ECO 386 FINAL PROJECT

DATA ANALYSIS ON CUSTOMERS SEGMENTATIONS

University of San Diego

Dr. Levkoff

November 15, 2020

1. Executive summary

This project is designed to analyze factors that may affect income. Especially, by

looking at the customer segment data, businesses can predict the behaviors for different groups of customers, and can target these specific groups of customers to effectively allocate marketing resources. The data set is searched and downloaded from Kaggle website as a unified data sheet. Based on pervious knowledge, 10 variables and 850 observations are selected and included in the data set. The 10 variables are customer id, age, education, years employed, income, card debt, other debt, defaulted, address and debt income ratio. (Figure 1) Among all variables, 3 variables are categorical variables, 6 are numerical variables, and 1 of them is the ranking. For the defaulted variable, if a customer is defaulted, it will be filled in with a number 1; if not, 0 will be filled in instead. There are empty values in this variable, which will be filled manually with NA later.

Graphical user interface, text, application, email

Description automatically generated

Figure 1 R Script for Basic Information

1. Data Cleaning and Graphs

The data cleaning process is relatively easy. Since the data set is organized, the only

processes are changing the names of the variables and fill-in NA’s for empty data. Line 27-28 shows the process for renaming variables. And lines 33-35 shows steps to fill in NA for empty data. After the data cleaning, there is no missing values for the data set, as the dimension of the data is still 850 observations.

Graphical user interface, text, application

Description automatically generated

Figure 2 R Script for Data Cleaning

In order to plot the data, few libraries and packages are installed with the coding. Package “ggplot2”, library plyr, library MASS and library mgcv are installed. (Figure 3) In order to find the distribution of the data, histograms and probability density curves are used. The relating coding is in Figure 4, and the graphs for all variables are in Figure 5-8. For all variables that are continuous, they are plotted with a probability density curve. For variables that are categorical, they are plotted using a histograms.

Graphical user interface, text, application

Description automatically generatedText

Description automatically generated

Figure 3 and 4 R Script for Installed Libraries and Distribution Graphs

Graphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generated

Figure 5 and 6 Probability Density Curves for Age, Years, Income and Debts

A picture containing chart, line chart

Description automatically generatedGraphical user interface, application

Description automatically generated

Figure 7 and 8 Probability Density Curves for Debts and Histograms for Education and Defaults

In order to find the relationships between income and other factors, related visualized graphs are used to determine the relationships. Lines 98-101 shows the steps to clean out some outliers, and lines 104-123 are used to determine the visualized relationships. According to Figure 10, age and years employed have a positive effect on income, which means the older a person or the more years employed of a person, the higher income they will get. But this positive relationship will start to decline when reaching the maximum point; this shows that if a person is beyond the age or beyond the years employed, he/she will have a downward slope for income. According to Figure 11, among all levels of education, card debt has a positive effect for people who have incomes lower than 100,000, and has a negative effect for people who have incomes higher than 100,000. According to Figure 12, people with defaulted history tend to have a positive relationship between income and debt income ratio; while people with no defaulted history tend to have a slightly negative relationship between income and debt income ratio. The violin graph in Figure 13 shows that people among all incomes tend to have a debt below 10 in other forms than card debt.

Graphical user interface, text, application

Description automatically generated

Figure 9 R Script for Scatter Plots

Graphical user interface, chart, scatter chart

Description automatically generated

Figure 10 Scatter Plots between Income and age or years employed

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 11 and 12 Scatter Plots between Income and Card Debt or Debt Income Ratio

Chart, line chart

Description automatically generated

Figure 13 Violin Graph between Income and Other Debt

1. Regression Models

In order to build regression model, library tseries is installed in line 131. Lines 134-

135 shows the nonlinear and logistic transformation of income in order to build nonlinear regression models. Lines 138-157 are used to partition the data. The data set is divided into training set and testing set. The training set is divided as 70% of the data set, and the testing data is 30%. The partition is shown in the dimension of training data.

Graphical user interface, text, application, email

Description automatically generatedA picture containing text

Description automatically generated

Figure 14 and 15 R Script for Data Partition and Dimensions for Training and Testing Data

There are 6 models built in this task. First four are linear models, and the last two are nonlinear models. For each model, the dependent variable is income because the models are used to predict the relationships between inflation and other factors. The first model is a linear model between income and years employed. The training data is used to build the model, and some indicators are showing whether this is a good model for the data. According to the summary in Figure 17, years employed has a positive effect on income. As 1 additional year employed, the income goes up by 3.47. The p-value for both the variable and the model is small enough, which means the model is significant. The R square states that about 44.01% of the data are explained by the model. Based on the histogram in Figure 18 and the Jarque Bera Test in Figure 17, the p-value shows that null hypothesis is rejected, which means that the residuals cannot be concluded as normally distributed. The confidence interval for years employed is (3.16, 3.78), which means that we are 95% confident that the mean years employed are between 3.16 and 3.78. And the in sample error and out sample error will be compared later in part 3 and 4 of this project.

