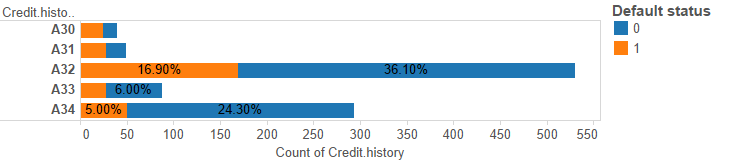
LOGISTIC REGRESSION SUBMISSION

# Checkpoint-1: Data Understanding and Data Exploration

* Display the plots and explain the insights

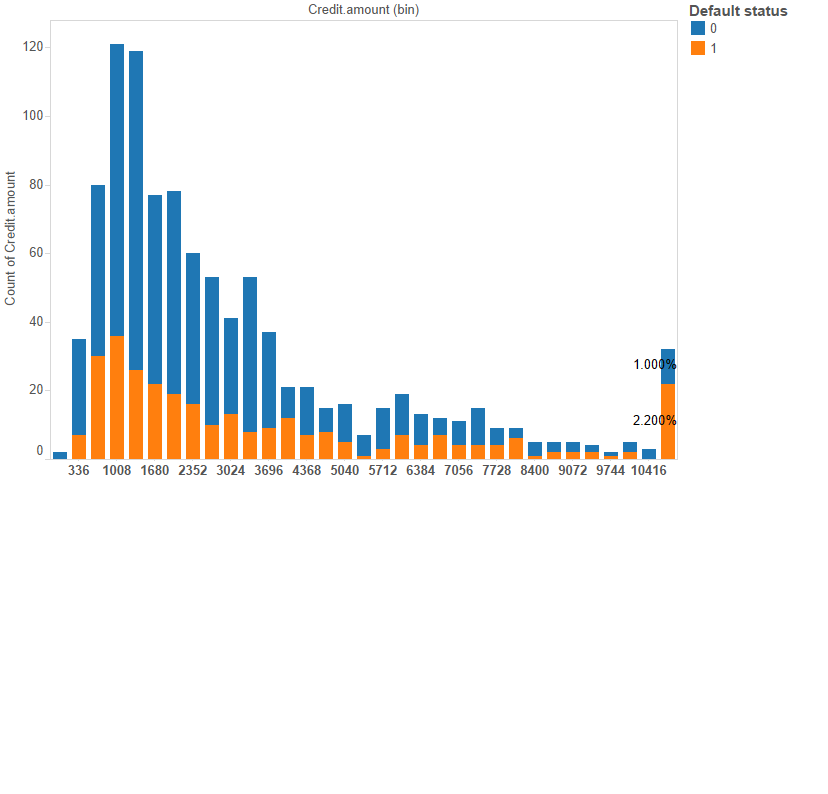
CREDIT HISTORY PLOT



Credit history plots indicate that people with critical accounts show the least tendency to default.

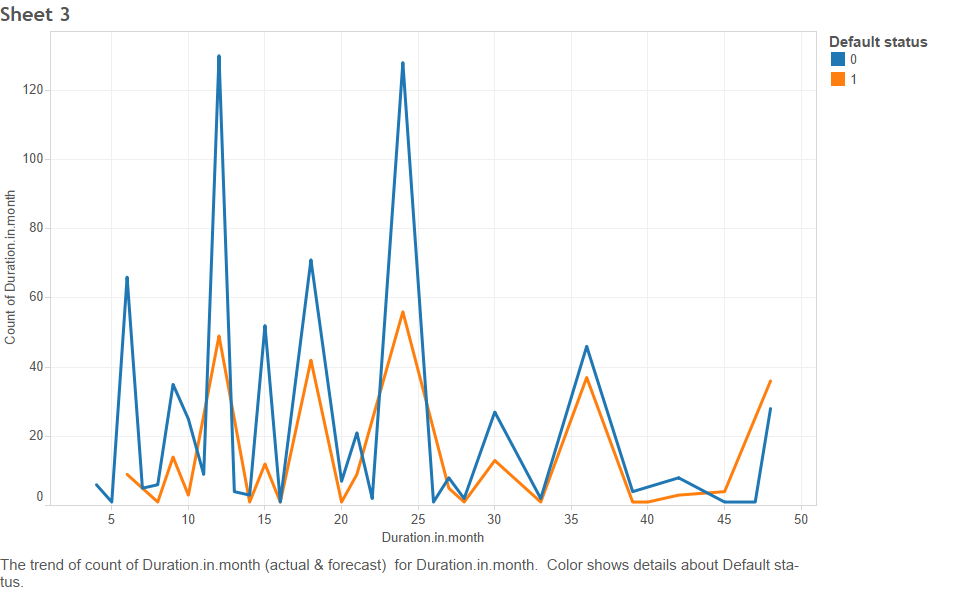
Quite contrastingly, people who have existing credits and have duly paid their dues seem to default more often.

