, the most centrally located point, in the cluster where they are called cluster medoids. This is very different with K-Means which uses mean value as the clustering centroid thus very sensitive to outliers in the dataset. K-Medoids algorithm is a more robust method when implementing clustering with outliers compared with K-Means algorithm.

In this report, we use K-Medoids algorithm to implement clustering. The process can be divided into two phases: build phase and swap phase. In build phase, the first step is to select k objects to become the medoids. After finding a set of k medoids, clusters are constructed by assigning each observation to the nearest medoid. In swap phase, all the algorithm does is to improve the quality of clustering by exchanging the selected object (the medoids) and non-selected objects.

The swap process will continue until the objective function cannot be reduced anymore. Then we can assume that the intra cluster distance and the inter cluster distance achieve a minimum and maximum respectively.

The pam() function in R can help us partition data into k clusters under K-Medoids algorithm. By grouping the clusters generated after pam() function, we can view the summary statistics for each cluster and derive some common patterns for clients within a cluster.

4.4.2 Similarity Measurement

In previous learning, the similarity calculation function like Euclidean distance only works for numeric values. However, in our dataset, there are 5 fields that are categorical combined with 2 numeric columns. Thus, the challenge we need to take is how to cluster on mixed type data. Among many distance metrics, one is actually quite useful to crack our case, the Gower distance (1971).

Gower distance computes the average of partial dissimilarities across the individuals, be it the numeric or categorical ones. Each partial dissimilarity (the Gower distance) ranges in [0,1]. The formula for Gower distance is shown as below:

Gower distance’s formula

represents partial dissimilarities that computes the distance across all features between two data points depending on the type of variable being evaluated. For numerical feature f, partial dissimilarity equals to the ratio between absolute difference of observations and and maximum range observed from all individuals (max\_N(x) — min\_N(x)). The formula is shown as below, N is the number of individuals in the dataset:

Partial dissimilarity computation formula for numeric feature

For categorical feature f, partial dissimilarity equals to 1 when observations and have different values and equals to 0 otherwise.

In R, we use daisy() function from “cluster” package to derive the Gower distance. By running the daisy() function, we can derive a matrix with all the distance values called “gower\_dist” which will then be passed to pam() function as an input argument.

4.4.3 Decide Number of Optimal Clusters, K

Clustering algorithm like K-Medoids requires us to predefine the number of clusters, K, like what we need to do in K-Means. It is very significant to find the optimal number of clusters in clustering analysis. If K is too high, we can only observe broadly scattered clusters with no meaning. If K is too low, the data points may be clustered incorrectly and thus cannot conclude each cluster with the most presentable features. Therefore, it is even more impossible to find our target group, high-value customers. There is no absolute answer for find the optimal number of clusters since it highly depends on the dataset distribution shape, scaling of the dataset and clustering number required by user (i.e. physical significance of K). The approach we use to examine the optimal K is called Silhouette Coefficient.

Silhouette Coefficient is a very popular method used to decide the number of clusters that a dataset should formulate. Silhouette Coefficient is computed on each instance and the formula is shown as below:

Silhouette Method

y depicts the mean distance to other instances in the same cluster. x depicts the mean distance to the nearest cluster (i.e. the next closest cluster). The calculated value for Silhouette Coefficient ranges from -1 to 1. When the coefficient closes to 1, it means that the instance the coefficient check is a part of the right cluster. When the coefficient closes to -1, it means that the instance is a part of a wrong cluster. For different K values specified, the system can eventually generate a final Silhouette Coefficient for each K. While choosing K with the highest Silhouette Coefficient, we need also consider K with almost good Silhouette value but simple clustering. Because Silhouette Coefficient calculate on each instance thus is very computation expensive, we only try K from 1 to 5. The plot of Silhouette Coefficient is shown as below:

Chart, line chart

Description automatically generated

Plot of Silhouette Coefficient

From the plot, since K=5 has the highest Silhouette Coefficient 0.27, we finally define the number of clusters to be 5.

4.4.4 Visualization on high-dimensional dataset

In our dataset we do clustering on high-dimensional data, thus we need to use t-Distributed Stochastic Neighbor Embedding (t-SNE) to perform dimensionality reduction which is suitable for visualization of high-dimensional datasets. The principal of t-SNE is to maximize the distance between observations that are most different in a high dimensional space. Because of this, observations that are similar in the high dimensional space may become clustered in a low dimensional space. And how t-SNE defines the similarity in high dimensional space is according to the generated Gower distance. The visualization of our five clustered customer groups is shown as below:

**Chart, scatter chart, bubble chart

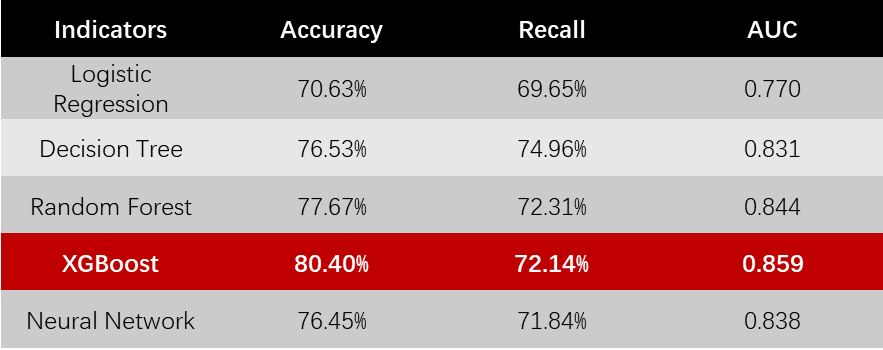
Description automatically generated**

Clusters observed in a lower dimensional space

# **5. Result and Performance Evaluation**

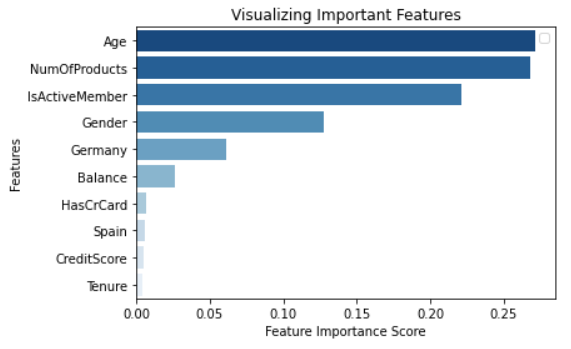
## 5.1 Customer Churn Prediction

For customer churn prediction, we fit 5 models on our dataset and now we compare results of each model. Graph below shows the accuracy, recall, and AUC score on tests et of each model, which stand for models’ ability to predict true positive and true negative.



From graph above, we can include that XGBoost, Random Forest and Decision Tree perform well on test set, which means they will do well in customer churn prediction. Specifically, we intend to choose **XGBoost** to be our optimal customer churn prediction, because it shows a highest accuracy, a high recall and an optimal AUC score, showing that XGBoost is a desirable classifier on customer churn prediction. Of course, in real business context, we will trade off importance of True Positive and True Negative rate according to requirement of enterprises and other real business factors to make an optimal choice.

For significant features identification, because we have chosen XGBoost as the optimal model, we calculate feature importance scores on XGBoost and then get a diagram below showing which features are most significant when conduct customer churn prediction, which means these features has an important impact on customers’ choice to close account or not. Specifically, we should focus on **Age, NumofProducts, IsActiveMember, Gender, Germany** (they contribute 95% to the model) and put forward some useful suggestions correspondingly.



## 5.2 Customer Segmentation

### 5.2.1 Customer Portrait

We partition the dataset with 45211 observations into 5 clusters and conclude each cluster with a specific pattern which is helpful in recognizing the high-value customer group we target with. Following is the statistical summary (please refer to Appendix4 for the detailed statistical result) and customer portrait for each subgroup:

**Cluster 1 (successful young people):**

This group of customers are the most highly educated (94% have tertiary degree) among the five clusters. Though second youngest, they have achieved great success in their career, with 65% management, 10% technician and 7% entrepreneur. The majority are married and all of them have house loans. This group has the highest growing value and the company will make great revenue if they are able to retain these customers.

**Cluster 2 (highest default risk group):**

This group of customers are the youngest and 70% of them are single, which can explain that only one third of them bear the burden of house loans. About half of the customers are technician, while the second and third highest are administration and services profession. 80% of them only got secondary degree, and their bank balance is among the lowest. Furthermore, this group has the highest default rate at 2.38%, while the other four groups are all leveled at around 1.87%. Therefore, we do not consider this cluster as high value customers.

**Cluster 3 (middle income group):**

Customers in this cluster are middle aged and almost all of them only have secondary degree. 40% of them are blue collars and the others work in administration (17%), services (15%) and technician (14%). The overwhelming majority has to pay house loans and their bank balance is not high. Overall, this cluster will contribute certain value, but not high, to the company.

