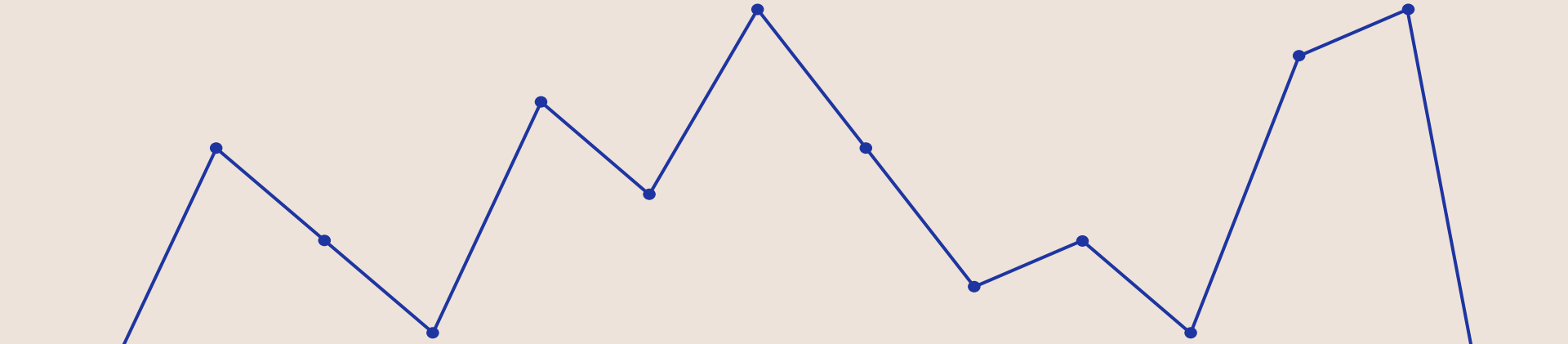
Group 5:
German Credit Data Analysis

By Kamerin Vesajd & Chiayu Tu

Table of Contents:

Exploratory Analysis of Data



Data Overview

Imported 1000 observations of German bank applicants profile data

---

Predictors:

20 Columns:

- *6 Numeric Columns*
- *14 Categorical Columns*

48 Columns:

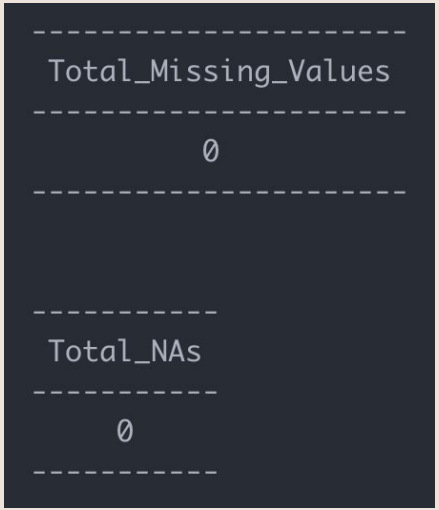
- ***After Dummy Variables:*** *41 Categorical Columns*

Response: Default

Account holder failed to repay debts or financial obligations

- **Default = 0 (Good)**
- **Default = 1 (Bad)**

No Missing/NA Values:



Objective:

- To minimize the bank's potential losses by evaluating the demographic and socio-economic characteristics of loan applicants in order to determine whether to approve or reject their loan application.

- Determine the best logistic regression model to improve risk management and predictions

Numerical Predictors:

- **Duration (Months):** Length of time the individual is expected to take to repay the credit.
- **Amount:** Credit amount requested (DM)
- **Residence:** Length of time lived in current residence
- **Age:** Age of applicant by years
- **Cards:** Numbers of existing credit cards at the bank
- **Liable:** Number of people liable to provide maintenance for

Categorical Predictors:

- **Checking account (4 Levels)**
- **Credit history (5 Levels)**
- **Purpose (4 Levels)**
- **Savings account/bonds (5 Levels)**
- **Employment Length (5 Levels)**
- **Personal status and sex (4 Levels)**
- **Installment (4 Levels)**
- **Property (4 Levels)**
- **Job (4 Levels)**
- . . .

Predictor List:

(Numerical:Int — Categorical:Factor with Levels)

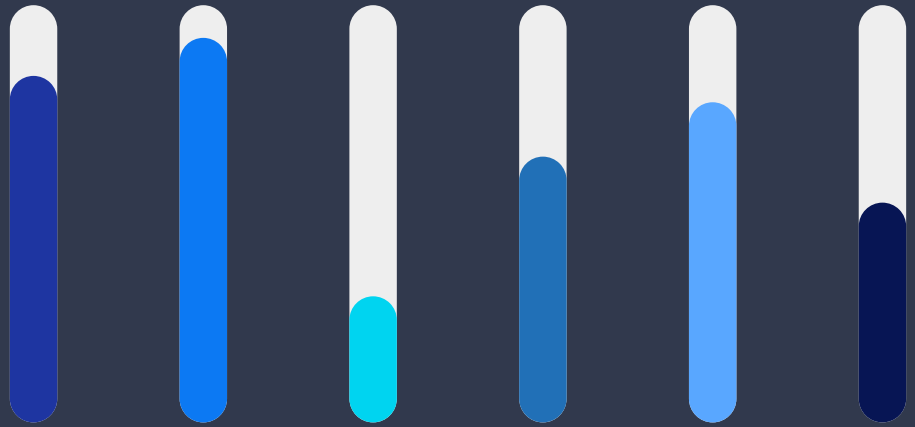
```
'data.frame':   1000 obs. of  21 variables:
 $ Default        : Factor w/ 2 levels "0","1": 1 2 1 1 2 1 1 1 1 2 ...
 $ checkingstatus1: Factor w/ 4 levels "A11","A12","A13",..: 1 2 4 1 1 4 4 2 4 2 ...
 $ duration       : int  6 48 12 42 24 36 24 36 12 30 ...
 $ history        : Factor w/ 5 levels "A30","A31","A32",..: 5 3 5 3 4 3 3 3 3 5 ...
 $ purpose        : Factor w/ 10 levels "A40","A41","A410",..: 5 5 8 4 1 8 4 2 5 1 ...
 $ amount         : int  1169 5951 2096 7882 4870 9055 2835 6948 3059 5234 ...
 $ savings        : Factor w/ 5 levels "A61","A62","A63",..: 5 1 1 1 1 5 3 1 4 1 ...
 $ employ         : Factor w/ 5 levels "A71","A72","A73",..: 5 3 4 4 3 3 5 3 4 1 ...
 $ installment    : int  4 2 2 2 3 2 3 2 2 4 ...
 $ status         : Factor w/ 4 levels "A91","A92","A93",..: 3 2 3 3 3 3 3 3 1 4 ...
 $ others         : Factor w/ 3 levels "A101","A102",..: 1 1 1 3 1 1 1 1 1 1 ...
 $ residence      : int  4 2 3 4 4 4 4 2 4 2 ...
 $ property       : Factor w/ 4 levels "A121","A122",..: 1 1 1 2 4 4 2 3 1 3 ...
 $ age            : int  67 22 49 45 53 35 53 35 61 28 ...
 $ otherplans     : Factor w/ 3 levels "A141","A142",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ housing        : Factor w/ 3 levels "A151","A152",..: 2 2 2 3 3 3 2 1 2 2 ...
 $ cards          : int  2 1 1 1 2 1 1 1 1 2 ...
 $ job            : Factor w/ 4 levels "A171","A172",..: 3 3 2 3 3 2 3 4 2 4 ...
 $ liable         : int  1 1 2 2 2 2 1 1 1 1 ...
 $ tele           : Factor w/ 2 levels "A191","A192": 2 1 1 1 1 2 1 2 1 1 ...
 $ foreign        : Factor w/ 2 levels "A201","A202": 1 1 1 1 1 1 1 1 1 1 ...
```

