KOGOD SCHOOL of BUSINESS
AMERICAN UNIVERSITY • WASHINGTON, DC

ITEC 621 Predictive Analytics Project

Project Name: Financial Participation in Kenya and Uganda

Class Section: Thursday 5:30-8PM (002)

Team Number: 3

Team Members: **Abhinav Nitesh, Ahmed Malik, Binh Minh An Nguyen, Farhaan S Haque**

Last updated: 04-28-2022

Deliverable Number: 5

Executive Summary: Our team's assignment was to develop models and methods to enhance data-driven decision-making at a particular organization. For this project, we selected the data set based on the ongoing competition at the Zindi Data Science Community platform. To focus our project we identified Specific, Measurable, Achievable, Realistic, and Time Frame (SMART) goals to drive our analysis. Specifically, our SMART goal was to identify what attributes were the main drivers for a person to open a bank account. This information could be used to better focus on specific attributes to better target consumers.

Based on the above SMART goal, our dependent variable is "bank_account". The dataset contains 10 independent variables that we investigated to determine which were the most important predictor of a person opening a bank account. In order to solve our business problem, we created multiple models and compared their accuracy and efficiency. The main models we used include logistic regression, random forest, and boosted tree model.

If deployed, the data mining solution should predict residents opening a bank account with at least 84.1% accuracy and be scalable to a larger amount of data without slowing down current business processes. It should also give insight into which variables to focus on to increase financial participation, but we understand the solution will not be an ultimate fix. Our hope is that the data mining solution that we provide will help the governments determine the key indicators to focus on, and in turn, will help optimize the financial participation and consumer experience.

1. Business Question: How to increase the financial participation of the population in Kenya and Uganda? Who should be the banks' targeted customers?

Business Case: Compared to other continents, Africa is the least developed in terms of financial participation. The financial participation status in a subset of African nations (based on the countries we are focusing on) is ~ 83.7% in Kenya *(Xinhua, 2021)* and 49% in Uganda *(Hamdan, Lehmann-Uschner and Menkhoff, 2021)*. These figures represent various financial services, ranging from formal to transitory mobile banking. Meanwhile, the rest of the population relies on paper/cash transactions. The participation of a population in the banking system has wide-ranging benefits. The government benefits from improved tax collection, enhanced economy monitoring, and streamlined accountability. Utilizing banking channels for transactions stimulates the economy, especially in the E-commerce sector. In addition to the ease of transactions, people who open bank accounts benefit from efficient financial management services. Financial Participation means allowing a large number of individuals to open an account with a formal banking institution and avail of banking services ranging from direct deposits to withdrawals and internet banking.

Who will benefit from this? Banks with a more extensive clientele base would benefit from increased revenue and can utilize it to invest in profitable ventures. Furthermore, banks can leverage their clients' data to derive insights and design refined marketing campaigns for different target audiences. Additionally, a solid database can help banks offer customized products to attract more diverse groups of customers.

2. Analytics Question: **Prediction goal**: Classifying whether a person will open a bank account or not, based on the attributes of a person **Interpretation goal**: Identifying the different characteristics of customers which can be utilized by the banks for customer targeting?

Based on our results, Banks can target the right customers. This will help banks cut costs, emphasize product customization, and increase revenue. Gradually, we also expect to see increases in financial participation, thus, improving the general economy, the well-being of people, and financial accountability.

3. Dataset: Based on an ongoing competition at the Zindi Data Science Community platform *(Africa, 2021)*. The dataset involves consolidated survey results by Finscope covering from 2016 to 2018. The original data contains a set of 23,525 rows and 10 possible predictor variables out of 13 columns to build a prediction model. We shall limit ourselves only to data from the year 2018 containing data for Kenya and Uganda. There are 6057 rows for Kenya and 2031 are for Uganda totaling up to 8094 rows. The typical vital predictors are the demographic information of a customer: *gender, age, employment status, education levels, marital status, household size, cell phone access, country, and relationship with the head of the household.*

4. Descriptive Analytics

Since seven out of nine variables in our data, including the binary outcome variable bank_account, are categorical variables, we use the ANOVA test for *bank_account* against numeric predictor variables *age* and *household_size*, and Chi-squared tests to explore the dependency degree of outcome variable against other categorical predictors and between each predictor variable.

ANOVA Test *[Appendix I]* The boxplot shows that means of *household_size* between 2 groups of those who open and don't open bank accounts are very close to each other. In addition, the boxes overlap each other. Thus, there is less variability in median *household_size*. On the other hand, for age, the variability of the mean *age* is almost the same for those who open and don't open bank accounts. Nevertheless, the ANOVA test result gives us a more precise result of the mean value yield between 2 groups of the outcome; as the p-value is very small (p-value < 0.0001) for both tests, it still shows that there are differences between the respective mean of *household_size* and *age* among two groups of the outcome.

Chi-square Test *[Appendix II]*

- **The outcome against categorical variables**: The Chi-squared tests are all significant given the magnitude of the extremely small p-value (< 0.0001) yielded. Thus, we can reject the NULL Hypothesis and conclude that there is sufficient evidence to prove that the response variable *bank_account* is dependent on all categorical predictor variables.

- **Predictor variable against each other**: The Chi-squared test results are significant, given that most of the p-values are very small, except for the case of *{location_type, gender}*. Thus, we can reject the NULL Hypothesis of independent predictor variables, except for the case of *{location_type, gender}* with a p-value of 0.899. This is a red flag of multicollinearity.

- **<u>Variables</u>:** For our models, we will be using the following variables:

1. **Gender** variable identifies if the person is Male or Female. **Female** is the reference level.
2. **Age** is a quantitative variable containing the age of the respondent. The minimum value is 16 and the maximum is 98.
3. **Employment_status** offers information about employment status and includes status such as:- Self-employed, Government dependent, Formally employed Private, Informally employed, Formally employed Government, Farming and Fishing, Remittance Dependent, Other Income, No Income. We have chosen **Self-employed** as the reference level.
4. **Education_level** provides information about the completed education level of a person. Education levels included in our dataset are:- Secondary School, Other, Vocational/Specialized training, Primary School, and Tertiary Education. We have chosen **Other** as our base case.
5. **Marital_status** indicates the marital status of a person. Statuses included in our dataset are: Married, Widowed, Single. We have chosen **Single** as the reference level.
6. **Household size** is a quantitative variable that shows the family size of a person opening a bank account. The lowest is 1 and the maximum is 21.
7. **Location_type** gives the location of the person either in a Rural or an Urban area. **Rural** is the reference level.
8. **Relationship with the head of the household** tells us the relationship between the respondent and the head of the household. Relationships included are Spouse, Head of the Household, another relative, Child, Parent, and another non-relative. We have chosen **Other non-relative** as the reference level.
9. **Cell Phone Access** indicates if the person has access to a cell phone. We have chosen **Yes** as the reference level.

<u>Data Pre-Processing and Transformations</u>

After carefully evaluating, we removed categories that don't add value to our analysis. In particular, they are: “marital_status = Don't know” and “job_type = Don't Know/Refuse to answer”.

Furthermore, to make data look more appealing and easy to read, we renamed and recoded category names:

marital_status: “Divorce/Separate” into “Divorce”, “Married/Living together” into “Married”, and “Single/Never Married” into “Single”. **Education_level:** “No formal education” and “Others/Don't Know/RTA” into “Others”, “Primary Education” into “Primary School”, and “Secondary Education” into “Secondary School”.

In addition, since there are only 7 observations for the "Government Dependent" *job_type* for Uganda, while it is 65 for Kenya, we randomly sampled three rows from the existing data and added them back to the 2018 dataset.

5. Modeling Method and Specification

5.1 Initial Logistic Regression Model: A logistic regression model was fit using all the 10 predictors part of the data, the results suggest that this is a good model to predict whether a person will open a bank account or not. The results indicate that all variables were significant, and help reduce the deviance in the model relative to the null model (deviance reduces from 8307 to 5875). The predictors that were significant at the 0.01 level included country, location_type, cellphone_access, education_level, job_type, age, and gender.

5.2 Assumptions tested: Based on the results of the logistical regression the predictors, in this case, were independent, the condition Index (CI) of 49.1 is high but it is below the threshold of 50, which would suggest that our model's multicollinearity is tolerable. In addition, all the predictor's GVIF were below 5 – the highest being 3.32 which is well below the 10 threshold. So, hence suggests that the logistic model with all the variables is a good model.

5.3 Model specifications evaluation: The first model specification used in our analysis was the initial logistical model containing all the 10 predictors (Model A). The second model specification was selected through stepwise variable selection with the p-value threshold being placed at 0.05. This was done to make sure only significant predictors were used. Based on the results of the stepwise variable selection, it suggests that we exclude the *household size* from the model as it was deemed not to be significant (Model B). Which made sense as *household size* should not play a role in whether a person opens a bank account or not. Hence, it was decided to exclude *household size* from our model.

5.3.a Assumptions tested *(Model B - excluding household_size)*: With the exclusion of the *household-size* from the model, the condition index (CI) drops from 49.1 to 46.6 and the highest GVIF was 3.0, and since all the other GVIF were less than 10 suggesting that the multicollinearity levels are tolerable for this specification. Hence, suggests that this model without *household size* is better than the null and the full model.
Note: Since we use Logistic Regression for interpretation, our team decided not to use cross-validation testing for this model.

5.4 Method Evaluated: We choose Random Forest and Boosted Trees for Predictive Accuracy. This aligns with our goals of predicting if a person will open a bank account or not. We selected Logistic Regression for Interpretation as explained above. We used 10 Fold CV to train the model. Random Forest and Boosted Trees were applied and ran on both Model A and Model B. Out of the two models, Model B performed better at both methods. Therefore, we select the specification Model B for further examining cross-validation.

For Random Forest, age is the most important variable, followed by job_type and education level (post-secondary school). At 50% threshold Random Forest model B gives 84.1% accuracy, 43.3% sensitivity and 94.9% Specificity. The model was better at predicting true negatives over true positives, in other words the model did a great job at predicting who will not open a bank account (not open 94.9% vs 43.3% will open) . This result is explainable as the dataset involves major observations of people who did not open a bank account. Given the selected threshold of 50%, we yield an ROCR curve with the area under the curve being 69.1%.

