PREDICTIVE ANALYTICS IN THE PERSONAL LOAN MARKET

BIG DATA & BUSINESS ANALYTICS

OVERVIEW

About the Data
Hypothesis
Cleaning the Data
Partitioning the Data
Regression Task
Classification Task
Conclusion

PREDICTIVE ANALYTICS IN THE PERSONAL LOAN MARKET

THE DATA

32,581
OBSERVATIONS

12
VARIABLES

Loan Holder Demographics

Person_age
Person_income
Person_home_ownership
Person_emp_length
Loan_intent
Cb_person_default_on_file
Cb_person_cred_hist_length

Loan Attributes

Loan_percent_income
Loan_grade
Loan_amnt
Loan_int_rate
Loan_status

Research Questions

REGRESSION

Using the available independent variables, to what accuracy can we build a linear regression model to predict person income?

CLASSIFICATION

Which supervised ML technique can most accurately predict default risk?

DELETION WHEN NECESSARY APPROACH

ERRONEOUS OUTLIERS

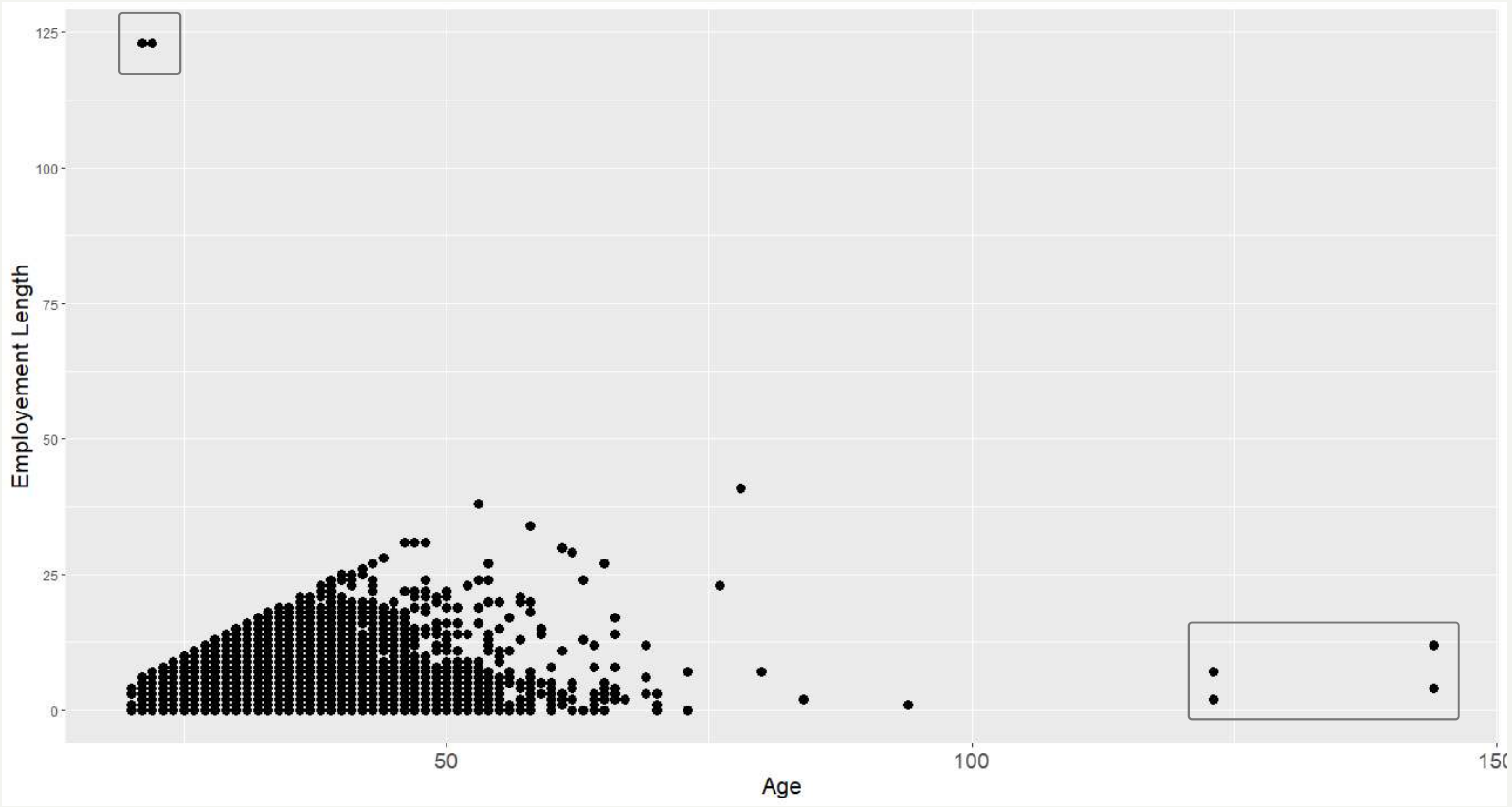
Age and years employmed were restricted by number of years

