**Credit Card Approval Prediction**

**-Business Risk Analysis with R**

**Content**

**[Abstract 2](#_Toc29606)**

**[Data Source 4](#_Toc4720)**

**[Data Understanding 5](#_Toc9374)**

**[Data Visualization 7](#_Toc14431)**

[1. Gender distribution 7](#_Toc7278)

[2.Car ownership 7](#_Toc22457)

[3.House ownership 7](#_Toc24353)

[4.Distribution of number of children 8](#_Toc3585)

[5.Distribution of Income 8](#_Toc8005)

[6.Distribution of income type 9](#_Toc2625)

[7.Distribution of level of education 9](#_Toc17322)

[8.Distribution of age 10](#_Toc25667)

[9.Distribution of work-experience 11](#_Toc10282)

[10.Understand the breakage of good and bad clients 11](#_Toc1746)

**[Data processing 12](#_Toc22891)**

[1.Missing data 12](#_Toc3294)

[2.challenge and solution: 12](#_Toc19454)

[Data modeling-line regression model and log model 14](#_Toc12451)

[Data modeling-classification model  17](#_Toc20819)

[Inferences and Conclusion 17](#_Toc9557)

[1.Inferences -Regression and Hypothesis test 17](#_Toc24910)

[2.Conclusion 18](#_Toc14040)

[(1)Product design and optimization ,marketing strategy 18](#_Toc17797)

[(2)Risk management strategy 18](#_Toc10384)

[(3) Balance strategy 18](#_Toc23406)

# 

# **Abstract**

A Credit card is an empowering financial tool that helps individuals take up monetary and financial decisions by increasing their liquidity. This is seen as a luxury by many, who use credit cards for transactions to better their financial planning in the near, and long term. However, credit cards aren’t accessible to everyone. Financial institutions have devised a mechanism to determine if an individual is credit worthy or not. What may seem like a simple yes/no decision is indeed complex.

The process of accepting or declining an individual for a credit card is facilitated by the individual’s credit score. If the credit score of an individual is above a certain threshold, which varies between financial institutions, the individual is then accepted into the program. Credit scores are based on historical data. This data may include various factors such as repayments, defaults, employment status, expected annual income, assets owned, and many more. These factors altogether decide the credit worthiness of an individual.

There have been many models that predict the creditworthiness of an individual, but due to untimely factors such as fluctuations in the economy, inflation and certain other macroeconomic factors, these models go obsolete.

Logistic and linear regression are preferred models, logistic more so, as it deals with binary classification tasks, and we have attempted a classification tree as well. We hope to use the various techniques we have at our disposal to better understand and predict whether the client is ‘good’ or ‘bad’ based on historical data.

# **Data Source**

We have sourced our data from Kaggle, which houses numerous databases for users to perform basic data analysis and modelling. We opted to go for the Credit Card Approval Rating database on Kaggle, this is an open-source dataset.

Please use the following the hyperlink to access the source website for our dataset - <https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction>

# **Data Understanding**

For understanding our dataset, we have performed basic functions on R to give us important information regarding our dataset. We have two datasets, application, and credit.

The dimensions of our application dataset are –

Text

Description automatically generated

We can see that our dataset has 18 columns, with 438557 rows.

The dimensions of credit dataset are –

Text

Description automatically generated

The credit dataset has 4 columns, with 1048575 rows. This large number is because we have multiple entries for the column ‘ID’.

We then proceed to solve for the unbalancing, which is to aggregate the ‘Status’ for each ID and then merge the two datasets. On doing so, we get the following dimensions for our merged dataset –

A picture containing text

Description automatically generated

There are several NA’s that arise due to the merging of the two datasets, the NA count is –

Text

Description automatically generated with medium confidence

We then proceed to remove NA’s with na.omit() function in R. We then get –

Text

Description automatically generated with low confidence

The dataset now contains 21 columns with 25134 rows.

Let us now generate a random sample of our dataset by using the sample() function in R –

A picture containing graphical user interface

Description automatically generated

Using the head() function, we shall now generate the first six observations as part of our dataset – Graphical user interface

Description automatically generated

Let us also look at the structure of our dataset, this can be achieved by using the str() function in R.

We have mainly numerical and character variables in our dataset.

