MGSC661 Final Project

BANK MARKETING ANALYSIS

Xintong Li & Ziye Zhang | MGSC661 | 2021-12-11

I. Intro

In this project, we will analyze data from the banking industry. The data is related to direct marketing campaigns, mainly phone calls, for term deposits of a Portuguese banking institution. A term deposit is a financial product where banks lock up customers' funds into an account for investing purposes. The way of executing a phone call to customers who will potentially subscribe to a term deposit brings returns that far outweigh the expenses. Nevertheless, it is time and cost consuming if banks blindly contact all the clients; some clients may even feel bothered when being consistently promoted products that they are not interested in, and thus banks will lose customer loyalty. Therefore, it is imperative and interesting to uncover the true potential clients and be able to achieve a high conversion rate using data-driven decisions.

Our team aims to build two types of models that will facilitate the decision-making of the bank. The first is the clustering models; one is on customer information aiming to help bank managers understand their customer segmentation, and the other is on economic data, such as the employment variation rate, and consumer price index, to help them understand the business cycle of the economy. The second is a classification model predicting if the client will subscribe to a term deposit using the client information and economic conditions provided in the dataset.

The following report consists of 4 parts. In the Descriptive Analysis section, we explore the underlying distribution of each predictor, as well as the relationship between the numerical predictors and the target variable which is a binary yes/no purchasing decision. By performing descriptive analysis, we can visually get a feeling for the data and examine the correlated predictors. In the Model Selection & Methodology section, we will explain the methodology and logic behind each model and how to select the best-performing model. In the Model Result Interpretation section, we interpret the result generated from our best-performing model. Finally, In the last Managerial Implications & Business Applications section, we identify the potential business applications as well as the underlying profits that can be generated by our models for the banking industry.

II. Descriptive Analysis

2.1 PERSONAL INFORMATION DATA EXPLORATION

Our dataset contains the data on the personal information of customers such as age, job, marital, education, and so on. By drawing histograms (Appendix 1-1 to Appendix 1-3) of these variables, we find that the majority of customers are between the ages of 30-50, and there are very few customers older than 75 or under 20 years old. Among all the customers, many people work in administration, blue-collar, services, technician industry and only a small number of customers works as housemaid or students. Among all the customers, most of them hold university degrees or high school degrees. The boxplot (Appendix 1-4) illustrates that the age distributions are not very different across different jobs except the distribution for students and for the retired, which are the youngest and oldest as expected.

We then draw a standardized stacked bar chart (Appendix 1-5 to Appendix 1-7) of these variables regarding the proportion of purchases. We can clearly see the shape of the distribution of age is non-linear, with the proportion of purchases decreasing dramatically from age 20 to 50, then increasing from age 50 to 65. The proportion of purchases reaches the minimum point at age 50. Furthermore, the distribution of age appears to be quadratic. This indicates setting its degree to 2 for the logistic regression model which requires the assumption of linearity. Besides, from the chart of the proportion of purchases by contact type (Appendix 1-8), we can also see that most customers do not purchase after being contacted by cellular or telephone, but the proportion of customers who are willing to purchase after being contacted by cellular is still higher than those who are contacted by telephone.

2.2 TIME VARIABLES DATA EXPLORATION

In this section, we explore the data about customers’ purchasing preferences regarding the month of the year and day of the week. By drawing the stacked bar chart of these two variables (Appendix 1-9 and Appendix 1-10), we can see that people are more likely to purchase in March, September, October, and December. Moreover, May is the period when the purchases happen the fewest. Besides, from the proportion of purchases by the day of the week, we can observe that customers’ buying preferences for each day of the week don't vary significantly.

2.3 FINANCIAL VARIABLES DATA EXPLORATION

The heatmap (Appendix 1-11) reveals the correlations among the financial variables and with the target variable, we observe that Consumer Confidence Index (CCI) has a low correlation (0.05) with the target variable. Employment Variation Rate (ER) and the number of employees have a relatively higher correlation with the target variable.

By performing Principal Component Analysis (PCA) (Appendix 1-12) on the financial variables as well as the target variable, we can see the three-month Euribor Rate (Interest Rate), Employment Variation Rate, Consumer Price Index (CPI), and Number of employees are all negatively correlated with the target variable. Moreover, these variables appear to have collinearity issues. Hence, we consider removing two of them or using the first two principal components of these variables when we develop the models.

III. Model Selection & Methodology

3.1 CLUSTERING MODEL

Using the dataset, two clustering models are developed for two distinct purposes. The first extracts the financial variables including 3-month Euribor Interest Rate, Employment Variation Rate, CPI, CCI, and the number of employees. The second one groups customer profile using their age, job, marital status, education level, etc.

K-means clustering is used for both clustering tasks because of its ability to handle large datasets. One of the drawbacks of K-means is we must specify the number of clusters using domain knowledge, which is the knowledge and experience in finance and banking in our case. However, we can use several metrics to evaluate clustering; for both models, we deploy the elbow method and silhouette score for comparing the number of clusters. We attempt to both minimize within-cluster variation and maximize between-cluster variation. For the elbow method, we seek to find the elbow position where adding one more cluster does not improve much within-cluster variation; for the silhouette method, we compare silhouette scores and focus on the highest scores among all coefficients. Meanwhile, we use financial and customer knowledge to help determine the number of clusters.

The final financial variable clustering model contains 2 main clusters which mainly represent contraction and expansion in the economy. The silhouette scores are above 0.8 and do not improve much from 2 clusters to 5 clusters (Appendix 1-13) while the elbow method (Appendix 1-14) clearly indicates 2 clusters is the optimal choice since little within-cluster variation can be reduced. As for the final customer profile clustering model, 3 clusters, representing junior, mid-age, and senior customers, are selected. This model is reasonably good as it has a silhouette score of 0.48 and has much within-cluster variation reduction (Appendix 1-15 and 1-16).