CREDIT AMOUNT PLOT



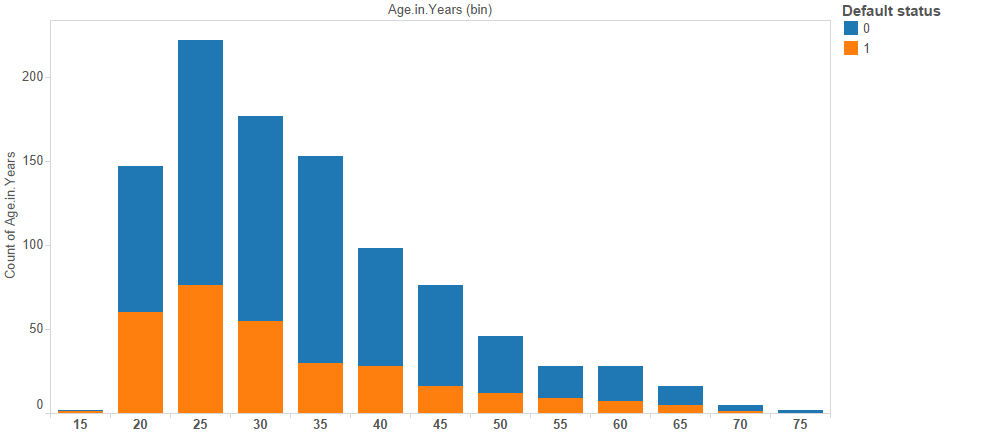
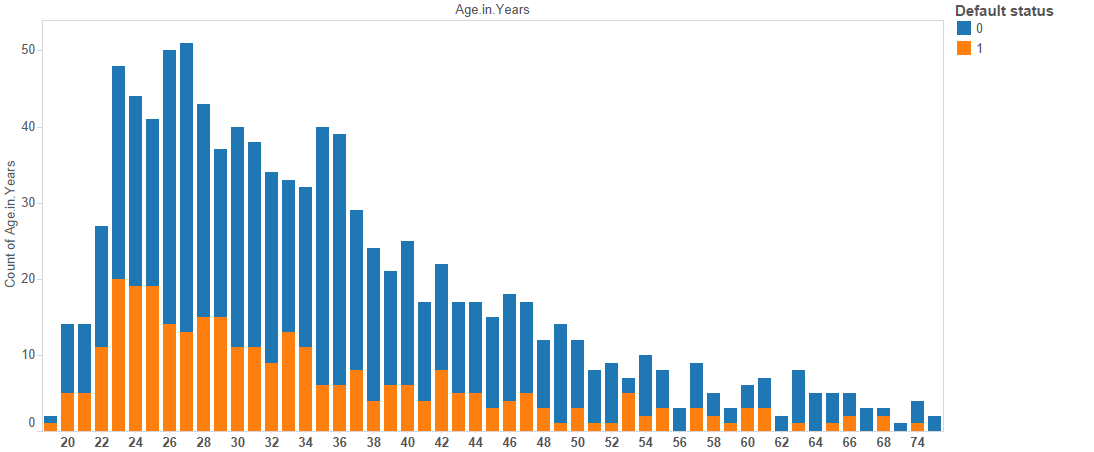
Credit amount plots show that regardless of the bucket, the ratio of defaulters to non defaulters is the same except in the highest bucket. This means that credit amount does not sufficiently explain defaulting behaviour of the consumer.

