

Home Credit: Final Report

James Matheson

February 24, 2019

Introduction

Assessing the Problem

Traditionally, borrowing costs have been tied to creditors' assessment of credit default risk based on simple, but broad financial criteria. Unfortunately, this leaves many potential borrowers out of the market, or paying higher interest rates. Conversely, these same criteria may not always be sufficient to reduce default risk by clients who may meet the criteria, but may nonetheless eventually default because of other, potentially foreseeable, reasons not sufficiently accounted for in the criteria.

Client Solution

Using historical client credit data, I will create a Machine Learning model capable of predicting, as accurately as possible, an individual applicant's likelihood of defaulting. This model can then be used by Home Credit to make more refined decisions as to whom they offer loans, the types of loans offered and the interest terms.

Steps Taken

1. Using R, cleaned and merged the "train" data to get a single data frame for incorporation into ML algorithms;
2. Examined the data using P-tests to determine significant features for incorporation into the ML algorithms;
3. Reserved 80% of the training data for training the algorithms and 20% for testing;
4. Using R's Caret package, trained several different models:
 - a. General Linear Model, using 100% of the apportioned train data
 - b. Naive Bayes model, using 100% of the apportioned train data
 - c. K Nearest Neighbor Model, using 100% of the apportioned train data
 - d. Random Forest Model, with the data segmented into 5%, 10%, 20%, 30%, 60% and 80% groups for comparison, and as a check against overfitting
5. Used visualization tools to demonstrate these correlations and the accuracy of the model.

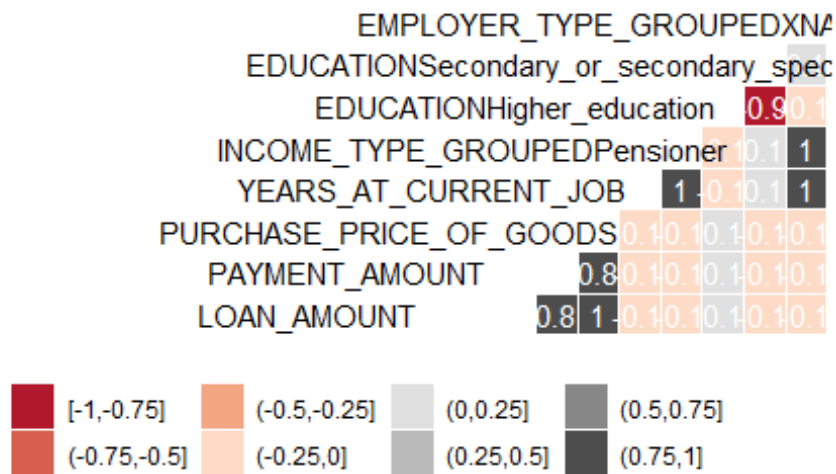
Data Cleanup and Feature Selection

1. I examined P values for each potential feature vs. TARGET to assist in determining its relevance and significance for use in ML models;
2. I used a correlation chart to determine highly correlated (and therefore redundant) features for removal so that they won't be overrepresented in the ML trainers at later steps;
3. I supplied a plot for each variable that demonstrates simple counts for each categorical variable or binned continuous variable;
4. I supplied a plot demonstrating the significance of each variable to the TARGET variable;
5. I supplied summary statistical information for each variable;
 - for continuous features, I used the Shapiro-Wilks Test (`shapiro.test()`) to test for distributive normality;
 - upon finding that my continuous features are not normally distributed, I applied the Wilcoxon Rank Sum Test (`wilcox.test()`) as the t-test assumption of normality is not met.

Correlation matrix with ONLY correlation values > .6

The full correlation matrix for the features of this data set is too large to legibly display here. Therefore, a simplified correlation matrix follows, which displays ONLY correlation values between features that are greater than 0.6.

Based on this correlation matrix, I removed three features from the model: INCOME_TYPE_GROUPEDPensioner, EMPLOYER_TYPE_GROUPEDXNA, and EDUCATIONHigher_education.



Summary of P values for each feature

The following table of P-values will be used for final feature selection in our ML models. Of particular note are the “Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1” which demonstrate the significance of the feature based on its P-value.

```
##
## Call:
## glm(formula = TARGET ~ ., family = "binomial", data = results_train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6034  -0.4391  -0.3328  -0.2487   3.2870
##
```

```

## Coefficients:
##
## Estimate Std. Error z value
## (Intercept) -1.976e+00 6.090e-01 -3.244
## SK_ID_CURR -4.553e-08 6.635e-08 -0.686
## LOAN_TYPERevolving loans -3.389e-01 2.879e-02 -11.773
## AGE -8.499e-03 8.860e-04 -9.593
## GENDERM 3.612e-01 1.553e-02 23.251
## GENDERXNA -6.500e+00 3.500e+01 -0.186
## OWNS_CARY -3.862e-01 2.066e-02 -18.688
## AGE_OF_CAR 6.053e-03 9.512e-04 6.363
## OWNS_REALTY 4.839e-02 1.539e-02 3.145
## CHILDREN -7.783e-03 1.008e-02 -0.772
## TOTAL_INCOME 1.362e-08 1.829e-08 0.745
## LOAN_AMOUNT 2.276e-06 1.265e-07 17.998
## PAYMENT_AMOUNT 1.011e-05 1.351e-06 7.478
## PURCHASE_PRICE_OF_GOODS -2.898e-06 1.219e-07 -23.784
## RATIO_LOAN_TO_ANNUITY 4.653e-04 2.136e-03 0.218
## EDUCATIONHigher education 1.194e+00 5.903e-01 2.022
## EDUCATIONIncomplete higher 1.357e+00 5.912e-01 2.295
## EDUCATIONLower secondary 1.732e+00 5.925e-01 2.924
## EDUCATIONSecondary / secondary special 1.597e+00 5.901e-01 2.706
## MARITAL_STATUSMarried -1.629e-01 2.220e-02 -7.336
## MARITAL_STATUSSeparated -2.428e-02 3.357e-02 -0.723
## MARITAL_STATUSSingle / not married -5.724e-02 2.626e-02 -2.180
## MARITAL_STATUSUnknown -8.218e+00 5.087e+01 -0.162
## MARITAL_STATUSWidow -1.487e-01 4.134e-02 -3.596
## HOUSING_STATUSHouse / apartment 3.332e-02 1.141e-01 0.292
## HOUSING_STATUSMunicipal apartment 1.771e-01 1.192e-01 1.485
## HOUSING_STATUSOffice apartment -1.728e-01 1.397e-01 -1.237
## HOUSING_STATUSRented apartment 1.811e-01 1.225e-01 1.478
## HOUSING_STATUSWith parents 8.831e-02 1.170e-01 0.755
## YEARS_AT_CURRENT_JOB -3.329e-02 1.566e-03 -21.267
## YEARS_SINCE_GETTING_IDENTITY_DOCUMENT -5.066e-03 8.003e-04 -6.330
## REGION_AND_CITY_RATING 1.757e-01 1.432e-02 12.271
## External.Score.1 2.317e-03 2.393e-02 0.097
## External.Score.2 -2.239e+00 3.436e-02 -65.146
## External.Score.3 -1.145e+00 2.570e-02 -44.541
## MAX_DAYS_LATE_BUREAU 1.015e-04 7.516e-05 1.350
## INCOME_TYPE_GROUPEDOther 1.042e-01 7.555e-01 0.138
## INCOME_TYPE_GROUPEDPensioner -2.756e+00 8.834e-01 -3.120
## INCOME_TYPE_GROUPEDState servant -6.391e-02 3.648e-02 -1.752
## INCOME_TYPE_GROUPEDWorking 1.276e-01 1.743e-02 7.318
## EMPLOYER_TYPE_GROUPEDBank -4.612e-01 1.153e-01 -4.000
## EMPLOYER_TYPE_GROUPEDBusiness Entity -7.900e-02 6.944e-02 -1.138
## EMPLOYER_TYPE_GROUPEDEducation -2.483e-01 7.596e-02 -3.269
## EMPLOYER_TYPE_GROUPEDElectricity -2.934e-01 1.507e-01 -1.947
## EMPLOYER_TYPE_GROUPEDGovt Services -2.546e-01 7.559e-02 -3.368
## EMPLOYER_TYPE_GROUPEDHousing 2.960e-02 7.643e-02 0.387
## EMPLOYER_TYPE_GROUPEDIndustry -1.897e-01 7.479e-02 -2.536
## EMPLOYER_TYPE_GROUPEDMedicine -1.920e-01 7.908e-02 -2.428

```

```

## EMPLOYER_TYPE_GROUPEDOther          -1.397e-01  7.458e-02  -1.873
## EMPLOYER_TYPE_GROUPEDSelf-employed   2.126e-02  7.048e-02   0.302
## EMPLOYER_TYPE_GROUPEDService         -1.604e-01  7.565e-02  -2.121
## EMPLOYER_TYPE_GROUPEDTrade           -1.171e-01  7.483e-02  -1.564
## EMPLOYER_TYPE_GROUPEDTransport       -4.745e-02  7.761e-02  -0.611
## EMPLOYER_TYPE_GROUPEDXNA             3.554e+01  1.787e+00  19.886
##                                     Pr(>|z|)
## (Intercept)                          0.001179 **
## SK_ID_CURR                           0.492578
## LOAN_TYPERevolving loans              < 2e-16 ***
## AGE                                  < 2e-16 ***
## GENDERM                              < 2e-16 ***
## GENDERXNA                            0.852673
## OWNS_CARY                            < 2e-16 ***
## AGE_OF_CAR                           1.98e-10 ***
## OWNS_REALTYT                          0.001659 **
## CHILDREN                             0.439990
## TOTAL_INCOME                         0.456509
## LOAN_AMOUNT                          < 2e-16 ***
## PAYMENT_AMOUNT                       7.55e-14 ***
## PURCHASE_PRICE_OF_GOODS              < 2e-16 ***
## RATIO_LOAN_TO_ANNUITY                 0.827594
## EDUCATIONHigher education             0.043168 *
## EDUCATIONIncomplete higher            0.021708 *
## EDUCATIONLower secondary              0.003455 **
## EDUCATIONSecondary / secondary special 0.006816 **
## MARITAL_STATUSMarried                 2.20e-13 ***
## MARITAL_STATUSSeparated               0.469508
## MARITAL_STATUSSingle / not married    0.029275 *
## MARITAL_STATUSUnknown                 0.871674
## MARITAL_STATUSWidow                   0.000323 ***
## HOUSING_STATUSHouse / apartment       0.770330
## HOUSING_STATUSMunicipal apartment     0.137599
## HOUSING_STATUSOffice apartment        0.216014
## HOUSING_STATUSRented apartment        0.139517
## HOUSING_STATUSWith parents             0.450204
## YEARS_AT_CURRENT_JOB                  < 2e-16 ***
## YEARS_SINCE_GETTING_IDENTITY_DOCUMENT 2.45e-10 ***
## REGION_AND_CITY_RATING                < 2e-16 ***
## External.Score.1                      0.922857
## External.Score.2                      < 2e-16 ***
## External.Score.3                      < 2e-16 ***
## MAX_DAYS_LATE_BUREAU                  0.177062
## INCOME_TYPE_GROUPEDOther              0.890308
## INCOME_TYPE_GROUPEDPensioner          0.001808 **
## INCOME_TYPE_GROUPEDState servant       0.079727 .
## INCOME_TYPE_GROUPEDWorking            2.52e-13 ***
## EMPLOYER_TYPE_GROUPEDBank             6.34e-05 ***
## EMPLOYER_TYPE_GROUPEDBusiness Entity  0.255276
## EMPLOYER_TYPE_GROUPEDEducation        0.001078 **

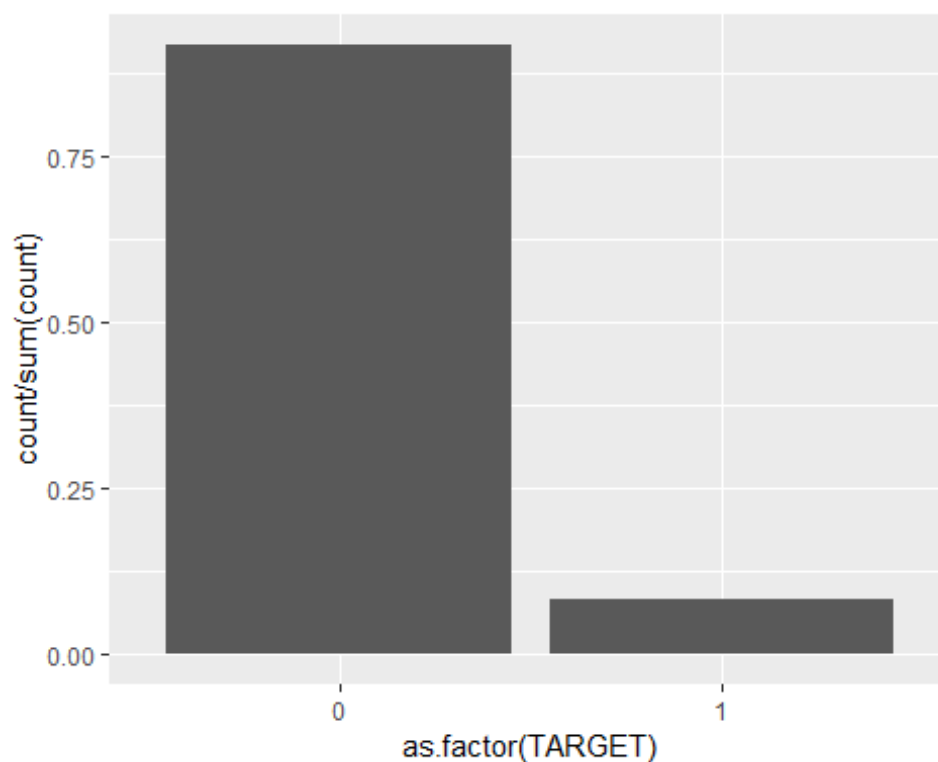
```

```
## EMPLOYER_TYPE_GROUPEDElectricity      0.051515 .
## EMPLOYER_TYPE_GROUPEDGovt Services    0.000757 ***
## EMPLOYER_TYPE_GROUPEDHousing          0.698551
## EMPLOYER_TYPE_GROUPEDIndustry         0.011197 *
## EMPLOYER_TYPE_GROUPEDMedicine         0.015178 *
## EMPLOYER_TYPE_GROUPEDOther            0.061041 .
## EMPLOYER_TYPE_GROUPEDSelf-employed    0.762938
## EMPLOYER_TYPE_GROUPEDService          0.033954 *
## EMPLOYER_TYPE_GROUPEDTrade            0.117726
## EMPLOYER_TYPE_GROUPEDTransport        0.540877
## EMPLOYER_TYPE_GROUPEDXNA              < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 172443  on 307383  degrees of freedom
## Residual deviance: 157879  on 307330  degrees of freedom
## (127 observations deleted due to missingness)
## AIC: 157987
##
## Number of Fisher Scoring iterations: 8
```

