**Statistical Analysis Report for Loan Approval**

This Analysis report contains the Model analysis and Prediction using different Machine Learning Classification tools based on the features selection using different techniques from different packages. Also, it contains the visual representation and analysis of the relationship between the features and the target using ggplots. The analysis, prediction and report generation was done using R language and R Studio and several classification Machine Learning packages in R. Source code(R) for this report is attached in this document.

1. **Analysis of Model Variance using ANOVA**

The following R code shows creating a model using logistic regression(glm) and analyzing the Deviance of different features and trying to hand pick only the important features and remove the features that doesn’t give much value and also noise to prediction .

glmmodel1 <- glm (ApprovedLog ~ LoanPayoffPeriodInMonths + RequestedAmount + CoApplicant + YearsAtCurrentEmployer + Age + TypeOfCurrentEmployment + CheckingAccountBalance + DebtsPaid + SavingsAccountBalance + CurrentOpenLoanApplications + InterestRate + LoanReason + NumberOfDependantsIncludingSelf + RentOrOwnHome + YearsInCurrentResidence, data=trainData)

anova(glmmodel1, test="Chisq")

Analysis of Deviance Table

Model: gaussian, link: identity

Response: ApprovedLog

Terms added sequentially (first to last)

Df Deviance Resid. Df Resid. Dev Pr(>Chi)

NULL 357 79.777

LoanPayoffPeriodInMonths 1 4.6170 356 75.160 3.610e-07 \*\*\*

RequestedAmount 1 0.3139 355 74.846 0.184580

CoApplicant 2 1.4468 353 73.399 0.017301 \*

YearsAtCurrentEmployer 4 1.4122 349 71.987 0.094560 .

Age 1 0.3136 348 71.673 0.184809

TypeOfCurrentEmployment 3 0.5446 345 71.128 0.383312

CheckingAccountBalance 3 7.1547 342 63.974 1.003e-08 \*\*\*

DebtsPaid 1 1.5998 341 62.374 0.002741 \*\*

SavingsAccountBalance 4 2.2601 337 60.114 0.012980 \*

CurrentOpenLoanApplications 1 0.1426 336 59.971 0.371132

InterestRate 1 0.1669 335 59.804 0.333379

LoanReason 5 1.4707 330 58.334 0.143105

NumberOfDependantsIncludingSelf 1 0.0697 329 58.264 0.531856

RentOrOwnHome 2 0.0748 327 58.189 0.810721

YearsInCurrentResidence 1 0.0587 326 58.130 0.566236

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The difference between the null deviance and the residual deviance shows how our model is doing against the null model (a model with only the intercept). The wider this gap, the better. Analyzing the table, we can see that there is not much significant drop in deviance when adding each variable one at a time. A large p-value here indicates that the model without the variable explains more or less the same amount of variation.

So we can see that the following variables are not of much importance

* CurrentOpenLoanApplications
* InterestRate
* LoanReason
* NumberOfDependantsIncludingSelf
* RentOrOwnHome
* YearsInCurrentResidence

1. **Check for Co-Linearity using VIF**

We can check for multi-collinearity to check if there is linear relationship between the features. I used VIF function from ‘fmsb’ package to check for co-linear relationship. Based on the analysis, we find the following variables are having co-linear relationship between the predictors.

* LoanReason
* CoApplicant
* YearsAtCurrentEmployer
* RentOrOwnHome
* TypeOfCurrentEmployment
* CheckingAccountBalance
* DebtsPaid
* SavingsAccountBalance
* TypeOfCurrentEmployment

So, for testing and model analysis, I designed a model excluding the above features and analyzed across some of the top Machine learning tools for Classification and found some interesting data prediction which is updated in Table 5a

1. **Choosing Best Models using STEP**

I try to choose the best model based on AIC value in STEP wise procedure.

We can see below based on the final AIC value which is the lowest and after that there is no improvement.

**step <- stepAIC(glmmodel1, direction="both")**

**Step: AIC=412.95**

ApprovedLog ~ LoanPayoffPeriodInMonths + RequestedAmount + CoApplicant + CheckingAccountBalance + DebtsPaid + SavingsAccountBalance

Df Deviance AIC

<none> 61.433 412.95

- RequestedAmount 1 61.871 413.50

+ CurrentOpenLoanApplications 1 61.291 414.13

+ NumberOfDependantsIncludingSelf 1 61.347 414.46

+ InterestRate 1 61.351 414.48

+ YearsInCurrentResidence 1 61.367 414.57

+ LoanReason 5 60.035 414.72

+ Age 1 61.409 414.81

+ TypeOfCurrentEmployment 3 60.946 416.10

+ RentOrOwnHome 2 61.347 416.45

+ YearsAtCurrentEmployer 4 60.733 416.85

- SavingsAccountBalance 4 63.757 418.25

- LoanPayoffPeriodInMonths 1 62.761 418.61

- CoApplicant 2 63.399 420.23

- DebtsPaid 1 63.477 422.67

- CheckingAccountBalance 3 66.449 435.05

So the final model feature comes as follows

* LoanPayoffPeriodInMonths
* RequestedAmount
* CoApplicant
* CheckingAccountBalance
* DebtsPaid
* SavingsAccountBalance

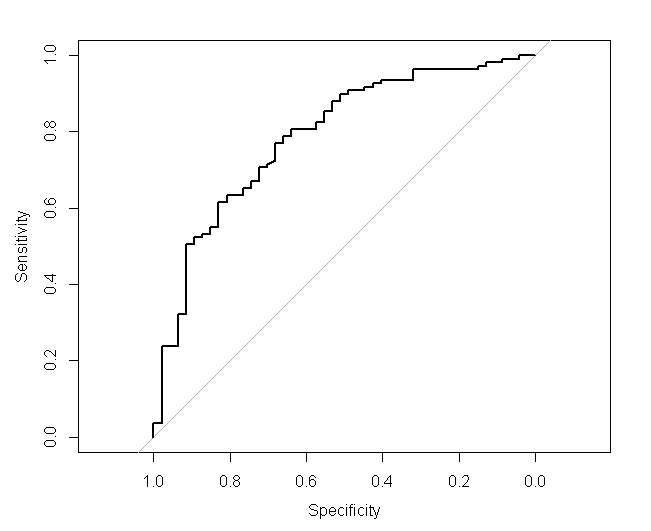
1. **Prediction Models using different Machine Learning tools**

I tried to design, evaluate and validate analytical models using different machine learning tools for predicting logical output. I used the following ML algorithms to predict Loan approval.

