**Logo, company name

Description automatically generated**

**DSA211-Statistical Learning with R-G1**

**Group Project Part 2\_Group 6**

**Instructor: Goh Jing Rong**

**Group Members:**

|  |
| --- |
| \*Censored\* |
| \*Censored\* |
| **Chong Gao Zhe** |
| \*Censored\* |
| \*Censored\* |

Table of Contents

[**1. Introduction** 3](#_Toc118558600)

[**1.1 Objective** 3](#_Toc118558601)

[**1.2 Selection *Criteria* / Variance bias trade off** 3](#_Toc118558602)

[**2. Method** 3](#_Toc118558603)

[**2.1.1 Best Subset Selection with Cross-Validation Approach** 3](#_Toc118558604)

[**2.1.2 Model 7 from Part 1 of Group Project with cross-validation approach** 4](#_Toc118558605)

[**2.2 Ridge Regression Approach** 4](#_Toc118558606)

[**2.2.1 Ridge Regression on (Balance ~ . + I(Income^2))** 4](#_Toc118558607)

[**2.3 Lasso Approach** 5](#_Toc118558608)

[**2.3.1 Lasso on (Balance ~ . + I(Income^2))** 5](#_Toc118558609)

[**2.3.2 Testing of Interaction Terms after Lasso** 5](#_Toc118558610)

[**2.3.3 Balance~poly(Income,2,raw = TRUE)\*Gender + Rating\*Cards + Rating\*Ethnicity + Cards\*Gender + Age + Education\*Gender + Married + Gender\*Ethnicity** 6](#_Toc118558611)

[**2.3.4 Balance~poly(Income,2,raw = TRUE)\*Gender + Rating\*Cards + Rating\*Ethnicity + Cards\*Gender + Education\*Gender + Gender\*Ethnicity** 6](#_Toc118558612)

[**2.4 Random Forest Regression Approach** 6](#_Toc118558613)

[**2.4.1 Random Forest with (Balance ~ . + I(Income^2))** 7](#_Toc118558614)

[**2.4.2 Random Forest with (Balance ~ . + I(Income^2) - Rating)** 7](#_Toc118558615)

[**3. Best Predictive Model (Balance~poly(Income,2,raw = TRUE)\*Gender + Rating\*Cards + Rating\*Ethnicity + Cards\*Gender + Education\*Gender + Gender\*Ethnicity)** 7](#_Toc118558616)

[**3.1 Best Predictive Model Assumptions Justification** 8](#_Toc118558617)

[**4. Conclusion** 8](#_Toc118558618)

[**5. Appendix** 9](#_Toc118558619)

# **1. Introduction**

## **1.1 Objective**

The objective of the report is to explore and construct a predictive model for bank account balance (*Balance*) of an individual using the following parameters: monthly salary (*Income*), credit limit (*Limit*), credit rating score (*Rating*), number of credit cards (*Cards*), age (*Age*), years of education (*Education*), gender (*Gender*), marital status (*Married*), and ethnic group (*Ethnicity*).

## **1.2 Selection *Criteria* / Variance bias trade off**

To generalise the model beyond the training data, it is essential that the prediction model does not have high variance or high bias indicating that the training data is either overfitted or underfitted in the model. To ensure that the model performs well on a test data set, it is necessary to achieve reasonably low variance and low bias due to the variance-bias trade off. Therefore, the model with lowest mean-squared error will be selected.

# **2. Method**

We have established that Income has a quadratic relationship with Balance in Part 1 of our Group Project. Therefore, the different methods we would explore will take as a variable in our data.

### **2.1.1 Best Subset Selection with Cross-Validation Approach**

Given the quadratic relationship between Income and Balance that we have found in part 1 of the group report, we employ the *dplyr* function to mutate (add) an additional column of to the Bank2022P.csv data.

We then run the Best Subset Selection to find the best equation that the machine will evaluate. We did a 10-fold cross-validation method. From this method, we found that the best model from the Best Subset Selection method consists of the following predictors: *Rating*, *Cards*, *GenderMale* and , for which k = 4.

**Test MSE**: 375,795.7

Moreover, we can also see a similarity between this model and Model 4 in part 1 of the group report. Both models have the same predictors except that the Best Subset Selection indicated *GenderMale* as opposed to *Gender* considered by the multi-linear regression model in Model 4 of the first project.

Given the same methodology of identifying causality in the model for the training set before building out by testing it against the testing set to gauge its predictive capability, we would yield similar predictors to Models 5, 6 and 7 from part 1 group report, if we were to consider interaction terms under the Best Subset Selection method.

### **2.1.2 Model 7 from Part 1 of Group Project with cross-validation approach**

Here, we investigate the predictive capability of part 1 of the group report. We chose Model 7 to investigate because that is our best model in the previous report. Also, because the model considered interaction terms (\**Gender* and *Cards*\**Gender*), which was not accounted for in our Best Subset Selection model with cross-validation approach.

**Test MSE:** 211,234

Based on the test MSE value, it turns out that the model that considered interaction terms is a better fit. However, we will also consider other regression approaches in the subsequent sections.