Attribute Appendix

```
Default: 0 (no) and 1 (yes)

Attribute 1: (qualitative) Status of existing checking account
   A11 : ... < 0 DM
   A12 : 0 <= ... < 200 DM
   A13 : ... >= 200 DM / salary assignments for at least 1 year
   A14 : no checking account

Attribute 2: (numerical)      Duration in month

Attribute 3: (qualitative)       Credit history
   A30 : no credits taken/ all credits paid back duly
   A31 : all credits at this bank paid back duly
   A32 : existing credits paid back duly till now
   A33 : delay in paying off in the past
   A34 : critical account/ other credits existing (not at this bank)

Attribute 4: (qualitative)            Purpose
   A40 : car (new)
   A41 : car (used)
   A42 : furniture/equipment
   A43 : radio/television
   A44 : domestic appliances
   A45 : repairs
   A46 : education
   A47 : (vacation - does not exist?)
   A48 : retraining
   A49 : business
   A410 : others

Attribute 5: (numerical)      Credit amount

Attibute 6: (qualitative)      Savings account/bonds
   A61 : ... < 100 DM
   A62 : 100 <= ... < 500 DM
   A63 : 500 <= ... < 1000 DM
   A64 : .. >= 1000 DM
   A65 : unknown/ no savings account

Attribute 7: (qualitative)            Present employment since
   A71 : unemployed
   A72 : ... < 1 year
   A73 : 1 <= ... < 4 years
   A74 : 4 <= ... < 7 years
   A75 : .. >= 7 years

Attribute 8: (numerical)      Installment rate in percentage of disposable
income

Attribute 9: (qualitative)            Personal status and sex
   A91 : male : divorced/separated
   A92 : female : divorced/separated/married
   A93 : male : single
   A94 : male : married/widowed
   A95 : female : single
```

```
   A102 : co-applicant
   A103 : guarantor

Attribute 11: (numerical)        Present residence since

Attribute 12: (qualitative)      Property
   A121 : real estate
   A122 : if not A121 : building society savings agreement/ life insurance
   A123 : if not A121/A122 : car or other, not in attribute 6
   A124 : unknown / no property

Attribute 13: (numerical)        Age in years

Attribute 14: (qualitative)      Other installment plans
   A141 : bank
   A142 : stores
   A143 : none

Attribute 15: (qualitative)      Housing
   A151 : rent
   A152 : own
   A153 : for free

Attribute 16: (numerical)        Number of existing credits at this bank

Attribute 17: (qualitative)      Job
   A171 : unemployed/ unskilled - non-resident
   A172 : unskilled - resident
   A173 : skilled employee / official
   A174 : management/ self-employed/
   highly qualified employee/ officer

Attribute 18: (numerical)  Number of people being liable to provide
maintenance for

Attribute 19: (qualitative)      Telephone
   A191 : none
   A192 : yes, registered under the customer name

Attribute 20: (qualitative)      foreign worker
   A201 : yes
   A202 : no
```

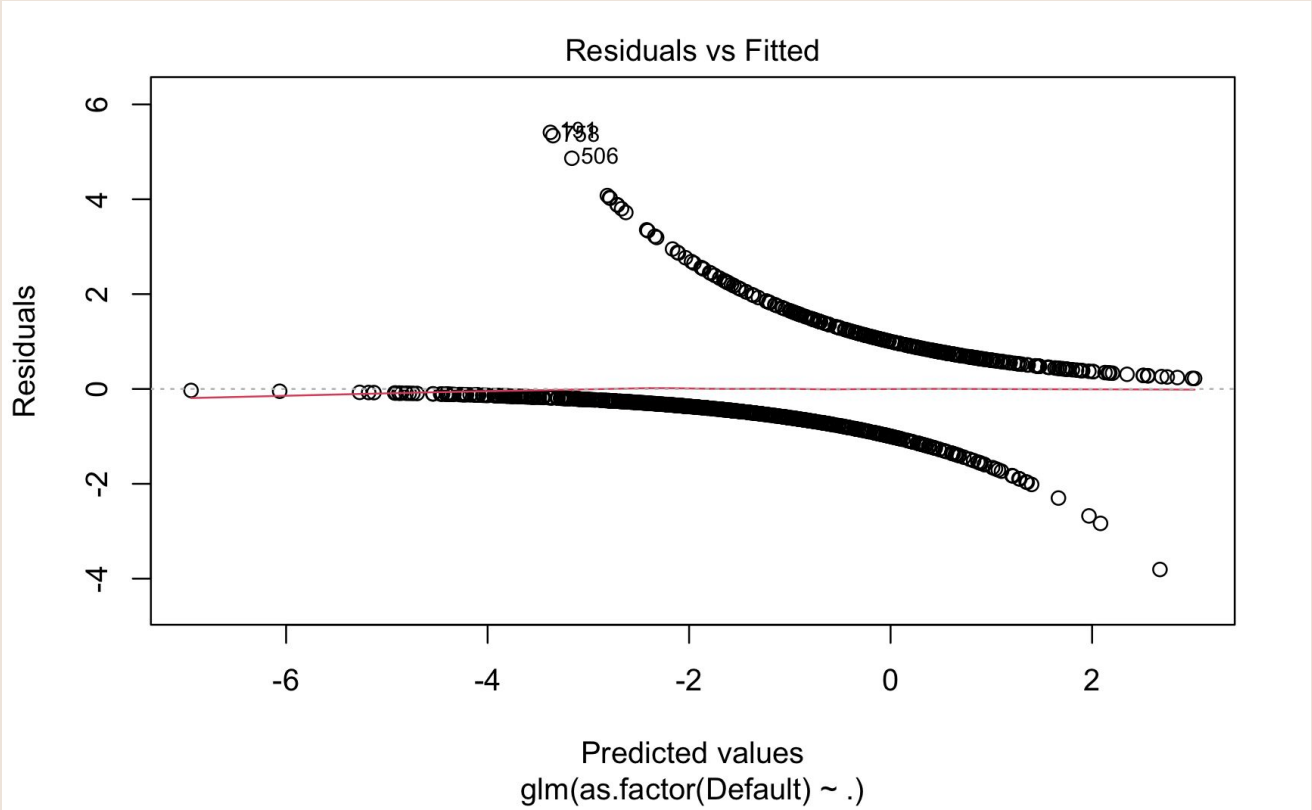
Observing Unusual Observations



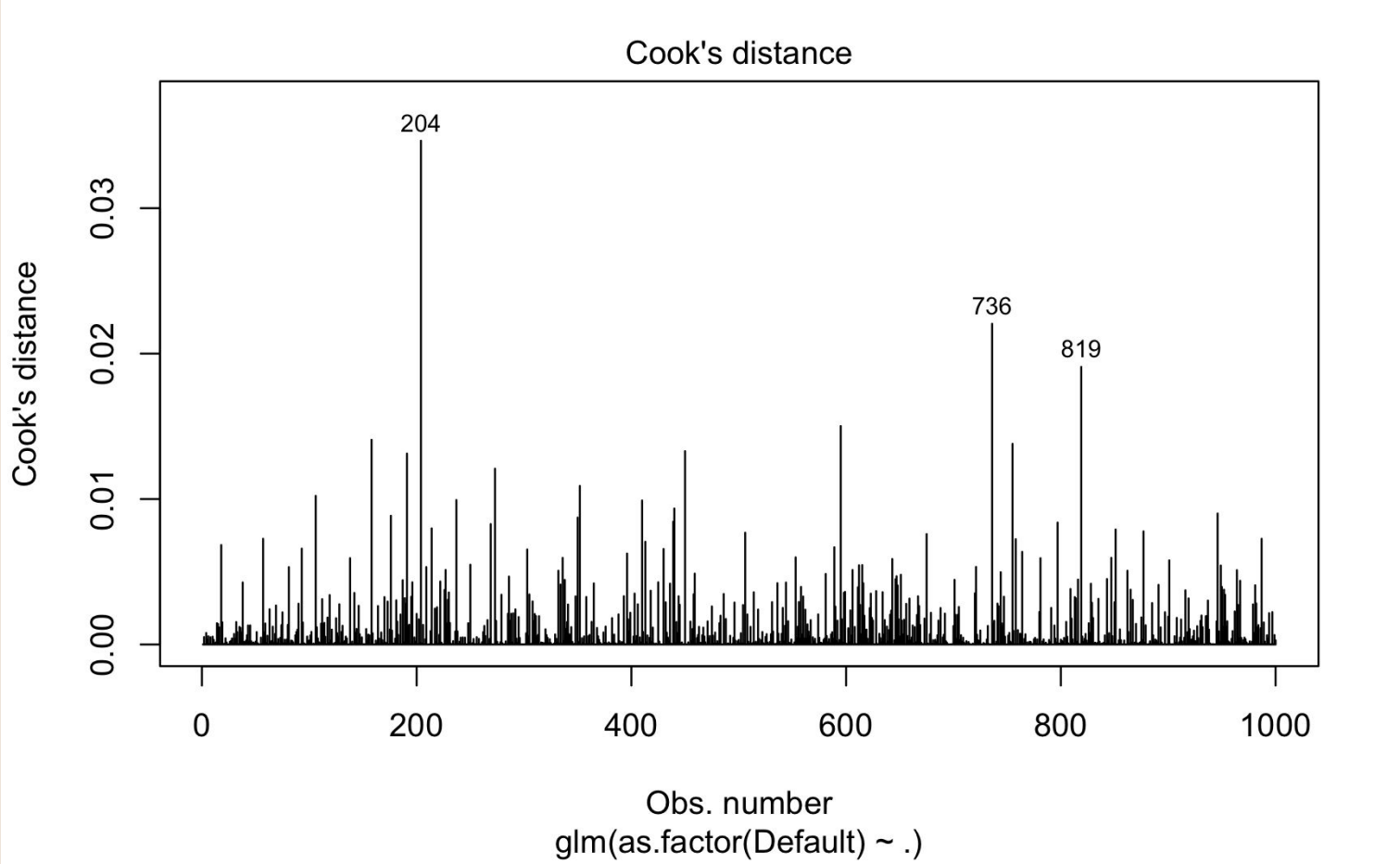
General Logistic Regression Model:

glm(Default ~ . , data=Credit, family= “binomial”)

Residual Plot of GLM Model:

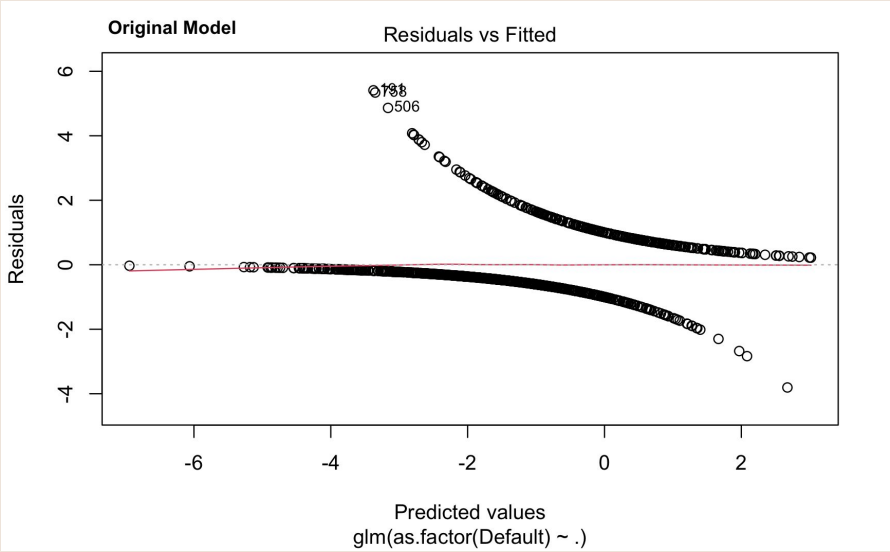


Influential Points’ Highest Leverages



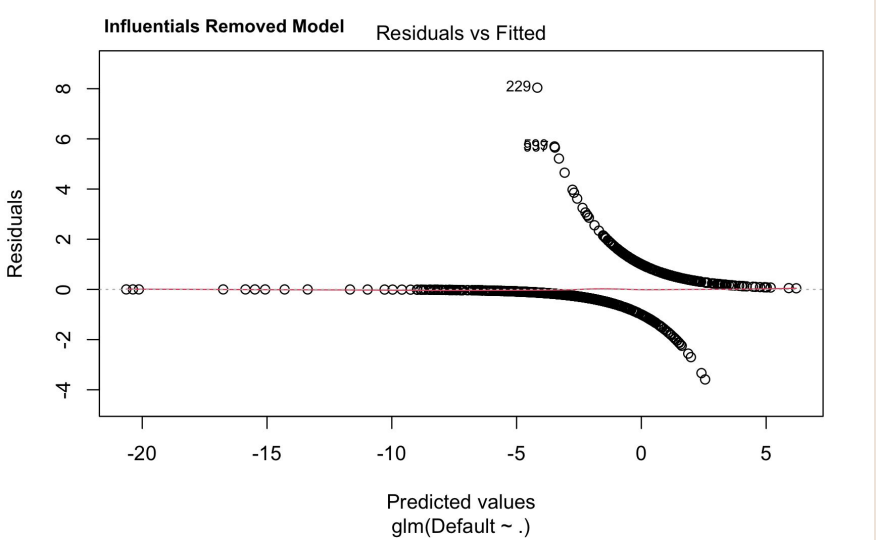
General Logistic Regression Model:
glm(Default ~ . , data=Credit, …)

Original Model:



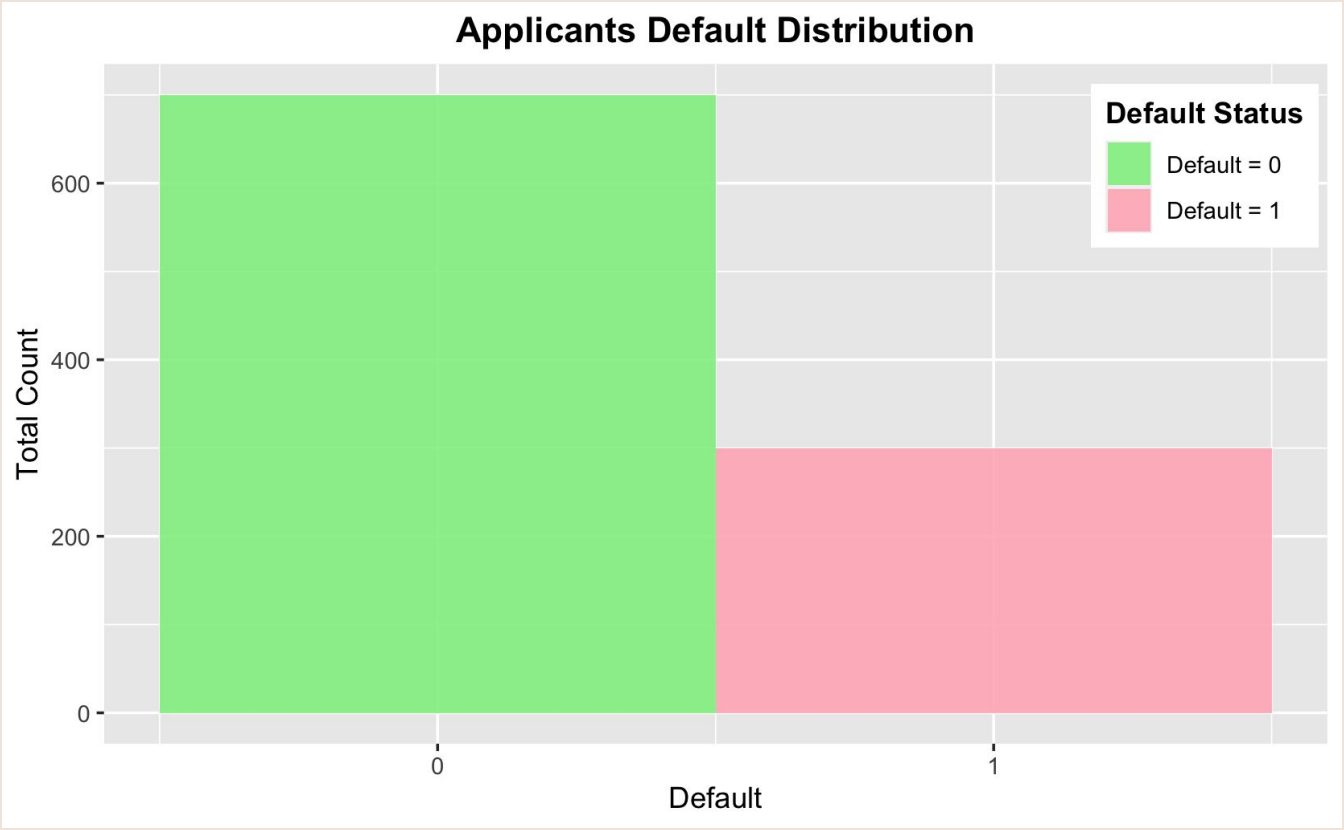
Influentials Removed Model:
glm(Default ~ . , data=Credit[-influential,] ,...)

New Model:

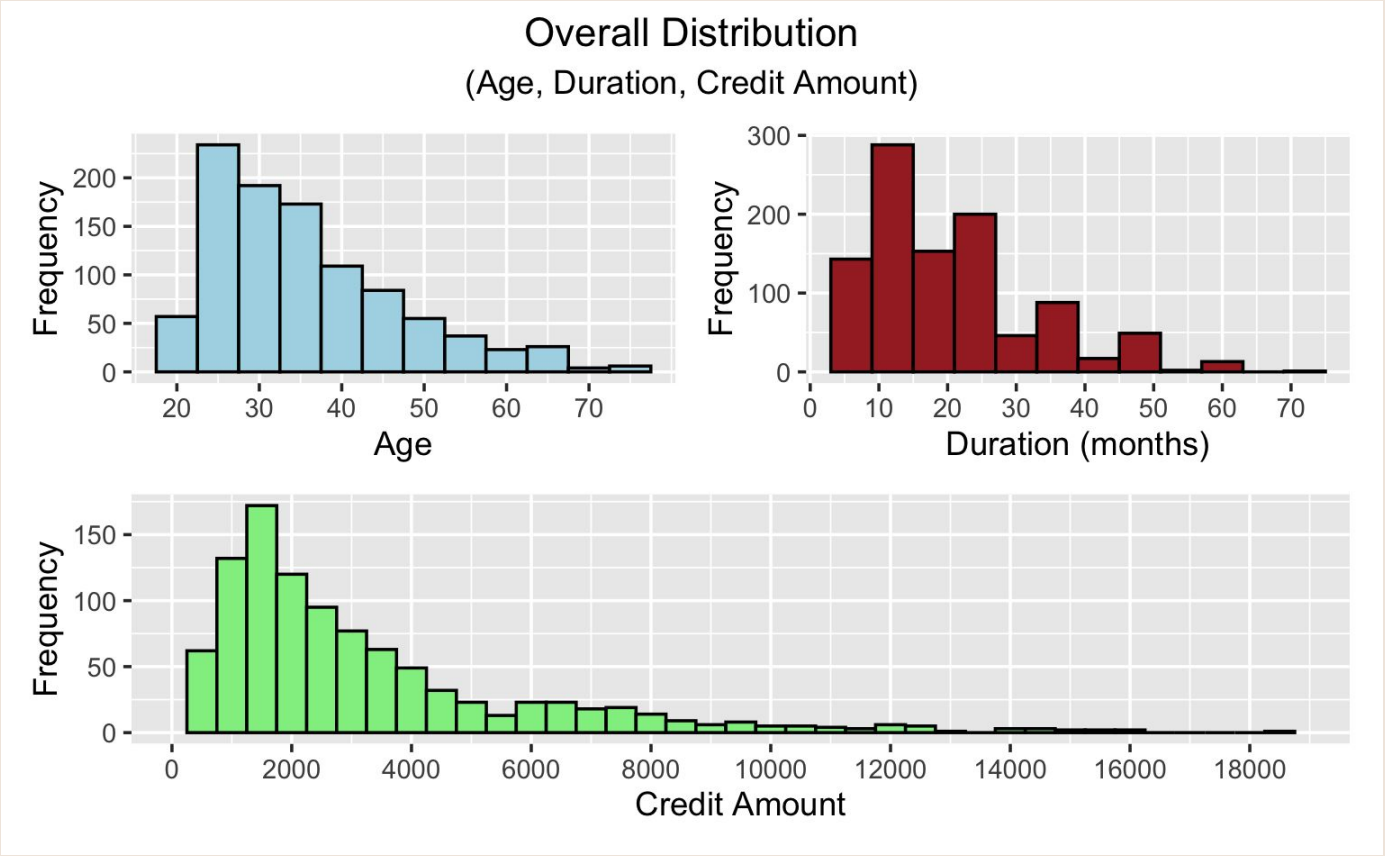


Examining Data Distribution & Relationship

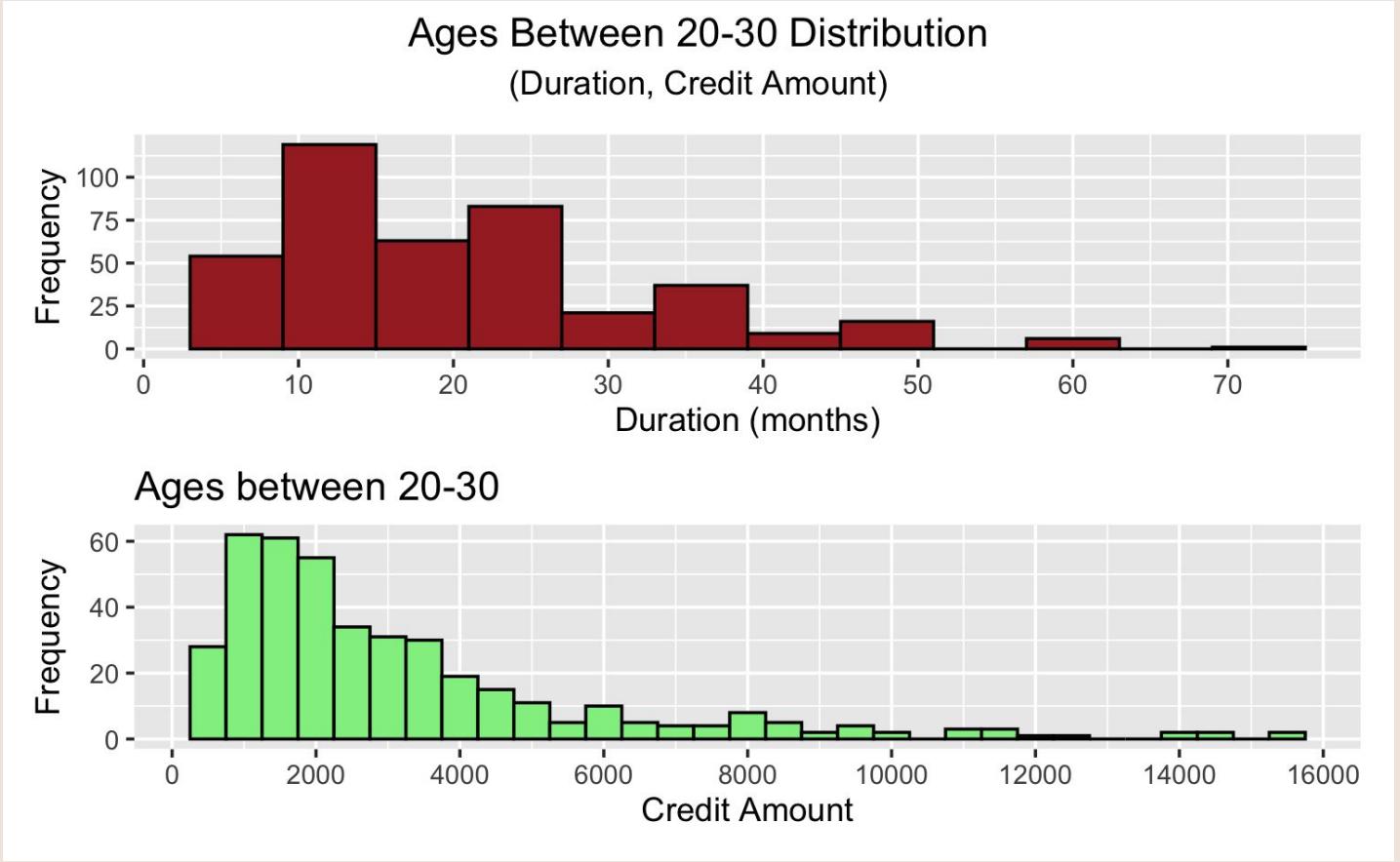
Default Response Frequency



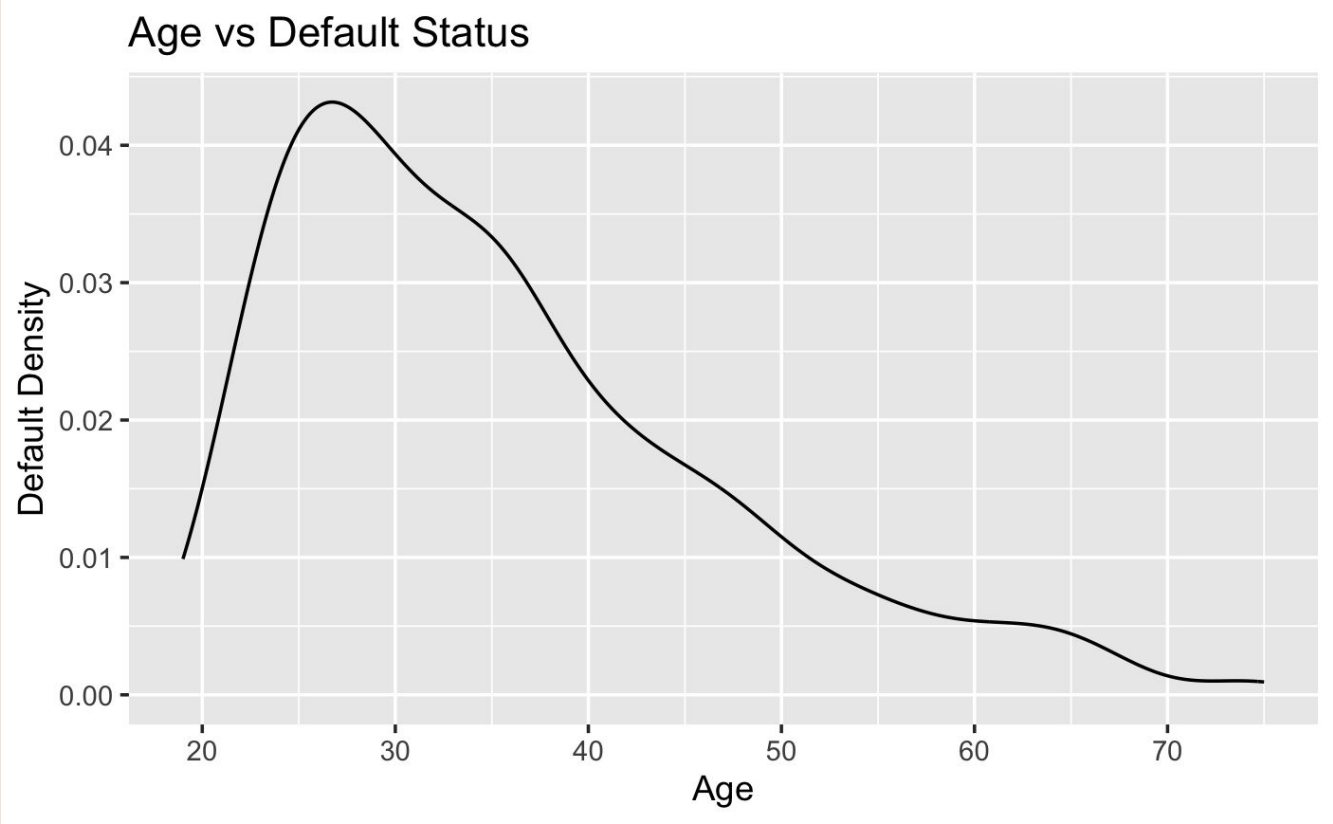
Age, Duration (Months), Credit Amount



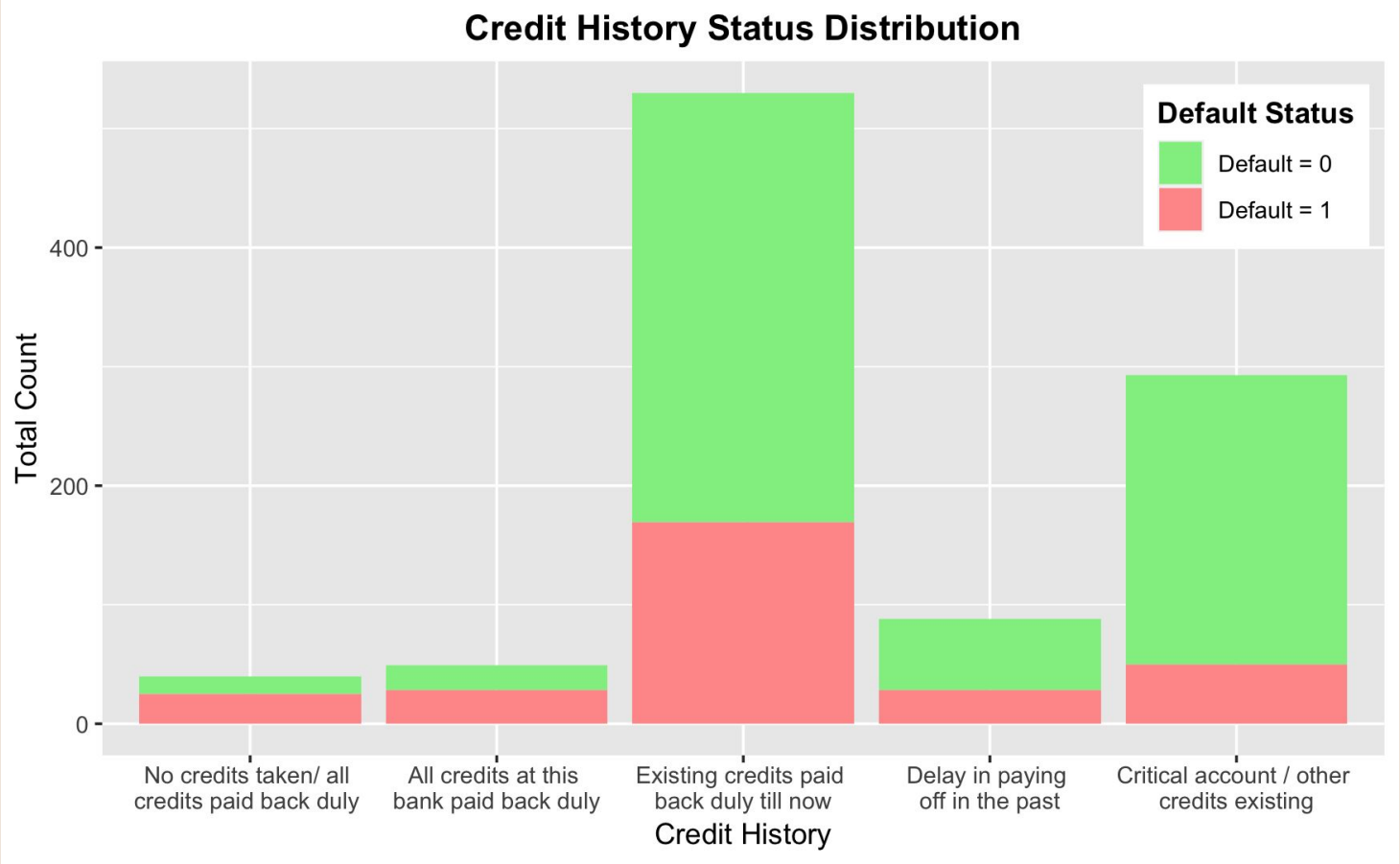
Filtered by Ages 20-30



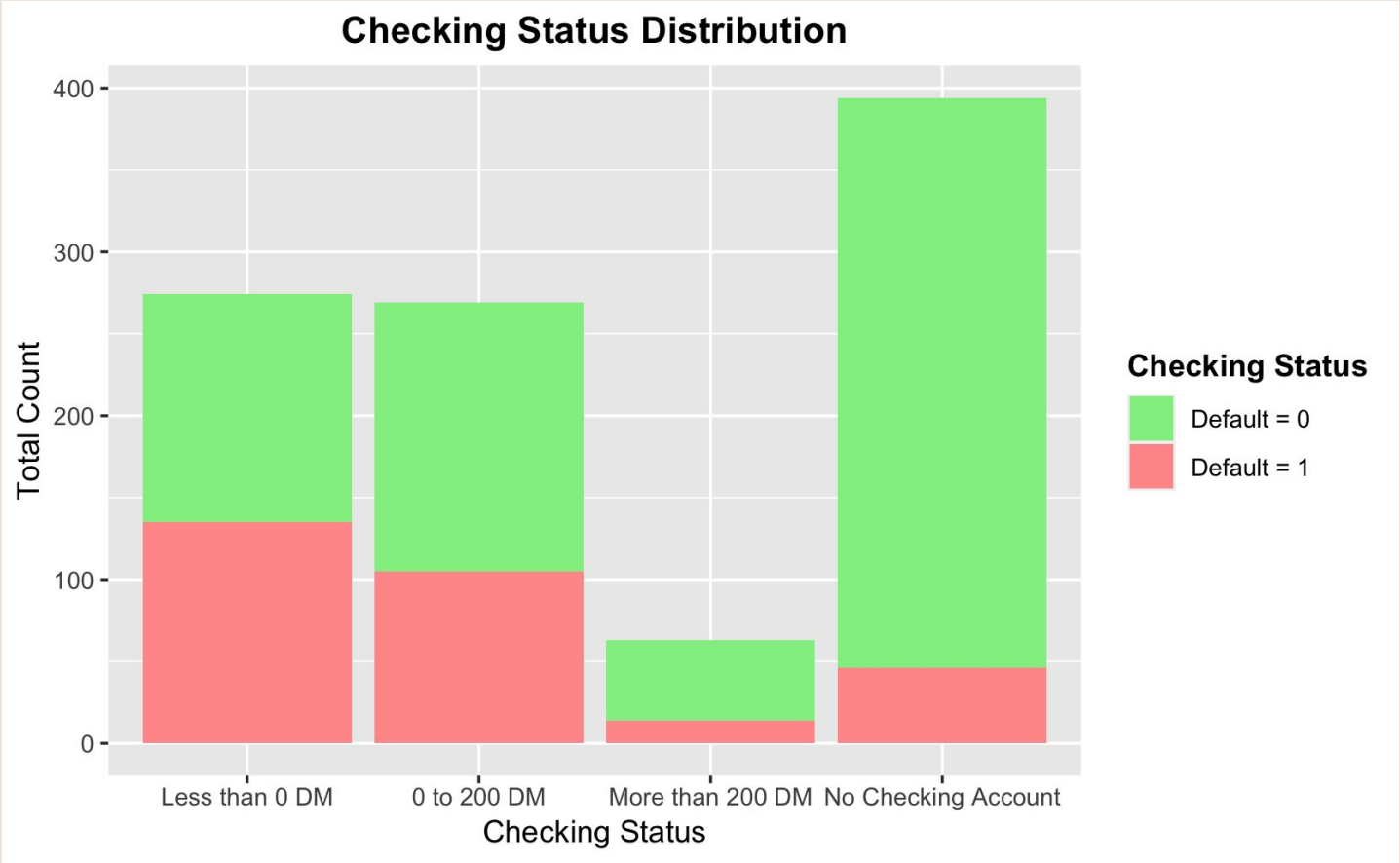
Age vs Default Status



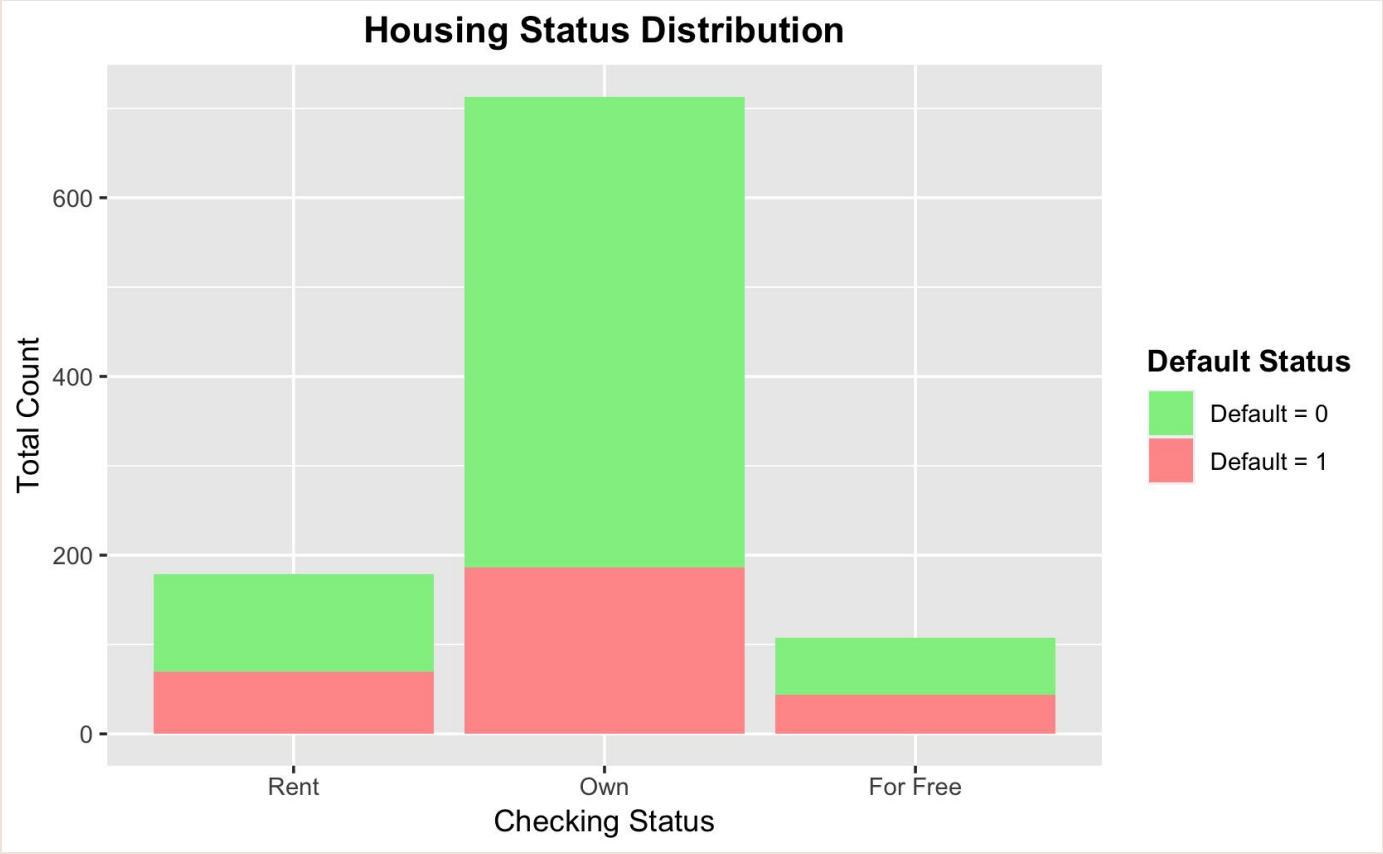
Credit History Status Results



Checking Status Results



Housing Status Results



Summary of Data Distributions

