Modelling Default Rates of Single-Family Mortgage Loans with Multinomial Logistic Regression

Chau, Jessica 998286554
Chen, Tianxiao 1001126890
Veitch, David 1004657917
Wang, Shirly 998695185

University of Toronto

Tuesday April 2, 2019

1 Introduction

Canada Pension Plan Investment Board (CPPIB) manages over C$368 billion in investment assets. Portfolios of residential single-family mortgages are one of the largest investments that CPPIB maintains, and as a result, it is important to develop a strong understanding of the underlying dynamics of these assets. CPPIB specifically wants to understand the factors that affect mortgage defaults. The primary goal of this project is to create a model that can predict the likelihood that a mortgage will default in a specified time period. For the purpose of this project, we defined a default as a mortgage experiencing a 90-day delinquency.

There are additional restrictions to the model as requested by CPPIB:

- The model cannot be a black-box model (e.g. random forests or neural networks)
- The model should show the magnitude of covariates' impacts on the score
- The model should include macroeconomic factors

2 Data

Loan-level origination and performance data was collected from Freddie Mac which includes data from 1999 to 2017 [1]. This dataset includes information from the origination of the mortgage such as the borrower's credit score, whether the borrower is a first time home buyer, origination interest rate, and more. The full Single-Family Loan-Level Dataset contains over 26.6 million fixed-rate mortgages. In order to process the data and run models within a reasonable timeframe, we used a sample dataset, provided also by Freddie Mac, which contains simple random samples of 50,000 loans selected from each origination year. In addition, we only considered loans that:

- Have a 30-year term
- Only prepaid, defaulted, or performed during any given period
- Contain no missing values with respect to the covariates we selected

After imposing these restrictions, 769,228 loans remained in our dataset.

In addition to loan-level data, we also collected data on several macroeconomic factors and real estate indicators that could potentially affect default risk. Data considered are summarized in the following table:

Monthly Unemployment Rates	Real and Nominal GDP
OECD Leading Indicator	1yr, 5yr, 10yr and 30yr Treasury Yields
State Level Home Prices	S&P 500 Returns
Consumer Price Index	Industrial Production Index
PMI	Average Disposable Income

[1] Freddie Mac. Single Family Loan-Level Dataset, 2019. Data retrieved from http://www.freddiemac.com/research/datasets/ .

Origination data, monthly performance data, and external factors were combined to a single dataframe, where each row contains the quarterly performance of a loan, loan-level origination data, and corresponding external factors at the begin of the quarter. Additionally, we considered the differences between certain macroeconomic variables at origination and in the current quarter. We used stratified sampling on default performance to obtain a smaller set of loans to decrease compute time.

3 Exploratory Data Analysis

To narrow down the list of covariates to be used in the model, we plotted all covariates and evaluated their impact on default and prepayment. For time series data (e.g. housing prices, Treasury rates, etc.), we selected the value of the series at mortgage origination, the current quarter, and also considered the difference between origination date and the current quarter date. When applicable, we used year-over-year change in values to ensure stationarity. Since this analysis was only used for an initial feature selection to rule out low impact factors, we did not consider using statistical tests.

If the distribution of a covariate differed markedly from loans that defaulted to loans that did not default, or loans that were prepaid to loans that did not, it was added to the model. For example, in Figure 1, the current 1yr Treasury rate, 1yr Treasury rate at origination, and change in 1yr Treasury rate (between origination and the current quarter), were assessed for their impacts on default. It appears that more loans tend to default (blue) in quarters when the current 1yr Treasury rate is high, and when the 1yr Treasury rate decreases from origination. Thses distributions are markedly different from those of loans that did not default.

