Data Analysis Report:
Predictive Modeling for Bank Deposit Subscriptions

(OCRUG Data Science Hackathon 2019)

Won *Second place* for
Best Model & Best Visualization

Project done by:
Yuan Liu
Chuyan Zhang
Takako Suzuki
Dora Yuan
York Fang

Report written by:
Takako Suzuki

November 10, 2019

ABSTRACT
Herein we presents the predictive modeling and the analysis results associated with the OCRUG 2019 Hackathon competition. The purpose of this report is to document the decision making phases of data preprocessing procedures and corresponding data modeling.

BACKGROUND
The original dataset is collected from a Portuguese bank marketing campaign related to bank deposit subscriptions. Our objective is to build a predictive model that explains the success of contacts, and to identify important attributes related to marketing success, which could be applied to further promote the conversion rate of the targeted marketing campaign.

Our analysis methodology for this project can be summarized to four stages (**Fig. 1**). In the first stage, we investigated into the business space of the targeted marketing campaign in banks as well as the data, which provided insights on approaches to data pre-processing. Then, the raw data was processed based upon both business contexts and statistical principals. With the prepared data, we built predictive models that explains the success of marketing campaign. Finally, business insights were generated based upon the modeling results.

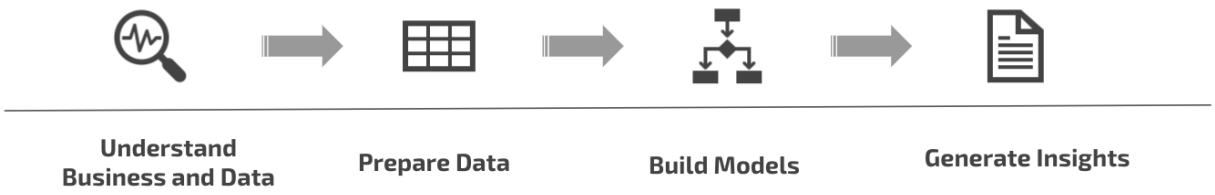


Figure 1. Overview of the analysis methodology

A. UNDERSTAND BUSINESS / DATA

In this dataset, the marketing campaign was conducted on phone calls, and the marketing success was measured by whether or not the client makes a deposit after the campaign. The attributes are categorized into three groups: (i) bank client data, (ii) data related to the last contact of the current campaign, and (iii) other attributes related to previous marketing campaign outcomes.
Generally speaking, every marketing campaign has one goal: conversion from leads to sales. Therefore, our team decided to start with the common inquiry of every marketing campaign - what factors contributes to a successful conversion through the phone campaign.

B. DATA PREPROCESSING

There was no missing data in this dataset, and thus we started by converting all categorical attributes to numeric data and exploring the correlations of all the independent attributes with the dependent variable "y" (whether the client makes a deposit or not) using linear regression. The results indicated that all attributes have very low correlation values. Our hypothesis for this outcome was that all the other

variables have up to 4 levels but in the "job" attribute, it has 12 levels and that factor might lead to desaturation of the correlation.

Part 1. Complexity Reduction. To reduce the number of levels in "job" attribute, we performed a clustering method on the "job" attribute by assigning occupations into groups based on the similarity in personal information (age, marital, education, default, balance, housing, and loan). As a result, we reduced 12 levels to 5 levels. Using K-means Nearest Neighbor we found the initial optimal number of clusters is 4 with SSE of 30,000 (**Fig. 2**).