**Cluster 4 (less educated blue collar):**

This cluster has the lowest education level – 89% of them only attended primary school. They’re the second oldest group and the majority of them work as blue collars. With low bank balance and high house loans to pay, we cannot see growth value in this group.

**Cluster 5 (wealthy class):**

This group of customers is the oldest among the five and well educated (59% obtained tertiary degree). Around half of them work in the management level and 76% are married. They are the only group without house loan burden and their bank balance is the highest. This group perfectly satisfies our definition of high value customers and the company shall make best efforts to retain these customers.

### 5.2.2 Customer Portrait Performance Evaluation

The data as a whole has a tendency to be clustered since the value of Hopkins statistic is far above the threshold 0.5. After clustering, the computed silhouette coefficient with number of clusters K=5 equals to 0.27, which means clusters are almost well apart from each other and are clearly distinguished.

# **6. Analytics strategy**

The previous sections have illustrated the methodology and performance of the customer churn prediction model and the customer segmentation model. In practice, the two models will be used together to reduce the cost of potential customer churn. When a new customer enters the company, firstly we will apply the customer segmentation model to label the customer as either high value, high growing value, or low value, so that tailor made marketing strategies can be applied to each group. High value customers are the main source of revenue for the company, and therefore we shall pay special attention to their activities and keep close contact with them. Although high growing value customers cannot generate great value for the company currently, it’s highly likely that they will become high value customers in the near future, and hence we shall try to retain this group of customers. Low value customers can neither contribute a lot to the company at the moment nor have much growth value, so there is no need to spend extra time on them other than the general marketing strategy.

Once we have identified our target customers—including both high value customers and high growing value customers, we shall use the customer churn prediction model on these two groups to see if there is any indication of customer churn. The customer churn prediction model not only predicts which customers are likely to churn, it also tells which variables have significant influence on whether the customer will close the account. The influencing factors selected by XGBoost are from two categories, one is the identity information, including gender, age, and whether the nationality is German, which we cannot modify; another is related to customers’ behavior, such as the number of products held by customers, and whether the customer is an active member, which can be influenced by the marketing strategy. Knowing that, the marketing team can keep a close eye on the activity level of the target customer and the number of products on hand; if the activity level suddenly changes or the customer abruptly withdraws some products, it’s likely that the customer is about to churn and the marketing team shall take actions in advance to prevent it from happening. Suggested strategies include communicating with the target customers face to face to show your respect to them and better understand their demands. Specifically, for high value customers, in addition to updating them the latest products in time, it will be helpful if you can provide supplementary services, such as family health care services, or products related to overseas study for their children. For high growing value customers, since they don’t have much disposable income at present, they may be more interested in small amount investments with high return rate.

Except what has been mentioned above, general strategies shall be considered to reduce customer churn rate. First of all, it’s important for an asset management company to remain a high credit rating to attract customers. The risk management department should take responsibility to evaluate the risk of each product, in case there is any defect that will cause downgrade. Providing systematical training to employees will also reduce illegal or improper actions that may affect the credit rating of the company and hence drive customers away. Secondly, many customers switched to the competitors after closing their accounts, so it’s crucial to pay close attention to the competitors’ products and upgrade your products and services accordingly. Another strategy is to distribute questionnaires to customers regularly, especially high value and high growing value customers, to understand their needs and know where you can improve.

# **7. Future Actions**

The test results and analytics strategies are based on two public datasets, which may not well fit the asset management industry. To make our models more useful and provide more precise recommendations, we need to gather the actual customer data from White Rock to train our model. The information needed contains identity information, such as customer gender, age, nationality, education level, occupation, salary, marital status, family size, and behavioral information, including credit scores, account balance, number of products held, active times per month, debt amount, and whether the customer has churned or not. The same data will be used to build both the customer churn prediction model and the customer segmentation model. Since our aim is to predict whether a customer will churn or not rather than whether he/she has churned, it will be helpful to provide the characteristics of the churned customer before the churn happened to make our model more predictive. The parameters of the prediction model can also be changed according to White Rock’s business needs. For example, if White Rock considers type II error (mistakenly classified the churned customer as not churn) more expensive, we can adjust the parameters to reduce the corresponding error. For the customer segmentation model, we still cluster the customers into four or five groups to see their common characteristics within each group and identify the high value and low value customers. Based on our findings, the marketing team can specify the potential customer churn in advance and make efforts to retain the high value and high growing value customers.

After implementing our method for a few months, we will evaluate the actual performance of our model and improve the model accuracy if needed. We will also compare the cost saved by retaining only the high value and high growing value customers with retaining customers randomly to see whether our method is cost effective.

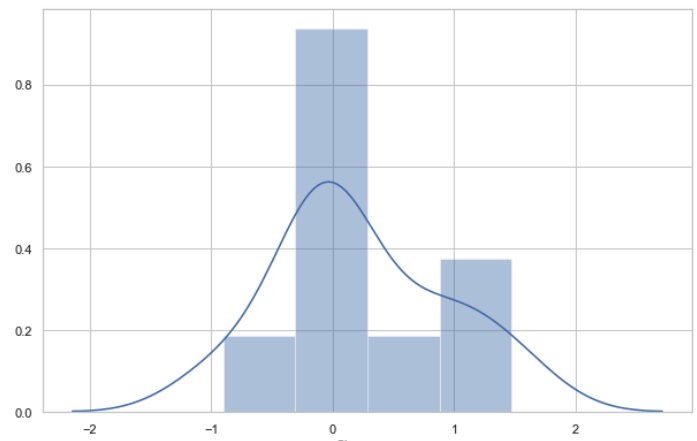
# **8. Conclusion**

The POC above has demonstrated that we are able to build models to predict customer churn with high accuracy and segment customers into different groups to make tailor made strategies. The uniqueness of our method is that we combine the supervised learning and unsupervised learning together to identify whether a potential churn customer is of high value to the company or not. In view of that the revenue generated by high value customers is much higher than that of low value customers, deploying our method will help White Rock save considerable cost and make great revenue. Firstly, our model can distinguish whether a newly acquired customer is of high value; if not, then there is no need to spend too much time on low value customers. Secondly, our model can predict whether a target customer (including both high value customers and high growing value customers) is about to churn with high accuracy, so that time will not be wasted on trying to retain customers who has no intention to churn. Finally, for the target customers who plan to churn, our model tells about the characteristics of this group so that the marketing team can apply specific strategies to retain the customers, which will increase the success rate to a large extent. The retained customers in turn will act as free word-of-mouth advertising and attract new high net worth customers to the company. To conclude, by selecting our model, White Rock will be one step closer to the leading digital asset management firm globally.

# **Appendices**

### Appendix 1: Customer Churn Prediction Dataset

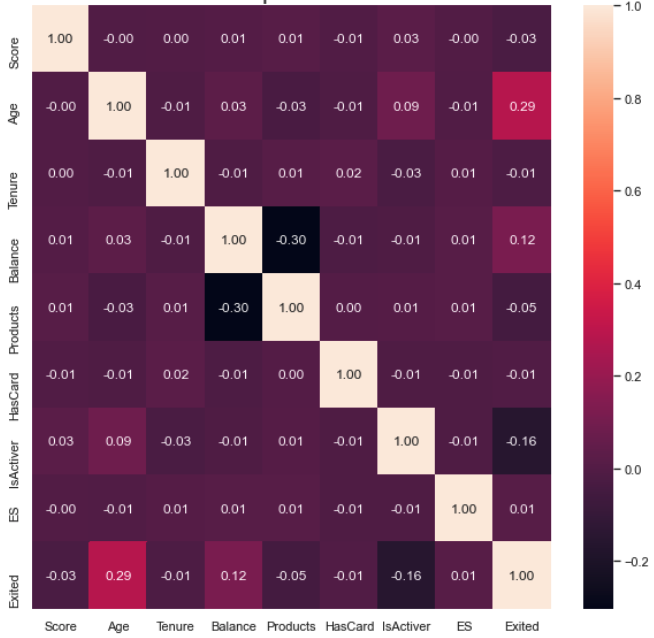
### Appendix 1.1.1: Skewness of all numeric data