Distribution of the Dependent Variable “Target”

The following plot displays a simple distribution of our TARGET values. TARGET = 1 means that the sample had some problem repaying their loan. TARGET = 0 means that the sample successfully repaid their loan without issue.

Note that because the TARGET data is highly unbalanced, I downsampled the data sets when training each model.

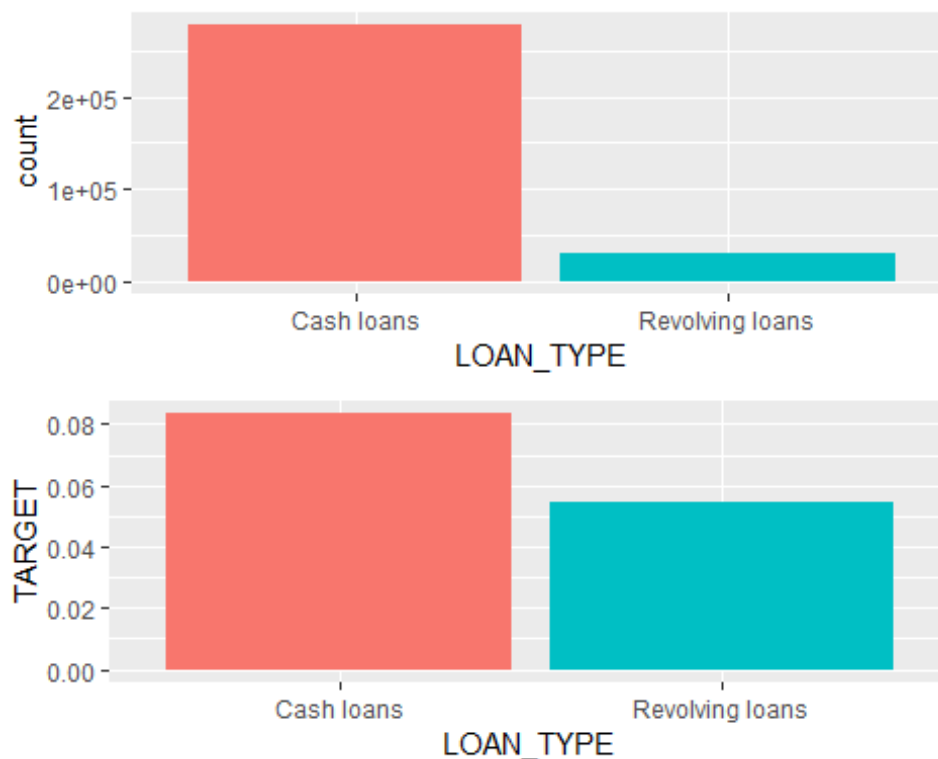


```
# A tibble: 2 x 2
  TARGET      n
  <int> <int>
1     0 282686
2     1  24825
```

Feature Significance

The following statistics and plots demonstrate both simple counts for each feature, as well as the significance of each feature to our ML models.

Loan Type and Loan Type v. Target



```
# A tibble: 2 x 2
```

```
  LOAN_TYPE      n
  <chr>      <int>
1 Cash loans 278232
2 Revolving loans 29279
```

Call:

```
glm(formula = TARGET ~ LOAN_TYPE, family = "binomial", data = results_train)
```

Deviance Residuals:

```
      Min       1Q   Median       3Q      Max
-0.4175 -0.4175 -0.4175 -0.4175  2.4101
```

Coefficients:

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.396250   0.006855  -349.6   <2e-16 ***
LOAN_TYPERevolving loans -0.451779   0.026581  -17.0   <2e-16 ***
---
```

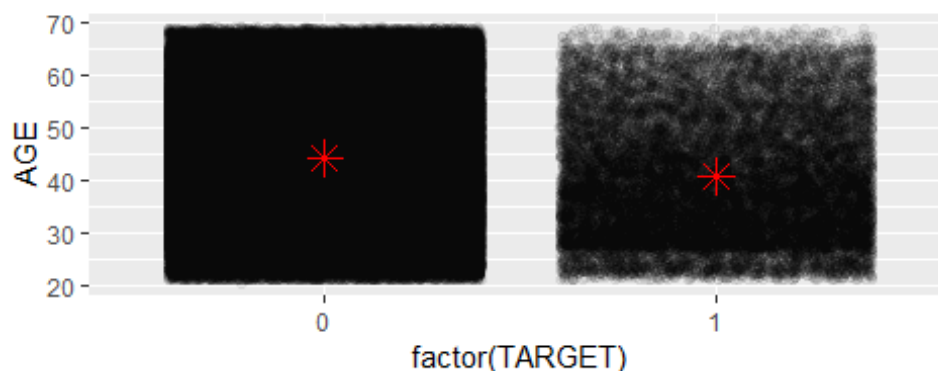
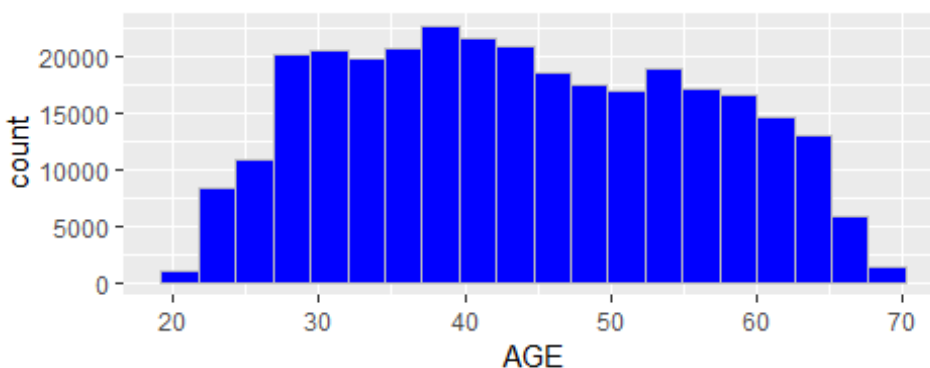

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 172542 on 307510 degrees of freedom
Residual deviance: 172217 on 307509 degrees of freedom
AIC: 172221

Number of Fisher Scoring iterations: 5

Age and Age v. Target



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
20.50	33.98	43.12	43.91	53.89	69.07

Anderson-Darling normality test

data: results_train\$AGE
A = 2381, p-value < 2.2e-16

Wilcoxon rank sum test with continuity correction

data: AGE by TARGET

W = 4091300000, p-value < 2.2e-16

alternative hypothesis: true location shift is not equal to 0

Gender and Gender v. Target

Note that, because the number of XNA values for gender was extremely low (total count of 4), they have been removed in the following.



```
# A tibble: 2 x 2
  GENDER      n
  <fct>    <int>
1 F      202373
2 M      105007
```

Call:

```
glm(formula = TARGET ~ GENDER, family = "binomial", data = results_train,
     na.action = na.omit)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.4625	-0.4625	-0.3810	-0.3810	2.3062

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.586793	0.008711	-296.95	<2e-16 ***
GENDERM	0.405238	0.013429	30.18	<2e-16 ***

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

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 172541 on 307506 degrees of freedom
Residual deviance: 171649 on 307505 degrees of freedom
(4 observations deleted due to missingness)
AIC: 171653
```

Number of Fisher Scoring iterations: 5

Owns a car? and Owns Car v. Target



```
# A tibble: 2 x 2
  OWNS_CAR      n
  <chr>    <int>
1 N      202924
2 Y      104587
```

```
Call:
glm(formula = TARGET ~ OWNS_CAR, family = "binomial", data = results_train)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-0.4215 -0.4215 -0.4215 -0.3878 2.2913
```

Coefficients:

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.37624    0.00796 -298.53  <2e-16 ***
OWNS_CARY   -0.17359    0.01434 -12.11  <2e-16 ***
```

```
---
```

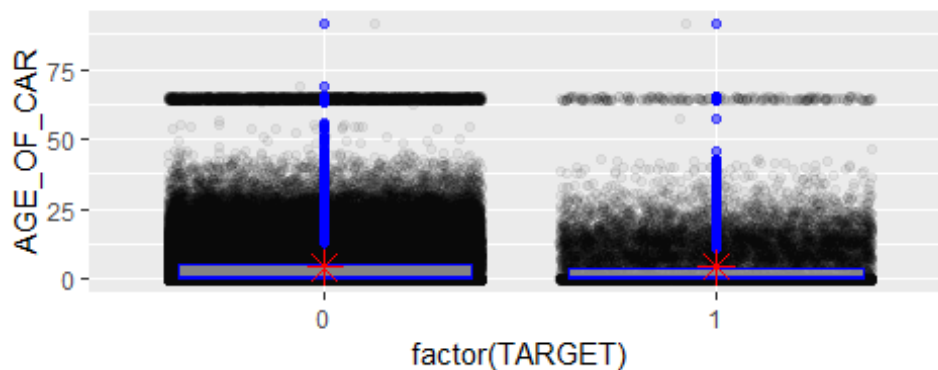
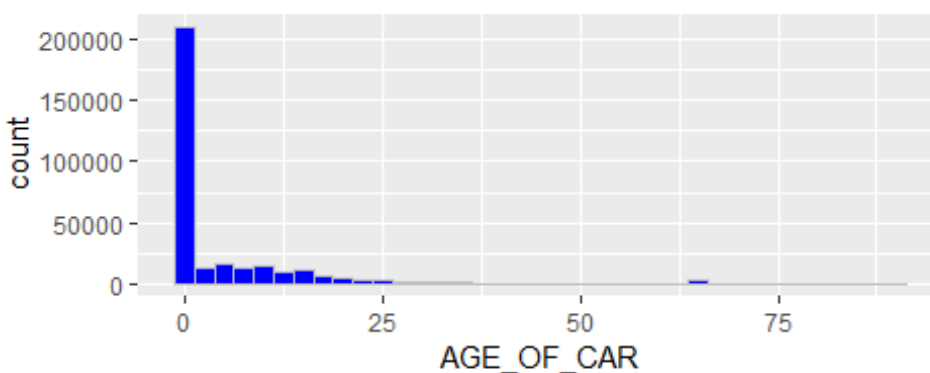
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 172542 on 307510 degrees of freedom
Residual deviance: 172393 on 307509 degrees of freedom
AIC: 172397
```

Number of Fisher Scoring iterations: 5

Age of car and Age of Car v. Target