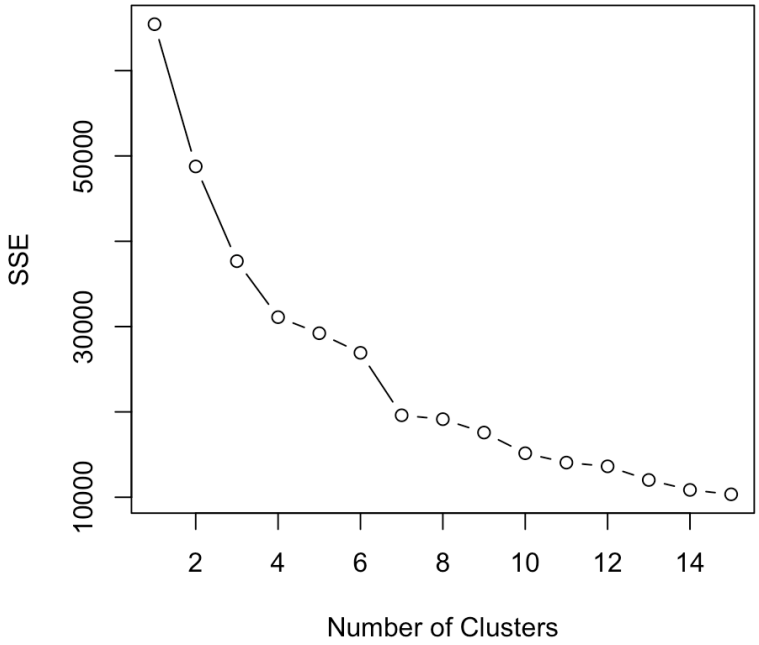


Figure 2. SSE Curve for cluster number optimization

We then categorized jobs into 4 different groups by assigning job to the cluster with highest count in each job; however, due to the imbalanced job distributions in the dataset, we were unable to assign any job with the highest count to cluster 3. Therefore, we decided to use a total of 5 clusters so that each cluster has at least one highest count of the job (**Fig. 3**).

Figure 3. Level reduction for "job" attribute via K-means clustering approach (N = 5)

Part 2. Attribute Conversion. With the in-depth research into literatures, we selected two tentative predictive models that are commonly used among marketing analysis: *Naive Bayes* and *Decision Tree*. Due to the limitation of Naive Bayes' model which works better with categorical variables, we converted numerical variables to categorical by binning using the *Smbinning* package. Optimal Binning analyzed the relationship with a binary target variable and identified the optimal cut points (**Table 1**).

```
> result <- smbinning(df=df, y="y", x="balance")
> result$ivtable
  Cutpoint CntRec CntGood CntBad CntCumRec CntCumGood CntCumBad PctRec GoodRate BadRate
1   <= -47   3193      166   3027      3193        166      3027 0.0706   0.0520  0.9480
2    <= 60   7628      594   7034     10821        760     10061 0.1687   0.0779  0.9221
3   <= 798  17577     1963  15614     28398       2723     25675 0.3888   0.1117  0.8883
4    > 798  16813     2566  14247     45211       5289     39922 0.3719   0.1526  0.8474
5  Missing      0        0      0     45211       5289     39922 0.0000      NaN     NaN
6    Total  45211     5289  39922        NA         NA        NA 1.0000   0.1170  0.8830
    Odds  LnOdds     WoE     IV
1 0.0548 -2.9033 -0.8820 0.0392
2 0.0844 -2.4716 -0.4503 0.0288
3 0.1257 -2.0737 -0.0524 0.0010
4 0.1801 -1.7142  0.3071 0.0394
5    NaN     NaN     NaN    NaN
6 0.1325 -2.0213  0.0000 0.1084
```

Table 1. Example of binning outputs from "balance" attribute

We were aware of the disadvantage of this binning method because converting the numeric data into categorical variables with limited a large number of levels may cause losing in accuracy due to the difference in weight percentage. However, our purpose was to avoid overfitting of our data when performing the predictive models and to decrease processing time.

Part 3. Confirmation of Significance. Our last step in data pre-processing is to confirm that the attributes we dropped were proven to be insignificant. One method was to compare the impact of different levels within each variable on the marketing success. For instance, the "Default" attribute in **Figure 4** was discarded because the rate of success for the two levels was very close. We noted that this graphical analysis may have flaws because we are disregarding the interaction of "Default" with other attributes. Hence, we ran linear regressions to examine each attribute and compare the significance value.

