GE Credit/Financing Risk Analytic Solution

The power and value of descriptive statistics and predictive modeling

Logistic Regression & Random Forest

Charles Adkins

Southern New Hampshire University

What do we currently know?

Dataset =

700 No Defaults & 300 Defaults

- 30% error rate

Goal =

minimize Default events by lowering error rate of application approvals

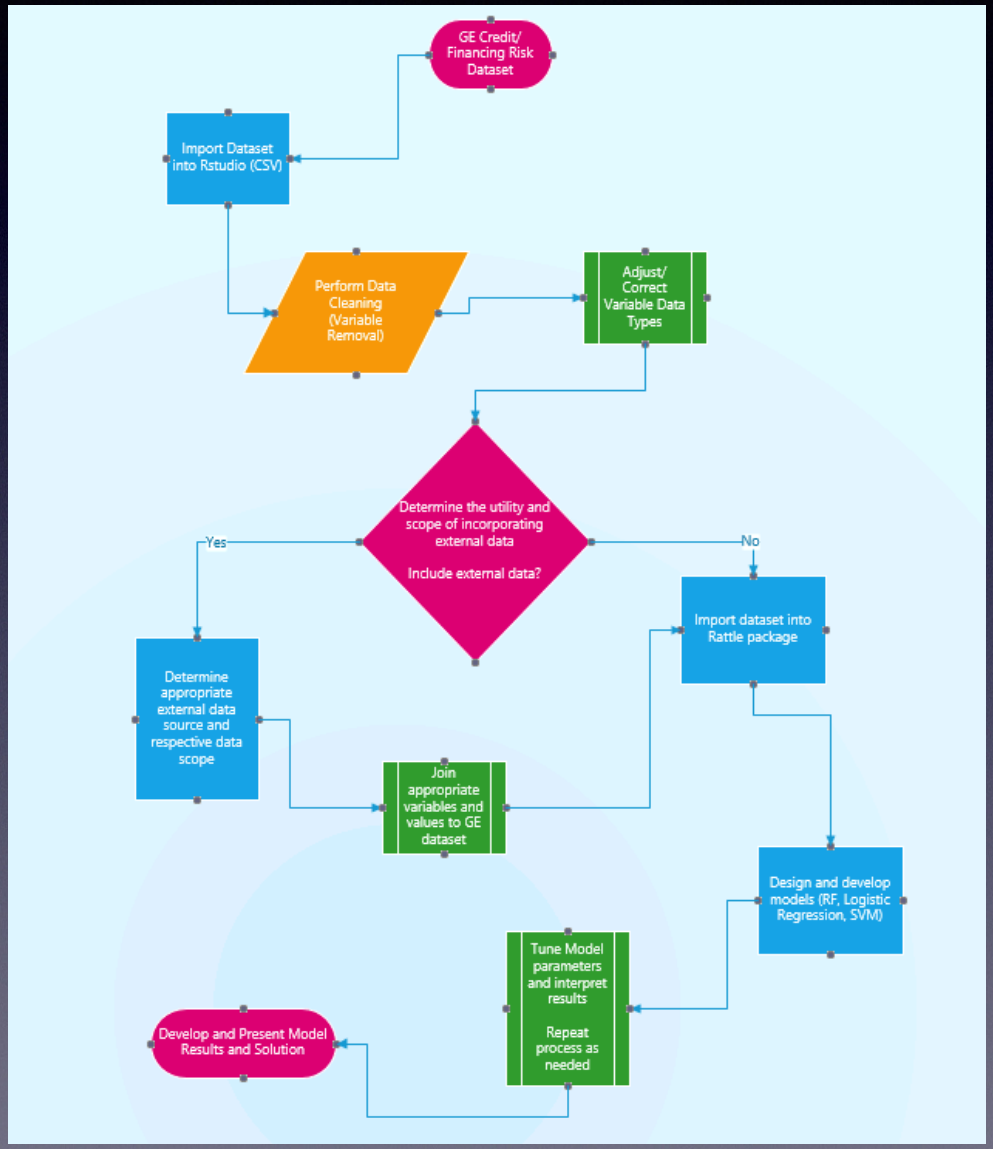
How do we approach the issue?

Utilizing the Random Forest model will provide us with reliable results by accounting for data bias or inaccurate influence and by identifying important variables.

By identifying the important variables, the model will establish criteria that is able to predict whether an applicant will default or not, based upon the values of those variables.

The prediction will assist the credit branches in the application decision making process by providing them with numerical data that supports the likelihood of a certain event, Default or No Default.

Analytical Approach



Data Exploration - Rcode

```r
## Import dataset into RStudio ## ------------------------------>

Credit_Data <- read.csv("~/Credit_Data.csv")




## View Dataset ## ------------------------------------------->

View(Credit_Data)




# Data Exploration and Cleaning - Check data quality (N/a, Null, missing values),
##data summary, descriptive statistics (Histograms), structure and variable datatypes ## --------------->


#Checks for N/a's
which(!complete.cases(Credit_Data))

#Provides summary of variable value distributions
summary(Credit_Data)

#Creates Histogram
hist(Credit_Data$variableName, breaks = #, xlim = c(beginning #, ending #),
       ylim = c(beginning #, ending #), main = "title of graphic", xlab = "x-axis label", ylab = "y-axis label")

#Displays structure (Datatype) and the first 10 values for each variable
str(Credit_Data)
```