3.2 CLASSIFICATION MODEL

Before we start modeling, the dataset is split into the training and test datasets. Since our data is an imbalanced dataset with only 10% of the customers purchasing term deposits. We apply the SMOTE method to resample and balance the training dataset and use it to train various models. Throughout the model selection process, each model's training accuracy is compared with its test accuracy to prevent overfitting. Hyper-parameter tuning is applied to required models (see Appendix 1-17 for an example of tuning for Gradient Boosting) Since identifying all purchasing customers is more of interest than identifying all non-purchasing customers, we use recall and precision to measure the robustness of our models. Precision is that among all predicted purchase customers, the probability of the model correctly identifying them as purchase, while recall is that among all those who truly want to purchase term deposits, the probability that our model would identify them as purchase. Finally, we will use a balanced metric, balanced accuracy, which accounts for the imbalance in classes.

Logistic Regression

In Logistic Regression, during model selection, we also apply the AIC criterion, which examines the model's goodness of fit while penalizing complexity. First of all, the baseline overfitted model where all predictors are regressed on the target variable is built. Multicollinearity is checked by running a VIF test. From the VIF result and the correlation matrix we discussed in the Data Description section, we remove the number of employees and employment variation rate, which are highly collinear with the other financial variables. We sequentially remove the predictors having an insignificant p-value, and compare the AIC of the models obtained in each step with the model selected by the Step AIC algorithm, which automatically removes predictors that do not contribute to the model. We also add age squared in the model based on our observation in Appendix 1-5. Finally, we compare the model with original financial variables and the one using the first two components of PCA on the variables, and it turns out that using PCA does not have improvement on the model.

Discriminant Analysis

Both linear and quadratic discriminant analyses are implemented to the training dataset with each predictor removed sequentially by examining the training and test balanced accuracy. The first two principal components are also used to replace the financial variables, but poorer predictive power is obtained.

Decision Tree

Firstly, a fully grown tree is built and the tree is visualized. There are many branches and leaves so we cut and prune the tree using the cost-complexity parameter, which is the minimum improvement needed for a node to split. Then, a pruned tree with the optimal cp is built and visualized. However, this tree still has many nodes, which makes the model hard to interpret. Therefore, we further simplify the tree by only including the important predictors and specifying the maximum depth and the minimum number of observations in leaf nodes. The decision tree model will be explained in detail in the Model Result Interpretation Section.

Random Forest

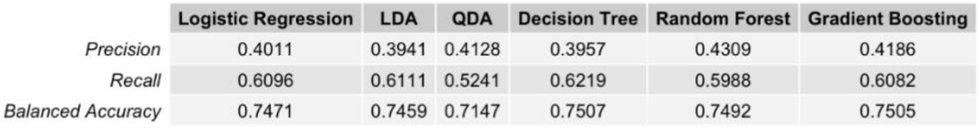
Random Forest is an algorithm that utilizes many small trees (weak learners) simultaneously and combines the results to reach the ultimate prediction. It is less prone to overfitting whereas the result is harder to interpret. Using random forest, we first select all predictors and then run a model using the top 9 predictors ranked by the variable importance (Appendix 1-18). Finally, we conduct hyperparameter tuning to find the optimal number of features used in building each individual tree.

Gradient Boosting

Gradient Boosting is an algorithm that builds many trees sequentially; each latter tree learns from the former trees and improves accordingly. Like the Random Forest, we first use all predictors and get the variable importance (Appendix 1-19). Next, we select the top 9 features in order to reduce model complexity. Similarly, we do cross-validation to grid-search the best hyperparameters including the number of trees, maximum depth of each tree, minimum number of observations in each leaf node, and learning rate.

Model Performance and Evaluation

Table 1 below summarizes the three metrics, precision, recall and balanced accuracy, for each model. Gradient Boosting has the highest precision score and the second highest balanced accuracy, and the recall score is moderately high. However, the Decision Tree surpasses Gradient Boosting in both recall and balanced accuracy, although its precision is not high. The other models have very similar performance except for QDA, which only has a balanced accuracy of 71.47%.



	Logistic Regression	LDA	QDA	Decision Tree	Random Forest	Gradient Boosting
Precision	0.4011	0.3941	0.4128	0.3957	0.4309	0.4186
Recall	0.6096	0.6111	0.5241	0.6219	0.5988	0.6082
Balanced Accuracy	0.7471	0.7459	0.7147	0.7507	0.7492	0.7505

Table 1: Model Performance

Our task is focused on identifying customers who want to purchase term deposits. Thus, the recall score is of greater interest than other metrics like precision. In other words, we would like to uncover the latent pool of truly interested customers as much as possible, because the labor cost and operation cost of performing calls is much less than the profit of selling term deposits given the nature of contacting customers. Therefore, we choose the decision tree, which gives us the highest recall score as well as highest balanced accuracy, as our final predictive model.

IV. Model Result Interpretation

4.1 CLUSTERING MODEL

In this subsection, we explore our two clustering models. Visualizing them is impossible since they have a high dimension. Therefore, they are interpreted by calculating the cluster centers.

