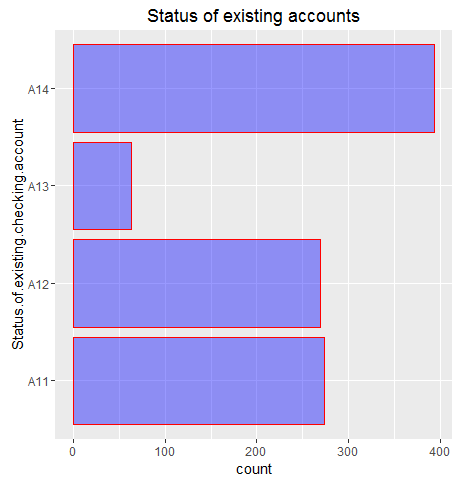
LOGISTIC REGRESSION SUBMISSION

**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

1. Status.of.existing.checking.account



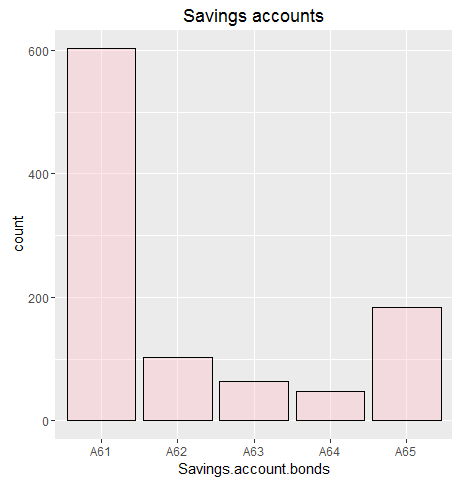
There are high number of A14 (no checking account). Very less number of accounts with 0 DM <=…< 200 DM. Bank is not considering existing checking accounts. If at all the customers have existing checking accounts money amount should be significant.

1. Credit History



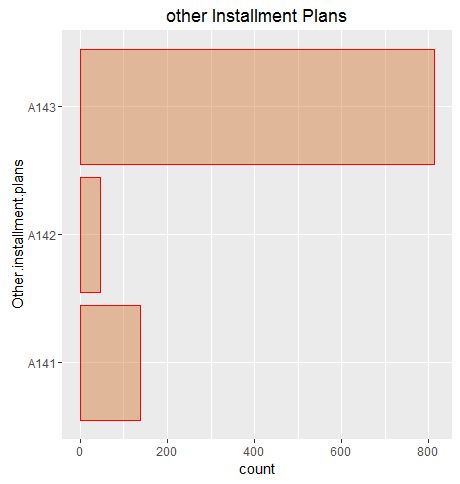
High number of people have their all their credits paid back duly at this bank (A32). But there is also a significant count of people whose accounts are critical (A34). Bank should take quick action on the critical accounts.

1. Savings account bonds



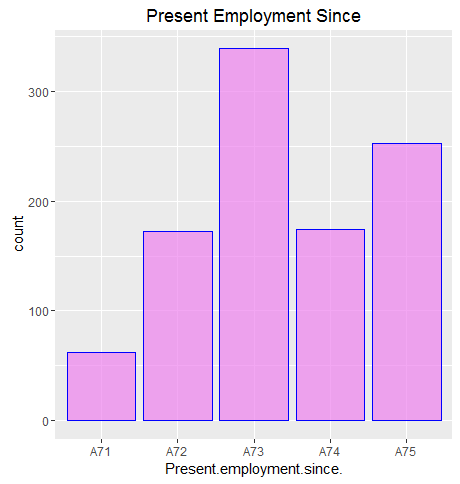
High number of people have <100 DM as their Savings account bonds. Percentage in 60% of the whole dataset. Close to 19% people who have taken credit have NO savings accounts bonds.  
Very less people have significant money in savings accounts bonds. Bank should look if it should consider people to have savings bonds accounts while availing credit.

1. Other Installment Plans



More than 80% of the customers of the bank have no other instalments, which is a good sign.

1. Present Employment Since



High number of customers availing credit with the bank fall in the range 1-4 years.

And less number of customers, with less than 1 yr of experience or no experience. This means bank is less encouraged to provide credit to customers who are unemployed or experience less than 1 yr. Good step from bank’s perspective to avoid risk.

# 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 above upper whisker value was capped with the upper whisker value.

Scaling was done on 3 variables to negate the impact of magnitude of the variable values.

* Explain the methodology of how did you created dummy variables.  
  Dummy variables was created for the factor variables using model.matrix function. Intercept was removed.

The number of dummy variables were maintained such that:  
“number of dummy variables = number of levels in the categorical variable – 1”

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

Binning is not done on any of the variables.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| Outlier treatment | Age.in.Years |
| Dummy creation | Status.of.existing.checking.account  Credit.history  Purpose  Savings.account.bonds  Present.employment.since.  Personal.status.and.sex  Other.debtors...guarantors  Property  Other.installment.plans  Housing.  Job\_status  Telephone.  foreign.worker |
| Binning of variables | NONE |

# Checkpoint 3: Splitting the Dataset into train and test

Seed was set to 100. And the dataset was divided in 70:30 ratio (train:test) keeping in mind the, proportion of the response variables is same in train and test.

# 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

Step function was used for finding out the optimised model. After that further model building was done by removing more number of least significant variables. It was also taken care that, there is no significant jump in the AIC values while removing the variables and the final model AIC was kept below.

The VIF value was targeted to be restricted to be below 3.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Significant variables in final model (add more rows if requires)** | **Coefficients value (Numeeric)** |
| Duration.in.month | 0.4880 |
| Installment.rate.in.percentage.of.disposable.income | 0.2437 |
| Status.of.existing.checking.accountA13 | -0.9426 |
| Status.of.existing.checking.accountA14 | -1.8497 |
| Credit.historyA32 | -1.1282 |
| Credit.historyA33 | -1.2715 |
| Credit.historyA34 | -1.7049 |
| PurposeA41 | -0.7334 |
| Other.installment.plansA143 | -0.4653 |
| Savings.account.bondsA64 | -1.1458 |
| Savings.account.bondsA65 | -0.5434 |
| Present.employment.since.A74 | -0.6592 |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 692.7 |
| Null deviance | 855.2 |
| Residual Deviance | 666.7 |

# Checkpoint 5: Model Evaluation

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

According to the C-statistics, C- Index for train as well as test are above 0.70 i.e. 70%. Indicating a good model.  
Ideally, for a good model c-statistics lie between 0.5 to 0.7.   
Currently, for our model C-statistics are above 0.70, indicating better discriminatory power.

According to KS-statistics, KS-stats are above .40 for both train and test datasets. And the cumulative percentages for both train and test lies in top 5 deciles.   
Value for Train lies in 3rd decile.  
Value for Test lies in 4th decile.

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 | 8.095675e-01 | C-statistic | 7.475661e-01 |
| KS-statistic | 0.4836735 | KS-statistic | 0.415873 |
| 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

Threshold value selected is 0.35.

Since we want to minimise the False Negative value, low threshold value has to be selected i.e we do not want to predict a customer to be a non-defaulter when he is actually a defaulter.

Additionally, fill the below table:  
**Train Dataset**

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.75 |
| Sensitivity | 0.7000 |
| Specificity | 0.7714 |

**Test Dataset**

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7067 |
| Sensitivity | 0.6222 |
| Specificity | 0.7429 |