OMITTING NA

Resulted in 12.12% data loss

CLEANING THE DATA

6 values
> 120 years



CLEANING THE DATA

Heatmap Matrix

High Correlation (>.6)

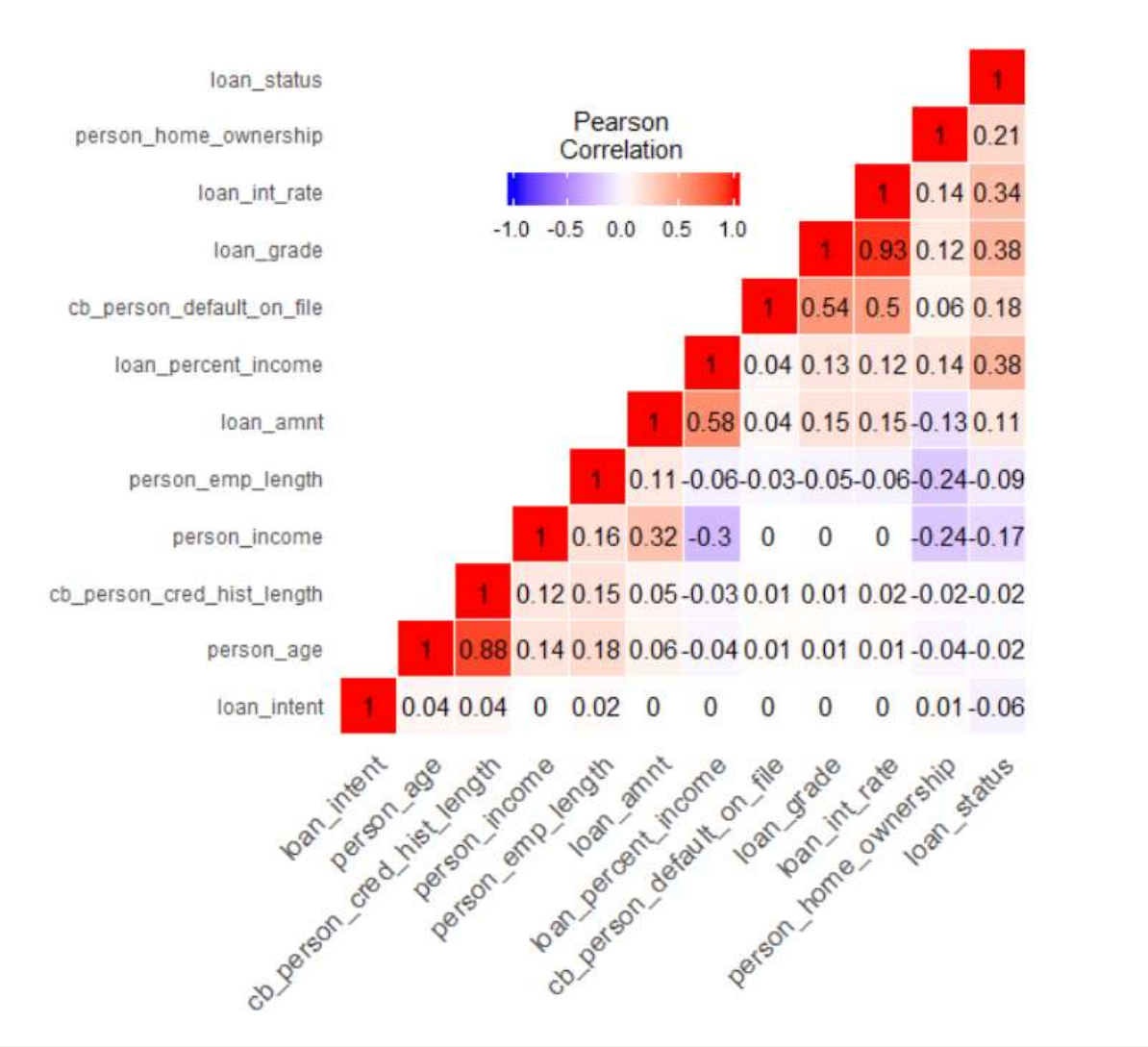
- Age & Credit History
- Loan Grade & Loan Interest Rate