Financial Variable Clustering Model

From the two cluster centers in Appendix 1-20, it is easy to notice that the variables in one cluster all have higher values than the ones in the other. Specifically, cluster 1 contains observations having a higher level of interest rate, employment rate, CPI, CCI, and the number of employees, while cluster 2 has a lower level for these variables. These two clusters are due to monetary policies conducted by the central bank. That is, the central bank increases the interest rate when the economy is in expansion and all statistics above are growing; it tries to boost the economy by decreasing the interest rate when the economy is declining.

Customer Personal Information Clustering Model

From the result of the three clusters (Appendix 1-21), we can see that customers are segmented into three groups: young people, middle-aged people, and elderly people. The first cluster is mainly composed of middle-aged people and the people in the first cluster also appear to have the second-highest divorce rate, marriage rate, they mainly work as self-employed, and very few of them got retired from work. The second cluster contains mainly elderly people; people in this cluster appear to have the highest retired rate, divorce rate, and marriage rate. The third cluster contains mainly younger people, and it mostly consists of students and single people who have completed a university degree. People in the third cluster also appear to have a lower divorce rate, lower marriage rate. Among these three groups, there is a positive trend in terms of education level, the younger generation has more years of education and more university degrees than the middle and elder generations.

4.2 CLASSIFICATION (DECISION TREE) MODEL

In this subsection, we will talk in detail about our final decision tree model. As discussed, the decision tree has the best recall score and balanced accuracy score. It is capable of handling non-linear relationships and interactions hidden in the data. Features are automatically handled by the decision tree with unimportant features being unused. Meanwhile, it is easy to interpret for both people with technical and non-technical backgrounds.

Figure 1 below plots the Decision Tree diagram. The features contributing to the model are Interest Rate, Month, Job, Day of Week, CPI, Contact Communication Type, and Number of Days that passed by after the client was last contacted from a previous campaign. Among these features, the Interest rate is the most frequently used and important feature. The tree starts with Interest Rate for making the split. Later, it is used in both intermediate nodes and leaf nodes for making the final decision.

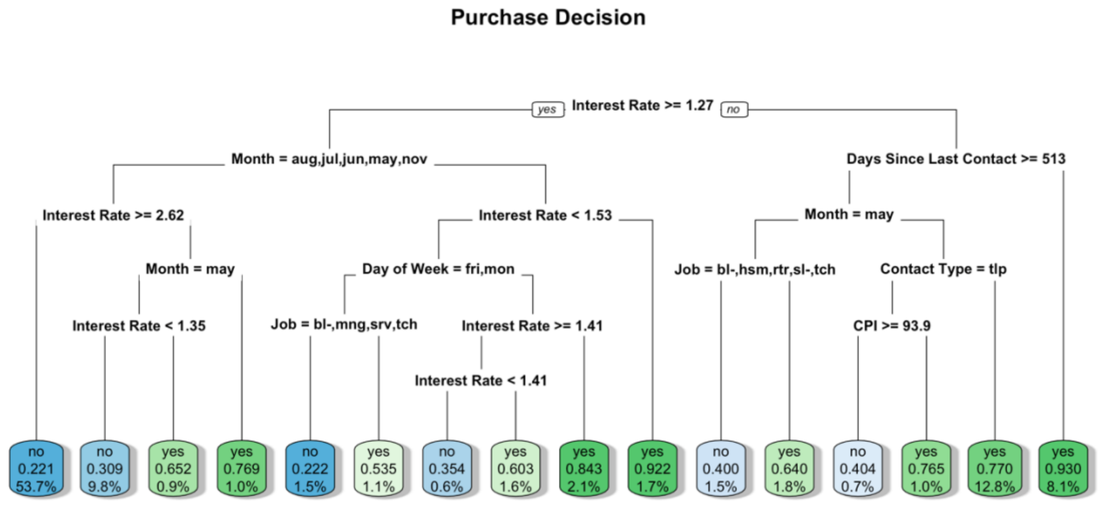


Figure 1: Decision Tree Model

Based on its result, we could observe that when the interest rate is greater than 2.6% and customers are contacted in month August, July, June, May, or November, these people will be predicted as not being able to subscribe to the term deposit since the probability of subscribing the term deposit is only 0.22. If the customer is contacted in May during which the interest rate is between 1.27% and 1.35%, they are likely to not purchase term deposits. On the other extreme, we could also observe that when the interest rate is less than 1.27% and the number of days since the last contact is less than 513 days, the clients will be predicted to be willing to subscribe to a term deposit since the probability of them to subscribe the term deposit is 0.93. If the number of days since the last contact is greater than or equal to 513 days, but customers are not contacted in May or contacted via telephone (i.e., via cellular), these customers will be more likely to purchase our product.

Generally speaking, the probability of purchasing term deposits by customers would be lower when they are contacted via telephone and in May, June, July, August, and November. Among these months, May is the worst month during which people tend to refuse term deposits. Interest rate plays a big role for customers to make purchase decisions. Most people prefer to purchase when the interest rate is low. In terms of occupation, blue collars, managers, people working in services, technicians, and housemaids tend to dislike term deposits. Some conservative people also consider inflation (i.e., CPI) when making their decisions. To them, higher inflation means money deposited in banks will depreciate faster so that they are unwilling to purchase bank products.

After all, only about 10% of customers will purchase during marketing campaigns. The fact that the majority of the customers will not purchase the term deposit reinforces the importance of our project, as the bank could save a huge amount of cost by only reaching out to those customers who are classified as willing to subscribe to a term deposit using the decision model we build.

V. Managerial Implications & Business Applications

Our work took a deep dive into the underlying economic trends, the bank's customer segments, and the indicators for customers' purchasing decisions of term deposits during marketing campaigns. We achieved this by building clustering models and a classification model. In summary, there are three findings that require further managerial attention and consideration.