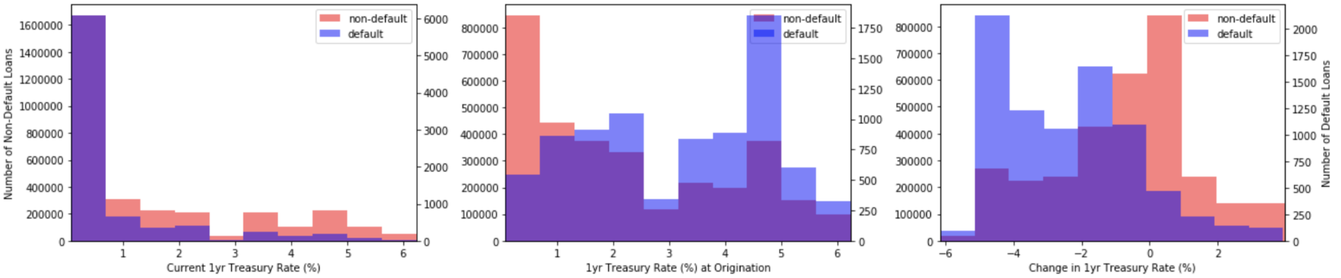


Figure 1: 1yr Treasury Rates

Figure 2 shows an example of a covariate that we did not consider in the model. Year-over-year (YoY) change in industrial production did not appear to have a significant impact on the probability of default.

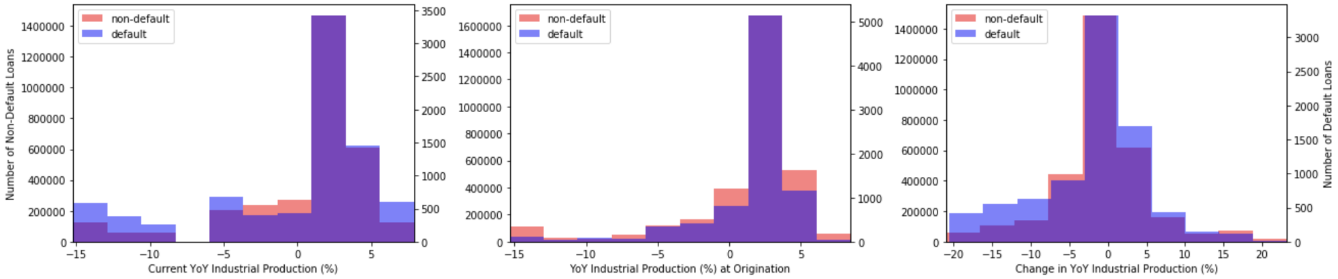


Figure 2: YoY % Change in Industrial Production

After selecting the relevant variables for the model, a correlation matrix was created. Highly correlated variables were removed to avoid multicollinearity. Multicollinearity can make the model unstable, in that a slightly different dataset can produce regression coefficients with markedly different magnitudes, or even different signs. Figure 3 shows how the 1yr, 5yr, 10yr and 30yr Treasury rates are highly correlated.