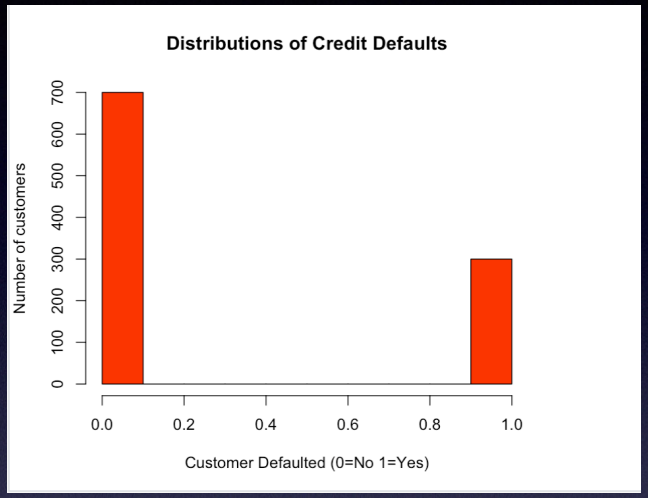
Data Exploration - Output

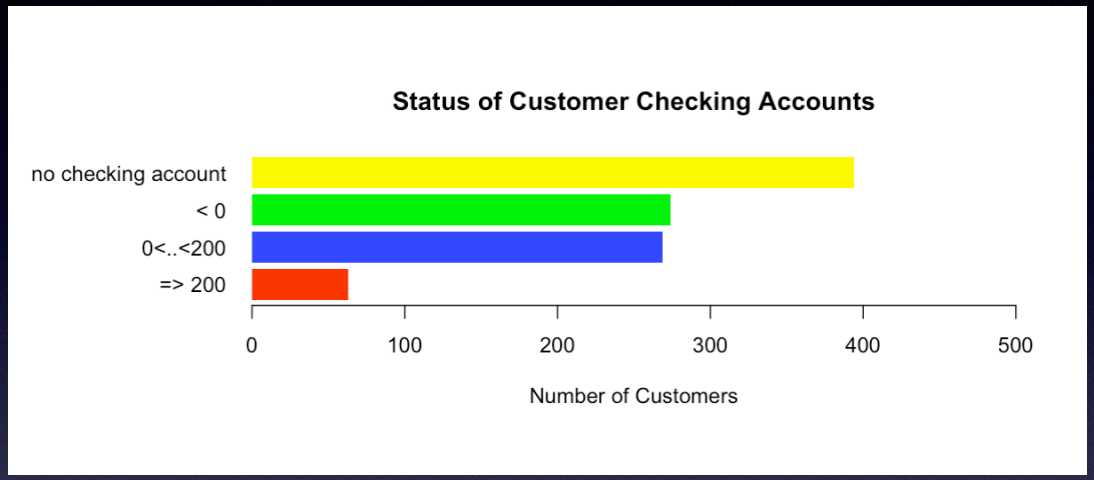
```
> which(!complete.cases(Credit_Data))
integer(0)
> complete.cases(Credit_Data)
   [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
  [11] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
  [21] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
  [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
  [41] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
  [51] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
  [61] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
  [71] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
  [81] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
  [91] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 [101] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

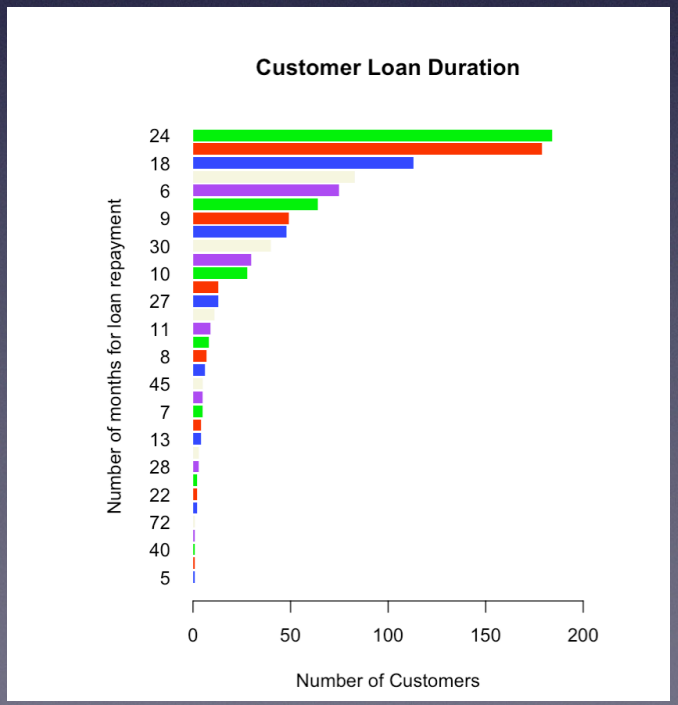
```
> str(Credit_Data)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       1000 obs. of  31 variables:
 $ CHK_ACCT        : Factor w/ 4 levels "< 0","0<..<200",..: 1 2 4 1 1 4 4 2 4 2
 ...
 $ DURATION        : num  6 48 12 42 24 36 24 36 12 30 ...
 $ HISTORY         : Factor w/ 5 levels "no credits taken",..: 5 3 5 3 4 3 3 3 3 5
 ...
 $ NEW_CAR         : num  0 0 0 0 1 0 0 0 0 1 ...
 $ USED_CAR        : num  0 0 0 0 0 0 0 1 0 0 ...
 $ FURNITURE       : num  0 0 0 1 0 0 1 0 0 0 ...
 $ RADIO/TV        : num  1 1 0 0 0 0 0 0 1 0 ...
 $ EDUCATION       : num  0 0 1 0 0 1 0 0 0 0 ...
 $ RETRAINING      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ AMOUNT          : num  1169 5951 2096 7882 4870 ...
 $ SAV_ACCT        : Factor w/ 5 levels "< 100","100<=..< 500",..: 5 1 1 1 1 5 3 1
 4 1 ...
 $ EMPLOYMENT      : Factor w/ 5 levels "unemployed","< 1 year",..: 5 3 4 4 3 3 5
 3 4 1 ...
 $ INSTALL_RATE    : num  4 2 2 2 3 2 3 2 2 4 ...
 $ MALE_DIV        : num  0 0 0 0 0 0 0 0 1 0 ...
 $ MALE_SINGLE     : num  1 0 1 1 1 1 1 1 0 0 ...
 $ MALE_MAR_or_WID : num  0 0 0 0 0 0 0 0 0 1 ...
 $ CO-APPLICANT    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ GUARANTOR       : num  0 0 0 1 0 0 0 0 0 0 ...
 $ PRESENT_RESIDENT: Factor w/ 4 levels "<= 1 year","1<..<= 2 years",..: 4 2 3 4 4
 4 4 2 4 2 ...
 $ REAL_ESTATE     : num  1 1 1 0 0 0 0 0 1 0 ...
 $ PROP_UNKN_NONE  : num  0 0 0 0 1 1 0 0 0 0 ...
 $ AGE             : num  67 22 49 45 53 35 53 35 61 28 ...
 $ OTHER_INSTALL   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ RENT            : num  0 0 0 0 0 0 0 1 0 0 ...
 $ OWN_RES         : num  1 1 1 0 0 0 1 0 1 1 ...
 $ NUM_CREDITS     : num  2 1 1 1 2 1 1 1 1 2 ...
 $ JOB             : num  2 2 1 2 2 1 2 3 1 3 ...
 $ NUM_DEPENDENTS  : num  1 1 2 2 2 2 1 1 1 1 ...
 $ TELEPHONE       : num  1 0 0 0 0 1 0 1 0 0 ...
 $ FOREIGN         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ DEFAULT         : num  0 1 0 0 1 0 0 0 0 1 ...
>
```

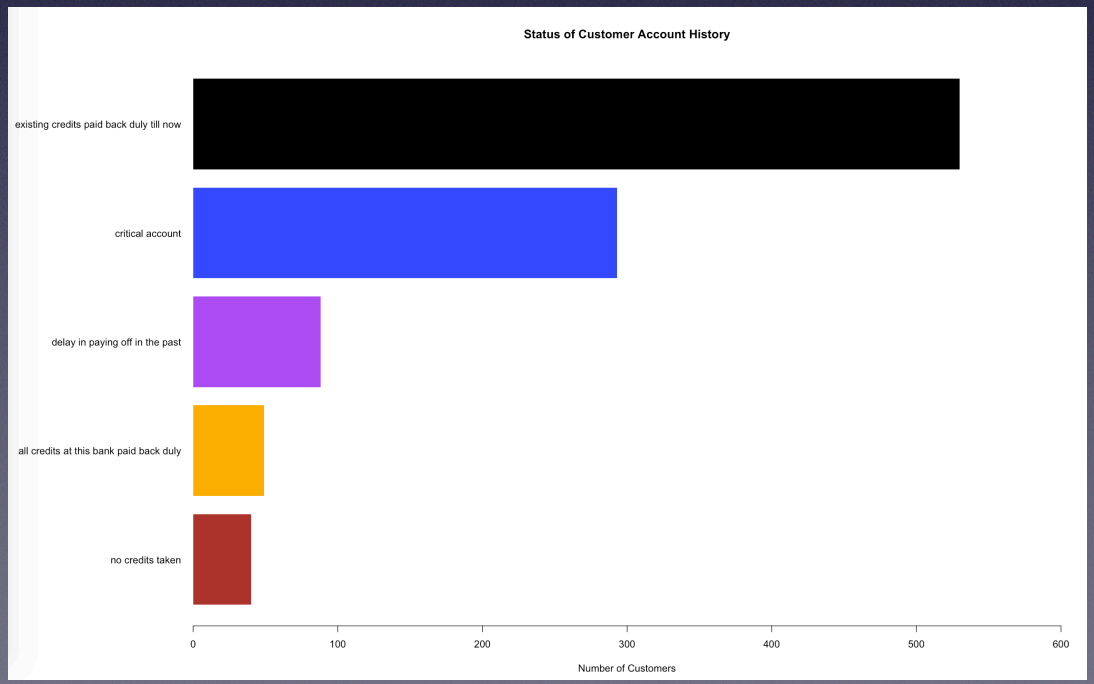
```
> summary(Credit_Data)
               CHK_ACCT      DURATION
 < 0                :274   Min.   : 4.0
 0<..<200           :269   1st Qu.:12.0
 => 200             : 63   Median :18.0
 no checking account:394   Mean   :20.9
                           3rd Qu.:24.0
                           Max.   :72.0
                                        HISTORY       NEW_CAR
 no credits taken                         : 40   Min.   :0.000
 all credits at this bank paid back duly  : 49   1st Qu.:0.000
 existing credits paid back duly till now:530   Median :0.000
 delay in paying off in the past          : 88   Mean   :0.234
 critical account                         :293   3rd Qu.:0.000
                                                 Max.   :1.000
    USED_CAR       FURNITURE        RADIO/TV      EDUCATION
 Min.   :0.000   Min.   :0.000   Min.   :0.00   Min.   :0.00
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.00   1st Qu.:0.00
 Median :0.000   Median :0.000   Median :0.00   Median :0.00
 Mean   :0.103   Mean   :0.181   Mean   :0.28   Mean   :0.05
 3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:1.00   3rd Qu.:0.00
 Max.   :1.000   Max.   :1.000   Max.   :1.00   Max.   :1.00
   RETRAINING        AMOUNT                          SAV_ACCT
 Min.   :0.000   Min.   :  250   < 100                  :603
 1st Qu.:0.000   1st Qu.: 1366   100<=..< 500           :103
 Median :0.000   Median : 2320   500<=..< 1000          : 63
 Mean   :0.097   Mean   : 3271   => 1000                : 48
 3rd Qu.:0.000   3rd Qu.: 3972   unknown/no savings account:183
 Max.   :1.000   Max.   :18424
           EMPLOYMENT   INSTALL_RATE      MALE_DIV     MALE_SINGLE
 unemployed     : 62   Min.   :1.000   Min.   :0.00   Min.   :0.000
 < 1 year       :172   1st Qu.:2.000   1st Qu.:0.00   1st Qu.:0.000
 1 <=..< 4 years:339   Median :3.000   Median :0.00   Median :1.000
 4 <=..< 7 years:174   Mean   :2.973   Mean   :0.05   Mean   :0.548
 >= 7 years     :253   3rd Qu.:4.000   3rd Qu.:0.00   3rd Qu.:1.000
                       Max.   :4.000   Max.   :1.00   Max.   :1.000
 MALE_MAR_or_WID  CO-APPLICANT     GUARANTOR           PRESENT_RESIDENT
 Min.   :0.000   Min.   :0.000   Min.   :0.000   <= 1 year     :130
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1<..<= 2 years:308
 Median :0.000   Median :0.000   Median :0.000   2<..<= 3 years:149
 Mean   :0.092   Mean   :0.041   Mean   :0.052   > 4 years     :413
 3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000
 Max.   :1.000   Max.   :1.000   Max.   :1.000
  REAL_ESTATE     PROP_UNKN_NONE       AGE         OTHER_INSTALL
 Min.   :0.000   Min.   :0.000   Min.   :19.00   Min.   :0.000
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:27.00   1st Qu.:0.000
 Median :0.000   Median :0.000   Median :33.00   Median :0.000
 Mean   :0.282   Mean   :0.154   Mean   :35.55   Mean   :0.186
 3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:42.00   3rd Qu.:0.000
 Max.   :1.000   Max.   :1.000   Max.   :75.00   Max.   :1.000
```

Data Visualizations

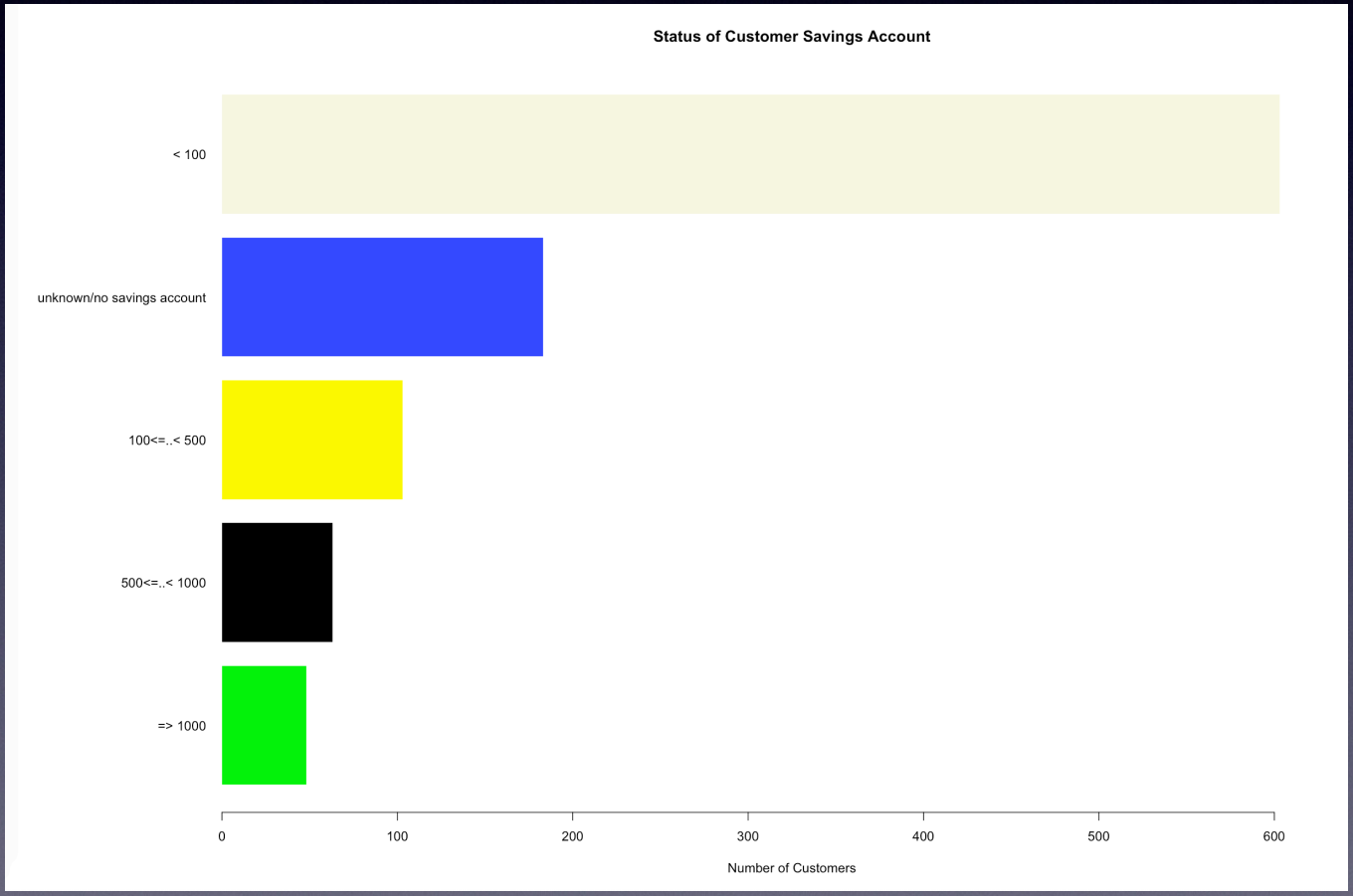




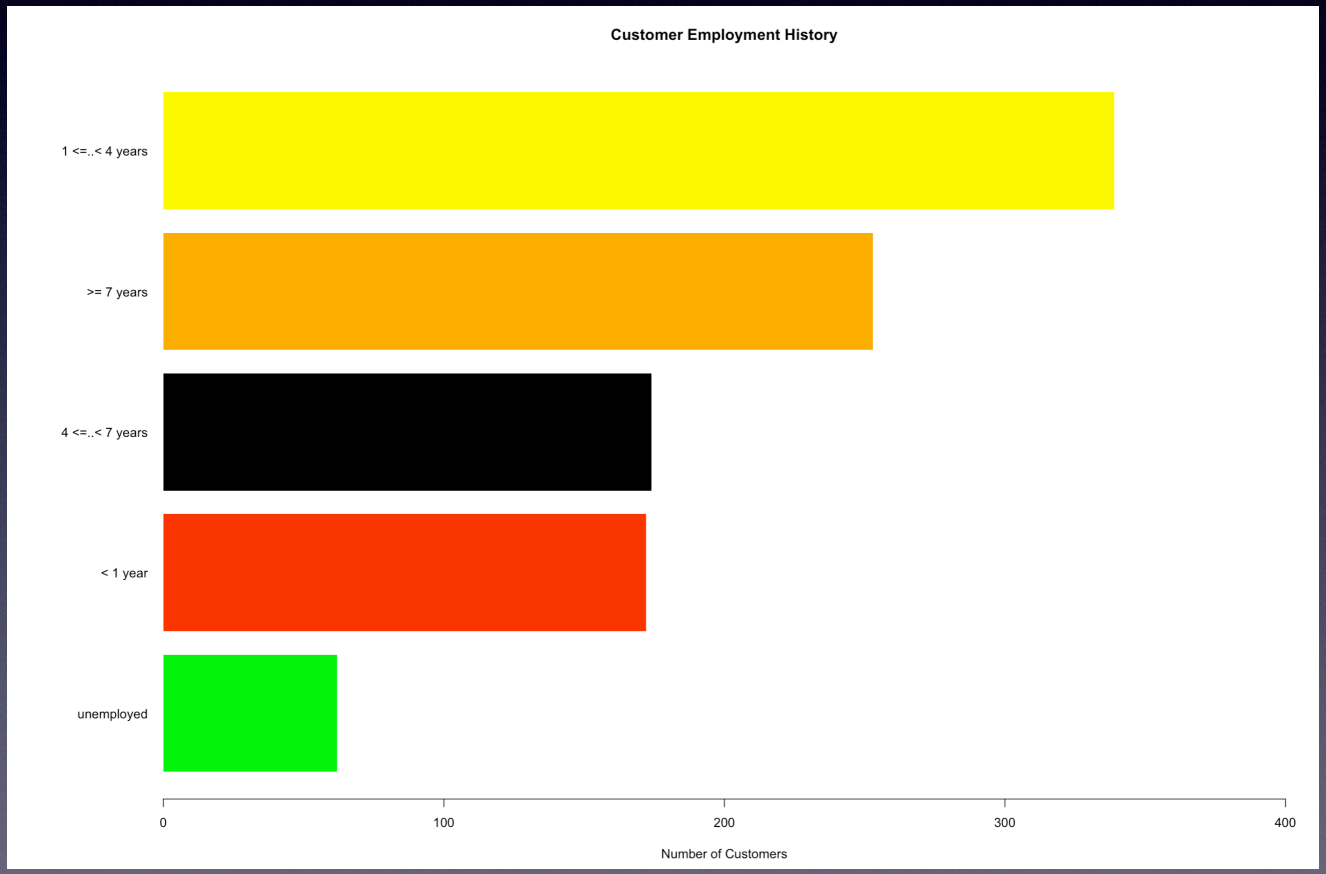




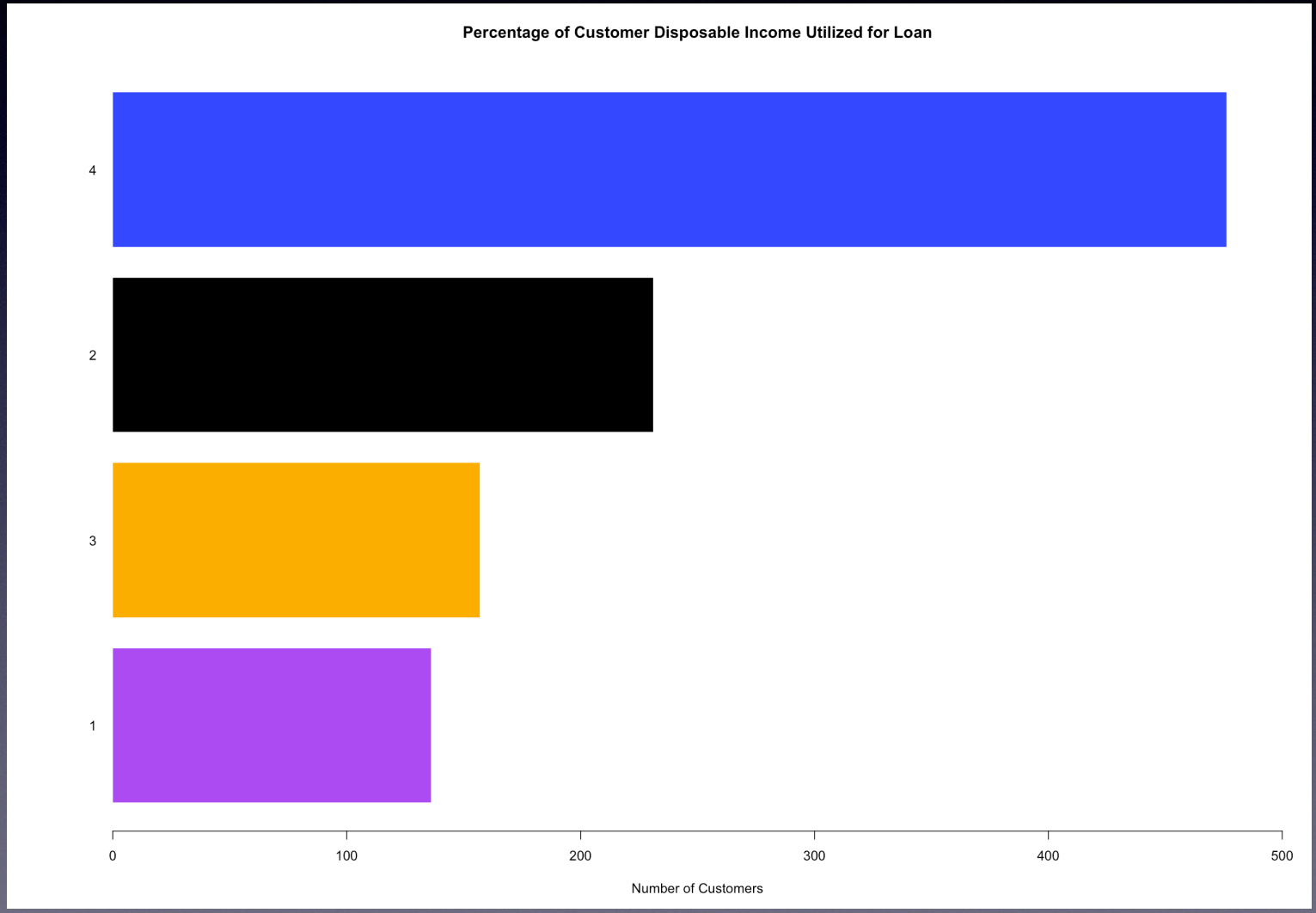
Data Visualizations - cont'd



Data Visualizations - cont'd 2



Data Visualizations - cont'd 3



Data Cleaning - Data Types Rcode