Text

Description automatically generated

# **Data Visualization**

First, we proceed to understand the demographics of the clients in our dataset, this involves understanding the gender breakup of clients, understanding if clients own certain assets such as realty, a vehicle, mobile, work-phones etc. Since these are binary in nature, we have used pie-charts to represent them.

1. Gender distribution

We first create a frequency distribution table to get the percentage breakup of the gender of our clients.

Chart, pie chart

Description automatically generatedWe can see that majority of clients in our dataset are female, accounting for 62.2% of the total clientele. Males occupy the other 37.8%.

Text

Description automatically generated

2.Car ownership

Creating a similar frequency distribution table yields the following result –

Text

Description automatically generatedAs we can see, 58.2% of the observations in the dataset do not own a car, with 41.8% accounting to clients who own a car.

Chart, pie chart

Description automatically generated

3.House ownership

The FLAG\_OWN\_REALTY variable gives us an idea about whether the client in question owns a property/house. A frequency distribution of the same yields -

We can see here that majority of clients own a house/property, with 65.5% of the clientele owning a house, while the other 34.5% do not seem to own a house/property

Chart, pie chart

Description automatically generated

Text

Description automatically generated

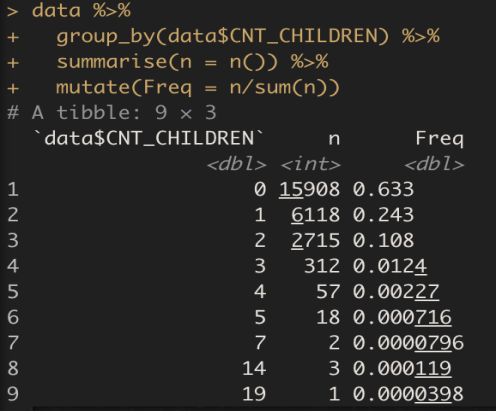
Moving onto some other plots, we now have some bar plots to explain the distribution of continuous variables.

4.Distribution of number of children

A quick frequency distribution table for the number of children for clients is as follows.  
Here, we can see that majority of the clients in our dataset do not have children, account for nearly 63.3%. Most of our clients are distributed between having 0-2 children, while we have several outliers including clients with as many as 12-19 children in their family.

Chart, histogram

Description automatically generated



5.Distribution of Income

This does seem to be the most important variable in our dataset, as income level by common sense will have an impact on the credit worthiness of the client. The understanding is that the higher the income level, the more the affinity to accumulating credit and more the chances of repaying, thereby being a positive factor in our analysis. We have tried to show this information via a histogram –

Chart

Description automatically generated

The Income distribution seems to be right skewed, mean of total income is greater than median of the total income, thereby giving it a right tail. We do appear to have some outliers as summary statistics for income gives the highest value of $6750000.

6.Distribution of income type

There are different occupations and income types in our dataset, and the next graph is to try and understand what the distribution of type of income is in our dataset.

A frequency distribution shows the following –

Text

Description automatically generatedChart, bar chart

Description automatically generated

As we can see, the majority income type comes from the working-class accounting for a massive 62.2% of our dataset. Other important income types are commercial associate and state servant, each accounting for 28.1% and 9.7% respectively.

7.Distribution of level of education

Education levels also play a huge role in determining your creditworthiness, it is assumed that the higher your education level, the better suited you are to credit and repayment. So, it makes sense to try and understand the distribution of education level in our dataset. A quick frequency distribution gives us the following .

Here, we can see that majority of our database has received up till Secondary and Secondary special education, amounting for nearly 66.9% of our dataset. Other prominent education type featuring is higher education.

Chart, bar chart

Description automatically generated

Text

Description automatically generated

8.Distribution of age

The penultimate bar-plot/histogram distribution we are interested in is to try and see what the distribution of age in our dataset is.

Chart

Description automatically generated with low confidence

Through this distribution, we can say that our dataset has a good mix of clients from different age groups, the distribution does seem to be normal with a few misspecifications due to the large number of datapoints, but the mean age is almost equal to median age. This could be an important factor in our analysis as age may be a determining variable, given that you are expected to have a steady income stream during your middle years for you to qualify for credit cards and your credit worthiness increases as you earn more. During your early years, you do not qualify for credit cards as you do not have a credit history, whereas later while you do have a lot of liquidity, your affinity to spend more goes down.