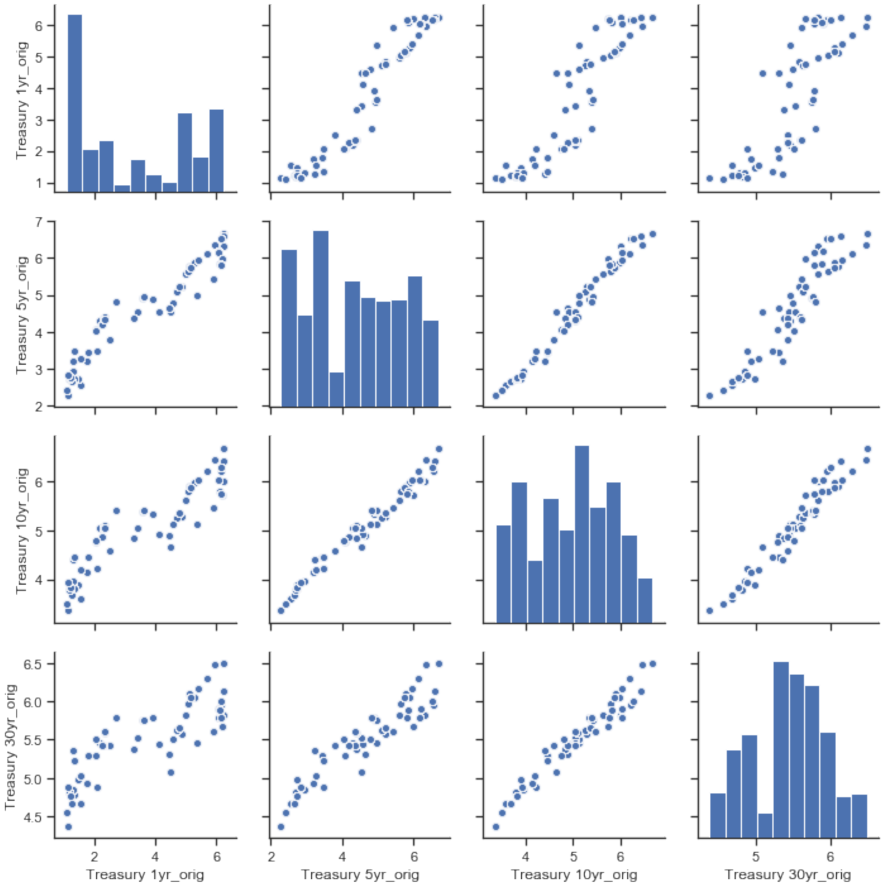


Figure 3: Correlation Matrix of 1yr, 5yr, 10yr, and 30yr Treasury Rates

It should also be noted that loan age was included as a covariate. Through our exploratory analysis we found that default and prepayment rates tend to change with a loan's age (e.g. recently originated loans are unlikely to default or prepay).

The list of covariates we included are shown in the table below:

Variable	Explanation
loan age	Age of a loan (capped at 10yrs)
credit score	The borrowers credit score at origination
curr upb	The mortgage's current unpaid balance
upb_prev	The mortgage's unpaid balance in the previous quarter
orig interest rate	The mortgage's interest rate at origination
channel	Whether a broker or correspondent was involved in the origination
num of borrowers	The number of borrower(s) who are obligated to repay the mortgage
loan purpose	The purpose of the mortgage loan
dti	Original debt-to-income ratio
ltv	Loan-to-value ratio in the current quarter
prev ltv	Loan-to-value ratio in the previous quarter
Housing Starts	The number of US new privately owned housing units started during the period
Building Permits	The number of permits that have been issued for new private housing construction
New Home Sales	Sales of newly constructed homes during the current period
Existing Home Sales	Sales of existing homes during the current period
OECD Leading Indicator YoY	Year-over-year change in OECD Leading Indicator in the current quarter
State Unemployment Rate	State Unemployment rate in the current quarter for the mortgage's state
Real GDP YoY	Year-over-year change in real GDP in the current year

SP500 YoY	Year-over-year change in S&P500 index in the current quarter
Treasury 1yr	1yr Treasury rate in the current quarter
Treasury 30yr	30yr Treasury rate in the current quarter
30yr Fixed Rate Mortgage Average	Average 30yr fixed rate mortgage rate in the current quarter
ZillowHouseValue	The state-level average house price for the current quarter
ZillowHouseChg	Current ZillowHouseValue divided by origination ZillowHouseValue

4 Modelling

4.1 Model Considerations

The desired output of the analysis was a model with predictive power and interpretability. With this in mind, machine learning techniques (e.g. neural nets and random forests) were not considered for their low interpretability.

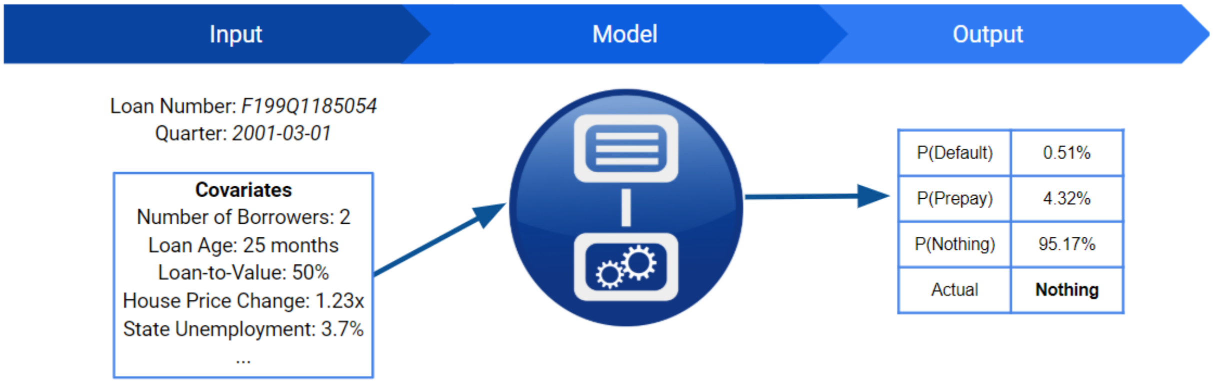
A generalized linear model (GLM) was chosen because of the high interpretability of this class of models. Specifically, these models produce coefficients, and their respective confidence intervals, which can help answer the question of what factors actually drive mortgage defaults and prepayments.