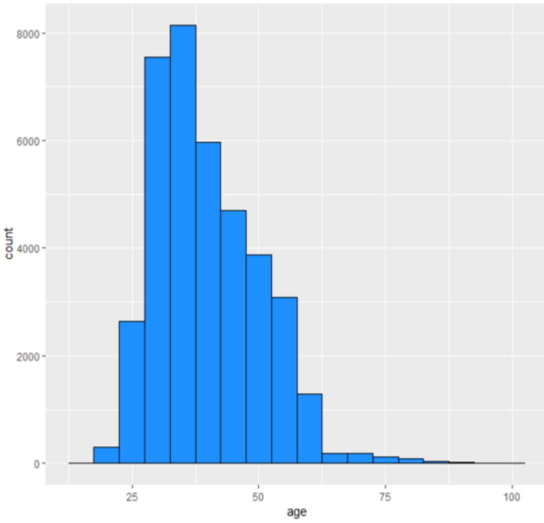
First, Interest Rate, CPI, Employment Rate, and Number of people being employed are highly related. This indicates when the economy is in expansion, people spend more money, consume more goods, make more investments and companies start to hire more people. The central bank would like to cool the economy by raising the interest rate. Conversely, when the economy is in recession, the unemployment rate increases, CPI decreases, and people tend to consume less, the central bank will decrease the interest rate to heat the economy. Banks could take advantage of the economic cycle and be proactive by looking at these statistics.

Second, using our model, the bank can segment its customers into three segments. Segmentation allows the bank to make better use of its marketing budget, gain a competitive edge over rival companies and, importantly, demonstrate better knowledge of their customers' needs and wants. For example, our model segments the customers into three groups, young people, middle-aged people, and elderly people. The bank could use this information to research the education level, property, job, and family situation of their customers based on their age. For the elderly people who work in more profitable industries, the bank could suggest them upgrade their purchases and increase their investment or send them advertisements of high-end financial products. For the younger people who have better financial and family situations, the bank can also send them advertisements of mortgage products.

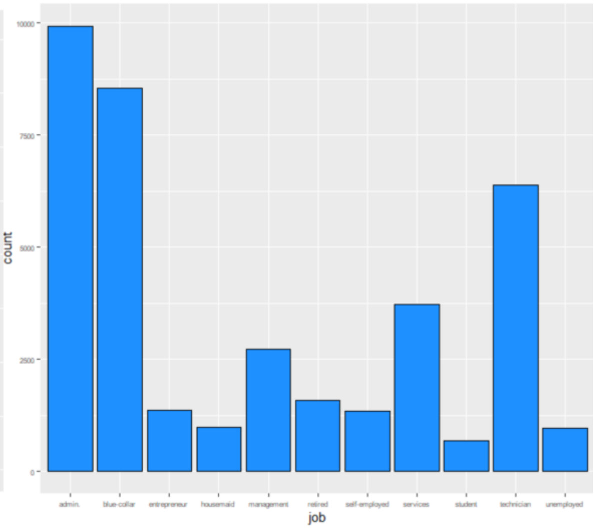
Third, if we want our marketing campaign to be successful, the classification model could prove useful. We need to uncover the pool of potential customers as much as possible and they must be contacted in the right manner. If the executives of the bank would like to launch a campaign for marketing its term deposits. They should consider launching it when the economy is not in expansion, e.g., the interest rate is lower than 1.27%. Also, people tend to purchase when they are contacted twice or more during the campaign as they get to know much more information about the product. And if possible, the bank should contact customers via cellular phone instead of telephone. Furthermore, the bank should avoid launching campaigns during May, June, July, August, and November, because the model suggests contacting customers in these months is less likely to be successful.

In conclusion, our clustering models classify economic conditions and capture the customer segments well. By using the result generated from our clustering model, the bank can customize advertisements to targeted customer segments and save advertisement costs. The decision tree model has a decent ability to detect potential customers. By letting the model go into production, the bank will save a great portion of the cost by targeting customers who tend to have a higher conversion rate while still maintaining a high level of revenue. Future work includes testing this predictive model with other bank products or building customized predictive models for each product in marketing campaigns.

