Georgia Tech

MGT 6203 Group Project Progress Report

ASSESS CREDIT INVESTOR RISK IN FIXED RATE MORTGAGE LOANS

TEAM 1:

HARIPRASAD RAMAMOORTHY
LAXMINARAYANAN KRISHNAN
SERDAR AYDINOGLU
UNNAMALAI SUPPIAH
BHAKTI PATEL

*For this project, the term default and delinquency will be used interchangeably.

Table of Contents

BACKGROUND

The U.S. mortgage industry is among the largest in the world. In fact, mortgage debt is one of the main sources of debt held by Americans ("Topic: Mortgage Industry in the U.S."). A mortgage is "a type of loan used to purchase or maintain a home, land, or other types of real estate" (Kagan). Borrowers of the mortgage agree to pay the lender over an agreed period –typically in a series of regular payments divided into principal and interest. The property then serves as collateral to secure the loan until the borrower pays back the full amount of the loan. Credit risk is the event that the borrower fails to make payments and defaults on the loan payments.

A typical mortgage in the US can be anywhere from 15 - 30 years and there is a significant risk involved for the lender when it comes to repayment. Lenders often leverage securitization as a common financial tool used to manage risk. Figure 1 below provides an illustration of mortgage securitization process with Government Sponsored Enterprises such as Fannie Mae. In securitization, a mortgage is typically originated by a mortgage bank and is subsequently sold to the GSEs. The cash flows of like mortgages are bundled as securities and sold to investors through mortgage-backed securities (MBS). The GSE assumes the credit risk of the mortgages by providing investors with a guarantee.

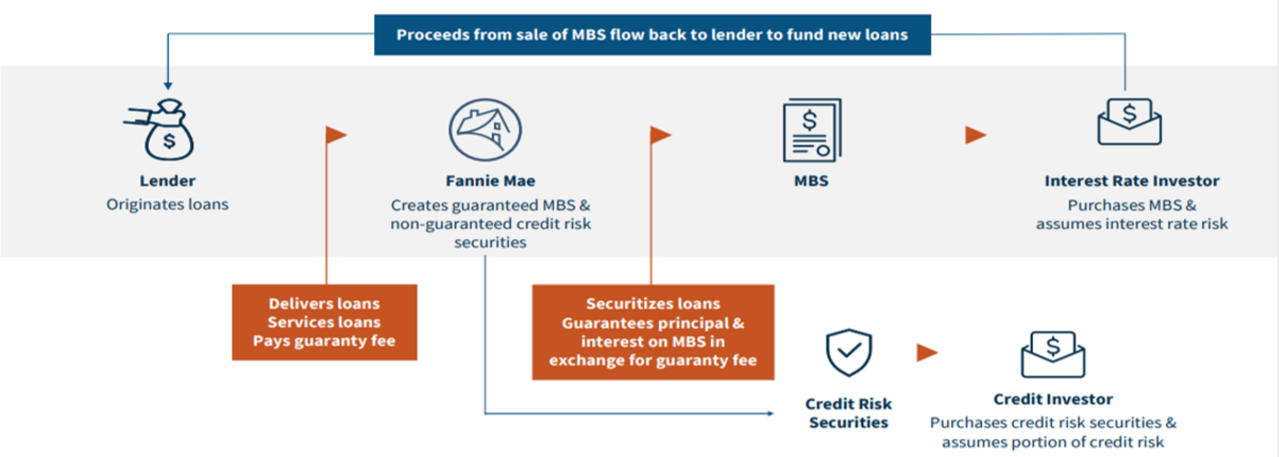
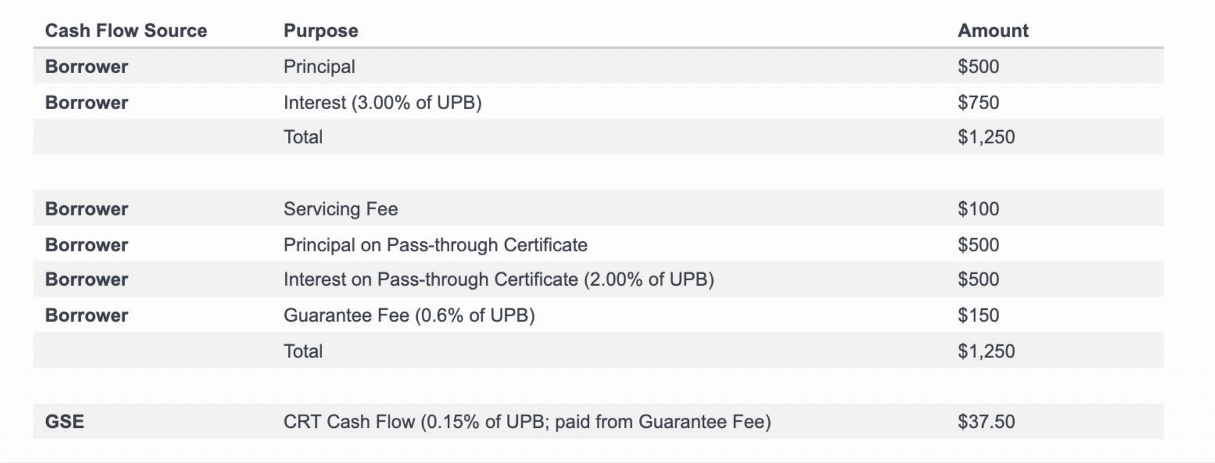

Proceeds from sale of MBS flow back to lender to fund new loans
Lender
Originates loans
Fannie Mae
Creates guaranteed MBS & non-guaranteed credit risk securities
MBS
Interest Rate Investor
Purchases MBS & assumes interest rate risk
Delivers loans
Services loans
Pays guaranty fee
Securitizes loans
Guarantees principal & interest on MBS in exchange for guaranty fee
Credit Risk Securities
Credit Investor
Purchases credit risk securities & assumes portion of credit risk

Figure 1: Illustration of the US Mortgage Market (Fannie Mae)

As seen in Figure 1, the GSEs may also securitize a portion of their guaranteed fees to Credit Risk Transfer (CRT) to private investors as a source of capital for potential losses. The practice of CRT securitization is part of a set of strategic objectives by the Federal Housing Finance Agency (FHFA) to encourage GSE's to share mortgage credit risk with private investors to minimize impact on taxpayer dollars. Figure 2 below illustrates a sample cash flow of how a monthly payment is spread across the securitization process. In a $1250 monthly payment, $100 goes to the mortgage servicer, $1,000 to pass-through certificate holders or MBS investors, and $150 to the GSE. The $37.50 of $150 paid to the GSE as part of the guarantee fee is passed to the CRT investors.



Cash Flow Source	Purpose	Amount
Borrower	Principal	$500
Borrower	Interest (3.00% of UPB)	$750
	Total	$1,250
Borrower	Servicing Fee	$100
Borrower	Principal on Pass-through Certificate	$500
Borrower	Interest on Pass-through Certificate (2.00% of UPB)	$500
Borrower	Guarantee Fee (0.6% of UPB)	$150
	Total	$1,250
GSE	CRT Cash Flow (0.15% of UPB; paid from Guarantee Fee)	$37.50

Figure 2: Illustration of Cash Flow within the Securitization Process (Glowacki)

OBJECTIVE

Upon considering the multi-faceted securitization process above, it becomes quite clear the risk CRT investors bear and the importance of having the appropriate tools to make investments in CRT. The risk of a mortgage non-payment involves significant capital. Default or delinquency is difficult to estimate based on simple credit rating scores, given other factors such as purpose of property purchase, its location, and market rates, debt to income ratio might have a significant impact on when the loan is either refinanced, paid-off (loss of revenue interest stream) or defaulted (left unpaid). Hence, a composite model that weighs these distinct factors to predict the probability of delinquency can help financial institutions such as Fannie Mae and private investors to minimize investment losses.

Our team proposes to build a model to determine the probability of default of a mortgage based upon various characteristics available when the mortgage is originated. The initial period of loan underwriting for mortgages may be the riskiest phase of the investment since no historic loan repayment data are available to guide the underwriting process. Having a predictor of the risk based on borrower, property, and market location data – including any potential for past due payments – could be used to predict the probability of default and thus loss of value from interest & not principal payments prior to selling it off by bundling mortgages in the form of Credit Risk Transfer (CTR) securities. The result of the model can be used to determine the "Credit Risk Insurance" minimum premium to be charged to each mortgage.

The model's objective is to establish the probability of default which can be paired with the outstanding loan balance to determine the credit risk size. Thus, the model would help put a price on credit risk. Knowing the fair price of credit risk would be helpful to whoever will eventually own that credit risk. Investors are one of the classes of risk owner; another would be mortgage insurers be it public like Fannie Mae and Freddie Mac, or private like Old Republic and MGIC.

INITIAL HYPOTHESIS

We believe loans with high debt-to-income (DTI) ratio, poor loan to value (LTV), poor FICO score, and poor history of repayment in recent months will present higher chances of delinquency. The hypothesis was derived from a combination of literary sources such as “Do Riskier Borrowers Borrow More” along with personal experience working in the field. The conclusion of the LTV to default risk analysis in the article “Do Riskier Borrowers Borrow More” was “lenders need to recognize that a borrower’s choice of LTV provides a qualified signal about that borrower’s default riskiness” (Harrison et al.).

Furthermore, an analysis conducted by the Urban Institute further confirmed our hypothesis as it concluded that both DTI (Debit to Income) and credit scores are good indicators of default as shown in Figure 3 and 4, respectively.

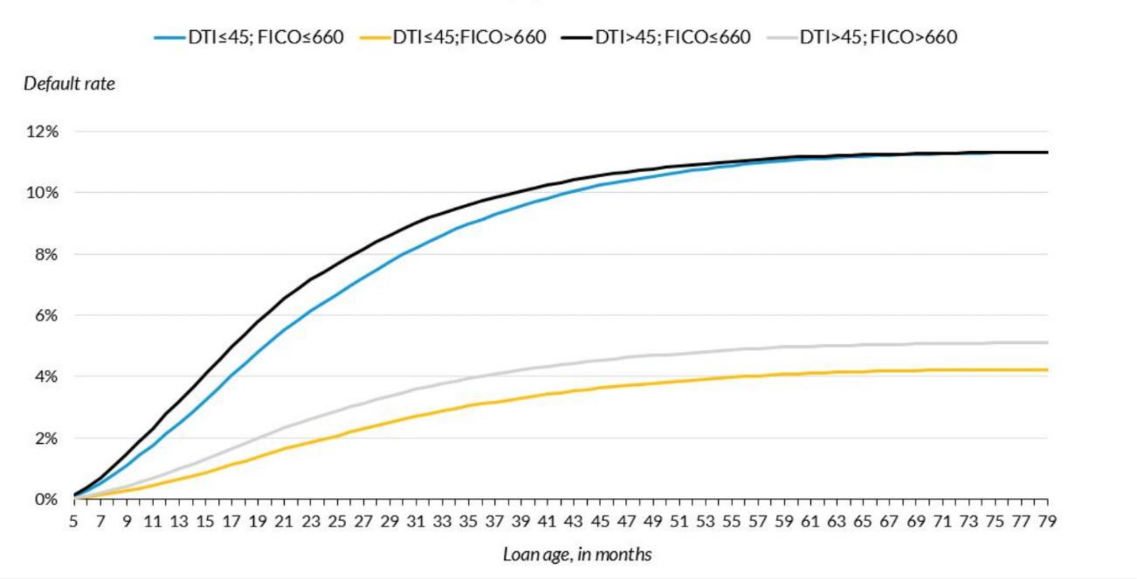


Figure 3: Default on Loans by DTI & Credit Score (Harrison et al.)