Text, letter

Description automatically generated

Figure 16 R Script for Model 1

Text

Description automatically generatedChart, histogram

Description automatically generated

Figure 17 and 18 Summary and Normality Test for Model 1 and Histogram for Model 1

Model 2 is a linear model with income as the dependent variable and years employed and age as the independent variables. According to summary in Figure 20, both years employed and age have positive relationships to income. As age goes up by 1 unit, income goes up by 0.88 unit. The p-values for both variables and the model is small enough, which means the model is significant. The R square is 46.62%, which tells us about 44.01% of the data are explained by the model. Based on the histogram in Figure 21 and the Jarque Bera Test in Figure 20, the p-value shows the null hypothesis is rejected, which means that the residuals cannot be concluded as normally distribution. The in sample error and out sample error will be compared later in part 3 and 4 of this project.

Text, letter

Description automatically generated

Figure 19 R Script for Model 2

Text

Description automatically generatedChart, histogram

Description automatically generated

Figure 20 and 21 Summary and Normality Test for Model 2 and Histogram for Model 2

Model 3 is a linear model with income as the dependent variable and years employed, age, and card debt as the independent variables. According to summary in Figure 23, all 3 variables have positive relationships to income. As card debt goes up by 1 unit, income goes up by 5.25 unit. The p-values for both variables and the model are small enough, which means the model is significant. The R square is 55.71%, which tells us about 55.71% of the data are explained by the model. Based on the histogram in Figure 22 and the Jarque Bera Test in Figure 23, the p-value shows the null hypothesis is rejected, which means that the residuals cannot be concluded as normally distribution. The in sample error and out sample error will be compared later in part 3 and 4 of this project.

Text, letter

Description automatically generated

Figure 22 R Script for Model 3

Text

Description automatically generatedChart, histogram

Description automatically generated

Figure 23 and 24 Summary and Normality Test for Model 3 and Histogram for Model 3

Model 4 is a linear model with income as the dependent variable and age and education as the independent variables. According to summary in Figure 26, both variables have positive relationships to income. As education goes up by 1 unit level, income goes up by 5.96 unit. The p-values for both variables and the model are small enough, which means the model is significant. The R square is 28.34%, which tells us about 28.34% of the data are explained by the model. Based on the histogram in Figure 25 and the Jarque Bera Test in Figure 26, the p-value shows the null hypothesis is rejected, which means that the residuals cannot be concluded as normally distribution. The in sample error and out sample error will be compared later in part 3 and 4 of this project.

Text

Description automatically generated

Figure 25 R Script for Model 4

Text

Description automatically generatedChart, histogram

Description automatically generated

Figure 26 and 27 Summary and Normality Test for Model 4 and Histogram for Model 4

Model 5 is a nonlinear model with income as the dependent variable and years employed, and years employed^2 as the independent variables. According to summary in Figure 29, both variables have positive relationships to income. As years employed^2 goes up by 1 unit, income goes up by 0.14 unit. The p-values for both years employed^2 and the model are small enough, which means the model is significant. But the p-value for years employed is too large to be significant. The R square is 49.78%, which tells us about 49.78% of the data are explained by the model. Based on the histogram in Figure 28 and the Jarque Bera Test in Figure 29, the p-value shows the null hypothesis is rejected, which means that the residuals cannot be concluded as normally distribution. The in sample error and out sample error will be compared later in part 3 and 4 of this project.

Graphical user interface, text

Description automatically generated

Figure 28 R Script for Model 5

Text

Description automatically generatedChart, histogram

Description automatically generated

Figure 29 and 30 Summary and Normality Test for Model 5 and Histogram for Model 5

Model 6 is a nonlinear model with income as the dependent variable and years employed, years employed^2 and years employed^3 as the independent variables. According to summary in Figure 32, years employed and years employed^3 have positive relationships to income; while years employed^2 has a negative relationship to income. As years employed^3 goes up by 1 unit, income goes up by 0.007 unit. The p-values for both years employed, years employed^3 and the model are small enough, which means the model is significant. But the p-value for years employed^2 is too big to be significant. The R square is 50.88%, which tells us about 50.88% of the data are explained by the model. Based on the histogram in Figure 31 and the Jarque Bera Test in Figure 32, the p-value shows the null hypothesis is rejected, which means that the residuals cannot be concluded as normally distribution. The in sample error and out sample error will be compared later in part 3 and 4 of this project.

Text

Description automatically generated

Figure 31 R Script for Model 6

Text

Description automatically generatedChart, histogram

Description automatically generated

Figure 32 and 33 Summary and Normality Test for Model 6 and Histogram for Model 6

1. Validate Regression Models

All of the above models shows some relationships between income and other factors. Overall, the income has a positive relationship with all variables found, which is also correlates with what were found in the scatter plots. According to Figure 35, the in-sample errors are shown. Among those in-sample errors, model 3, with a value of 23.61, has the lowest in-sample error, which shows that model 3 is the model that fits the data best. With the test data partition, the RMSE are calculated for each model. The out-of-sample errors are calculated for all 6 models in Figure 37. According to the out-of-sample errors, model 3 has the lowest number for out-of-sample error, which shows that model 3 is the most against testing model. As mentioned above, model 3 has both the lowest in-sample error and the lowest out-of-sample error. Model 3 is the model picked as the best model under our experiments.