For Boosted Trees, education level, job type, and age are the top 3 most important variables to split Boosted Trees. However, for Boosted Tree, education level plays the most significant roles, followed by job type and age, while age was the most important in Random Forest. While comparing the prediction results using three levels of threshold, the model yields the best result at a threshold of 50%. The corresponding area under the ROCR curve is 69.8%. At 50% Threshold Boosted Trees model gives 84.4% accuracy, 44.8% sensitivity and 94.8% Specificity. Similar to the case for Random Forest, Boosted Tree model B did a better job at predicting true negatives over true positives, the model was significantly better at predicting when a person will not open a bank account (94.8%) over opening one (44.8%). Again, such results are expected given the skewness in our dataset.

For Logistic Regression, given the threshold of 50%, the model returns the best prediction accuracy of 84.6%, Specificity of 95%, and Sensitivity of 44.8%. Similar to Boosted Trees, the model gains a result of 66.1% for Sensitivity when we pull down the threshold to 30%, and a dramatic sacrification of Sensitivity for Specificity when the threshold is 70%. Additionally, the area under the ROCR curve, at 50% threshold, is 69.9%. At the first glance, the Logistic Regression model outperformed all regression methods. Nevertheless, given the Condition Index of 46.6, not yet crossing the threshold of 50 for intolerable multicollinearity, the CI falls to the higher end of consideration range. Thus, a non-parametric method would be more preferable in our case.

5.5 Cross Validation and Revalidation: We use the confusion matrix and statistical indexes to conduct the cross-validation on our models. Based on the results described above, given the chosen threshold of 50%, Boosted Trees outperformed Random Forest with a higher general accuracy rate by 0.3%, while we observe that the Specificity from Boosted Tree is 0.1% less than that from Random Forest, it's probably due to rounding issues. Moreover, at a threshold of 50%, both Random Forest and Boosted Trees return an ROCR curve that does not hug the top left corner tightly, since both models did a great job at predicting the negatives than the positives. However, Boosted Trees returned a higher value of the area under the curve by 0.7% and it is much closer to the acceptable area of 70%. Thus, we select Boosted Tree as our best model for prediction accuracy.

6. Revalidation Result Analysis

Based on the model selection for prediction accuracy, we revalidate Boosted Trees using 10-FCV technique on the full dataset. At 50% Threshold we got an accuracy of 84.1%, a sensitivity of 41.6%, a specificity of 95.4%. MSE is 0.773 at 2898 Trees. Our best model does extremely well in predicting True Negatives. Compared to the Boosted Trees on the train set that we selected above, this new Boosted Tree performs a slightly higher Specificity rate (by 0.6%), but less accurately by 0.3% in general. Thus, depending on the business requirement, either model can be used for future prediction. If banks want to focus more on predicting who won't open a bank account, they can apply the new Boosted Trees on the whole dataset, otherwise, utilizing the Boosted Trees on the training set.

7. Interpretation of Results

The team decided to use the logistical regression model for interpretation, and only focus on education level and job type as these predictors played a major role in predicting whether a person will open a bank account or not, so hence, the effect of job type and education level on whether a person will or will not open a bank account were significant.

On average and holding everything else constant, compared to the job type *'self employed'*, the people part of the *Fishing & Farming, No Income,* and *Informally employed/Remittance dependent* groups reduce the log-odds of opening a bank account by -0.359, -0.903, and -0.897 units respectively and the odds of that happening decrease by a factor of 0.699, 0.405 and 0.408 respectively.

Coming to education level, on average and holding everything else, compared to the education level 'others', the people that completed their Secondary School, Primary School, Vocational/Specialized Training and Tertiary Education increase the log-odds of opening a bank account by 1.412, 0.508, 2.322, and 3.156 units respectively and the odds of that happening increased by a factor of 4.106, 1.662, 10.198 and 23.469 respectively.

8. Conclusion

Of the different models used, the best model for deployment for predictive accuracy seems to be Boosted Trees, along with the logistical regression model for interpretation. In order to assess projected improvements the banks should consider implementing the boosted tree and logistic regression model, on a trial basis to evaluate how effective the models are in predicting future bank account openings or non-openings, Banks shall gain valuable insights into the attributes of the potential customers. This trial run can help banks determine if the boosted trees and logistical model is worth the initial investment and demonstrate the value of such data analysis for them. It seems likely that a trial will demonstrate the value of this sort of modeling and justify the investment.

Appendix

Data ETL

```
library(tidyverse)

bank_acct <- read_csv("./data/Train.csv")

## Rows: 23524 Columns: 13
```

We are only interested in 2018 data points. Below is Data ETL:

```
bank_acct_clean <- bank_acct %>%
  filter(year == 2018) %>%
  filter(job_type != "Dont Know/Refuse to answer", marital_status != "Dont know") %>%
  mutate(
    marital_status = case_when(marital_status == "Divorce/Separate" ~ "Divorce",
                               marital_status == "Married/Living together" ~ "Married",
                               marital_status == "Single/Never Married" ~ "Single",
                               TRUE ~ "Widowed"),
    education_level = case_when(
      education_level %in% c("No formal education", "Other/Dont know/RTA") ~ "Others",
      education_level == "Primary education" ~ "Primary School",
      education_level == "Secondary education" ~ "Secondary School",
      TRUE ~ education_level),
    age = age_of_respondent,
    gender = gender_of_respondent,
    bank_account = case_when(bank_account == "Yes" ~ 1,
                             TRUE ~ 0)
  ) %>%
  select(-c(age_of_respondent, gender_of_respondent))
```

Since there are only 7 observations for `Government Dependent` job_type for Uganda, while it is 65 for Kenya. We randomly sample 3 rows from the existing data. After that, we combine the additional sample data with the initial bank_account dataset.

```
# Sampling the Uganda - Government Dependent job_type
RNGkind(sample.kind="default")
set.seed(1)
uganda_gvt <- bank_acct_clean %>%
  filter(country == "Uganda", job_type == "Government Dependent")
sample <- sample(nrow(uganda_gvt), 0.5 * nrow(uganda_gvt))
uganda_gvt_sample <- uganda_gvt[sample,]

# Add the sample data to bank_account data
bank_acct_smp <- bank_acct_clean %>% rbind(uganda_gvt_sample)
```