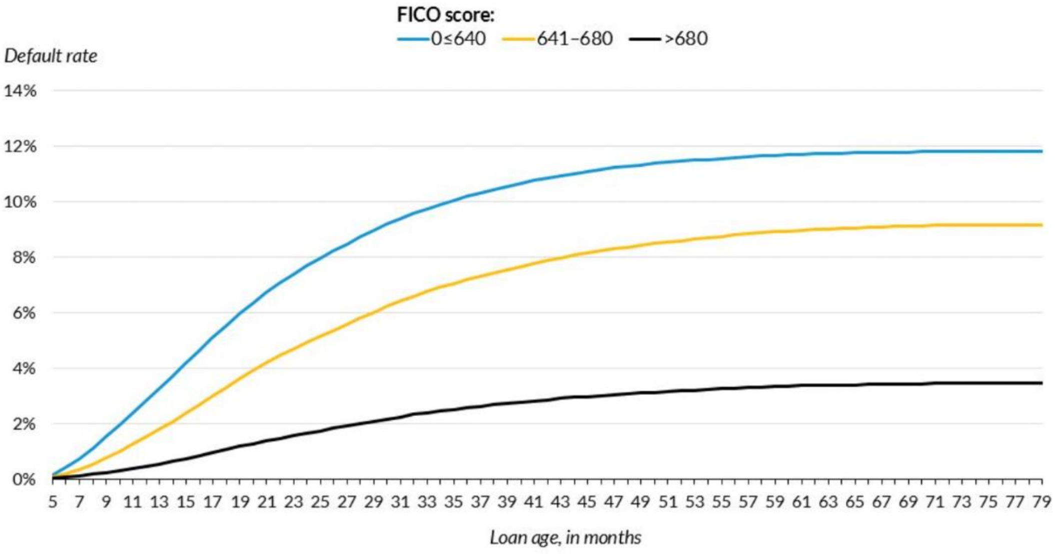


Figure 4: Default on Loan Credit Score (Harrison et al.)

Please note that there are different FICO models like FICO 2, FICO 4, etc., FICO 9 is the newest version, but is not widely used. Different lenders use different models depending on the type of loan. For example: FICO 8 is mostly used for Auto and Bankcard lending processes, while FICO 2,4 and 5 are extensively used for mortgage lending. The main difference between the FICO models is the credit bureau from which the data is derived for the models - Experian, TransUnion, and Equifax. Mortgage lenders pull one of each and compile the reports in a document called "Residential Mortgage Credit Report". The mean of the three is used to represent the worthiness and payback of the mortgage. For our project, the credit scores provided by Fannie Mae are already an aggregate score across all the bureaus.

METHODOLOGY

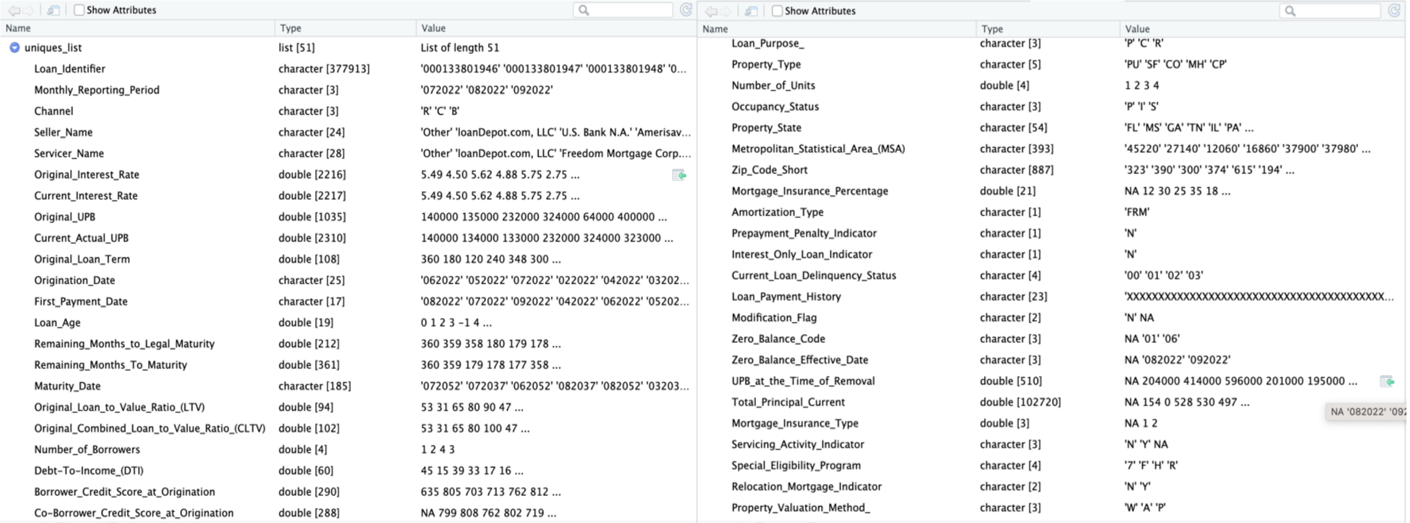
DATA SOURCE

The primary data source for the project was originally Fannie Mae's quarterly loan characteristics report for Yr2022-Q3. The data set consists of Fixed Rate Mortgage (FRM) loans data. We have since expanded our new data source to include data for the following period: Yr2022-Q1, Q2, Q3 and Yr2021-Q1. Since the volume of files was big, we could not upload to GitHub. On average, the file size was 70GB. The team uploaded the files and final code to our Team's Channel; the link is available in the data source section at the end.

Although Fixed Rate Mortgage (FRM) data will be the focus of this project in the interest of time, the team also plans to investigate differences in geography (property state and zip-code affluence) data and 10-year average US treasury rates to add dimension to model when predicting the probability of default. The data was acquired from the IRS and Wall Street Journal, respectively. The zip code data will be used to determine if loan defaults are more prominent in some areas of the country in comparison to another. Additionally, the US treasury rates will be used to analyze loans whose mortgage rates deviate significantly and therefore represent a riskier borrower with a greater risk of default. If needed, manual weighting of treasury data will be conducted to minimize impact of last two years of pandemic and might be a companion model examined in single factor regression comparisons.

DATA CLEANSING & TRANSFORMATION

Preparing the primary dataset containing the loan data for one reporting quarter to be model ready was our first focus. The data originally contained 108 attributes of which only 51 are available for the project due to data sharing or other restrictions (PII). Hence, we first began by removing all the data that was unavailable leaving on the 51 available attributes shown below in Figure 5.

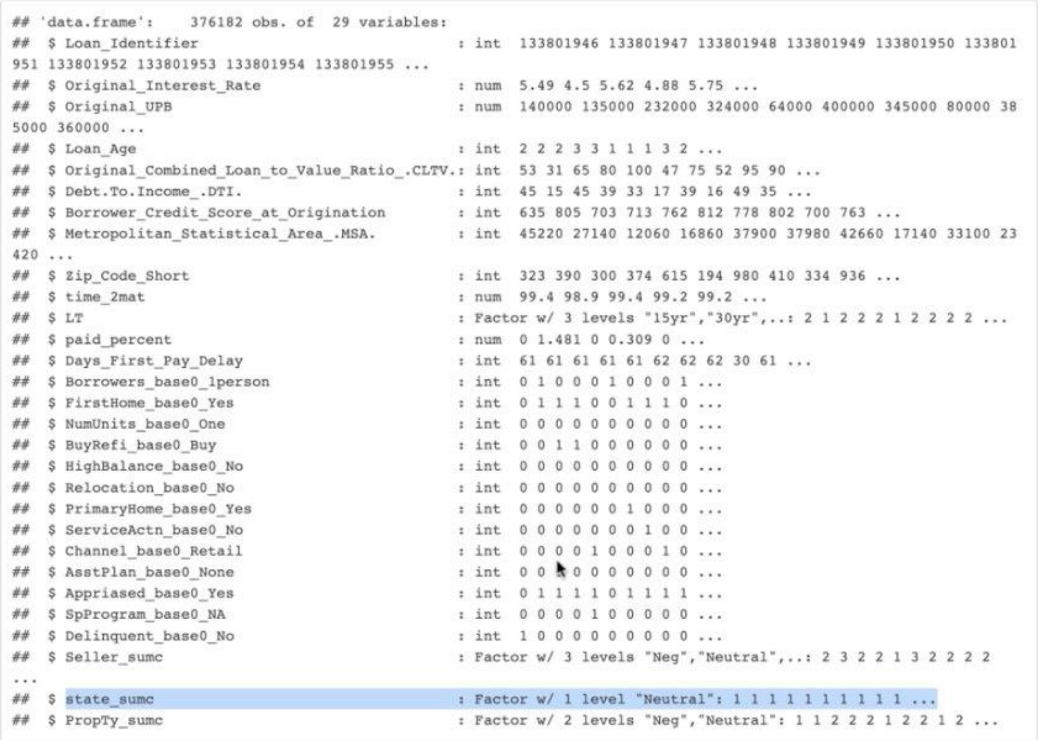


Name	Type	Value
uniques_list	list [51]	List of length 51
Loan_Identifier	character [377913]	'000133801946' '000133801947' '000133801948' '0...
Monthly_Reporting_Period	character [3]	'072022' '082022' '092022'
Channel	character [3]	'R' 'C' 'B'
Seller_Name	character [24]	'Other' 'loanDepot.com, LLC' 'U.S. Bank N.A.' 'Amerisav...
Servicer_Name	character [28]	'Other' 'loanDepot.com, LLC' 'Freedom Mortgage Corp....
Original_Interest_Rate	double [2216]	5.49 4.50 5.62 4.88 5.75 2.75 ...
Current_Interest_Rate	double [2217]	5.49 4.50 5.62 4.88 5.75 2.75 ...
Original_UPB	double [1035]	140000 135000 232000 324000 64000 400000 ...
Current_Actual_UPB	double [2310]	140000 134000 133000 232000 324000 323000 ...
Original_Loan_Term	double [108]	360 180 120 240 348 300 ...
Origination_Date	character [25]	'062022' '052022' '072022' '022022' '042022' '03202...
First_Payment_Date	character [17]	'082022' '072022' '092022' '042022' '062022' '05202...
Loan_Age	double [19]	0 1 2 3 -1 4 ...
Remaining_Months_to_Legal_Maturity	double [212]	360 359 358 180 179 178 ...
Remaining_Months_To_Maturity	double [361]	360 359 179 178 177 358 ...
Maturity_Date	character [185]	'072052' '072037' '062052' '082037' '082052' '03203...
Original_Loan_to_Value_Ratio_(LTV)	double [94]	53 31 65 80 90 47 ...
Original_Combined_Loan_to_Value_Ratio_(CLTV)	double [102]	53 31 65 80 100 47 ...
Number_of_Borrowers	double [4]	1 2 4 3
Debt-To-Income_(DTI)	double [60]	45 15 39 33 17 16 ...
Borrower_Credit_Score_at_Origination	double [290]	635 805 703 713 762 812 ...
Co-Borrower_Credit_Score_at_Origination	double [288]	NA 799 808 762 802 719 ...