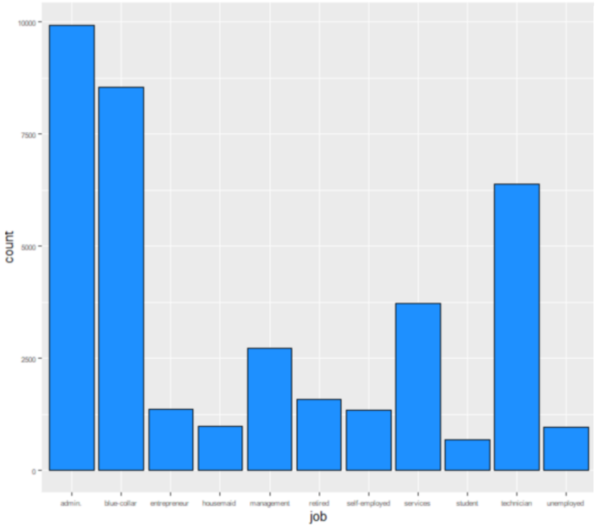
VI. APPENDIX



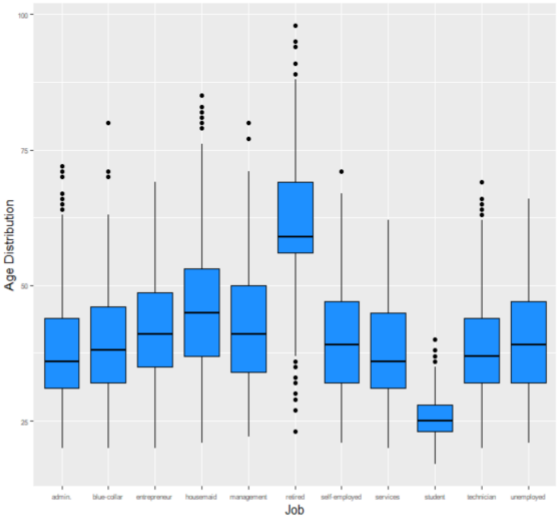
Appendix 1-1: Histogram of Age



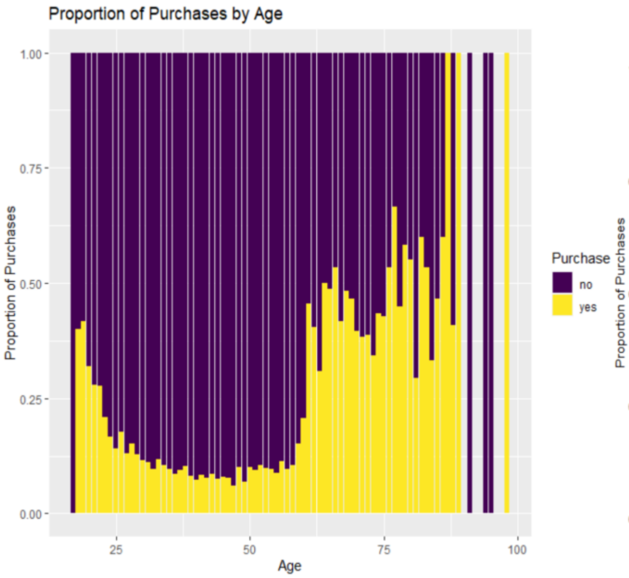
Appendix 1-2: Histogram of Job



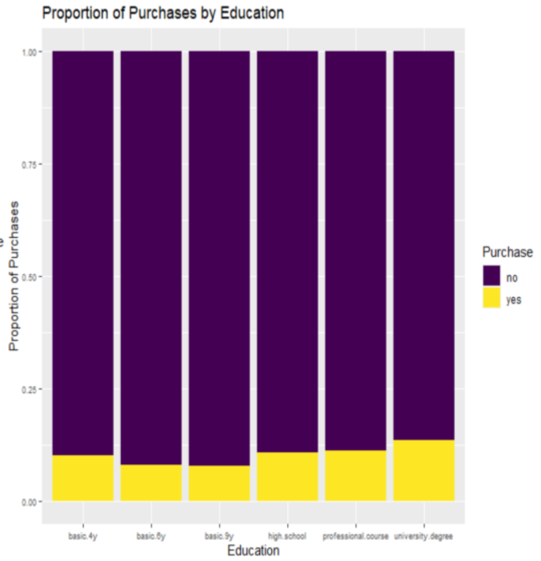
Appendix 1-3: Histogram of Education



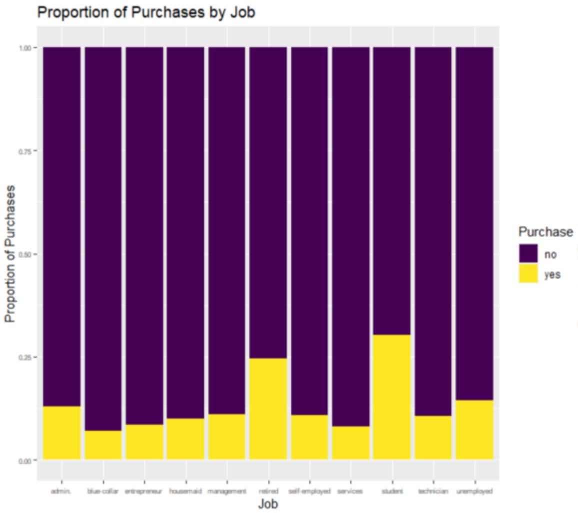
Appendix 1-4: Boxplot of The Age Distribution for Different Jobs



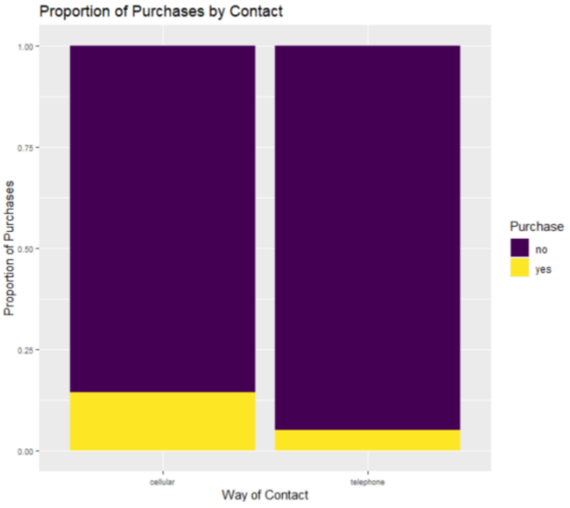
Appendix 1-5: Proportion of Purchases by Age