4.2 Train-Test Split

We implemented stratified sampling on default events to create a training set consisting of 70% of loans (i.e. the training set contained 70% of the total defaulted loans). The rest of the loans were set aside as a test set.

4.3 Model Selection

A multinomial logistic regression model was ultimately selected and fit to the data. This model can be used to determine the probability that a given mortgage m at time t will default, prepay, or experience neither of these events. Using a multinomial logistic regression model, as opposed to a plain logistic regression model, was necessary to account for the fact that at any given time a mortgage could either default, prepay, or have nothing occur. It was important to model these multiple events because for any given month, a mortgage having a higher probability of prepayment will likely, by virtue of the fact that the probabilities of all events must sum to one, have a lower probability of default. Using the model output we were then able to, for any given quarter, predict the expected number of mortgages that would default or prepay as shown in Figure 4.



Loan Number: *F199Q1185054*
Quarter: *2001-03-01*

Covariates
Number of Borrowers: 2
Loan Age: 25 months
Loan-to-Value: 50%
House Price Change: 1.23x
State Unemployment: 3.7%
...

P(Default)	0.51%
P(Prepay)	4.32%
P(Nothing)	95.17%
Actual	**Nothing**

Figure 4: Illustration of Model

This model seeks to minimize the following loss function:

$$\begin{aligned}\text{Loss} &= -(\text{Log-Likelihood}) + \text{Regularization Penalty}\\ &= -(\text{Log-Likelihood}) + \sum_{k=1}^{K} \alpha_k|\beta_k| + \nu_k|\gamma_k|\end{aligned}$$

$$\text{Log-Likelihood} = \sum_{m=1}^{M}\sum_{t=1}^{T_m} \mathbf{1}_{Y_{m;t}=\text{Def}} \log(p(Y_{m;t}=\text{Def})) + \mathbf{1}_{Y_{m;t}=\text{Pre}} \log(p(Y_{m;t}=\text{Pre})) + \mathbf{1}_{Y_{m;t}=\text{Nothing}} \log(p(Y_{m;t}=\text{Nothing}))$$

- $\mathbf{1}_{Y_{m;t}=\text{Event}}$ is an indicator function which is equal to 1 if mortgage m experiences the event (default, prepay, or nothing) at time t, and 0 otherwise
- α_k and ν_k, $k \in \{1, \ldots, K\}$, are scalars which represent the regularization penalty applied to the absolute value of the β_k and γ_k parameters respectively
- T_m represents the total number of periods data exists for loan m

The model minimizes this loss function by learning the coefficients β and γ:

$$\begin{aligned}\ln \frac{\Pr(Y_{m;t}=\text{Def})}{\Pr(Y_{m;t}=\text{Nothing})} &= \beta_0 + \beta_1 x_{m;t;1} + \cdots + \beta_K x_{m;t;K}\\ \ln \frac{\Pr(Y_{m;t}=\text{Pre})}{\Pr(Y_{m;t}=\text{Nothing})} &= \gamma_0 + \gamma_1 x_{m;t;1} + \cdots + \gamma_K x_{m;t;K}\\ \Pr(Y_{m;t}=\text{Def}) &= \frac{e^{\beta_0+\beta_1 x_{m;t;1}+\cdots+\beta_K x_{m;t;K}}}{1+e^{\beta_0+\beta_1 x_{m;t;1}+\cdots+\beta_K x_{m;t;K}}+e^{\gamma_0+\gamma_1 x_{m;t;1}+\cdots+\gamma_K x_{m;t;K}}}\\ \Pr(Y_{m;t}=\text{Pre}) &= \frac{e^{\gamma_0+\gamma_1 x_{m;t;1}+\cdots+\gamma_K x_{m;t;K}}}{1+e^{\beta_0+\beta_1 x_{m;t;1}+\cdots+\beta_K x_{m;t;K}}+e^{\gamma_0+\gamma_1 x_{m;t;1}+\cdots+\gamma_K x_{m;t;K}}}\\ \Pr(Y_{m;t}=\text{Nothing}) &= 1 - \Pr(Y_{m;t}=\text{Def}) - \Pr(Y_{m;t}=\text{Pre})\end{aligned}$$

4.4 Model Output

Figure 5 shows the parameters of the model after it was fit on the training set. We implemented a standard normalization on all covariates, except the intercept, so that the magnitude of coefficients gives us a sense of feature importance.