Name	Type	Value
Loan_Purpose_	character [3]	'P' 'C' 'R'
Property_Type	character [5]	'PU' 'SF' 'CO' 'MH' 'CP'
Number_of_Units	double [4]	1 2 3 4
Occupancy_Status	character [3]	'P' 'I' 'S'
Property_State	character [54]	'FL' 'MS' 'GA' 'TN' 'IL' 'PA' ...
Metropolitan_Statistical_Area_(MSA)	character [393]	'45220' '27140' '12060' '16860' '37900' '37980' ...
Zip_Code_Short	character [887]	'323' '390' '300' '374' '615' '194' ...
Mortgage_Insurance_Percentage	double [21]	NA 12 30 25 35 18 ...
Amortization_Type	character [1]	'FRM'
Prepayment_Penalty_Indicator	character [1]	'N'
Interest_Only_Loan_Indicator	character [1]	'N'
Current_Loan_Delinquency_Status	character [4]	'00' '01' '02' '03'
Loan_Payment_History	character [23]	'XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX...
Modification_Flag	character [2]	'N' NA
Zero_Balance_Code	character [3]	NA '01' '06'
Zero_Balance_Effective_Date	character [3]	NA '082022' '092022'
UPB_at_the_Time_of_Removal	double [510]	NA 204000 414000 596000 201000 195000 ...
Total_Principal_Current	double [102720]	NA 154 0 528 530 497 ...
Mortgage_Insurance_Type	double [3]	NA 1 2
Servicing_Activity_Indicator	character [3]	'N' 'Y' NA
Special_Eligibility_Program	character [4]	'7' 'F' 'H' 'R'
Relocation_Mortgage_Indicator	character [2]	'N' 'Y'
Property_Valuation_Method_	character [3]	'W' 'A' 'P'

Figure 5: Cleaned Data with 51 Attributes

Next, we ensured that the data within the available attributes had less than 5% missing values and that all loan IDs were unique. Reports were scrutinized over the quarterly monthly record to ensure that any unique Loan is included in the final data. Once the data were cleaned, we began to evaluate each multi-level categorical attribute to see how we can collapse them meaningfully to allow for significance detection. We binarized attributes with two-character levels with defined base-case conditions and collapsed factors with greater than three levels for modeling efficiency directly by interpreting the meaning of the levels as it applies to our modeling effort or by using single factor regressions using differential or contrast coding instead of the typical reference level coding scheme. The final data once cleansed and prepped is shown in Figure 6 below.



```
## 'data.frame':    376182 obs. of  29 variables:
##  $ Loan_Identifier                          : int  133801946 133801947 133801948 133801949 133801950 133801
951 133801952 133801953 133801954 133801955 ...
##  $ Original_Interest_Rate                   : num  5.49 4.5 5.62 4.88 5.75 ...
##  $ Original_UPB                             : num  140000 135000 232000 324000 64000 400000 345000 80000 38
5000 360000 ...
##  $ Loan_Age                                 : int  2 2 2 3 3 1 1 1 3 2 ...
##  $ Original_Combined_Loan_to_Value_Ratio_.CLTV.: int  53 31 65 80 100 47 75 52 95 90 ...
##  $ Debt.To.Income_.DTI.                     : int  45 15 45 39 33 17 39 16 49 35 ...
##  $ Borrower_Credit_Score_at_Origination     : int  635 805 703 713 762 812 778 802 700 763 ...
##  $ Metropolitan_Statistical_Area_.MSA.      : int  45220 27140 12060 16860 37900 37980 42660 17140 33100 23
420 ...
##  $ Zip_Code_Short                           : int  323 390 300 374 615 194 980 410 334 936 ...
##  $ time_2mat                                : num  99.4 98.9 99.4 99.2 99.2 ...
##  $ LT                                       : Factor w/ 3 levels "15yr","30yr",..: 2 1 2 2 2 1 2 2 2 2 ...
##  $ paid_percent                             : num  0 1.481 0 0.309 0 ...
##  $ Days_First_Pay_Delay                     : int  61 61 61 61 61 62 62 62 30 61 ...
##  $ Borrowers_base0_1person                  : int  0 1 0 0 0 1 0 0 0 1 ...
##  $ FirstHome_base0_Yes                      : int  0 1 1 1 0 0 1 1 1 0 ...
##  $ NumUnits_base0_One                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ BuyRefi_base0_Buy                        : int  0 0 1 1 0 0 0 0 0 0 ...
##  $ HighBalance_base0_No                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Relocation_base0_No                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PrimaryHome_base0_Yes                    : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ ServiceActn_base0_No                     : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Channel_base0_Retail                     : int  0 0 0 0 1 0 0 0 1 0 ...
##  $ AsstPlan_base0_None                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Appriased_base0_Yes                      : int  0 1 1 1 1 0 1 1 1 1 ...
##  $ SpProgram_base0_NA                       : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ Delinquent_base0_No                      : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ Seller_sumc                              : Factor w/ 3 levels "Neg","Neutral",..: 2 3 2 2 1 3 2 2 2 2
...
##  $ state_sumc                               : Factor w/ 1 level "Neutral": 1 1 1 1 1 1 1 1 1 1 ...
##  $ PropTy_sumc                              : Factor w/ 2 levels "Neg","Neutral": 1 1 2 2 2 1 2 2 1 2 ...
```

Figure 6: Clean & Prepped Data

EXPLORATION & INITIAL DISCOVERIES

Our research aims to examine the factors that contribute to delinquency in loan repayments, using a logistic regression model based on a cleaned dataset with factors collapsed. The first logistic regression model was created with only numeric variables to examine multicollinearity. We then removed correlated factors by checking the VIF values. Using the selected variables, we built another logistic regression model with delinquency label as the response variable. The initial model results indicated that the statistically significant factors that affect delinquency include the original interest rate, loan age, borrower credit score at origination, paid percent, first home, buy refinance, and primary home. Deeper analysis required model cross-validation and estimation of accuracy of predictions.

The correlations between features before and after the multicollinearity check are plotted below.

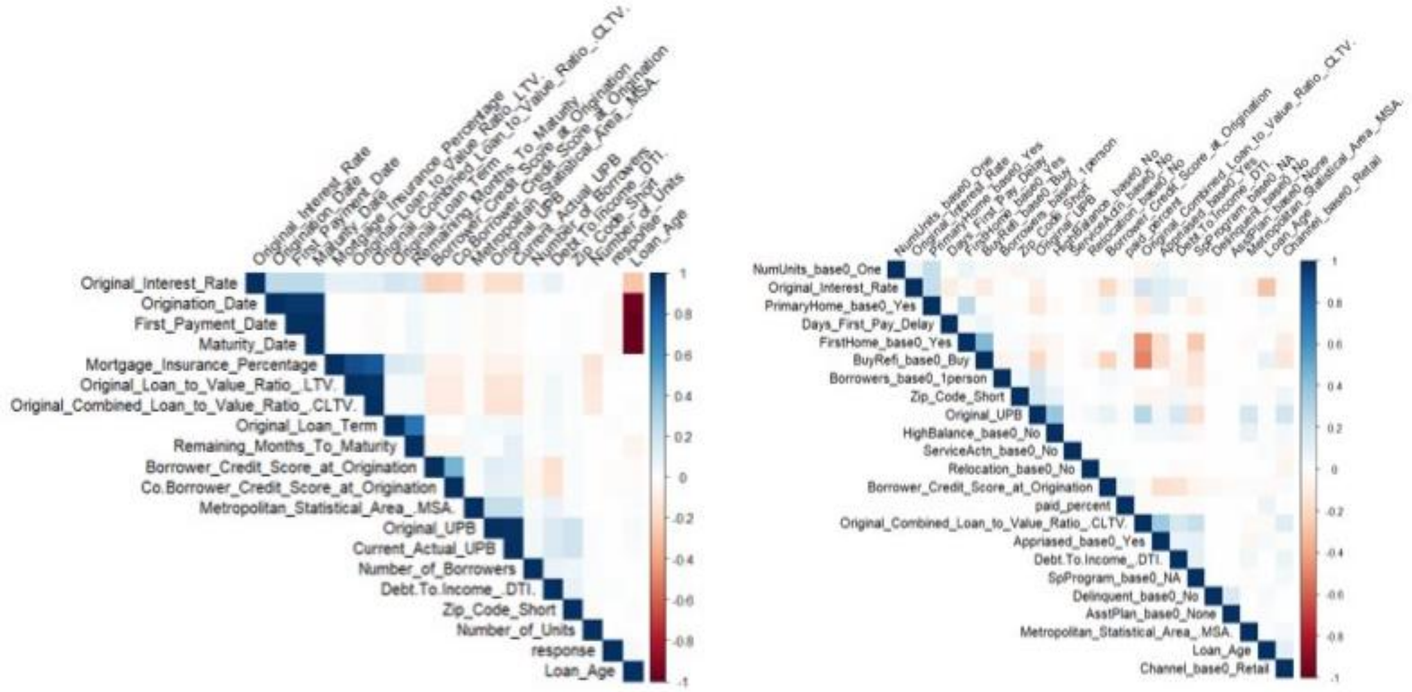


Figure 7: Before (Left) & After (Right) Correlation Plot

The VIF values also show that there is no stronger correlation between the features inside the final data set. All the VIF values are plotted in Figure 8 below.

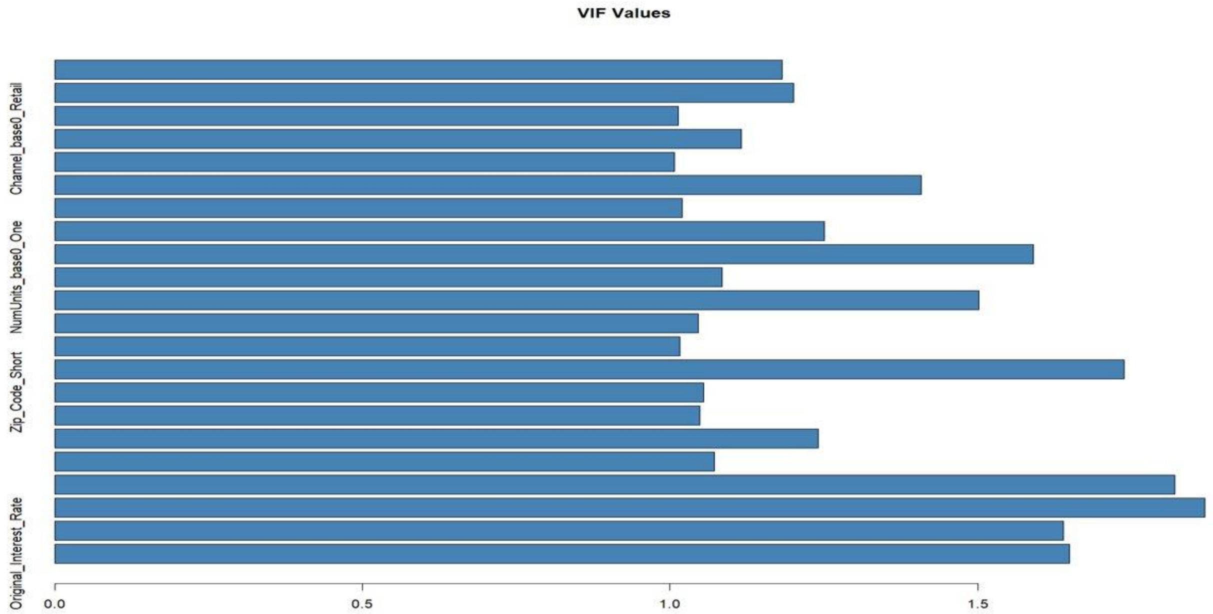
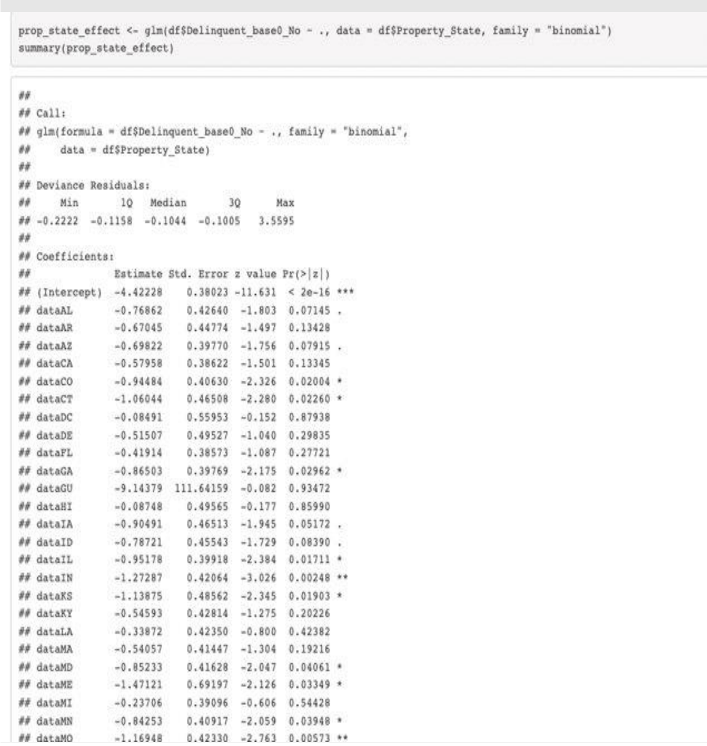