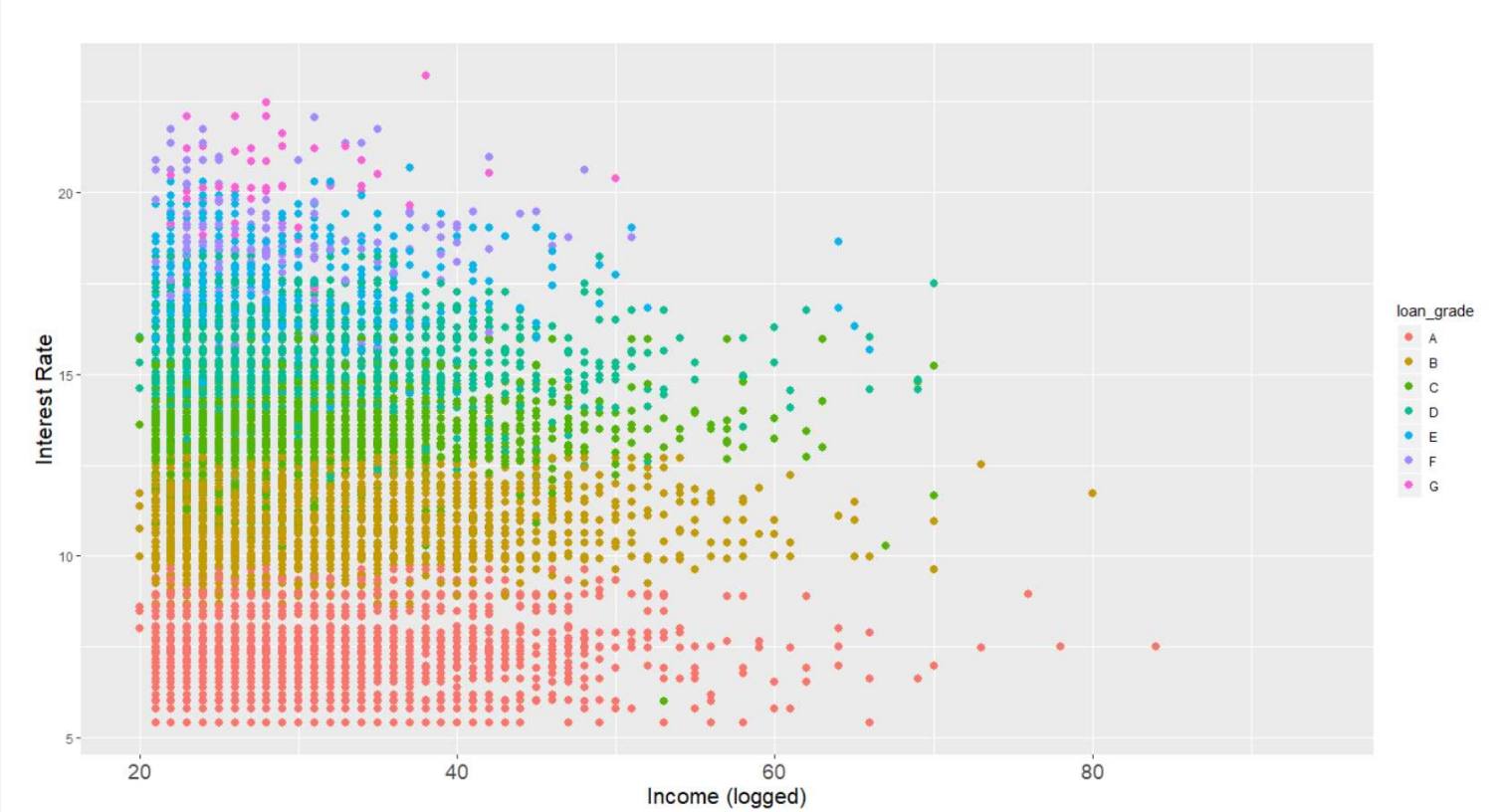
Moderate Correlation (.4-.6)

- Loan Amount & % of Income
- Default on File & Loan Grade
- Default on File & Interest Rate

