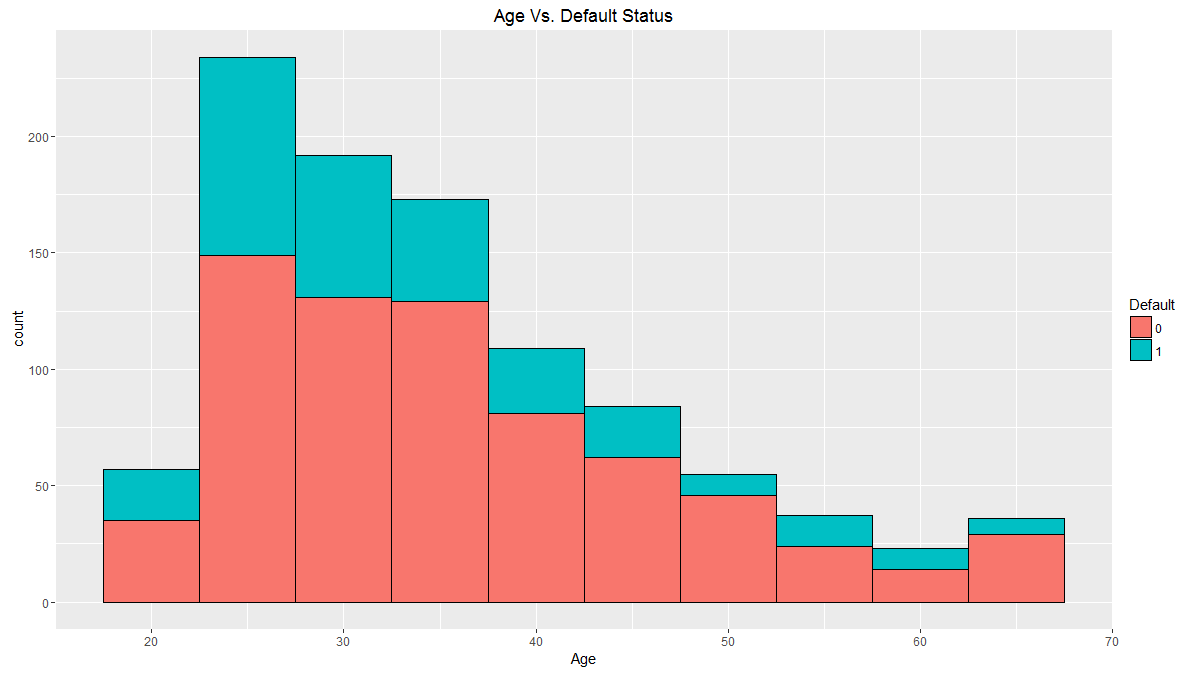
LOGISTIC REGRESSION SUBMISSION

**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

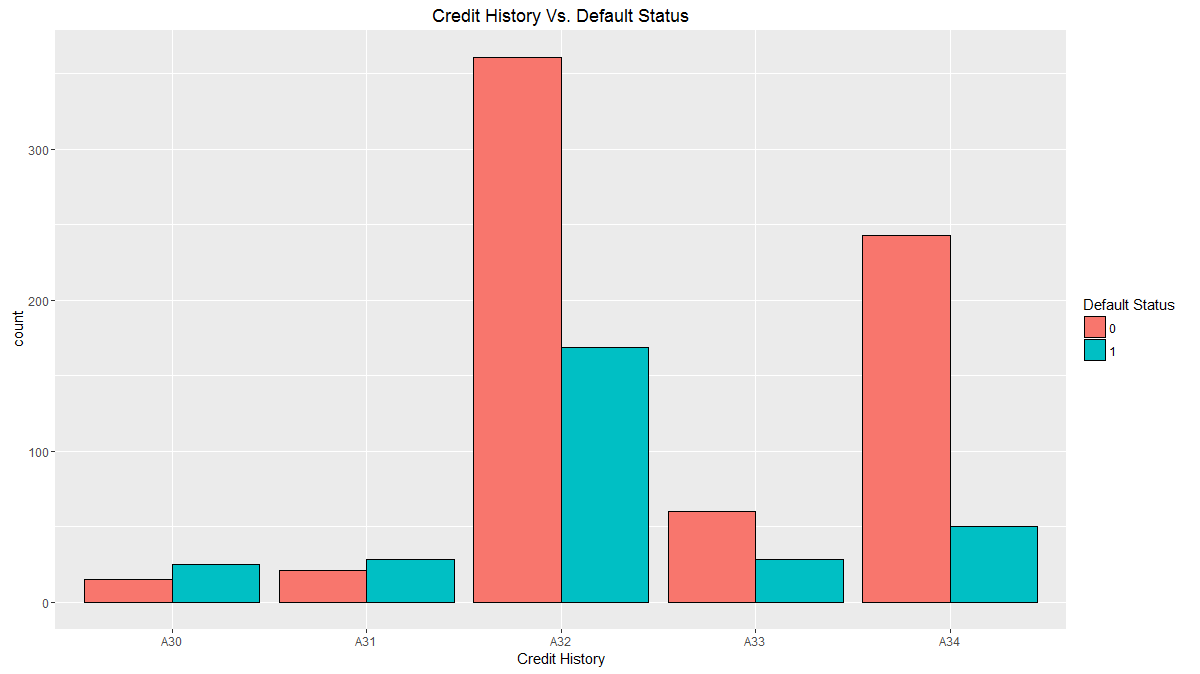
* *Assumption: Default status=1 is for customers who have defaulted.*