Figure 8: VIF Plot

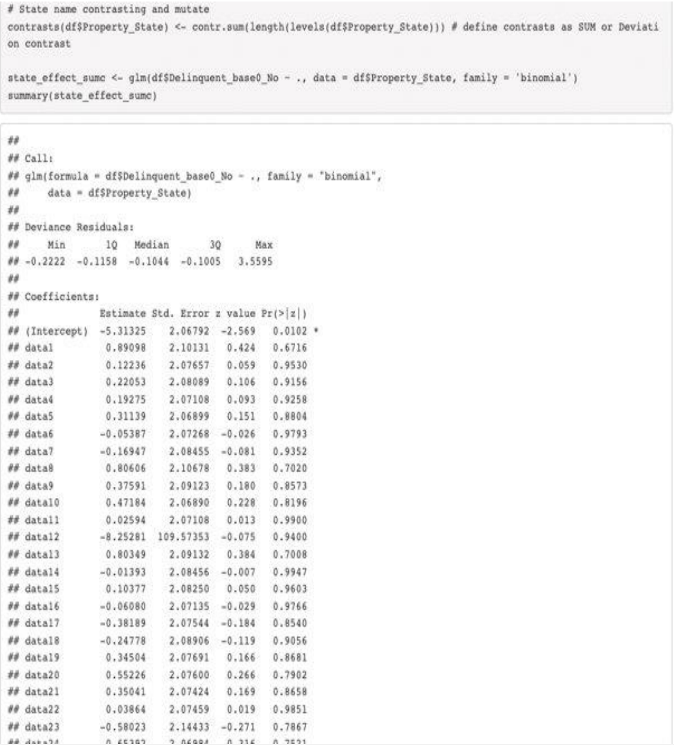
We ran preliminary single-factor generalized linear models to see the effects of state, seller, and property type on delinquency. Initially, the results were difficult to interpret because the model used an arbitrary value as the default reference case to determine significance. Hence, we pivoted to using Sum Contrast, which allowed for the use of population means to determine significance. We found that the use of Sum Contrast was much more effective in this case because we were not interested in marginal effects and wanted to see if there are main effects of state, seller, and property. As mentioned earlier, a default regression model uses dummy coding –meaning that if there are 2 levels within a factor, they would be coded as 0 (first or reference level) and 1 (second level). However, sum contrast allows for these numerical coding to change such that the first level becomes -0.5 (or -1) and the second level becomes 0.5 (or 1). Like continuous predictors in regression models, sum contrasts are like mean-centering a particular continuous predictor.

Using Sum Contrasts, the team was able to do some interesting analysis. We found that state was not at all a significant attribute in our data set in comparison to running the default model (see Figure 9 and 10 below), so geographical factors including zip-code or affluence data will not be examined further in this model. In the case of property type, and seller, the Sum Contrast allowed the team to better pinpoint which types of properties and banks (sellers) were significant when determining risk of default. This coding method allowed us to collapse multi-level variables to just 3 levels (positive & significant, negative & significant, and neutral or no significant influence) based on the results of the contrast coding regression model for each of these factors.



```
prop_state_effect <- glm(df$Delinquent_base0_No ~ ., data = df$Property_State, family = "binomial")
summary(prop_state_effect)
```

```
##
## Call:
## glm(formula = df$Delinquent_base0_No ~ ., family = "binomial",
##     data = df$Property_State)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.2222  -0.1158  -0.1044  -0.1005   3.5595
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -4.42228    0.38023 -11.631  < 2e-16 ***
## dataAL       -0.76862    0.42640  -1.803  0.07145 .
## dataAR       -0.67045    0.44774  -1.497  0.13428
## dataAZ       -0.69822    0.39770  -1.756  0.07915 .
## dataCA       -0.57958    0.38622  -1.501  0.13345
## dataCO       -0.94484    0.40630  -2.326  0.02004 *
## dataCT       -1.06044    0.46508  -2.280  0.02260 *
## dataDC       -0.08491    0.55953  -0.152  0.87938
## dataDE       -0.51507    0.49527  -1.040  0.29835
## dataFL       -0.41914    0.38573  -1.087  0.27721
## dataGA       -0.86503    0.39769  -2.175  0.02962 *
## dataGU       -9.14379  111.64159  -0.082  0.93472
## dataHI       -0.08748    0.49565  -0.177  0.85990
## dataIA       -0.90491    0.46513  -1.945  0.05172 .
## dataID       -0.78721    0.45543  -1.729  0.08390 .
## dataIL       -0.95178    0.39918  -2.384  0.01711 *
## dataIN       -1.27287    0.42064  -3.026  0.00248 **
## dataKS       -1.13875    0.48562  -2.345  0.01903 *
## dataKY       -0.54593    0.42814  -1.275  0.20226
## dataLA       -0.33872    0.42350  -0.800  0.42382
## dataMA       -0.54057    0.41447  -1.304  0.19216
## dataMD       -0.85233    0.41628  -2.047  0.04061 *
## dataME       -1.47121    0.69197  -2.126  0.03349 *
## dataMI       -0.23706    0.39096  -0.606  0.54428
## dataMN       -0.84253    0.40917  -2.059  0.03948 *
## dataMO       -1.16948    0.42330  -2.763  0.00573 **
```



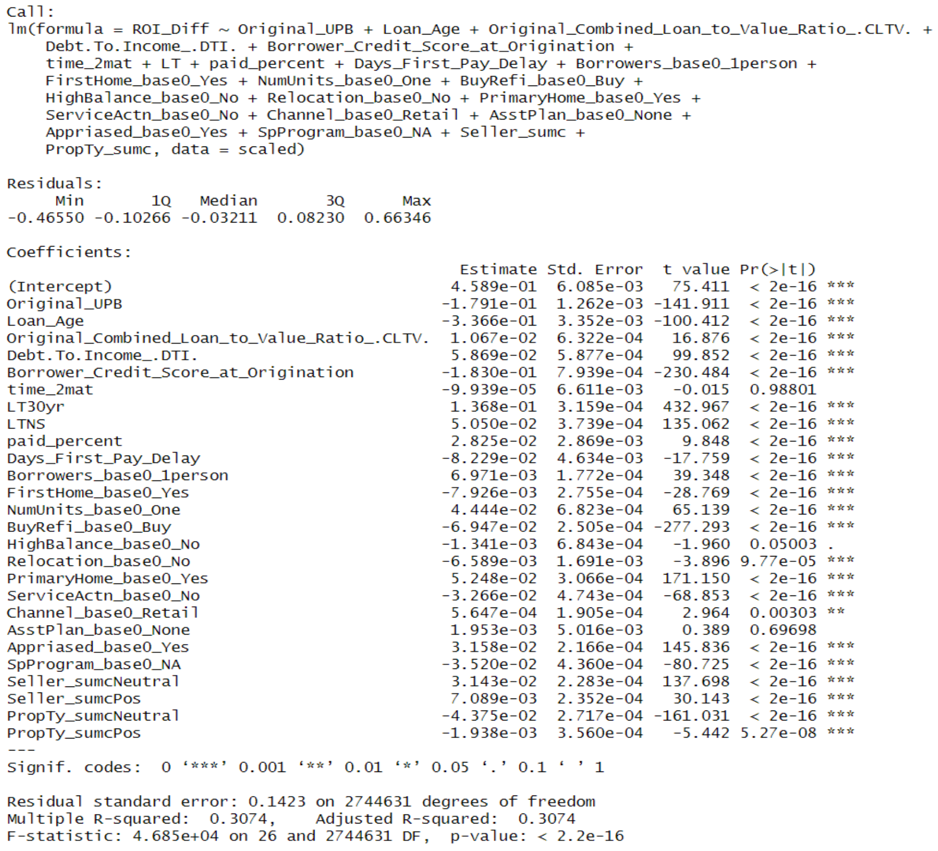
```
# State name contrasting and mutate
contrasts(df$Property_State) <- contr.sum(length(levels(df$Property_State))) # define contrasts as SUM or Deviati
on contrast

state_effect_sumc <- glm(df$Delinquent_base0_No ~ ., data = df$Property_State, family = 'binomial')
summary(state_effect_sumc)
```

```
##
## Call:
## glm(formula = df$Delinquent_base0_No ~ ., family = "binomial",
##     data = df$Property_State)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.2222  -0.1158  -0.1044  -0.1005   3.5595
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.31325    2.06792  -2.569   0.0102 *
## data1         0.89098    2.10131   0.424   0.6716
## data2         0.12236    2.07657   0.059   0.9530
## data3         0.22053    2.08089   0.106   0.9156
## data4         0.19275    2.07108   0.093   0.9258
## data5         0.31139    2.06899   0.151   0.8804
## data6        -0.05387    2.07268  -0.026   0.9793
## data7        -0.16947    2.08455  -0.081   0.9352
## data8         0.80606    2.10678   0.383   0.7020
## data9         0.37591    2.09123   0.180   0.8573
## data10        0.47184    2.06890   0.228   0.8196
## data11        0.02594    2.07108   0.013   0.9900
## data12       -8.25281  109.57353  -0.075   0.9400
## data13        0.80349    2.09132   0.384   0.7008
## data14       -0.01393    2.08456  -0.007   0.9947
## data15        0.10377    2.08250   0.050   0.9603
## data16       -0.06080    2.07135  -0.029   0.9766
## data17       -0.38189    2.07544  -0.184   0.8540
## data18       -0.24778    2.08906  -0.119   0.9056
## data19        0.34504    2.07691   0.166   0.8681
## data20        0.55226    2.07600   0.266   0.7902
## data21        0.35041    2.07424   0.169   0.8658
## data22        0.03864    2.07459   0.019   0.9851
## data23       -0.58023    2.14433  -0.271   0.7867
## data24        0.65392    2.06984   0.316   0.7521
```