```
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.000 0.000 0.000 4.102 5.000 91.000
```

```
# A tibble: 62 x 2
```

```
  `factor(AGE_OF_CAR)`      n
  <fct>                <int>
1 0                    205063
```

```

2 1 5280
3 2 5852
4 3 6370
5 4 5557
6 5 3595
7 6 6382
8 7 7424
9 8 5887
10 9 5020
# ... with 52 more rows

```

Anderson-Darling normality test

```

data: results_train$AGE_OF_CAR
A = 50400, p-value < 2.2e-16

```

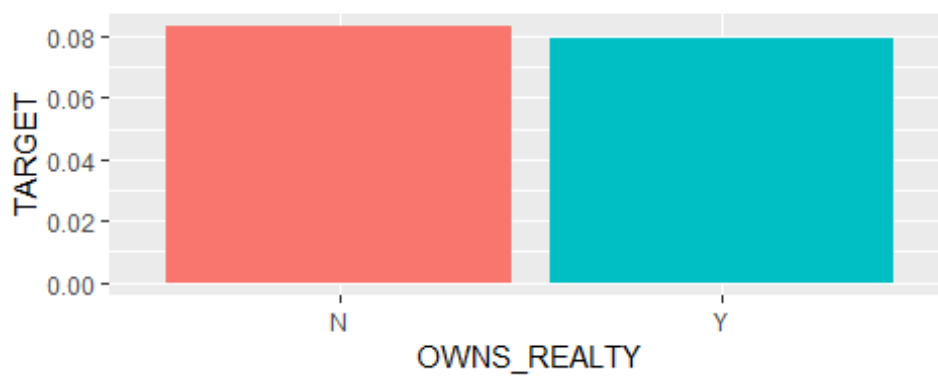
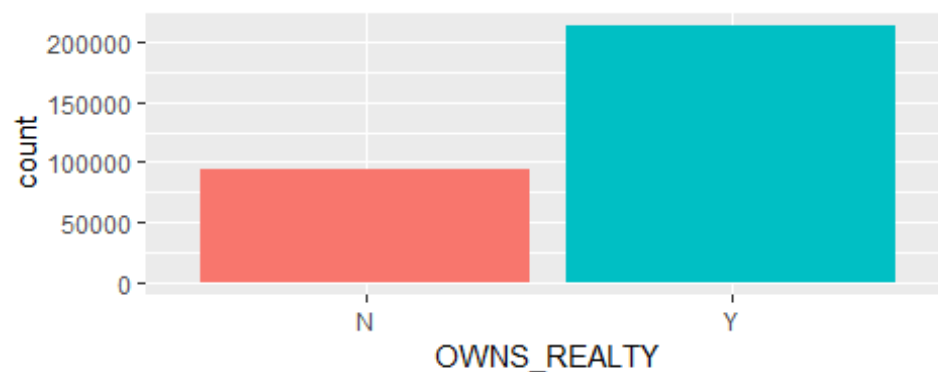
Wilcoxon rank sum test with continuity correction

```

data: AGE_OF_CAR by TARGET
W = 3595300000, p-value = 1.451e-14
alternative hypothesis: true location shift is not equal to 0

```

Owns Real Estate? and Owns RE v. Target



```
# A tibble: 2 x 2
  OWNS_REALTY      n
  <chr>         <int>
1 N             94199
2 Y             213312
```

Call:
 glm(formula = TARGET ~ OWNS_REALTY, family = "binomial", data = results_train)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.4169	-0.4169	-0.4073	-0.4073	2.2497

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.39900	0.01179	-203.408	< 2e-16 ***
OWNS_REALTY	-0.04858	0.01425	-3.409	0.000651 ***

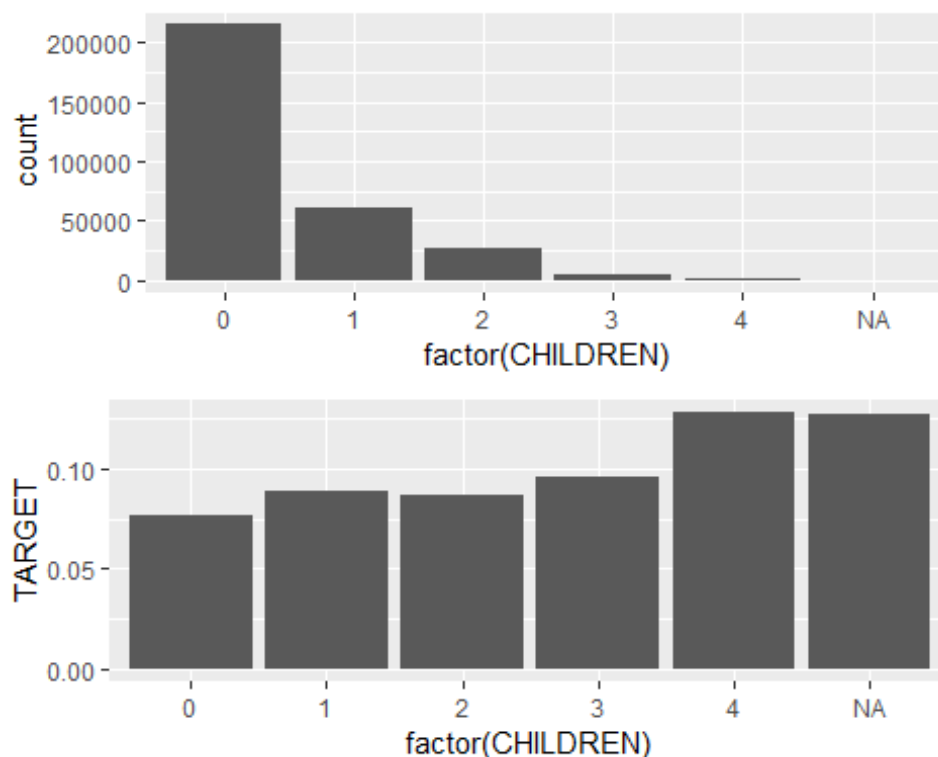
 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 172542 on 307510 degrees of freedom
 Residual deviance: 172530 on 307509 degrees of freedom
 AIC: 172534

Number of Fisher Scoring iterations: 5

Number of Children and Number of Children v. Target



```
# A tibble: 6 x 2
  CHILDREN     n
  <int> <int>
1      0 215371
2      1  61119
3      2  26749
4      3   3717
5      4    429
6     NA    126
```

Call:

```
glm(formula = TARGET ~ CHILDREN, family = "binomial", data = results_train)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.4804	-0.4207	-0.4024	-0.4024	2.2602

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.473205	0.007764	-318.54	<2e-16 ***
CHILDREN	0.093030	0.008893	10.46	<2e-16 ***

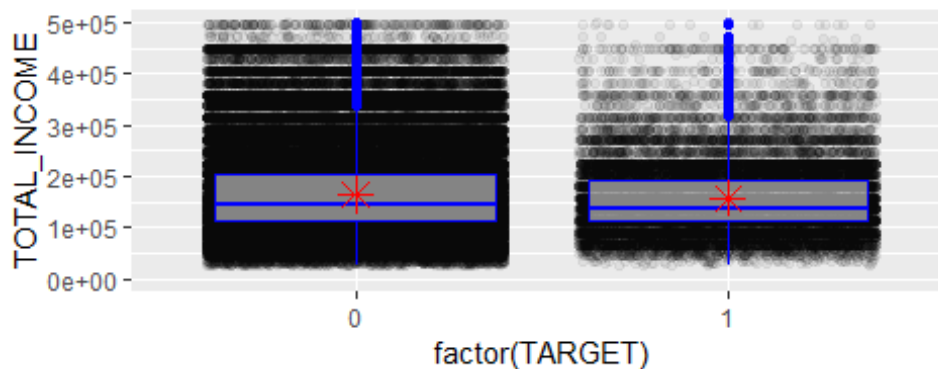
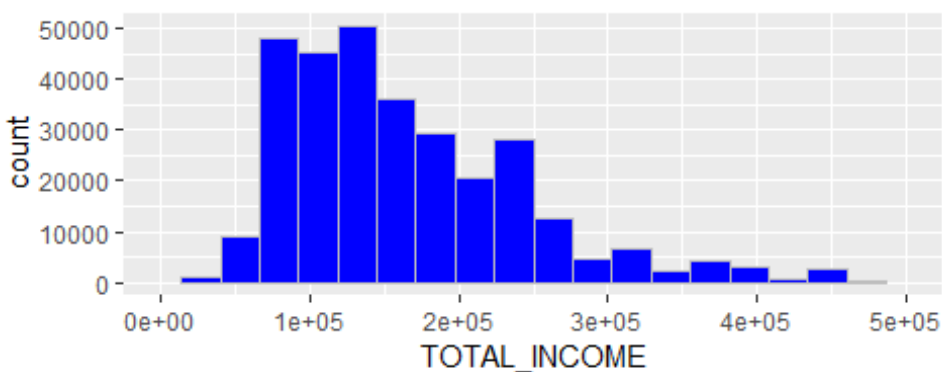
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 172443 on 307384 degrees of freedom
 Residual deviance: 172337 on 307383 degrees of freedom
 (126 observations deleted due to missingness)
 AIC: 172341

Number of Fisher Scoring iterations: 5

Total Income and Total Income v. Target



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
25650	112500	147150	168798	202500	117000000

Anderson-Darling normality test

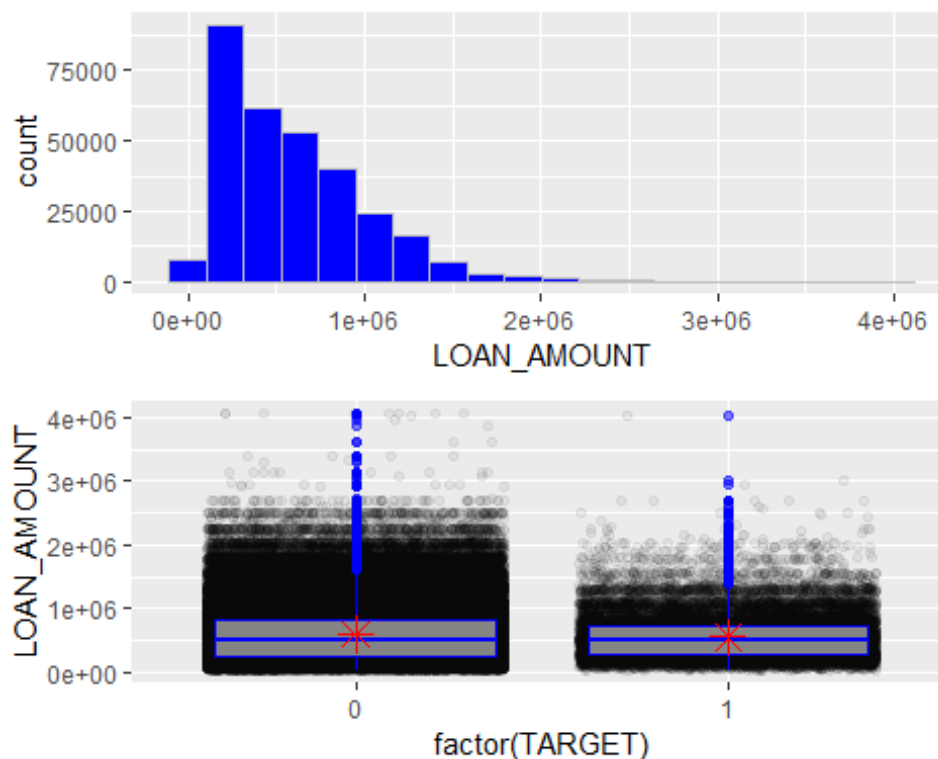
data: results_train\$TOTAL_INCOME
 A = 49822, p-value < 2.2e-16