Applicant's characteristics that are least likely to Default:

- Own their own home
- No Checking Account
- Pays Back their Credit on Time
- Age pass 30

Logistic Regression Model Overview

glm(Default ~ . , data=Credit, family= “binomial”)

	Significant_P_Values
checkingstatus1A13	0.008905
checkingstatus1A14	1.664e-13
duration	0.002724
historyA34	0.001099
purposeA41	8.508e-06
purposeA42	0.002421
purposeA43	0.0003078
purposeA49	0.02667
amount	0.003894
savingsA64	0.01073
savingsA65	0.00031
installment	0.0001846
statusA93	0.03172
othersA103	0.02107
otherplansA143	0.006871
foreignA202	0.02609

	Non_Collinear_Variables
checkingstatus1A12	9.338
checkingstatus1A13	8.046
purposeA410	7.146
purposeA44	6.889
purposeA45	6.509
purposeA46	7.467
purposeA49	9.767
savingsA62	7.562
savingsA63	9.498
installment	9.743
othersA102	6.614
othersA103	8.873
residence	9.086
otherplansA142	7.6
otherplansA143	8.657

Correlation Plot of Credit Data (Numerical)

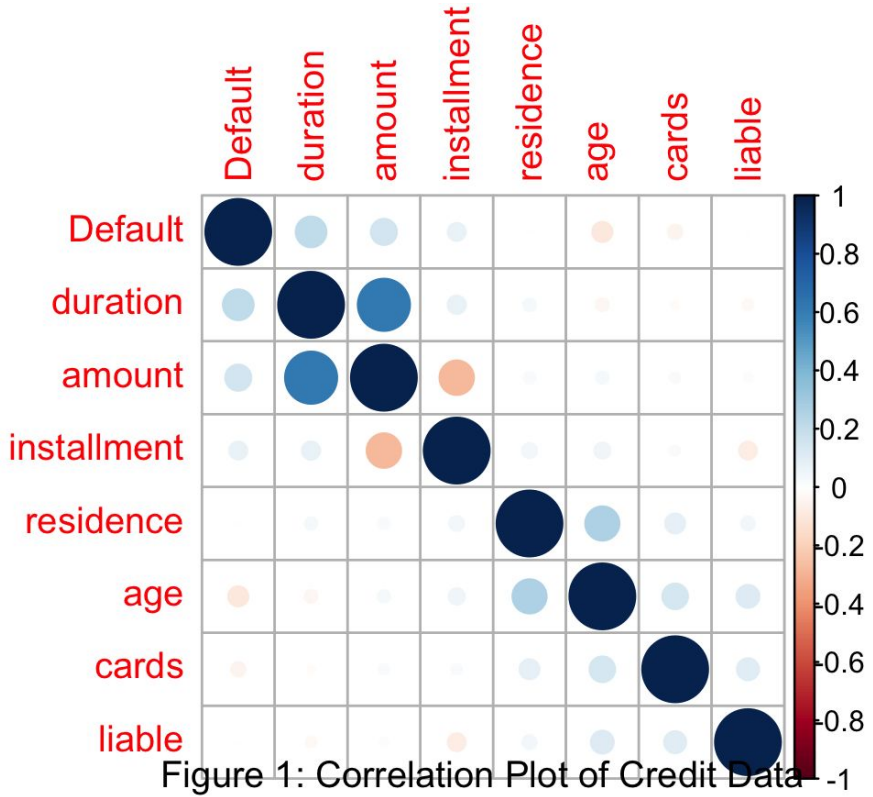
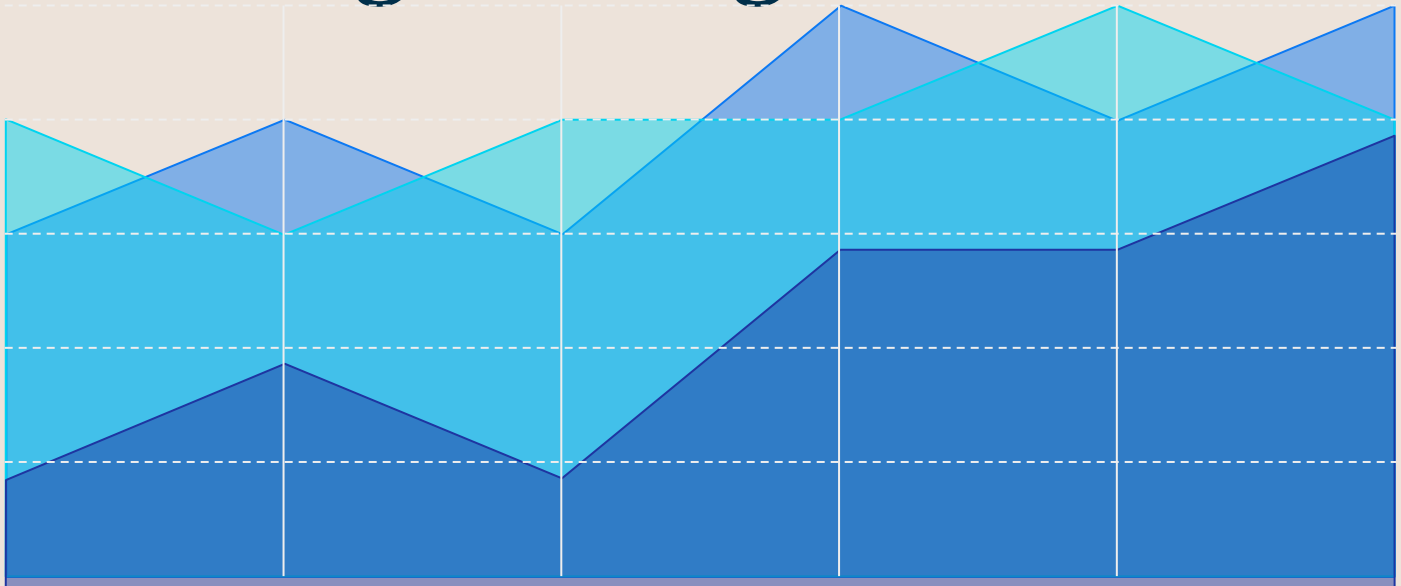


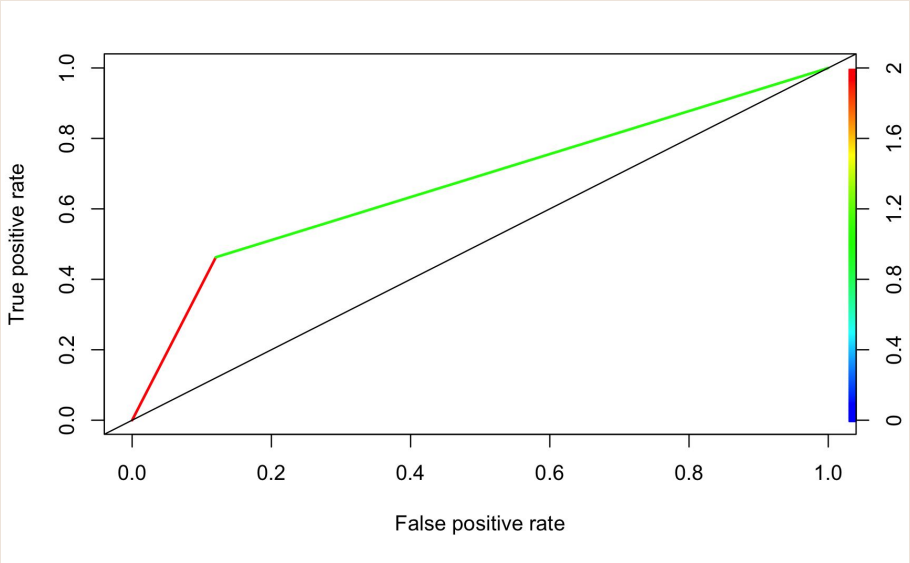
Figure 1: Correlation Plot of Credit Data

