Freddie Mac Initial Data Analysis

Anonymous CVPR submission

Paper ID ****

1. Introduction

Freddie Mac holds a large portion of the United States of America's home mortgages. As such looking for trends in the values and performance of these loans is crucial. Our dataset consists of data collected at the time loans were issued and a calculated Net Present Value (NPV) that acts as a proxy for the total value of each loan. This paper summarizes a subset of the different variables found in the dataset. The dataset can be found on my GitHub at the following address: https://tinyurl.com/ycpr8lrj

2. Linear Regression

2.1. Simple Linear

We begin our analysis with a simple linear regression using all the variables our dataset provides to us. Two variables have been removed due to a lack of contrast. These columns only contained one value so they are not useful for prediction. Additional categorical variables such as zip code and MSA code were removed due to a high number of insignificant dummy variables being produced. The resulting summary of our first linear regression can be found in Figure 1. This shows that some of our variables have very little impact on predicting NPV. Particularly the number of borrowers associated with the loan. The introduction of a second borrower was only shown to lower the value of the NPV by around 11 cents which is very insignificant. However, were many variables that showed correlations that were very statistically significant. However, our R_a^2 value is very low at only .01026. Additionally the residual plots found in Figure 2 show strong heteroscedasticity in our dataset particularly around the mean. In an attempt to boost predictive power we look to remove some variables.

2.2. Variable Reduction

We begin a simple variable reduction by removing variables that are not statistically significant in our first regression. A summary of this regression can be shown in Figure 3 and the residual plots can be found in Figure 4. This summary shows that our R_a^2 value actually decreased and our residual plots look largely the same. However, this should be useful for the next step where we check for interactions between these variables.

2.3. Interaction Terms

Due to screen size limitations Figure 5 does not show all variables included in the regression, however the R_a^2 value jumped to .01152. Other interesting finds were that variables that were previously significant such as Credit Score now become insignificant as most of the predictive power was wrapped up in interactions with other variables in the dataset. Again, the residuals shows in Figure 6 show similar trends to the previous residual plots.

3. Conclusion

These models show that there is a lot of discretion available for variable selection when creating a linear model for this dataset. Although this is not a comprehensive review, it provides a good baseline to build upon with more advanced statistical techniques.