**a. Prediction using GBM (Gradient boost model)**

**All Features**: I am evaluating the accuracy of the model using roc curve and used all the features in the data to GBM function. The following graph shows the ROC curve for gbm.

**Area under the curve: 0.7881**



**Confusion Matrix for GBM using ALL features:**

**Reference**

**Prediction N Y**

**N 18 7**

**Y 29 102**

Accuracy : 0.7692

95% CI : (0.6951, 0.8328)

No Information Rate : 0.6987

P-Value [Acc > NIR] : 0.0311498

Kappa : 0.3677

Mcnemar's Test P-Value : 0.0004653

Sensitivity : 0.3830

Specificity : 0.9358

Pos Pred Value : 0.7200

Neg Pred Value : 0.7786

Prevalence : 0.3013

Detection Rate : 0.1154

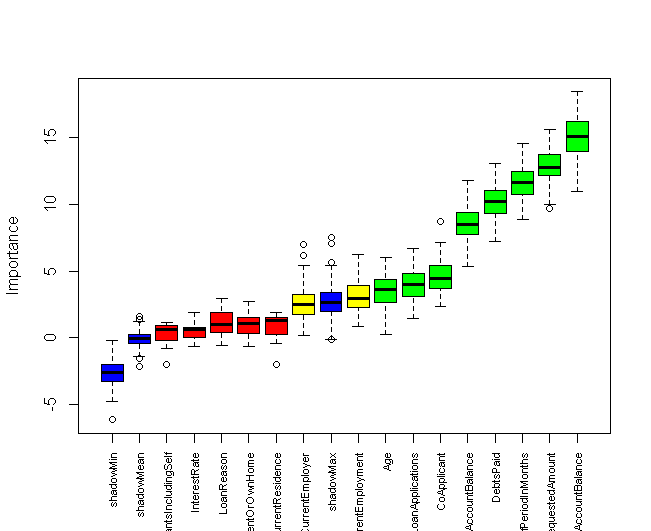
Detection Prevalence : 0.1603

Balanced Accuracy : 0.6594

'Positive' Class : N

**b. Model using Barota Ranking:**

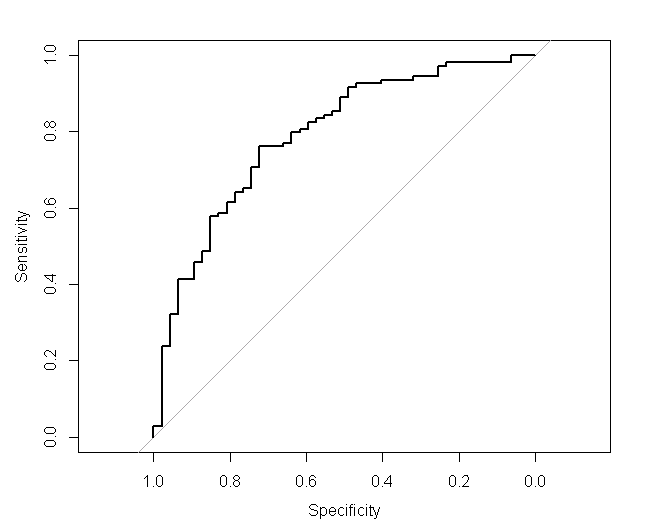
I am evaluating the accuracy of the model using roc curve and used the features that was selected from feature ranking using Barota. So, I used only the important features based on feature ranking. The below graph shows the important features in GREEN and un-important features in RED.

Feature Ranking using Barota 

* **CheckingAccountBalance**
* **RequestedAmount**
* **LoanPayoffPeriodInMonths**
* **DebtsPaid**
* **SavingsAccountBalance**
* **CoApplicant**
* **CurrentOpenLoanApplications**
* **Age**
* **TypeOfCurrentEmployment**
* **YearsAtCurrentEmployer**
* **YearsInCurrentResidence**
* **RentOrOwnHome**
* **LoanReason**
* **InterestRate**
* **NumberOfDependantsIncludingSelf**

The following graph shows the ROC curve for gbm using feature ranking.

**Area under the curve: 0.7904**



**Confusion Matrix for GBM using Barota Features:**

Reference

Prediction N Y

N 22 9

Y 25 100

Accuracy : 0.7821

95% CI : (0.709, 0.8441)

No Information Rate : 0.6987

P-Value [Acc > NIR] : 0.01276

Kappa : 0.4268

Mcnemar's Test P-Value : 0.01010

Sensitivity : 0.4681

Specificity : 0.9174

Pos Pred Value : 0.7097

Neg Pred Value : 0.8000

Prevalence : 0.3013

Detection Rate : 0.1410

Detection Prevalence : 0.1987

Balanced Accuracy : 0.6928

'Positive' Class : N

**c. Step Model Features:**

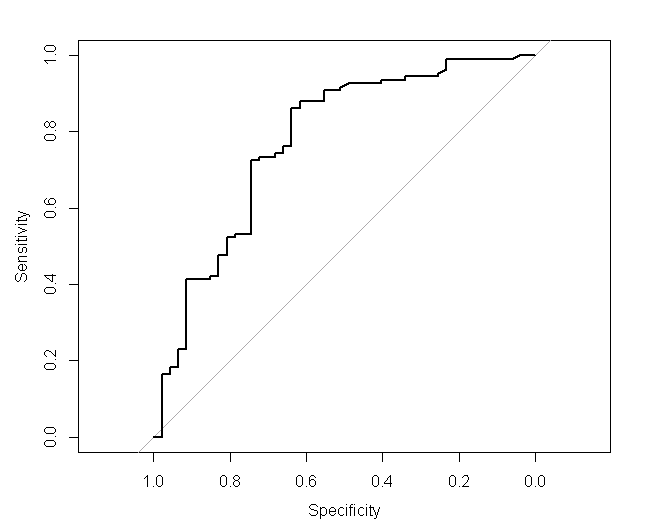
I am evaluating the accuracy of the model using ROC curve and used the features that was selected from using STEP based on AIC.

**Model Features:**

* LoanPayoffPeriodInMonths
* RequestedAmount
* CoApplicant
* CheckingAccountBalance
* DebtsPaid
* SavingsAccountBalance

The following graph shows the ROC curve for gbm using features from STEP.