```
                ## Ensure that specific variable values are applicable
                    ## in terms of the metadata provided,
                       ## as well as their respective datatypes ##


## Fix datatypes IF APPLICABLE, integer to factors ## ------------------------------------------------->

Credit_Data$CHK_ACCT <- factor(Credit_Data$CHK_ACCT,
                               levels = c("0", "1", "2", "3"),
                               labels=c("< 0", "0<..<200", "=> 200", "no checking account")),

Credit_Data$HISTORY <- factor(Credit_Data$HISTORY,
                              levels = c("0", "1", "2", "3", "4"),
                              labels=c("no credits taken", "all credits at this bank paid back duly", "existing credits paid back duly till now",
                                       "delay in paying off in the past", "critical account")),

Credit_Data$SAV_ACCT <- factor(Credit_Data$SAV_ACCT,
                               levels = c("0", "1", "2", "3", "4"),
                               labels=c("< 100", "100<=..< 500", "500<=..< 1000", "=> 1000", "unknown/no savings account")),

Credit_Data$EMPLOYMENT <- factor(Credit_Data$EMPLOYMENT,
                                 levels = c("0", "1", "2", "3", "4"),
                                 labels=c("unemployed", "< 1 year", "1 <=..< 4 years", "4 <=..< 7 years", ">= 7 years")),

Credit_Data$PRESENT_RESIDENT <- factor(Credit_Data$PRESENT_RESIDENT,
                                       levels = c("1", "2", "3", "4"),
                                       labels= c("<= 1 year", "1<..<= 2 years", "2<..<= 3 years", "> 4 years"))



## Save changes made to the dataset by writing a new csv file ## ---------------->

write.csv(Credit_Data, file = "Creditdata_fixed.csv")
```

Data Exploration - Rcode

```r
## Gain an understanding of variable interactions by cross tabluation ## ------------------------------>

newObjectName <- table(Credit_Data$variableName, Credit_Data$otherVariableName)


## EXAMPLE = Specific to this data dataset, see below - cross tabulation for variables CHK_ACCT and DEFAULT (target variable) ------>

checking_default.tab <- table(Credit_Data$CHK_ACCT, Credit_Data$DEFAULT)


                ## This type of analysis helps you identify interactions between specific variables,
                  ## which can sometimes uncover hidden patterns within the data,
                    ## further assisting you in your model development process ##
```

Data Exploration -

Cross Tabulation