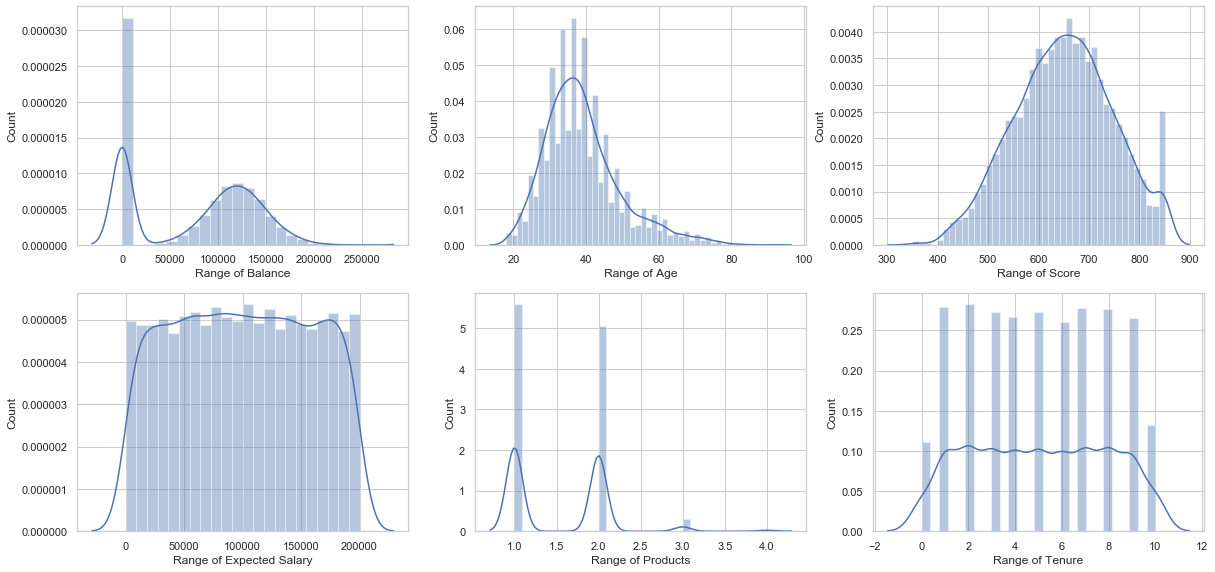
### Appendix 1.1.2: Kurtosis of all numeric data



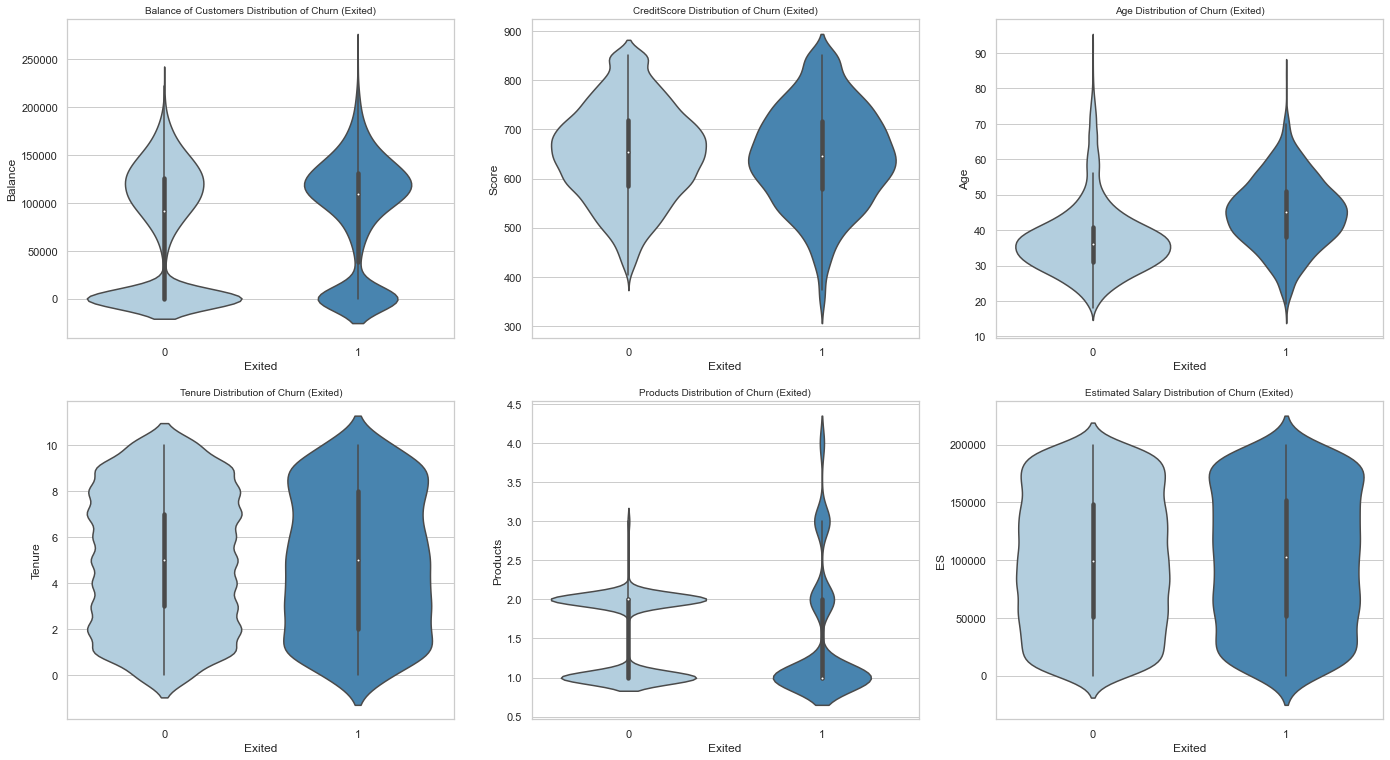
### Appendix 1.2: Heat map of data correlation



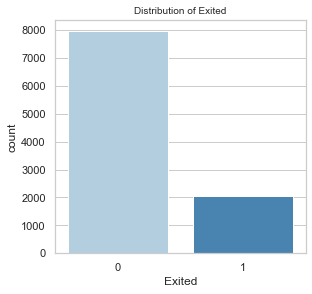
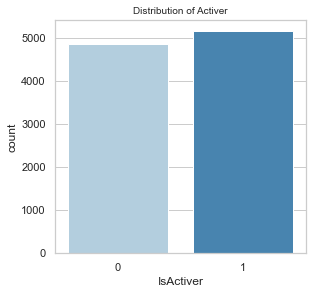
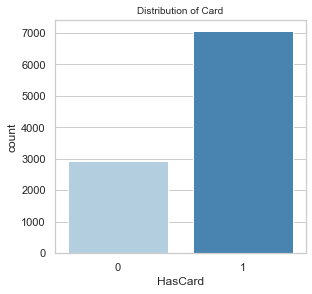
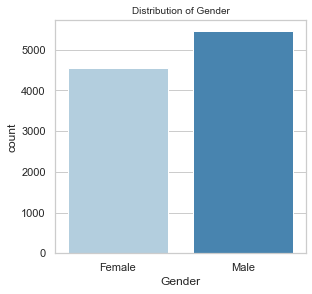
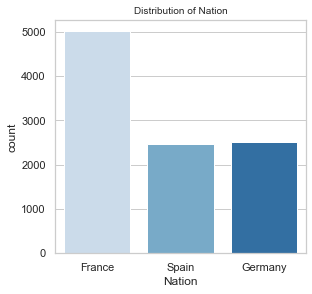
### Appendix 1.3.1: Simulated distribution curve of numerical variables



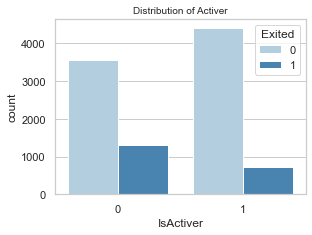
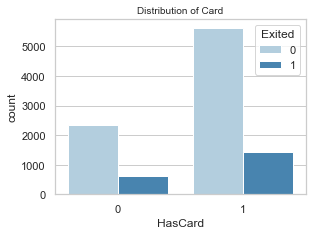
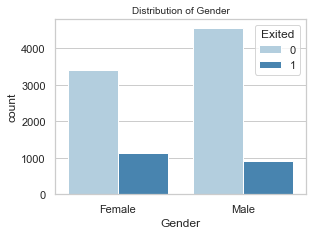
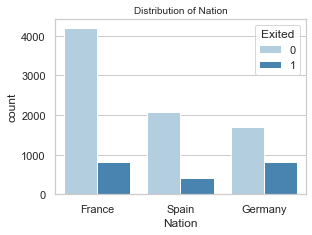
### Appendix 1.3.2: Distribution of numerical variables given different exited situation