**Area under the curve: 0.7748**



**Confusion Matrix for GBM using STEP(AIC) Features:**

Reference

Prediction N Y

N 24 13

Y 23 96

Accuracy : 0.7692

95% CI : (0.6951, 0.8328)

No Information Rate : 0.6987

P-Value [Acc > NIR] : 0.03115

Kappa : 0.4166

Mcnemar's Test P-Value : 0.13361

Sensitivity : 0.5106

Specificity : 0.8807

Pos Pred Value : 0.6486

Neg Pred Value : 0.8067

Prevalence : 0.3013

Detection Rate : 0.1538

Detection Prevalence : 0.2372

Balanced Accuracy : 0.6957

'Positive' Class : N

1. **ROC Comparison with Model Features and ML.**

The following table shows AUC value for different ML algorithms with different model features.

We can see that Logistic regression does better than other algorithms and it’s very interesting to see that removing model features using STEP and Barota doesn’t give any significant accuracy.

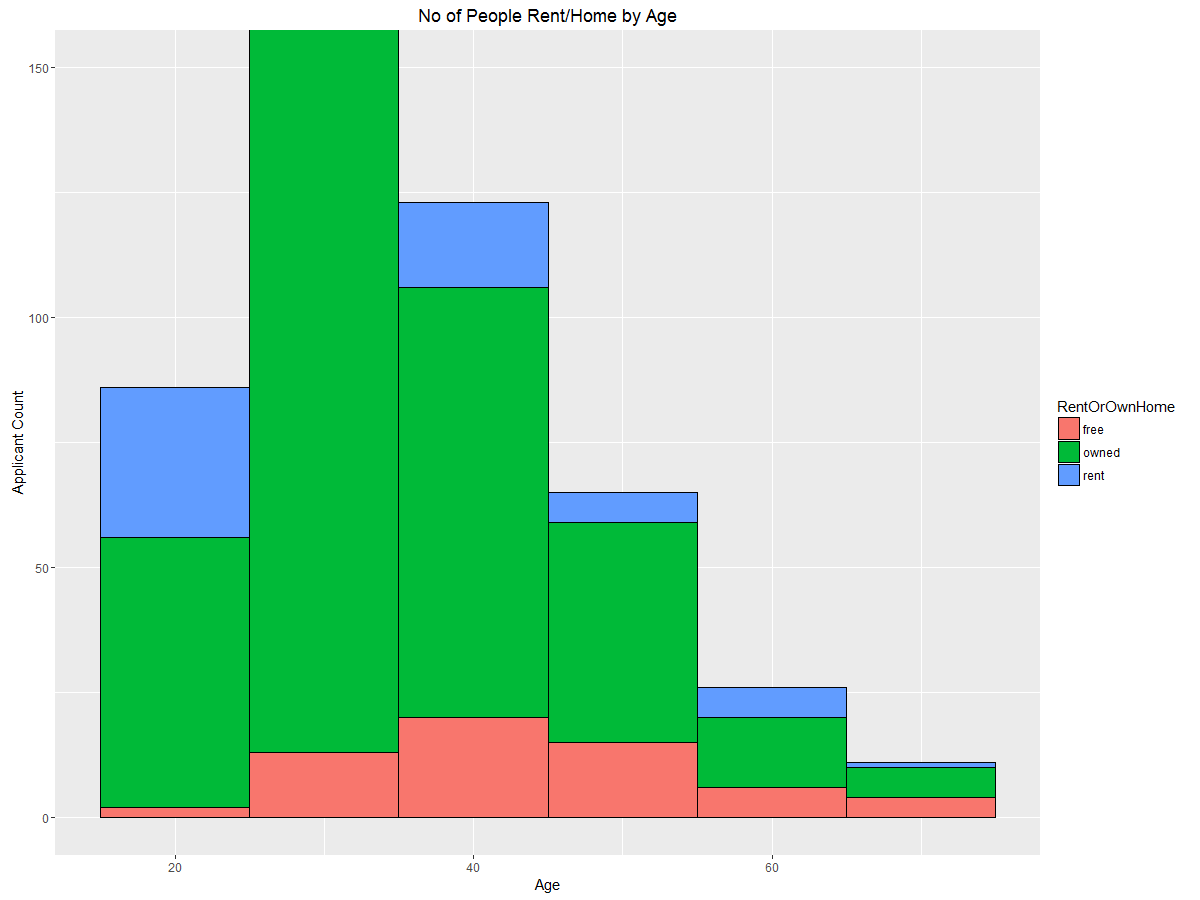
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ML / Features | All Features | Model Selection using STEP | Barota Feature Ranking | VIF Selection | ANOVA  Selection |
| GBM | 0.7863 | 0.7866 | 0.7742 | 0.6197 | 0.7753 |
| Logistic regression | 0.8198 | 0.8071 | 0.8097 | 0.633 | 0.8036 |
| Random Forest | 0.8172 | 0.7709 | 0.798 | 0.6173 | 0.7986 |
| XGboost | 0.7346 | 0.7346 | 0.7346 | 0.7346 | 0.7346 |

Table 5a.