Examining the “good” logistic regression

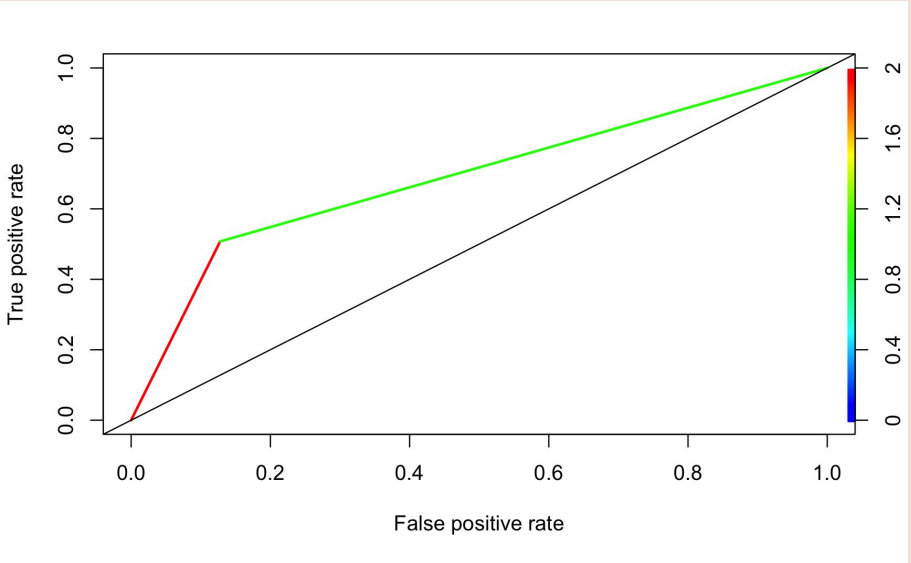


Training Set: 80% of Data
Testing Set: 20% of Data

ROC Curve: Initial Model



ROC Curve: StepAIC Model



Performance: Initial Model VS StepAIC Model

```
[1] "Initial Model Confusion Matrix"
          Reference
Prediction   0   1
         0 117  36
         1  16  31
[1] "StepAIC Model Confusion Matrix"
          Reference
Prediction   0   1
         0 116  33
         1  17  34
```

	Initial_Model_Rates	Step_Model_Rates
Sensitivity	0.8797	0.8722
Specificity	0.4627	0.5075

	Initial_Model_Acc	Step_Model_Acc
Accuracy	0.74	0.75

AUC

Initial_Model_AUC	Step_Model_AUC
0.6712	0.6898

Final Equation (StepAIC) Model Equation

$$\hat{Y} = 1.95 - .61\ checkingstatus1A12 - 1.16\ checkingstatus1A13 - 1.76\ checkingstatus1A14$$
$$+ .0245\ duration - 1.27\ historyA34 - 1.62\ purposeA41 - 2.38\ purposeA410 - .94\ purposeA43$$
$$+ .0001\ amount - 1.26\ savingsA64 - .91 savingsA65 + .38\ installment$$
$$- .84\ statusA93 - .021\ age - .81\ otherplansA143 - .51\ housingA152 - 1.61\ foreignA202$$

Step_Model_Error_Rate
0.3812

Applying Odd Ratio

1 - exp(Coefficient Estimate)

Odd Ratio Concept:

+ Regression Coefficient:

Predictor **Increase** | Probability of Response **Increase**

- Regression Coefficient:

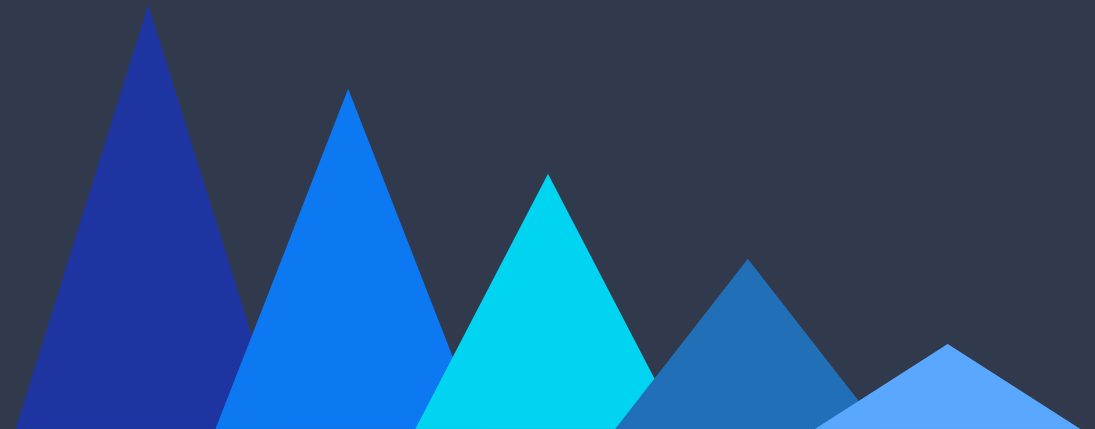
Predictor **Increase** | Probability of Response **Decrease**

Interpreting Regression Coefficients:

- **Default ~ Age (Year) :** 2% Decrease odds of Default for every 1 Year of age increase

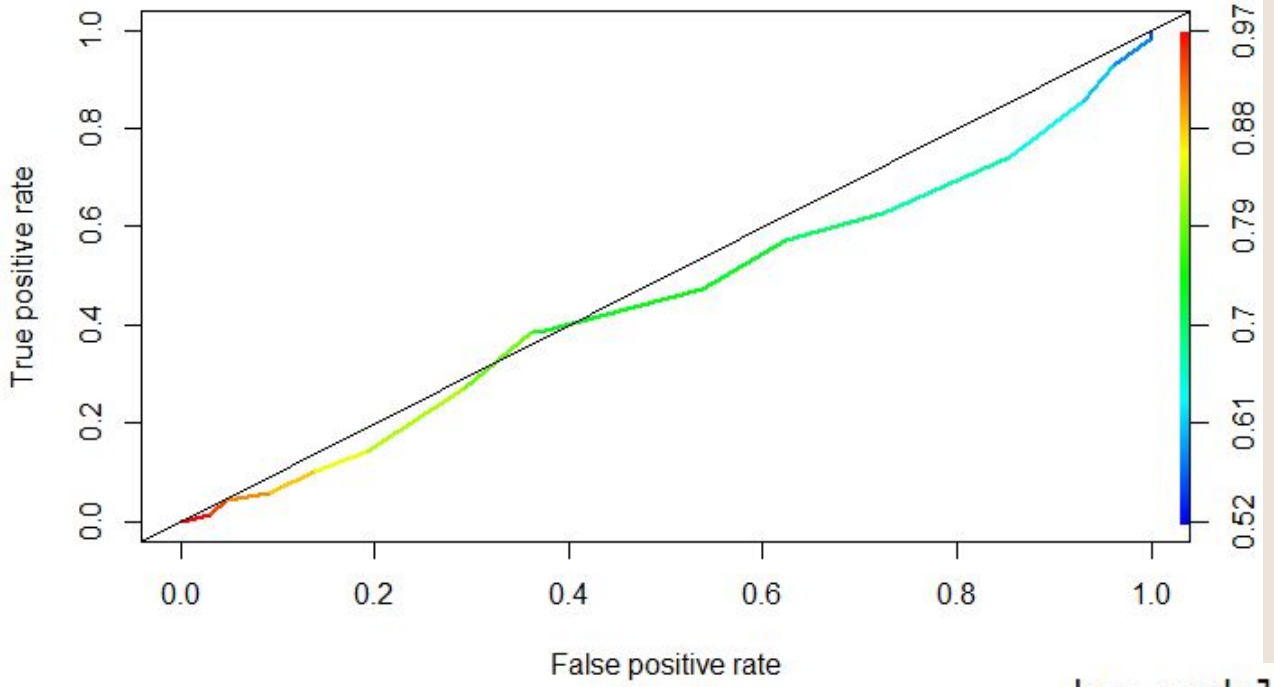
- **Default ~ Duration (Months):** 2% Increase odds of Default for every 1 month increase of duration