```
> checking_default.tab <- table(Credit_Data$CHK_ACCT, Credit_Data$DEFAULT)
> checking_default.tab

                       0   1
  < 0                139 135
  0<..<200           164 105
  => 200              49  14
  no checking account 348  46
```

```
> savings_default.tab <- table(Credit_Data$SAV_ACCT, Credit_Data$DEFAULT)
> savings_default.tab

                           0   1
  < 100                  386 217
  100<=..< 500            69  34
  500<=..< 1000           52  11
  => 1000                 42   6
  unknown/no savings account 151  32
```

```
> duration_default.tab <- table(Credit_Data$DURATION, Credit_Data$DEFAULT)
> duration_default.tab

     0   1
  4  6   0
  5  1   0
  6 66   9
  7  5   0
  8  6   1
  9 35  14
  10 25  3
  11  9  0
  12 130 49
  13  4  0
  14  3  1
  15 52 12
  16  1  1
  18 71 42
  20  7  1
  21 21  9
  22  2  0
  24 128 56
  26  1  0
  27  8  5
  28  2  1
  30 27 13
  33  2  1
  36 46 37
  39  4  1
  40  0  1
  42  8  3
  45  1  4
  47  1  0
  48 20 28
  54  1  1
  60  7  6
  72  0  1
```

```
> history_default.tab <- table(Credit_Data$HISTORY, Credit_Data$DEFAULT)
> history_default.tab

                                          0   1
  no credits taken                       15  25
  all credits at this bank paid back duly  21  28
  existing credits paid back duly till now 361 169
  delay in paying off in the past        60  28
  critical account                       243  50
```

```
> employment_default.tab <- table(Credit_Data$EMPLOYMENT, Credit_Data$DEFAULT)
> employment_default.tab

                   0   1
  unemployed      39  23
  < 1 year       102  70
  1 <=..< 4 years 235 104
  4 <=..< 7 years 135  39
  >= 7 years      189  64
```

```
> install_default <- table(Credit_Data$INSTALL_RATE, Credit_Data$DEFAULT)
> install_default

     0   1
  1 102  34
  2 169  62
  3 112  45
  4 317 159
```

Data Exploration -

Cross Tabulation cont'd

```
> age_default <- table(Credit_Data$AGE, Credit_Data$DEFAULT)
> age_default

     0  1
 19  1  1
 20  9  5
 21  9  5
 22 16 11
 23 28 20
 24 25 19
 25 22 19
 26 36 14
 27 38 13
 28 28 15
 29 22 15
 30 29 11
 31 27 11
 32 25  9
 33 20 13
 34 21 11
 35 34  6
 36 33  6
 37 21  8
 38 20  4
 39 15  6
 40 19  6
 41 13  4
 42 14  8
 43 12  5
 44 12  5
 45 12  3
 46 14  4
 47 12  5
 48  9  3
 49 13  1
 50  9  3
 51  7  1
 52  8  1
 53  2  5
 54  8  2
 55  5  3
 56  3  0
 57  6  3
 58  3  2
 59  2  1
 60  3  3
 61  4  3
 62  2  0
 63  7  1
 64  5  0
 65  4  1
 66  3  2
 67  3  0
 68  1  2
 70  1  0
 74  3  1
 75  2  0
```

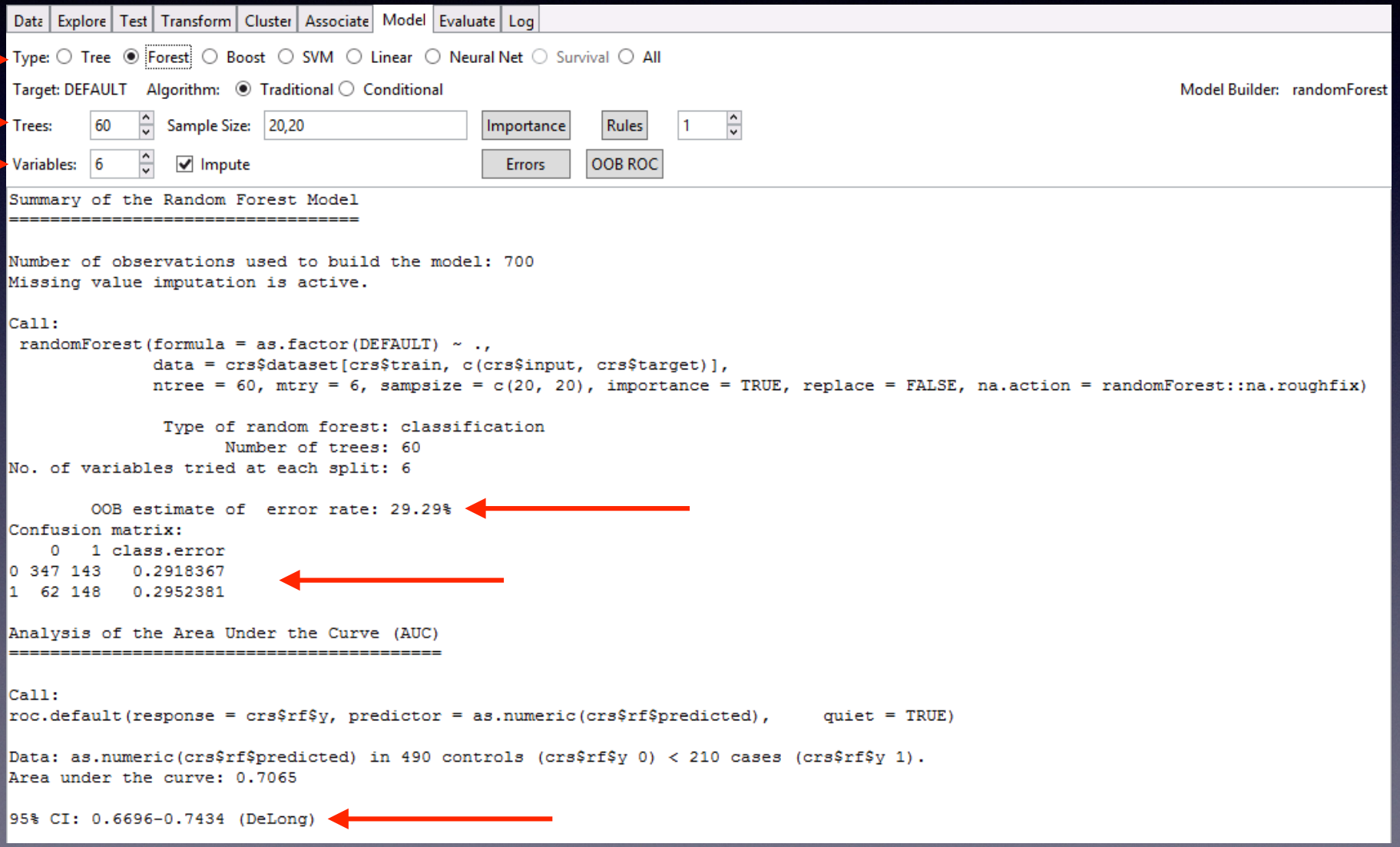
DAT-650 Pilot Model -

Rattle Rcode

```
## Once you've gained an understanding of the data and made all necessary corrections,
##  model development can begin, in this case through the Rattle package ###


## Install rattle, if you have not done so already ------------------------------------------------------------------>

install.packages("rattle")

## Engage Rattle utilizing Library() ---------------------------------->

library(rattle)