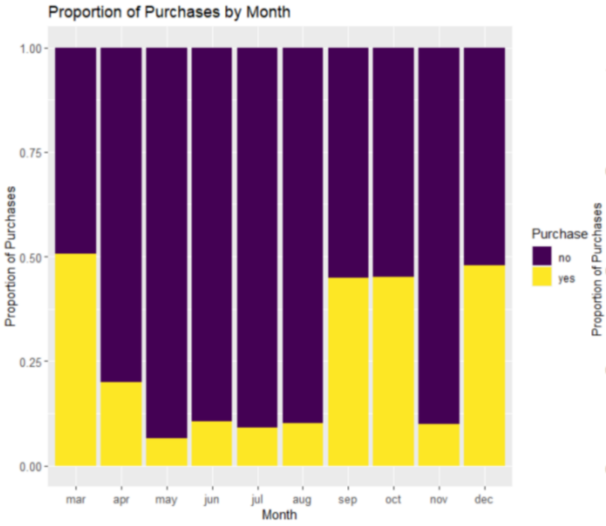
Appendix 1-6: Proportion of Purchases by Education

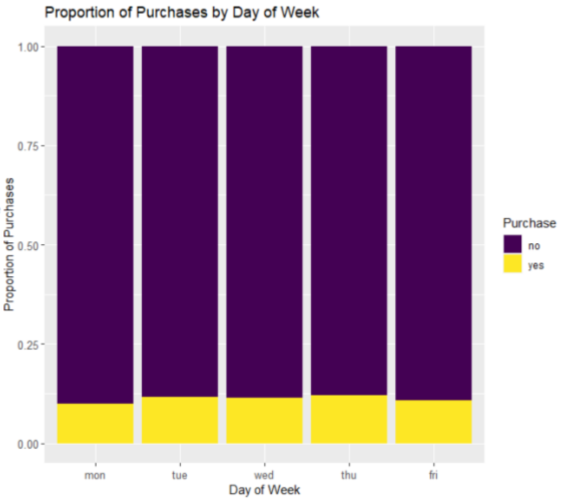


Appendix 1-7: Proportion of Purchases by Job



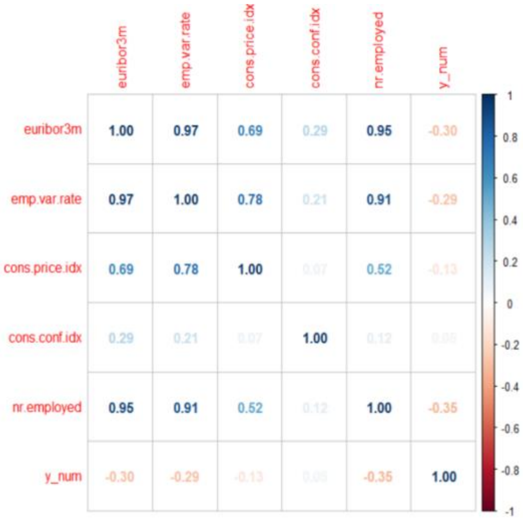
Appendix 1-8: Proportion of Purchases by Contact

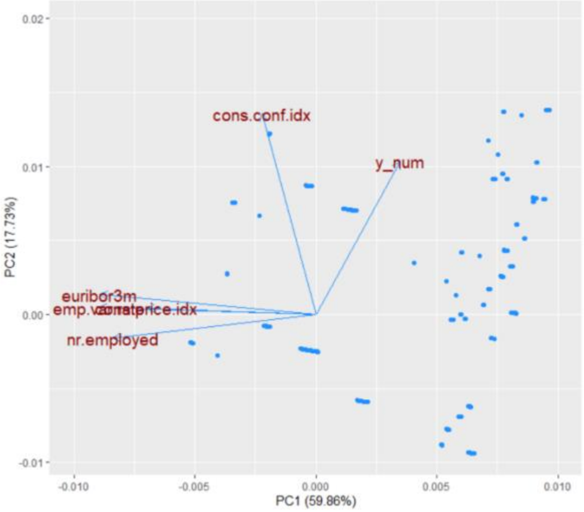




Appendix 1-9: Proportion of Purchases by Month

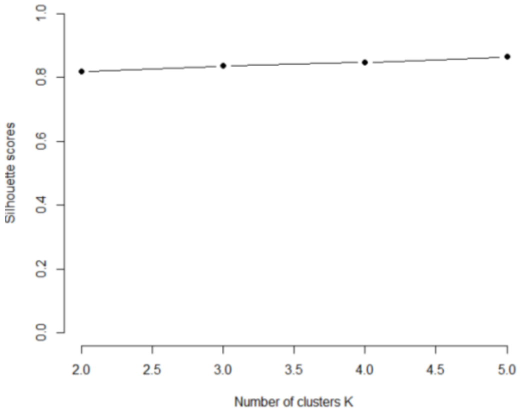
Appendix 1-10: Proportion of Purchases by Day of Week



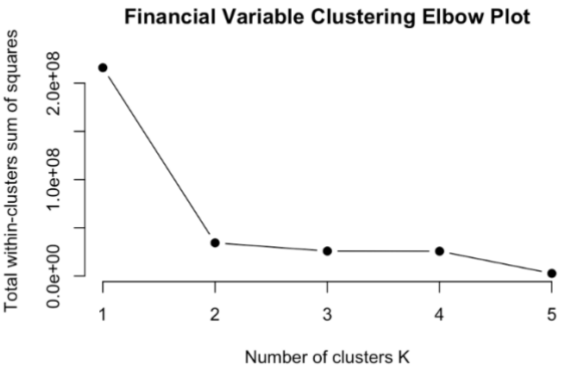


Appendix 1-11: Correlation Matrix (Heatmap) of Financial Variables

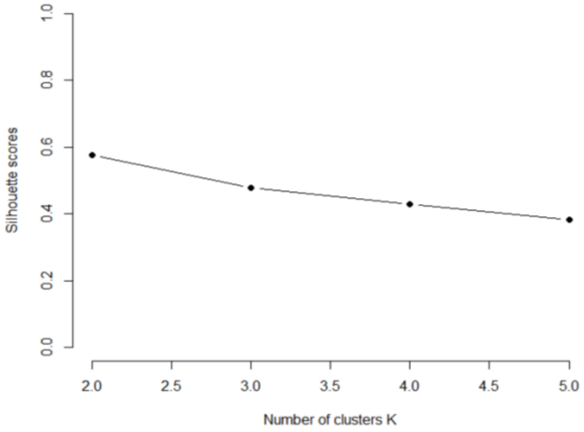
Appendix 1-12: PCA on Financial Variables