## **2.2 Ridge Regression Approach**

It would still be preferable for us to further reduce the MSE through other regression methods that improves the predictive capability. Hence, we use ridge regression as it enables us to reduce the MSE through increasing Lambda (λ), shrinking the coefficients to achieve lower variance and emphasis on the regressors. Lowering emphasis on the regressors helps us to reduce variance albeit it increases the biasedness of our models.

### **2.2.1 Ridge Regression on (Balance ~ . + I(Income^2))**

Given that we had established earlier that could be a good predictor of Balance, we decided to conduct a Ridge Regression on the model as well. To do so, we created a matrix consisting of the independent variables and their corresponding values. After which, we sampled 80% of the data into the training set and 20% into the test set. Next, we used the *glmnet* function to create a model based on the training set and used *cv.glmnet* function on the training set to obtain the best λ through 10-fold cross validation. We found that the λ that provided the lowest MSE was 339.4212. Thus, using the lowest λ and the model created earlier, we found the MSE to be 458,014.3 when applied on the test set. Reapplying all the data to refit the model, we found the coefficients to be as such: Balance = 586.10463310 + 0.82809011 Income + 0.01697751 Limit + 0.31660454 Rating - 47.45859296 Cards + 0.11498616 Age - 19.75151887 Education - 398.60574960 GenderFemale - 68.01367780 MarriedYes. As expected, all coefficients of the independent variables are non-zeros, which could have led to overfitting hence, a higher test MSE.

**Test MSE:** 458,014.3

## **2.3 Lasso Approach**

While ridge regression shrinks the coefficient estimates of our predictors, it included all the predictors in our data which resulted in overfitting. Hence, we have the need to perform another type of variable selection just like the best subset selection to obtain a parsimonious model. This led us to explore the lasso approach, which also shrinks the coefficient estimates, but could potentially shrink some estimates towards zero. After which, we would do a multiple regression model to explore interaction terms of the remaining variables by including those that are significant to the model, i.e. reduce MSE of our model.

### **2.3.1 Lasso on (Balance ~ . + I(Income^2))**

Using the library *glmnet*, we first create a *model.matrix* of (*Balance*~.+I()) and a Balance vector. We then split the data into 80% training data and 20% test data, sampled randomly. Subsequently, we performed 10-fold cross validation on the lasso training data fits to obtain the best(min) value of tuning parameter λ = 6.071055. Thereafter, we use this λ to obtain our MSE and variables whose coefficient estimates are not zero. The predictors we obtained are *Income, , Rating, Cards, Age, Education, Gender, Married and Ethnicity*.

**Test MSE:** 358,778.1

### **2.3.2 Testing of Interaction Terms after Lasso**

By regressing *Balance* on *(poly(Income,2,raw = TRUE) + Rating + Cards + Age + Education + Gender + Married + Ethnicity)^2*, we test for all interaction terms among these predictors and choose the ones that are significant to our model based on p-value<0.05. From the R output, we see that *Income:Gender, Rating:Cards, Rating:Ethnicity, Cards:Gender, Education:Gender and Gender:Ethnicity* are significant.

### **2.3.3 Balance~poly(Income,2,raw = TRUE)\*Gender + Rating\*Cards + Rating\*Ethnicity + Cards\*Gender + Age + Education\*Gender + Married + Gender\*Ethnicity**

By adding all the significant interaction terms found previously, we then perform a 10-fold cross validation on the multiple regression model to obtain the test MSE. However, we note that *Age* and *Married* are not significant predictors as they have high p-values and are not included in any interaction terms, which indicates that we can consider dropping them if it decreases the MSE and improves prediction accuracy.

**Test MSE**: 207,160.2

### **2.3.4 Balance~poly(Income,2,raw = TRUE)\*Gender + Rating\*Cards + Rating\*Ethnicity + Cards\*Gender + Education\*Gender + Gender\*Ethnicity**

For this model, we drop the predictors *Age* and *Married* and do the multiple regression again. This time, we obtain a lower test MSE which suggests that the two predictors are indeed insignificant to the model and by dropping them, we can improve the predictive accuracy of the model. This MSE is the lowest one so far, which indicates that this approach could be the most predictive model.

**Test MSE:** 206,612.9

## **2.4 Random Forest Regression Approach**

The next approach we explored was the Random Forest Regression Approach. We wanted to see if the use of a Random Forest Regression would provide a better way of predicting values for *Balance* given our data set. We went straight into this rather than trying out tree regression or pruned tree regression because we would like to find a model that has a good predictive accuracy. Tree regression and pruned tree regression are not comparable to other supervised learning methods like the ones we have done like multiple linear regression, ridge regression and lasso in this aspect. Hence, we deploy a random forest algorithm to use the qualities features of multiple decision trees to improve predictive accuracy.

### **2.4.1 Random Forest with (Balance ~ . + I(Income^2))**

Just like with other approaches, we split the data into 80% training and 20% testing. Since this is a regression, we use mtry = p/3, which means mtry rounds off to 3 as we have 10 predictors, including , as well as using ntree = 1000. We then perform the random forest on the training set and use this to evaluate our test MSE by comparing predicted and observed *Balance* for the test set.