Figure 9: Default Model

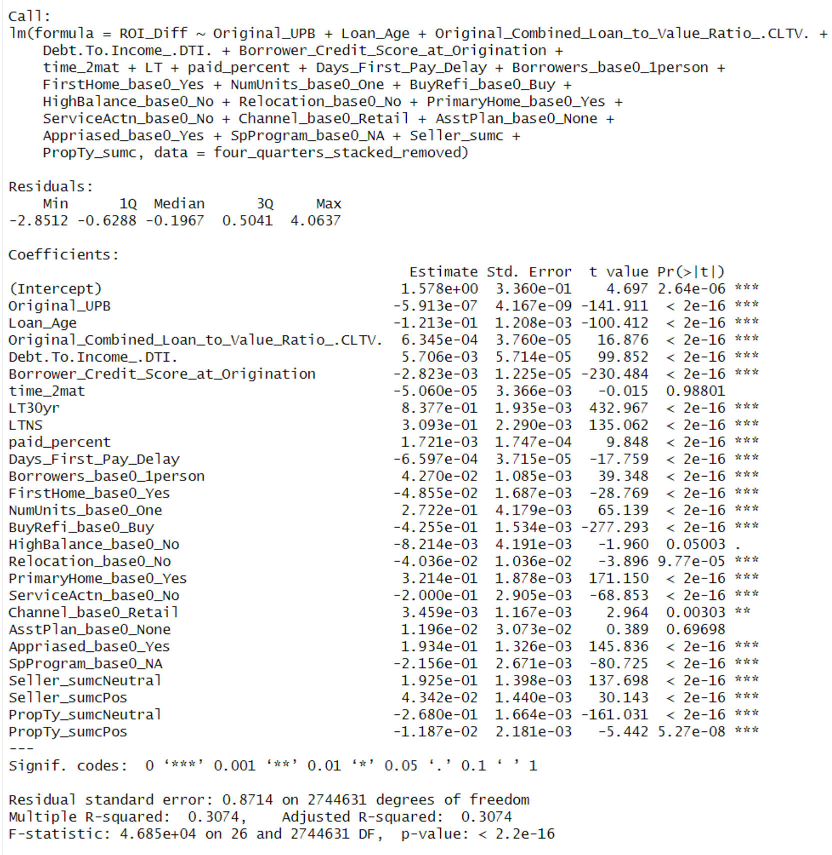
Figure 10: Model with Sum Contrast

Once the team cleaned the data and had chance to conduct some exploration on the attributes, a linear regression model was run using the clean data to forecast the difference between original interest rate and historical treasury rate data. The objective is to determine if the variation between the interest rate and the treasury rate can be approximated as a solitary factor that serves as a surrogate for delinquency. Both scaled and unscaled data were examined here to evaluate if data scaling can confound interpretations. We found that as shown in Figure 11 –12, the coefficients are different between scaled and unscaled models, but the influence directions and R squared values are same. Subsequent comparisons will use only scaled data (min-max scaling) for detailed results interpretations in all models.



```
Call:
lm(formula = ROI_Diff ~ Original_UPB + Loan_Age + Original_Combined_Loan_to_Value_Ratio_.CLTV. +
    Debt.To.Income_.DTI. + Borrower_Credit_Score_at_Origination +
    time_2mat + LT + paid_percent + Days_First_Pay_Delay + Borrowers_base0_1person +
    FirstHome_base0_Yes + NumUnits_base0_One + BuyRefi_base0_Buy +
    HighBalance_base0_No + Relocation_base0_No + PrimaryHome_base0_Yes +
    ServiceActn_base0_No + Channel_base0_Retail + AsstPlan_base0_None +
    Appriased_base0_Yes + SpProgram_base0_NA + Seller_sumc +
    PropTy_sumc, data = scaled)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.46550 -0.10266 -0.03211  0.08230  0.66346 

Coefficients:
                                               Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                                   4.589e-01  6.085e-03   75.411  < 2e-16 ***
Original_UPB                                 -1.791e-01  1.262e-03 -141.911  < 2e-16 ***
Loan_Age                                     -3.366e-01  3.352e-03 -100.412  < 2e-16 ***
Original_Combined_Loan_to_Value_Ratio_.CLTV.  1.067e-02  6.322e-04   16.876  < 2e-16 ***
Debt.To.Income_.DTI.                          5.869e-02  5.877e-04   99.852  < 2e-16 ***
Borrower_Credit_Score_at_Origination         -1.830e-01  7.939e-04 -230.484  < 2e-16 ***
time_2mat                                    -9.939e-05  6.611e-03   -0.015  0.98801    
LT30yr                                        1.368e-01  3.159e-04  432.967  < 2e-16 ***
LTNS                                          5.050e-02  3.739e-04  135.062  < 2e-16 ***
paid_percent                                  2.825e-02  2.869e-03    9.848  < 2e-16 ***
Days_First_Pay_Delay                         -8.229e-02  4.634e-03  -17.759  < 2e-16 ***
Borrowers_base0_1person                       6.971e-03  1.772e-04   39.348  < 2e-16 ***
FirstHome_base0_Yes                          -7.926e-03  2.755e-04  -28.769  < 2e-16 ***
NumUnits_base0_One                            4.444e-02  6.823e-04   65.139  < 2e-16 ***
BuyRefi_base0_Buy                            -6.947e-02  2.505e-04 -277.293  < 2e-16 ***
HighBalance_base0_No                         -1.341e-03  6.843e-04   -1.960  0.05003 .  
Relocation_base0_No                          -6.589e-03  1.691e-03   -3.896 9.77e-05 ***
PrimaryHome_base0_Yes                         5.248e-02  3.066e-04  171.150  < 2e-16 ***
ServiceActn_base0_No                         -3.266e-02  4.743e-04  -68.853  < 2e-16 ***
Channel_base0_Retail                          5.647e-04  1.905e-04    2.964  0.00303 ** 
AsstPlan_base0_None                           1.953e-03  5.016e-03    0.389  0.69698    
Appriased_base0_Yes                           3.158e-02  2.166e-04  145.836  < 2e-16 ***
SpProgram_base0_NA                           -3.520e-02  4.360e-04  -80.725  < 2e-16 ***
Seller_sumcNeutral                            3.143e-02  2.283e-04  137.698  < 2e-16 ***
Seller_sumcPos                                7.089e-03  2.352e-04   30.143  < 2e-16 ***
PropTy_sumcNeutral                           -4.375e-02  2.717e-04 -161.031  < 2e-16 ***
PropTy_sumcPos                               -1.938e-03  3.560e-04   -5.442 5.27e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1423 on 2744631 degrees of freedom
Multiple R-squared:  0.3074,	Adjusted R-squared:  0.3074 
F-statistic: 4.685e+04 on 26 and 2744631 DF,  p-value: < 2.2e-16
```

Figure 11: Scaled Linear Regression Model



```
Call:
lm(formula = ROI_Diff ~ Original_UPB + Loan_Age + Original_Combined_Loan_to_Value_Ratio_.CLTV. +
    Debt.To.Income_.DTI. + Borrower_Credit_Score_at_Origination +
    time_2mat + LT + paid_percent + Days_First_Pay_Delay + Borrowers_base0_1person +
    FirstHome_base0_Yes + NumUnits_base0_One + BuyRefi_base0_Buy +
    HighBalance_base0_No + Relocation_base0_No + PrimaryHome_base0_Yes +
    ServiceActn_base0_No + Channel_base0_Retail + AsstPlan_base0_None +
    Appriased_base0_Yes + SpProgram_base0_NA + Seller_sumc +
    PropTy_sumc, data = four_quarters_stacked_removed)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8512 -0.6288 -0.1967  0.5041  4.0637

Coefficients:
                                              Estimate Std. Error  t value Pr(>|t|)
(Intercept)                                  1.578e+00  3.360e-01    4.697 2.64e-06 ***
Original_UPB                                -5.913e-07  4.167e-09 -141.911  < 2e-16 ***
Loan_Age                                    -1.213e-01  1.208e-03 -100.412  < 2e-16 ***
Original_Combined_Loan_to_Value_Ratio_.CLTV.  6.345e-04  3.760e-05   16.876  < 2e-16 ***
Debt.To.Income_.DTI.                         5.706e-03  5.714e-05   99.852  < 2e-16 ***
Borrower_Credit_Score_at_Origination        -2.823e-03  1.225e-05 -230.484  < 2e-16 ***
time_2mat                                   -5.060e-05  3.366e-03   -0.015  0.98801
LT30yr                                       8.377e-01  1.935e-03  432.967  < 2e-16 ***
LTNS                                         3.093e-01  2.290e-03  135.062  < 2e-16 ***
paid_percent                                 1.721e-03  1.747e-04    9.848  < 2e-16 ***
Days_First_Pay_Delay                        -6.597e-04  3.715e-05  -17.759  < 2e-16 ***
Borrowers_base0_1person                      4.270e-02  1.085e-03   39.348  < 2e-16 ***
FirstHome_base0_Yes                         -4.855e-02  1.687e-03  -28.769  < 2e-16 ***
NumUnits_base0_One                           2.722e-01  4.179e-03   65.139  < 2e-16 ***
BuyRefi_base0_Buy                           -4.255e-01  1.534e-03 -277.293  < 2e-16 ***
HighBalance_base0_No                        -8.214e-03  4.191e-03   -1.960  0.05003 .
Relocation_base0_No                         -4.036e-02  1.036e-02   -3.896 9.77e-05 ***
PrimaryHome_base0_Yes                        3.214e-01  1.878e-03  171.150  < 2e-16 ***
ServiceActn_base0_No                        -2.000e-01  2.905e-03  -68.853  < 2e-16 ***
Channel_base0_Retail                         3.459e-03  1.167e-03    2.964  0.00303 **
AsstPlan_base0_None                          1.196e-02  3.073e-02    0.389  0.69698
Appriased_base0_Yes                          1.934e-01  1.326e-03  145.836  < 2e-16 ***
SpProgram_base0_NA                          -2.156e-01  2.671e-03  -80.725  < 2e-16 ***
Seller_sumcNeutral                           1.925e-01  1.398e-03  137.698  < 2e-16 ***
Seller_sumcPos                               4.342e-02  1.440e-03   30.143  < 2e-16 ***
PropTy_sumcNeutral                          -2.680e-01  1.664e-03 -161.031  < 2e-16 ***
PropTy_sumcPos                              -1.187e-02  2.181e-03   -5.442 5.27e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8714 on 2744631 degrees of freedom
Multiple R-squared:  0.3074,	Adjusted R-squared:  0.3074
F-statistic: 4.685e+04 on 26 and 2744631 DF,  p-value: < 2.2e-16
```