Wilcoxon rank sum test with continuity correction

data: TOTAL_INCOME by TARGET


```
W = 3643100000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Amount of Loan and Amount of Loan v. Target



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
45000	270000	513531	599026	808650	4050000

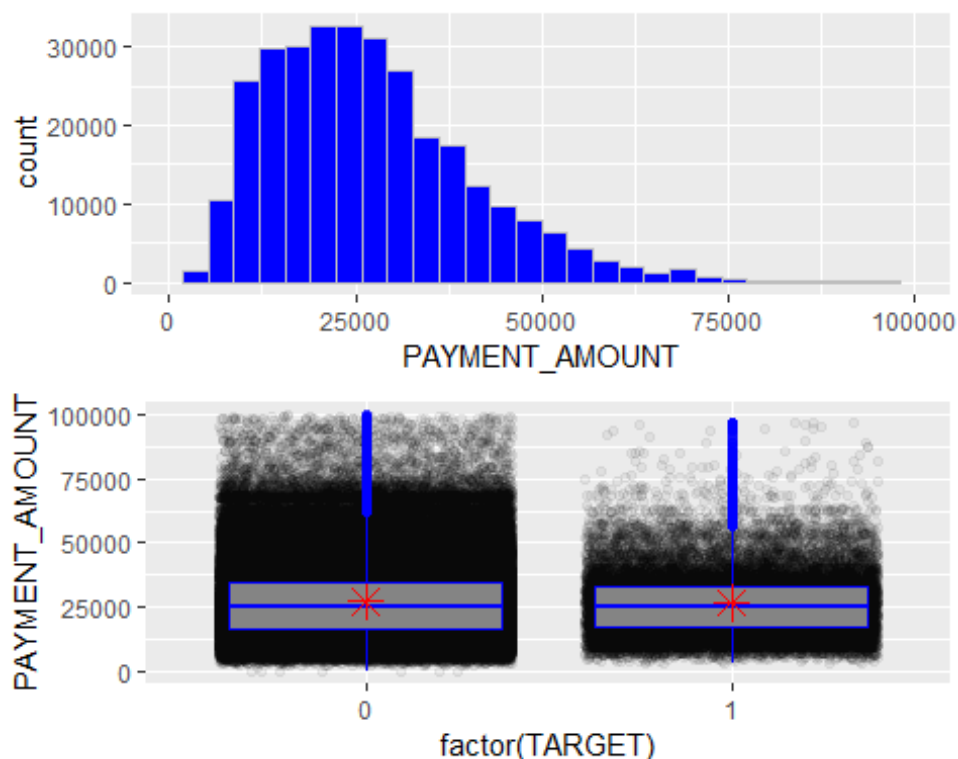
Anderson-Darling normality test

```
data: results_train$LOAN_AMOUNT
A = 7249.4, p-value < 2.2e-16
```

Wilcoxon rank sum test with continuity correction

```
data: LOAN_AMOUNT by TARGET
W = 3639200000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Monthly Payment and Monthly Payment v. Target



	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	16524	24903	27108	34596	258026	

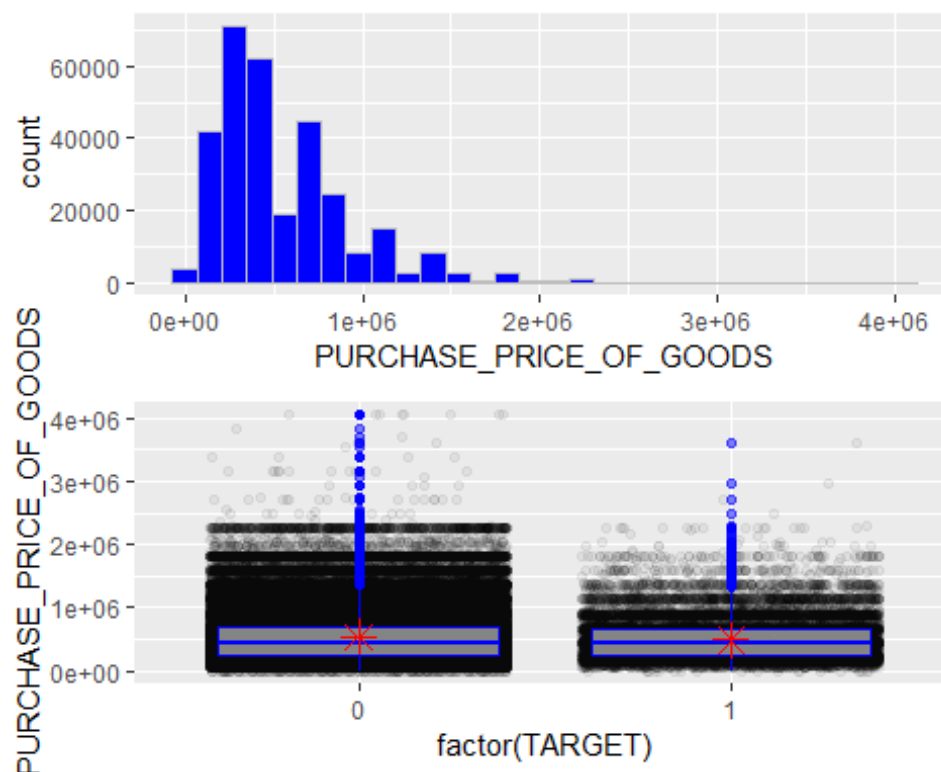
Anderson-Darling normality test

data: results_train\$PAYMENT_AMOUNT
A = 4118.6, p-value < 2.2e-16

Wilcoxon rank sum test with continuity correction

data: PAYMENT_AMOUNT by TARGET
W = 3509200000, p-value = 0.9764
alternative hypothesis: true location shift is not equal to 0

Price of Goods Purchased with the Loan and Price v. Target



	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	238500	450000	537909	679500	4050000	

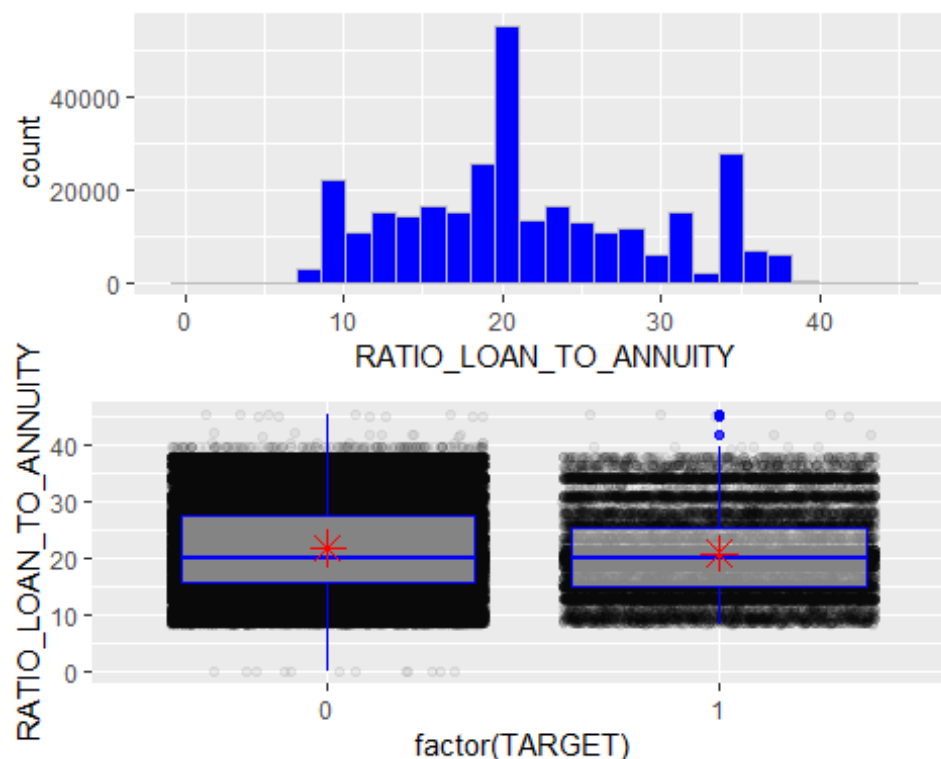
Anderson-Darling normality test

data: results_train\$PURCHASE_PRICE_OF_GOODS
A = 8872.1, p-value < 2.2e-16

Wilcoxon rank sum test with continuity correction

data: PURCHASE_PRICE_OF_GOODS by TARGET
W = 3742200000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

Ratio of Loan to Payment Amount and Ratio v. Target



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	15.61	20.00	21.61	27.10	45.31

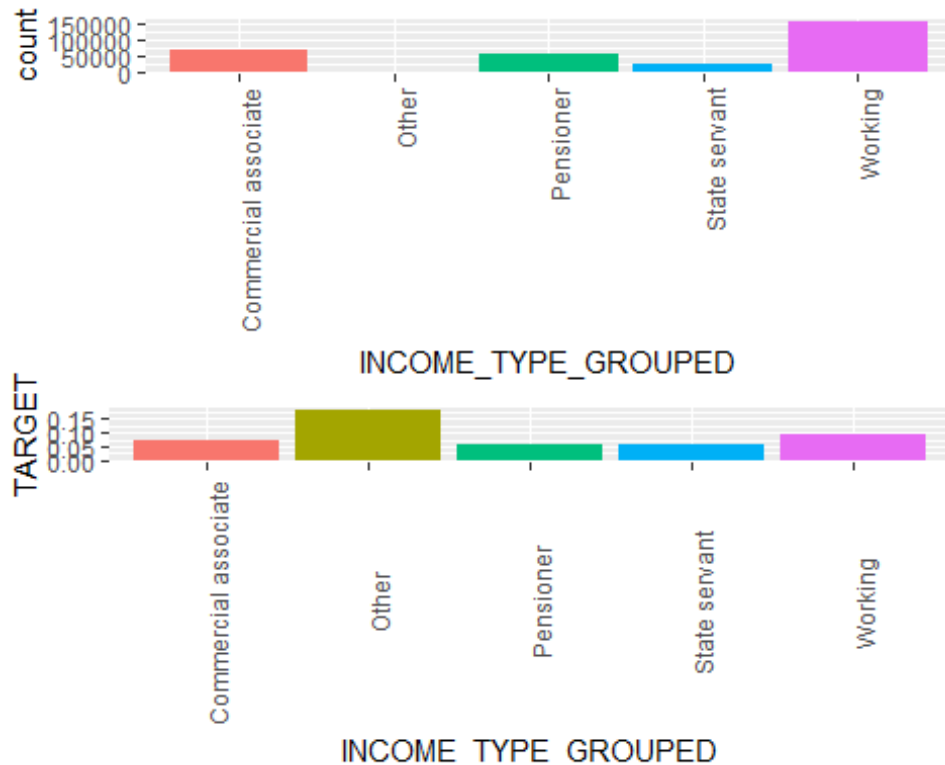
Anderson-Darling normality test

data: results_train\$RATIO_LOAN_TO_ANNUIITY
A = 3810.3, p-value < 2.2e-16

Wilcoxon rank sum test with continuity correction

data: RATIO_LOAN_TO_ANNUIITY by TARGET
W = 3733600000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

Type of Income and Type of Income v. Target



A tibble: 5 x 2

INCOME_TYPE_GROUPED	n
1 Commercial associate	71617
2 Other	55
3 Pensioner	55362
4 State servant	21703
5 Working	158774

Call:

```
glm(formula = TARGET ~ INCOME_TYPE_GROUPED, family = "binomial",
     data = results_train)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.6335	-0.4490	-0.3944	-0.3328	2.4171

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.51458	0.01420	-177.074	< 2e-16 ***
INCOME_TYPE_GROUPEDOther	1.01050	0.34989	2.888	0.00388 **
INCOME_TYPE_GROUPEDPensioner	-0.35135	0.02358	-14.900	< 2e-16 ***
INCOME_TYPE_GROUPEDState servant	-0.28126	0.03242	-8.675	< 2e-16 ***