**Test MSE:** 281,432.2

### **2.4.2 Random Forest with (Balance ~ . + I(Income^2) - Rating)**

Note from the importance measure plot in 2.4.1 appendix that the predictor *Limit* is more important than *Rating* based on %IncMSE. From the correlation table of Group Project Part 1, we know that *Limit* and *Rating* are multicollinear with a high correlation. Therefore, a good model would not include both of these predictors. Since *Limit* is more important than *Rating* from the plot, we now drop *Rating* and indeed obtain a lower test MSE. This would be our final model using the Random Forest Regression approach.

**Test MSE:** 260,594.9

However, we observe that this test MSE is still higher than the model we get from Lasso Approach followed by multiple regression in 2.3.4. This could be due to the fact that random forests pose a major challenge, which is they cannot extrapolate outside unseen data. When we train our random forests model on the training set, it will be difficult for the model to predict the testing set if the values lie beyond the boundaries of the training set. On the contrary, for multiple regression models, the trained model can easily extrapolate by identifying patterns in the model to improve prediction of the testing set, which leads to a lower MSE.

# **3. Best Predictive Model (Balance~poly(Income,2,raw = TRUE)\*Gender + Rating\*Cards + Rating\*Ethnicity + Cards\*Gender + Education\*Gender + Gender\*Ethnicity)**

Out of all 4 approaches we have explored, the best predictive model to predict Balance was obtained from the lasso approach in 2.3.4 as it has the lowest MSE, which indicates high predictive accuracy. The coefficient estimates are shown in Appendix Section 3.

## **3.1 Best Predictive Model Assumptions Justification**

Since our best predictive model makes use of multiple regression, we would have to ensure the assumptions for the regression hold. To check for error constant variance assumption of the model, we performed residual analyses of our best predictive model with its residuals against fitted values (*Balance*), as well as all independent variables involved. From the residual plots in the appendix, the constant variance assumption is not violated as their errors are fairly distributed across the values of our variables.

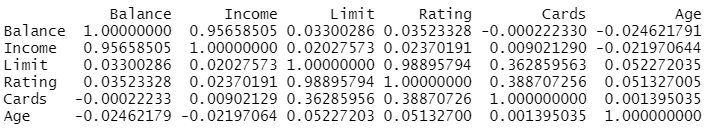
Additionally, to check for normal assumptions for our predictive model, we fitted the residuals of the regression model to a normal distribution and plotted it. From the Q-Q and P-P plots, the residuals generally follow a normal distribution and does not make us doubt our model deviates from normality.

# **4. Conclusion**

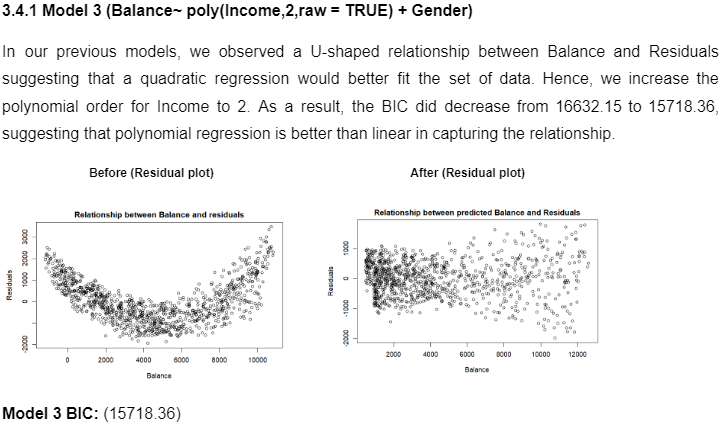
The lasso approach followed by multiple regression was the most optimum predictive model based on test MSE, which is ***Balance~poly(Income,2,raw = TRUE)\*Gender + Rating\*Cards + Rating\*Ethnicity + Cards\*Gender + Education\*Gender + Gender\*Ethnicity***. As discussed in section 2.3, we managed to reduce the variance significantly which also performs variable selection, leading to a parsimonious model by regularising the data. After this, we included significant interaction terms for non-zero coefficient variables in the model and regressed it to get better predictors to predict *Balance*. As a result, the MSE obtained for this would be improved over simply regressing using the best subset selection with interaction terms. In contrast, the ridge regression model shrunk the coefficients of the regressor due to which it over-fitted the training data and had higher prediction error than the Lasso. Hence, after comparing the models obtained after carrying out the 4 methods, we can conclude that the Lasso model (2.3.4) is the best model for predicting the bank account balance.