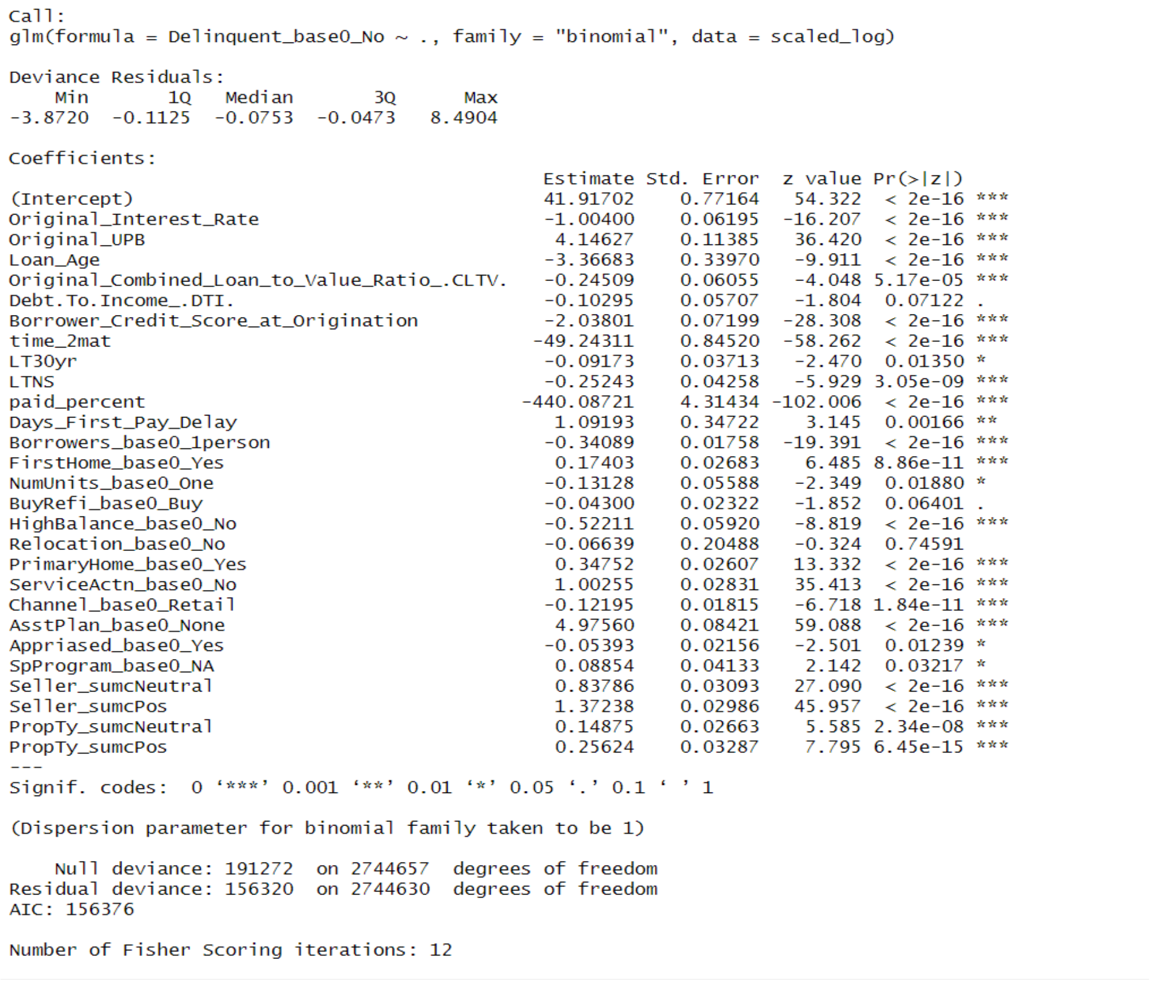
Figure 12: Unscaled Linear Regression Model

Upon examining the boxplots illustrating the distribution of the interest rate difference for the two delinquency cases, we concluded that it is not a suitable factor to include in our model as it is distributed identically for both cases. Thus, both of our additional external data inclusion plans for this work could not be carried out further due to a lack of significant model impact when examined in single factor regressions itself. A logistic regression model with cleaned and recoded attributes and delinquency as the response would be our best approach forward and scaled data will be used for all models from here on.

MODEL SELECTION & INTERPRETATION

All-Factors Logistic Regression Model

Based on what the team discovered in the exploratory and initial discoveries section, we first decided to use an all-factors logistic model for our project as seen in Figure 13 using recoded factors of interest and after removal of factors that did not show a significant impact in single-factor regressions (e.g., State, Zip-code, etc.). As earlier with linear regression, we see that the weights of coefficients are different, but the impact direction and AIC values are same in both the scaled and unscaled models.



```
Call:
glm(formula = Delinquent_base0_No ~ ., family = "binomial", data = scaled_log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8720  -0.1125  -0.0753  -0.0473   8.4904

Coefficients:
                                              Estimate Std. Error  z value Pr(>|z|)
(Intercept)                                   41.91702    0.77164   54.322  < 2e-16 ***
Original_Interest_Rate                        -1.00400    0.06195  -16.207  < 2e-16 ***
Original_UPB                                   4.14627    0.11385   36.420  < 2e-16 ***
Loan_Age                                      -3.36683    0.33970   -9.911  < 2e-16 ***
Original_Combined_Loan_to_Value_Ratio_.CLTV.  -0.24509    0.06055   -4.048 5.17e-05 ***
Debt.To.Income_.DTI.                          -0.10295    0.05707   -1.804  0.07122 .
Borrower_Credit_Score_at_Origination          -2.03801    0.07199  -28.308  < 2e-16 ***
time_2mat                                    -49.24311    0.84520  -58.262  < 2e-16 ***
LT30yr                                        -0.09173    0.03713   -2.470  0.01350 *
LTNS                                          -0.25243    0.04258   -5.929 3.05e-09 ***
paid_percent                                -440.08721    4.31434 -102.006  < 2e-16 ***
Days_First_Pay_Delay                           1.09193    0.34722    3.145  0.00166 **
Borrowers_base0_1person                       -0.34089    0.01758  -19.391  < 2e-16 ***
FirstHome_base0_Yes                            0.17403    0.02683    6.485 8.86e-11 ***
NumUnits_base0_One                            -0.13128    0.05588   -2.349  0.01880 *
BuyRefi_base0_Buy                             -0.04300    0.02322   -1.852  0.06401 .
HighBalance_base0_No                          -0.52211    0.05920   -8.819  < 2e-16 ***
Relocation_base0_No                           -0.06639    0.20488   -0.324  0.74591
PrimaryHome_base0_Yes                          0.34752    0.02607   13.332  < 2e-16 ***
ServiceActn_base0_No                           1.00255    0.02831   35.413  < 2e-16 ***
Channel_base0_Retail                          -0.12195    0.01815   -6.718 1.84e-11 ***
AsstPlan_base0_None                            4.97560    0.08421   59.088  < 2e-16 ***
Appriased_base0_Yes                           -0.05393    0.02156   -2.501  0.01239 *
SpProgram_base0_NA                             0.08854    0.04133    2.142  0.03217 *
Seller_sumcNeutral                             0.83786    0.03093   27.090  < 2e-16 ***
Seller_sumcPos                                 1.37238    0.02986   45.957  < 2e-16 ***
PropTy_sumcNeutral                             0.14875    0.02663    5.585 2.34e-08 ***
PropTy_sumcPos                                 0.25624    0.03287    7.795 6.45e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 191272  on 2744657  degrees of freedom
Residual deviance: 156320  on 2744630  degrees of freedom
AIC: 156376

Number of Fisher Scoring iterations: 12
```

Figure 13: Scaled Logistic Regression Model

We interpreted the results of the scaled all-factor logistic model first to understand the impact of the several factors. Firstly, it is easy to see that all factors (and sub-levels) have a significant impact on the model. Next, looking at a general overview of the coefficients and significance: The largest positive integer coefficients (other than intercept) are for Original unpaid balance (4.1), delay to first payment (1.1), Not having Service action (1.00), Not having assistance plan (4.9), and belonging to a sub-group of Sellers that have a positive significant influence in a single factor regression model (1.3). This can be interpreted as these attributes increasing the log-odds of default by the respective coefficient magnitude with other factors being the same. On the other hand, the largest negative and significant coefficients were for percent of paid balance (-441), time to loan maturity (-49), loan age (-3.3) etc. suggesting that an increase in these factors reduces the log odds of delinquency by the magnitude of the coefficients respectively, with all other attributed remaining the same. While these interpretations make logical sense in some cases (i.e., high credit score reducing odds of default), they are confounding in some cases like when time to loan maturity is higher, a lower chance of default is assessed.

Though we can interpret these influences on the delinquency outcome, these do not offer sufficiently nuanced separation to allow predictions if all factors have a significant influence. With these preliminary observations, it was decided to manually reduce the number of variables to only ones that were integer and influential or character variables (binary or recoded).

A manual, but straightforward method to conduct a significance test for other numeric variables such as LTV, original interest rates, LTV, and credit score average split (and distribution) with respect to delinquent and non-delinquent observations was to create boxplots and look at median values between the two response groups for confirmation. As indicated by Figure 14-17 there is no significant separator between delinquent and non-delinquent cases except for FICO credit score (irrespective of if the model is scaled or not, graphs for unscaled data not shown). While Credit score does have the largest separator between cases, it is still not drastic.