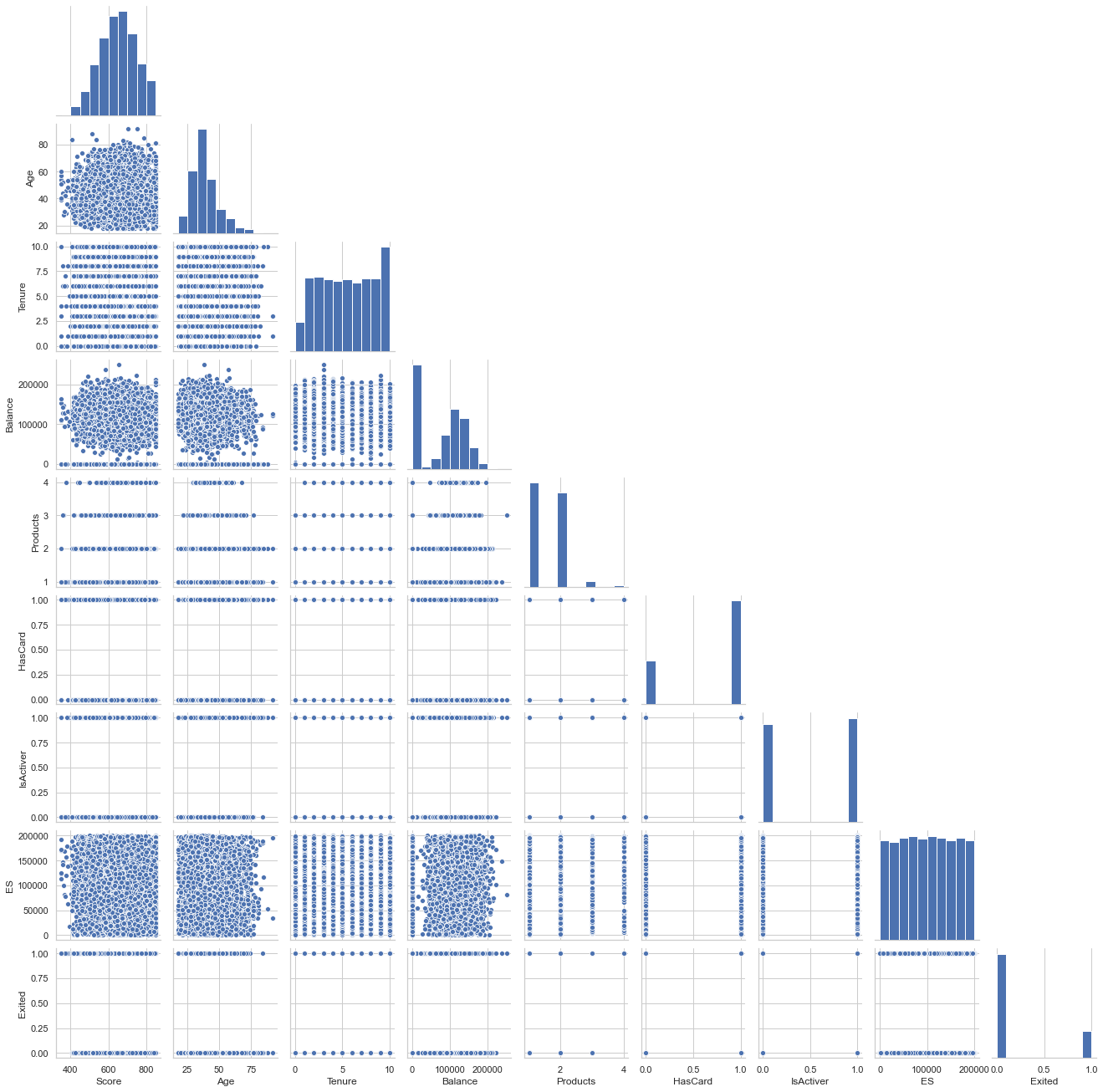
### Appendix 1.4.1: Distribution of categorical variables

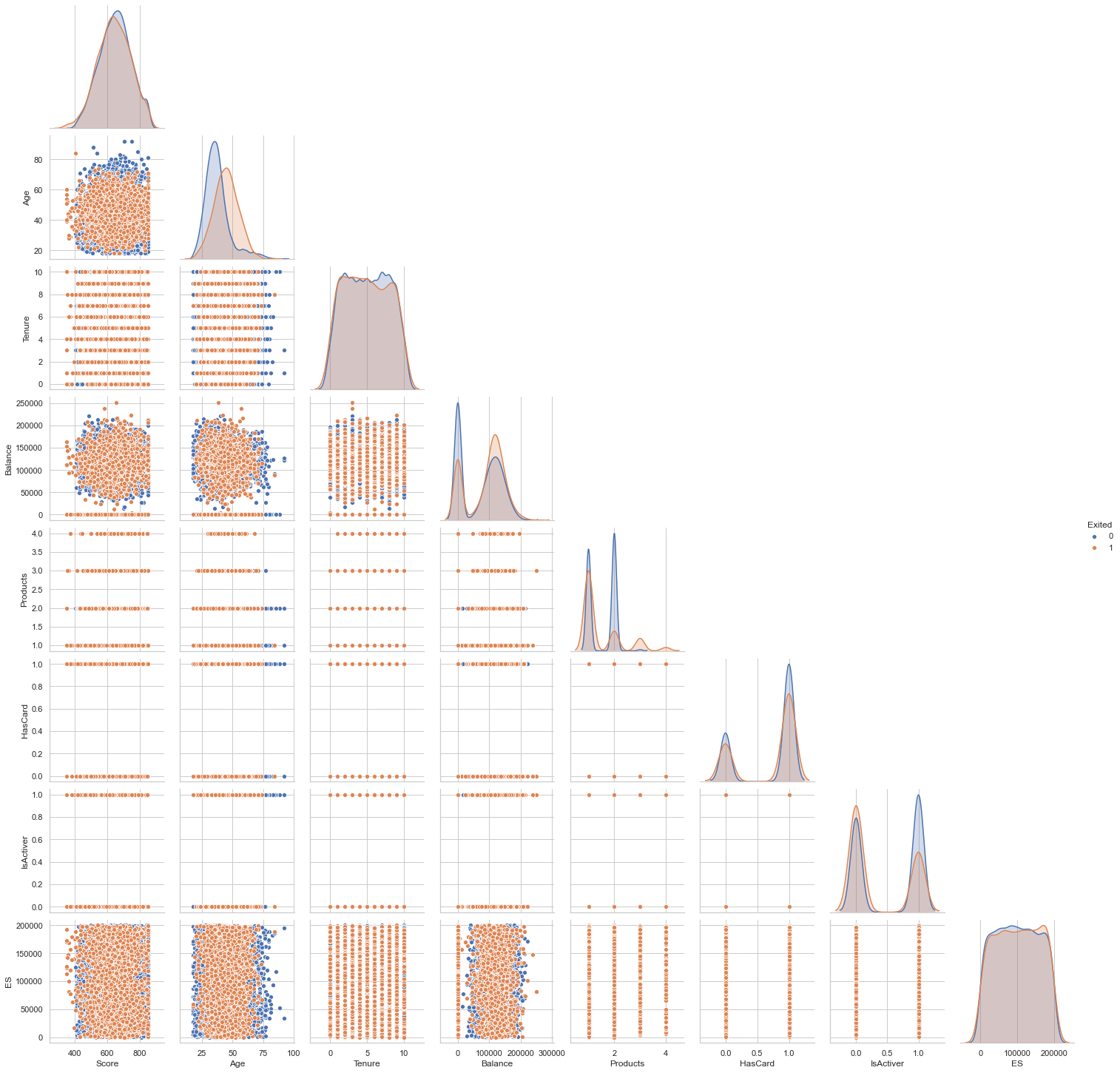


### Appendix 1.4.2: Distribution of categorical variables given different exited situation



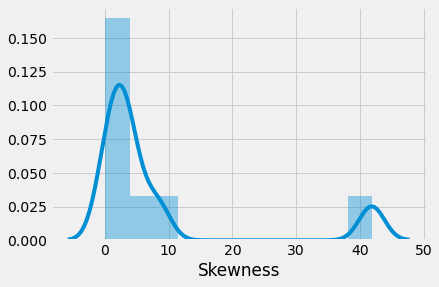
### Appendix 1.5.1: Pairplot



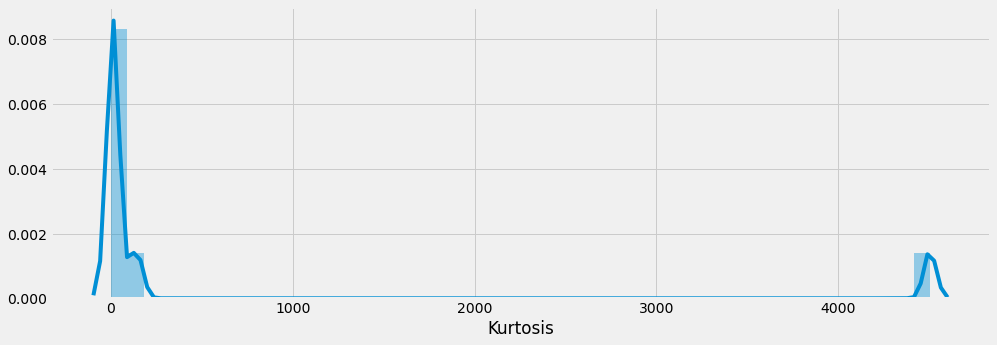


### Appendix 2: Customer Segmentation Dataset

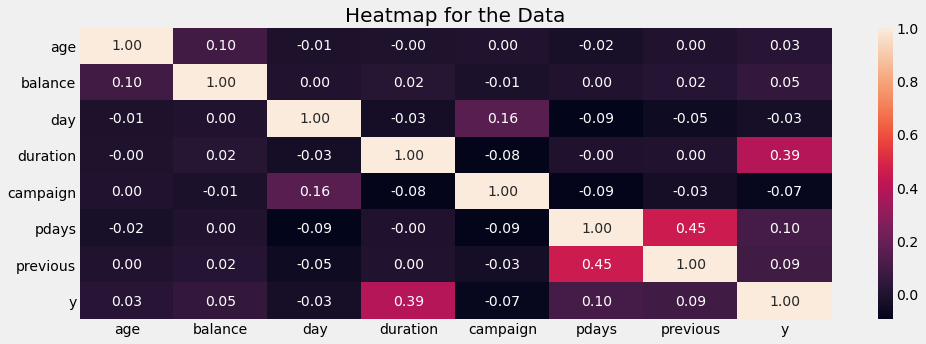
### Appendix 2.1.1: Skewness of all numeric data