## Open Rattle ------------------------------------------------------->

rattle()
```

DAT-650 Pilot Run -

Random Forest



DAT-650 Pilot Run - Random Forest

Variable Importance

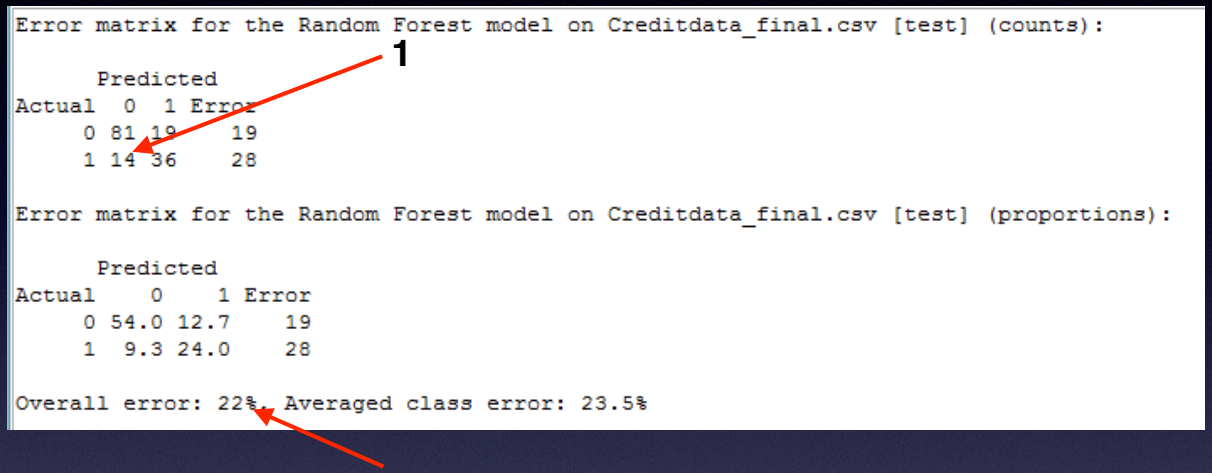
- Important variables establish which variables are most influential in determining the outcome of Default or No Default
- By identifying those variables, we can more quickly identify the presence of these variables and make a determination of application approval or disapproval

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
CHK_ACCT	5.29	5.72	5.87	2.41
HISTORY	2.86	3.17	3.50	1.65
DURATION	2.12	2.53	3.03	1.48
EMPLOYMENT	2.31	2.50	3.02	1.38
NEW_CAR	1.71	2.41	2.85	0.44
SAV_ACCT	1.60	3.40	2.85	1.26
PROP_UNKN_NONE	2.06	1.30	2.59	0.53
MALE_SINGLE	2.39	-0.59	2.34	0.25
RADIO.TV	1.99	1.71	2.22	0.21
GUARANTOR	2.03	1.96	2.07	0.15
AGE	1.55	0.44	1.70	2.31
MALE_DIV	1.44	1.01	1.47	0.11
AMOUNT	0.34	1.77	1.28	2.02
CO.APPLICANT	1.61	0.49	1.22	0.18
NUM_DEPENDENTS	0.16	2.22	1.20	0.21
OTHER_INSTALL	1.39	0.35	1.16	0.47
FOREIGN	0.54	0.86	1.11	0.09
RENT	0.78	-0.16	1.03	0.22
RETRAINING	0.77	0.54	0.88	0.19
REAL_ESTATE	0.62	-0.05	0.75	0.43
NUM_CREDITS	0.91	-0.41	0.70	0.20
PRESENT_RESIDENT	0.62	-0.26	0.55	0.84
JOB	0.30	0.07	0.34	0.51
OWN_RES	0.54	-0.75	0.29	0.37
EDUCATION	-0.45	0.76	0.26	0.10
INSTALL_RATE	0.49	-0.69	0.13	0.77
MALE_MAR_or_WID	-0.21	0.63	0.00	0.33
USED_CAR	-0.09	-0.23	-0.18	0.26
FURNITURE	-0.57	0.12	-0.54	0.24
TELEPHONE	-0.91	-0.07	-1.38	0.29

1

2

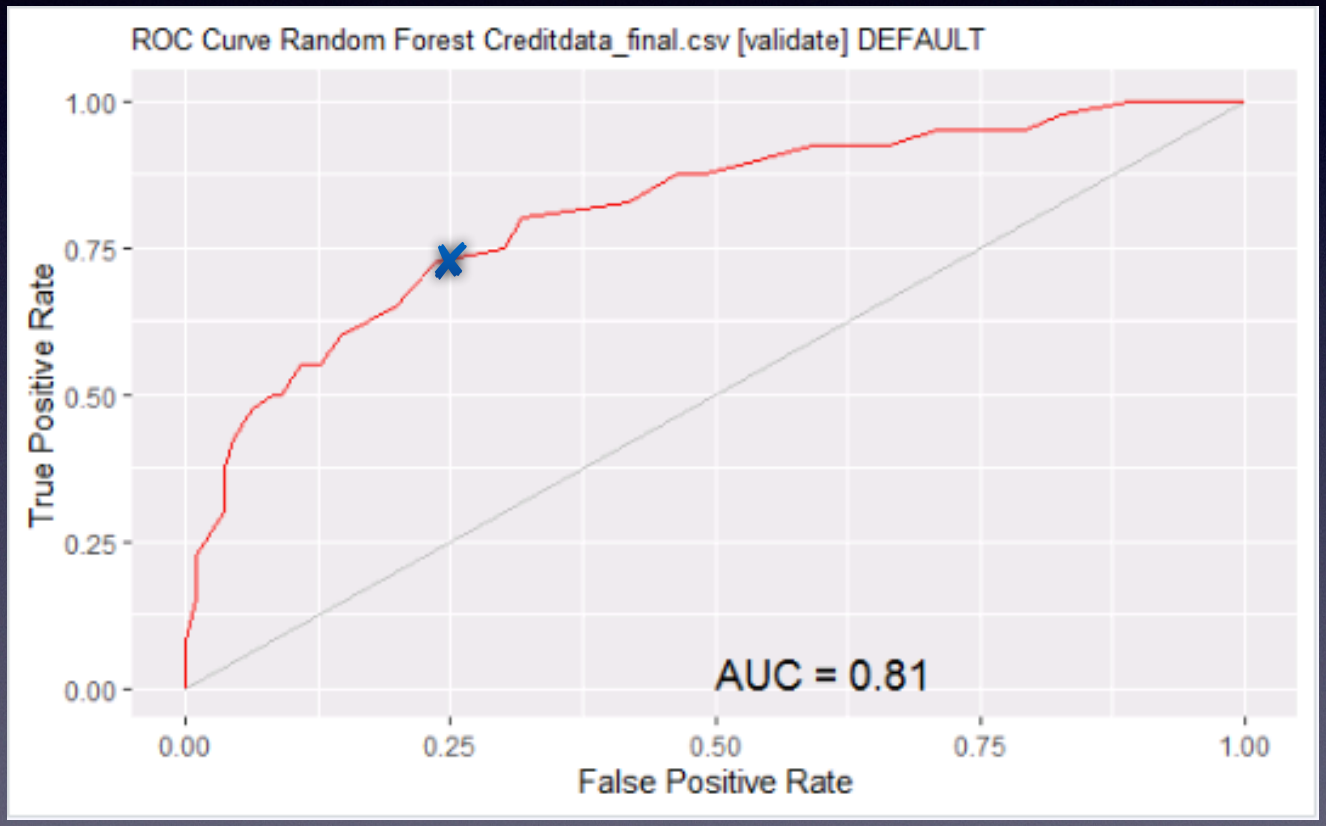
Pilot Run - Model Error



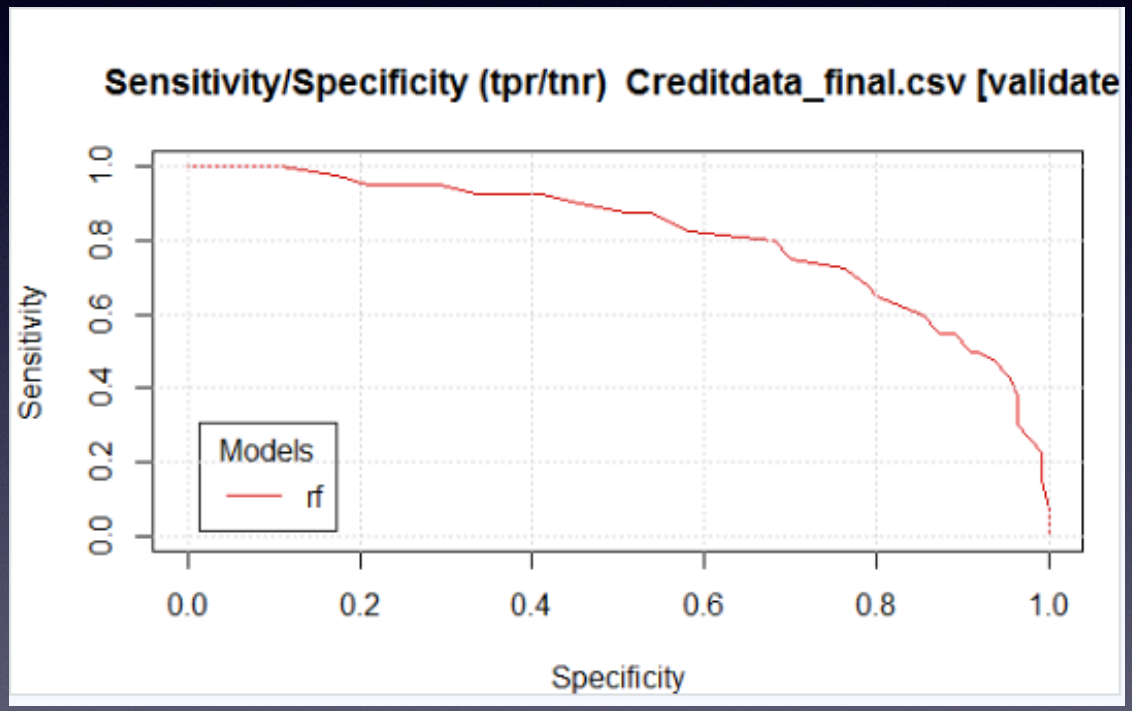
- This error matrix displays the error based upon the test dataset. It displays the disagreement between the final model's predictions and the actual outcomes of the testing observations.

- When referring back to the model summary, we remember the error rate was ~29% for the training set, while this finalized report displays the overall error at 22%, a 7% error decrease.

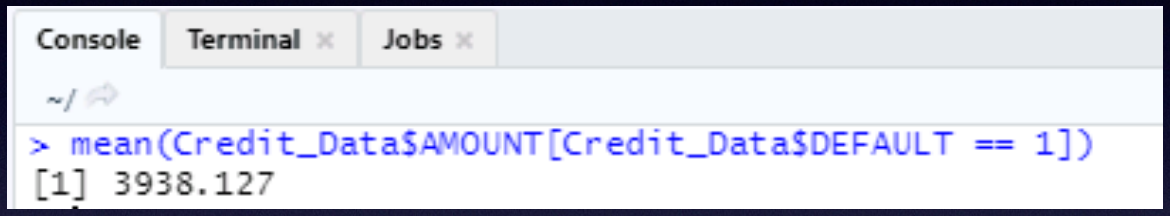
Pilot Run - Accuracy



Pilot Run - Model Accuracy Cont'd



Amount Financed -
Default Applicants



$3938.13 avg. Amount of Credit
X
300 Default Applicants

-$1,181,439 financed to these 300 applicants

Pilot Run -

Failures, Successes and Areas of Concern

Failures:

The pilot run helped us establish a baseline understanding of our data and how it behaves when utilized in a Random Forest Model. However, the results of that said model are limited. We need to derive more concrete, data-driven evidence from the results that the GE credit branch associates can utilize during the financing application process. This specifically includes reducing the error rate of which occurred without the implementation of an analytic solution, resulting in a 30% default rate.

