Cary Wang

Lab 5 – Lending Club

Part 2: Descriptive Statistics

<u>Proportion of *highgrade*</u>: 41.6%

<u>Median Income t-test</u>

```
Welch Two Sample t-test

data:  loan_data$highgrade and loan_data$med_income
t = -39.411, df = 471180, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.05979526 -0.05412959
sample estimates:
mean of x mean of y 
0.4160905 0.4730530
```

<u>Median Loan Amount t-test</u>

```
Welch Two Sample t-test

data:  loan_data$highgrade and loan_data$req_above
t = -57.177, df = 471160, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.08553696 -0.07986713
sample estimates:
mean of x mean of y 
0.4160905 0.4987926
```

<u>Home Ownership t-test</u>

```
Welch Two Sample t-test

data:  loan_data$highgrade and loan_data$home_rent
t = 15.909, df = 471220, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01994136 0.02554541
sample estimates:
mean of x mean of y 
0.4160905 0.3933472
```

Part 3: Logistic Classifier

<u>GLM Summary</u>

```
Call:
glm(formula = highgrade ~ annual_inc + home_ownership + loan_amnt, 
    data = loan_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-8.9559  -0.4179  -0.3304   0.5563   0.7912
```

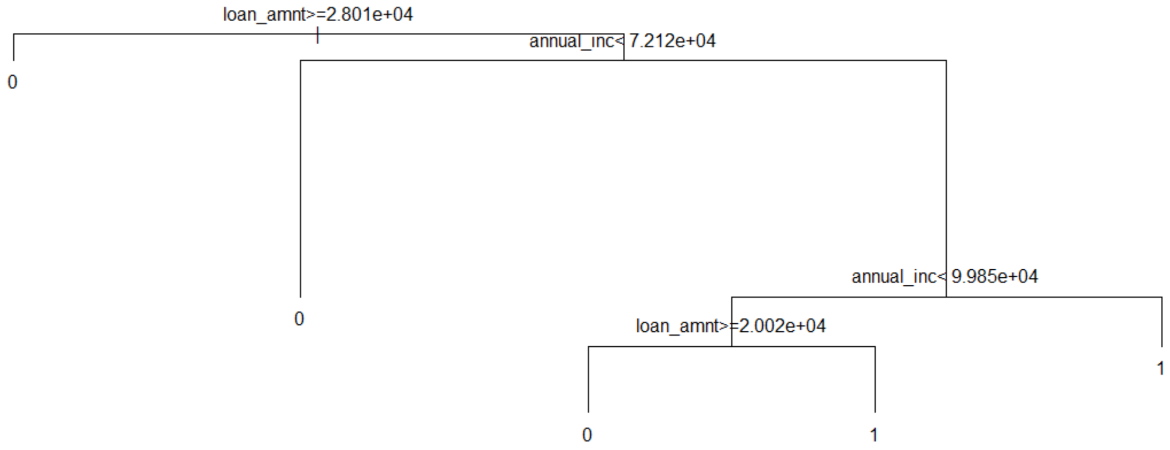
```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)             1.003e+00  4.866e-01   2.062   0.0392 *  
annual_inc              1.148e-06  1.973e-08  58.156   <2e-16 ***
home_ownershipMORTGAGE -5.224e-01  4.867e-01  -1.073   0.2831    
home_ownershipOWN      -5.463e-01  4.867e-01  -1.123   0.2616    
home_ownershipRENT     -5.653e-01  4.866e-01  -1.162   0.2454    
loan_amnt              -8.842e-06  1.307e-07 -67.642   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.2368242)

    Null deviance: 57248  on 235628  degrees of freedom
Residual deviance: 55801  on 235623  degrees of freedom
AIC: 329285

Number of Fisher Scoring iterations: 2
```

Accuracy: 41.48% error

Random Benchmark: 50.21%

All-Zero Benchmark: 41.6%

Part 4: Supervised Learning



Model Accuracy: 39% error – it is more accurate than the logistic model by 2%

Part 5: Test Data

Logistic Model Error: 44.9%

Classification Tree Error: 38.2%

Random Benchmark: 50.0%

All-Zero Benchmark: 45.3%