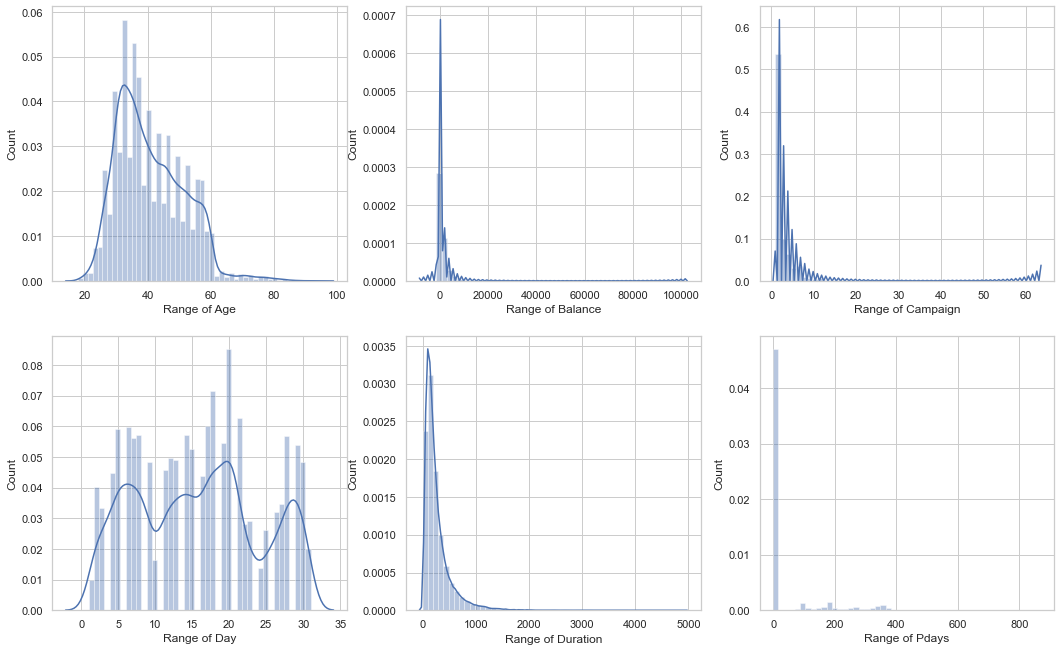
### Appendix 2.1.2: Kurtosis of all numeric data



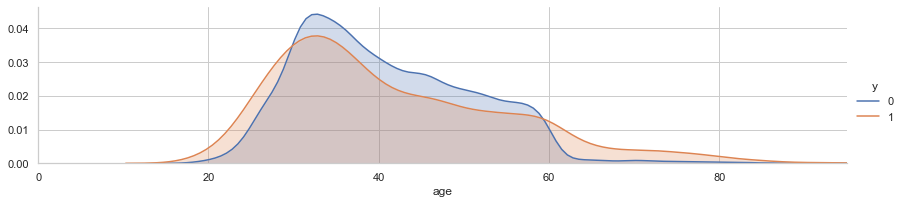
### Appendix 2.2: Heat map of data correlation

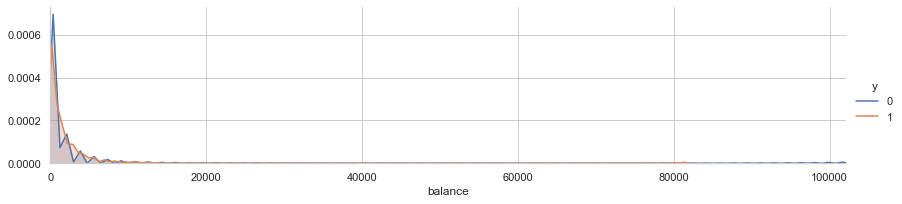


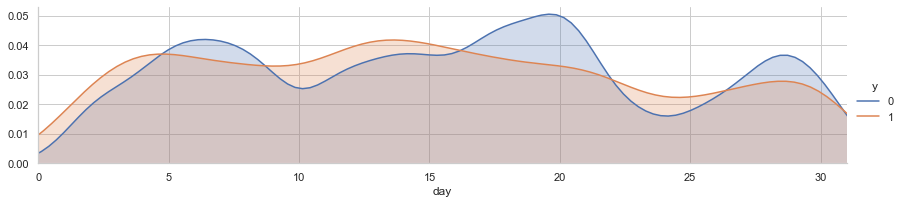
### Appendix 2.3.1: Simulated distribution curve of numerical variables

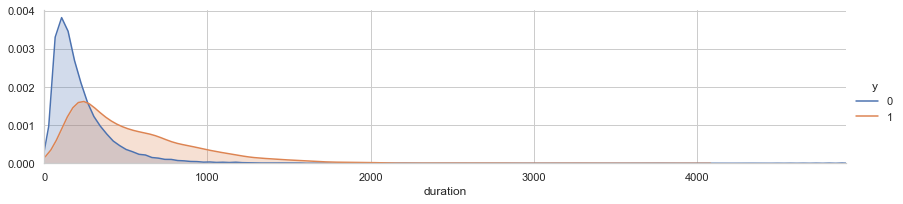


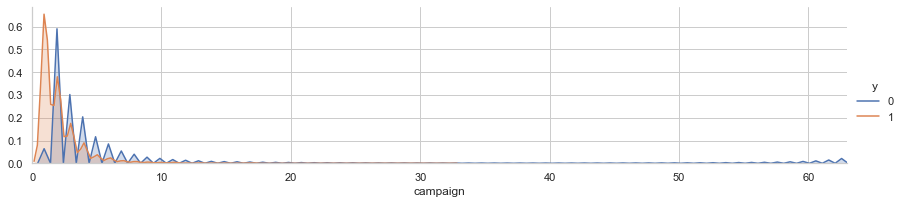
### Appendix 2.3.2: Distribution of numerical variables given different response to campaign

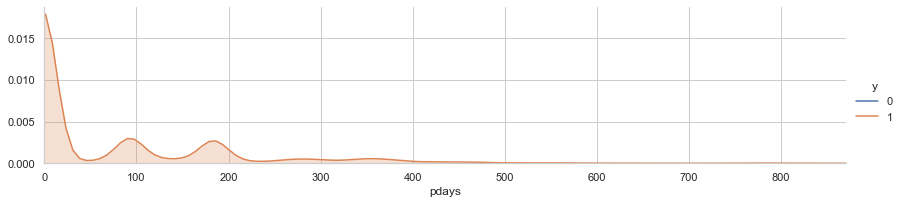


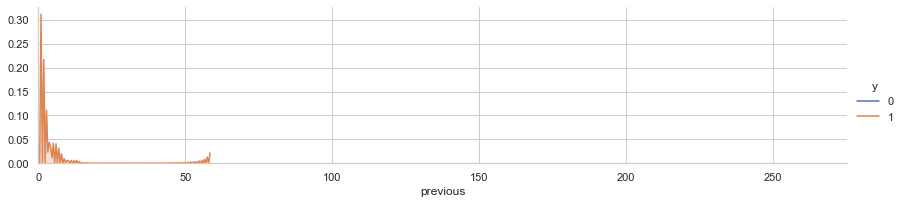


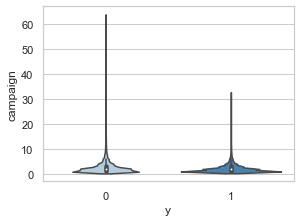
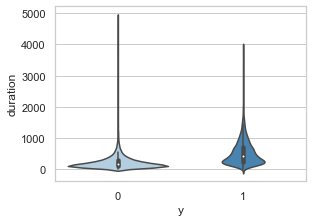
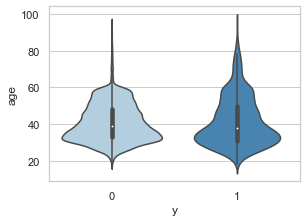