4.5 Model Validation

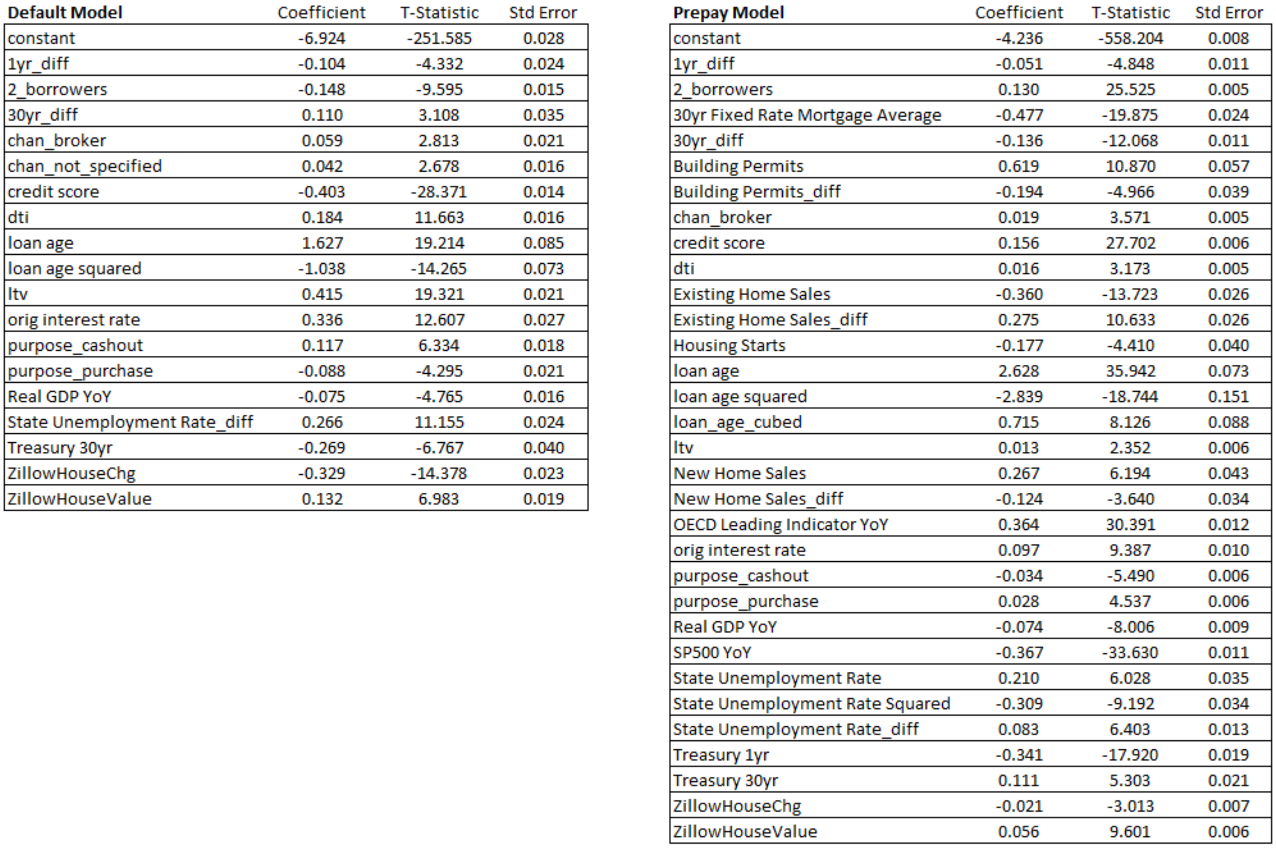
Validating the model on the test set we are able to see how predicted default and prepayment rates and amounts compare to reality in Figure 6. One numerical assessment of the model that can be made in this regard is the mean absolute difference between the expected and actual default rate (the main variable of interest) across all quarterly periods. On the test set this mean difference is .08%. As can be seen from Figure 6 the time periods where the difference is largest are 2002-2004 and 2009-2011.