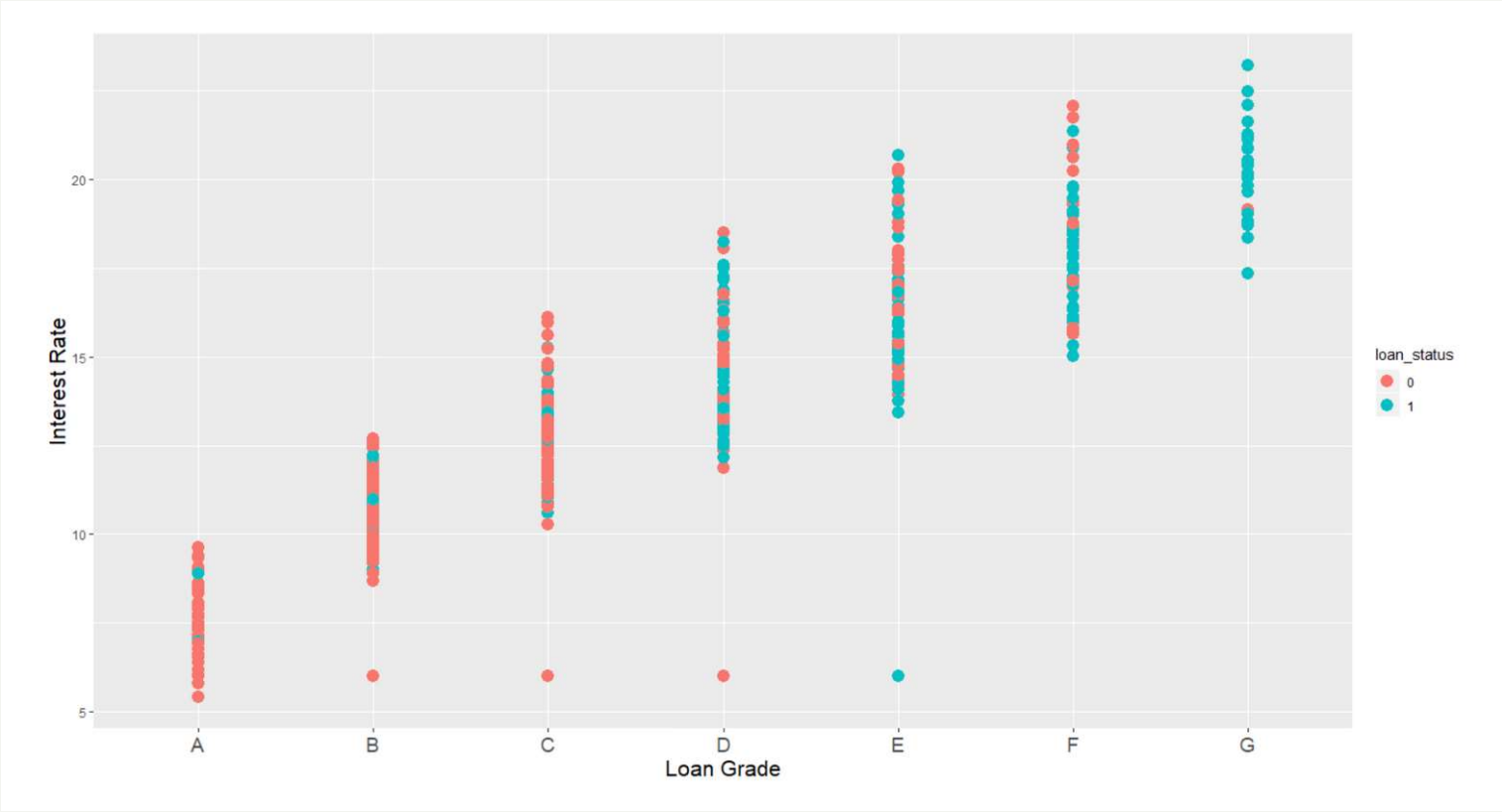
Weak Correlation (.2-.4)

- Loan Status & Loan % of Income
- Loan Status & Loan Grade
- Loan Status and Loan Interest Rate
- Income & loan amount
- Income & Home Ownership





INCOME X INTEREST RATE X LOAN GRADE



INTEREST RATE X LOAN GRADE X LOAN STATUS

70%

Training Set
- 22,808 obs

15%

Validation Set
- 4,887

15%

Testing Set
- 4,886

Partitioning the Data

MODEL 1

Removed

- log_person_income
- cb_person_default_on_file
- cb_person_cred_hist_length
- loan_int_rate
- loan_percent_income