4. Figures

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.033e+05  7.152e+03 -14.447  < 2e-16 ***
CreditScore             -5.628e+00  7.624e-01  -7.382 1.57e-13 ***
FirstTimeHomebuyerY     -7.164e+02  1.152e+02  -6.217 5.09e-10 ***
MI.Percentage            3.518e+01  4.130e+00   8.519  < 2e-16 ***
Number.of.Units         -4.557e+02  2.259e+02  -2.017 0.043692 *
Occupancy.StatusO        2.709e+02  2.699e+02   1.004 0.315556
Occupancy.StatusS        1.485e+02  4.160e+02   0.357 0.721125
CLTV                     6.637e+01  6.767e+01   0.981 0.326661
DTI                     -2.008e+01  3.452e+00  -5.816 6.04e-09 ***
UPB                      1.557e-02  7.387e-04  21.077  < 2e-16 ***
LTV                     -5.854e+01  6.784e+01  -0.863 0.388199
Interest.Rate           -9.437e+02  1.134e+02  -8.318  < 2e-16 ***
ChannelC                 2.123e+03  2.005e+03   1.059 0.289607
ChannelR                 3.720e+02  1.521e+03   0.245 0.806816
ChannelT                 1.512e+03  1.521e+03   0.994 0.320088
PPMY                     9.370e+02  2.612e+02   3.587 0.000334 ***
Property.TypeCP         -1.537e+03  2.095e+03  -0.734 0.463091
Property.TypeLH         -6.241e+03  2.346e+03  -2.660 0.007804 **
Property.TypeMH          1.053e+03  9.341e+02   1.127 0.259666
Property.TypePU          1.926e+03  1.812e+02  10.631  < 2e-16 ***
Property.TypeSF          1.001e+03  1.524e+02   6.567 5.14e-11 ***
Loan.PurposeN            4.882e+01  1.162e+02   0.420 0.674292
Loan.PurposeP            1.374e+02  1.216e+02   1.130 0.258366
Original.Term            3.561e+02  1.909e+01  18.651  < 2e-16 ***
Borrower.Num            -1.102e+01  7.962e+01  -0.138 0.889927
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15350 on 178033 degrees of freedom
  (213361 observations deleted due to missingness)
Multiple R-squared:  0.01039,   Adjusted R-squared:  0.01026
F-statistic: 77.92 on 24 and 178033 DF,  p-value: < 2.2e-16
```

Figure 1. Summary from simple linear regression

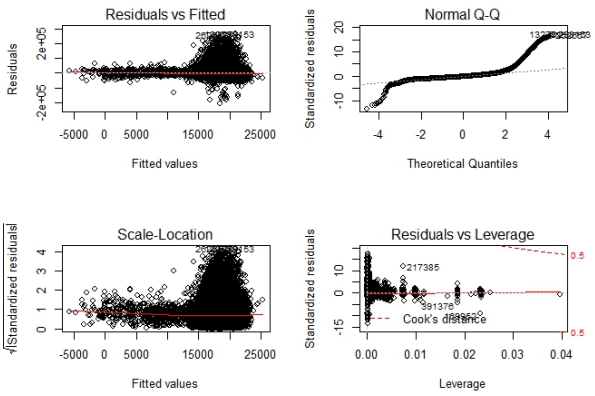

Figure 2. Residual plots of simple linear regression

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.001e+05  5.977e+03 -16.753  < 2e-16 ***
CreditScore             -5.967e+00  6.852e-01  -8.708  < 2e-16 ***
FirstTimeHomebuyerY     -7.824e+02  1.015e+02  -7.707 1.29e-14 ***
MI.Percentage            3.985e+01  2.720e+00  14.654  < 2e-16 ***
Number.of.Units         -5.377e+02  1.828e+02  -2.941  0.00327 **
DTI                     -1.847e+01  3.136e+00  -5.889 3.88e-09 ***
UPB                      1.774e-02  6.479e-04  27.379  < 2e-16 ***
Interest.Rate           -1.060e+03  1.014e+02 -10.446  < 2e-16 ***
Property.TypeCP         -1.063e+03  2.075e+03  -0.512  0.60870
Property.TypeLH         -5.435e+03  2.021e+03  -2.690  0.00715 **
Property.TypeMH          1.535e+03  7.189e+02   2.135  0.03273 *
Property.TypePU          1.812e+03  1.702e+02  10.647  < 2e-16 ***
Property.TypeSF          9.330e+02  1.399e+02   6.669 2.57e-11 ***
Original.Term            3.549e+02  1.638e+01  21.662  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15360 on 209114 degrees of freedom
  (182291 observations deleted due to missingness)
Multiple R-squared:  0.009676,   Adjusted R-squared:  0.009615
F-statistic: 157.2 on 13 and 209114 DF,  p-value: < 2.2e-16
```