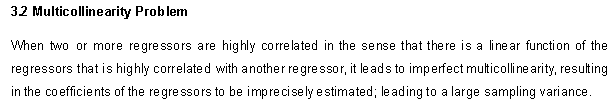
# **5. Appendix**

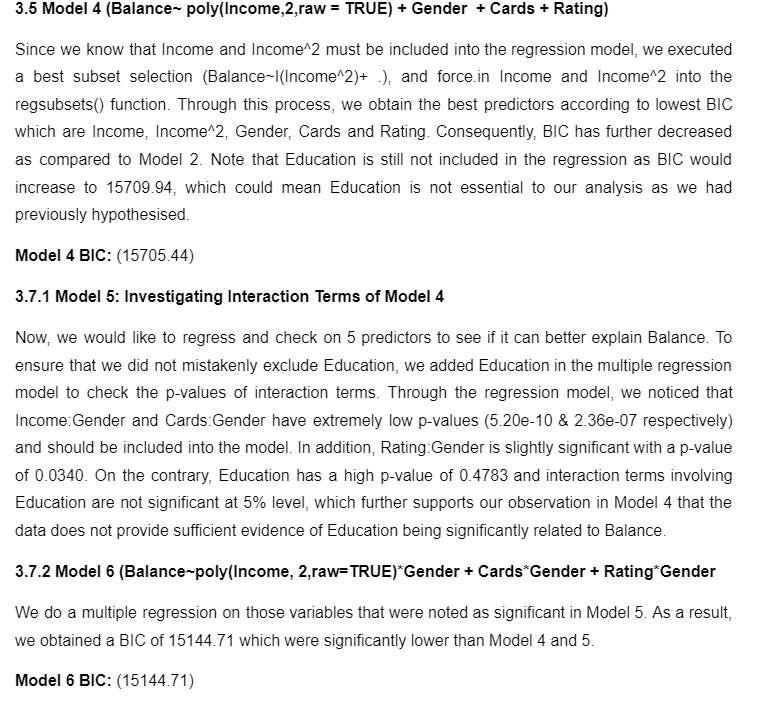
**Group Project Part 1 Reference**

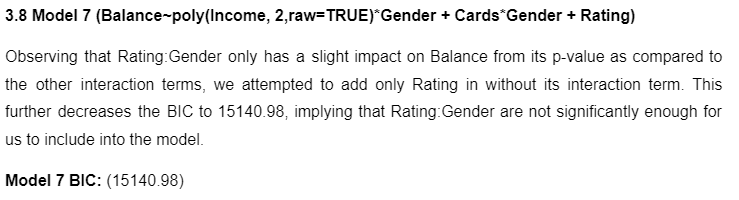
**Correlation**

**Quadratic relationship of Income with Balance**

**Part 1 Model References**

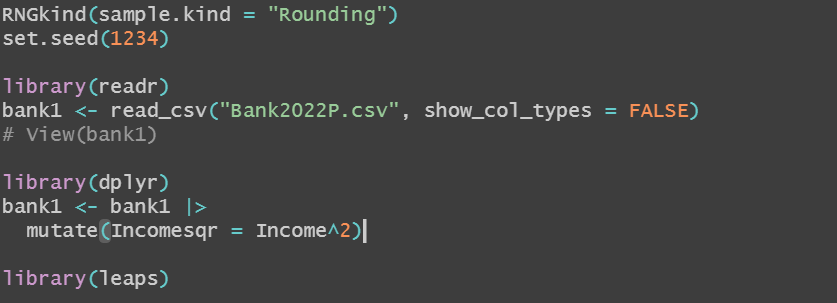
****

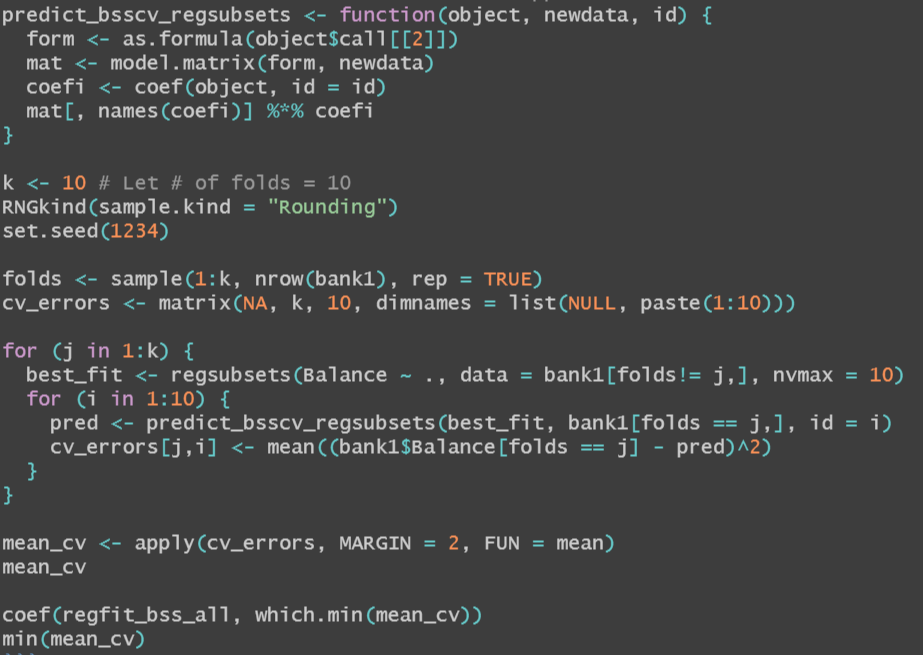
****

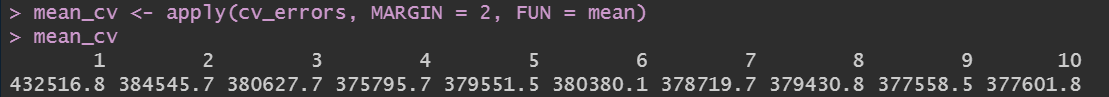
****

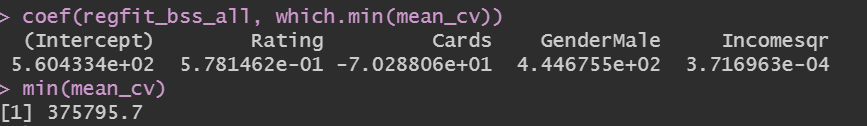
**Best Subset Selection with Cross-Validation Approach**