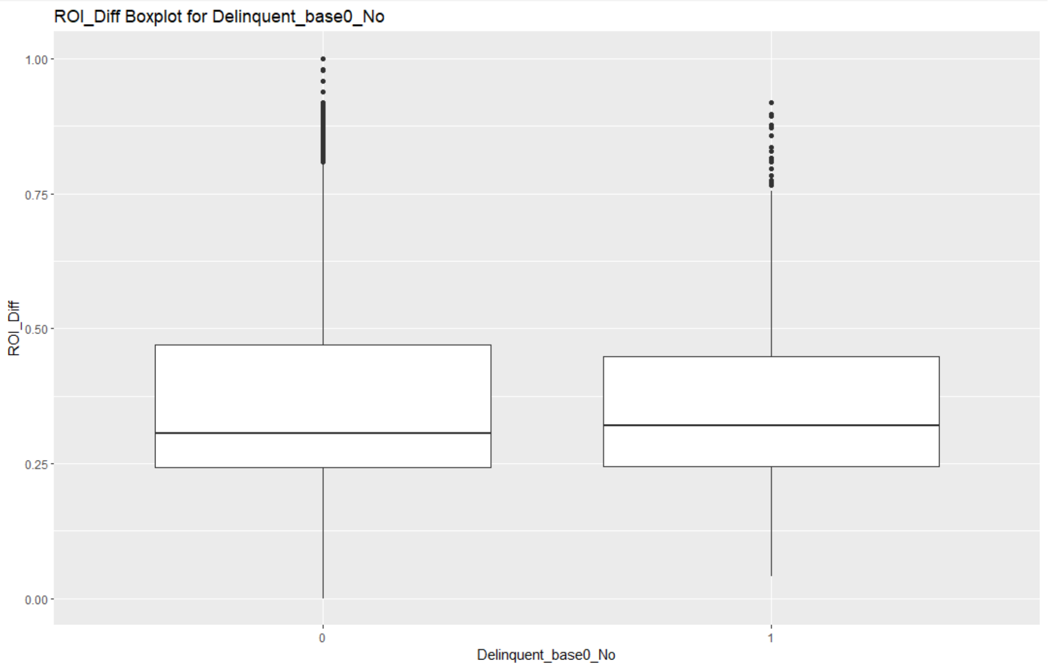


Figure 14: Scaled Treasury/Original Interest Rate Difference v. Delinquency

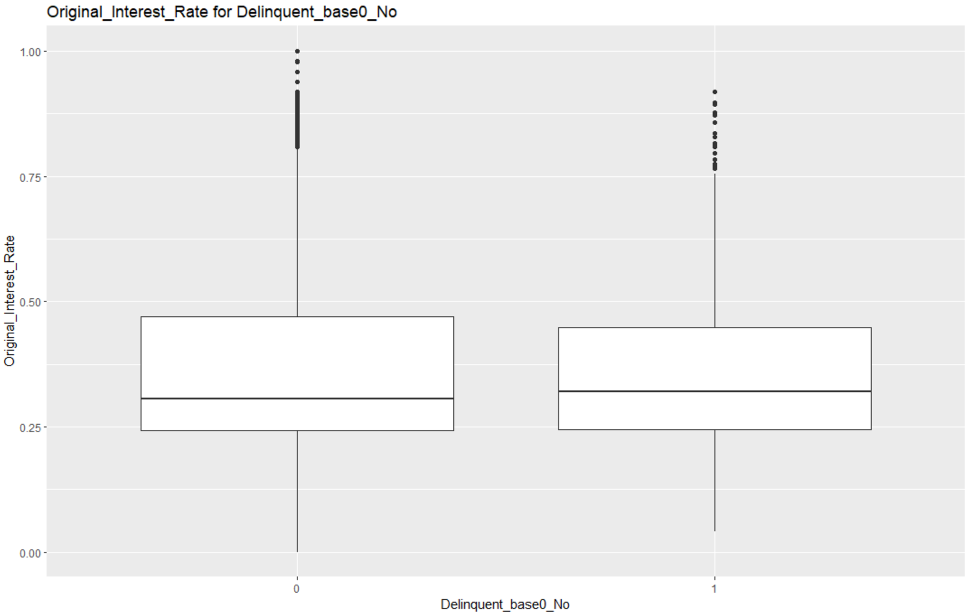


Figure 15 Scaled Original Interest Rate v. Delinquency

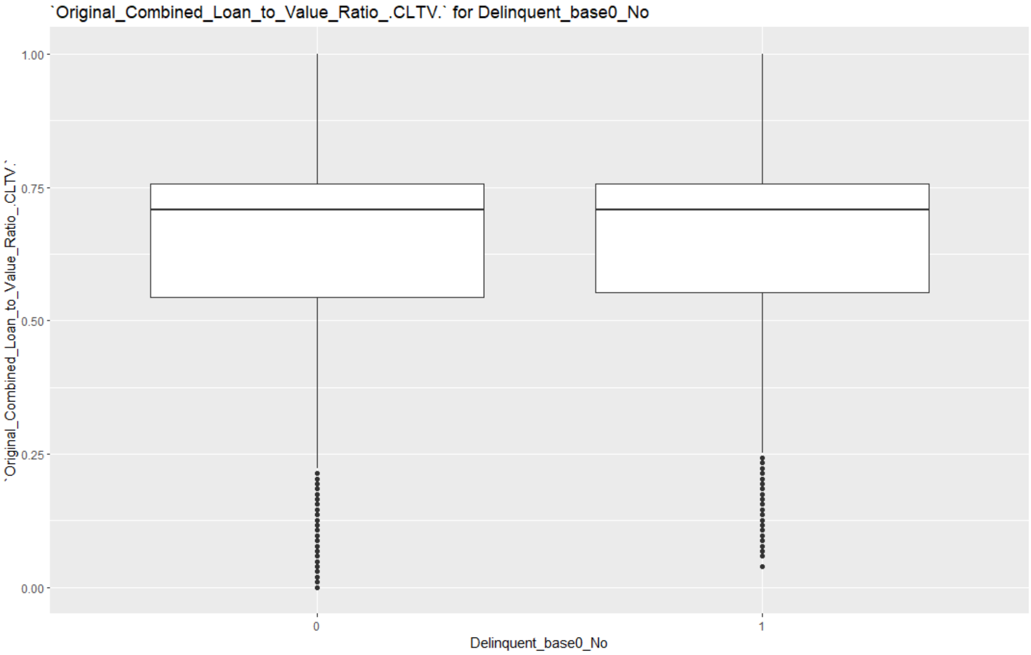


Figure 16: Scaled LTV v. Delinquency

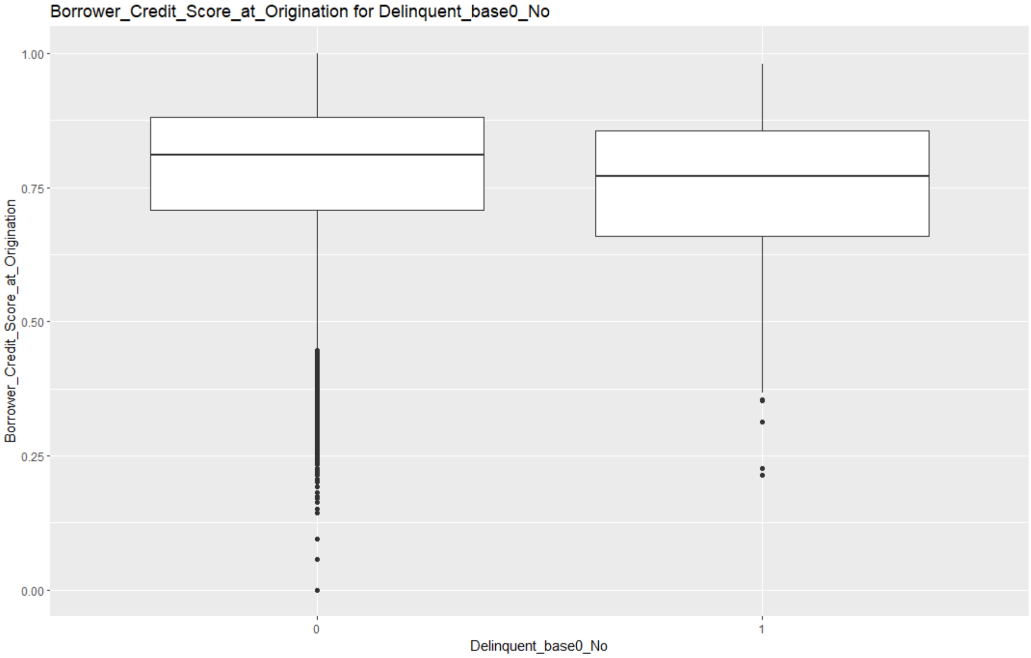
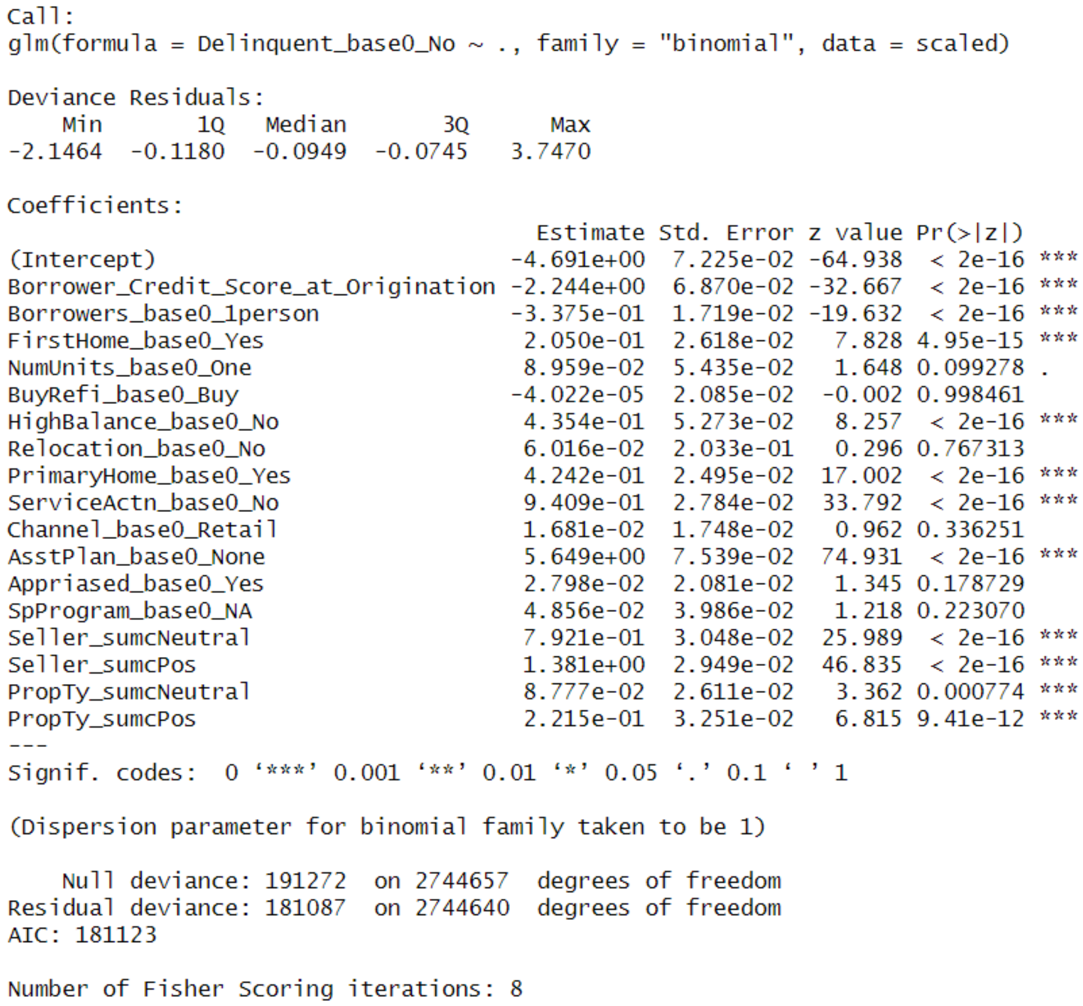


Figure 17 Scaled FICO Credit Score v. Delinquency

Thus, our initial hypothesis that high LTV and DTI play a significant role in prediction of delinquency is not convincingly proven by our analysis—plausibly due to an inherent bias in "approved loan" data that we have here from Fannie-Mae's database. Considering all loan applicants go through a rigorous vetting process with their lenders prior to approval, it is possible that the vetting process itself introduces bias in the data. The are often generally accepted criteria for loan approval among lenders. These criteria subsequently limit the variety of candidates seen in the loan data, as all approved candidates share a strong set of characteristics that meet the lending criteria. Nevertheless, we continued our journey toward selecting a refined model, as seen in the next section.

Logistic Regression with Selected Variables (FICO credit score + all other character/level variables)

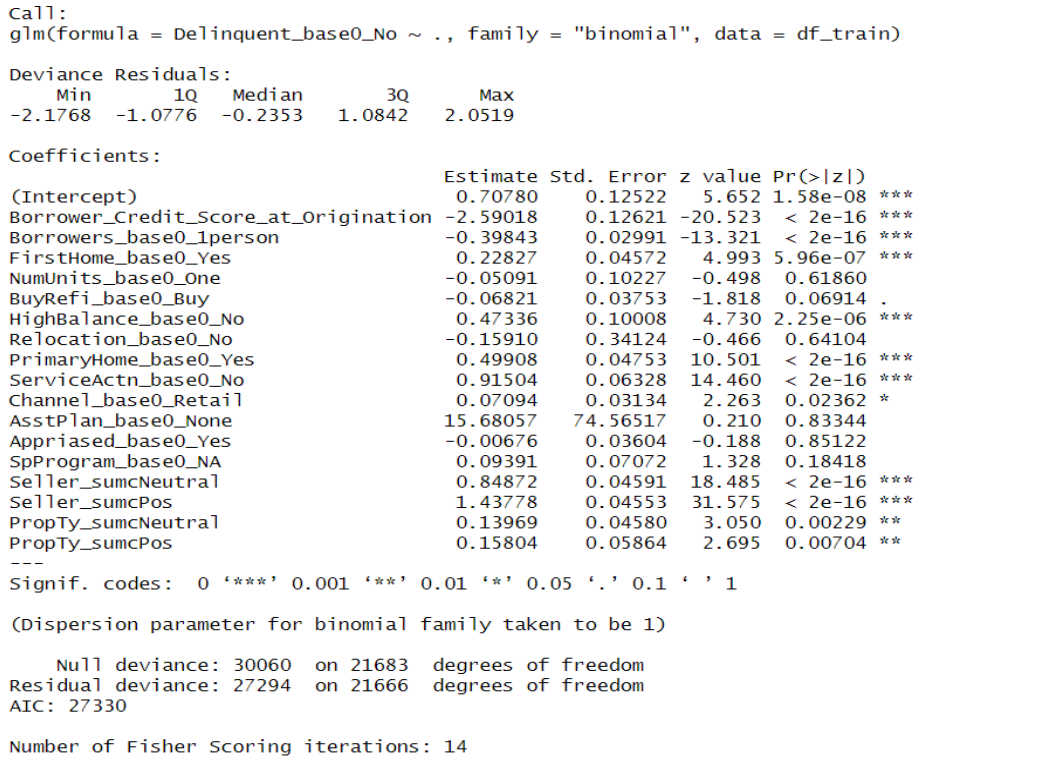
To provide a clearer predictive capability, we decided to reduce the number of attributes used in the model. Since the VIF scores would not be helpful, we explored the distribution of important integer attributes with respect to delinquency and chose to retain only ones that demonstrated clear difference in the distribution. The team pivoted and ran another logistic regression model with a selected group of variables – particularly FICO credit score and all other binary and recoded character/level variables. As expected, AIC value increased, which means the accuracy decreased as well. Please see Figure 18 for results.



```
Call:
glm(formula = Delinquent_base0_No ~ ., family = "binomial", data = scaled)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1464  -0.1180  -0.0949  -0.0745   3.7470

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -4.691e+00  7.225e-02 -64.938  < 2e-16 ***
Borrower_Credit_Score_at_Origination -2.244e+00  6.870e-02 -32.667  < 2e-16 ***
Borrowers_base0_1person              -3.375e-01  1.719e-02 -19.632  < 2e-16 ***
FirstHome_base0_Yes                   2.050e-01  2.618e-02   7.828 4.95e-15 ***
NumUnits_base0_One                    8.959e-02  5.435e-02   1.648 0.099278 .
BuyRefi_base0_Buy                    -4.022e-05  2.085e-02  -0.002 0.998461
HighBalance_base0_No                  4.354e-01  5.273e-02   8.257  < 2e-16 ***
Relocation_base0_No                   6.016e-02  2.033e-01   0.296 0.767313
PrimaryHome_base0_Yes                 4.242e-01  2.495e-02  17.002  < 2e-16 ***
ServiceActn_base0_No                  9.409e-01  2.784e-02  33.792  < 2e-16 ***
Channel_base0_Retail                  1.681e-02  1.748e-02   0.962 0.336251
AsstPlan_base0_None                   5.649e+00  7.539e-02  74.931  < 2e-16 ***
Appriased_base0_Yes                   2.798e-02  2.081e-02   1.345 0.178729
SpProgram_base0_NA                    4.856e-02  3.986e-02   1.218 0.223070
Seller_sumcNeutral                    7.921e-01  3.048e-02  25.989  < 2e-16 ***
Seller_sumcPos                        1.381e+00  2.949e-02  46.835  < 2e-16 ***
PropTy_sumcNeutral                    8.777e-02  2.611e-02   3.362 0.000774 ***
PropTy_sumcPos                        2.215e-01  3.251e-02   6.815 9.41e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 191272  on 2744657  degrees of freedom
Residual deviance: 181087  on 2744640  degrees of freedom
AIC: 181123

Number of Fisher Scoring iterations: 8
```