```
INCOME_TYPE_GROUPEDWorking      0.27077    0.01656   16.348 < 2e-16 ***
```

```
---
```

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

```
(Dispersion parameter for binomial family taken to be 1)
```

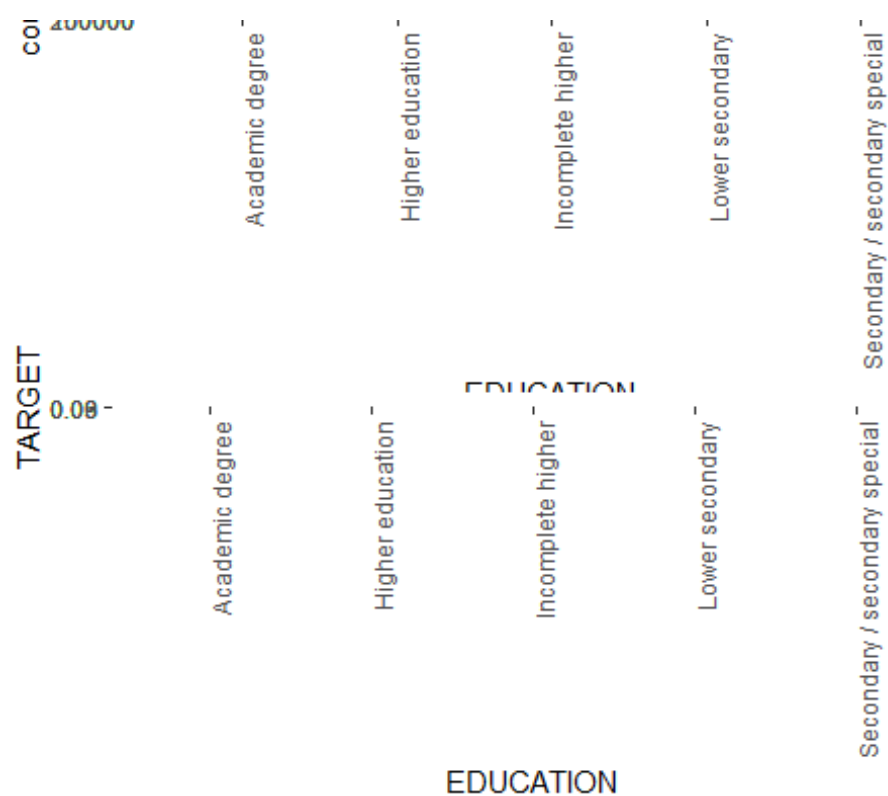
```
Null deviance: 172542 on 307510 degrees of freedom
```

```
Residual deviance: 171258 on 307506 degrees of freedom
```

```
AIC: 171268
```

```
Number of Fisher Scoring iterations: 5
```

Education Level and Education Level v. Target



```
# A tibble: 5 x 2
```

```
  EDUCATION      n
  <chr>      <int>
1 Academic degree    164
2 Higher education  74863
3 Incomplete higher  10277
4 Lower secondary   3816
5 Secondary / secondary special 218391
```

```
Call:
```

```
glm(formula = TARGET ~ EDUCATION, family = "binomial", data = results_train)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.4811	-0.4328	-0.4328	-0.3318	2.8288

Coefficients:

	Estimate	Std. Error	z value
(Intercept)	-3.9826	0.5765	-6.908
EDUCATIONHigher education	1.1105	0.5767	1.925
EDUCATIONIncomplete higher	1.6043	0.5776	2.778
EDUCATIONLower secondary	1.8844	0.5788	3.255
EDUCATIONSecondary / secondary special	1.6616	0.5766	2.882

Pr(>|z|)

(Intercept)	4.91e-12	***
EDUCATIONHigher education	0.05417	.
EDUCATIONIncomplete higher	0.00548	**
EDUCATIONLower secondary	0.00113	**
EDUCATIONSecondary / secondary special	0.00395	**

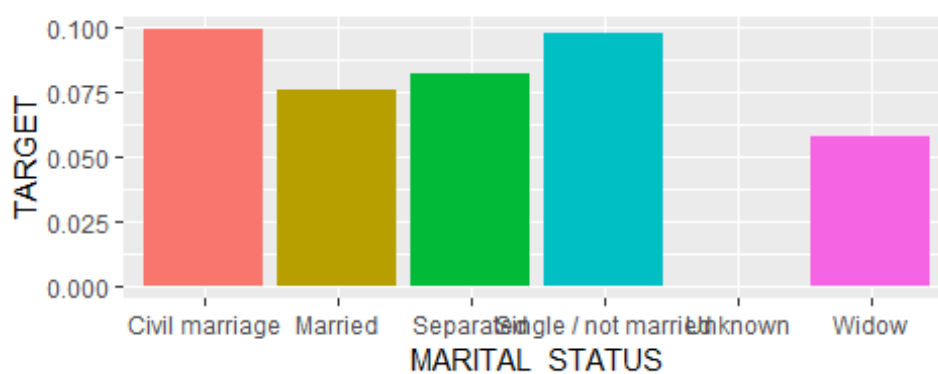
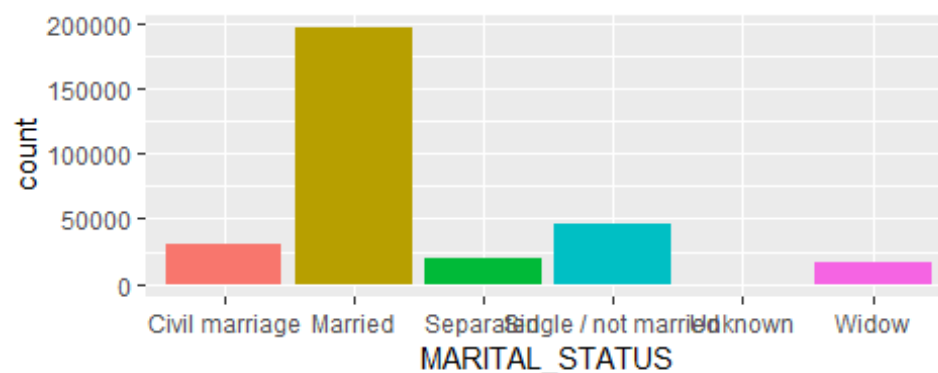
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 172542 on 307510 degrees of freedom
 Residual deviance: 171437 on 307506 degrees of freedom
 AIC: 171447

Number of Fisher Scoring iterations: 5

Marital Status and Marital Status v. Target



A tibble: 6 x 2

MARITAL_STATUS	n
1 Civil marriage	29775
2 Married	196432
3 Separated	19770
4 Single / not married	45444
5 Unknown	2
6 Widow	16088

Call:

```
glm(formula = TARGET ~ MARITAL_STATUS, family = "binomial", data = results_train)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.4577	-0.4135	-0.3965	-0.3965	2.3846

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.20340	0.01937	-113.780	< 2e-16
MARITAL_STATUSMarried	-0.30031	0.02116	-14.190	< 2e-16
MARITAL_STATUSSeparated	-0.21285	0.03236	-6.577	4.81e-11


```
MARITAL_STATUSSingle / not married -0.01538    0.02498    -0.616    0.538
MARITAL_STATUSUnknown                -6.36237    31.08014    -0.205    0.838
MARITAL_STATUSWidow                  -0.57974    0.03884   -14.928    < 2e-16
```

```
(Intercept)                        ***
```

```
MARITAL_STATUSMarried              ***
```

```
MARITAL_STATUSSeparated            ***
```

```
MARITAL_STATUSSingle / not married
```

```
MARITAL_STATUSUnknown
```

```
MARITAL_STATUSWidow                ***
```

```
---
```

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

```
(Dispersion parameter for binomial family taken to be 1)
```

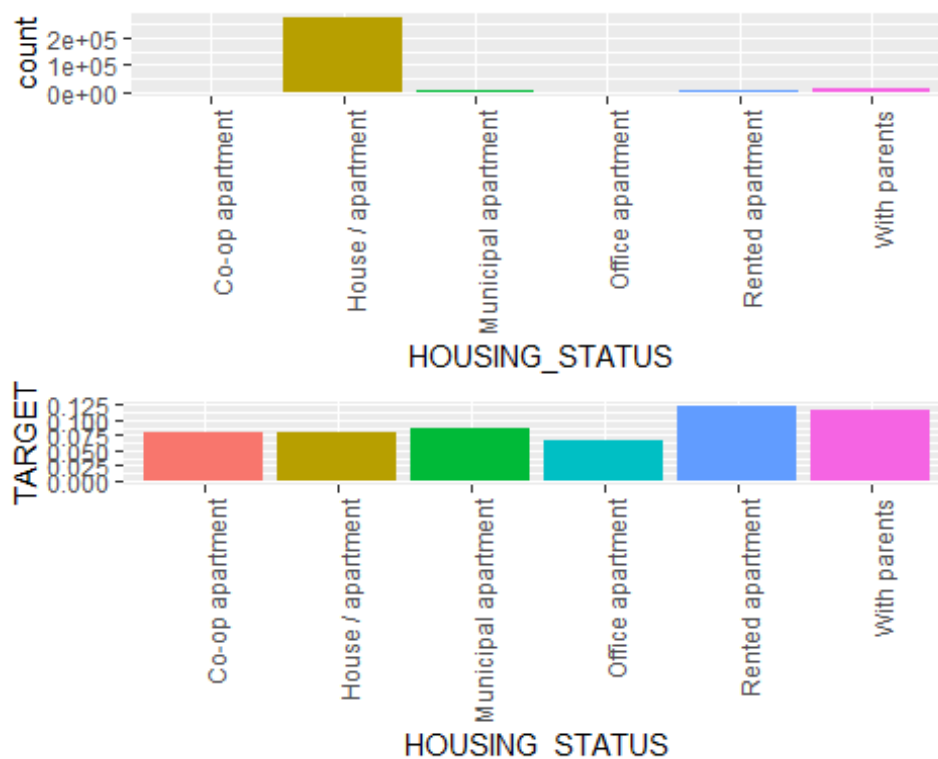
```
Null deviance: 172542  on 307510  degrees of freedom
```

```
Residual deviance: 172045  on 307505  degrees of freedom
```

```
AIC: 172057
```

```
Number of Fisher Scoring iterations: 7
```

Housing Status and Housing Status v. Target



A tibble: 6 x 2

HOUSING_STATUS	n
1 Co-op apartment	1122
2 House / apartment	272868
3 Municipal apartment	11183
4 Office apartment	2617
5 Rented apartment	4881
6 With parents	14840

Call:

```
glm(formula = TARGET ~ HOUSING_STATUS, family = "binomial", data = results_train)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.5126	-0.4029	-0.4029	-0.4029	2.3334

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.45159	0.11047	-22.192	< 2e-16 ***
HOUSING_STATUSHouse / apartment	-0.01885	0.11070	-0.170	0.864815
HOUSING_STATISMunicipal apartment	0.08041	0.11554	0.696	0.486434

```
HOUSING_STATUSOffice apartment    -0.20272    0.13575   -1.493  0.135338
HOUSING_STATUSRented apartment    0.48847    0.11875    4.113  3.9e-05 ***
HOUSING_STATUSWith parents        0.43025    0.11339    3.795  0.000148 ***
```

```
---
```

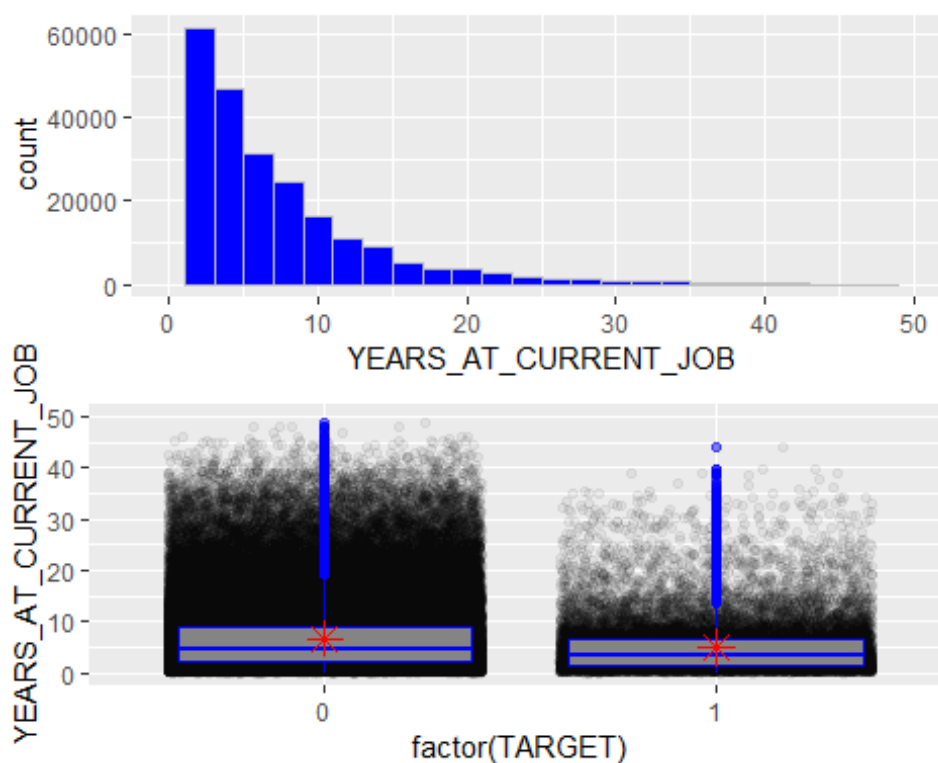
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 172542  on 307510  degrees of freedom
Residual deviance: 172165  on 307505  degrees of freedom
AIC: 172177
```

```
Number of Fisher Scoring iterations: 5
```

Years at Current Job and Years v. Target



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	2.554	6.075	185.420	15.625	999.981

Anderson-Darling normality test

```
data: results_train$YEARS_AT_CURRENT_JOB
A = 82723, p-value < 2.2e-16
```

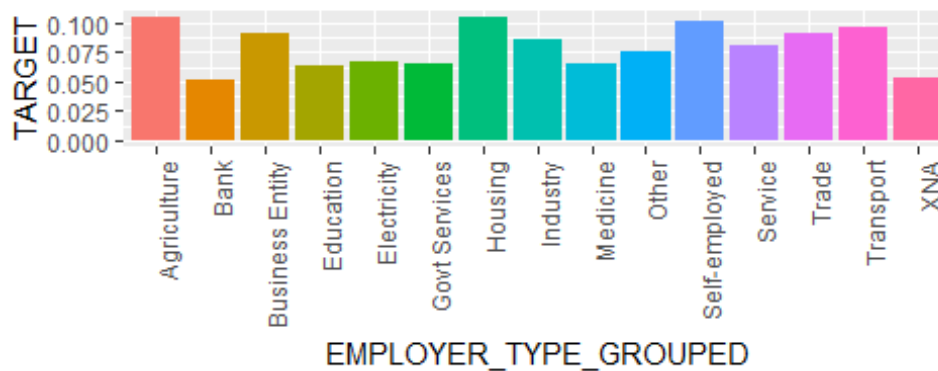
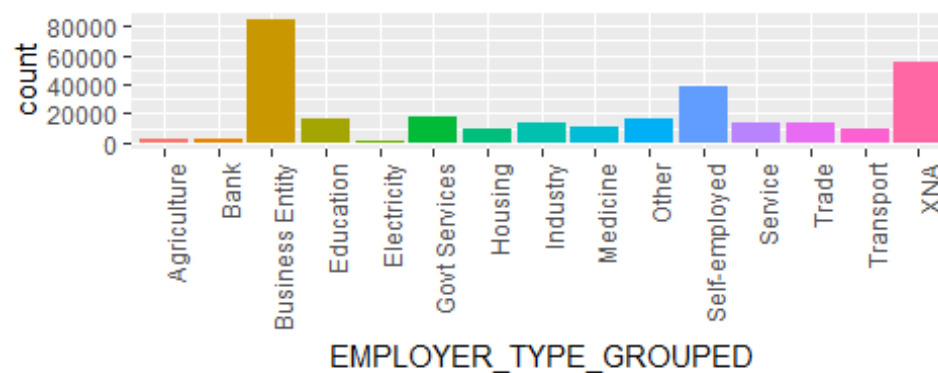
Wilcoxon rank sum test with continuity correction

data: YEARS_AT_CURRENT_JOB by TARGET

W = 4150800000, p-value < 2.2e-16

alternative hypothesis: true location shift is not equal to 0

Employer Organization Type and Type v. Target



A tibble: 15 x 2

	EMPLOYER_TYPE_GROUPED	n
	<chr>	<int>
1	Agriculture	2454
2	Bank	2507
3	Business Entity	84529
4	Education	17100
5	Electricity	950
6	Govt Services	17536
7	Housing	9679
8	Industry	14311
9	Medicine	11193
10	Other	16683
11	Self-employed	38412
12	Service	13478
13	Trade	14315

```
14 Transport      8990
15 XNA            55374
```

Call:

```
glm(formula = TARGET ~ EMPLOYER_TYPE_GROUPED, family = "binomial",
     data = results_train)
```

Deviance Residuals:

```
      Min       1Q   Median       3Q      Max
-0.4719 -0.4374 -0.4241 -0.3332  2.4328
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.145772	0.065926	-32.548	< 2e-16
EMPLOYER_TYPE_GROUPEDBank	-0.760288	0.111618	-6.812	9.66e-12
EMPLOYER_TYPE_GROUPEDBusiness Entity	-0.153162	0.067000	-2.286	0.022253
EMPLOYER_TYPE_GROUPEDEducation	-0.556057	0.073067	-7.610	2.74e-14
EMPLOYER_TYPE_GROUPEDElectricity	-0.498938	0.146104	-3.415	0.000638
EMPLOYER_TYPE_GROUPEDGovt Services	-0.501625	0.072590	-6.910	4.83e-12
EMPLOYER_TYPE_GROUPEDHousing	0.006975	0.073771	0.095	0.924668
EMPLOYER_TYPE_GROUPEDIndustry	-0.217486	0.072353	-3.006	0.002648
EMPLOYER_TYPE_GROUPEDMedicine	-0.506571	0.076149	-6.652	2.88e-11
EMPLOYER_TYPE_GROUPEDOther	-0.346169	0.072079	-4.803	1.57e-06
EMPLOYER_TYPE_GROUPEDSelf-employed	-0.032278	0.068052	-0.474	0.635281
EMPLOYER_TYPE_GROUPEDService	-0.273849	0.073043	-3.749	0.000177
EMPLOYER_TYPE_GROUPEDTrade	-0.158813	0.072062	-2.204	0.027535
EMPLOYER_TYPE_GROUPEDTransport	-0.089093	0.074967	-1.188	0.234664
EMPLOYER_TYPE_GROUPEDXNA	-0.717556	0.068555	-10.467	< 2e-16

```
(Intercept)      ***
EMPLOYER_TYPE_GROUPEDBank      ***
EMPLOYER_TYPE_GROUPEDBusiness Entity *
EMPLOYER_TYPE_GROUPEDEducation  ***
EMPLOYER_TYPE_GROUPEDElectricity ***
EMPLOYER_TYPE_GROUPEDGovt Services ***
EMPLOYER_TYPE_GROUPEDHousing
EMPLOYER_TYPE_GROUPEDIndustry   **
EMPLOYER_TYPE_GROUPEDMedicine   ***
EMPLOYER_TYPE_GROUPEDOther      ***
EMPLOYER_TYPE_GROUPEDSelf-employed
EMPLOYER_TYPE_GROUPEDService    ***
EMPLOYER_TYPE_GROUPEDTrade      *
EMPLOYER_TYPE_GROUPEDTransport
EMPLOYER_TYPE_GROUPEDXNA        ***
```

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 172542 on 307510 degrees of freedom
 Residual deviance: 171260 on 307496 degrees of freedom
 AIC: 171290

Number of Fisher Scoring iterations: 5

Years Since Getting Current Identity Document and Years v. Target



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	5.503	12.331	13.651	20.478	67.548

Anderson-Darling normality test

data: results_train\$YEARS_SINCE_GETTING_IDENTITY_DOCUMENT
 A = 3532.2, p-value < 2.2e-16

Wilcoxon rank sum test with continuity correction

data: YEARS_SINCE_GETTING_IDENTITY_DOCUMENT by TARGET
 W = 3807600000, p-value < 2.2e-16
 alternative hypothesis: true location shift is not equal to 0

Rating of Region and Rating v. Target

The meaning of this variable is not reported in the materials made available by HOME CREDIT. It may relate to population density, per capita wealth of the region, but this information is not supplied.



```
# A tibble: 3 x 2
  REGION_AND_CITY_RATING      n
      <int>    <int>
1           1    34167
2           2   229484
3           3    43860
```

Call:

```
glm(formula = TARGET ~ REGION_AND_CITY_RATING, family = "binomial",
     data = results_train)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.4979	-0.4035	-0.4035	-0.4035	2.4340

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.35108	0.02859	-117.20	<2e-16 ***
REGION_AND_CITY_RATING	0.44198	0.01309	33.77	<2e-16 ***

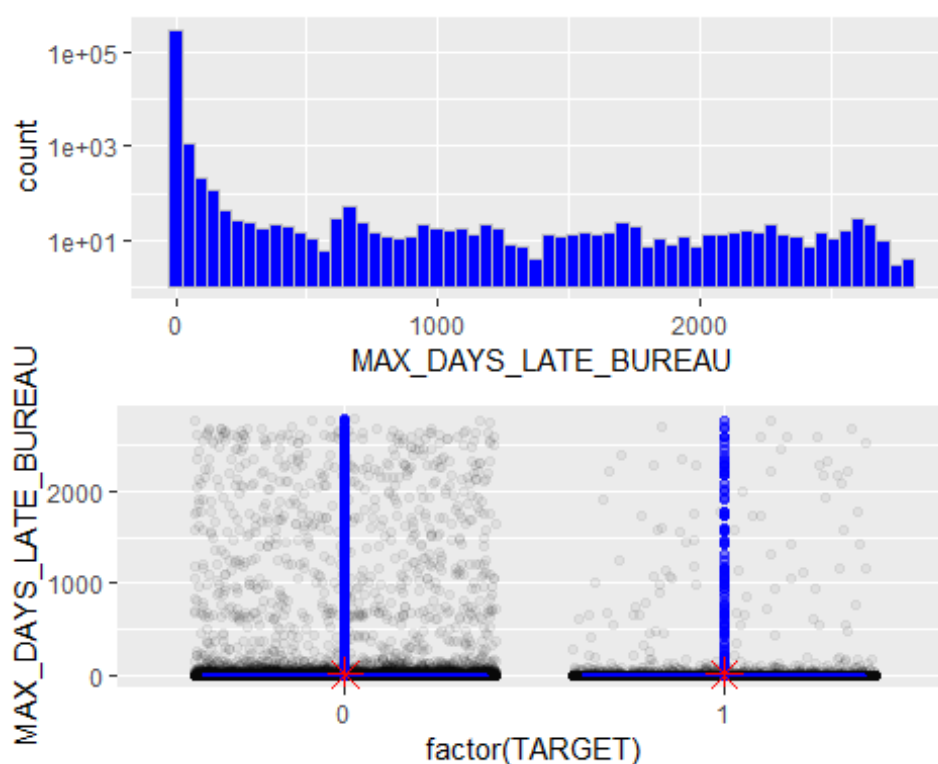
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 172542 on 307510 degrees of freedom
Residual deviance: 171406 on 307509 degrees of freedom
AIC: 171410

Number of Fisher Scoring iterations: 5

Maximum dates late payment as reported to HOME CREDIT by and outside credit bureau



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.000	0.000	0.000	4.086	0.000	2792.000	1

Anderson-Darling normality test

data: results_train\$MAX_DAYS_LATE_BUREAU
A = 116840, p-value < 2.2e-16

Wilcoxon rank sum test with continuity correction

data: MAX_DAYS_LATE_BUREAU by TARGET