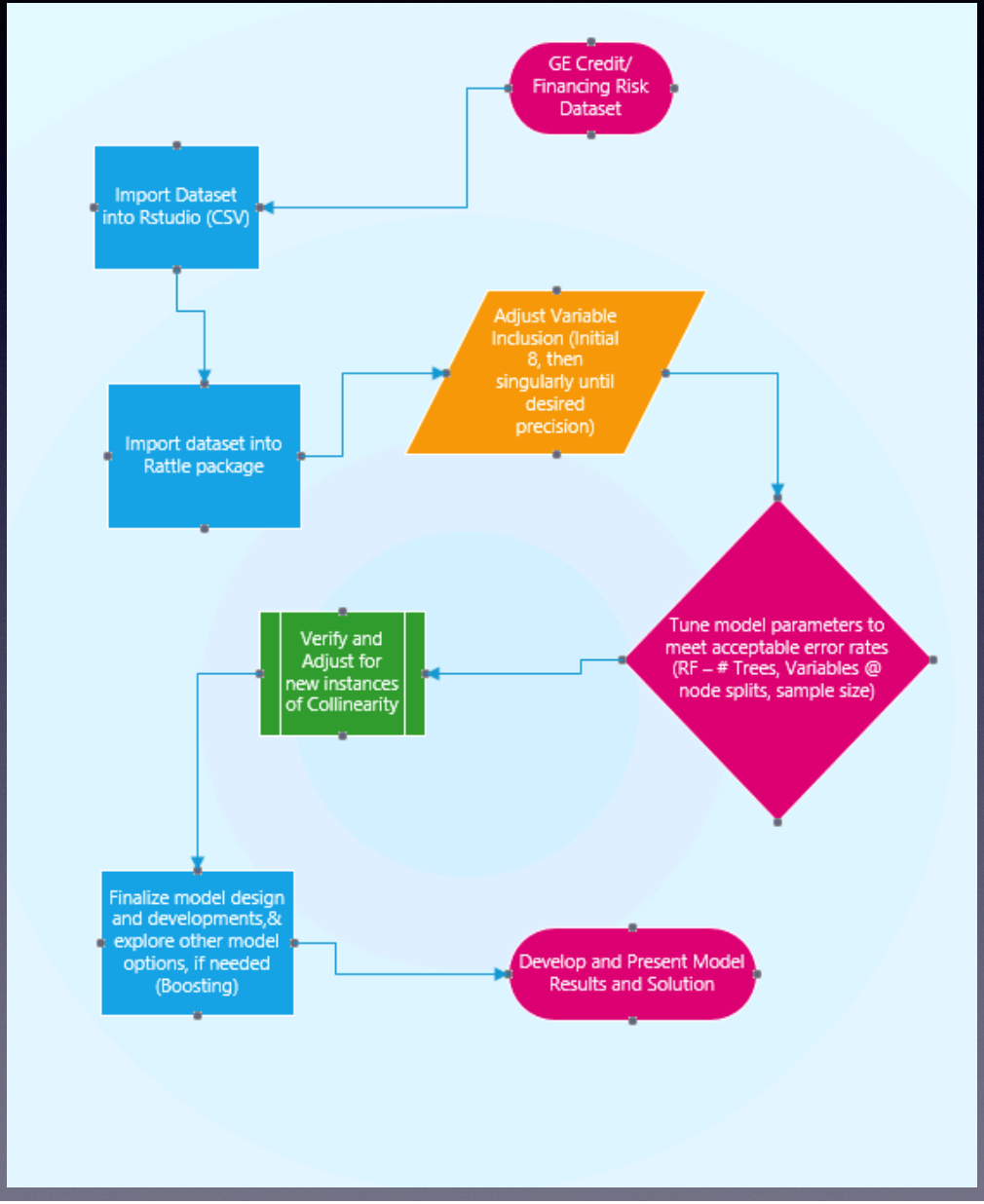
Successes:

We have successfully uncovered the most significantly influential variables, as they relate to the target variable. Through this, we can establish a logistic regression model by utilizing those variables. The results we obtained through our preliminary pilot run helps us establish a new process flowchart and create more well-tuned models.

Areas of Concern:

Over-fitting the model can create skewed results and complicate the value of the solution more so than having no solution at all. More specifically, we must mitigate model error and understand how variables relate to one another, as well as the target variable, to ensure that we're not simplifying the power certain variables have on the output of accurate results. This must be addressed by ensuring that our model development is taking a holistic, objective approach.

Pilot Modifications



Plan Implementation

Type: ○ Tree ◉ Forest ○ Boost ○ SVM ○ Linear ○ Neural Net ○ Survival ○ All

Target: DEFAULT Algorithm: ◉ Traditional ○ Conditional Model Builder: randomForest

Trees: 60 Sample Size: Importance Rules 1

Variables: 6 ☑ Impute Errors OOB ROC

Variable Importance

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
CHK_ACCT	3.80	11.58	9.60	23.88
HISTORY	3.07	4.12	4.97	11.95
DURATION	3.09	2.10	3.86	16.60
AGE	2.03	2.50	3.16	18.08
GUARANTOR	3.31	0.98	3.07	2.01
PROP_UNKN_NONE	1.95	1.91	2.84	3.19
AMOUNT	2.65	0.43	2.63	22.66
EDUCATION	1.70	0.62	1.94	1.70
USED_CAR	1.23	1.12	1.77	1.25
NUM_CREDITS	2.40	-0.43	1.71	3.09
REAL_ESTATE	1.83	0.36	1.60	3.55
MALE_MAR_or_WID	2.85	-0.83	1.56	2.00
RADIO.TV	0.55	1.65	1.55	2.69
EMPLOYMENT	0.16	2.19	1.54	12.75
SAV_ACCT	0.16	2.10	1.50	10.25
JOB	1.67	-0.04	1.43	5.46
INSTALL_RATE	0.77	0.55	1.10	6.83
FOREIGN	1.54	0.04	1.08	0.63
CO.APPLICANT	1.59	-0.31	0.87	1.57
RENT	0.10	0.97	0.75	2.04
OWN_RES	-0.32	1.18	0.42	3.62
OTHER_INSTALL	0.49	0.08	0.37	3.46
NEW_CAR	1.18	-2.30	-0.44	3.28
FURNITURE	0.34	-1.06	-0.44	2.33
MALE_SINGLE	-0.43	-0.74	-0.71	3.63
NUM_DEPENDENTS	0.28	-1.50	-0.83	2.29
MALE_DIV	-0.87	-0.83	-1.17	1.21
TELEPHONE	-1.03	-0.92	-1.25	2.58
RETRAINING	-1.70	0.20	-1.38	1.69
PRESENT_RESIDENT	-1.61	-1.24	-1.81	8.31

Logistic Regression Model Coefficients

Data | Explore | Test | Transform | Cluster | Associate | Model | Evaluate | Log

Type: ○ Tree ○ Forest ○ Boost ○ SVM ◉ Linear ○ Neural Net ○ Survival ○ All