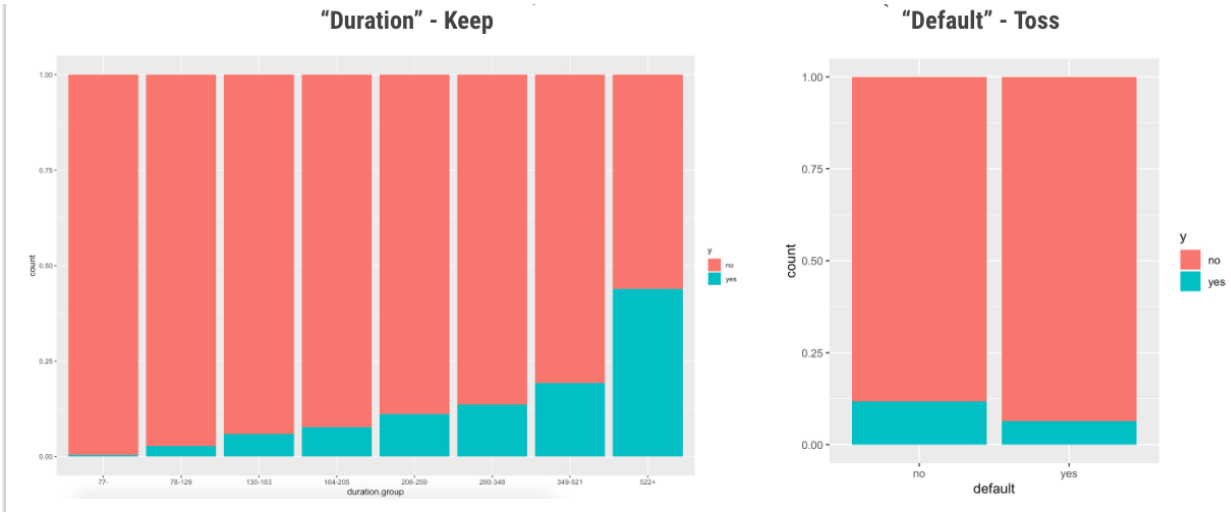


Figure 4. Examples of comparing the significance of attributes on marketing success. "y" = no: failed marketing. "y" = yes: successful marketing

Finally, we found out that "Pday", "Previous", "Default", and "Contacted" have the least impact on the dependent variable outcome and we dropped these four low impact variables.

C. DATA MODELING

The dependent variable was significantly skewed (**Fig. 5.** Left image), so we used a stratified random sampling method to ensure the ratio of "yes" and "no" is identical (**Fig. 5.** Right image).