Investigating Classifiers: KNN, LDA, QDA, Naive Bayes



KNN

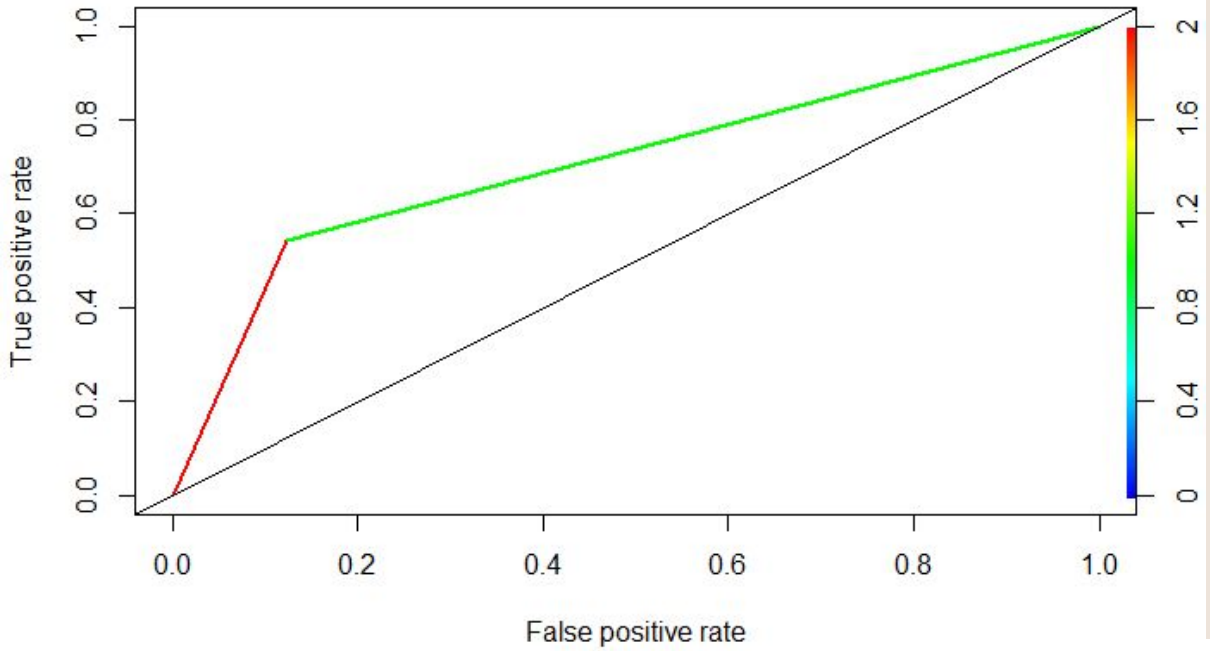
K-Nearest Neighbors Algorithm




```
knn.model   0   1
         0 129  68
         1   1   2
[1] "Error rate: 0.345"
[1] "Sensitivity: 0.992307692307692"
[1] "Specificity: 0.0285714285714286"
[1] "AUC: 0.451703296703297"
```

LDA

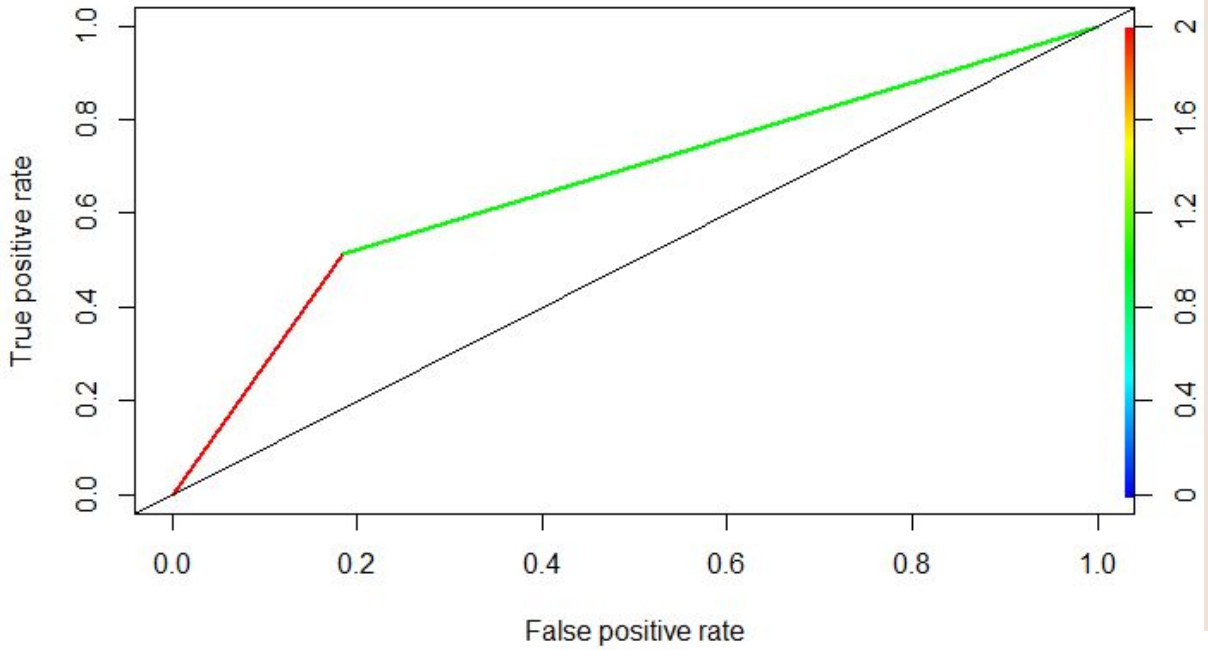
Linear discriminant analysis




```
          Reference
Prediction   0   1
         0 114  32
         1  16  38
[1] "Error rate: 0.24"
[1] "Sensitivity: 0.876923076923077"
[1] "Specificity: 0.542857142857143"
[1] "AUC: 0.709890109890011"
```

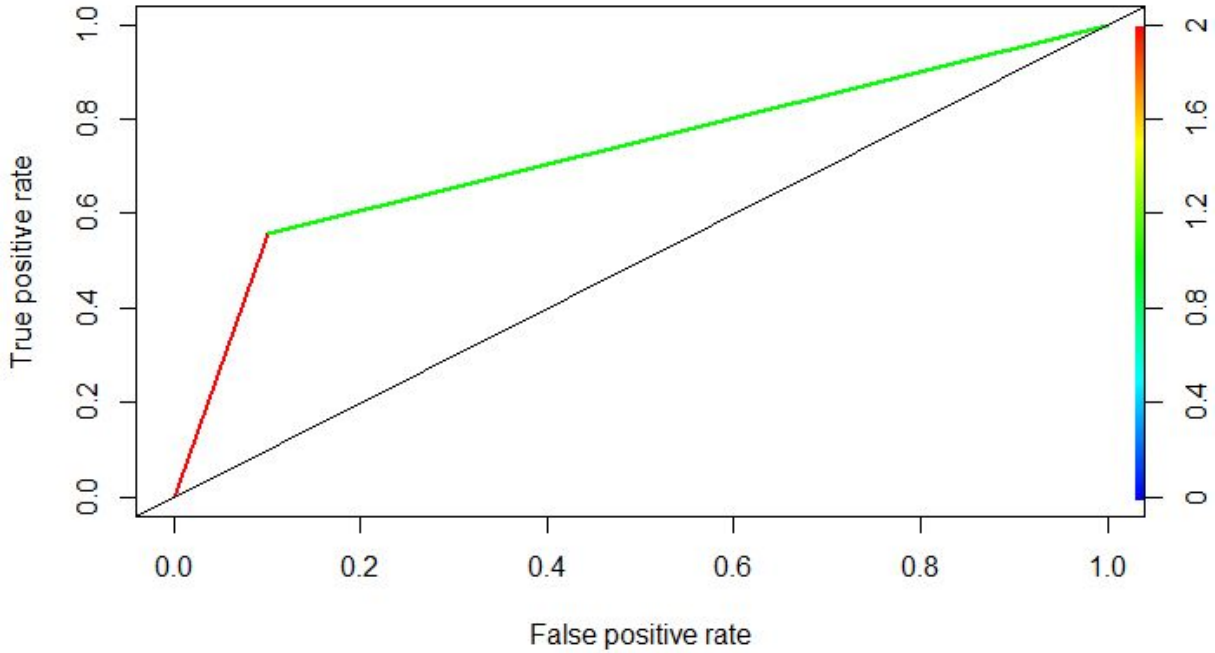
QDA

Qualitative Data Analysis




```
          Reference
Prediction   0   1
         0 106  34
         1  24  36
[1] "Error rate: 0.29"
[1] "Sensitivity: 0.815384615384615"
[1] "Specificity: 0.514285714285714"
[1] "AUC: 0.664835164835165"
```

Naive-Bayes




```
          Reference
Prediction   0   1
         0 117  31
         1  13  39
[1] "Error rate: 0.22"
[1] "Sensitivity: 0.9"
[1] "Specificity: 0.557142857142857"
[1] "AUC: 0.728571428571429"
```

Compare all models

	ER	SENS	SPEC	AUC
KNN	0.3450	0.9923077	0.02857143	0.4517033
LDA	0.2400	0.8769231	0.54285714	0.7098901
QDA	0.2900	0.8153846	0.51428571	0.6648352
LOG(IMR)	0.2433	0.8798000	0.47830000	0.6712000
LOG(SMR)	0.2667	0.8462000	0.47830000	0.6898000
NAIVE-B	0.2200	0.9000000	0.55714286	0.7285714

Finish