Text

Description automatically generatedText

Description automatically generated

Figure 34 and 35 R Script and Results for In-Sample Error

Text

Description automatically generatedText

Description automatically generated

Figure 36 and 37 R Script and Results for Out-Sample Error

1. Classification Models

In order to build classification models, library lattice, caret, rpart, e1071 and

randomRorest are installed into the code as lines 302-307 shows. In order to build classification models, categorical variables, such as defaulted and education, are converted to factors in lines 309-311. Lines 314-329 repeated the data partition steps.

Graphical user interface, text, application

Description automatically generated

Figure 38 R Script for Classification Models

There are 4 models built in this task. The first model is the CART model 1. Lines 332-338 shows the building steps. Categorical variable, which is the defaulted variable, is used in this model. After building the confusion matrix, the accuracy for this model is 73.61%, which shows that 73.61% of the time is the classifier correct. The confidence interval is between (0.672, 0.794), which means that 95% of the time the mean defaulted value is between 0.672 and 0.794. The p-value is 0.4137, which is too big to be significant.

Text

Description automatically generated

Figure 39 R Scripts for CART Model

Table

Description automatically generatedChart, line chart

Description automatically generated

Figure 40 and 41 Reference for CART Model 1 and Plotting for Complexity Parameter

The second model is the CART model 2. Lines 341-346 shows the building steps. Categorical variable, which is the education variable, is used in this model. After building the confusion matrix, the accuracy for this model is 51.76%, which shows that 51.76% of the time is the classifier correct. The confidence interval is between (0.4545, 0.5804), which means that 95% of the time the mean defaulted value is between 0.4545 and 0.5804. The p-value is 0.5253, which is too big to be significant.

Graphical user interface, text

Description automatically generated

Figure 42 R Scripts for CART Model 2

A close up of text on a white background

Description automatically generatedChart, line chart

Description automatically generated

Figure 43 and 44 Reference for CART Model 2 and Plotting for Complexity Parameter

The third model is the random forest model 1. Lines 348-353 shows the building steps. Categorical variable, which is the defaulted variable, is used in this model. After building the confusion matrix, the accuracy for this model is 77.78%, which shows that 77.78% of the time is the classifier correct. The confidence interval is between (0.7164, 0.8314), which means that 95% of the time the mean defaulted value is between 0.7164 and 0.8314. The p-value is 0.0522, which is slightly bigger than needed to be significant.

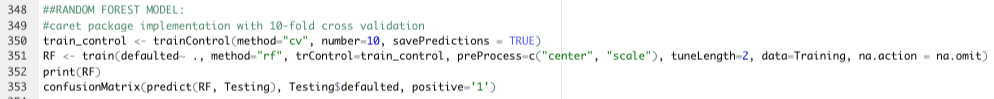


Figure 45 R Scripts for Random Forest Model 1

Table

Description automatically generated

Figure 46 Reference for Random Forest Model 1

The fourth model is the random forest model 2. Lines 356-358 shows the building steps. Categorical variable, which is the education variable, is used in this model. After building the confusion matrix, the accuracy for this model is 51.39%, which shows that 51.39% of the time is the classifier correct. The confidence interval is between (0.4451, 0.5823), which means that 95% of the time the mean defaulted value is between 0.4451 and 0.5823. The p-value is 0.6836, which is too big to be significant.

A picture containing text

Description automatically generated

Figure 47 R Scripts for Random Forest Model 2

A close up of text on a white background

Description automatically generated

Figure 48 Reference for Random Forest Model 2

1. Validate Classification Models

Based on the above accuracy, CART model 1 has an accuracy of 73.61%; CART model 2

has an accuracy of 51.76%; random forest model 1 has an accuracy of 77.78%; and random forest model 2 has an accuracy of 51.39%. Because random forest model 1 (the third model) has the highest accuracy, it is picked as the best model. The data is divided with the same method as the regression model: by dividing data set into 70% as the training data and 30% as the testing data. There are values missing in the defaulted variable. So when running the confusion matrix, those NA values are passed by R with our code demanding.

1. Conclusion

In this project, we determined the relationships between income and other factors to our

customer base. After the visualization and the models built, it is shown that income, age, education level, years employed have a positive relationship on income in the short term, but may have a negative effect in the long term, after a certain point passes. Based on our regression model evaluation, model 3 is both the best fitting model and the most against testing model, with an in-sample error of 23.61 and an out-of-sample error of 32.91. Based on our classification model evaluation, the third model, which is the random forest model 1, has a highest accuracy of 77.78%; it is chosen as the best model among the four. Although the best models are decided based on our models, there is need for further investigations on other models that may fit the data better. For example, the regression model that is chosen may not be the best fit model since the R square is only 55.71%. There may be better fitting model that is not covered in this report. Therefore, there is need to involve more data from different customer segments and time period in order to further build the models and analyze the relationships between income and other factors.