9.Distribution of work-experience

The last histogram we shall be looking at is the distribution of work-experience in our dataset. The histogram looks as follows –

Chart, histogram

Description automatically generated

As we can see, the graph is right skewed, with majority of clients in our dataset are newly employed and are getting into the groove of their work-life. This is an important distribution as it states that most of the clients are in that phase of life where they wish to have secondary disposable income readily available to them to fulfil their urgent liquidity needs.

10.Understand the breakage of good and bad clients

Furthermore, we have several side-by-side bar plots to understand the breakage of good and bad clients across various demographics, such as age, employment type, education type among many more. Please find the graphs for the same below –

Chart, waterfall chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, histogram

Description automatically generatedChart, bar chart

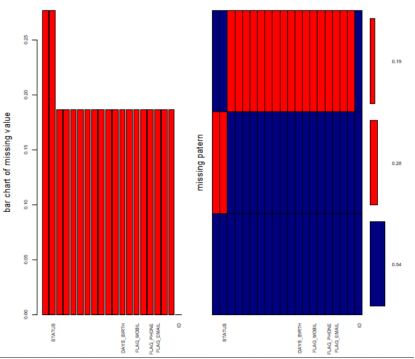
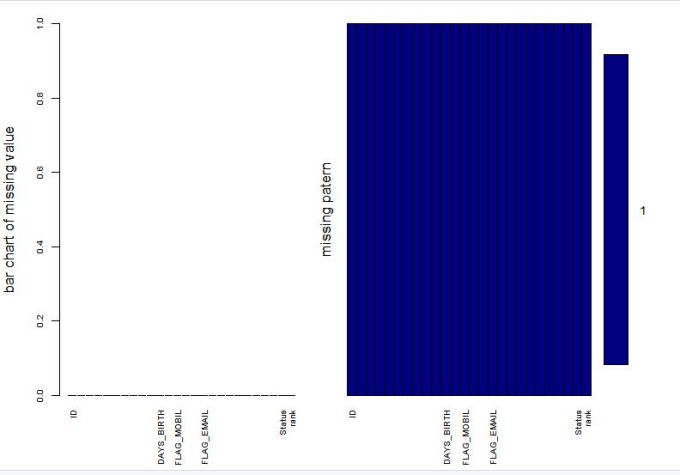
Description automatically generated

Chart, bar chart

Description automatically generated

# **Data processing**

1.Missing data

Missing data-before cleaning data Missing data-after cleaning data

Since there are two data tables in the data set, one of which is the basic information about the applicant, from which we can obtain information about the applicant's gender, age, education, income, property status, marital status, family members, working years, etc. , which can help banks obtain the basic information of credit card applicants, so as to construct the basic image of credit card applicants.

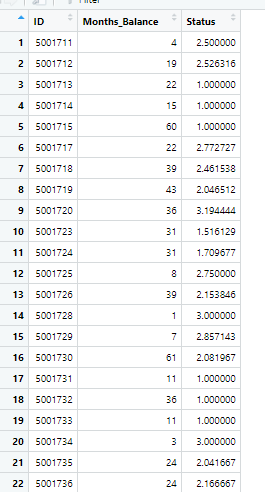
And this information provides information convenience for banks to understand customer needs and attract potential customers. This will also help banks formulate more targeted credit card advertisements and promotion strategies, and will also help optimize credit card products.

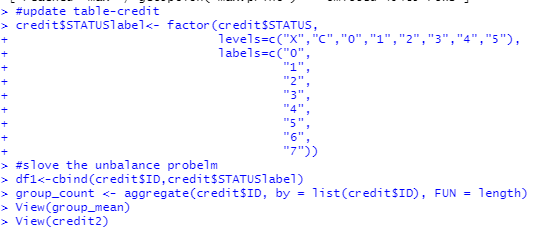
Another data table has information about credit card holders, from which we can see the statistical information of credit card holders' repayment bills, which includes the statistics of the month and the basic status of each month.