MODEL 2

Removed

- Factor level "Other" from person_home_ownership

MODEL 3

Removed

- Factor level "C" from Loan_grade

MODEL 4

Removed

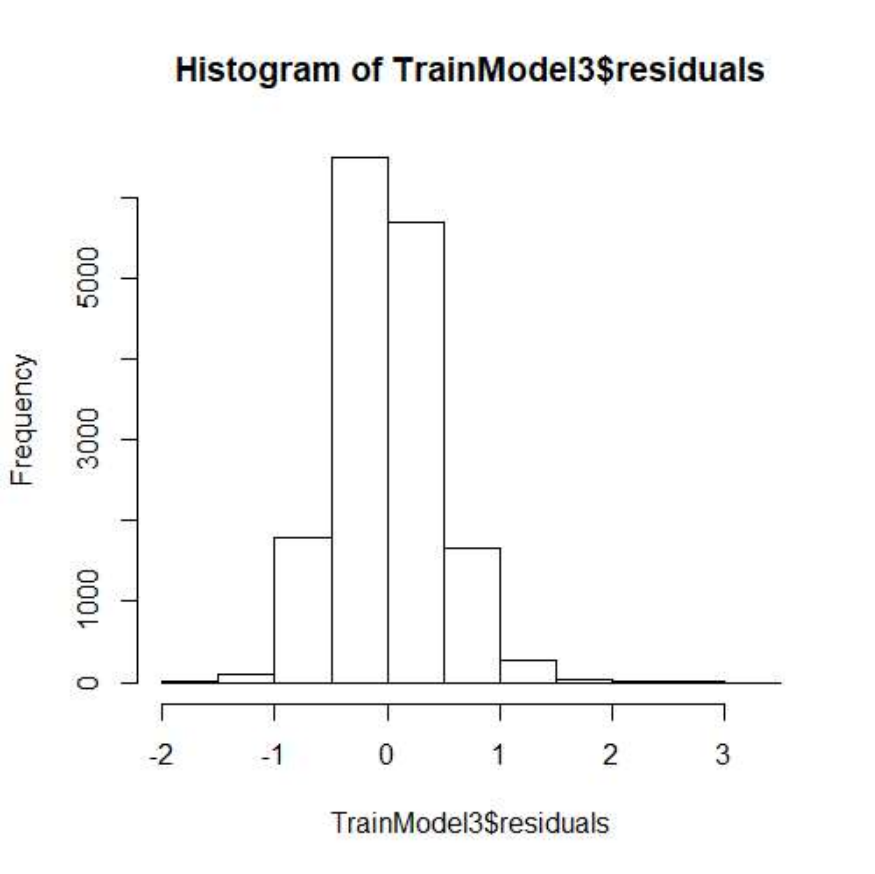
- Factor level "Other" from person_home_ownership & factor level "C" from Loan_grade

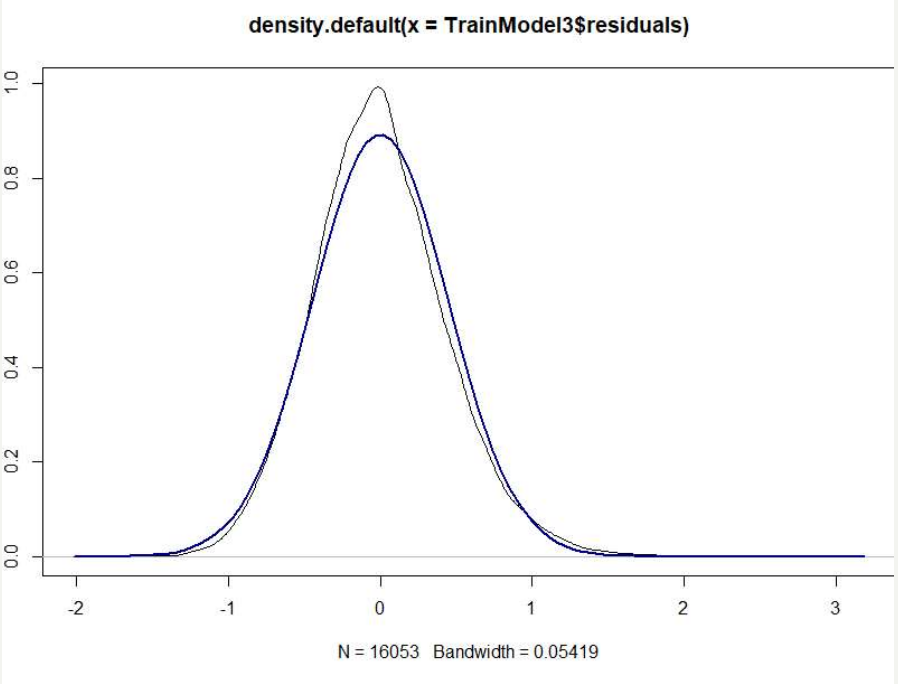
OLS Regression Tasks

Residual Standard Error

Model	In-Sample RSE	Out-of-Sample RSE
Model 1	.447	.4476
Model 2	.447	.448
Model 3	.4471	.4458
Model 4	.4471	.4463