1. Age vs. Default Status



We can observe from this plot is that more loans are taken by people in the age group 25-35 years. Default ratio is also high in that same age group.

2. Credit History vs. Default Status



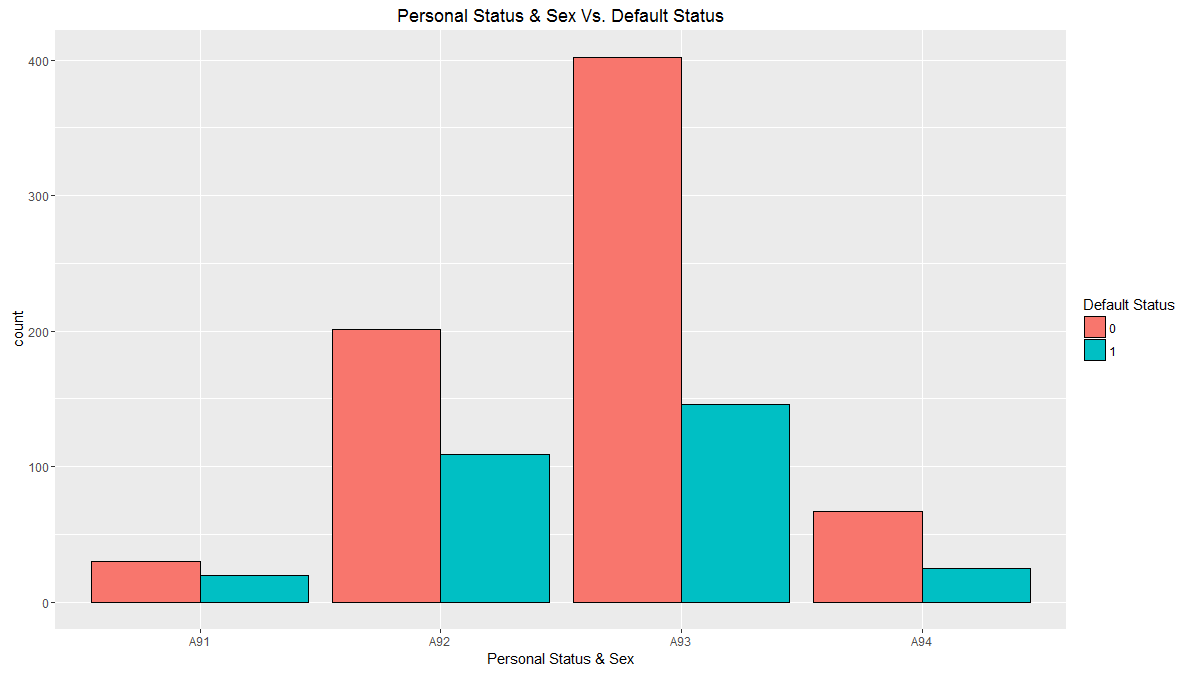
We observe that default is more than half of all loans to customers where credit history is -

A30: no credits taken/all credits paid back duly

A31: all credits at this bank paid back duly

Contrary to intuition, customers with credit history A34: critical account/other credits existing (not at this bank) have defaulted the least.