2.challenge and solution:

Challenge 1 :

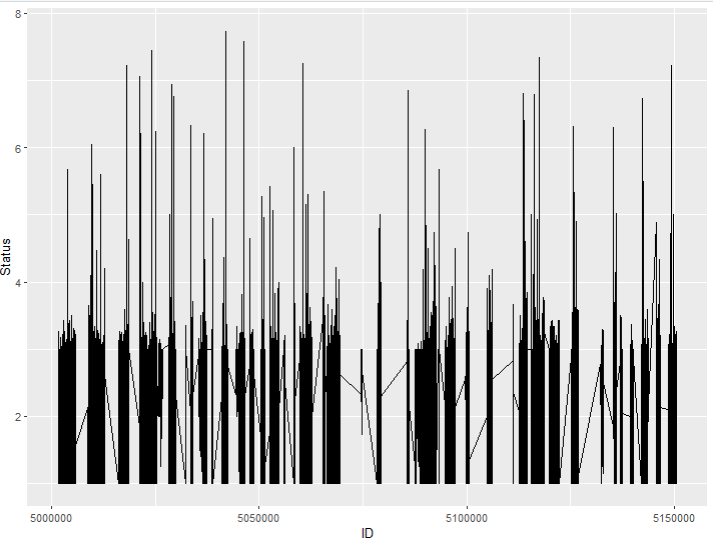
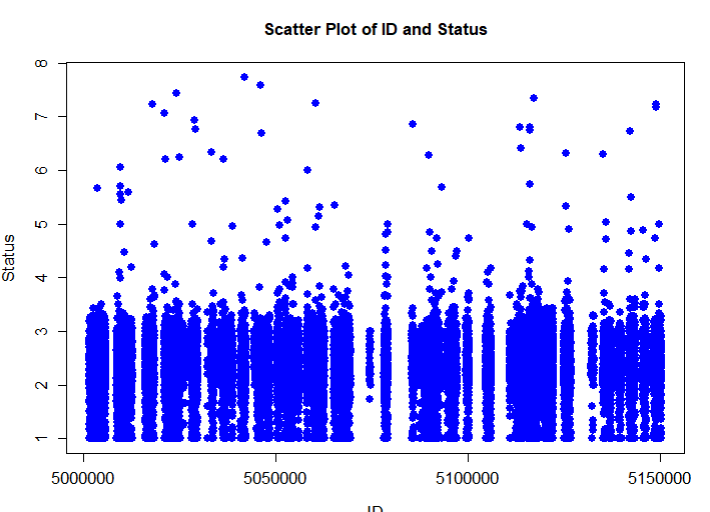
Processing the credit table, the first challenge: each of the customers, due to the existence of billing status information for multiple months, there may be multiple rows of data, so there are multiple data of the same customer ID. This is bad news for analyzing individual customers as well as for joint analysis with the first table.

In order to solve this problem, the first step is to assign "X", "0", "1" in the table to values (from 1 to 7, the larger the value, the longer time for repaying the bill,thus the worse the credit status, and the higher risk), the second step is to use the ID as the grouping basis to calculate the value of the average bill status (the total bill status value divided by the number of months, the higher the value, the higher the risk of default). Then we can build a new data table, including the unique ID, month, and average status.



Challenge 2:

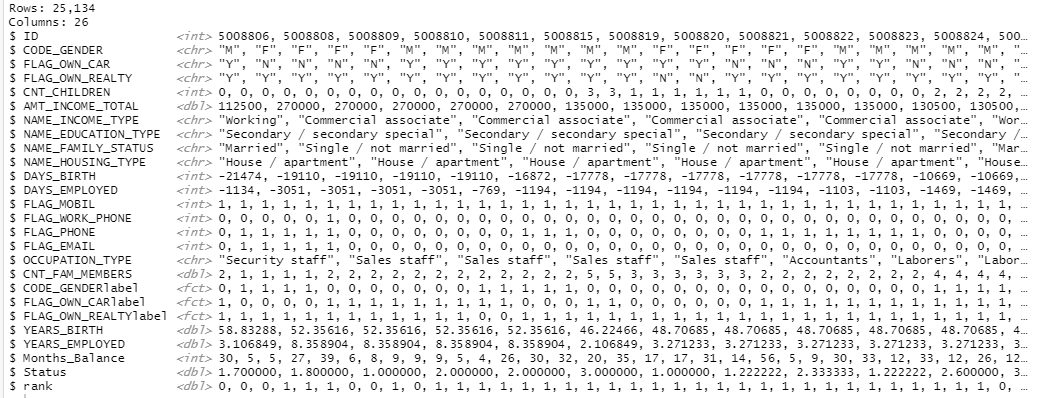
The second challenge: how to set the risk threshold. On the one hand, more customers can bring more operating income and profits, but on the other hand, customers with high default risk will also increase the default risk and corresponding operating costs. How to balance the relationship between the two and find the best data point is very important. By visualizing the data, we found that the average distribution of most billing statuses is between 1-3. Here, we choose 2 as the risk threshold (of course, you can also choose other thresholds. The higher the risk sensitivity, the lower the risk threshold) , so about half of the customers will be judged to be at risk of default. For high-risk customers, the bank can send emails and other methods to inform the bill status; for low-risk customers, the bank can devote more energy to maintain customer relationships, appropriately increase the authorized amount, stimulate consumption, and increase profits.