**5.Visual Analysis of Data**

1. The below graph shows the number of people who have rented /owned home by age category.

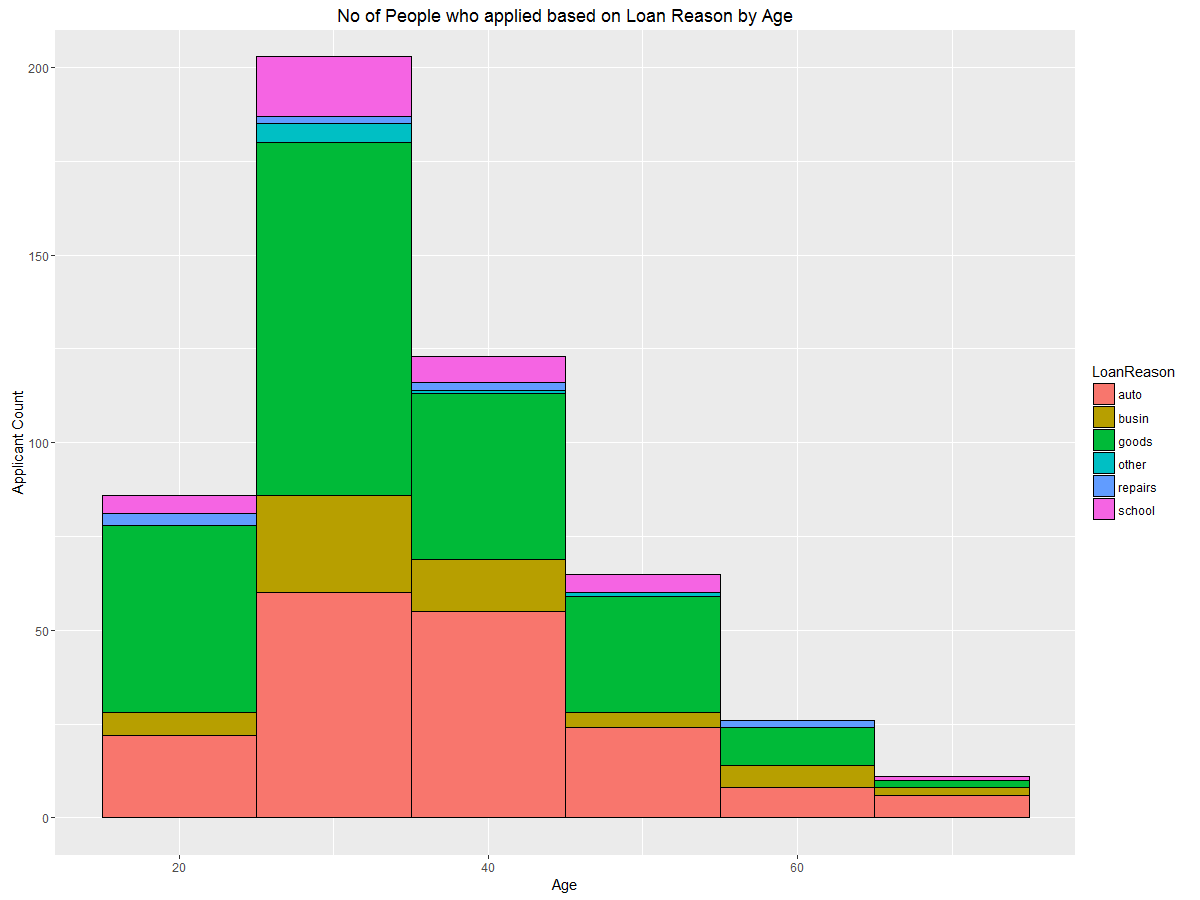
This graph shows that most % of the people do own home between the age of 30-45 yrs. compared to people who rent.



**Graph 1A.**

1. The below graph shows the number of people who have applied for loans based on the loan reason code by age category.

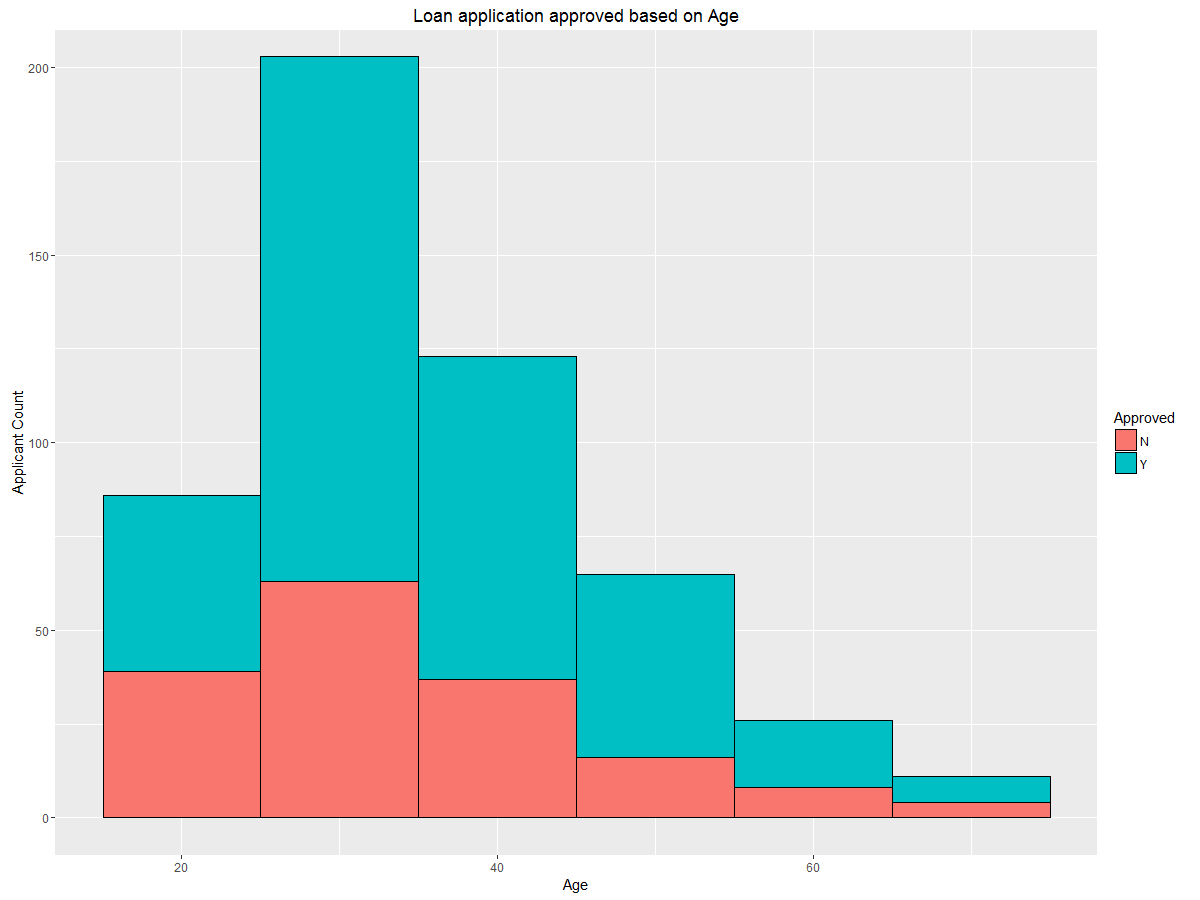
This graph shows that most % of the people have applied loans for Auto and Goods reason compared to school and repairs which makes sense comparing to statistical data.