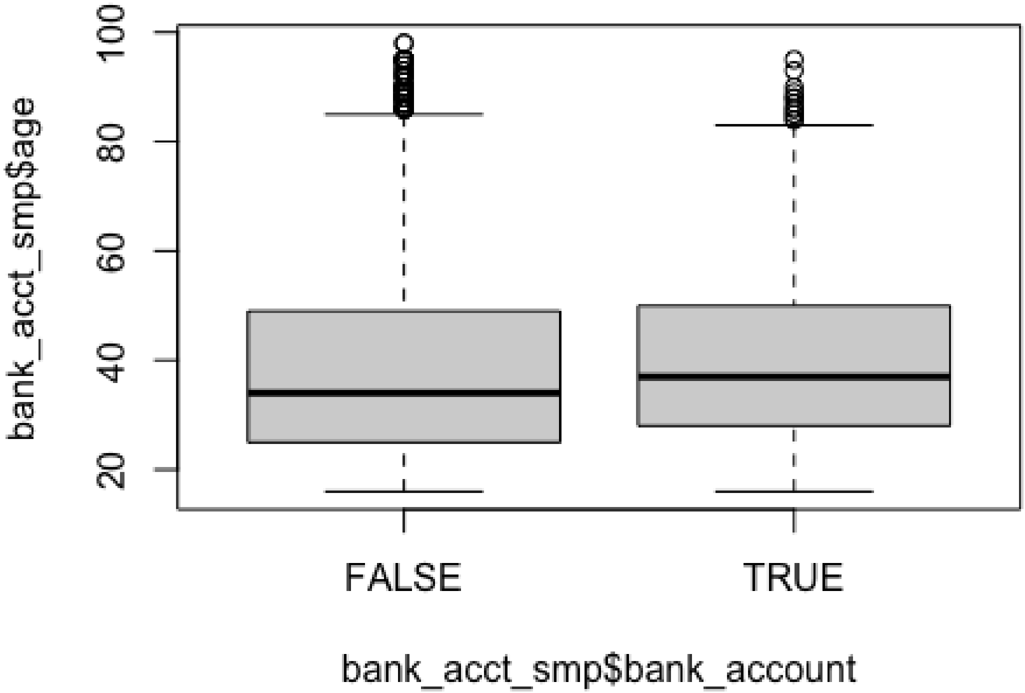
Preliminary Analysis

Assumption: Predictors and Response variables have a linear relationship

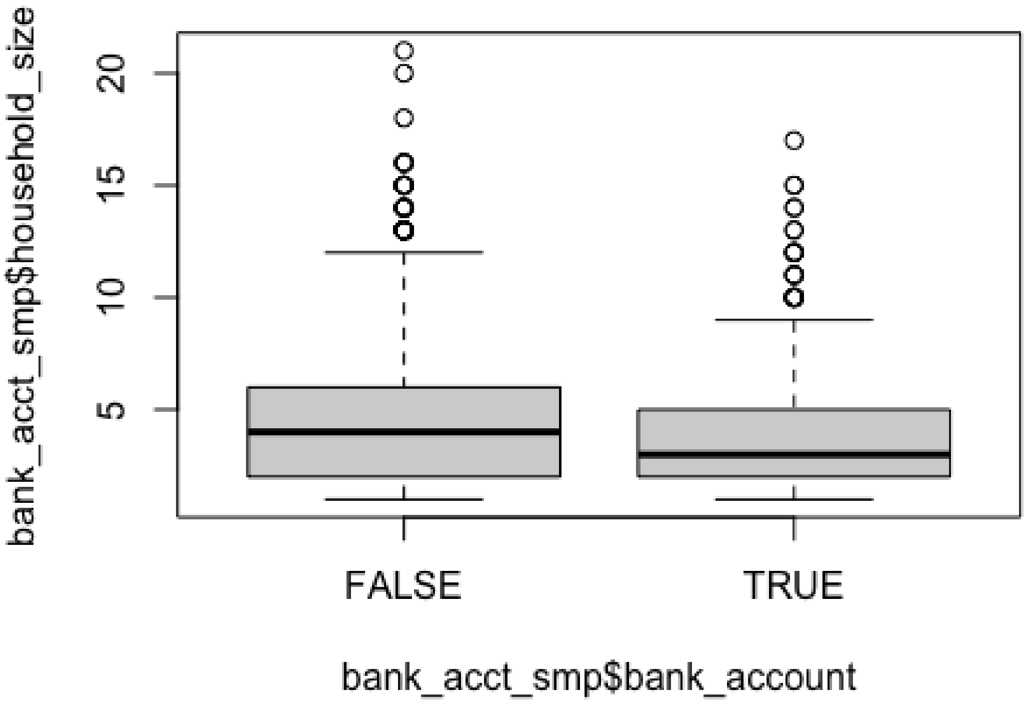
ANOVA Test

Inspect outcome variable **bank_account** and 2 numeric predictor variables **age** and **household_size**.

```
bank_acct_smp$bank_account <- bank_acct_smp$bank_account > 0.5

# Visualization - boxplot
boxplot(bank_acct_smp$age~bank_acct_smp$bank_account)
```



```
boxplot(bank_acct_smp$household_size ~ bank_acct_smp$bank_account)
```



```
# ANOVA Test
summary(aov(bank_account ~ age, data = as.data.frame(bank_acct_smp)))

##                Df Sum Sq Mean Sq F value     Pr(>F)    
## age             1    3.5   3.547   21.48 0.00000364 ***
## Residuals    8092 1336.5   0.165                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(aov(bank_account ~ household_size, data = as.data.frame(bank_acct_smp)))

##                  Df Sum Sq Mean Sq F value Pr(>F)    
## household_size    1   26.8  26.822   165.3 <2e-16 ***
## Residuals      8092 1313.2   0.162                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Inspect linear relationship between logit of outcome and numeric predictors

1. Between **bank_account** and **age**

```
quantile(bank_acct_smp$age, probs = c(0, 0.25, 0.5, 0.75, 1))

##   0%  25%  50%  75% 100% 
##   16   25   35   49   98

summary(bank_acct_smp$age)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   16.00   25.00   35.00   38.57   49.00   98.00
```

```
# Linear Spline with 3 knots for Age
library(ISLR)
library(splines)
spline.lm.age <- glm(bank_account~bs(age, knots = c(25, 35, 49), degree = 1),
              data = bank_acct_smp,
              family = binomial(link = "logit"))
summary(spline.lm.age)
```

```
##
## Call:
## glm(formula = bank_account ~ bs(age, knots = c(25, 35, 49), degree = 1),
##     family = binomial(link = "logit"), data = bank_acct_smp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7841  -0.7532  -0.6986  -0.3233   2.5243
##
## Coefficients:
##                                               Estimate Std. Error z value
## (Intercept)                                    -3.1438     0.1632 -19.269
## bs(age, knots = c(25, 35, 49), degree = 1)1   1.9669     0.1968   9.997
## bs(age, knots = c(25, 35, 49), degree = 1)2   2.1218     0.1703  12.461
## bs(age, knots = c(25, 35, 49), degree = 1)3   1.9864     0.1783  11.138
## bs(age, knots = c(25, 35, 49), degree = 1)4   1.4618     0.2448   5.971
##                                               Pr(>|z|)
## (Intercept)                                    < 2e-16 ***
## bs(age, knots = c(25, 35, 49), degree = 1)1  < 2e-16 ***
## bs(age, knots = c(25, 35, 49), degree = 1)2  < 2e-16 ***
## bs(age, knots = c(25, 35, 49), degree = 1)3  < 2e-16 ***
## bs(age, knots = c(25, 35, 49), degree = 1)4 2.35e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8307.4  on 8093  degrees of freedom
## Residual deviance: 8078.4  on 8089  degrees of freedom
## AIC: 8088.4
##
## Number of Fisher Scoring iterations: 5
```

Analysis

2. Between **bank_account** and **household_size**

```
# For household_size
quantile(bank_acct_smp$household_size, probs = c(0, 0.25, 0.5, 0.75, 1))
```

```
##   0%  25%  50%  75% 100%
##    1    2    4    6   21
```

```
summary(bank_acct_smp$household_size)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   4.000   4.243   6.000  21.000
```

```
spline.lm.hh <- glm(bank_account~bs(household_size, knots = c(2, 4, 6), degree = 1),
              data = bank_acct_smp,
              family = binomial(link = "logit"))
summary(spline.lm.hh)
```

```
##
## Call:
## glm(formula = bank_account ~ bs(household_size, knots = c(2,
##     4, 6), degree = 1), family = binomial(link = "logit"), data = bank_acct_smp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8666  -0.7305  -0.6267  -0.5051   2.2805
##
## Coefficients:
##                                                     Estimate Std. Error z value
## (Intercept)                                         -0.78596    0.06315 -12.446
## bs(household_size, knots = c(2, 4, 6), degree = 1)1 -0.32476    0.08934  -3.635
## bs(household_size, knots = c(2, 4, 6), degree = 1)2 -0.47283    0.08615  -5.489
## bs(household_size, knots = c(2, 4, 6), degree = 1)3 -1.01080    0.09620 -10.507
## bs(household_size, knots = c(2, 4, 6), degree = 1)4 -2.00129    0.50559  -3.958
##                                                         Pr(>|z|)
## (Intercept)                                          < 2e-16 ***
## bs(household_size, knots = c(2, 4, 6), degree = 1)1     0.000278 ***
## bs(household_size, knots = c(2, 4, 6), degree = 1)2 0.0000000405 ***
## bs(household_size, knots = c(2, 4, 6), degree = 1)3      < 2e-16 ***
## bs(household_size, knots = c(2, 4, 6), degree = 1)4 0.0000754831 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8307.4  on 8093  degrees of freedom
## Residual deviance: 8121.5  on 8089  degrees of freedom
## AIC: 8131.5
##
## Number of Fisher Scoring iterations: 4
```

Analysis

Chi-squared Test

1. Data Preparation

Since our data contains many categorical variables, we will conduct Chi-squared statistic test to inspect the possible relationship among these variables.

As the first step, we extract only categorical columns from the initial dataset.

```
library(magrittr)

library(tidyverse)
bank_acct_smp2 <- bank_acct_smp %>% mutate(bank_account = as.character(bank_account))

bank_acct_smp2_chisq <- bank_acct_smp2 %>%
  select(bank_account, location_type, cellphone_access, relationship_with_head,
marital_status, education_level, job_type, gender) %>%
  mutate_all(~as.factor(.))
```

2. Chi-squared Test Statistic

- Outcome variable **bank_account** against all predictors

```
##                         statistic p.value
## location_type           437.5763  3.647001e-97
## cellphone_access        440.2351  9.622234e-98
## relationship_with_head 186.3332  2.373989e-38
## marital_status          35.58392  0.0000000187521
## education_level         1378.213  3.661546e-297
## job_type                1124.838  1.654915e-237
## gender                  144.216   3.186948e-33
```

- **location_type** versus other variables

```
##                         statistic  p.value
## bank_account            437.5763   3.647001e-97
## cellphone_access        263.0944   3.631011e-59
## relationship_with_head 50.06058   1.346799e-09
## marital_status          167.6987   3.843091e-37
## education_level         749.5602   6.45671e-161
## job_type                660.7277   2.030365e-137
## gender                  0.01591681 0.8996038
```

- **cellphone_access** versus other variables

```
##                         statistic p.value
## bank_account            440.2351  9.622234e-98
## location_type           263.0944  3.631011e-59
## relationship_with_head 245.9606  4.043241e-51
## marital_status          190.0031  5.512546e-42
## education_level         658.1207  4.069102e-141
## job_type                545.0162  1.5276e-112
## gender                  41.46626  1.199227e-10
```

- **relationship_with_head** versus other variables

```
##                  statistic p.value
## bank_account     186.3332  2.373989e-38
## location_type    50.06058  1.346799e-09
## cellphone_access 245.9606  4.043241e-51
## marital_status   4388.564  0
## education_level  436.3508  5.692894e-80
## job_type         1668.315  0
## gender           1860.181  0
```

- **marital_status** versus other variables

```
##                         statistic p.value
## bank_account            35.58392  0.0000000187521
## location_type           167.6987  3.843091e-37
## cellphone_access        190.0031  5.512546e-42
## relationship_with_head 4388.564  0
## education_level         1030.314  4.268971e-217
## job_type                918.707   2.783284e-185
## gender                  290.0999  1.012983e-63
```

- **education_level** versus other variables

```
##                        statistic p.value
## bank_account           1378.213  3.661546e-297
## location_type          749.5602  6.45671e-161
## cellphone_access       658.1207  4.069102e-141
## relationship_with_head 436.3508  5.692894e-80
## marital_status         1030.314  4.268971e-217
## job_type               1612.213  2.511978e-319
## gender                 174.6268  1.062441e-36
```

- **job_type** versus other variables

```
##                        statistic p.value
## bank_account           1124.838  1.654915e-237
## location_type          660.7277  2.030365e-137
## cellphone_access       545.0162  1.5276e-112
## relationship_with_head 1668.315  0
## marital_status         918.707   2.783284e-185
## education_level        1612.213  2.511978e-319
## gender                 242.9802  5.293585e-48
```

- **gender** versus other variables

```


##                        statistic  p.value
## bank_account           144.216    3.186948e-33
## location_type          0.01591681 0.8996038
## cellphone_access       41.46626   1.199227e-10
## relationship_with_head 1860.181   0
## marital_status         290.0999   1.012983e-63
## education_level        174.6268   1.062441e-36
## job_type               242.9802   5.293585e-48
```

Preliminary Analysis

Logistic Regression on full dataset

1. Run Logistic Regression model on full dataset after releveling

Now, let’s relevel the reference for categorical variables:

```
bank_acct_smp$relationship_with_head <- relevel(bank_acct_smp$relationship_with_head, "Other 
non-relatives")
bank_acct_smp$education_level <- relevel(bank_acct_smp$education_level, "Others")
bank_acct_smp$marital_status <- relevel(bank_acct_smp$marital_status, "Single")

bank_acct_smp$country <- relevel(bank_acct_smp$country, “Uganda”)
```

```
# Fit the model again using full dataset
Call:
glm(formula = bank_account ~ ., family = binomial(link = "logit"), 
    data = bank.train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max
```

