LOGISTIC REGRESSION SUBMISSION

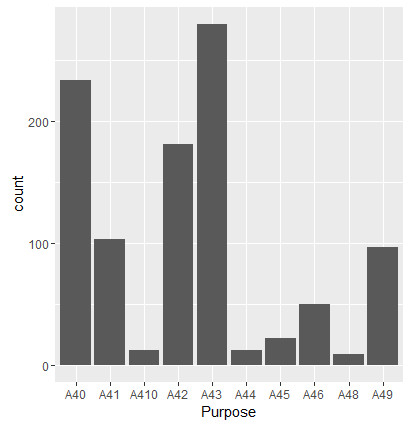
**NOTE:** This should briefly describe the important results and recommendations. The structure is suggestive; make sure to not exceed 7 pages**.**

# Checkpoint-1: Data Understanding and Data Exploration

* Display the plots and explain the insights

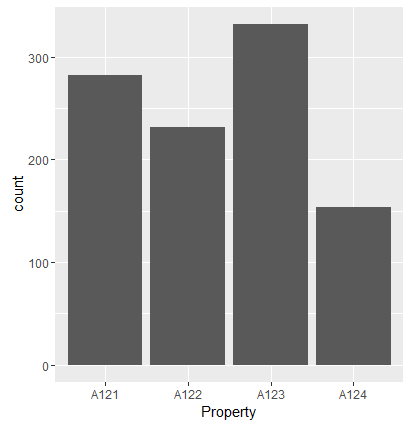
Plot 1: The below plot shows the distribution of various Purposes for which the loan was taken.

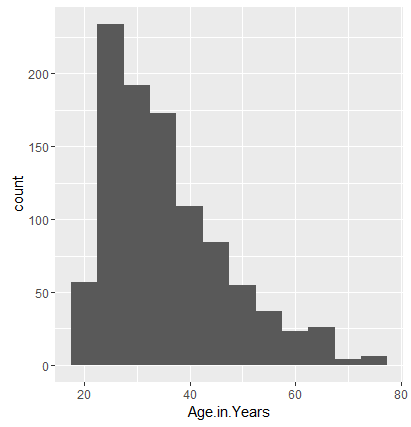
Most of the loans are taken for the purpose A43(radio/television), A40(new car) and A42(furniture/equipment) as can be seen in the univariate graph below.



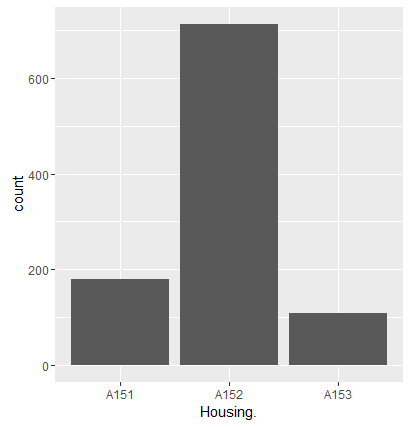
Plot 2: The below plot shows the distribution of various Purposes for which the loan was taken.

Most of the loans are taken for the property type A123(car or other)

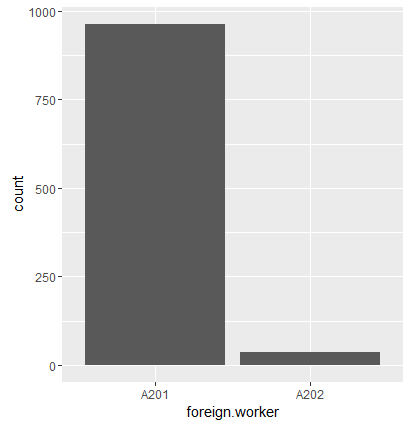


Plot 3: The below plot shows the distribution of various Age groups of people who have taken the loans. Most of the loans are taken within the age group 25-40.

Plot 4: The below plot shows the distribution of housing status of people who have taken the loans. Most of the loans are taken by the people with own houses.