**Graph 2A.**

1. The below graph shows the count of people who have applied for loans and got approved based on the age.

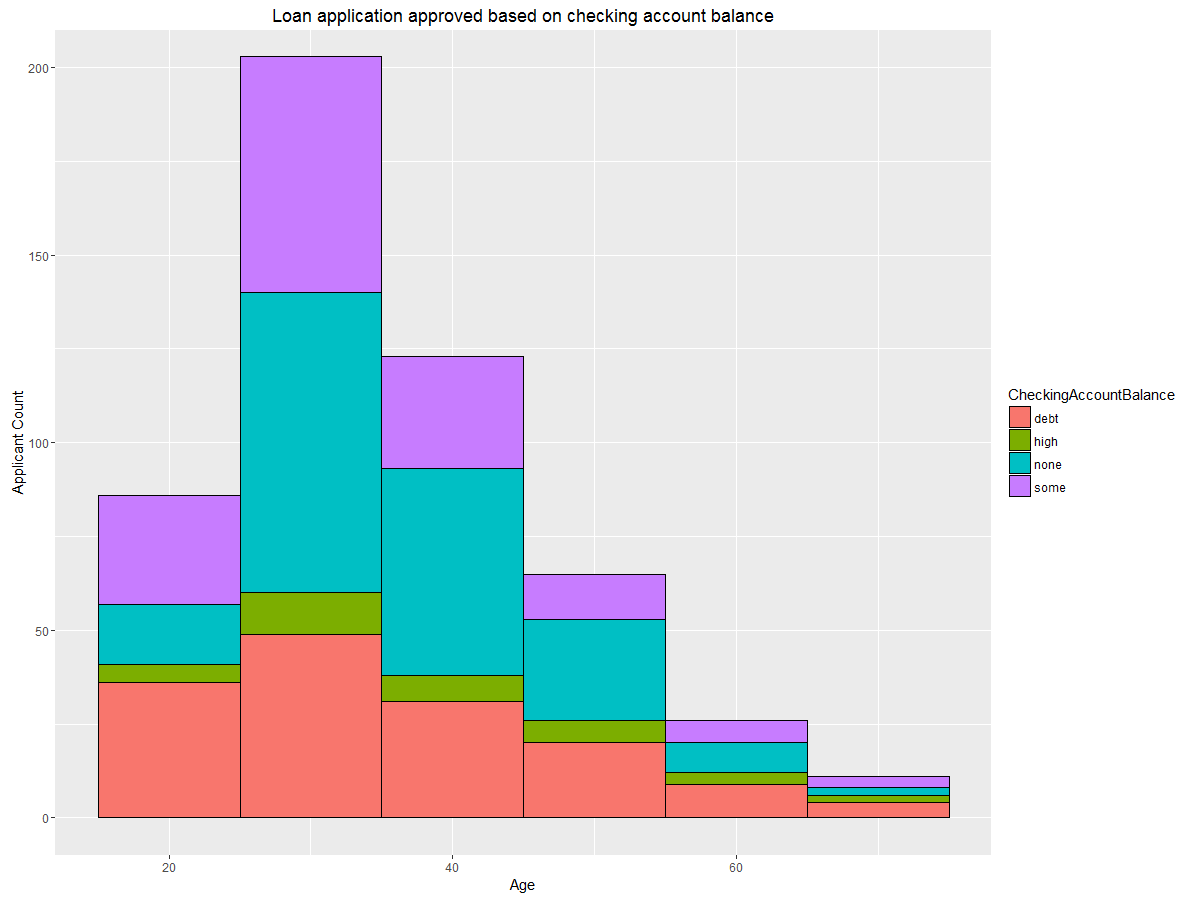
The graph shows number of loan applications in the middle age people between 30 and 50 yrs. and who got loan approved are more than old age application.



**Graph 3A.**

1. The below graph shows the number of people who have applied for loans and got approved based on the checking account balance.

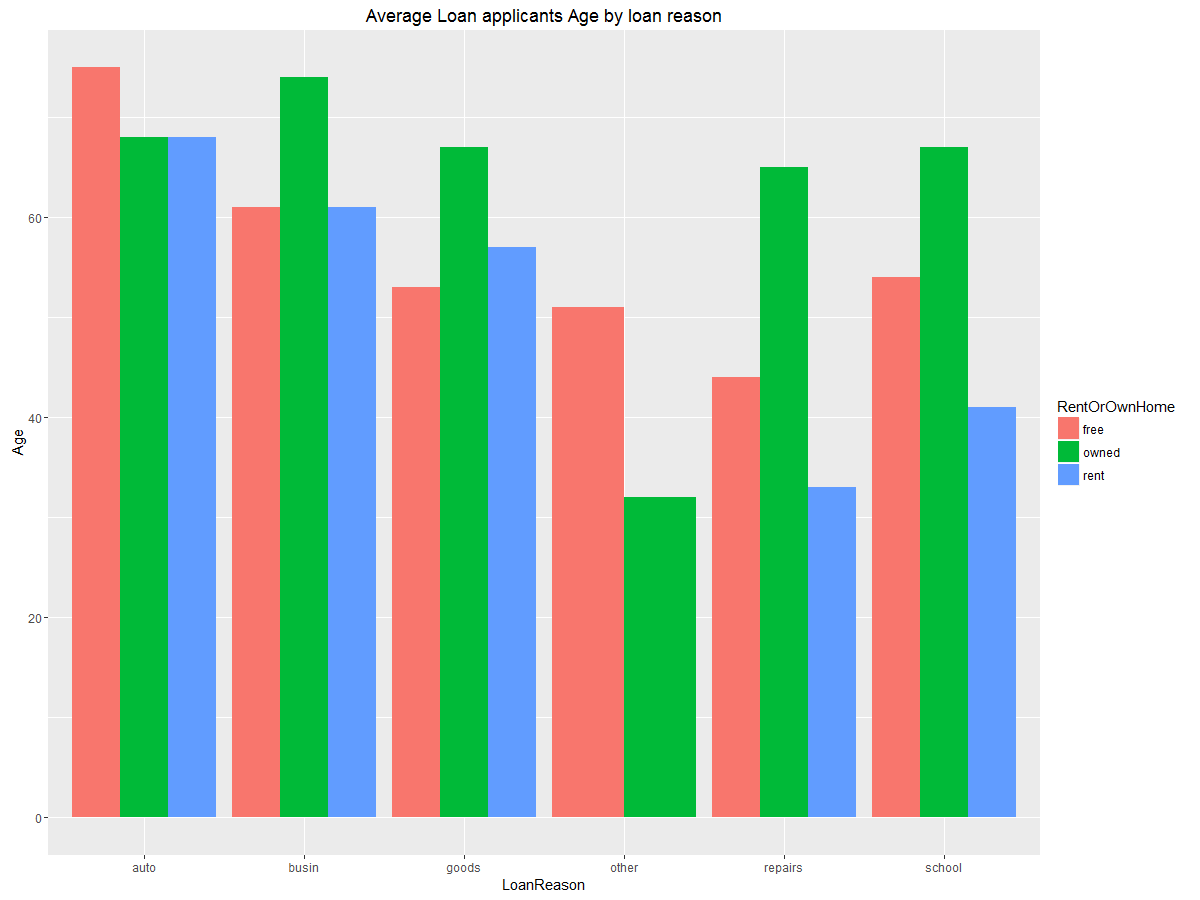
By comparing graph 4a and 3a shows that shows no of loan applications in the middle age people between 30 and 50 yrs and who got loan approved have none or little account balance, which is very interesting.