Residual Analysis





Jarque Bera Test

data: TrainModel3$residuals
X-squared = 2502.3, df = 2, p-value < 2.2e-16

TESTING SET

MODEL 3

RSE: .4433

Adj. R-Squared: .3728

Per one unit increase in the loan_status variable there will be a 44.67% decrease in income.

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.057e+01  4.168e-02 253.603  < 2e-16 ***
person_age                 6.878e-03  1.245e-03   5.526 3.52e-08 ***
person_home_ownershipOTHER 1.878e-01  1.118e-01   1.681  0.09292 .
person_home_ownershipOWN  -3.196e-01  3.021e-02 -10.580  < 2e-16 ***
person_home_ownershipRENT -2.049e-01  1.699e-02 -12.061  < 2e-16 ***
person_emp_length          1.220e-02  1.916e-03   6.366 2.20e-10 ***
loan_intentEDUCATION      -1.626e-02  2.557e-02  -0.636  0.52504
loan_intentHOMEIMPROVEMENT 4.518e-02  2.976e-02   1.518  0.12906
loan_intentMEDICAL        -2.364e-02  2.596e-02  -0.911  0.36258
loan_intentPERSONAL       -1.867e-02  2.611e-02  -0.715  0.47448
loan_intentVENTURE        -1.944e-02  2.629e-02  -0.740  0.45960
loan_gradeB               -4.059e-02  1.680e-02  -2.415  0.01577 *
loan_gradeD                7.454e-02  2.594e-02   2.874  0.00408 **
loan_gradeE                1.127e-01  4.024e-02   2.801  0.00512 **
loan_gradeF                2.122e-01  8.933e-02   2.376  0.01756 *
loan_gradeG                7.957e-02  2.575e-01   0.309  0.75732
loan_amnt                  3.543e-05  1.202e-06  29.461  < 2e-16 ***
loan_status1              -4.467e-01  2.134e-02 -20.935  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4433 on 3475 degrees of freedom
  (456 observations deleted due to missingness)
Multiple R-squared:  0.3758,    Adjusted R-squared:  0.3728
F-statistic: 123.1 on 17 and 3475 DF,  p-value: < 2.2e-16
```

LOSGISTIC REGRESSION MODEL 1

Most effective with all variables

LOGISTIC REGRESSION MODEL 2

Removed income, loan amount, credit history, default history, and age

CART

Eight internal nodes and five layers

SUPPORT VECTOR MACHINE

Low p-value allows us to reject null hypothesis that accuracy is greater than the no information rate

Classification Tasks

LOGISTIC REGRESSION MODELS

Logistic Regression 1

all variables

Increases Likelihood Multiplier

- loan_grade - "G"
- loan_percent_income

Decreases Likelihood Multiplier

- loan_intent "Venture"
- person_home_ownership "Own"

AIC = 13,510

		2.5 %	97.5 %
(Intercept)	1.237450e+03	1.540866e+02	9.995935e+03
person_age	9.878295e-01	9.733147e-01	1.002477e+00
person_home_ownershipOTHER	1.400623e+00	6.512250e-01	2.858723e+00
person_home_ownershipOWN	1.623927e-01	1.241138e-01	2.102208e-01