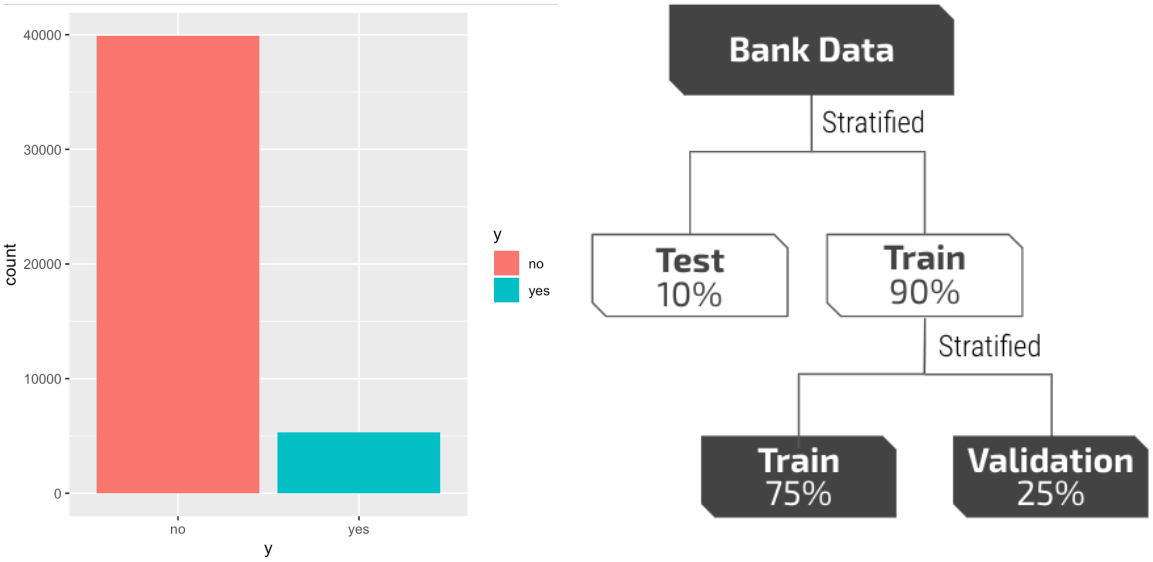


Figure 5. Left: visualization of marketing outcome distribution. Right: data partition for modeling

We explored the level of importance of each attribute using *Naive Bayes* and *Decision Tree* and found out that the most important attribute for both models is "duration". (**Fig. 6a** and **Fig.6b**., respectively)

```
### Level of importance
x.nb <- varImp(nb)
impTab <- x.nb$importance
ggplot(impTab, aes(x= reorder(row.names(impTab), +yes), y=yes)) +
  geom_bar(stat = 'identity', aes(fill = row.names(impTab))) +
  labs(title = "Variable in predicting term deposit", x = "Variables", y = "Importance") +
  scale_fill_brewer(palette = "Set3") + coord_flip() +
  theme_classic() +
  theme(legend.position = "none")
```