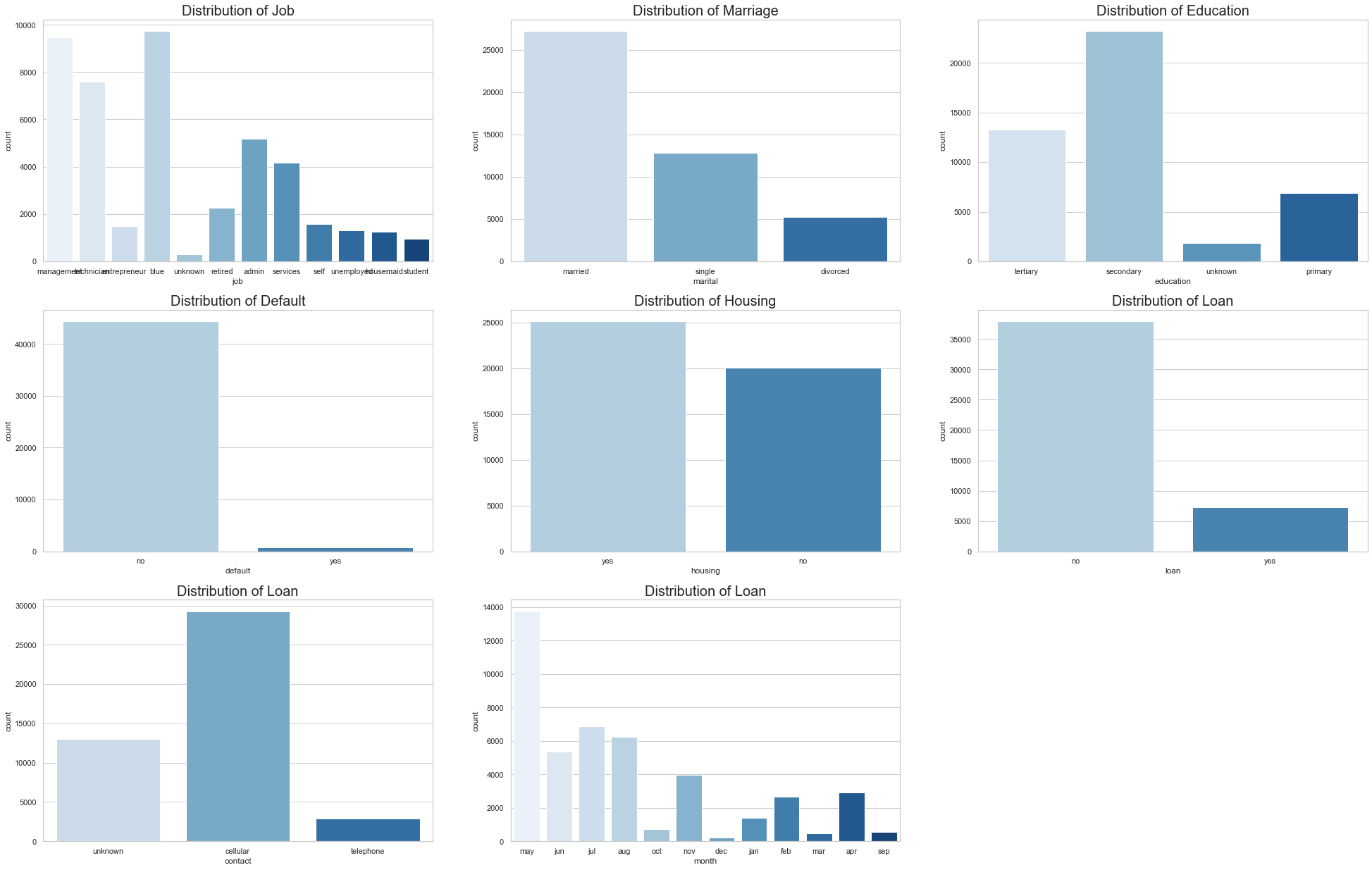




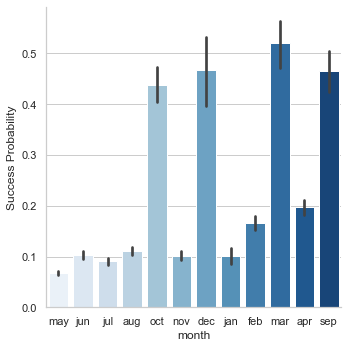
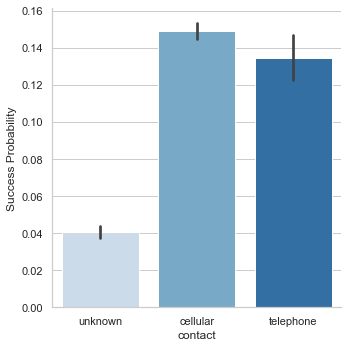
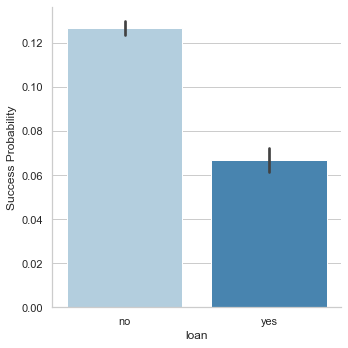
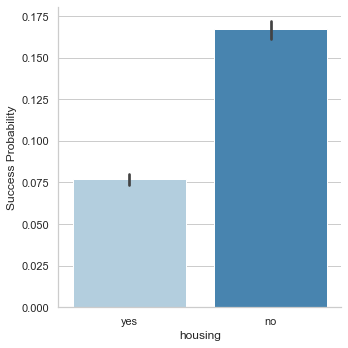
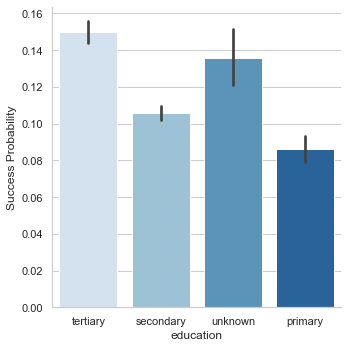
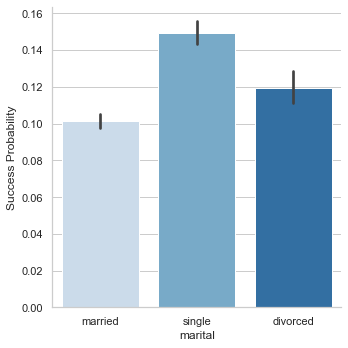
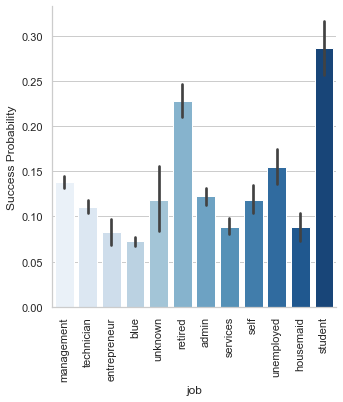




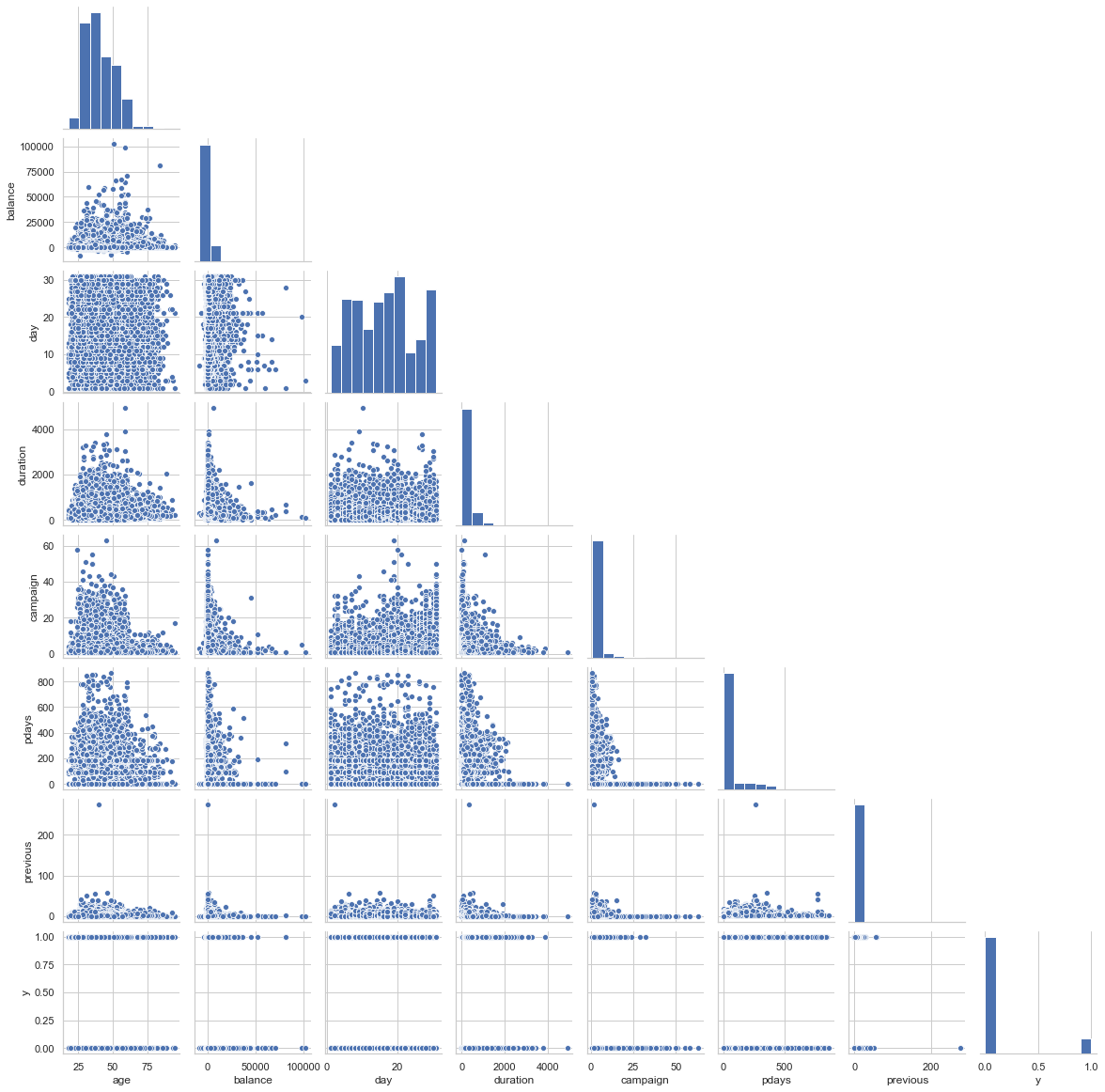
### Appendix 2.4.1: Distribution of categorical variables