**2.1.1**

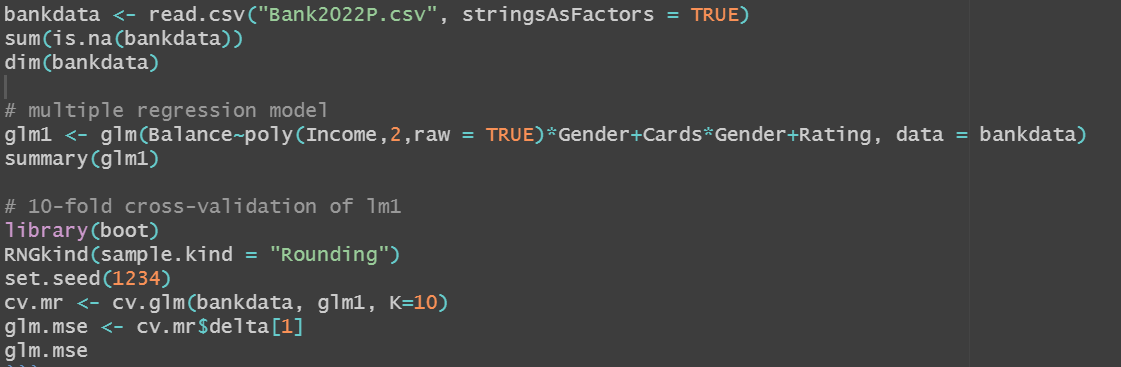
****

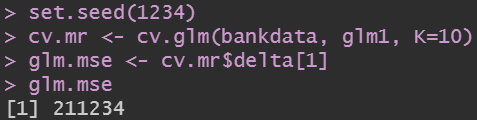
****

****

****

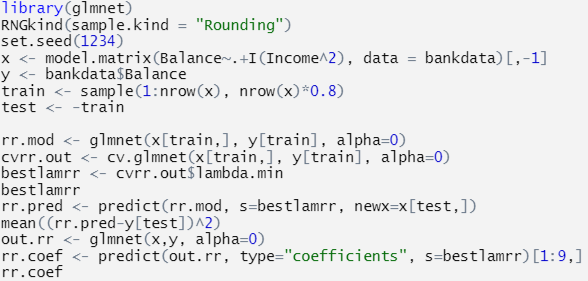
**2.1.2**

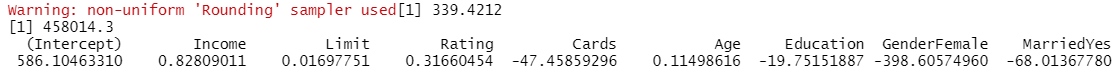
****

****

**Ridge Regression**

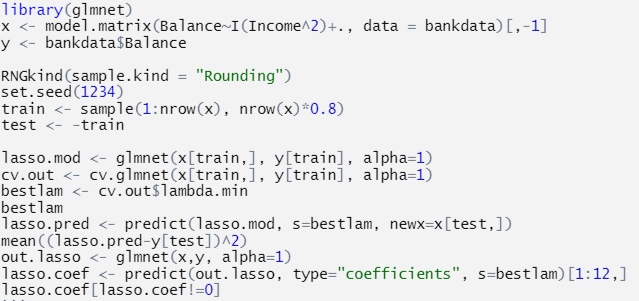