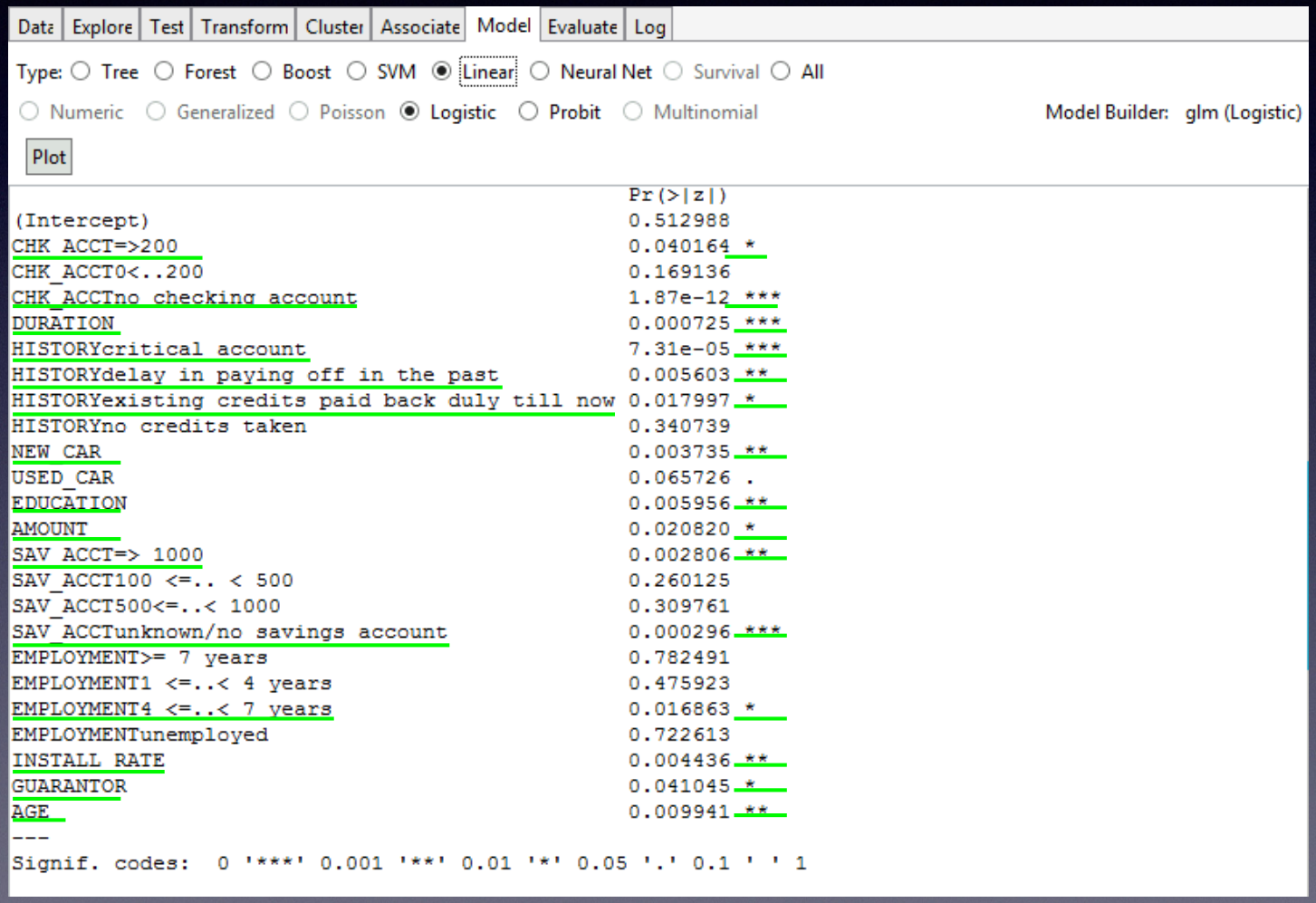
○ Numeric ○ Generalized ○ Poisson ◉ Logistic ○ Probit ○ Multinomial Model Builder: glm (Logistic)

Plot

Coefficients:

	Estimate	Std. Error	z value
(Intercept)	0.44675271	0.68290673	0.654
CHK_ACCT=>200	-0.82769149	0.40334664	-2.052
CHK_ACCT0<..200	-0.33752618	0.24547602	-1.375
CHK_ACCTno checking account	-1.97486627	0.28036868	-7.044
DURATION	0.03490477	0.01032684	3.380
HISTORYcritical account	-1.85881529	0.46868925	-3.966
HISTORYdelay in paying off in the past	-1.50374232	0.54284120	-2.770
HISTORYexisting credits paid back duly till now	-1.03813450	0.43883186	-2.366
HISTORYno credits taken	-0.58702452	0.61616484	-0.953
NEW_CAR	0.71060247	0.24506094	2.900
USED_CAR	-0.75937531	0.41263915	-1.840
EDUCATION	1.26326660	0.45934180	2.750
AMOUNT	0.00011345	0.00004909	2.311
SAV_ACCT=> 1000	-2.43412953	0.81455994	-2.988
SAV_ACCT100 <=.. < 500	-0.37635235	0.33420996	-1.126
SAV_ACCT500<=..< 1000	-0.49118505	0.48358091	-1.016
SAV_ACCTunknown/no savings account	-1.15275421	0.31851855	-3.619
EMPLOYMENT>= 7 years	-0.09150228	0.33144062	-0.276
EMPLOYMENT1 <=..< 4 years	-0.19878537	0.27885023	-0.713
EMPLOYMENT4 <=..< 7 years	-0.83093098	0.34771557	-2.390
EMPLOYMENTunemployed	0.15954032	0.44944935	0.355
INSTALL_RATE	0.27868051	0.09794265	2.845
GUARANTOR	-0.97825531	0.47881405	-2.043
AGE	-0.02701741	0.01048049	-2.578

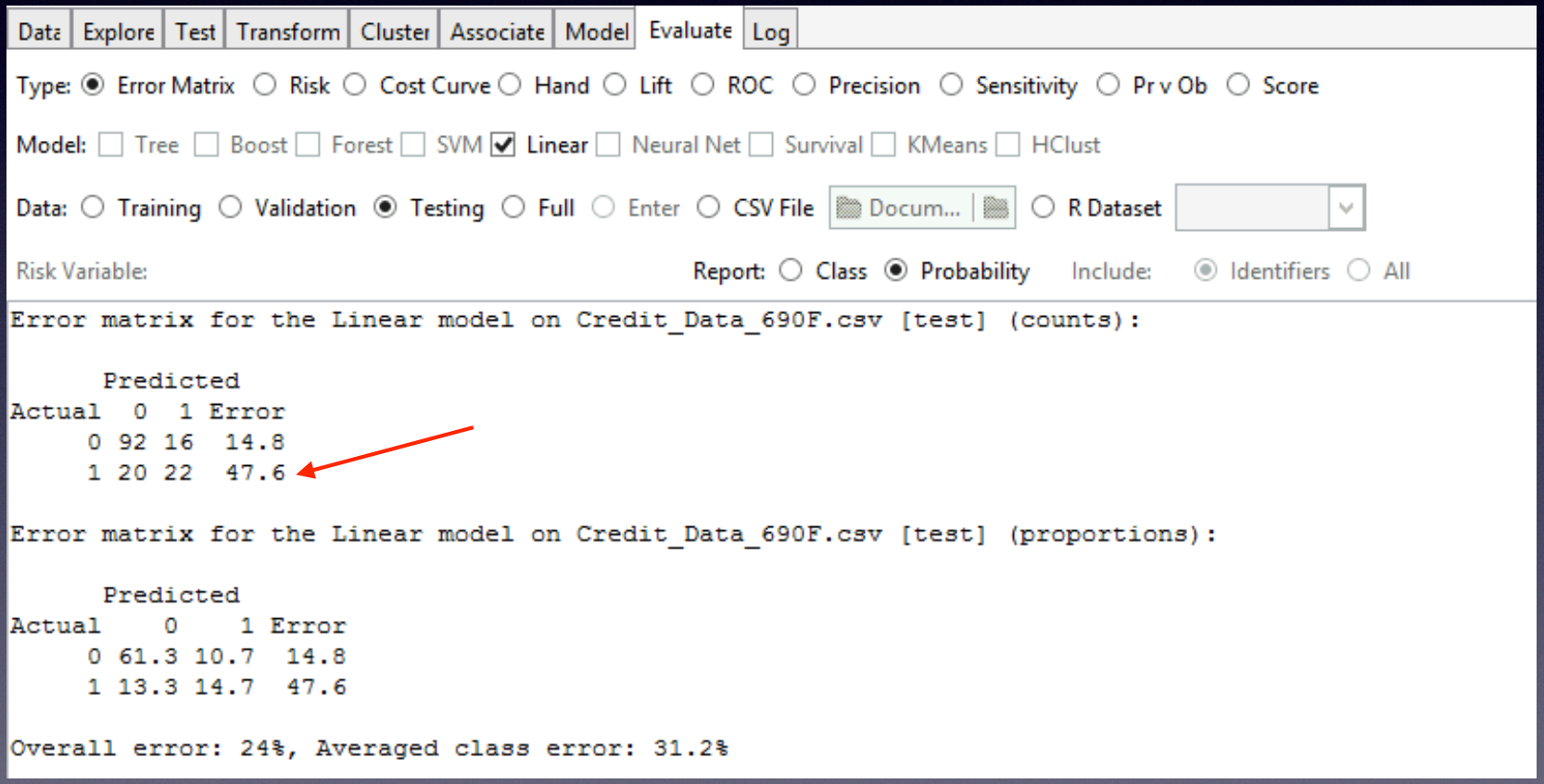
Logistic Regression Model P-values



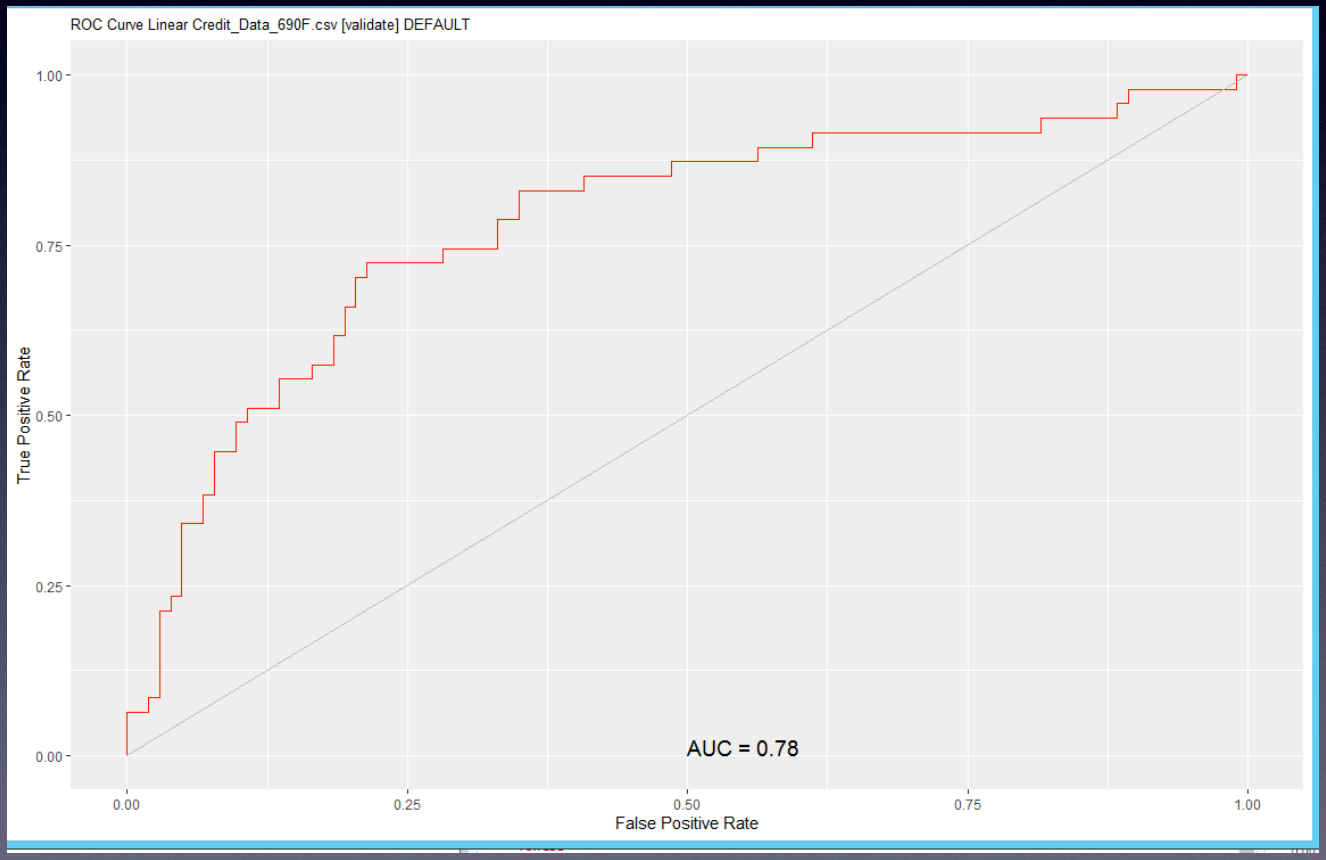
	Pr(>|z|)	
(Intercept)	0.512988	
CHK_ACCT=>200	0.040164	*
CHK_ACCT0<..200	0.169136	
CHK_ACCTno checking account	1.87e-12	***
DURATION	0.000725	***
HISTORYcritical account	7.31e-05	***
HISTORYdelay in paying off in the past	0.005603	**
HISTORYexisting credits paid back duly till now	0.017997	*
HISTORYno credits taken	0.340739	
NEW_CAR	0.003735	**
USED_CAR	0.065726	.
EDUCATION	0.005956	**
AMOUNT	0.020820	*
SAV_ACCT=> 1000	0.002806	**
SAV_ACCT100 <=.. < 500	0.260125	
SAV_ACCT500<=..< 1000	0.309761	
SAV_ACCTunknown/no savings account	0.000296	***
EMPLOYMENT>= 7 years	0.782491	
EMPLOYMENT1 <=..< 4 years	0.475923	
EMPLOYMENT4 <=..< 7 years	0.016863	*
EMPLOYMENTunemployed	0.722613	
INSTALL_RATE	0.004436	**
GUARANTOR	0.041045	*
AGE	0.009941	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

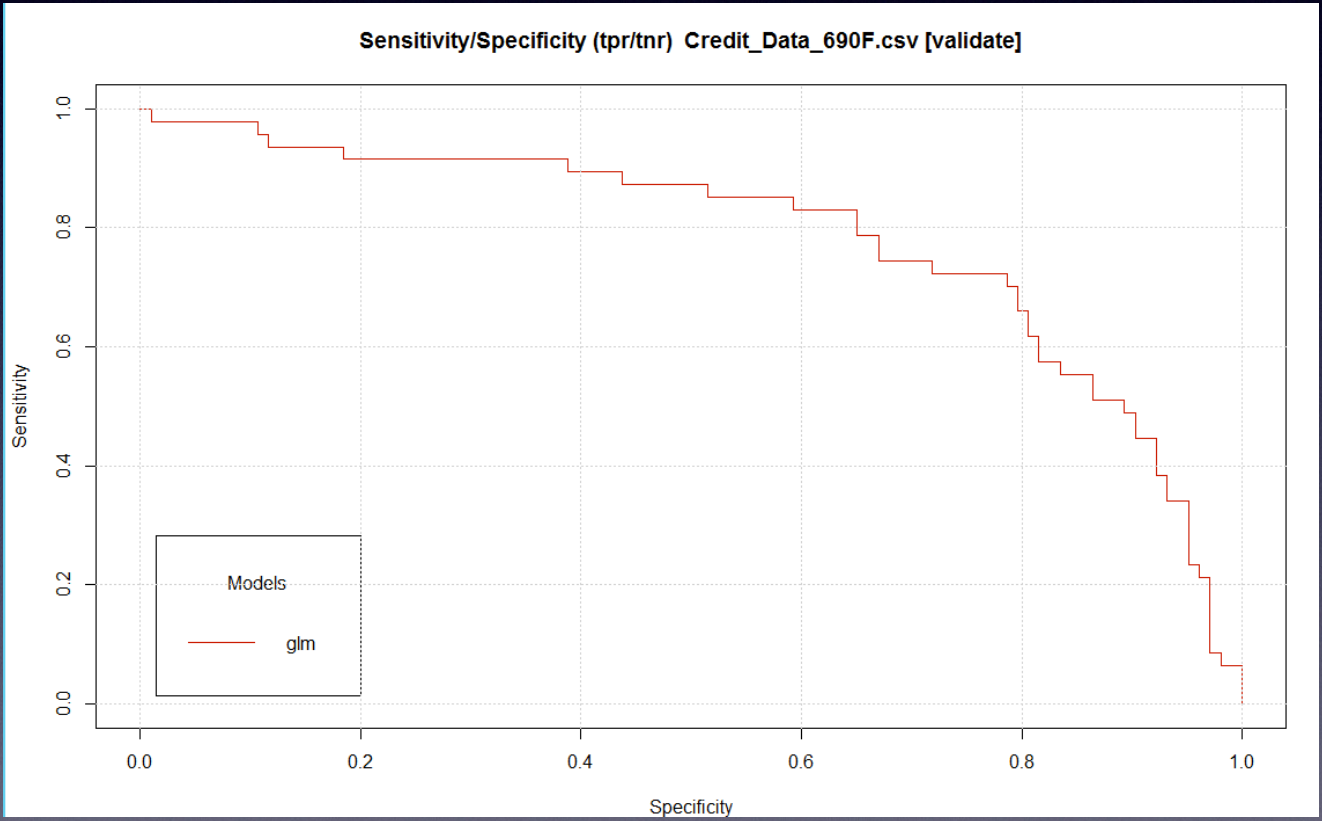
Logistic Regression Model - Error Rates



Logistic Regression Accuracy - ROC & AUC



Logistic Regression Accuracy - Sensitivity & Specificity



Logistic Regression -

Model Collinearity and Interactions