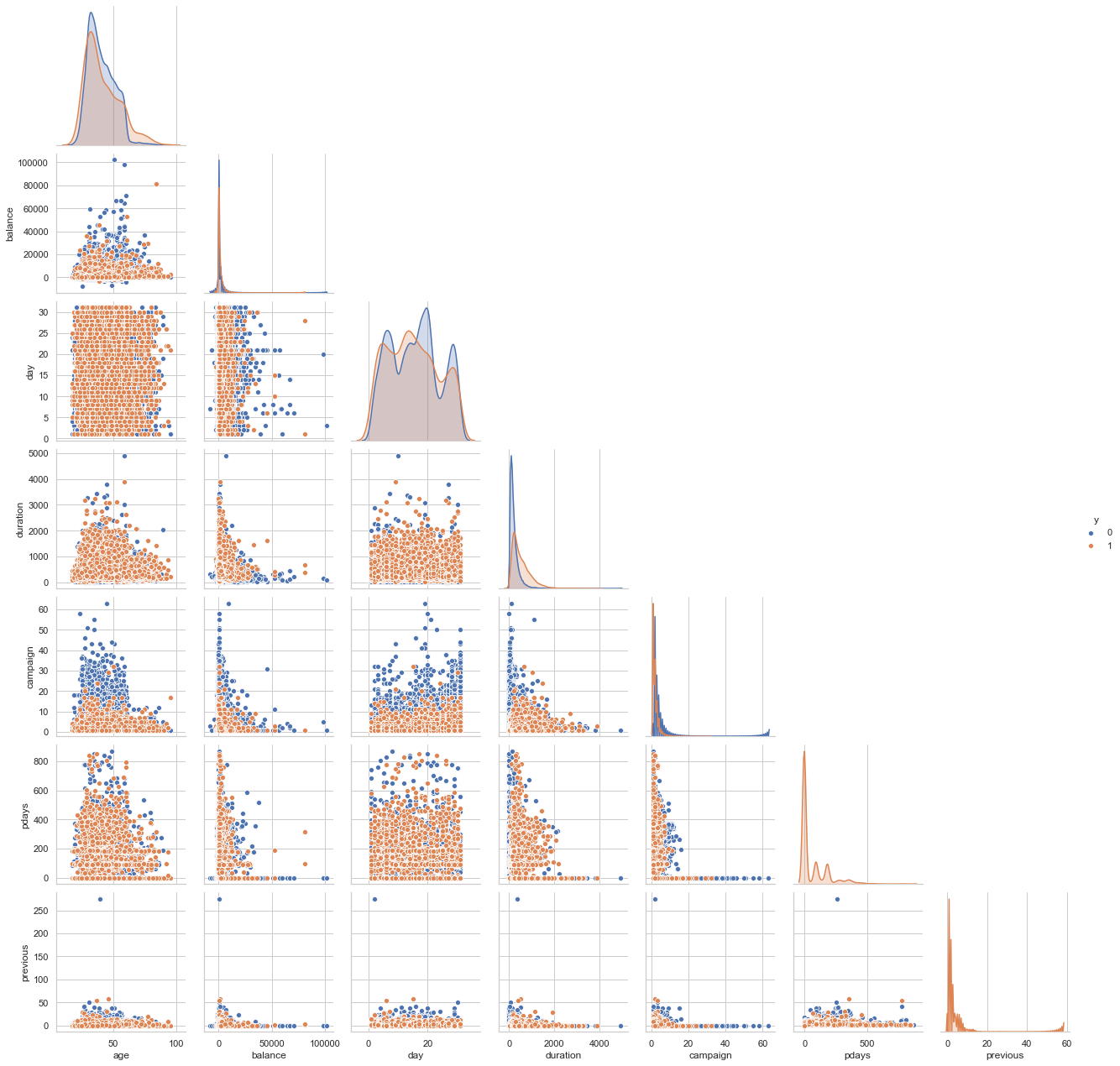


### Appendix 2.4.2: Success probability of categorical variables



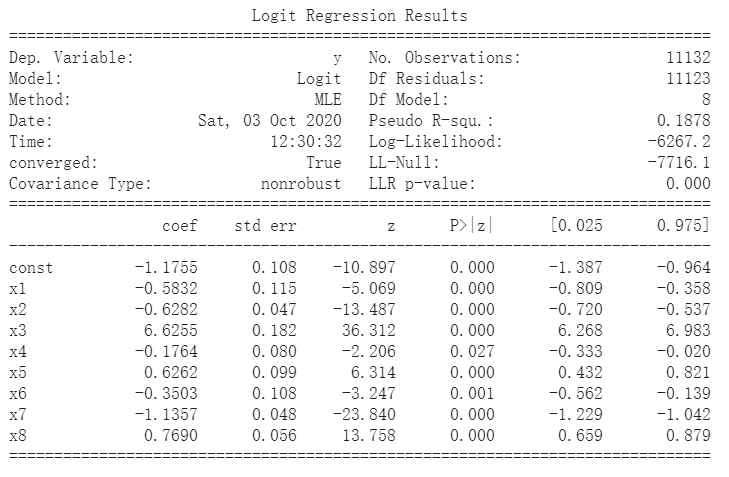
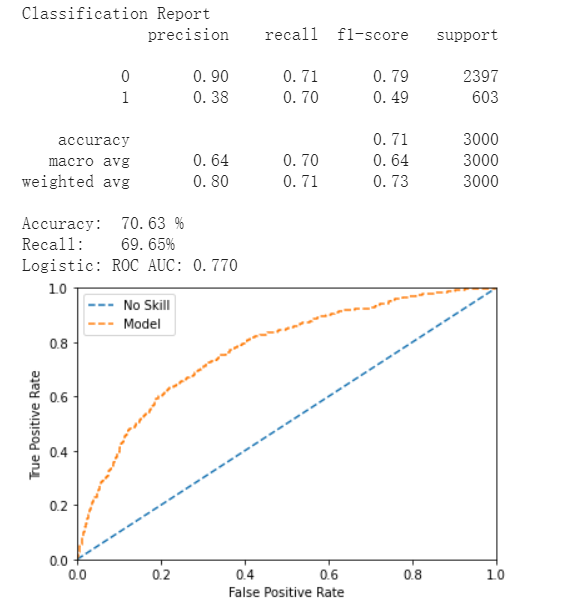
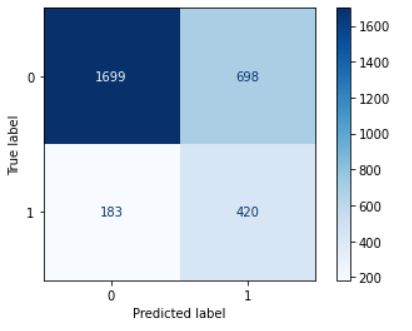
### Appendix 2.5.1: Pairplot