Challenge 3:

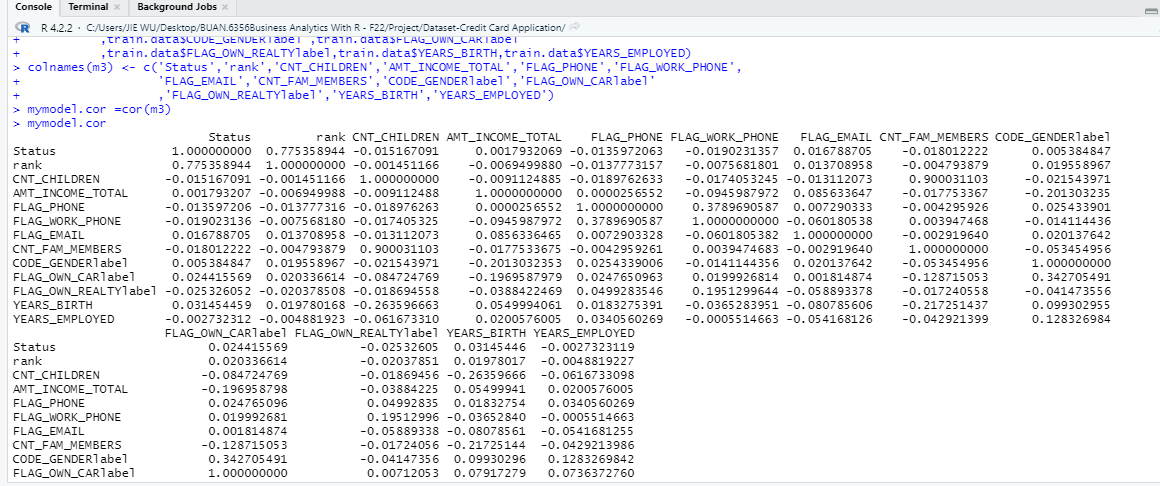
Another challenge in dealing with monolithic datasets is how to deal with data imbalance. We have two data sets. Although the two tables can be connected by ID, there are a lot of unmatched data and many missing values. Due to the large data set, by visualizing the status of missing values, we set standards to remove missing values and build new tables for analysis.

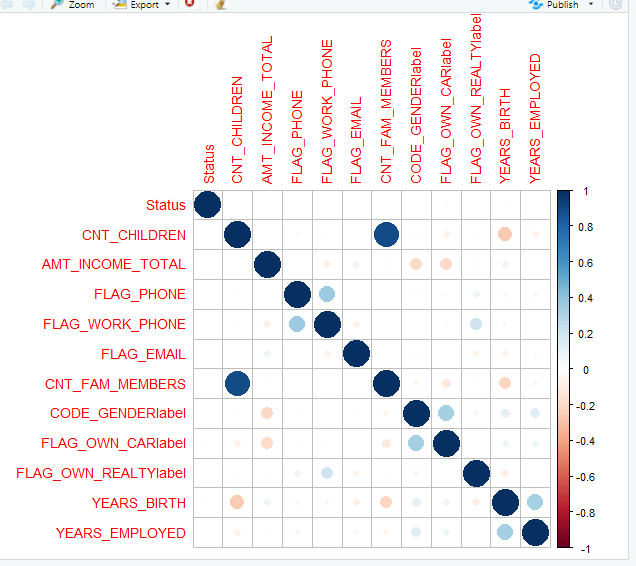


# **Data modeling-line regression model and log model**

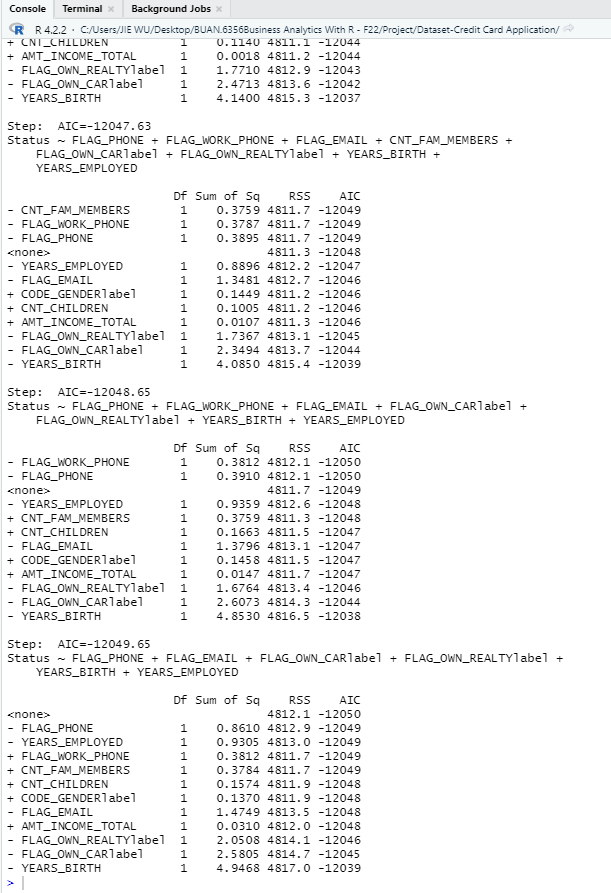
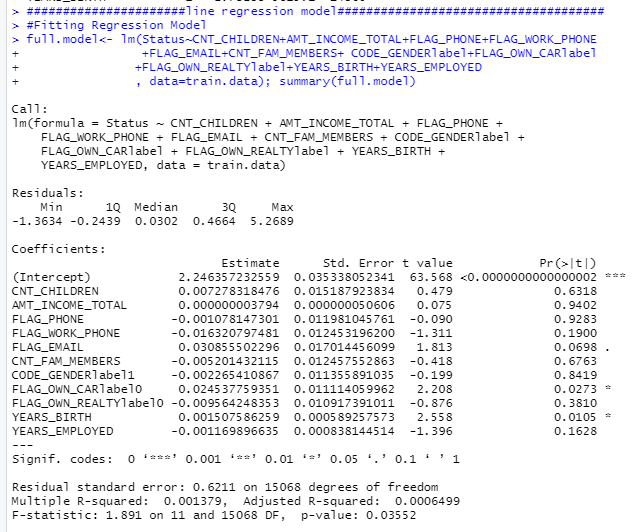
After completing the data processing, we found that there are correlations between different variables, especially we focus on the relationship between the credit status and the applicant's basic information. Such as the relationship between credit status and income, family status, property, work status, etc.

First of all, we focus on correlation, which will help us analyze the relationship between variables and also facilitate model building.

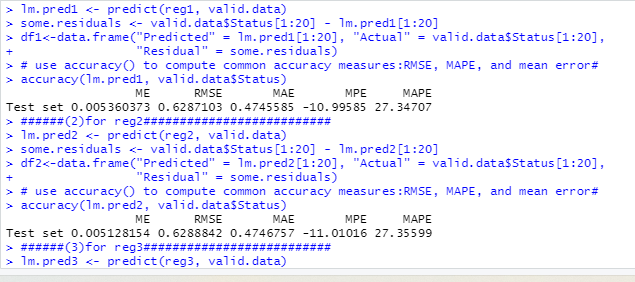




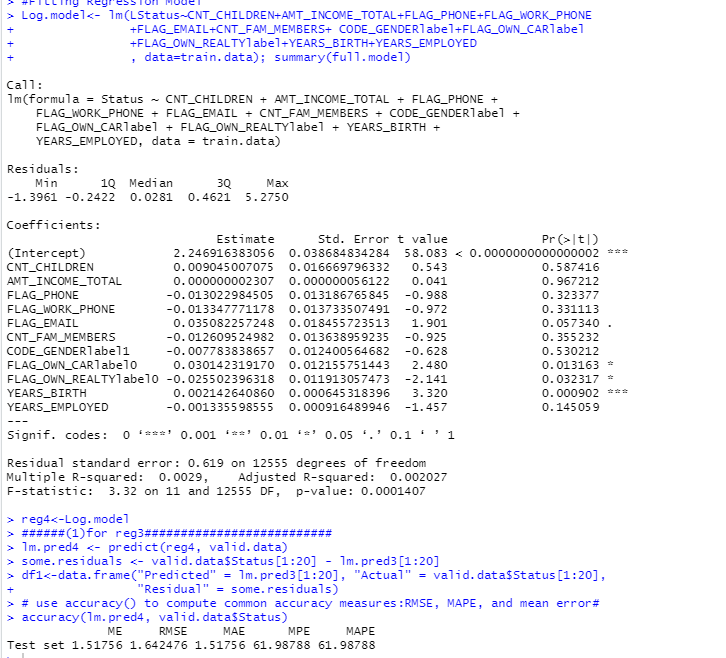
Then, we tried to use the linear regression method, while adopting the AIC method, to select the model



Next, we make predictions and evaluate the accuracy of the model (using predictions and RMSE etc.)



Finally, we also try to use the logarithmic transformation method to try to optimize the model.



# **Data modeling-classification model**

The fit tree obtained is as follows Diagram, timeline

Description automatically generated

There are various decision rules in this tree. The cp value used in the code is optimal, as it has the least cross-validation error, thus producing the above tree which is the best pruned tree.

The confusion matrix generated by running the model on our validation data set is :

Graphical user interface, application

Description automatically generated

This gives us an accuracy measure of 80.9628 ~ 81%.

# **Inferences and Conclusion**

**1.Inferences -Regression and Hypothesis test**

If we estimate the effect of different variables on the risk of default and the relationships regarding the values of the coefficients of the regression model and other analysis, we can find that:

(1)When holing other factors,the relative default risk of women is relatively small ;

(2)There is a certain correlation between the family member variable and the number of children;

(3)Owning property, such as real estate, has relatively low default risk, but owning a car cannot be presumed to reduce default risk;

(4)With more work experience and better communication status, the relative default risk is lower.

From the perspective of hypothesis test, T test and F test, individual factors, such as income, mail, and phone are not significant for predicting risk status, but property status, such as real estate is significant.

**2.Conclusion**

For the bank's credit card product formulation and publicity strategy, we can draw image for the clients:

Age: 30-50 years old

Gender: mostly female

Education Level: mostly without academic degrees

Family status: 1-2 children mainly

Income level: about 100000-200000 per year

Source of income: mainly work salary

Property status: most own real estate property

(1)Product design and optimization ,marketing strategy

1. Since most of the credit card products are mainly middle-aged women, hence,in the product design of the credit card, women's preferences should be considered.
2. Most of the sources of income are from work and most client own real estate property, and traditional mailing advertisements can be considered as good method .
3. At the same time, most families have children and multiple family members. Cooperating with manufacturers of maternal and child products and daily necessities to promote credit cards may be a good choice.

(2)Risk management strategy

Applicants who own real estate, have more work experience and better work and communication status have a relatively lower risk of default;

whether they own real estate or not has no significant impact on the status of presumed default status.

(3) Balance strategy

In the future , how to balance profits and risks is big concern.

a.We can appropriately adjust the risk threshold based on the economic status.

b.Segment management .

If the risk is too high, reduce the credit amount or no grant amount;

if the risk is medium, pay more attention on risk management;

if the risk is low, appropriately increase the credit amount so as to stimulate consumption and increase profits.

c.Dynamic management

We can analyze the customer's monthly repayment status and dynamically assess the default risk. For example, when the customer has reach at the certain level of default risk, it will trigger a warning system.Bank can send bills to the customer regularly and notifying current account status and reminding the repayment deadline.