DURATION IN MONTHS PLOT



For durations less than 10 months, tendency to default is less. After that, it appears to have a trend, with defaults peaking regularly at certain intervals of time. Higher the duration, greater the tendency to default.

AGE PLOT



# 

# Except for ages beyond 55, there doesn’t appear to be any significant trend w.r.t default (orange) when it comes to age or age groups (2nd graph, binned with 5 years).

# **PURPOSE PLOT**

# 

# Tendency to default is least when the people take a loan for getting a used car(A41) or retraining A(48). The highest tendency for default is when the loan is for a new car (A40).

# 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:

A two pronged approach was used for outlier treatment

* Box Plot : For numeric variables, box plots were first created to gauge the density of outliers. If the outliers were significant in magnitude, then the next approach was adopted. Boxplot function in R was used for this.
* Capping and Flooring: After dividing the dataset into quantiles, each quantile was observed for significant jumps. Once identified, the value at the quantile was benchmarked as the upper or lower bound, beyond which the corresponding values were either reduced or increased to the upper or lower bounds respectively. Quantile function in R was used for this.

* Explain the methodology of how did you created dummy variables

Factors were split into various columns and the presence of any factor was indicated by 1 and absence indicated by 0. Since by this method, n-1 columns are sufficient to indicate n factors, the first column was removed. These dummy columns were then combined into a separate data frame, excluding the original column. Column.matrix function in R was used for this.

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

No binning performed.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| Outlier treatment | 1. Duration.in.month  2. Credit.amount  3. Number.of.existing.credits.at.this.bank.  4. Number.of.people.being.liable.to.  provide.maintenance.for. |
| Dummy creation | 1. Status.of.existing.checking.account 2. Credit.history 3. Purpose 4. Savings.account.bonds 5. Present.employment.since. 6. Personal.status.and.sex 7. Other.debtors...guarantors 8. Property 9. Other.installment.plans 10. Housing. 11. Job\_status 12. Telephone. 13. foreign.worker |
| Binning of variables | NA |

# Checkpoint 3: Splitting the Dataset into train and test

Training data set : 700 observations with 0 to 1 ratio of 3:1

Testing data set : 300 observations with 0 to 1 ration of 3:1

# 

# 

# 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

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Significant variables in final model (add more rows if requires)** | **Coefficients value (Numeeric)** |
| Duration in months | 0.039606 |
| Checking Account status A12 | -0.364056 |
| Checking Account status A13 | -1.049760 |
| Checking Account status A14 | -2.031559 |
| Credit History A31 | -0.020208 |
| Credit History A32 | -1.299937 |
| Credit History A33 | -1.377065 |
| Credit History A34 | -1.881052 |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 711.37 |
| Null deviance | 855.21 on 699 degrees of freedom |
| Residual Deviance | 693.37 on 691 degrees of freedom |

# Checkpoint 5: Model Evaluation

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

C-statistic = 72.8% (test) and 79.16(train) indicates that the model produces a high number of concordant pairs.

K- Score = 102/700 (training) and 55/300 (test) indicates that the model is good since in both cases, it lies in the first 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 | 79.16 | C-statistic | 72.8 |
| KS-statistic | 102 | KS-statistic | 55 |
| 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 0.25

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 67%(train) and 63%(test) |
| Sensitivity | 82.38(train) and 76.67(test) |
| Specificity | 60.41 (train) and 57.14%(test) |