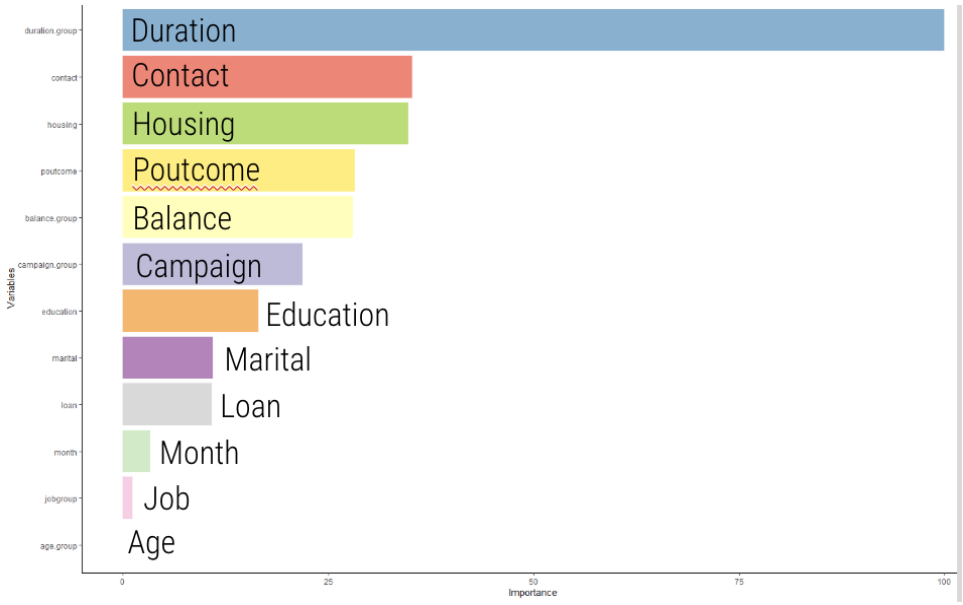


Figure 6a. Variable importance level from Naive Bayes model

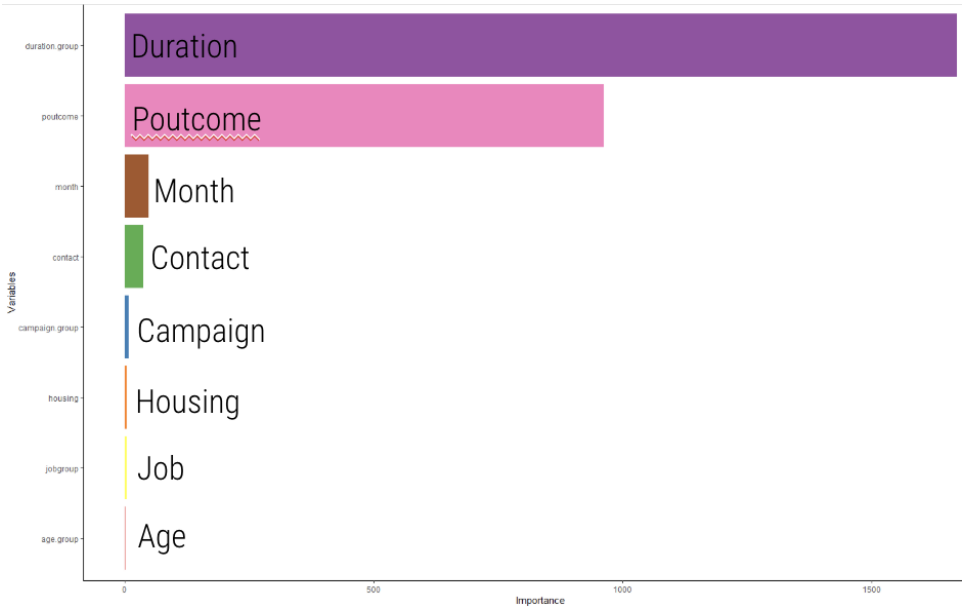


Figure 6b. Variable importance level from Decision Tree model

In the modeling phase, we successfully tested the two models and gathered the accuracy rate in each model. (see Table 2a and 2b, respectively)

```
> confusionMatrix(nb.pred, validate.data$y)
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  8579  698
       yes  403  492

               Accuracy : 0.8918
                 95% CI : (0.8856, 0.8977)
    No Information Rate : 0.883
    P-Value [Acc > NIR] : 0.002921

                  Kappa : 0.413

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.9551
            Specificity : 0.4134
         Pos Pred Value : 0.9248
         Neg Pred Value : 0.5497
             Prevalence : 0.8830
         Detection Rate : 0.8434
   Detection Prevalence : 0.9120
      Balanced Accuracy : 0.6843

       'Positive' Class : no
```

Table 2a. Naive Bayes

```
> confusionMatrix(dtree.pred, validate.data$y)
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  8784  825
       yes  198  365

               Accuracy : 0.8994
                 95% CI : (0.8934, 0.9052)
    No Information Rate : 0.883
    P-Value [Acc > NIR] : 0.00000007894

                  Kappa : 0.369

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.9780
            Specificity : 0.3067
         Pos Pred Value : 0.9141
         Neg Pred Value : 0.6483
             Prevalence : 0.8830
         Detection Rate : 0.8635
   Detection Prevalence : 0.9447
      Balanced Accuracy : 0.6423

       'Positive' Class : no
```

Table 2b. Decision Tree

Using the validation dataset, we compared the two models' accuracy rate and observed that Decision Tree is higher than Naive Bayes by 0.008. Although the two models show a comparably high accuracy, we chose Decision Tree as our final model because it takes care of various issues such as outliers and intercorrelation between variables which were present in our dataset. Table 3 showed an accuracy of 0.8936 for our final Decision Tree model with the testing dataset.

CONCLUSION

In conclusion, we presented here the systematic approach to the analysis of the bank marketing data. Specifically, we highlighted our approach to data pre-processing, modeling and result analysis.