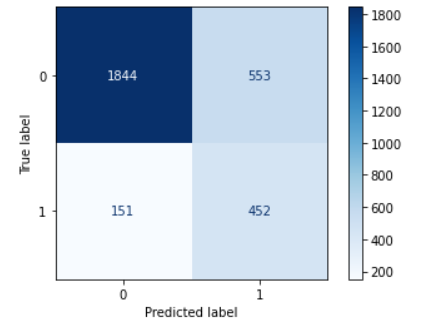
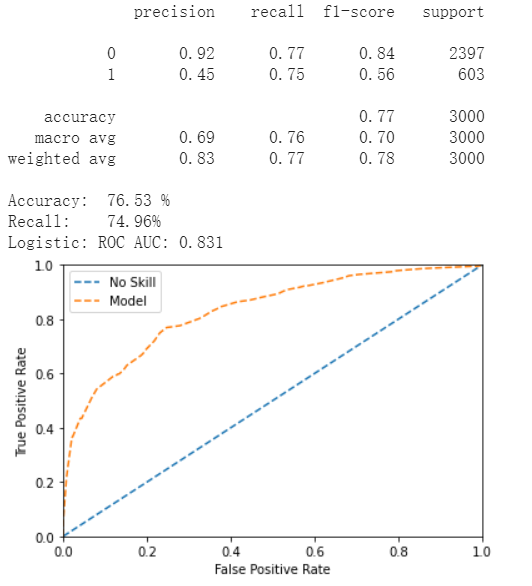
### Appendix 3: Statistical summary for prediction results

### Appendix 3.1: Logistical Regression

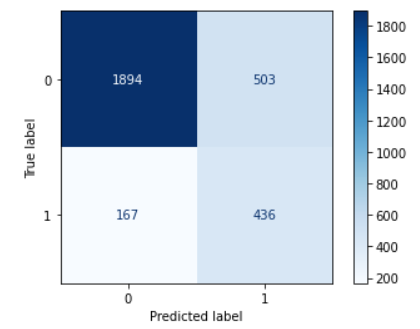
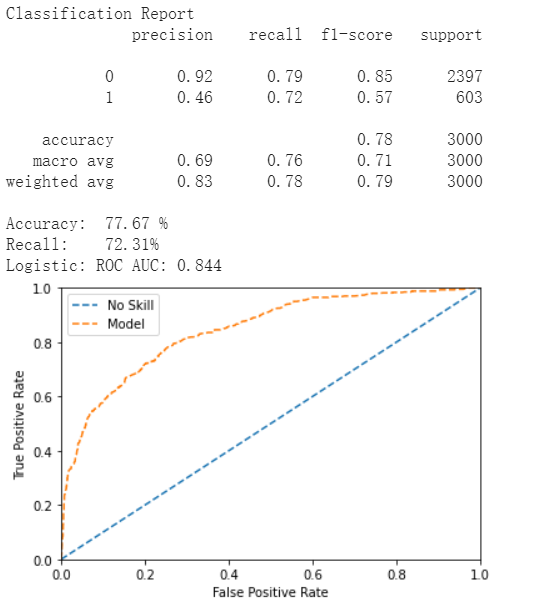
 

### Appendix 3.2: Decision Tree

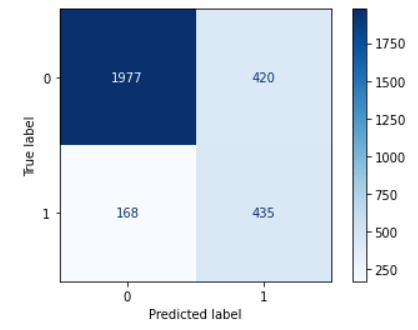
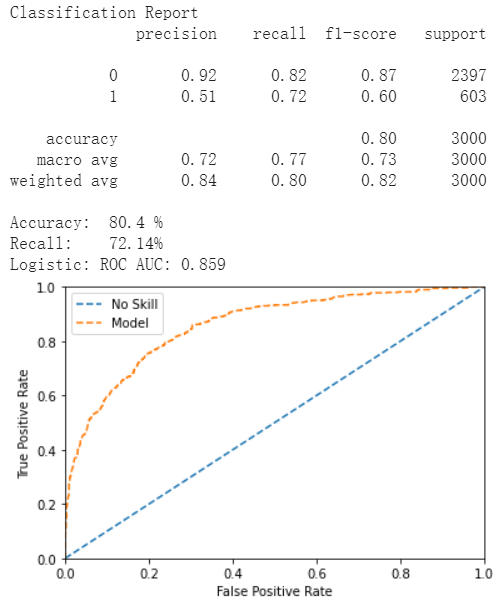


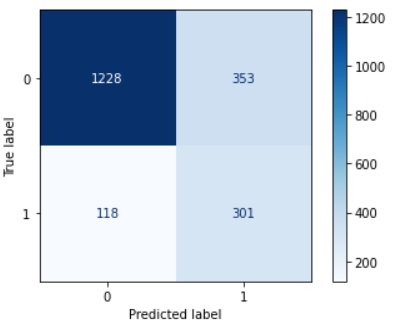
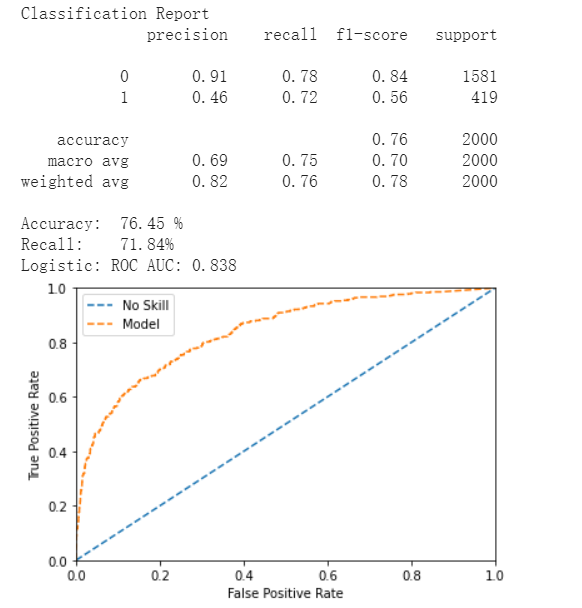
### Appendix 3.3: Random Forest

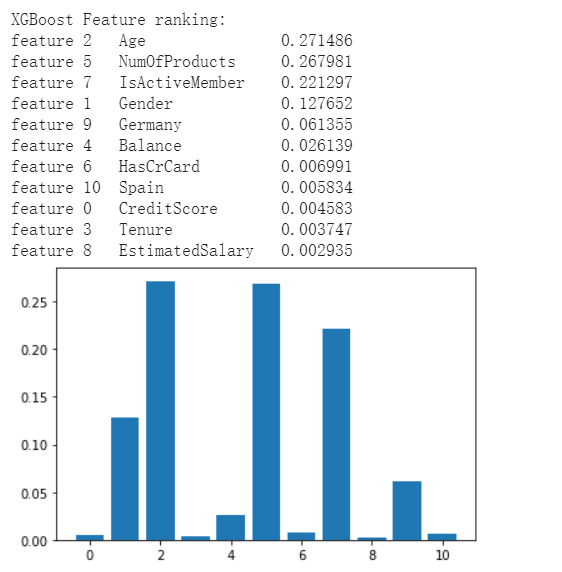
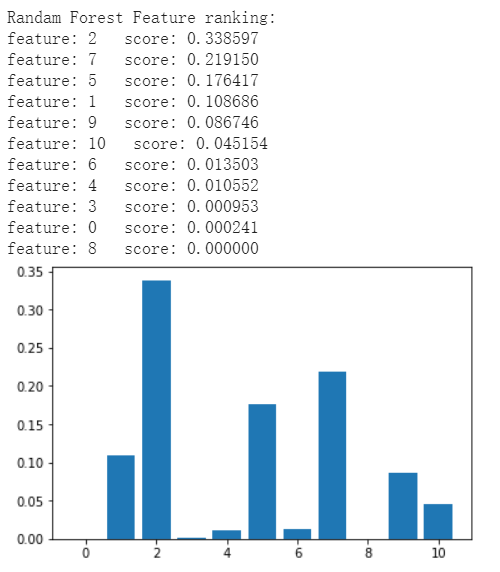
### Appendix 3.4: XGBoost

### Appendix 3.5: Neural Network

### Appendix 3.6: Feature Importance

### Appendix 4: Statistical summary for clustering results

Text, letter

Description automatically generated

Text, letter

Description automatically generated

Text, letter

Description automatically generated

Text, letter

Description automatically generated

A screenshot of a cell phone

Description automatically generated

# **References**

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert systems with applications*, *38*(3), 2354-2364.

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, *38*(12), 15273-15285.

Bilal Zorić, A. (2016). Predicting customer churn in banking industry using neural networks. *Interdisciplinary Description of Complex Systems: INDECS*, *14*(2), 116-124.

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research*, *34*(10), 2902-2917.

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, *36*(3), 5445-5449.