**Graph 4A.**

1. The below graph shows the average loan applicants age based on the loan reason.

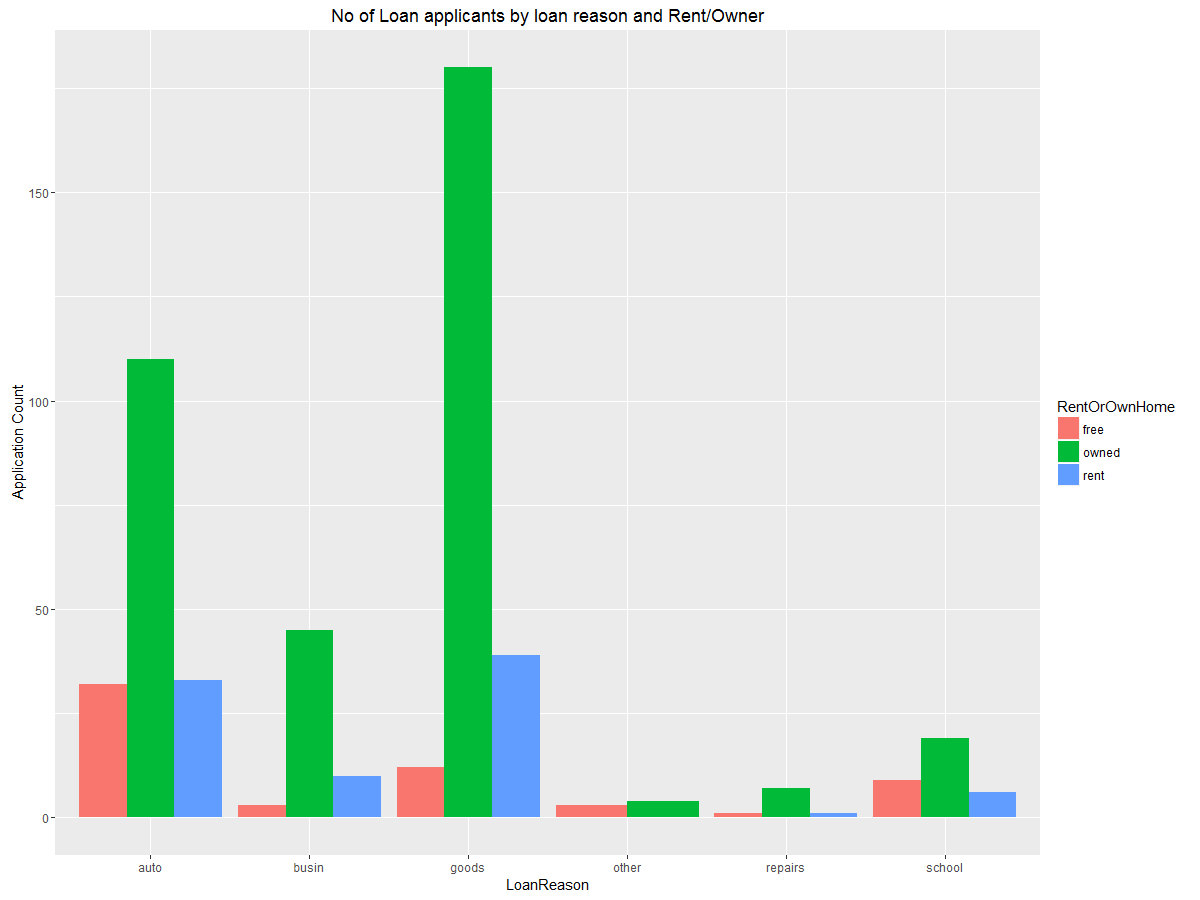
By comparing graph 2a and 5a shows that most of the loan applicants are in the auto and goods business and have average age between 45 to 60.



**Graph 5A.**

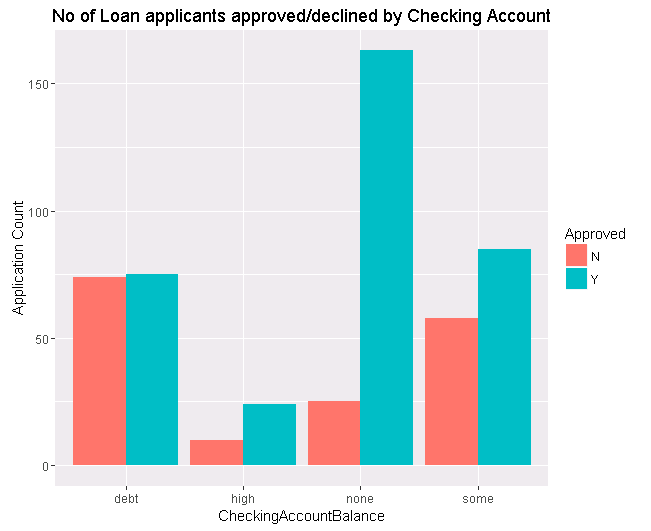
1. The below graph shows the no. of loan applicants age based on the loan Reason and Home ownership.

The graph also shows that the applicants who have Auto and Goods business and who also own Home tops the list than other category.