```
W = 3468100000, p-value < 2.2e-16  
alternative hypothesis: true location shift is not equal to 0
```

Findings and Recommendations

The Area Under ROC Curve was used as the primary statistic for evaluating model performance. The most successful model was the Random Forest model, trained on 80% of the data, with an AUC score of 0.7297.

The following table summarizes the overall performance of the different models.

Pct of train data used and model type	Training time in seconds	Sensitivity: TP/(TP+FN)	Accuracy: (TP+TN)/N	AUC Score	Specificity: (TN/N)	Precision: TP/(TP+FP)	Comment
100% GLM	51.71	0.6402820	0.6677669	0.7110466	0.9549854	0.1456520	Downsampled training set
100% Naive Bayes	2304.37	0.5268882	0.2640890	0.7110466	0.8529577	0.0574604	Downsampled training set
100% KNN	15074.17	0.3822759	0.3465741	0.6844597	0.8635919	0.0486442	Downsampled training set
5% Random Forest	665.32	0.6580060	0.6561087	0.7108947	0.9562179	0.1438003	Downsampled training set
10% Random Forest	1554.28	0.6374622	0.6721245	0.7139183	0.9549685	0.1470042	Downsampled training set
20% Random Forest	3440.52	0.6501511	0.6765146	0.7220000	0.9567006	0.1509399	Downsampled training set
30% Random Forest	5852.06	0.6539778	0.6761569	0.7268011	0.9571100	0.1514035	Downsampled training set
60% Random Forest	17222.51	0.6547835	0.6782381	0.7285197	0.9573377	0.1524430	Downsampled training set
80% Random Forest	65159.25	0.6521652	0.6829534	0.7297375	0.9573496	0.1541171	Downsampled training set

Specific Statistics for Each Model with AUC Graphs

More detailed statistical analysis for the performance of each model is given below, accompanied by graphs demonstrating the AUC performance of each mode. Again, overall, the Random Forest model trained on 80% of the data performed best overall. It's overall Accuracy, measured as True Positives + True Negatives / Total Negatives was the highest,

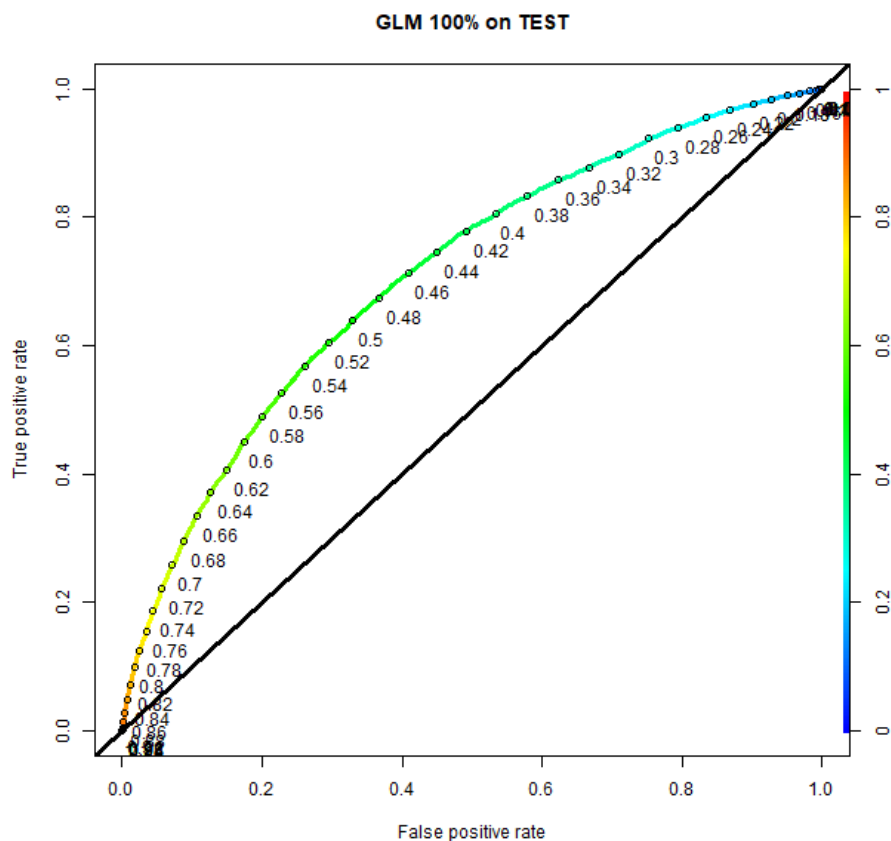
at 0.6829. Its Specificity (TN/N) was also highest overall at 0.9573. Its Precision (TP/TP+FP) 0.1541 was also the highest of the model tested here.

The specifics for the performance of each model follow:

Using a Generalized Linear Model

Trained on 100% of the train data

Area under ROC Curve = 0.7110466



Confusion Matrix

##	Predicted 1	Predicted 0
## Actual 1	3179	1786
## Actual 0	18647	37890

Sensitivity: $TP/(TP+FN) = 0.640282$

Specificity: $TN/N = 0.6701806$

Precision: $TP/(TP+FP) = 0.145652$

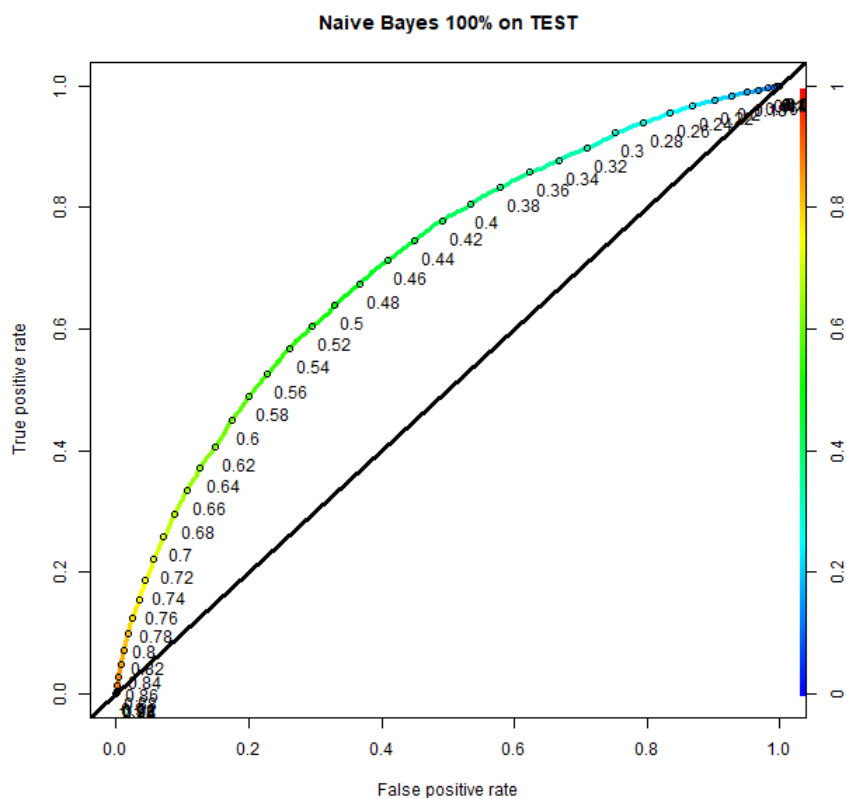
Accuracy = 3179.6160775

AUC Score = 0.7110466

Using a Naive Bayes Model

Trained on 100% of the train data

Area under ROC Curve = 0.7110466



Confusion Matrix

##	Predicted 1	Predicted 0
## Actual 1	2616	2349
## Actual 0	42911	13626

Sensitivity: $TP/(TP+FN) = 0.5268882$

Specificity: $TN/N = 0.2410103$

Precision: $TP/(TP+FP) = 0.0574604$

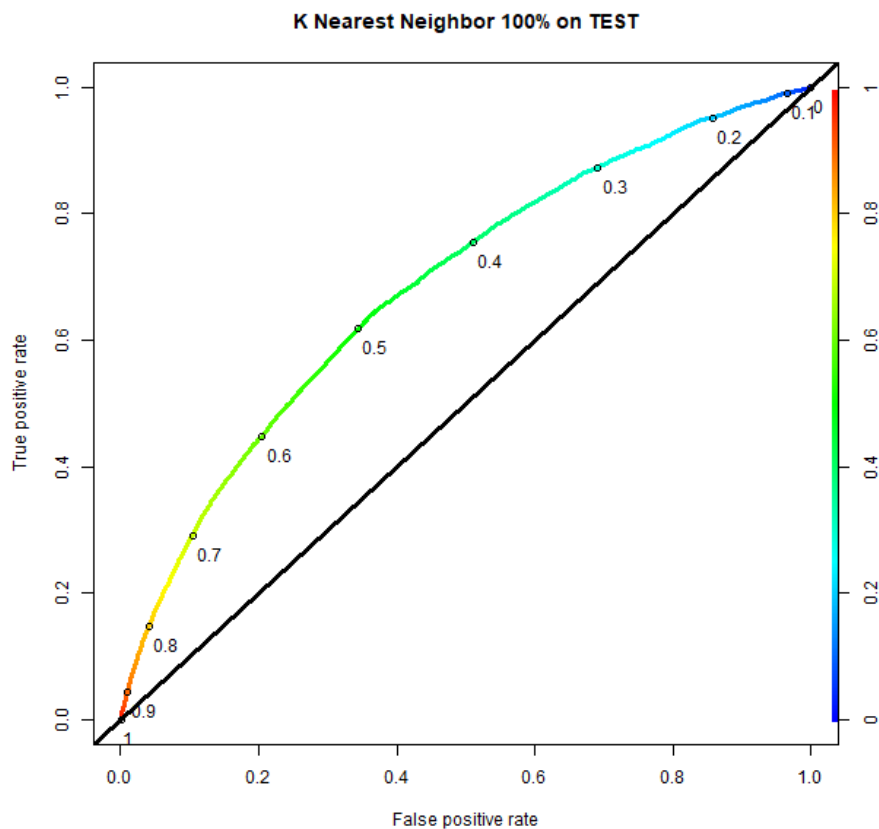
Accuracy = 2616.2215538

AUC Score = 0.7110466

Using a K Nearest Neighbor Model

Trained on 100% of the train data

Area under ROC Curve = 0.6844597



Confusion Matrix