```
-2.5078  -0.5760  -0.3488  -0.1447   3.1472

Coefficients: (3 not defined because of singularities)
                                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                                    -5.978209   0.615363  -9.715  < 2e-16 ***
countryKenya                                    1.327651   0.136709   9.712  < 2e-16 ***
year                                                  NA         NA      NA       NA    
location_typeUrban                              0.734275   0.087286   8.412  < 2e-16 ***
cellphone_accessNo                             -1.280056   0.157908  -8.106 5.22e-16 ***
age_of_respondent                               0.016426   0.003237   5.075 3.87e-07 ***
gender_of_respondentMale                        0.478750   0.101820   4.702 2.58e-06 ***
relationship_with_headSpouse                    1.514107   0.582559   2.599  0.00935 ** 
relationship_with_headHead of Household         1.609068   0.569274   2.827  0.00471 ** 
relationship_with_headOther relative            1.322573   0.609656   2.169  0.03005 *  
relationship_with_headChild                     1.001841   0.586587   1.708  0.08765 .  
relationship_with_headParent                    1.670256   0.645488   2.588  0.00967 ** 
marital_statusMarried                           0.253047   0.127261   1.988  0.04677 *  
marital_statusWidowed                           0.399846   0.163351   2.448  0.01437 *  
education_levelSecondary School                 1.370248   0.161848   8.466  < 2e-16 ***
education_levelVocational/Specialised training  2.333255   0.197148  11.835  < 2e-16 ***
education_levelPrimary School                   0.472457   0.155177   3.045  0.00233 ** 
education_levelTertiary education               2.943509   0.234021  12.578  < 2e-16 ***
job_typeGovernment Dependent                    0.795842   0.369711   2.153  0.03135 *  
job_typeFormally employed Private               0.605333   0.146832   4.123 3.75e-05 ***
job_typeInformally employed                    -0.896846   0.131291  -6.831 8.44e-12 ***
job_typeFormally employed Government            1.574702   0.237696   6.625 3.48e-11 ***
job_typeFarming and Fishing                    -0.358582   0.125702  -2.853  0.00434 ** 
job_typeRemittance Dependent                   -0.872524   0.150694  -5.790 7.04e-09 ***
job_typeOther Income                            0.052994   0.170052   0.312  0.75532    
job_typeNo Income                              -0.903209   0.484733  -1.863  0.06242 .  
age                                                   NA         NA      NA       NA    
genderMale                                            NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5826.7  on 5664  degrees of freedom
Residual deviance: 4151.1  on 5640  degrees of freedom
AIC: 4201.1

Number of Fisher Scoring iterations: 6
```

```
# Coefficients
```

```
# Coefficients
> log.odds.train.hh <- coef(fit.train.hh)
> odds.train.hh <- exp(log.odds.train.hh)
> prob.train.hh <- odds.train.hh/(1+odds.train.hh)
> cbind(log.odds.train.hh, odds.train.hh, prob.train.hh)
                                               log.odds.train.hh odds.train.hh prob.train.hh
(Intercept)                                          -5.97820867    0.00253336   0.002526959
countryKenya                                          1.32765145    3.77217384   0.790451892
year                                                          NA            NA            NA
location_typeUrban                                    0.73427539    2.08397137   0.675742775
```

```
cellphone_accessNo                                   -1.28005560    0.27802184   0.217540760
age_of_respondent                                     0.01642590    1.01656155   0.504106383
gender_of_respondentMale                              0.47875021    1.61405591   0.617452711
relationship_with_headSpouse                          1.51410706    4.54536057   0.819669075
relationship_with_headHead of Household               1.60906829    4.99815222   0.833281990
relationship_with_headOther relative                  1.32257306    3.75306584   0.789609479
relationship_with_headChild                           1.00184098    2.72329074   0.731420383
relationship_with_headParent                          1.67025556    5.31352554   0.841609891
marital_statusMarried                                 0.25304704    1.28794386   0.562926338
marital_statusWidowed                                 0.39984561    1.49159439   0.598650565
education_levelSecondary School                       1.37024807    3.93632704   0.797420229
education_levelVocational/Specialised training        2.33325542   10.31145503   0.911594044
education_levelPrimary School                         0.47245658    1.60392954   0.615965030
education_levelTertiary education                     2.94350898   18.98233825   0.949955807
job_typeGovernment Dependent                          0.79584173    2.21630575   0.689084285
job_typeFormally employed Private                     0.60533331    1.83186269   0.646875533
job_typeInformally employed                          -0.89684594    0.40785403   0.289699090
job_typeFormally employed Government                  1.57470210    4.82930277   0.828452898
job_typeFarming and Fishing                          -0.35858191    0.69866639   0.411302888
job_typeRemittance Dependent                         -0.87252428    0.41789533   0.294729323
job_typeOther Income                                  0.05299449    1.05442383   0.513245523
job_typeNo Income                                    -0.90320921    0.40526698   0.288391450
age                                                           NA            NA            NA
genderMale                                                    NA            NA            NA
```

```
# Compute R-sq - deviance
```

```
> # Compute R-sq - deviance
> dev.dif <- abs(fit.train.hh$null.deviance - fit.train.hh$deviance)
> dev.Rsq <- (fit.train.hh$null.deviance - fit.train.hh$deviance)/fit.train.hh$null.deviance
> dev.Rsq
```

Check Multicollinearity:

```
library(klaR)

ci <- cond.index(bank.fit.full, data = bank_acct_smp)
# Extract the largest CI
ci[(length(ci))]

## [1] 49.099977

library(car)

vif(bank.fit.full)

##                            GVIF Df GVIF^(1/(2*Df))
## country              1.567849  1       1.252138
## location_type        1.234101  1       1.110901
## cellphone_access     1.100471  1       1.049033
## household_size       1.301650  1       1.140899
## relationship_with_head 3.360581  5     1.128864
## marital_status       2.771278  2       1.290239
## education_level      1.471370  4       1.049458
## job_type             2.339542  8       1.054558
## age                  1.794639  1       1.339641
## gender               1.669776  1       1.292198
```

Logistic Regression Variable Selection

Chi-square Tests show that country and *household_size* are significantly correlated to each other. Accessing the full model, *country* is significant given a very small p-value < 0.001, while the *household_size is* insignificant at the magnitude of p-value = 0.051 > threshold of 0.05. In fact, as we investigated, *household_size* is a trait of a national population, which can explain the collinearity between these 2 variables. However, *country* contributes much more explanatory power to the outcome, as it also presents other socio-economic factors of the country that directly or indirectly affect the possession of opening a bank account.

Therefore, we select *country* while excluding *household_size* in our second model specification.

First, we fit the model again using the reduced specification:

```
bank.fit.hh <- glm(bank_account~.,
                  data = bank_acct_smp,
                  family = binomial(link = "logit"))
summary(bank.fit.hh)
##
## Call:
## glm(formula = bank_account ~ ., family = binomial(link = "logit"),
##    data = bank_acct_smp)
##
## Deviance Residuals:
##    Min         1Q   Median       3Q    Max
## -2.6558  -0.5693  -0.3451  -0.1417   3.1308
##
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                    -5.892797   0.511710 -11.516
## countryKenya                                 1.334429   0.115469  11.557
## location_typeUrban                           0.649119   0.073229   8.864
## cellphone_accessNo                               -1.310156   0.135808  -9.647
## relationship_with_headSpouse                 1.379325   0.482941   2.856
## relationship_with_headHead of Household      1.488584   0.472749   3.149
## relationship_with_headOther relative         1.175744   0.502459   2.340
## relationship_with_headChild                  0.870641   0.487073   1.787
## relationship_with_headParent                 1.575652   0.535136   2.944
## marital_statusMarried                        0.264336   0.108152   2.444
## marital_statusWidowed                        0.313988   0.138551   2.266
## education_levelSecondary School              1.412455   0.137173  10.297
## education_levelVocational/Specialised training  2.322202   0.163999  14.160
## education_levelPrimary School                0.508262   0.131783   3.857
## education_levelTertiary education            3.155719   0.201215  15.683
## job_typeGovernment Dependent                 0.742306   0.304021   2.442
## job_typeFormally employed Private            0.669656   0.122547   5.464
## job_typeInformally employed                  -0.963166   0.110958  -8.680
## job_typeFormally employed Government         1.478973   0.193379   7.648
## job_typeFarming and Fishing                  -0.358068   0.104975  -3.411
## job_typeRemittance Dependent                 -0.904679   0.126887  -7.130
## job_typeOther Income                         0.135326   0.144085   0.939
## job_typeNo Income                            -1.225673   0.473752  -2.587
## age                                          0.017687   0.002712   6.521
## genderMale                                   0.433804   0.085650   5.065
##                                              Pr(>|z|)
## (Intercept)                                    < 2e-16 ***
## countryKenya                                 < 2e-16 ***
## location_typeUrban                           < 2e-16 ***
## cellphone_accessNo                           < 2e-16 ***
## relationship_with_headSpouse                 0.004289 **
## relationship_with_headHead of Household      0.001640 **
## relationship_with_headOther relative         0.019285 *
## relationship_with_headChild                  0.073857 .
## relationship_with_headParent                 0.003236 **
## marital_statusMarried                        0.014521 *
## marital_statusWidowed                        0.023438 *
## education_levelSecondary School              < 2e-16 ***
## education_levelVocational/Specialised training  < 2e-16 ***
## education_levelPrimary School                0.000115 ***
## education_levelTertiary education            < 2e-16 ***
## job_typeGovernment Dependent                 0.014621 *
## job_typeFormally employed Private            4.64e-08 ***
## job_typeInformally employed                  < 2e-16 ***
## job_typeFormally employed Government         2.04e-14 ***
## job_typeFarming and Fishing                  0.000647 ***
## job_typeRemittance Dependent                 1.01e-12 ***
## job_typeOther Income                         0.347624
## job_typeNo Income                            0.009677 **
## age                                          7.00e-11 ***
## genderMale                                   4.09e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 8307.4  on 8093  degrees of freedom
## Residual deviance: 5879.1  on 8069  degrees of freedom
## AIC: 5929.1
##
## Number of Fisher Scoring iterations: 6
```

Coefficients of the reduced model:

```
##                                                   log.odds           odds
## (Intercept)                                  -5.89279677  0.002759249
## countryKenya                                 1.33442933  3.797828006
## location_typeUrban                           0.64911931  1.913854580
## cellphone_accessNo                           -1.31015609  0.269777943
## relationship_with_headSpouse                 1.37932469  3.972218262
## relationship_with_headHead of Household      1.48858383  4.430816293
## relationship_with_headOther relative         1.17574391  3.240552736
## relationship_with_headChild                  0.87064083  2.388440941
## relationship_with_headParent                 1.57565199  4.833892228
## marital_statusMarried                        0.26433583  1.302565565
## marital_statusWidowed                        0.31398752  1.368872654
## education_levelSecondary School              1.41245490  4.106022908
## education_levelVocational/Specialised training  2.32220187 10.198104469
## education_levelPrimary School                0.50826168  1.662398902
## education_levelTertiary education            3.15571942 23.469915774
## job_typeGovernment Dependent                 0.74230559  2.100773455
## job_typeFormally employed Private            0.66965601  1.953565203
## job_typeInformally employed                  -0.96316611  0.381682524
## job_typeFormally employed Government         1.47897321  4.388437383
## job_typeFarming and Fishing                  -0.35806789  0.699025615
## job_typeRemittance Dependent                 -0.90467942  0.404671595
## job_typeOther Income                         0.13532602  1.144909982
## job_typeNo Income                            -1.22567326  0.293559993
## age                                               0.01768708  1.017844423
## genderMale                                   0.43380352  1.543115649
##                                                     prob
## (Intercept)                                  0.002751656
## countryKenya                                     0.791572353
## location_typeUrban                           0.656811975
## cellphone_accessNo                           0.212460726
## relationship_with_headSpouse                 0.798882521
## relationship_with_headHead of Household      0.815865618
## relationship_with_headOther relative         0.764181685
## relationship_with_headChild                  0.704879023
## relationship_with_headParent                 0.828587852
## marital_statusMarried                        0.565701835
## marital_statusWidowed                        0.577858270
## education_levelSecondary School              0.804152857
## education_levelVocational/Specialised training 0.910699172
## education_levelPrimary School                    0.624398884
## education_levelTertiary education            0.959133492
## job_typeGovernment Dependent                 0.677499819
## job_typeFormally employed Private            0.661426130
## job_typeInformally employed                  0.276244736
## job_typeFormally employed Government         0.814417441
## job_typeFarming and Fishing                  0.411427355
## job_typeRemittance Dependent                 0.288089825
## job_typeOther Income                         0.533779968
## job_typeNo Income                            0.226939604
## age                                          0.504421655
## genderMale                                   0.606781548
```

```
# Compute R-sq - deviance

## [1] 0.2922973
```

Check Multicollinearity:

```
library(klaR)

ci <- cond.index(bank.fit.full, data = bank_acct_smp)
# Extract the Largest CI

## [1] 46.581929

# VIF

##                             GVIF Df GVIF^(1/(2*Df))
## country                1.537302  1        1.239880
## location_type          1.224008  1        1.106349
## cellphone_access       1.098486  1        1.048087
## relationship_with_head 3.024975  5     1.117049
## marital_status         2.529628  2     1.261143
## education_level        1.450003  4        1.047541
## job_type               2.312514  8        1.053793
## age                    1.788493  1        1.337346
## gender                 1.673479  1        1.293630
```

Modeling & Testing

Note: Since our data contains more than 8000 observations, it is a constraint to conduct cross-validation using Leave-One-Out method. Thus, for each model, we perform RSCV and 10FCV methods to compare and select the best model.

Logistic Regression model

Random split cross-validation

1. Create a train and test subset

```
set.seed(1)
tr.prop <- 0.7
train <- sample(nrow(bank_acct_smp), tr.prop * nrow(bank_acct_smp))
bank.train <- bank_acct_smp[train, ]
bank.test <- bank_acct_smp[-train, ]
```

2. Build model on Train subset

```
fit.train <- glm(bank_account~.,
                 data = bank.train,
                 family = binomial(link = "logit"))
summary(fit.train)
##
## Call:
```

```
## glm(formula = bank_account ~ ., family = binomial(link = "logit"), 
##    data = bank.train)
## 
## Deviance Residuals: 
##    Min        1Q    Median       3Q     Max  
## -2.5078  -0.5760  -0.3488  -0.1447   3.1472  
## 
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                      -5.978209   0.615363  -9.715
## countryKenya                                 1.327651   0.136709   9.712
## location_typeUrban                           0.734275   0.087286   8.412
## cellphone_accessNo                           -1.280056   0.157908  -8.106
## relationship_with_headSpouse                 1.514107   0.582559   2.599
## relationship_with_headHead of Household      1.609068   0.569274   2.827
## relationship_with_headOther relative         1.322573   0.609656   2.169
## relationship_with_headChild                  1.001841   0.586587   1.708
## relationship_with_headParent                 1.670256   0.645488   2.588
## marital_statusMarried                        0.253047   0.127261   1.988
## marital_statusWidowed                        0.399846   0.163351   2.448
## education_levelSecondary School              1.370248   0.161848   8.466
## education_levelVocational/Specialised training  2.333255   0.197148  11.835
## education_levelPrimary School                0.472457   0.155177   3.045
## education_levelTertiary education            2.943509   0.234021  12.578
## job_typeGovernment Dependent                 0.795842   0.369711   2.153
## job_typeFormally employed Private             0.605333   0.146832   4.123
## job_typeInformally employed                  -0.896846   0.131291  -6.831
## job_typeFormally employed Government         1.574702   0.237696   6.625
## job_typeFarming and Fishing                  -0.358582   0.125702  -2.853
## job_typeRemittance Dependent                 -0.872524   0.150694  -5.790
## job_typeOther Income                         0.052994   0.170052   0.312
## job_typeNo Income                            -0.903209   0.484733  -1.863
## age                                             0.016426   0.003237   5.075
## genderMale                                   0.478750   0.101820   4.702
##                                              Pr(>|z|)    
## (Intercept)                                   < 2e-16 ***
## countryKenya                                 < 2e-16 ***
## location_typeUrban                           < 2e-16 ***
## cellphone_accessNo                           5.22e-16 ***
## relationship_with_headSpouse                    0.00935 ** 
## relationship_with_headHead of Household      0.00471 ** 
## relationship_with_headOther relative         0.03005 *  
## relationship_with_headChild                  0.08765 .  
## relationship_with_headParent                 0.00967 ** 
## marital_statusMarried                        0.04677 *  
## marital_statusWidowed                        0.01437 *  
## education_levelSecondary School              < 2e-16 ***
## education_levelVocational/Specialised training  < 2e-16 ***
## education_levelPrimary School                0.00233 ** 
## education_levelTertiary education            < 2e-16 ***
## job_typeGovernment Dependent                 0.03135 *  
## job_typeFormally employed Private            3.75e-05 ***
## job_typeInformally employed                  8.44e-12 ***
## job_typeFormally employed Government         3.48e-11 ***
## job_typeFarming and Fishing                  0.00434 ** 
## job_typeRemittance Dependent                 7.04e-09 ***
## job_typeOther Income                         0.75532    
## job_typeNo Income                            0.06242 .  
## age                                            3.87e-07 ***
## genderMale                                   2.58e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5826.7  on 5664  degrees of freedom
## Residual deviance: 4151.1  on 5640  degrees of freedom
## AIC: 4201.1
##
## Number of Fisher Scoring iterations: 6
```

```
# Compute R-sq - deviance
```

```
## [1] 0.2875672
```

```
# Coefficients
```

```
##                                               log.odds.train.hh odds.train.hh
## (Intercept)                                        -5.97820867 0.00253336
## countryKenya                                       1.32765145  3.77217384
## location_typeUrban                                 0.73427539  2.08397137
## cellphone_accessNo                                 -1.28005560 0.27802184
## relationship_with_headSpouse                       1.51410706  4.54536057
## relationship_with_headHead of Household            1.60906829  4.99815222
## relationship_with_headOther relative               1.32257306  3.75306584
## relationship_with_headChild                        1.00184098  2.72329074
## relationship_with_headParent                       1.67025556  5.31352554
## marital_statusMarried                              0.25304704  1.28794386
## marital_statusWidowed                                   0.39984561   1.49159439
## education_levelSecondary School                    1.37024807  3.93632704
## education_levelVocational/Specialised training     2.33325542   10.31145503
## education_levelPrimary School                       0.47245658 1.60392954
## education_levelTertiary education                  2.94350898   18.98233825
## job_typeGovernment Dependent                       0.79584173  2.21630575
## job_typeFormally employed Private                     0.60533331     1.83186269
## job_typeInformally employed                        -0.89684594 0.40785403
## job_typeFormally employed Government               1.57470210  4.82930277
## job_typeFarming and Fishing                        -0.35858191 0.69866639
## job_typeRemittance Dependent                       -0.87252428 0.41789533
## job_typeOther Income                               0.05299449  1.05442383
## job_typeNo Income                                  -0.90320921 0.40526698
## age                                                0.01642590  1.01656155
## genderMale                                         0.47875021  1.61405591
##                                               prob.train.hh
## (Intercept)                                     0.002526959
## countryKenya                                       0.790451892
## location_typeUrban                                 0.675742775
## cellphone_accessNo                                 0.217540760
## relationship_with_headSpouse                       0.819669075
## relationship_with_headHead of Household            0.833281990
## relationship_with_headOther relative               0.789609479
## relationship_with_headChild                        0.731420383
## relationship_with_headParent                       0.841609891
## marital_statusMarried                              0.562926338
## marital_statusWidowed                              0.598650565
## education_levelSecondary School                    0.797420229
## education_levelVocational/Specialised training   0.911594044
## education_levelPrimary School                      0.615965030
## education_levelTertiary education                  0.949955807
## job_typeGovernment Dependent                       0.689084285
## job_typeFormally employed Private                  0.646875533
## job_typeInformally employed                        0.289699090
## job_typeFormally employed Government               0.828452898
## job_typeFarming and Fishing                        0.411302888
## job_typeRemittance Dependent                       0.294729323
## job_typeOther Income                               0.513245523
## job_typeNo Income                                  0.288391450
## age                                                0.504106383
## genderMale                                         0.617452711