**2.2**

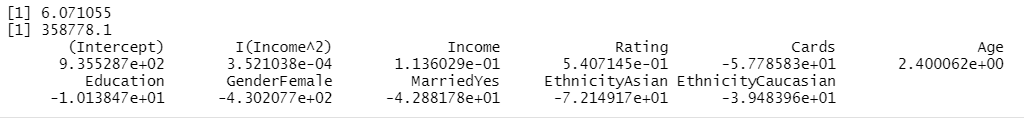
****

****

**Lasso**

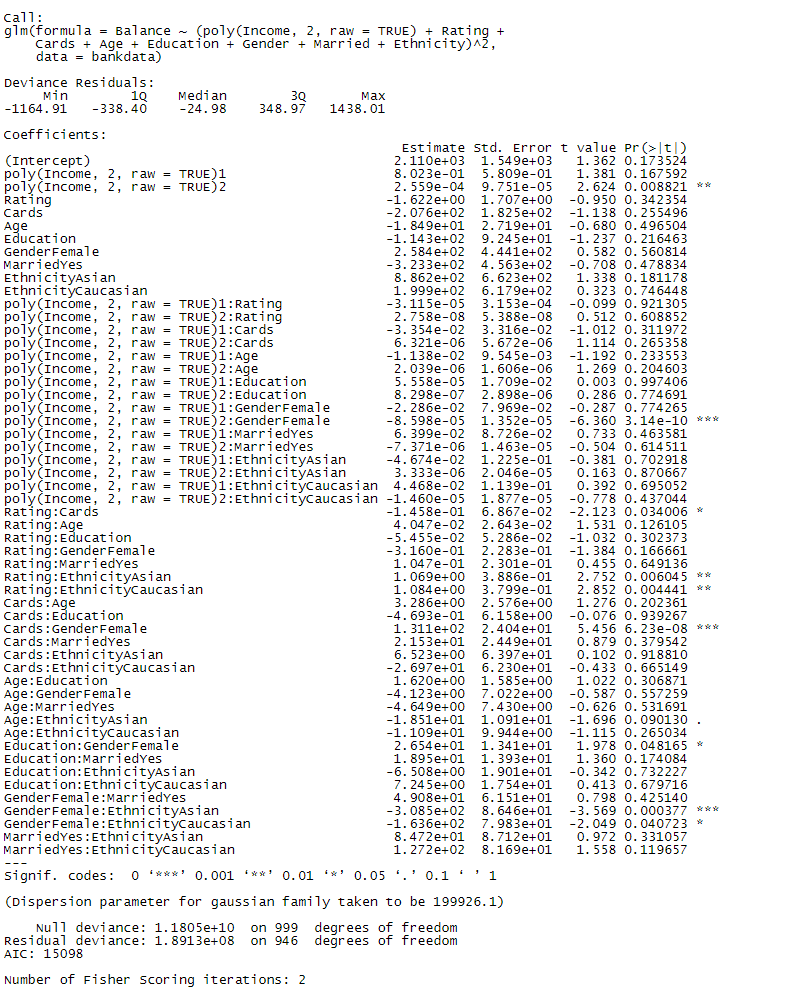
**2.3.1**

****

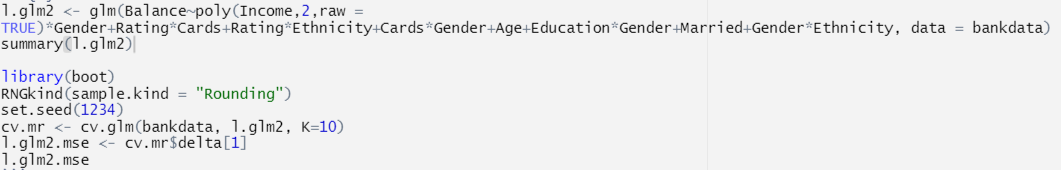
****

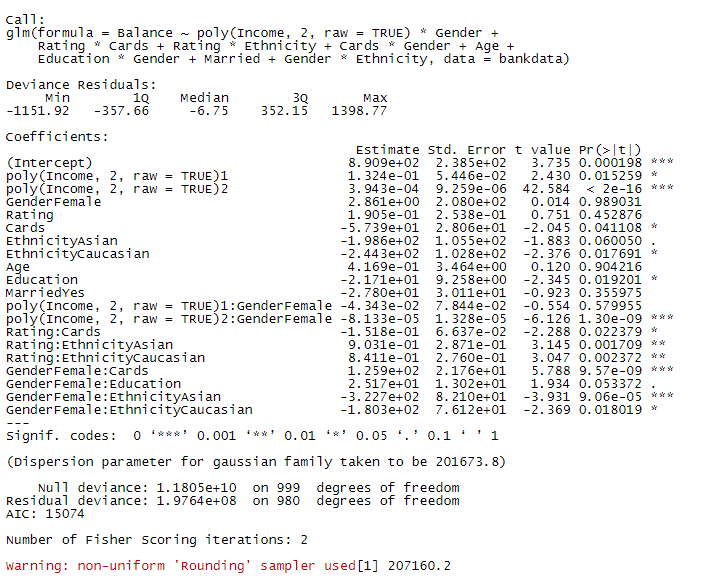
**2.3.2**

****

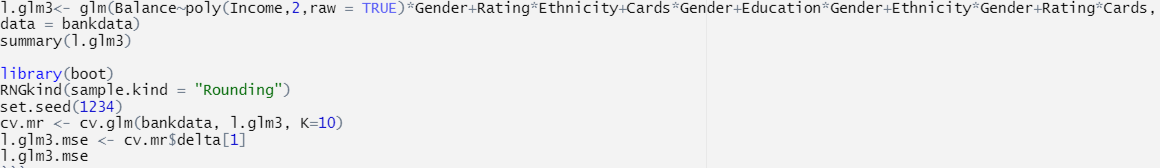