1. Personal Status vs. Sex

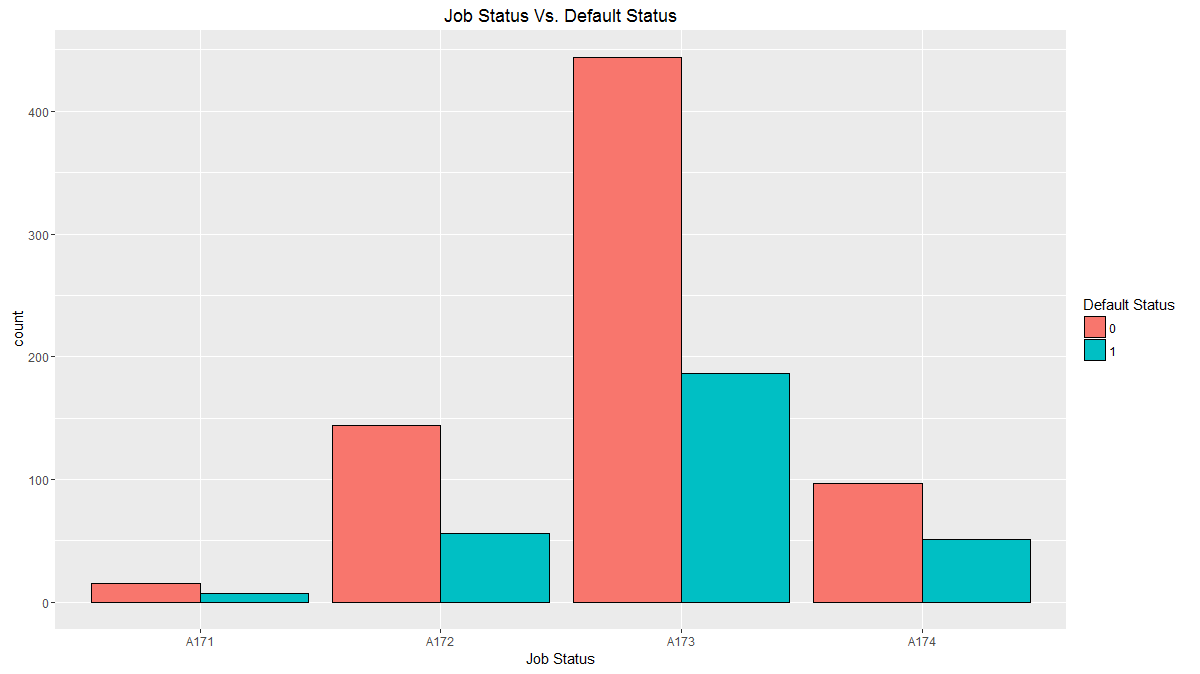


Most loans are taken by A93: Single:Male but more percentage of defaults are seen in-

A91: male: divorced/separated

A92: female: divorced/separated/married

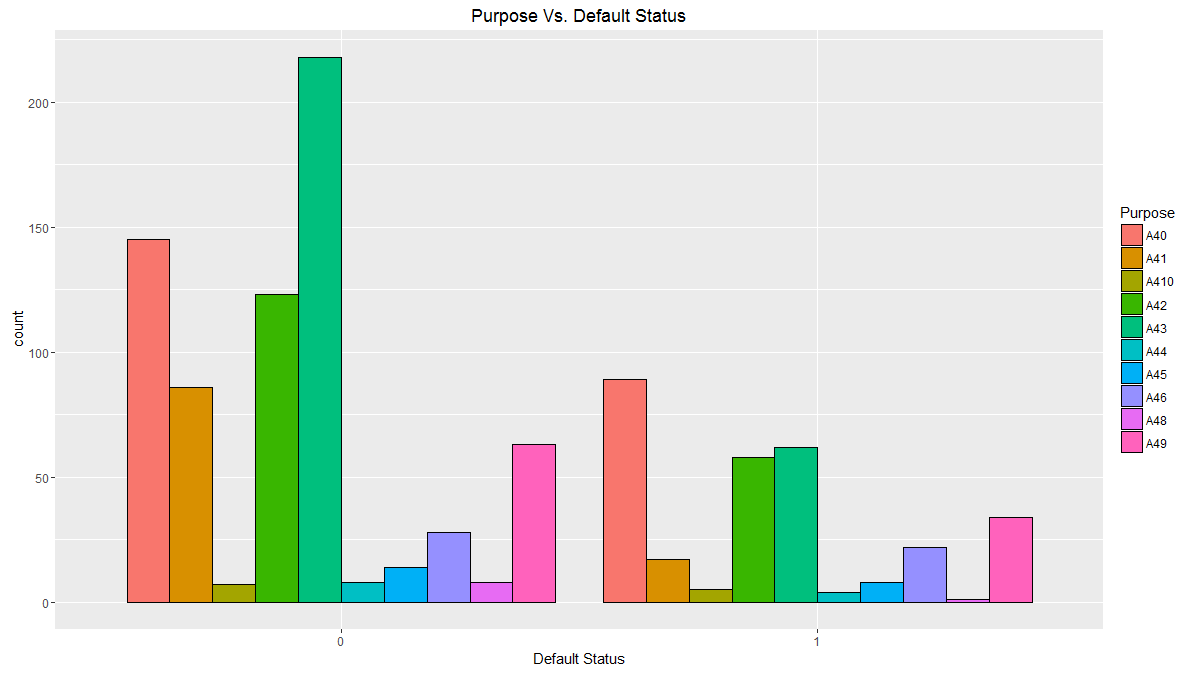
1. Job Status vs. Default Status



Though most loans are taken by customers in job A173: skilled employee / official, more percentage of defaults are seen in with customers with job status:

A174: management/ self-employed/highly qualified employee/ officer

1. Purpose vs. Default Status



We observe here is that even though purpose A40: Car(New) is not biggest in terms of purpose of load, it clearly stands out in terms of defaults. Car loans for a new car are the biggest category of defaulted loans.