# Train MSE and RMSE
train.mse <- mean(fit.train$residuals ^ 2)
train.rmse <- sqrt(train.mse)
cbind("MSE Train" = train.mse, "RMSE Train" = train.rmse)
##    MSE Train RMSE Train
## [1,]  17.90787   4.231769
```

2. Model validation without household_size

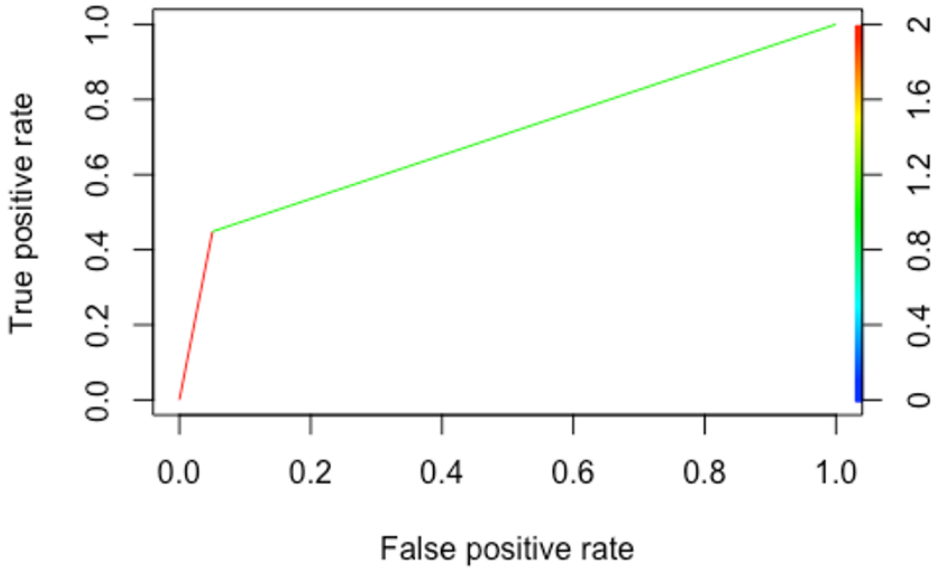
```
pred.hh.test <- predict(fit.train, bank.test, type = "response")
head(pred.hh.test)
##          1          2          3          4          5          6
## 0.12174118 0.02382138 0.91831403 0.39866595 0.17777087 0.24605958
test.hh.mse.rscv <- mean((bank.test$bank_account - pred.hh.test)^2)
test.hh.rmse.rscv <- sqrt(test.hh.mse.rscv)
cbind("RSCV Test MSE" = test.hh.mse.rscv,
      "RSCV Test RMSE" = test.hh.rmse.rscv)
##    RSCV Test MSE RSCV Test RMSE
## [1,]     0.11187     0.3344697
```

Confusion Matrix for Cross-Validation

```
library(ROCR)
thresh <- 0.5
bank.hh.prob.test <-ifelse(pred.hh.test > thresh, 1, 0)
conf.mat <- table("Predicted" = bank.hh.prob.test, "Actual" = bank.test$bank_account)
colnames(conf.mat)<-c("No","Yes")
rownames(conf.mat)<-c("No","Yes")
conf.mat
##          Actual
## Predicted   No  Yes
##    No  1828  278
##    Yes   97  226
```

```
               Accuracy Rate Error Rate Sensitivity Specificity False Positives
Threshold = 30%         0.818      0.182       0.661       0.860           0.140
Threshold = 50%         0.846      0.154       0.448       0.950           0.050
Threshold = 70%         0.834      0.166       0.260       0.984           0.016
```

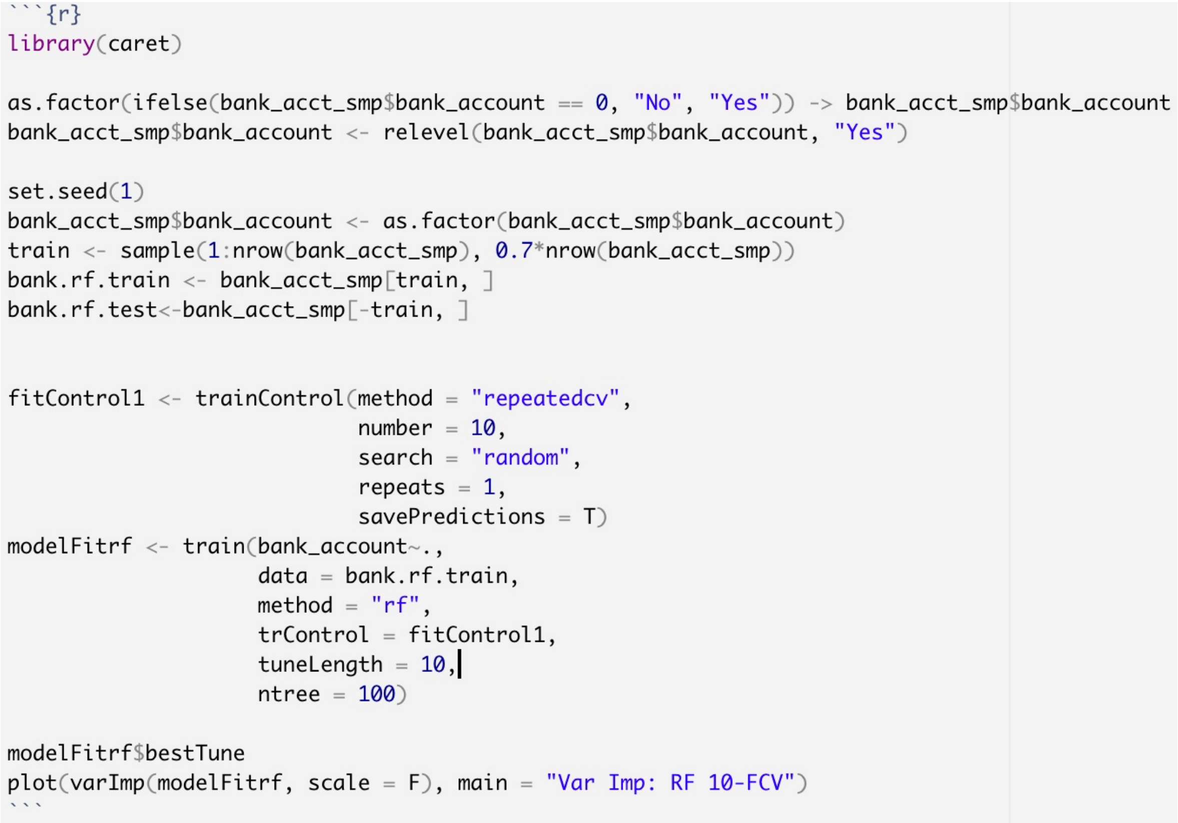
```
# ROC Curve
pred.hh <- prediction(bank.hh.prob.test, bank.test$bank_account)
perf.hh <- performance(pred.hh, "tpr", "fpr")
plot(perf.hh, colorize = T)
```



```
auc.hh <- performance(pred.hh, "auc")
auc.name.hh <- auc.hh@y.name[[1]]
auc.val.hh <- round(auc.hh@y.values[[1]], digits = 3)
paste(auc.name.hh, auc.val.hh)
## [1] "Area under the ROC curve 0.699"
```

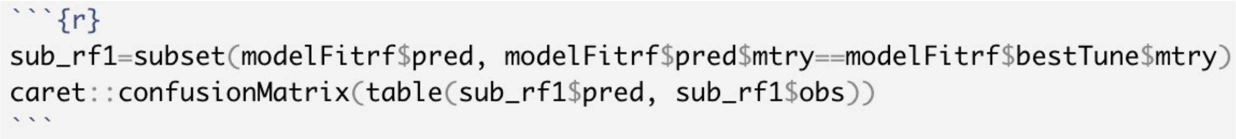
Random Forest

Build model on using Train set and 10-FCV method

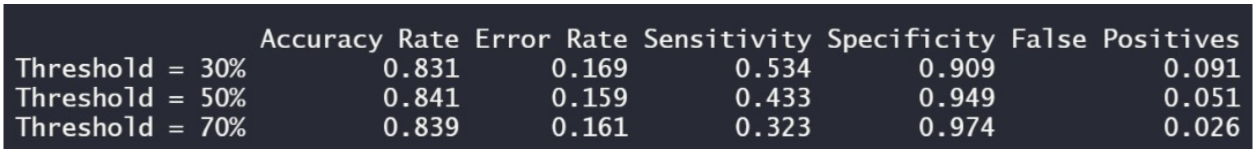