Figure 3. Summary from regression after variable reduction

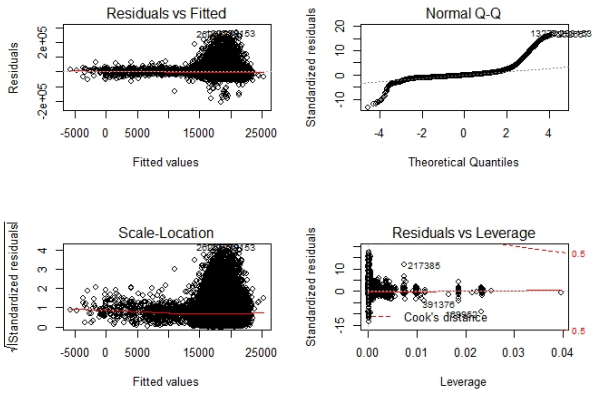

Figure 4. Residual plots after variable reduction

```
Coefficients: (1 not defined because of singularities)
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                   -5.612e+05  1.288e+05  -4.356 1.32e-05 ***
CreditScore                                   -4.644e+00  1.076e+02  -0.043 0.965580
FirstTimeHomebuyerY                           -1.035e+05  2.946e+04  -3.514 0.000441 ***
MI.Percentage                                  3.753e+02  5.878e+02   0.638 0.523165
Number.of.Units                                4.806e+04  4.223e+04   1.138 0.255087
DTI                                           -1.828e+03  5.563e+02  -3.287 0.001013 **
UPB                                           -4.404e-01  1.034e-01  -4.258 2.06e-05 ***
Interest.Rate                                  8.206e+04  9.999e+03   8.206 2.29e-16 ***
Property.TypeCP                                1.323e+05  1.283e+05   1.031 0.302539
Property.TypeLH                                6.404e+03  1.752e+05   0.037 0.970846
Property.TypeMH                                2.200e+04  1.695e+05   0.130 0.896731
Property.TypePU                               -6.092e+04  4.359e+04  -1.397 0.162265
Property.TypeSF                               -6.180e+03  3.144e+04  -0.197 0.844160
Original.Term                                  1.563e+03  3.557e+02   4.393 1.12e-05 ***
CreditScore:FirstTimeHomebuyerY                5.908e+00  2.071e+00   2.853 0.004335 **
CreditScore:MI.Percentage                      1.060e-01  5.452e-02   1.944 0.051899 .
CreditScore:Number.of.Units                    1.256e+01  3.835e+00   3.276 0.001053 **
CreditScore:DTI                                6.089e-02  6.393e-02   0.953 0.340843
CreditScore:UPB                               -3.541e-05  1.335e-05  -2.651 0.008023 **
CreditScore:Interest.Rate                     -2.445e+00  1.927e+00  -1.269 0.204418
CreditScore:Property.TypeCP                   -4.281e+01  4.351e+01  -0.984 0.325212
CreditScore:Property.TypeLH                    3.010e+01  4.976e+01   0.605 0.545319
CreditScore:Property.TypeMH                   -3.504e+00  1.455e+01  -0.241 0.809675
CreditScore:Property.TypePU                   -2.594e+00  3.501e+00  -0.741 0.458857
CreditScore:Property.TypeSF                   -2.276e+00  2.829e+00  -0.804 0.421230
CreditScore:Original.Term                      1.541e-02  2.959e-01   0.052 0.958456
FirstTimeHomebuyerY:MI.Percentage              1.011e+01  7.731e+00   1.308 0.190766
FirstTimeHomebuyerY:Number.of.Units            1.374e+02  6.410e+02   0.214 0.830250
FirstTimeHomebuyerY:DTI                       -3.242e+00  1.035e+01  -0.313 0.754171
FirstTimeHomebuyerY:UPB                        6.227e-03  2.037e-03   3.057 0.002233 **
FirstTimeHomebuyerY:Interest.Rate              1.965e+03  2.919e+02   6.732 1.68e-11 ***
FirstTimeHomebuyerY:Property.TypeCP            6.334e+03  5.076e+03   1.248 0.212041
```

Figure 5. Head of the summary of regression with interactions

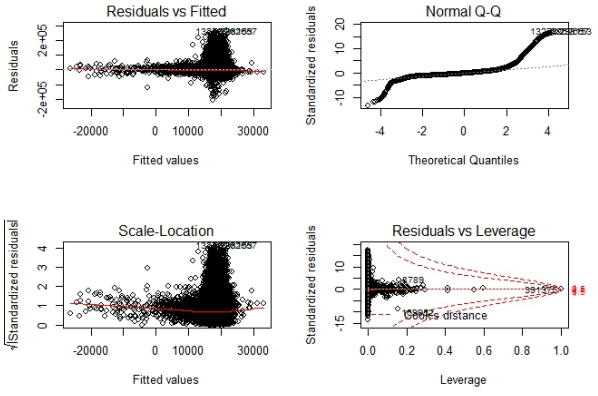

Figure 6. Residual plots with interaction terms