Figure 7 shows the probabilities the model assigns to mortgages in a given period, and how these probabilities differ based on if the mortgage experienced the event in question or not. Clearly the model tends to assign higher default probabilities to mortgages that actually default, showing the model does have some ability to discriminate between mortgages that are likely and unlikely to default.

Finally, we can see in Figure 8 actual versus expected default rates by mortgage origination year. The model tends to perform better on certain origination years, potentially suggesting the dynamics of what drives defaults differ by origination year.

4.6 Regularization and Variable Selection

We opted to use L_1 regularization as this form of regularization encourages sparsity (i.e. coefficients in the model to become exactly 0) which helped discard covariates that lacked predictive power. We found in fitting the model



Default Model	Coefficient	T-Statistic	Std Error
constant	-6.924	-251.585	0.028
1yr_diff	-0.104	-4.332	0.024
2_borrowers	-0.148	-9.595	0.015
30yr_diff	0.110	3.108	0.035
chan_broker	0.059	2.813	0.021
chan_not_specified	0.042	2.678	0.016
credit score	-0.403	-28.371	0.014
dti	0.184	11.663	0.016
loan age	1.627	19.214	0.085
loan age squared	-1.038	-14.265	0.073
ltv	0.415	19.321	0.021
orig interest rate	0.336	12.607	0.027
purpose_cashout	0.117	6.334	0.018
purpose_purchase	-0.088	-4.295	0.021
Real GDP YoY	-0.075	-4.765	0.016
State Unemployment Rate_diff	0.266	11.155	0.024
Treasury 30yr	-0.269	-6.767	0.040
ZillowHouseChg	-0.329	-14.378	0.023
ZillowHouseValue	0.132	6.983	0.019

Prepay Model	Coefficient	T-Statistic	Std Error
constant	-4.236	-558.204	0.008
1yr_diff	-0.051	-4.848	0.011
2_borrowers	0.130	25.525	0.005
30yr Fixed Rate Mortgage Average	-0.477	-19.875	0.024
30yr_diff	-0.136	-12.068	0.011
Building Permits	0.619	10.870	0.057
Building Permits_diff	-0.194	-4.966	0.039
chan_broker	0.019	3.571	0.005
credit score	0.156	27.702	0.006
dti	0.016	3.173	0.005
Existing Home Sales	-0.360	-13.723	0.026
Existing Home Sales_diff	0.275	10.633	0.026
Housing Starts	-0.177	-4.410	0.040
loan age	2.628	35.942	0.073
loan age squared	-2.839	-18.744	0.151
loan_age_cubed	0.715	8.126	0.088
ltv	0.013	2.352	0.006
New Home Sales	0.267	6.194	0.043
New Home Sales_diff	-0.124	-3.640	0.034
OECD Leading Indicator YoY	0.364	30.391	0.012
orig interest rate	0.097	9.387	0.010
purpose_cashout	-0.034	-5.490	0.006
purpose_purchase	0.028	4.537	0.006
Real GDP YoY	-0.074	-8.006	0.009
SP500 YoY	-0.367	-33.630	0.011
State Unemployment Rate	0.210	6.028	0.035
State Unemployment Rate Squared	-0.309	-9.192	0.034
State Unemployment Rate_diff	0.083	6.403	0.013
Treasury 1yr	-0.341	-17.920	0.019
Treasury 30yr	0.111	5.303	0.021
ZillowHouseChg	-0.021	-3.013	0.007
ZillowHouseValue	0.056	9.601	0.006

Figure 5: Model Parameters

that the level of α and ν that were used did not have much of an effect on performance, except for very large levels of α and ν which forced all covariates to become 0 and dramatically reduced performance.

To simplify the model and decrease the total number of variables included, a two step model fitting process was used. In the first step the model was fit with all of the variables, and any variables without statistical significance (t-values $< 2 \Rightarrow$ p-values > 0.05) were thrown out. In the second step the model was fit again with regularization (to help with generalization on the test set).