~~~
```{r}
library(caret)

as.factor(ifelse(bank_acct_smp$bank_account == 0, "No", "Yes")) -> bank_acct_smp$bank_account
bank_acct_smp$bank_account <- relevel(bank_acct_smp$bank_account, "Yes")

set.seed(1)
bank_acct_smp$bank_account <- as.factor(bank_acct_smp$bank_account)
train <- sample(1:nrow(bank_acct_smp), 0.7*nrow(bank_acct_smp))
bank.rf.train <- bank_acct_smp[train, ]
bank.rf.test<-bank_acct_smp[-train, ]


fitControl1 <- trainControl(method = "repeatedcv",
                            number = 10,
                            search = "random",
                            repeats = 1,
                            savePredictions = T)
modelFitrf <- train(bank_account~.,
                    data = bank.rf.train,
                    method = "rf",
                    trControl = fitControl1,
                    tuneLength = 10,
                    ntree = 100)

modelFitrf$bestTune
plot(varImp(modelFitrf, scale = F), main = "Var Imp: RF 10-FCV")
```
~~~

Confusion Matrix based on the Test set

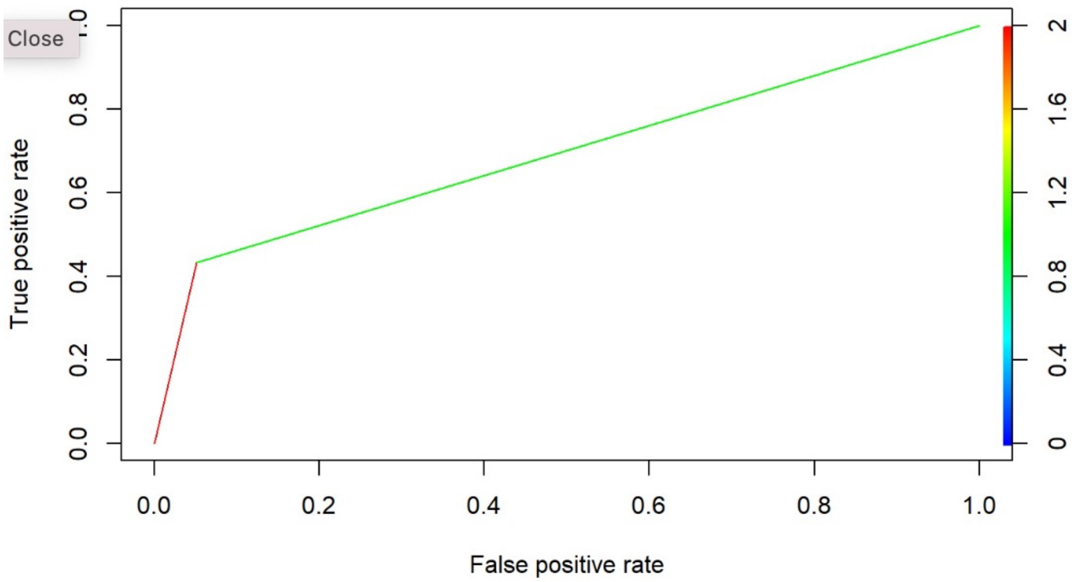


~~~
```{r}
sub_rf1=subset(modelFitrf$pred, modelFitrf$pred$mtry==modelFitrf$bestTune$mtry)
caret::confusionMatrix(table(sub_rf1$pred, sub_rf1$obs))
```
~~~

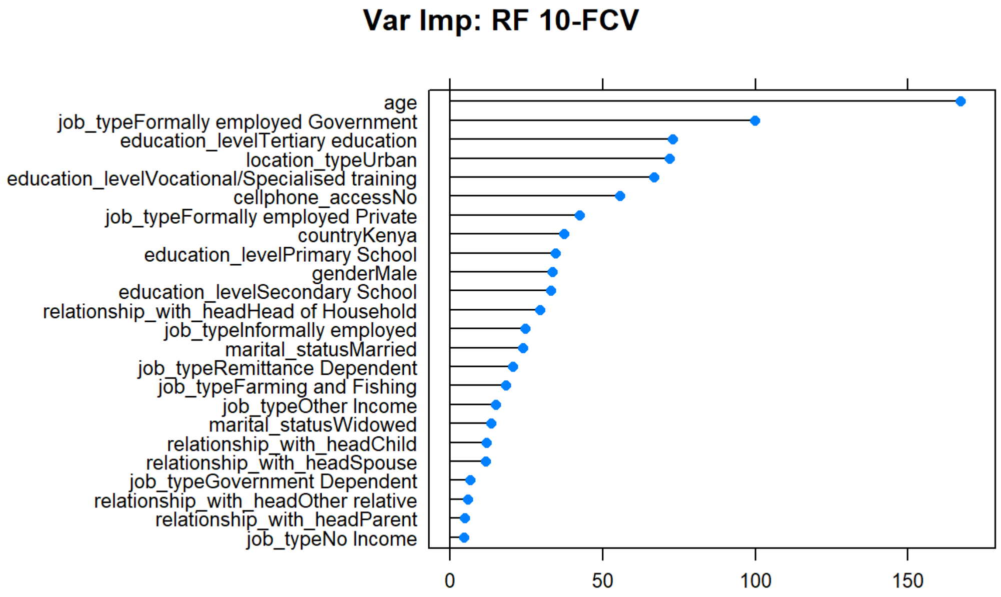


```
                Accuracy Rate Error Rate Sensitivity Specificity False Positives
Threshold = 30%         0.831      0.169       0.534       0.909           0.091
Threshold = 50%         0.841      0.159       0.433       0.949           0.051
Threshold = 70%         0.839      0.161       0.323       0.974           0.026
```

ROCR Curve at threshold = 0.5

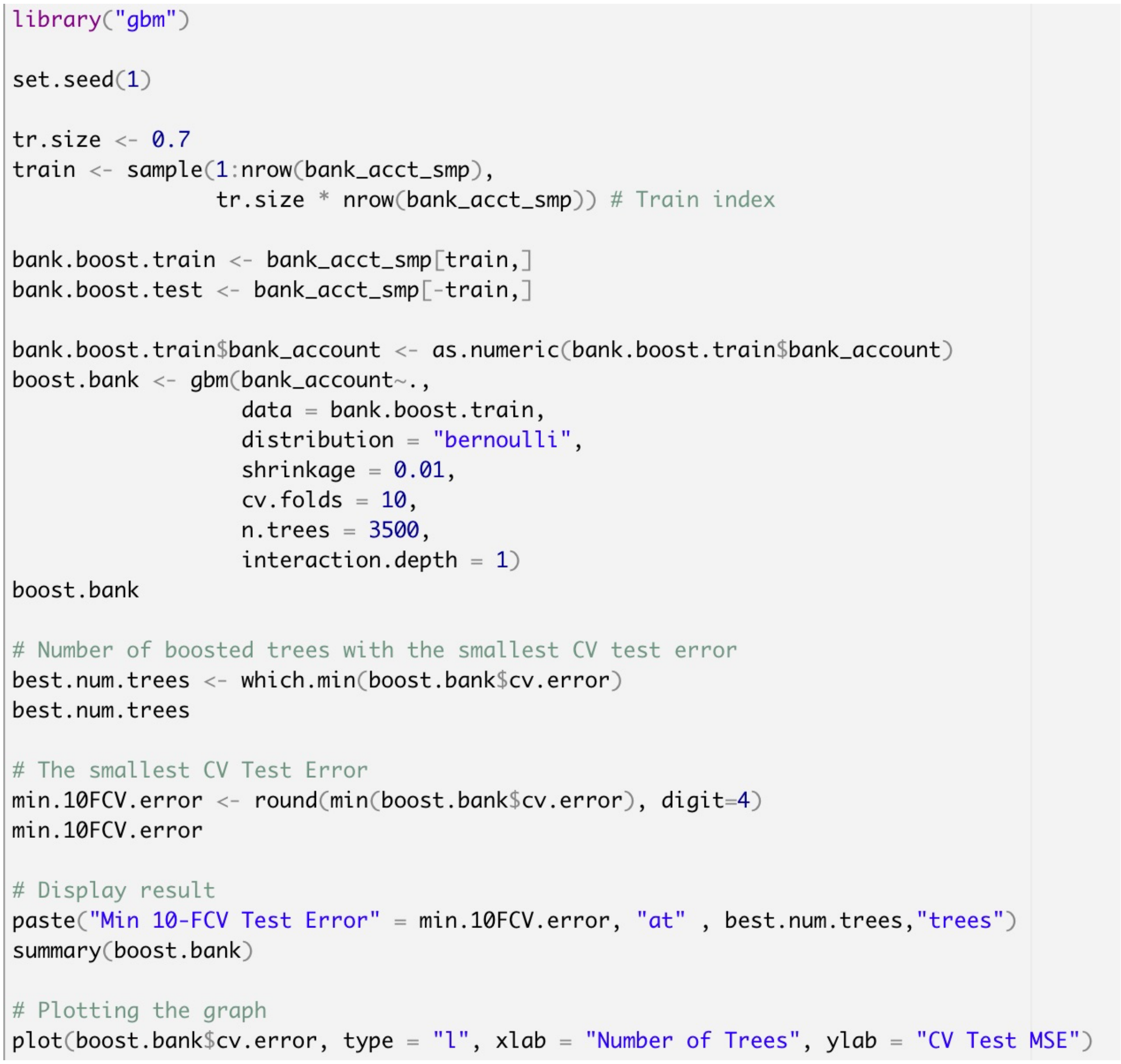


Variable Importance

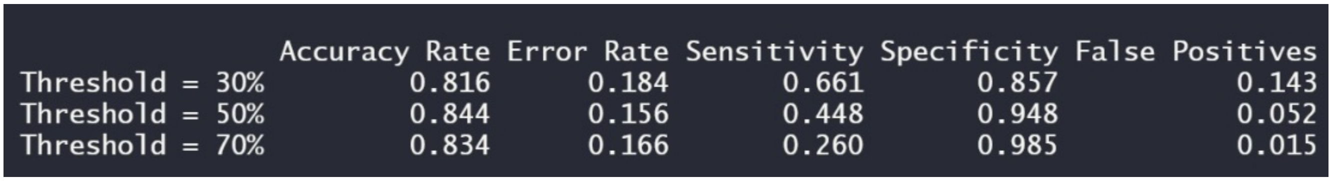


Boosted Trees

Cross-Validation on the Test set

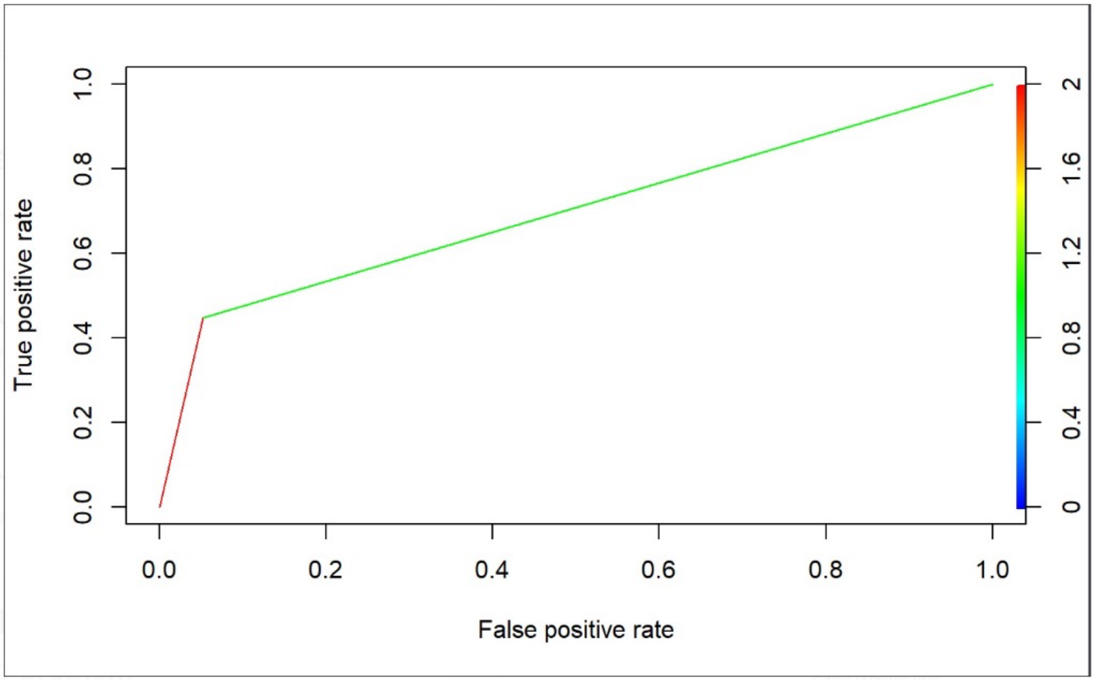