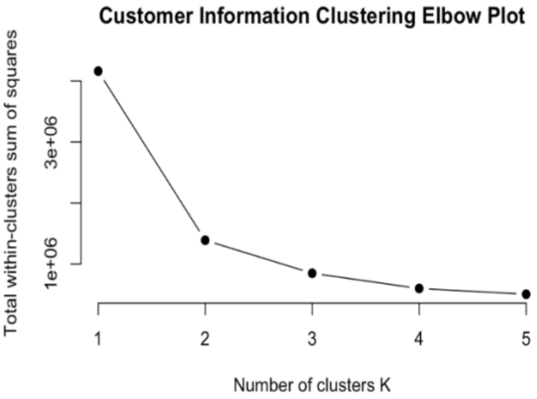
Appendix 1-13: Silhouette Scores for Financial Variable Clustering Model



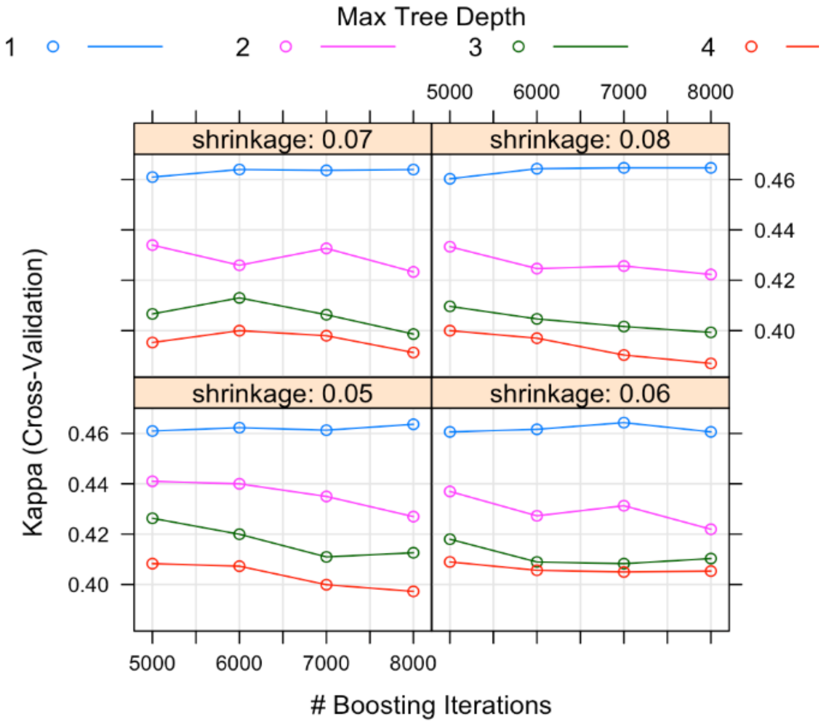
Appendix 1-14: Elbow Plot of Financial Variable Clustering Model



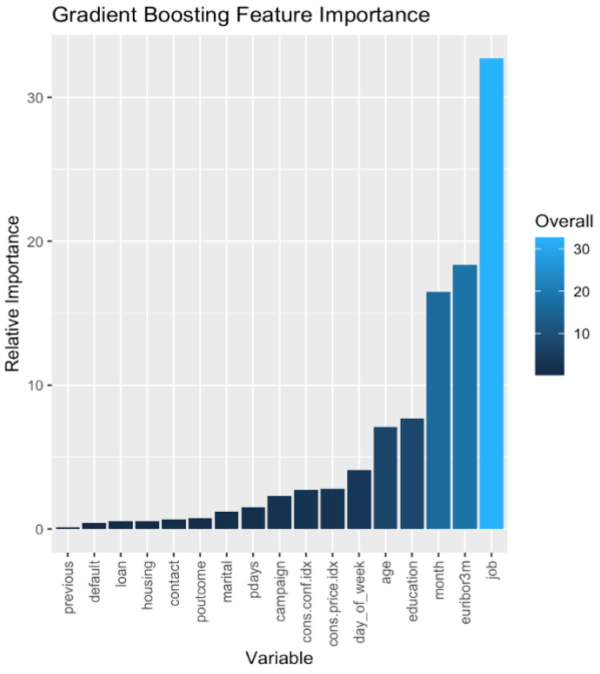
Appendix 1-15: Silhouette Scores for Customer Information Clustering Model