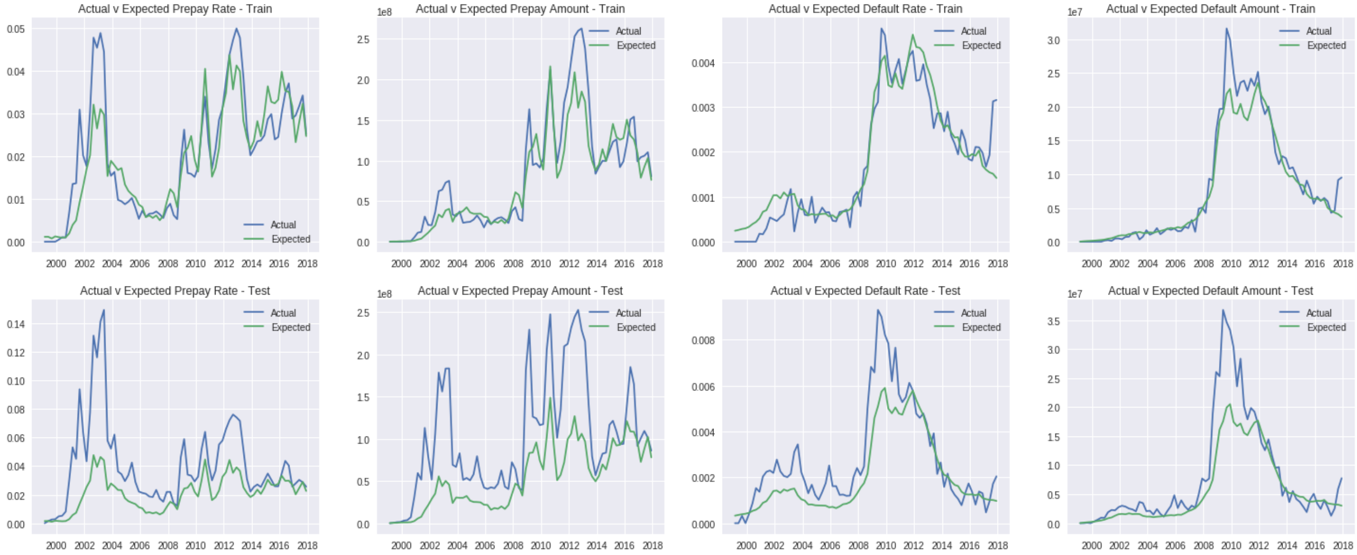


Figure 6: Actual Versus Expected Default and Prepay Rates Over Time

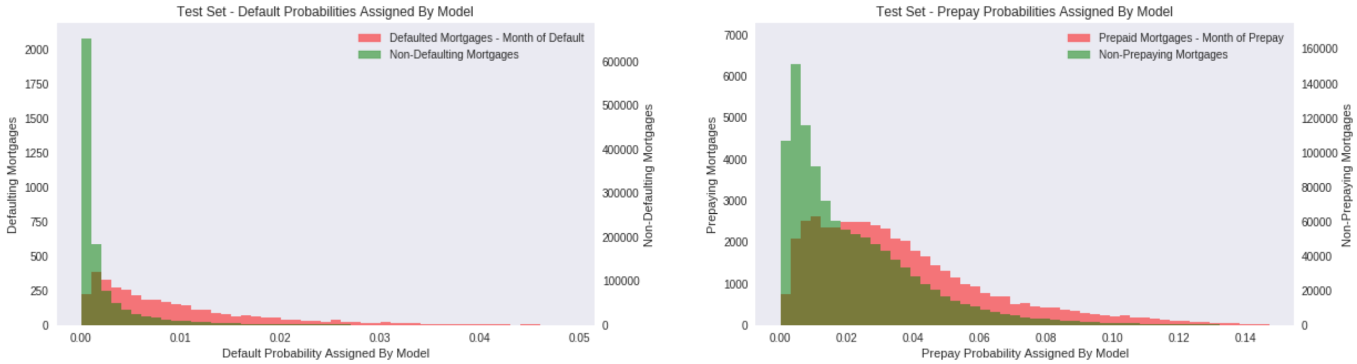


Figure 7: Probabilities Assigned By Model Versus Outcome

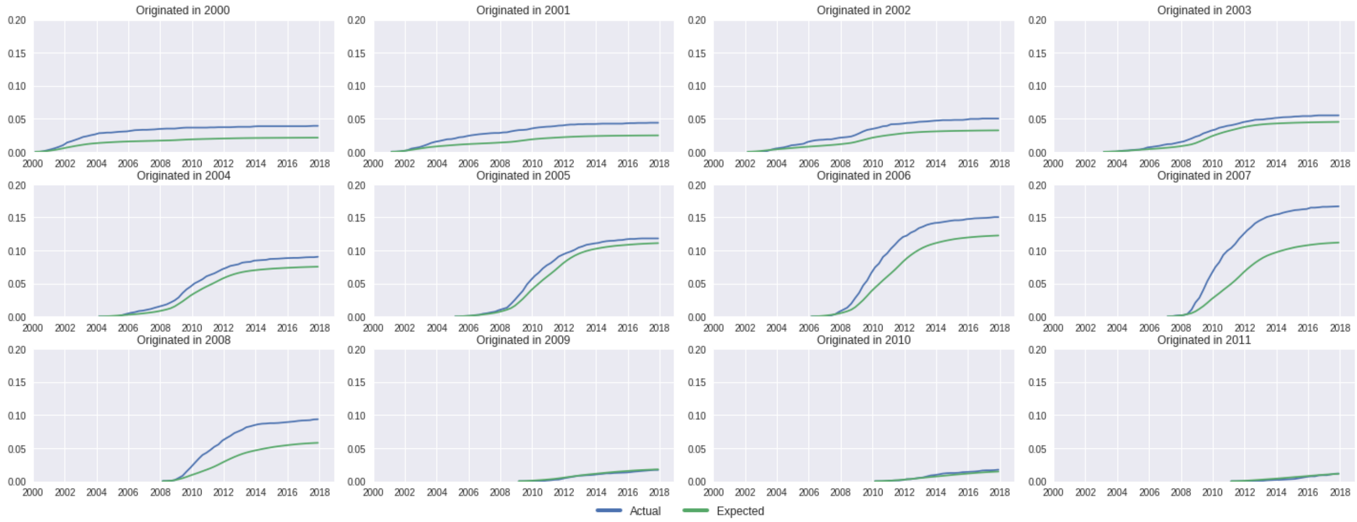


Figure 8: Actual Versus Expected Default Rates by Origination Year

5 Conclusion

Figure 5 displays the covariates' impact on the probability of default and prepayment. A positive coefficient indicates a higher likelihood of default/prepayment.

Since a multinomial logistic regression is implemented, the impact of the covariates is the impact on default in comparison to no default or prepayment. In the default model, the covariates that cause a higher default probability are:

- DTI (debt to income), a higher debt ratio is associated with a higher likelihood of default
- The larger the difference between current 30yr Treasury rate and mortgage origination 30yr Treasury rate, the higher the likelihood of default
- Broker and unspecified channels have a higher likelihood of default compared to retail/correspondent
- As loan age increases, the probability of default also increases
- Loan to value (LTV) is the ratio of the loan divided by its purchase price, the higher this ratio, the more likely a default occurs
- The higher the origination interest rate, the higher the likelihood of default
- The loan purpose indicates if the mortgage is a cash-out refinance, purchase mortgage or no cash-out. When the mortgage is cashout, the likelihood of default is higher than no cash-out refinance. When the mortgage loan is purchase, it has a lower default probability than no cash-out.
- When the unemployment rate is higher in the current period than at origination, there is a higher likelihood of default