```
> predictions_train2 <- predict(model1, train.data2)
> data.frame(
+     RMSE = RMSE(predictions_train2, train.data2$DEFAULT)
+ )
       RMSE
1 0.3990064
```

```
Credit_Data_Revised <- Credit_Data[ , -c(4:10,14:21,23:30)]
set.seed(123)

training.samples2 <- createDataPartition(Credit_Data_Revised$DEFAULT, p = 0.7, list = FALSE)
train.data2 <- Credit_Data_Revised[training.samples2, ]
test.data2 <- Credit_Data_Revised[-training.samples2, ]

model1 <- glm(DEFAULT ~., data = train.data2)

model1.predictions <- predict(model1, test.data2)


data.frame(
  +     RMSE = RMSE(predictions2, test.data2$DEFAULT)
  + )
RMSE
1 0.5157517

install.packages("car")
trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.6/car_3.0-8.tgz'
Content type 'application/x-gzip' length 1562838 bytes (1.5 MB)
==================================================
  downloaded 1.5 MB


The downloaded binary packages are in
/var/folders/79/zlmlj1bx025b6j6jl65_51200000gn/T//RtmpeE9Phu/downloaded_packages
>
  > car::vif(model1)
GVIF Df GVIF^(1/(2*Df))
CHK_ACCT     1.246218  3        1.037367
DURATION     1.095330  1        1.046580
HISTORY      1.227862  4        1.025991
SAV_ACCT     1.204535  4        1.023534
EMPLOYMENT   1.354156  4        1.038625
INSTALL_RATE 1.073877  1        1.036280
AGE          1.213424  1        1.101555
```

Analytic Conclusion

Based upon the results of both models, we have obtained a detailed understanding of our data and how it can be leveraged in the decision making process at GE credit branches.

More specifically, prior to the idea of initiating an analytic solution, associates were making their approval/disapproval decision upon intuitive notions, rarely, if ever based in supportive data.

Through the development and evaluation of both a random forest and logistic regression model, we have determined that 12 application variables are significantly responsible for influencing the likelihood of an applicant default on their loan.

These 12 variables:
CHK_ACCT, DURATION, AMOUNT, NEW_CAR, USED_CAR, EDUCATION, AMOUNT, SAV_ACCT, EMPLOYMENT, INSTALL_RATE, GUARANTOR, AND AGE
Can now be given the majority of attention during the application determination process, rather than the original 30.

Analytic Conclusion Details

Credit branch associates can utilize the model architecture with each application, by analyzing that particular applications values, in reference to both the model variable coefficient values and any values that reside within specific levels of those values, on the application.

If the values for the application fall within those specific ranges, they can be referenced to the models output, as to whether that value influences an increase in the likelihood of default or not.

Implementation Costs

Implementing this solution will require both monetary and time-invested allotments.

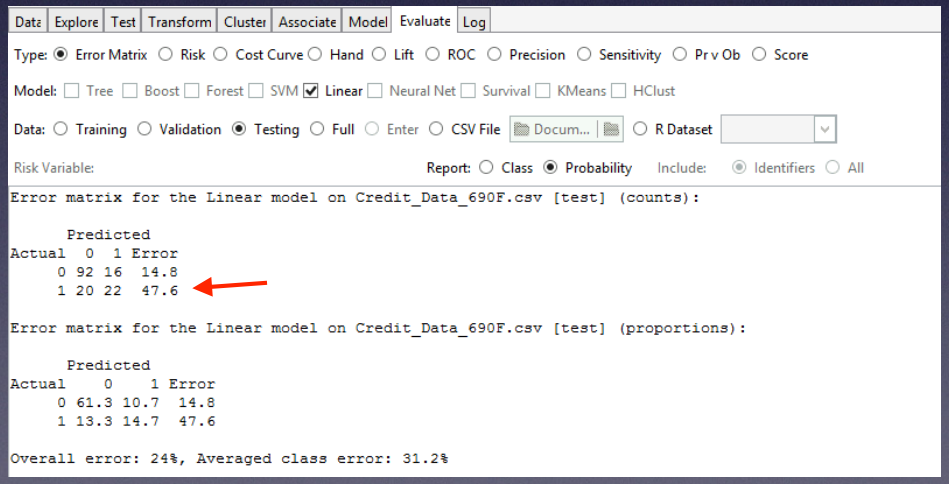
GE will need to invest in the appropriate software, to engage for any instances of future model development and evolution cycles, as well as hiring professionals or outsourcing such talent to conduct such evolutions. The amount required for such ranges in terms of software, licensing and the scope of the professionals salary requirements.

In terms of time-invested, credit branch associates will need to be trained on evaluated the applications based upon the newly established criteria. This training could be outsourced or brought in-house and conducted by the hired data professionals. The more familiar the credit associates are with the logic and the applicability of the model and its utility, the more successful they will be in making accurate, data-driven application decisions.

Analytic Solution ROI

$3938.13 avg. Amount of Credit
X
300 Default Applicants

-$1,181,439 financed to these 300 applicants



Based upon the ~48% error rate established by the logistic regression model testing dataset, we can say that ~52% of them were accurate.

If GE utilizes the analytic solution and the model architecture to make data-driven application processing decisions, there is potential to avoid approving 52% of applications that might've otherwise defaulted.

Analytic Solution ROI Cont'd

To accurate predict or avoid rather, 52% of possible defaulting applications, would save GE incredible amounts of financed funding.

$3938.13 avg. Amount of Credit
X
300 Default Applicants

-$1,181,439 financed to these 300 applicants

-$1,181,439 financed
X
0.52 accurate identification rate

$614,348.28 in avoided funding for potential default-prone applicants

Transferrable Value

In today's advancing markets, making decisions based upon intuitions, hunches or tradition, will no longer serve organizations or individuals well. With the access to information we have today, we are able to provide each industry with data that details the past, the present and the potential for future environments.

Leveraging such data to exploit value in specific situations and circumstances helps reduce organizational risk exposure, maximize financial gain, lower the instability within productivity sectors and make better, more advancing decisions for all parties involved.

Understanding where we have been, what caused events, where we are and what the key points of influencers are currently, will ultimately help us visualize what the future holds and how to effectively prepare for it.

Being unprepared, where the impact of unforeseen loss is incredibly detrimental to forward momentum, will overshadow ones ability to grow, learn and adapt.

The Value of Analytic Solutions

Descriptive statistics and Predictive modeling provide a highly flexible and comprehensive approach to maximizing our understanding of decisions and the impact of such decisions. They manage risk, value and help us prepare for what is to come. Such advantages are highly valuable in any and all industries and markets today and as well progress into the future, they will become more accurate, advanced and more proliferating.

Both descriptive statistics and predictive modeling are capable of not only benefiting large organizations and industries, but also individuals and customers themselves. By collecting, storing, analyzing and leverage data by way of analysis and modeling, they can be protected from malicious discrimination, cyberattacks, financial hijacking and much more. The transparency and insight gained through data analysis provides information to its users and onlookers in bright light, when otherwise, would've remained hidden from view.

References

Bhalia, D. (n.d.). A Complete Guide To Random Forest in R. Retrieved July 24, 2020 from https://

www.listendata.com/2014/11/random-forest-with-r.html#Random-Forest-R-Code

Frost, J. (2020). Understand Precision in Predictive Analytics to Avoid Costly Mistakes. Retrieved July 23, 2020 from https://statisticsbyjim.com/regression/prediction-precision-applied-regression/

Frost, J. (2020). Multicollinearity in Regression Analysis: Problems, Detection, and Solutions. Retrieved July 24, 2020 from https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/#:~:text=Fortunately%2C%20there%20is%20a%20very,VIF%20for%20each%20independent%20variabl e.

Enders, F. (2020). Encyclopedia Britannica. Collinearity. Retrieved July 22, 2020 from https:// www.britannica.com/topic/collinearity-statistics

Grace-Martin, K. (2020). Assessing the Fit of Regression Models. The Analysis Factor. Retrieved July 24, 2020 from https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/

Grace-Martin, K. (2020). Interpreting Regression Coefficients. Retrieved July 24, 2020 from https:// www.theanalysisfactor.com/interpreting-regression-coefficients/

Grace-Martin, K. (2020). Interpreting Interactions in Regression. Retrieved July 24, 2020 from https:// www.theanalysisfactor.com/interpreting-interactions-in-regression/

Grace-Martin, K. (2020). Clarifications on Interpreting Interactions in Regression. Retrieved July 25, 2020 from https://www.theanalysisfactor.com/clarifications-on-interpreting-interactions-in-regression/

Grace-Martin, K. (2020). 7 Practical Guidelines for Accurate Statistical Model Building. Retrieved July 22, 2020 from https://www.theanalysisfactor.com/7-guidelines-model-building/

Machine Learning Crash Course. (2020). Classification: ROC Curve and AUC. Retrieved July 25, 2020 from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Narkhede, S. (2018). Understanding AUC - ROC curve. Retrieved May 25, 2020 from https:// towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Kassambara. (2018). STHDA. Multicollinearity Essentials and VIF in R. Retrieved July 23, 2020 from http://www.sthda.com/english/articles/39-regression-model-diagnostics/160-multicollinearity-essentials- and-vif-in-r/

SydneyF. (2018). Alteryx - Help!... Mean Decrease in Gini for dummies. Retrieved May 27, 2020 from https://community.alteryx.com/t5/Alteryx-Designer-Discussions/Help-Mean-Decrease-in-Gini-for-dummies/td-p/197223

Tozzi, C. (2020). Does Your Data Measure Up? How to Assess Data Quality. Precisely. Retrieved July 23, 2020 from https://www.precisely.com/blog/data-quality/does-your-data-measure-up-assess-data-quality? utm_medium=Redirect-Syncsort&utm_source=Direct-Traffic

Widjaja, J. (2017). How do you explain ‘mean decrease accuracy’ and ‘mean decrease gini’ in layman’s terms? Retrieved May 27, 2020 from https://www.quora.com/How-do-you-explain-%E2%80%98mean- decrease-accuracy%E2%80%99-and-%E2%80%98mean-decrease-gini%E2%80%99-in- layman%E2%80%99s-terms

Williams, G. (2011). Data Mining with Rattle and R. The Art of Excavating Data for Knowledge Discovery. Springer Science+Business Media, LLC. Random Forests. Retrieved May 24, 2020.