person_home_ownershipRENT	2.147484e+00	1.942817e+00	2.375161e+00
person_emp_length	9.923475e-01	9.805062e-01	1.004268e+00
loan_intentEDUCATION	4.127189e-01	3.581520e-01	4.753305e-01
loan_intentHOMEIMPROVEMENT	1.052785e+00	8.978146e-01	1.233907e+00
loan_intentMEDICAL	8.165763e-01	7.133002e-01	9.347691e-01
loan_intentPERSONAL	5.222743e-01	4.513816e-01	6.039533e-01
loan_intentVENTURE	3.128907e-01	2.675887e-01	3.654189e-01
loan_gradeB	1.102801e+00	9.094777e-01	1.337575e+00
loan_gradeC	1.314644e+00	9.837714e-01	1.757049e+00
loan_gradeD	1.046948e+01	7.271875e+00	1.509484e+01
loan_gradeE	1.105719e+01	6.999744e+00	1.750370e+01
loan_gradeF	1.346250e+01	7.198694e+00	2.529985e+01
loan_gradeG	1.933259e+07	1.115589e+05	7.789203e+23
loan_amnt	9.999884e-01	9.999706e-01	1.000006e+00
loan_int_rate	1.084738e+00	1.040048e+00	1.131414e+00
loan_percent_income	4.405388e+03	1.741268e+03	1.124236e+04
cb_person_default_on_fileY	1.043188e+00	9.212406e-01	1.181297e+00
cb_person_cred_hist_length	1.017055e+00	9.945631e-01	1.040030e+00
log_person_income	3.583285e-01	2.966703e-01	4.324434e-01

LOGISTIC REGRESSION MODELS

Logistic Regression 2

all variables

Increases Likelihood Multiplier

- loan_grade - "G"
- loan_percent_income

Decreases Likelihood Multiplier

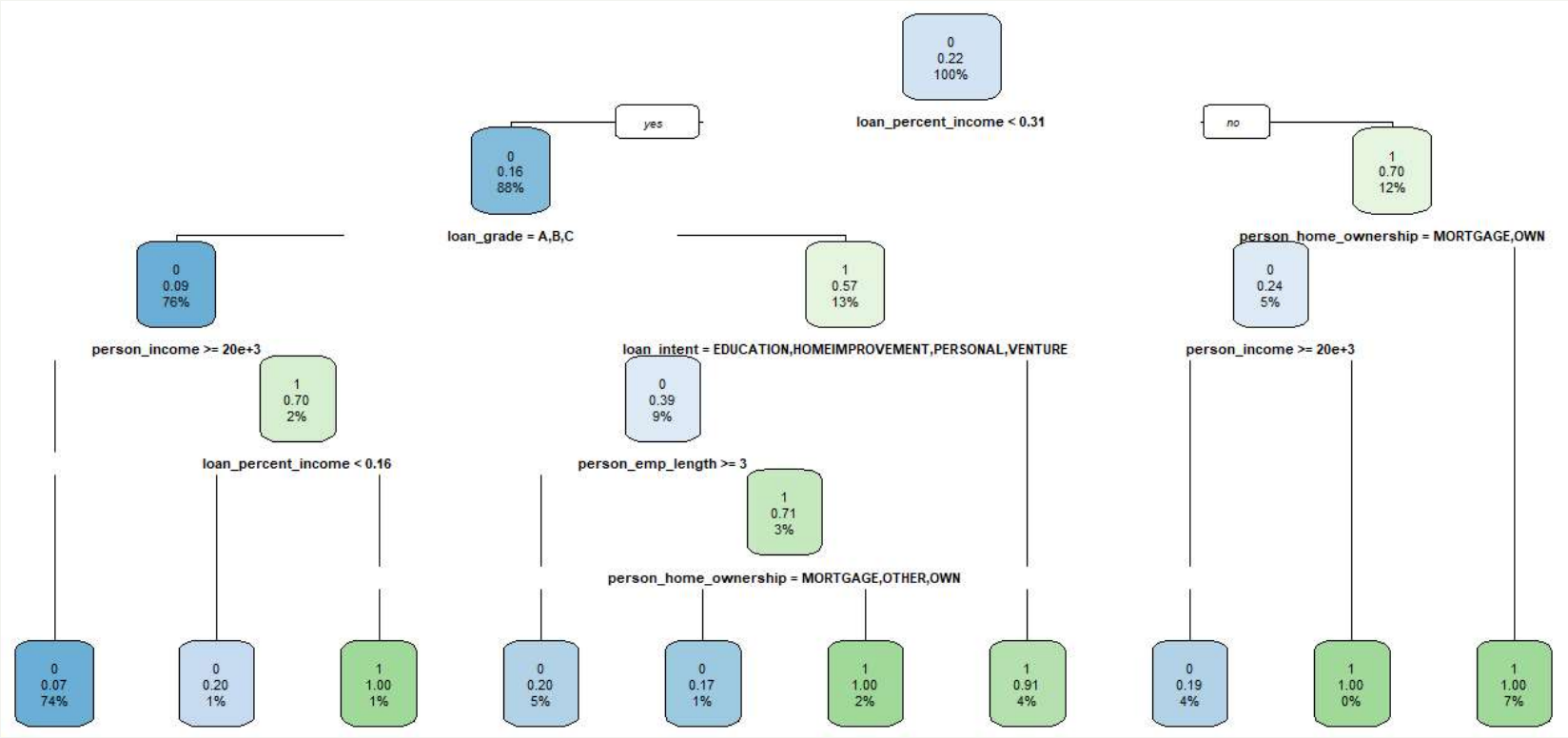
- loan_intent "Venture"
- person_home_ownership "Own"

AIC = 13,423

		2.5 %	97.5 %
(Intercept)	3.375122e+05	7.780920e+04	1.465944e+06
person_age	9.953121e-01	9.881600e-01	1.002447e+00
person_income	1.000007e+00	1.000006e+00	1.000009e+00
person_home_ownershipOTHER	1.455204e+00	6.756266e-01	2.972222e+00
person_home_ownershipOWN	1.555992e-01	1.188087e-01	2.016251e-01
person_home_ownershipRENT	2.166515e+00	1.962116e+00	2.393682e+00
loan_intentEDUCATION	4.128771e-01	3.581022e-01	4.757590e-01
loan_intentHOMEIMPROVEMENT	1.053874e+00	8.983290e-01	1.235753e+00
loan_intentMEDICAL	8.194469e-01	7.154877e-01	9.384745e-01
loan_intentPERSONAL	5.236995e-01	4.523882e-01	6.059028e-01
loan_intentVENTURE	3.112790e-01	2.660659e-01	3.637323e-01
loan_gradeB	1.117901e+00	9.209745e-01	1.357330e+00
loan_gradeC	1.362163e+00	1.025141e+00	1.810457e+00
loan_gradeD	1.092304e+01	7.622775e+00	1.567587e+01
loan_gradeE	1.136748e+01	7.215760e+00	1.794612e+01
loan_gradeF	1.338353e+01	7.197553e+00	2.500832e+01
loan_gradeG	1.694447e+07	8.608624e+04	1.737335e+24
loan_int_rate	1.083083e+00	1.038341e+00	1.129810e+00
loan_percent_income	2.766584e+03	1.807494e+03	4.252061e+03
log_person_income	2.017398e-01	1.760807e-01	2.311082e-01

CART

DENDOGRAM



SVM

Confusion Matrices