```
library("gbm")

set.seed(1)

tr.size <- 0.7
train <- sample(1:nrow(bank_acct_smp),
                tr.size * nrow(bank_acct_smp)) # Train index

bank.boost.train <- bank_acct_smp[train,]
bank.boost.test <- bank_acct_smp[-train,]

bank.boost.train$bank_account <- as.numeric(bank.boost.train$bank_account)
boost.bank <- gbm(bank_account~.,
                  data = bank.boost.train,
                  distribution = "bernoulli",
                  shrinkage = 0.01,
                  cv.folds = 10,
                  n.trees = 3500,
                  interaction.depth = 1)
boost.bank

# Number of boosted trees with the smallest CV test error
best.num.trees <- which.min(boost.bank$cv.error)
best.num.trees

# The smallest CV Test Error
min.10FCV.error <- round(min(boost.bank$cv.error), digit=4)
min.10FCV.error

# Display result
paste("Min 10-FCV Test Error" = min.10FCV.error, "at" , best.num.trees,"trees")
summary(boost.bank)

# Plotting the graph
plot(boost.bank$cv.error, type = "l", xlab = "Number of Trees", ylab = "CV Test MSE")
```

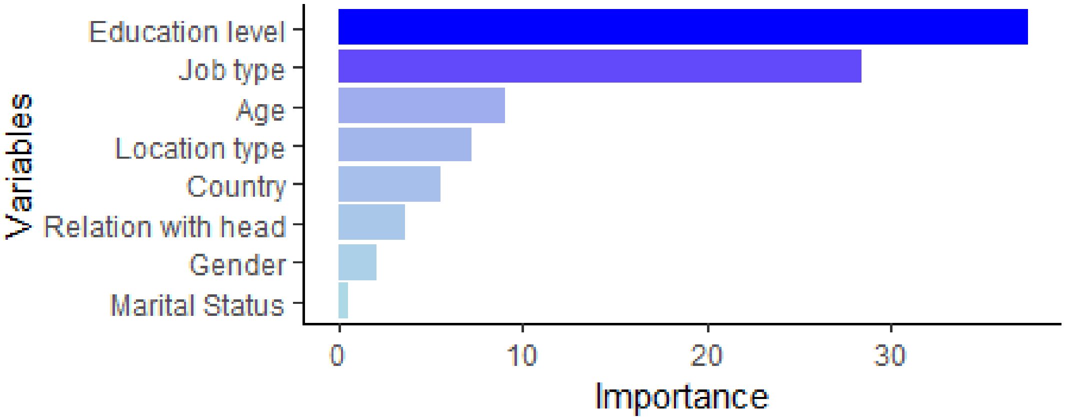


```
                 Accuracy Rate Error Rate Sensitivity Specificity False Positives
Threshold = 30%          0.816      0.184       0.661       0.857           0.143
Threshold = 50%          0.844      0.156       0.448       0.948           0.052
Threshold = 70%          0.834      0.166       0.260       0.985           0.015
```

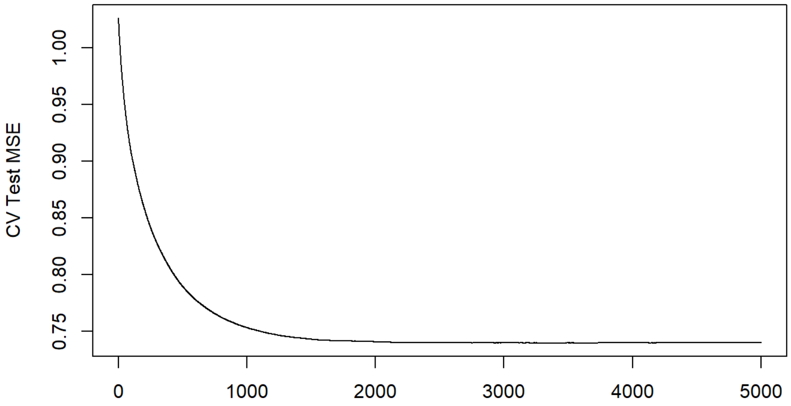
ROCR Curve at threshold = 0.5



Variable Importance

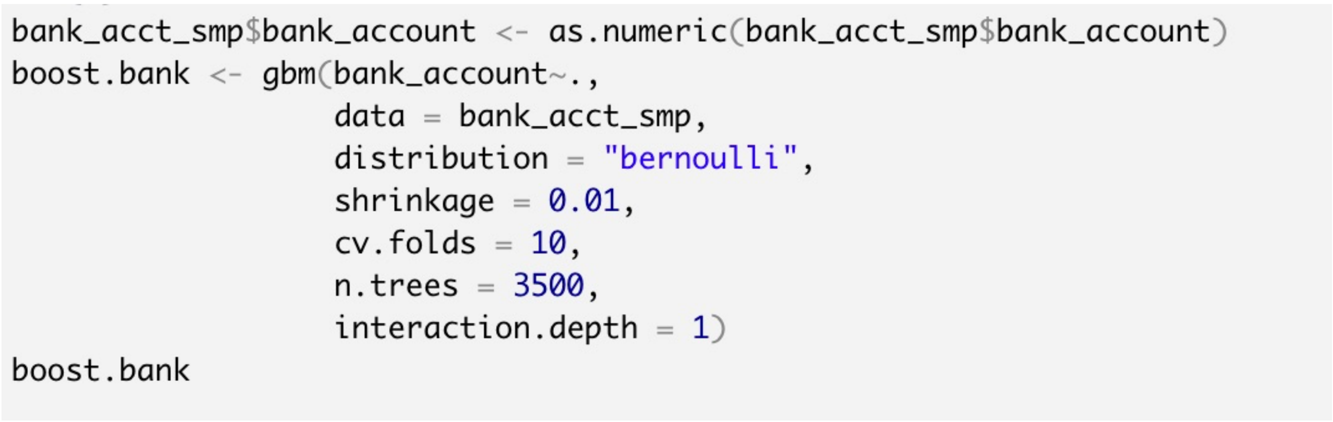


CV Test MSE vs Number of Trees

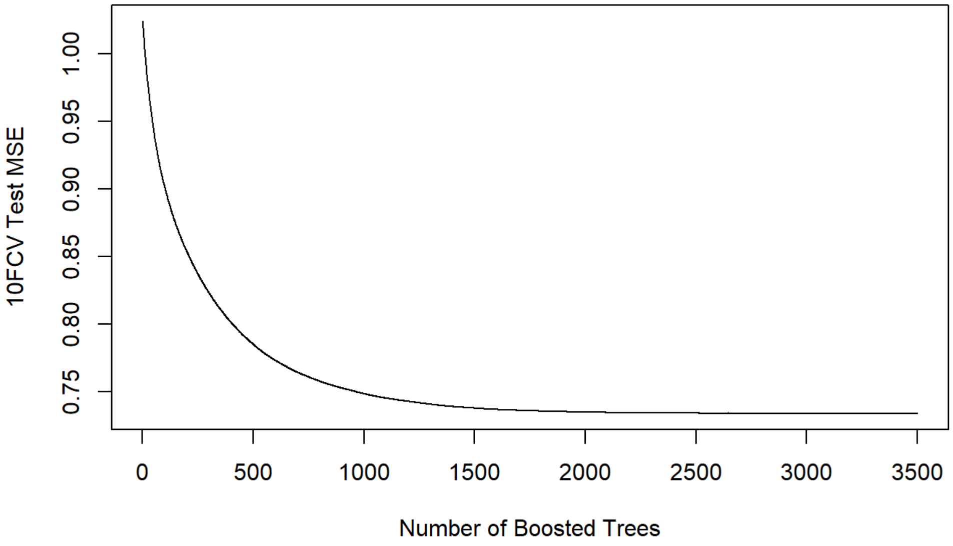


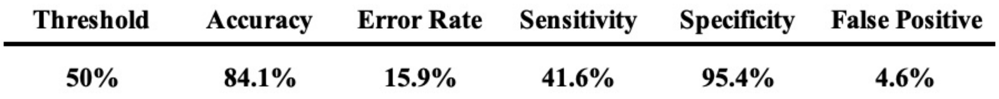
Model Revalidation

Boosted Trees using the whole dataset



```
bank_acct_smp$bank_account <- as.numeric(bank_acct_smp$bank_account)
boost.bank <- gbm(bank_account~.,
                  data = bank_acct_smp,
                  distribution = "bernoulli",
                  shrinkage = 0.01,
                  cv.folds = 10,
                  n.trees = 3500,
                  interaction.depth = 1)
boost.bank
```





Threshold	Accuracy	Error Rate	Sensitivity	Specificity	False Positive
50%	84.1%	15.9%	41.6%	95.4%	4.6%