****

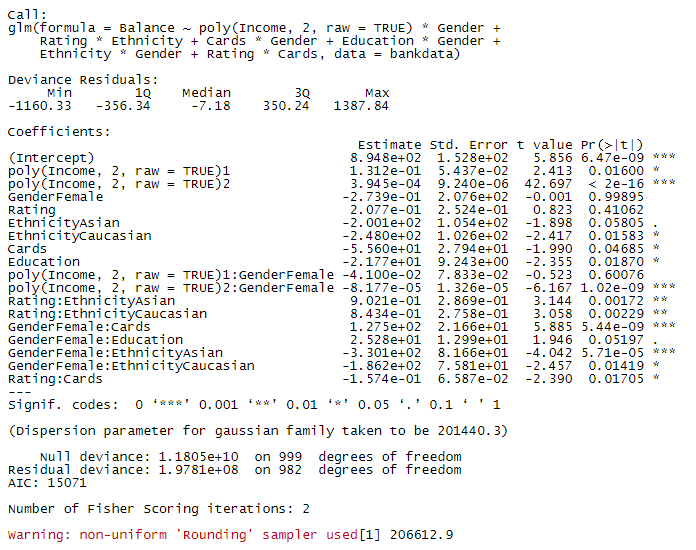
**2.3.3**

****

****

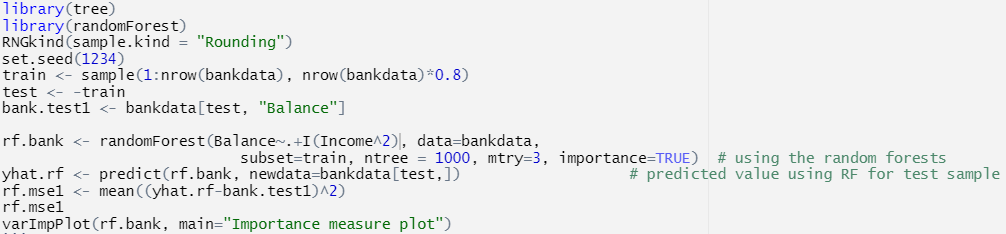
**2.3.4**

****

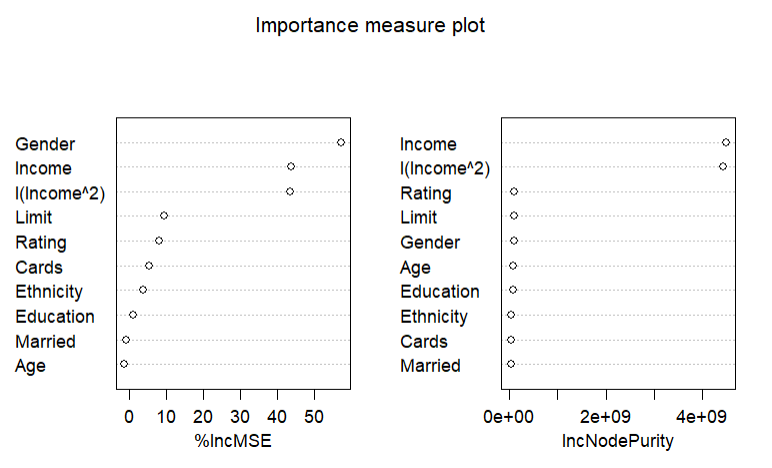
****

**Random Forest Regression**

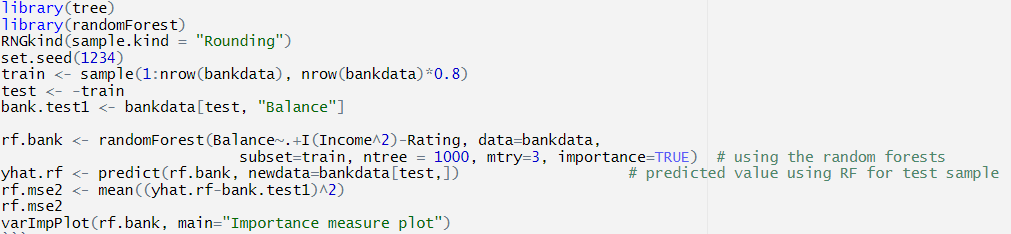
**2.4.1**

****

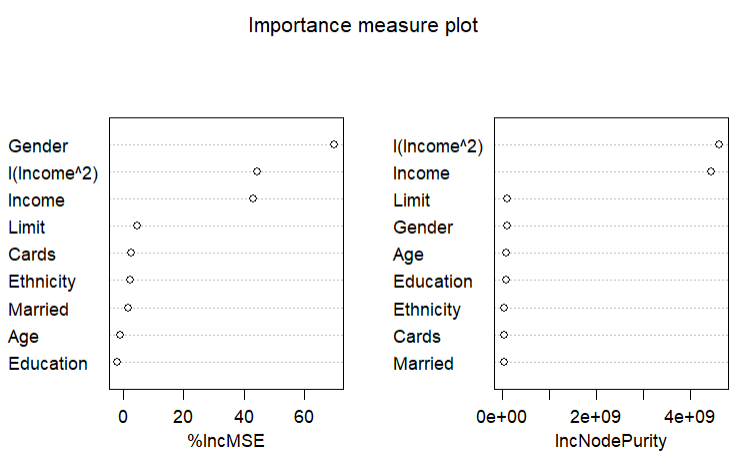