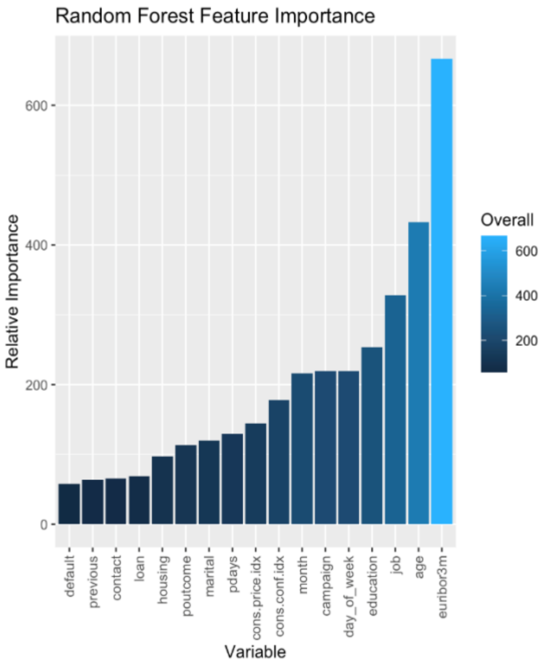


Appendix 1-16: Elbow Plot of Customer Information Clustering Model



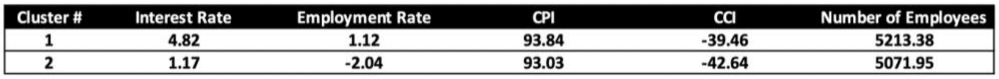
Appendix 1-17: Hyper-parameter Tuning for Gradient Boosting Model





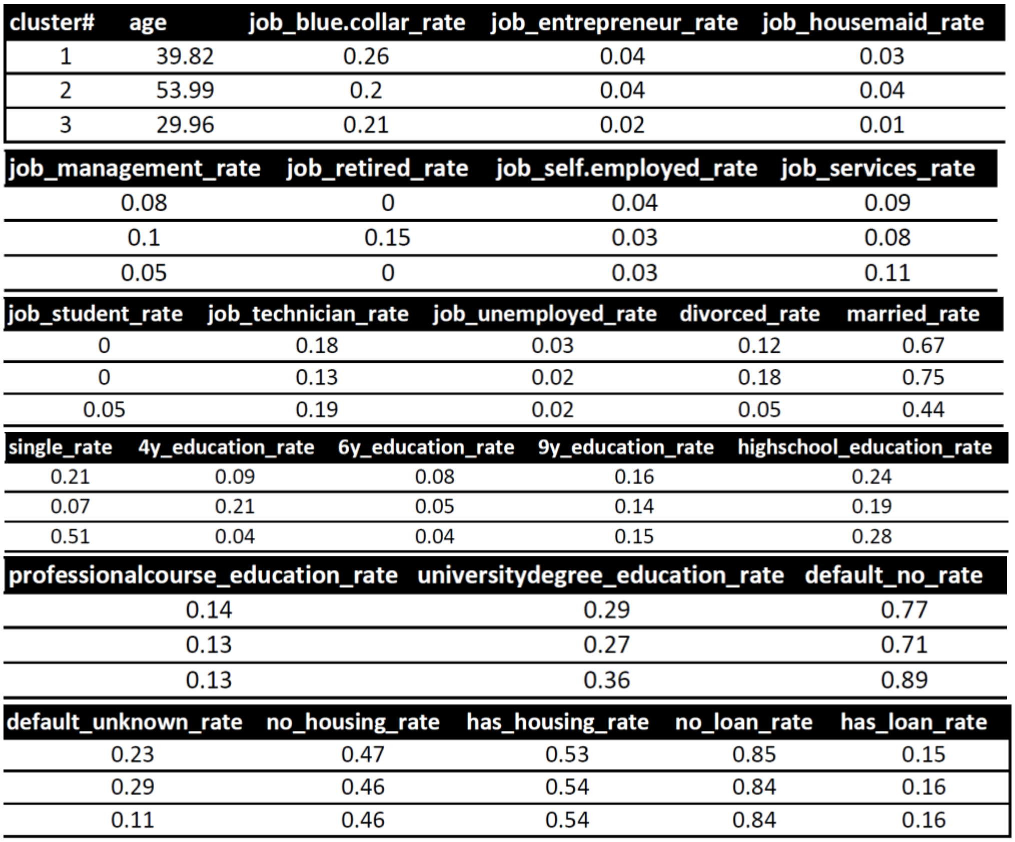
Appendix 1-18: Gradient Boosting Feature Importance

Appendix 1-19: Random Forest Feature Importance



Cluster #	Interest Rate	Employment Rate	CPI	CCI	Number of Employees
1	4.82	1.12	93.84	-39.46	5213.38
2	1.17	-2.04	93.03	-42.64	5071.95

Appendix 1-20: Financial Variable Clustering Model Centers



cluster#	age	job_blue.collar_rate	job_entrepreneur_rate	job_housemaid_rate
1	39.82	0.26	0.04	0.03
2	53.99	0.2	0.04	0.04
3	29.96	0.21	0.02	0.01

job_management_rate	job_retired_rate	job_self.employed_rate	job_services_rate
0.08	0	0.04	0.09
0.1	0.15	0.03	0.08
0.05	0	0.03	0.11

job_student_rate	job_technician_rate	job_unemployed_rate	divorced_rate	married_rate
0	0.18	0.03	0.12	0.67
0	0.13	0.02	0.18	0.75
0.05	0.19	0.02	0.05	0.44

single_rate	4y_education_rate	6y_education_rate	9y_education_rate	highschool_education_rate
0.21	0.09	0.08	0.16	0.24
0.07	0.21	0.05	0.14	0.19
0.51	0.04	0.04	0.15	0.28

professionalcourse_education_rate	universitydegree_education_rate	default_no_rate
0.14	0.29	0.77
0.13	0.27	0.71
0.13	0.36	0.89

default_unknown_rate	no_housing_rate	has_housing_rate	no_loan_rate	has_loan_rate
0.23	0.47	0.53	0.85	0.15
0.29	0.46	0.54	0.84	0.16
0.11	0.46	0.54	0.84	0.16

Appendix 1-21: Customer Information Clustering Model Centers