- A higher house value (according to Zillow) indicates a higher likelihood of default.

When the mortgage is paid out before the mortgage due date, it is generally due to the borrower refinancing with another mortgage broker. Since the model is multinomial logistic, the impact that the covariates is with respect to the baseline (the borrower does not default or prepay the mortgage). The covariates that cause a higher prepay probability are:

- As loan age increases, the more likely for the mortgage to be prepaid compared to no default, but after a certain age this probability decreases (as demonstrated by the quadratic and cubic terms)
- When there are more building permits there is a higher chance of prepayment
- The higher the borrower's credit score, the higher prepayment probability
- When there are two borrowers, there is a higher likelihood of prepaying the mortgage
- The higher the number of current homes sold (Existing Home Sales) compared to the amount at origination has a positive impact on prepayment of the mortgage
- The OECD Leading Indicator shows fluctuations of economic activity. A higher OECD Leading Indicator indicates a higher likelihood of prepay.
- When newly constructed home sales (New Home Sales) are higher, the borrower is more likely to prepay their mortgage
- The higher the origination interest rate, the more likely the home will be prepaid (most likely the borrower can refinance the mortgage at a lower rate)
- A higher unemployment rate indicates a higher likelihood of prepayment
- A higher 30yr Treasury rate indicates that the borrower is more likely to prepay the mortgage
- House value also has a positive effect on the probability of prepayment