**Graph 6A.**

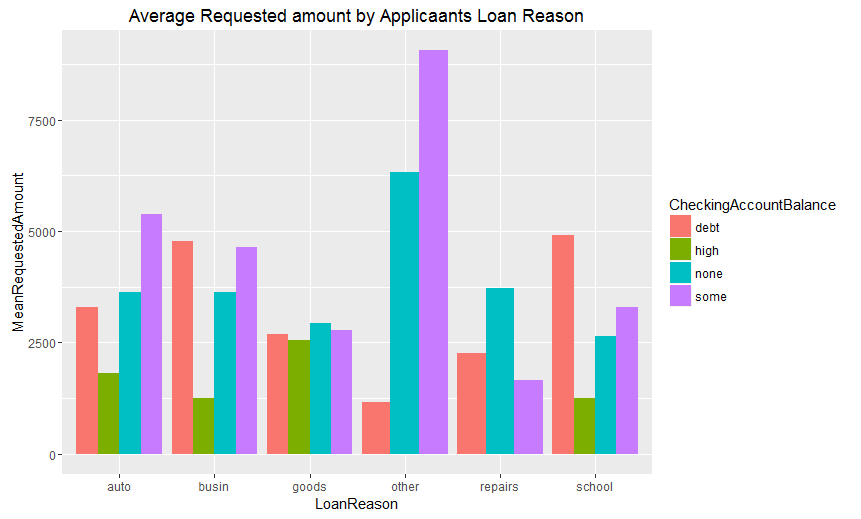
1. The below graph shows the number of loan applicants who got their loan approved based on their checking account balance. The below graph concedes with the graph 4a where loan applicants with some or none account balance have the most loan approved.



**Graph 7a**

1. The below graph shows the Mean Requested amount by Loan Reason and checking account balance.

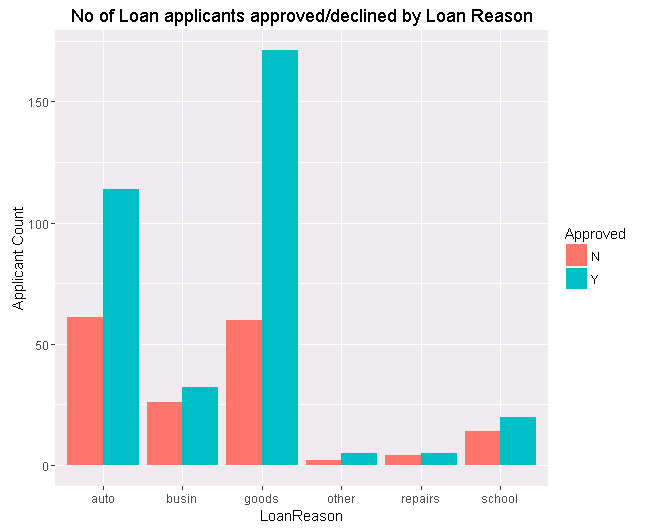
The below graph also shows Mean Requesting Amount is higher for applicant’s who do ‘Other’ business reason and only ‘Some Account Balance’ are higher than other categories.



**Graph 8A.**

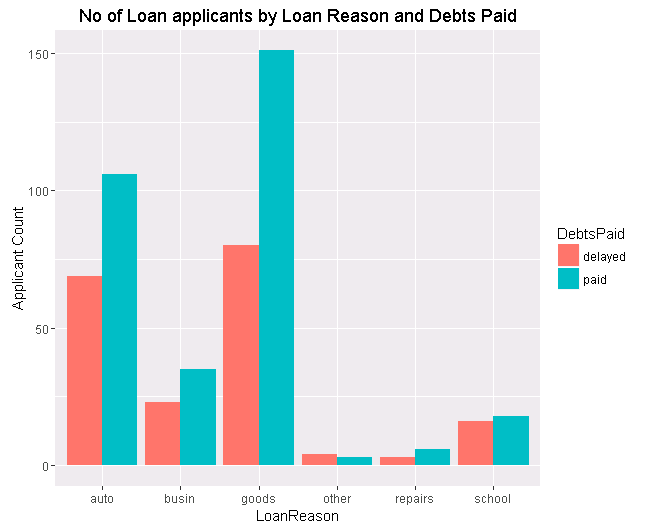
1. The below graph shows the loan applicant’s approved or declined based on the Loan Reason.

The below graph also shows applicants who have goods and auto business have higher ratio of loan approved than other business.



**Graph 9A.**

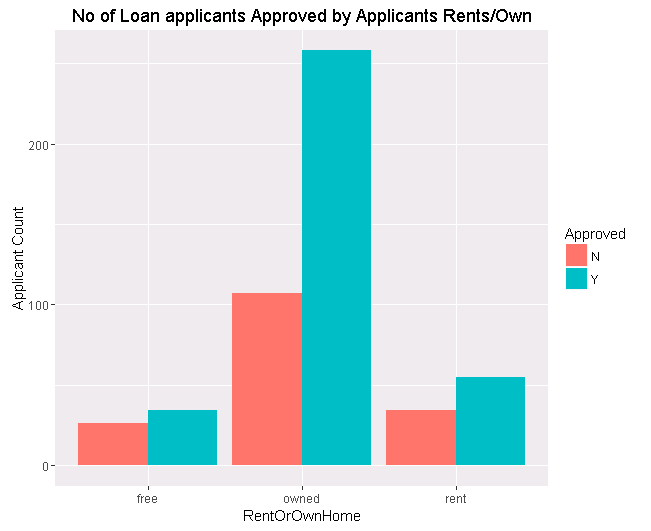
1. The below graph shows the count of loan applicant’s approved or declined based on the Loan Reason and Debt Paid. The below graph also shows applicants who have good debt payment have more loan approval rate in goods and auto business.