Plot 5: The below plot shows the distribution of foreign status of personal who have taken the loans. Most of the loans are taken by the people who are foreign workers.



# Checkpoint 2: Data Cleaning and Transformation

* Explain the methodology of Missing value treatment and additionally fill the below table:

|  |  |
| --- | --- |
| **Questions** | **Results(Numeric)** |
| Total number of observations in the dataset | 1000 |
| Total number of variables in the dataset | 21 |
| Total missing values in the dataset | 0 |

* Explain the methodology of Outlier treatment and fill the below table:

Outliers in this case are treated using the flooring or capping principles. We check for every percentile of the value using the quantile function, if there is a jump. If we observe a drastic improvement in the values at a difference of 1%, then we decide to floor or cap at that percentile value for the rest of the values above or below it respectively.

* Explain the methodology of how did you create dummy variables

The dummy variables are created using the model.matrix function. This function generates n columns for a variable with n different factor types. We remove one column as n-1 columns are enough to represent the data. We then column bind these variables to the main dataset removing the factor column and adding these continuous value columns.

* If binning for numerical variables done explain why it was required?

There is no binning done for any numeric variables.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| Outlier treatment | Duration.in.month  Credit.amount  Age.in.Years |
| Dummy creation | Status.of.existing.checking.account  Credit.history  Purpose  Savings.account.bonds  Present.employment.since.  Installment.rate.in.percentage.of.disposable.income  Personal.status.and.sex  Other.debtors...guarantors  Present.residence.since  Property  Other.installment.plans  Housing.  Number.of.existing.credits.at.this.bank.  Job\_status  Number.of.people.being.liable.to.provide.maintenance.for.  Telephone.  foreign.worker |
| Binning of variables | None |

# Checkpoint 3: Splitting the Dataset into train and test

The data contains 1000 observations. The target variable Default\_status has 30% values with Default\_status=1. If we randomly choose 30% points as test data, it may not contain or contain very less Default\_status with values = 1. we'll use the sample.split() function from the caTools package which ensures splitting of the test and train data in same proportion as master data.

# Checkpoint 4: Modelling

* Explain the methodology of building the model? In the final model, interpret what the coefficients of the variable imply. Check if the coefficients make business sense

We first build an initial model with all the available variables. Then using the stepAIC function on this model we filter the initial set of variables to proceed with model creation.

Then we perform variable selection starting with the model given by stepAIC. Then perform it iteratively which is called Feature selection through stepwise variable selection process, also use vif to check multicollinearity and iteratively eliminate the insignificant variables. The threshold for the VIF function is used as 3 in this assignment.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Significant variables in final model (add more rows if requires)** | **Coefficients value (Numeric)** |
| Intercept | -1.261428 |
| Duration.in.month | 0.050362 |
| Status.of.existing.checking.accountA14 | -1.752497 |
| Credit.historyA34 | -0.515349 |

Duration.in.month : This variable positively impacts, as duration increases the probability of defaulting loan decreases as it may most probably have smaller instalments.

Status.of.existing.checking.accountA14 : This variable impacts negatively. People with no checking account tend to default more.

Credit.historyA34: This variable impacts negatively. People who have critical account/other credits existing in other bank tend to default more.

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 734.4 |
| Null deviance | 855.21 on 699 degrees of freedom |
| Residual Deviance | 726.40 on 696 degrees of freedom |

# Checkpoint 5: Model Evaluation

* Calculate c-statistic and KS-statistic. What can you tell about the model based on their values?

The **c-statistic**is an implicit measure of the fraction of concordant pairs.  More the concordant pairs higher the value of c-statistic. The c-statistic for both the train and the test data set are 75.58% and 73.18% respectively which are satisfactory values and it says that model is predicting well on both the train and the test data sets.

KS-statistic is a way to measure the goodness of fit of logistic regression models. KS- Statistic is a measure of the degree of separation between the positive and negative distributions. The KS-statistic values obtained from the final model for the test and train data sets are 0.3952381(39) and 0.4095238(~41) which are satisfactory values and it says that model is predicting well on both the train and the test data sets.