```
          Reference
Prediction     0     1
         0 15412  1498
         1   253  2860

               Accuracy : 0.9126
                 95% CI : (0.9086, 0.9164)
    No Information Rate : 0.7824
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7137

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6563
            Specificity : 0.9838
         Pos Pred Value : 0.9187
         Neg Pred Value : 0.9114
             Prevalence : 0.2176
         Detection Rate : 0.1428
   Detection Prevalence : 0.1555
      Balanced Accuracy : 0.8201
```

In Sample

```
          Reference
Prediction    0    1
         0 3311  316
         1   65  606

               Accuracy : 0.9114
                 95% CI : (0.9025, 0.9197)
    No Information Rate : 0.7855
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7081

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6573
            Specificity : 0.9807
         Pos Pred Value : 0.9031
         Neg Pred Value : 0.9129
             Prevalence : 0.2145
         Detection Rate : 0.1410
   Detection Prevalence : 0.1561
      Balanced Accuracy : 0.8190
```

Out of Sample

Accuracy

Model	In-Sample Accuracy	Out-of-Sample Accuracy
Logistic Regression 1	0.8676	0.8751
Logistic Regression 2	0.867	0.8749
CART	0.9224	0.9235
Support Vector Model	0.9126	0.9114

92.9%

accurate predictive model.

```
          Reference
Prediction    0    1
         0 3799  327
         1   21  739

               Accuracy : 0.9288
                 95% CI : (0.9212, 0.9358)
    No Information Rate : 0.7818
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7671

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6932
            Specificity : 0.9945
         Pos Pred Value : 0.9724
         Neg Pred Value : 0.9207
             Prevalence : 0.2182
         Detection Rate : 0.1512
   Detection Prevalence : 0.1555
      Balanced Accuracy : 0.8439
```

TESTING SET

CART

-CART is the most accurate followed by SVM

Limitations: CPU performance

- Next Step:

- Fine tuning the models
- Cross Validation
- Risk Curve creation

Revisiting the Hypothesis:

- Regression- Disappointing RSE results. Likely missing exogenous variables
- Classification- High accuracy on imbalanced partitioned data

CONCLUSION