**Graph 10A.**

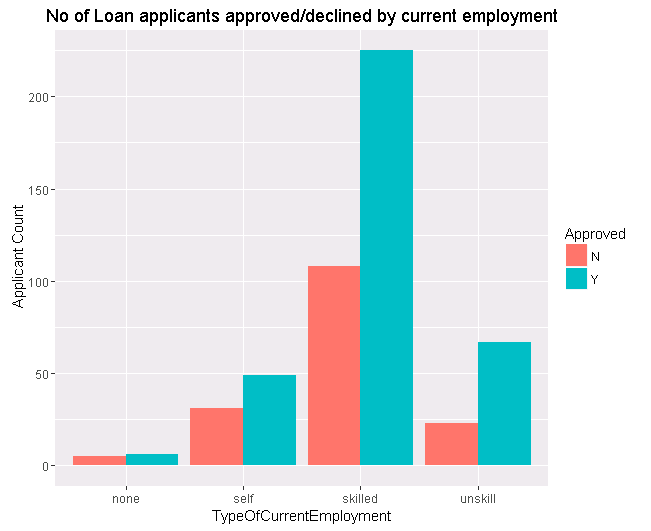
1. The below graph shows the count of loan applicant’s approved or declined based on the Home Ownership.

The below graph also shows applicants who have good home ownership have more loan approval rate. This comparing with Graph 1A shows data that home ownership plays important role in Model selection but the model selection doesn’t include RentOrOwnHome feature which is very interesting.



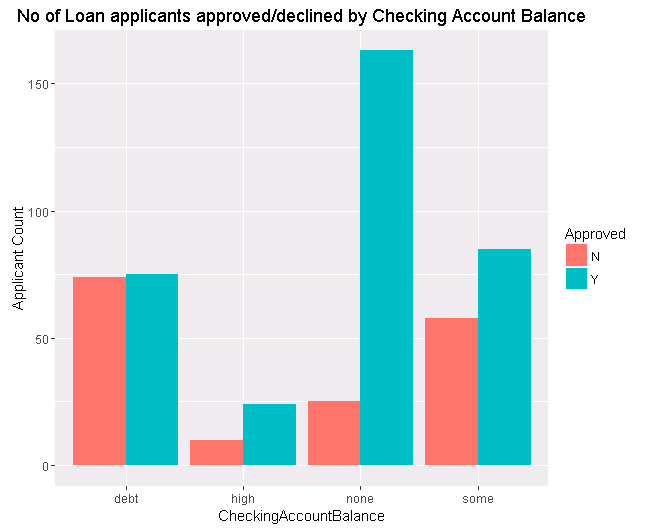
**Graph 11A.**

1. The below graph shows the count of loan applicant’s approved or declined based on the Employers skill. The graph also shows skilled applicants have more loan approval compared to unskilled and self-workers.



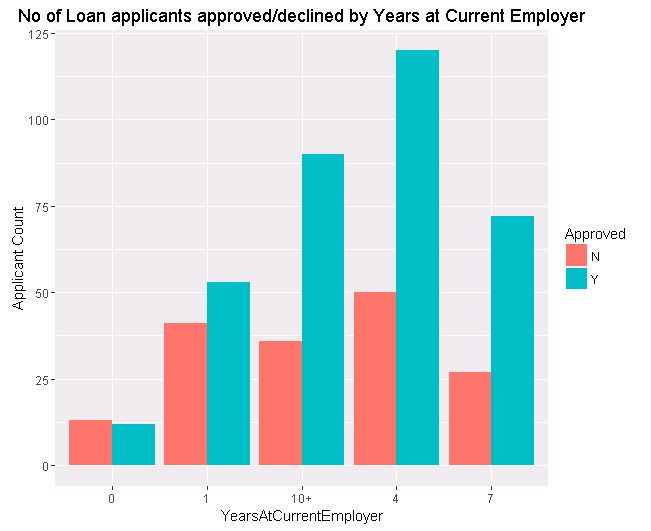
**Graph 12A.**

1. The below graph shows the count of loan applicant’s approved or declined based on applicants Checking Account Balance. Based on the Barota feature ranking, we also know that ‘CheckingAccountBalance’ which remains the top feature in the rank. This graph confirms that Checking account balance plays an important role in Loan approval.



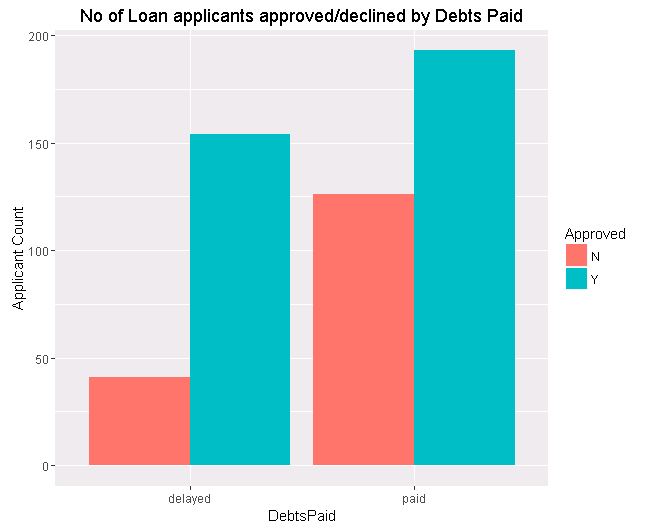
**Graph 13A.**

1. The below graph shows the count of loan applicant’s approved or declined based on number of years at current employer. Based on the below data, we see that there is not much relation to the loan approval. Because applicants with even 10 yrs. plus are also not approved. This conforms to the Ranking that YearsAtCurrentEmployer is not a significant feature in Loan Approval decision.



**Graph 14A.**

1. The below graph shows the count of loan applicant’s approved or declined based on Debt paid. Based on the below data, we see that their loan % approval is higher for applicants who have paid than delayed. This conforms to the Ranking that DebtsPaid is a significant feature in Loan Approval decision.



**Graph 15A.**

1. **Accuracy with Different ML Tools**

|  |  |  |
| --- | --- | --- |
| **Machine Learning** | **Features** | **Accuracy** |
| Gradient Boost Model | All | 76.92% |
| Logistic regression Model | All | 77.56% |
| Random Forest Model | All | 77.56% |
| Extreme Boost GBM | All | 76.28% |

Based on the above model analysis and prediction accuracy we can see that there is not much significance between the top classification ML tools, almost all of them give more or less the same accuracy.

**7. Conclusion**

Based on the model analysis and validation, we can conclude that our Model is valid with all the features included because of the fact that we tried tuning model parameters with different techniques from caret package and tried choosing model parameters from Barota feature ranking and STEP AIC procedure. And we found that we don’t see any significant difference in the AUC value and also we saw the predictions from different classification algorithms and they all give around the same accuracy.

Source Code:

