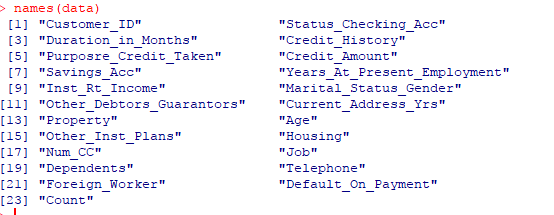
The Task:

The purpose of this analysis and case study is to predict the customers who are likely to be defaulters of a Bank and strategize accordingly so that it can retain their customers.Defaulter someone who does not pay interest or other money that they owe, or who fails to do something that they should do by law. So, to solve this problem, the nature of the relationship between of each variable with Default on Payment must be understood along with the individual characteristic of each variable. A statistical model is then adopted to further the analysis and arrive at the results and interpretation.

The Dataset:

The dataset contains the following variables:

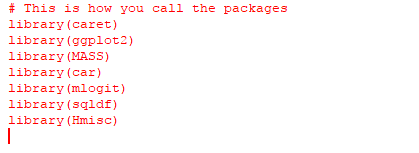


In the given dataset, the first column which is named as “Customer\_ID” and “Count” is of no relevance to our model and is thus excluded.

The statistical model:

We have adopted the logistic regression model analysis in this case. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). In the following pages of documentation, the approach steps have been clearly outlined.

Setting up the R model by loading the required libraries:



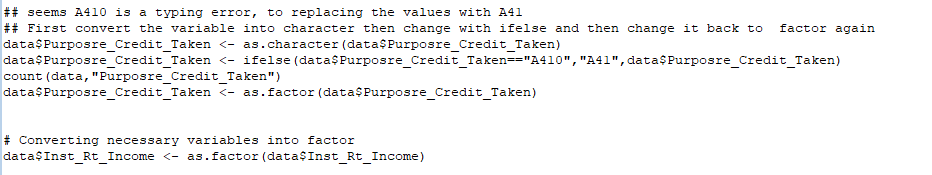
Data:

In the next step, the data is read into the R environment from the file.

C:\Users\ADMIN\Desktop\Bank Defaulters\data fitting.PNG

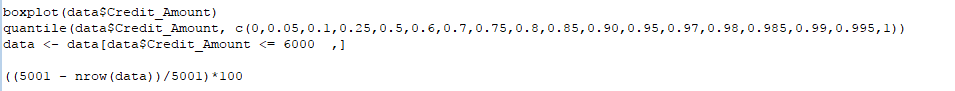
Data selection and data type modification:

In the given dataset, due to some typing error we had to convert “Purpose\_Credit\_Taken” into character then rectified by using iflese and then again convert it back to factor. Another necessary variable “Inst\_Rt\_Income” is converted into factor.

C:\Users\ADMIN\Desktop\Bank Defaulters\str and summary.PNG 

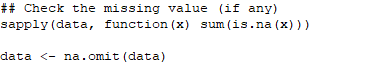
Data Cleaning:

In this step all the numeric variables are checked for the presence of outliers and using boxplots and quantile functions, they are removed. In each case, care is taken to keep at least 80% of the data. Only the data points that lie above or below the maximum and minimum



entries of the boxplot are removed from the dataset.

Checking for missing values:



After the data has been cleaned off all the outliers, it is then checked for any missing values in the following manner. No missing values were found.

**Running the linear regression model:**

Once the data has been cleaned, a logistic regression has been performed with the value as the dependent variable. Once the model has been run, the anova value for each individual variable is checked. The variables with p value< 0.05 are removed from the model one by one such that only the statistically significant ones remain.

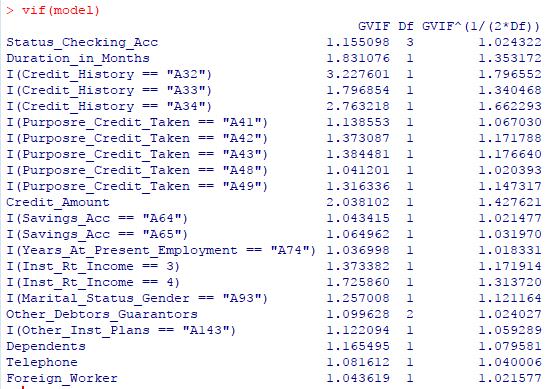


Goodness of fit of the model:

Calculate the C statistic (equivalent to the area under the Receiver Operating Characteristic Curver ROC) for a logistic regression model, a measure of goodness of fit for binary outcomes in a logistic regression model. Values for this measure range from 0.5 to 1.0. A value of 0.5 indicates that the model is no better than chance at making a prediction of membership in a group and a value of 1.0 indicates that the model perfectly identifies those within a group and those not. Models are typically considered reasonable when the C-statistic is higher than 0.7 and strong when C exceeds 0.8. Here it is 83% which indicates a high amount of goodness of fit of the model.

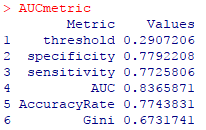
**Performance of the Logistic Regression :**

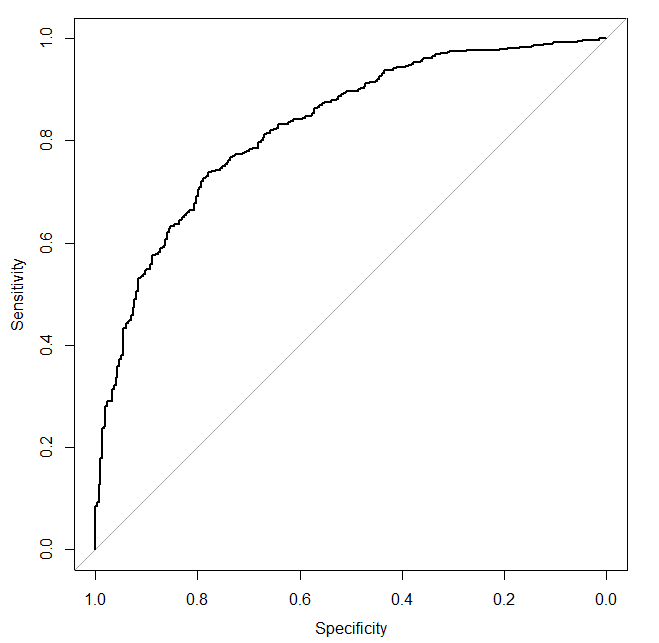
1. ***Assumption of multicollinearity:*** - This is the most important of the assumptions of a linear model and it states that there should be no perfect linear relationship between two or more of the predictors or independent variables. This is tested with the vif function and any variable with a value of GVIF should be within 2. If it is greater than 10 then serious problem. In our case multicollinearity between independent variables was absent.



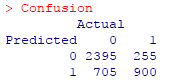
***2> Gini Coefficient:-*** It is the most commonly used measure of inequality. The range of the Gini coefficient goes from 0 (no concentration) to √(n−1n) (maximal concentration). The bias corrected Gini coefficient goes from 0 to 1. The small sample variance properties of the Gini coefficient are not known, and large sample approximations to the variance of the coefficient are poor. In this case it is approximately 0.67.

***3> Accuracy Rate:-*** Accuracy is one of the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. Higher accuracy means model is preforming better. In this case it is approximately 77% which implies to be a better model.

******



***4> Confusion Matrix:-*** It is nothing but a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting. This is how it looks like:



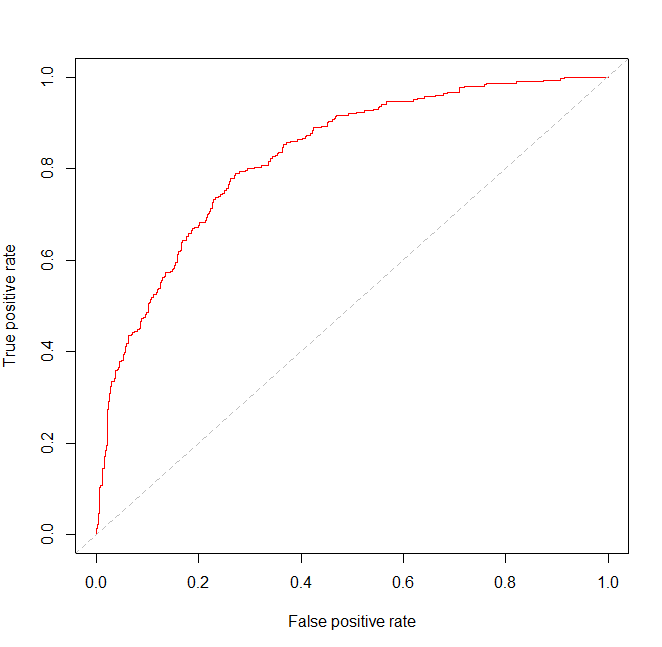
**Validation of the model:**

The predictions for the value variable are done using the validation part of our data. The predictions are then saved in a csv file for reference.

***C:\Users\ADMIN\Desktop\Bank Defaulters\save data.PNG***

***KS Statistics:*** Kolmogorov-Smirnov (KS) statistics is one of the commonly used measures to assess predictive power for marketing or credit risk models. The KS statistic is usually published for logistic regression problems to give an indication of the quality of the model.The value for KS ranges from 0 to 1 and the closer to 1 the KS, the better is the model. The calculation is :

C:\Users\ADMIN\Desktop\Bank Defaulters\KS.PNG



We can assume the model to be ok because of above result.

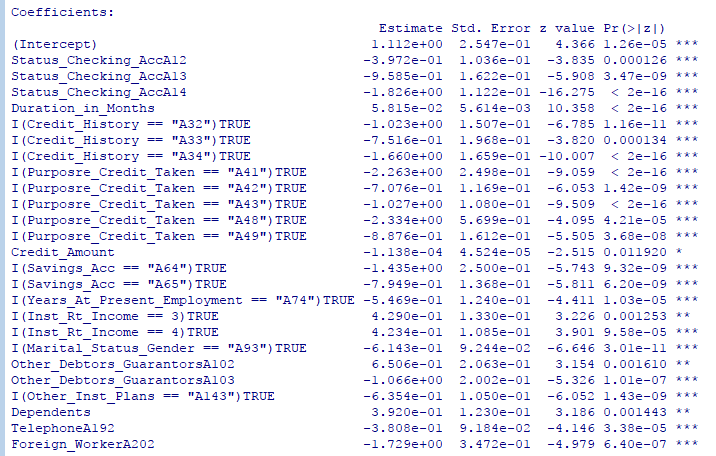
The significant variables and their significance:

The following image shows the variables that are significant to our model and the image below shows us the relationship of these variables with the dependent variable. The following are the significant variables with a positive relationship with customer lifetime value:

* Duration\_in\_Months
* Inst\_Rt\_Income
* Other\_Debtors\_GuarantorsA102
* Dependents

However, the following are the significant variables with a negative relationship with the dependent variable:

* Status\_Checking\_Acc
* Credit\_History
* Purpose\_Credit\_Taken
* Credit\_Amount
* Savings\_Acc
* Years\_At\_Present\_Employment
* Marital\_Status\_Gender
* Other\_Debtors\_GuarantorsA103
* Other\_Inst\_Plans
* Telephone
* Foreign\_WorkerA202



The Business Interpretation

For the Bank to prosper they should focus on the following targets:

* The Bank should focus on number **3** and **4** of the variable **Inst\_Rt\_Income**. These customers are potential defaulters when it comes to Default\_on\_Payment as they have positive relationship as inferred by the model.
* Customers with high **Duration in Months** should also be targeted. They also have a strong positive relationship with churn.
* The Bank must look into the **Other\_Debtors\_Guarantee** as number **A102** is likely to be a defaulter whereas number **A103** unlikely to be a defaulter.
* **Dependents** can be fruitful and this variable should be kept within the fold.
* Number (**A12-A14**) of **Status\_Checking\_Acc** have the highest chances not to be defaulters as these have quite negative relationship with Default\_on\_Payment.
* **Credit\_History** numbers (**A32-A34**) must not be pursued as it has significantly negative relationship with Default\_on\_Payment.
* **Purpose\_Credit\_Taken** numbers (**A41, A42, A43, A48, A49**) are not so important as it have much negative relationship.
* Customers with higer **credit amount** are not likely to be defaulters have significantly negative relationship.
* **A64** and **A65** of **saving accounts** must not be pursued as these customers are negatively related to Defaulter\_on\_Payment.
* The Bank should not focus on the customer **A74** of **Years\_At\_Present\_Employment** as it has significantly negative relationship with Defaulter\_on\_Payement.
* Variable **Marital\_Status\_GenderA93** has lesser chance to be a defaulter.
* **A143** of **Other\_Inst\_Plans** also has lesser chance to be a defaulter.
* **TelephoneA92** and **Foreign\_WorkerA202** are unlikely to be a defaulter as these two has negative relationship with Default\_on\_Payment.