Ideally the k-statistic should be greater than 40 for a given model. Here for train it is 39.5 and test it is almost 41. Also, both the KS statistics lie in the 1st decile which is a very good sign as it lies in the top deciles. So, we can accept this model for the prediction of the loan defaulting.

Additionally, fill the below table:

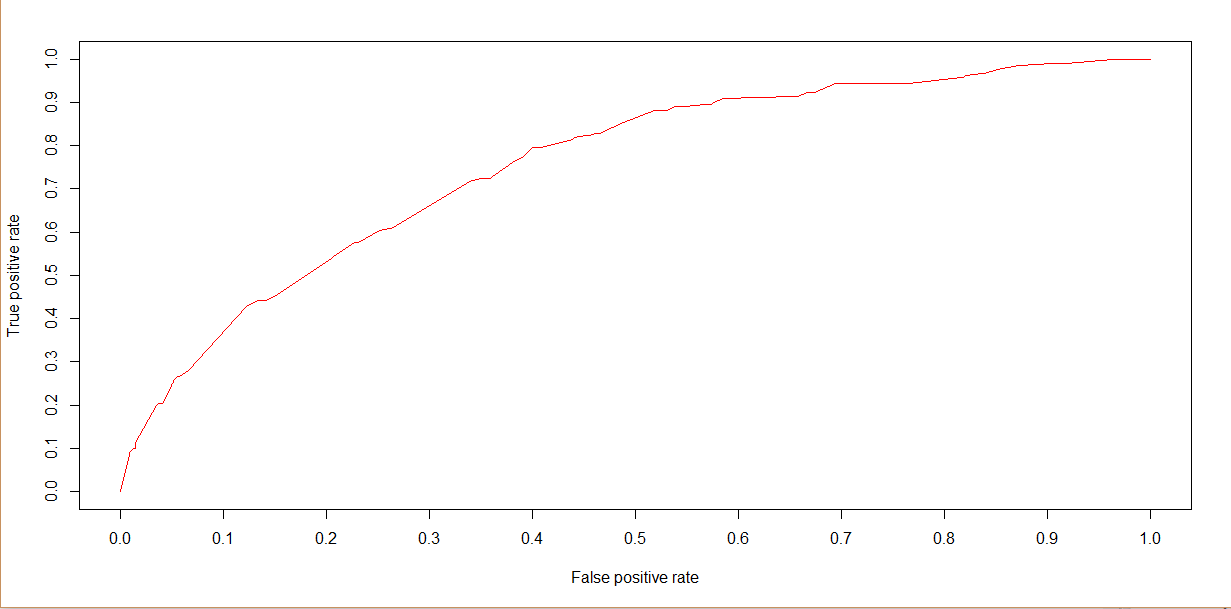
**Note**: Write the numeric value of c-statistic and KS-statistic after applying your final model to the train dataset and test dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| **Train Dataset** | | **Test Dataset** | |
| C-statistic | 75.58% | C-statistic | 73.18% |
| KS-statistic | 0.3952381 (1st decile) | KS-statistic | 0.4095328(1st decile) |
| Model Evaluation (write Accept or Reject) | | Accept | |

# Checkpoint 6: Threshold value

* Select an appropriate threshold value and calculate the confusion matrix and overall accuracy, sensitivity and specificity

We can find the correct threshold from the ROC curve plotted for model\_perf(train data set) we can see that the threshold value of 0.5 is appropriate for this model:



Additionally, fill the below table:

Train data set

|  |  |
| --- | --- |
| **Parameter** | **Values (Numeric)** |
| Threshold value | 0.5 |
| Overall Accuracy | 73.86% |
| Sensitivity | 28.09% |
| Specificity | 93.46% |

Test data set

|  |  |
| --- | --- |
| **Parameter** | **Values (Numeric)** |
| Threshold value | 0.5 |
| Overall Accuracy | 69% |
| Sensitivity | 26.67% |
| Specificity | 87.14% |