****

****

**2.4.2**

****

****

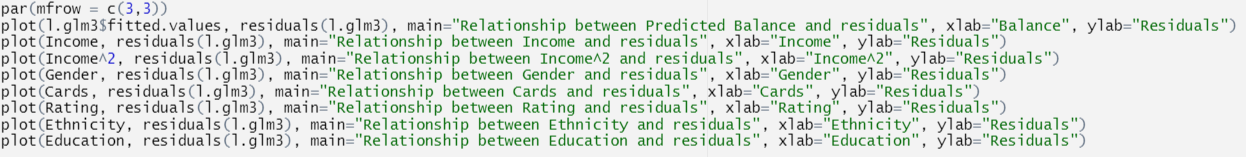
****

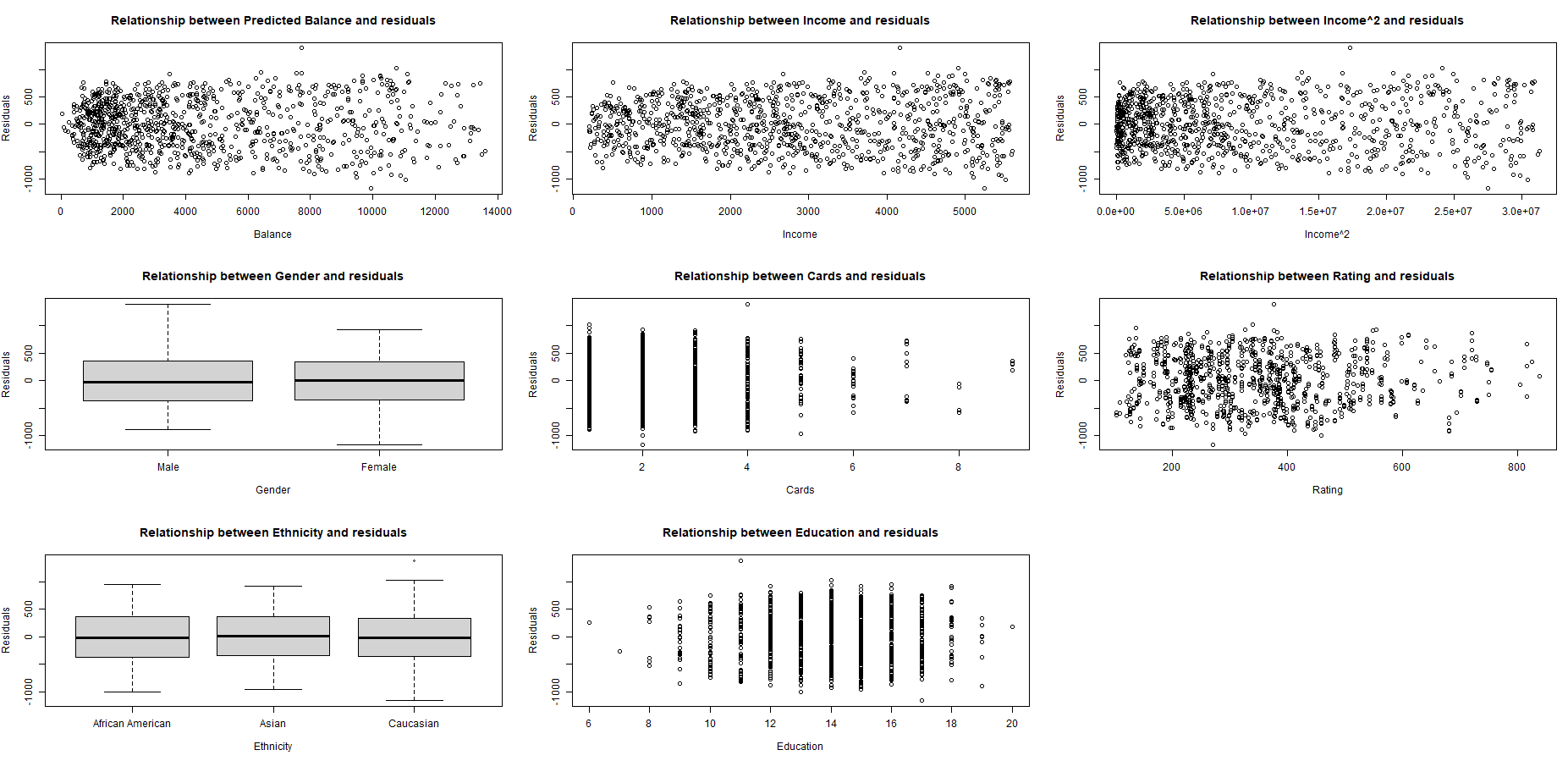
**3. Best Predictive Model’s Assumptions Justification**

Text

Description automatically generated

**Residual Analyses**

****



**Normality Assumption Justification**

****