We found that duration is the most relevant feature, suggesting that the longer the call representative spends with a customer, the higher the conversion rate would be. The second feature is Poutcome which indicates that customers who have deposited before having a higher chance to deposit again. This is very

Decision Tree-Test

```
> confusionMatrix(dtree.test, test.data$y)
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  3902  391
       yes   90  137

               Accuracy : 0.8936
                 95% CI : (0.8842, 0.9024)
    No Information Rate : 0.8832
    P-Value [Acc > NIR] : 0.01478

                  Kappa : 0.3148

 Mcnemar's Test P-Value : < 0.0000000000000002

            Sensitivity : 0.9775
            Specificity : 0.2595
         Pos Pred Value : 0.9089
         Neg Pred Value : 0.6035
             Prevalence : 0.8832
         Detection Rate : 0.8633
   Detection Prevalence : 0.9498
      Balanced Accuracy : 0.6185

       'Positive' Class : no
```

Accuracy: 0.8936

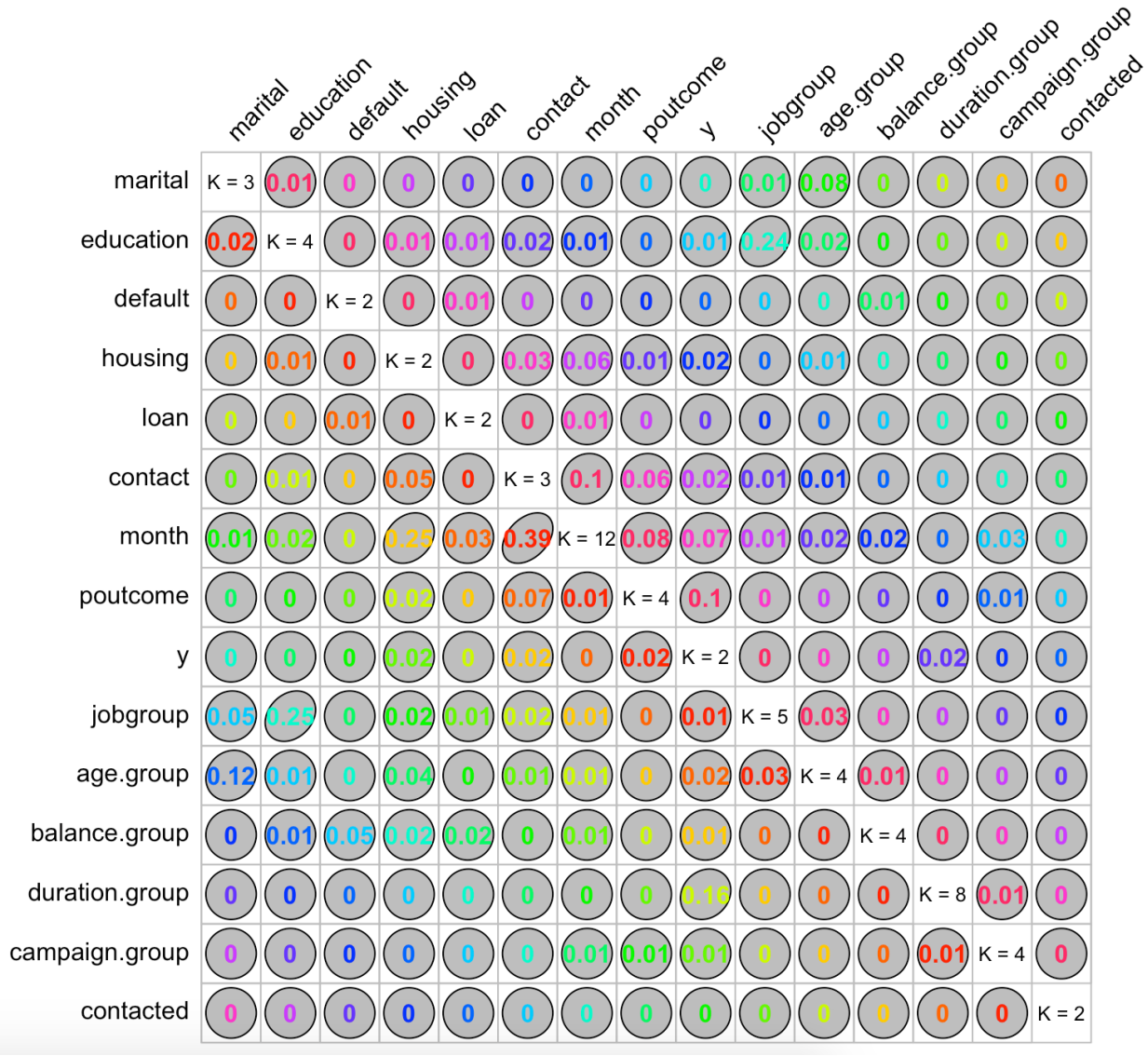
Table 3. *Testing final model*

common in marketing when a customer has already established customer loyalty with the company. A loyal customer remains loyal when offering lower prices and better discounts; therefore, a phone campaign with alluring deals will achieve a higher conversion rate. A new direction we wish to approach in the future is to segment customers into two groups: old customers and new customers (**Fig. 7**). This is because the two groups present different purchasing behaviors, leading to different outcomes.

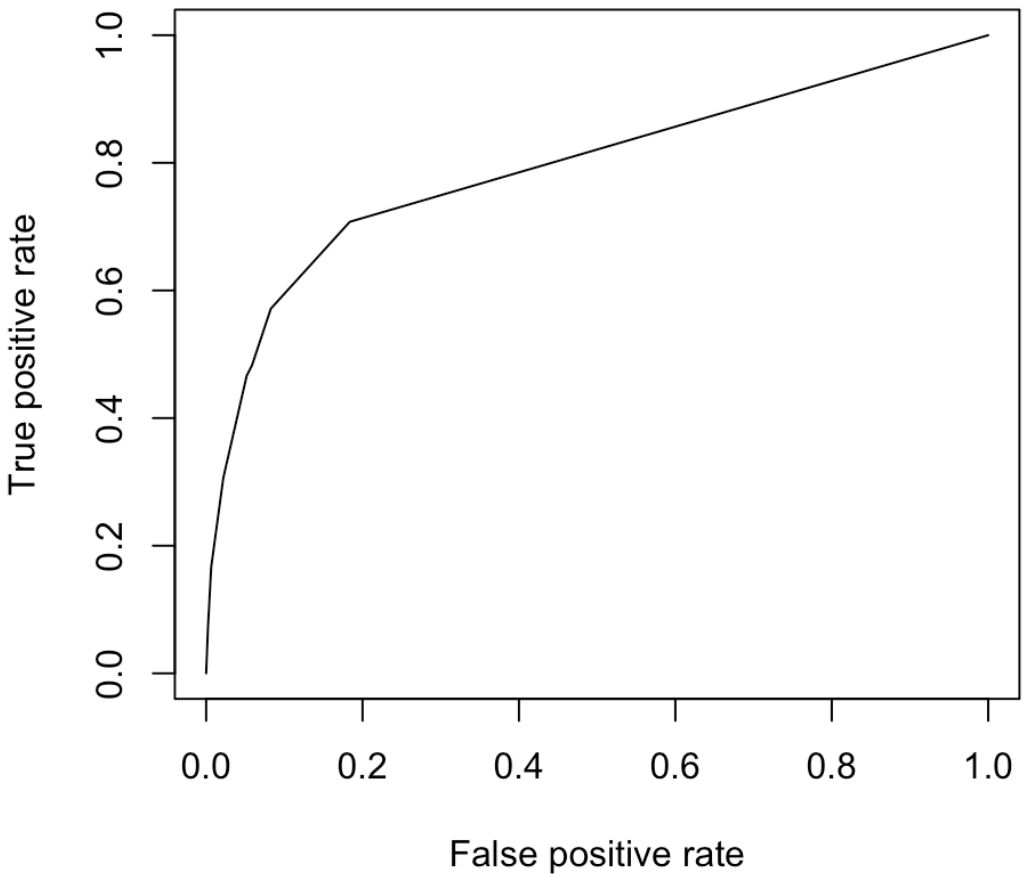
Figure 7. Summary and outloook

Appendix

Appendix A: Correlation Matrix after dropping low impact variables



Appendix B: ROC graph of Decision Tree Model (Test dataset)


True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0