##		Predicted 1	Predicted 0
## Actual 1		1898	3067
## Actual 0		37120	19417

Sensitivity: $TP/(TP+FN) = 0.3822759$

Specificity: $TN/N = 0.3434388$

Precision: $TP/(TP+FP) = 0.0486442$

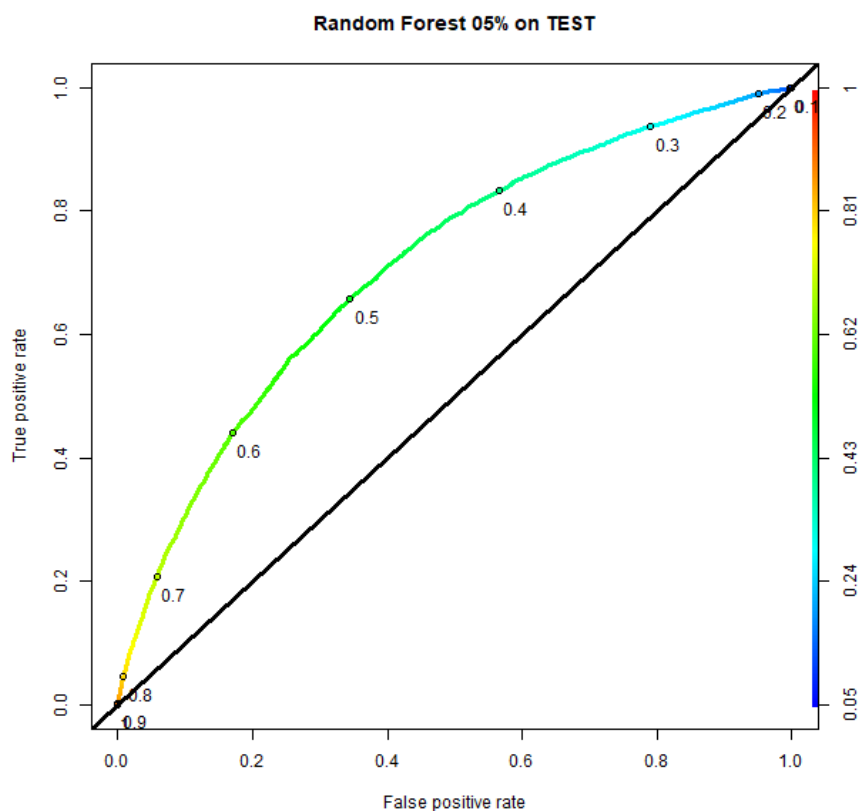
Accuracy = 1898.3157133

AUC Score = 0.6844597

Using a Random Forest Model

Trained on 5% of the train data

Area under ROC Curve = 0.7108947



Confusion Matrix

##	Predicted 1	Predicted 0
## Actual 1	3267	1698
## Actual 0	19452	37085

Sensitivity: $TP/(TP+FN) = 0.658006$

Specificity: $TN/N = 0.6559421$

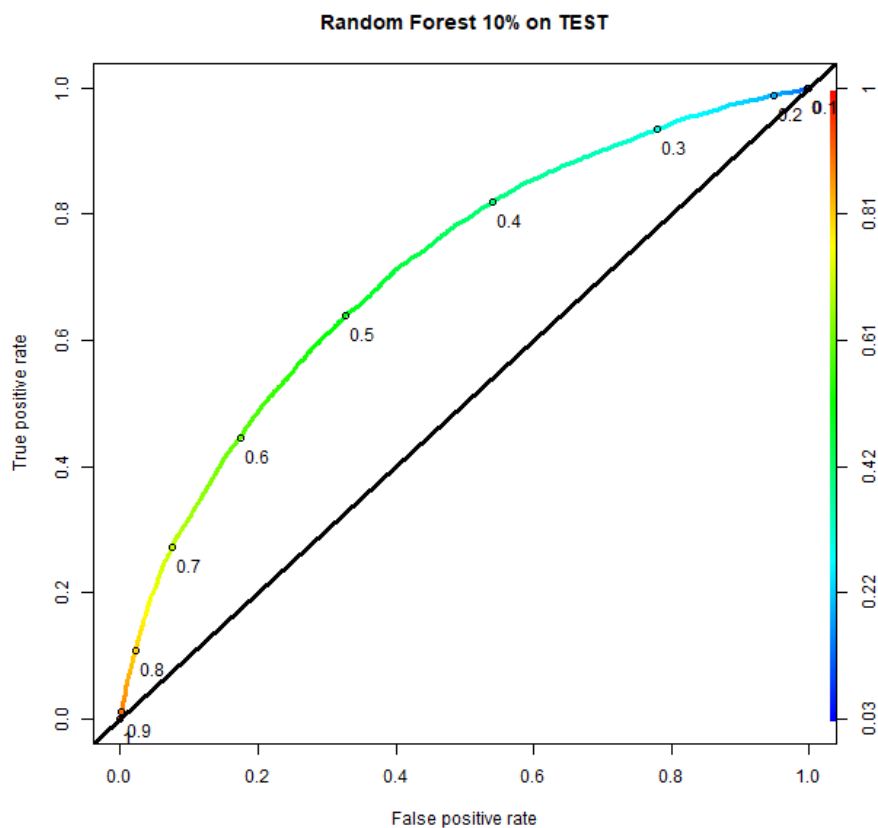
Precision: $TP/(TP+FP) = 0.1438003$

Accuracy = 3267.6029885

AUC Score = 0.7108947

Trained on 10% of the train data

Area under ROC Curve = 0.7139183



Confusion Matrix

##	Predicted 1	Predicted 0
## Actual 1	3165	1800
## Actual 0	18365	38172

Sensitivity: $TP/(TP+FN) = 0.6374622$

Specificity: $TN/N = 0.6751685$

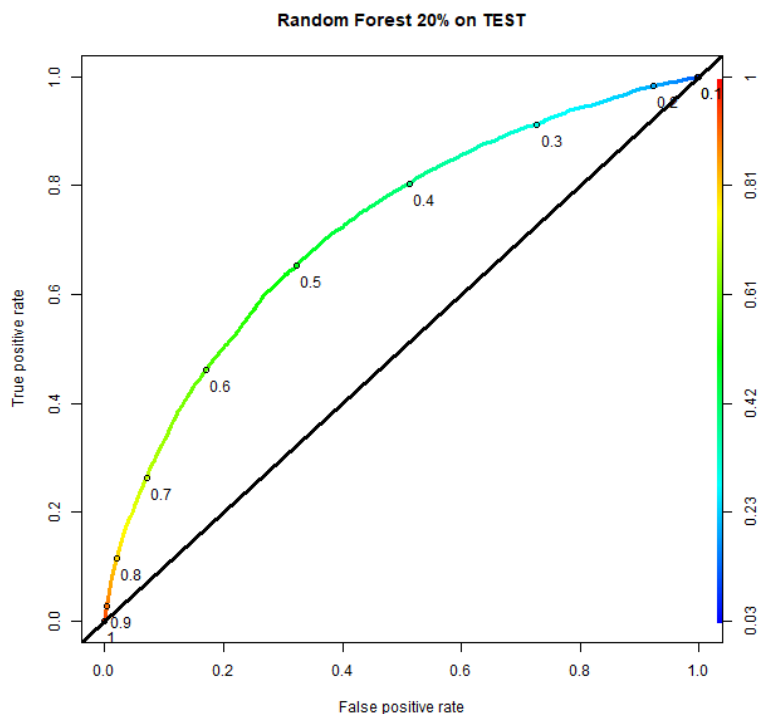
Precision: $TP/(TP+FP) = 0.1470042$

Accuracy = 3165.6206627

AUC Score = 0.7139183

Trained on 20% of the train data

Area under ROC Curve = 0.722



Confusion Matrix

##	Predicted 1	Predicted 0
## Actual 1	3228	1737
## Actual 0	18158	38379

Sensitivity: $TP/(TP+FN) = 0.6501511$

Specificity: $TN/N = 0.6788298$

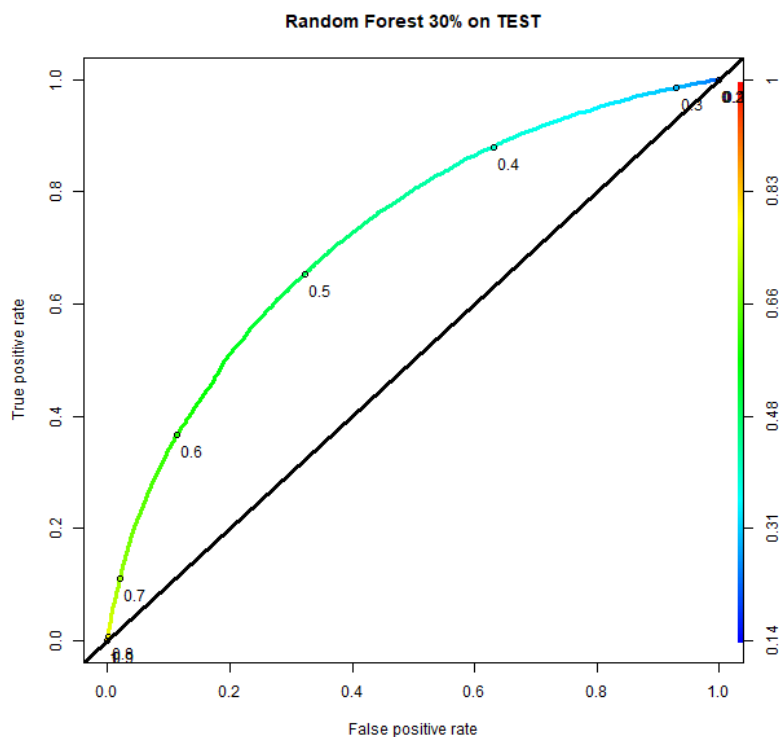
Precision: $TP/(TP+FP) = 0.1509399$

Accuracy = 3228.6240285

AUC Score = 0.722

Trained on 30% of the train data

Area under ROC Curve = 0.7268011



Confusion Matrix

##		Predicted 1	Predicted 0
## Actual 1		3247	1718
## Actual 0		18199	38338

Sensitivity: $TP/(TP+FN) = 0.6539778$

Specificity: $TN/N = 0.6781046$

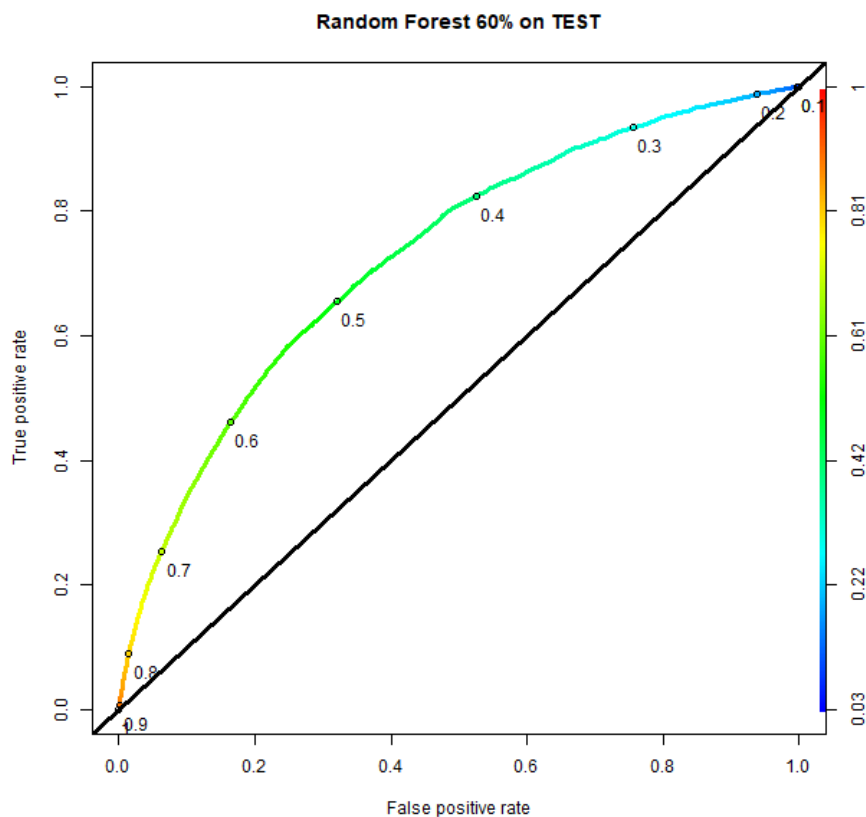
Precision: $TP/(TP+FP) = 0.1514035$

Accuracy = 3247.6233618

AUC Score = 0.7268011

Trained on 60% of the train data

Area under ROC Curve = 0.7285197



Confusion Matrix

##	Predicted 1	Predicted 0
## Actual 1	3251	1714
## Actual 0	18075	38462

Sensitivity: $TP/(TP+FN) = 0.6547835$

Specificity: $TN/N = 0.6802979$

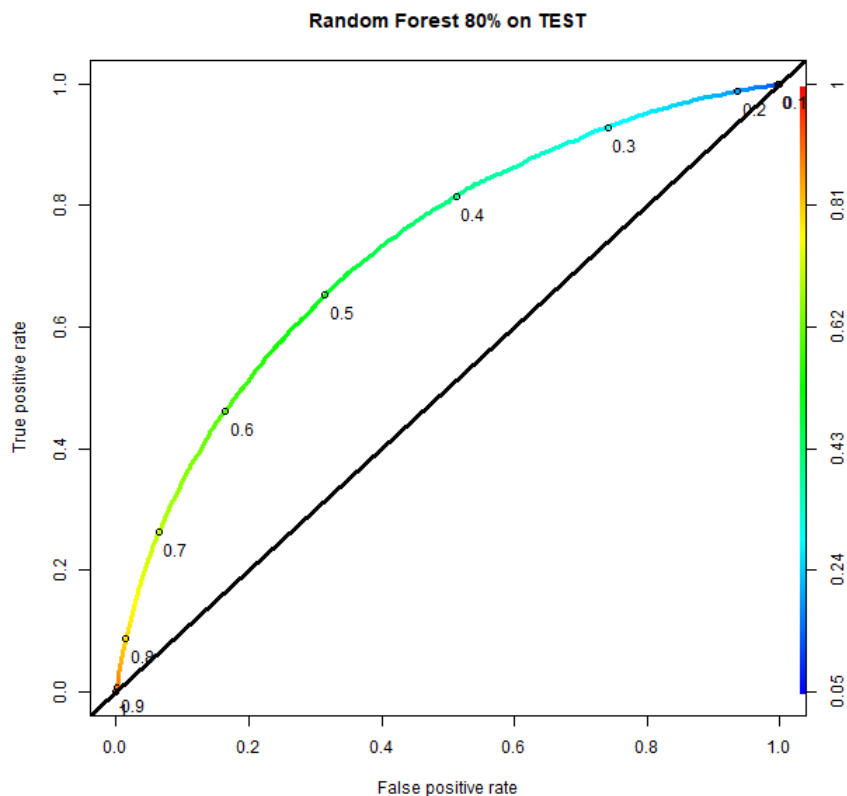
Precision: $TP/(TP+FP) = 0.152443$

Accuracy = 3251.625378

AUC Score = 0.7285197

Trained on 80% of the train data

Area under ROC Curve = 0.7297375



Confusion Matrix

##	Predicted 1	Predicted 0
## Actual 1	3238	1727
## Actual 0	17772	38765

Sensitivity: $TP/(TP+FN) = 0.6521652$

Specificity: $TN/N = 0.6856572$

Precision: $TP/(TP+FP) = 0.1541171$

Accuracy = 3238.6303047

AUC Score = 0.7297375

Conclusion and Suggestions

As one might anticipate, accurately predicting the likelihood of a client defaulting or struggling to repay debt is challenging. Changing life circumstances, and changing environmental or political conditions can significantly alter overall outcomes.

Area Under a ROC Curve is a useful measurement of a model's performance vs. random selection, however. Given that random selection would typically yield an AUC score of .5, the Random Forest model trained on 80% of the data is a significant improvement upon this. With an AUC score of 0.7297 this model would be best used in combination with Home Credit's current decision processes to decrease risk and maximize profits.