# 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 are capped at Upper hinge + 1.5 times Inter Quartile Range(IQR) for outliers on the higher side and Lower Hinge – 1.5 times IQR for outliers on the lower side.**
* Explain the methodology of how did you created dummy variables: **Dummy variables are created using model.matrix function and then combining the resulting data frame with the original data frame.**
* If binning for numerical variables done explain why it was required? **Binning was not done.**

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| Outlier treatment | 1. Duration.in.month  2. Credit.amount  3. Age.in.Years |
| Dummy creation | 1. Status.of.existing.checking.account  2. Credit.history  3. Purpose  4. Savings.account.bonds  5. Present.employment.since.  6. Installment.rate.in.percentage.of.disposable.income  7. Personal.status.and.sex  8. Other.debtors...guarantors  9. Property  10. Other.installment.plans  11. Present.residence.since  12. Housing.  13. Number.of.existing.credits.at.this.bank.  14. Job\_status |
| Binning of variables | None |

# Checkpoint 3: Splitting the Dataset into train and test

# Creating Train and Test data in the ratio 7:3 using sample split on Default\_status variable

set.seed(100)

split\_german\_credit = sample.split(german\_credit\_final$Default\_status, SplitRatio = 0.7)

table(split\_german\_credit)

german\_credit\_final.train = german\_credit\_final[split\_german\_credit,]

german\_credit\_final.test = german\_credit\_final[!(split\_german\_credit),]

# 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 :

Steps -

* + Initial model is build using glm function with Default value as the predicted variable and all other variables as predictor variables.
  + The resulting model is refined using stepAIC function on the model where insignificant variables are dropped from the model.
  + Then the refined model is further refined by analysing p-value, VIF of the variables and AIC of the model.
  + Variables with high VIF and high p-value are dropped and AIC is kept as low as possible.

Additionally, fill the below table:

|  |  |  |
| --- | --- | --- |
| **Significant variables in final model (add more rows if requires)** | **Coefficients value (Numeric)** | **Interpretation** |
| Duration.in.month | 0.040659 | Longer the duration more chances of default. Intuitive. |
| Status.of.existing.checking.accountA13 | -0.89428 | More money in checking account less chances of default. Intuitive. |
| Status.of.existing.checking.accountA14 | -1.89719 | No checking account also reduces changes of default |
| Credit.historyA32 | -0.59294 | Timely payback of credits increases the chances of default. Counter-intuitive. |
| Credit.historyA34 | -1.08736 | If other credits exists, chances of default also reduces. |
| Savings.account.bondsA64 | -1.21999 | Larger the bonds held, less are the chances of default. Intuitive. |
| Savings.account.bondsA65 | -0.62838 | If no bonds are held or information is not available chances of default decreases. |
| Installment.rate.in.percentage.of.disposable.income4 | 0.658082 | More is the instalment percentage of disposable income, more are the chances of default. Intuitive. |
| Personal.status.and.sexA93 | -0.50851 | If loan is taken by a single male, chances of default decreases. |
| Other.installment.plansA143 | -0.63388 | If there are no other instalment plans, chances of default decreases. Intuitive. |
| Present.residence.since2 | 0.538635 | If a customer is living in his current residence for last two years, his chances of default increases. |
| Housing.A152 | -0.48821 | Customer who own their houses are less likely to default. Intuitive. |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 693.42 |
| Null deviance | 855.21 |
| Residual Deviance | 667.42 |

# Checkpoint 5: Model Evaluation

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

**C statistics of more than 76% shows that proportion of concordant pairs is high in the model. This indicates that model has high discriminating capability.**

**KS-Statistics for training and test data lie in 4th and 5th decile respectively.**

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 | 0.8094 | C-statistic | 0.7659 |
| KS-statistic | 0.529932 | KS-statistic | 0.4063492 |
| Decile | 4th | Decile | 5th |
| Model Evaluation (write Accept or Reject) | | **We accept the model.** | |

# Checkpoint 6: Threshold value

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

**We assume that value of 1 in the default status to be customers who have defaulted. Therefore, the positive class is 1. Aim of our model is to reduce the chances of giving loan to customers who are likely to default. Hence, false positive are to be minimised. Sensitivity, which is the ratio of true positives to total positives, should be high for a model in this scenario.**

**Using probability threshold of 0.3, we get the best sensitivity, specificity and accuracy of the model.**

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.6833 |
| Sensitivity | 0.7333 |
| Specificity | 0.6619 |