Figure 18: Scaled Logistic Regression with Selected Variables

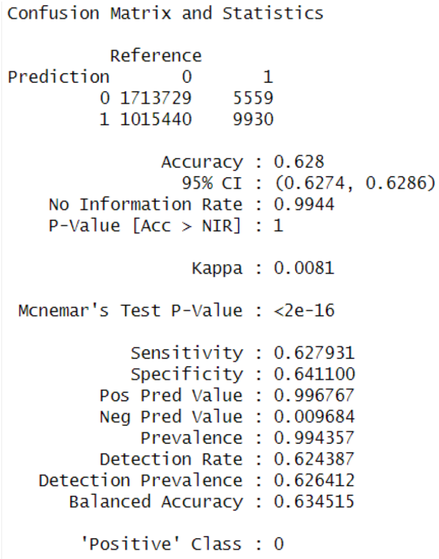
At this point, the team explored ways to adjust the model. However, considering around 99.5% of the data set can be categorized as non-delinquent, it poses a challenge to train an accurate model. The team used the existing 15,000 delinquency cases to create two new data frames—one with only delinquent data points, and a second with randomly selected, but the same number of non-delinquent cases. We ran 20 iterations in which each time we selected 70% of both delinquent and non-delinquent observations from the respective data frames and bound the two data frames together to train the model and used the remaining 30% for testing to ensure a balanced data with both cases included equally in the model training and evaluation. The iterations for both scaled and unscaled data resulted in a similar AIC value around 27,000 for both as seen in Figure 19.



```
Call:
glm(formula = Delinquent_base0_No ~ ., family = "binomial", data = df_train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.1768  -1.0776  -0.2353   1.0842   2.0519  

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)                         0.70780    0.12522   5.652 1.58e-08 ***
Borrower_Credit_Score_at_Origination -2.59018    0.12621 -20.523  < 2e-16 ***
Borrowers_base0_1person            -0.39843    0.02991 -13.321  < 2e-16 ***
FirstHome_base0_Yes                 0.22827    0.04572   4.993 5.96e-07 ***
NumUnits_base0_One                 -0.05091    0.10227  -0.498  0.61860    
BuyRefi_base0_Buy                  -0.06821    0.03753  -1.818  0.06914 .  
HighBalance_base0_No                0.47336    0.10008   4.730 2.25e-06 ***
Relocation_base0_No                -0.15910    0.34124  -0.466  0.64104    
PrimaryHome_base0_Yes               0.49908    0.04753  10.501  < 2e-16 ***
ServiceActn_base0_No                0.91504    0.06328  14.460  < 2e-16 ***
Channel_base0_Retail                0.07094    0.03134   2.263  0.02362 *  
AsstPlan_base0_None                15.68057   74.56517   0.210  0.83344    
Appriased_base0_Yes                -0.00676    0.03604  -0.188  0.85122    
SpProgram_base0_NA                  0.09391    0.07072   1.328  0.18418    
Seller_sumcNeutral                  0.84872    0.04591  18.485  < 2e-16 ***
Seller_sumcPos                      1.43778    0.04553  31.575  < 2e-16 ***
PropTy_sumcNeutral                  0.13969    0.04580   3.050  0.00229 ** 
PropTy_sumcPos                      0.15804    0.05864   2.695  0.00704 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30060  on 21683  degrees of freedom
Residual deviance: 27294  on 21666  degrees of freedom
AIC: 27330

Number of Fisher Scoring iterations: 14
```

Figure 19: Scaled Logistic Regression with Selected Variables – Adjusted

When applying both models on entire dataset to see the forecast accuracy, the unscaled model is 62.74% accurate and scaled model can forecast 62.8% of the labels correctly – Figure 20. The scaled and unscaled model look similar, but due to the benefits inherent to scaling methods, the team considered the marginal improvement in the scaled model and will use it to interpret final project results.



```
Confusion Matrix and Statistics

          Reference
Prediction       0       1
         0 1713729    5559
         1 1015440    9930

               Accuracy : 0.628
                 95% CI : (0.6274, 0.6286)
    No Information Rate : 0.9944
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0081

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.627931
            Specificity : 0.641100
         Pos Pred Value : 0.996767
         Neg Pred Value : 0.009684
             Prevalence : 0.994357
         Detection Rate : 0.624387
   Detection Prevalence : 0.626412
      Balanced Accuracy : 0.634515

       'Positive' Class : 0
```

Figure 20: Confusion Matrix for Scaled Final Model

The accuracy of 62.86% can be evaluated as either satisfactory or unsatisfactory, depending on the current cost of delinquency and prevailing market prices. Incorporating additional data may improve the precision of the models. Further, this accuracy is based on using a split of 0.5 to split the predictions into delinquent or non-delinquent cases. Raising or lowering this cut-off threshold could be another method to improve the accuracy of our predictions. The accuracy of 0.25 split is 97.21% and a 0.75 split is 14.17%.

After analyzing the coefficient values of the scaled model, we found that among the statistically significant variables, the borrower's credit score is the most crucial factor that affects the probability of delinquency. It is not surprising that as the credit score increases, the likelihood of delinquency decreases. Another important variable is the number of borrowers, where a single borrower is associated with a lower chance of delinquency. Additionally, the seller of the mortgage is also a significant factor, with specific sellers that do not have a significant negative influence in single factor regression having a higher chance of delinquency. A similar but much smaller impact was observed for the type of property that the loan was used to finance. The fact that the loan was for a primary home or was for an account that did not have a Servicing activity indicator or did not have a high balance were all associated with a smaller but significant increase in likelihood of delinquency. Furthermore, if the property was not the borrower's primary residence or their first home, the chances of delinquency also increase slightly but significantly. Most of these factors follow a logical pattern of expected impacts when significant, and when they are indeed confounding, as in the case of default when the balance was not high, could indicate more a 'by-stander' chance impact rather than a primary impact. Thus, the high balance value as a factor is not a representative indicator of predictability.

CONCLUSION

Through our analysis, we developed a logistic regression model that can effectively predict a loan delinquency with about 63% accuracy. Among the key indicators, the model shows that credit score is a primary factor when predicting delinquency. However, the team felt that credit score is not the only significant indicator. The other indicators in our hypothesis, based on common knowledge about loans and repayments from daily life experiences: debt-to-income (DTI) ratio, loan to value (LTV), and history of repayment were simply not seen as significant as we thought by the model. However, other attributes like the mortgage seller, the type of property, and the use of the property as primary residence or not, all have a significant impact on the delinquency rate despite these likelihoods being small in magnitude. This aspect could not have been predicted without the use of a granular modeling approach like the one undertaken for this project. Overall, it is thought that risk spread is indeed a game of small chances represented by these character/level variables in addition to the credit score—if not this would not be a commercially viable business venture.

Thus, we conclude that our modeling efforts have proved an existing and conceivable hypothesis of delinquency prediction based on consumer credit score, and more importantly have uncovered additional non-traditional indicators of delinquency that were not hypothesized by us, nor can be assumed to be common knowledge in the field. We believe that this modeling effort will have a significant impact on the current CTR decisions, especially if risk buyers want to change the classification threshold to better suit the magnitude of Type I and Type II errors, they are comfortable dealing within the risk assumption process.

Furthermore, the teams' exploratory efforts also helped identify some bias in the data that may not have been considered at the project's inception. Hence, future iterations of the data should consider applications data prior to approval rather than approved loan data to see if a more accurate model can be derived using less biased data.

WORKS CITED

Fannie Mae. (2023). *Capital Markets Activities & News | Fannie Mae*. Fannie Mae. Retrieved March 25, 2023, from https://capitalmarkets.fanniemae.com/media/22751/display

Glowacki, Jonathan. “CRT 101: Everything You Need to Know about Freddie Mac and Fannie Mae Credit Risk Transfer.” *Www.milliman.com*, Oct. 2021, www.milliman.com/en/insight/crt-101-everything-you-need-to-know-about-freddie-mac-and-fannie-mae-credit-risk-transfer. Accessed 3 Apr. 2023.

Goodman, Laurie, et al. “How Debt Burden Affects FHA Mortgage Repayment, in Six Charts.” *Urban Institute*, 12 Sept. 2019, www.urban.org/urban-wire/how-debt-burden-affects-fha-mortgage-repayment-six-charts. Accessed 3 Apr. 2023.

Harrison, David M., et al. “Do Riskier Borrowers Borrow More?” *Real Estate Economics*, vol. 32, no. 3, Sept. 2004, pp. 385–411, https://doi.org/10.1111/j.1080-8620.2004.00096.x. Accessed 3 Apr. 2023.

Kagan, Julia. “Mortgage.” *Investopedia*, 25 Feb. 2021, www.investopedia.com/terms/m/mortgage.asp.

“Topic: Mortgage Industry in the U.S.” *Statista*, Statista Research Department, 21 Nov. 2022, https://www.statista.com/topics/1685/mortgage-industry-of-the-united-states/#editorsPicks.

DATA SOURCE

Teams Channel Link for Data & Final Code *(Copy Paste URL in Browser)*:
https://gtvault.sharepoint.com/:f:/s/MGT6203ULSBH/EnSANW5qy6tCtxWj1Q8gNboB9Y0H_QcVhq5o2vywqMz1YQ?e=u8xPRj