30 Year Fixed-Rate Agency Mortgage Default Risk Prediction

Honors Research Distinction in Finance at Fisher College of Business

By

Yiming Xu

Bachelor of Science in Business Administration in Finance

Bachelor of Science in Mathematics, minor in statistics

The Ohio State University

2019

Committee:

Professor Kewei Hou, Advisor

Abstract

Prior to 1980's, people evaluated mortgage default risk established on rule of thumb and their experience towards risk ratings. The collapse of mortgage market in 2008 stimulated people to quantitatively assess mortgage default risk hence different statistical models are applied to consider different and specific business requirements. A good measure of Probability of Default (PD) benefits financial institutions in assessing loan loss and some insights from modelling default risk can guide a competitive mortgage pricing and better underwriting practice. This project provides some quantitative methods to help institutions assess the default risk on a pool of mortgage loans using statistical tools.

Several models such as logistic regression, gradient boosting and decision trees were used to evaluate risks. HPI files were matched with original data to provide better predict ability. Through GBM feature selection and covariance matrix analysis, eight variables were totally selected. Cross validation was conducted to choose best hyper parameters and rare event model was applied in logistic regression as well to reduce suffering from small-sample bias. The ROC tests for these models reached 0.81 and KS scores were greater than 0.45.

3.5 million Mortgage observations were used from Freddie Mac (issued between Q42009 to Q32012 with five years' performance). HPI (Housing Price Index) file from FRED is also used to combine with mortgage data.

This paper adds machine learning techniques to traditional statistical models to predict single-family mortgages' default risk within five years of the issuance. Strategies such as mortgages swap in/out can be applied on application side together with models, which can improve profitability and system efficiency. We can also have a better intuition towards how credit markets evolve after the financial crisis from feature importance of different models.

Acknowledgments

This research was made possible by support from

Dr. Kewei Hou

Table of Contents

List of Figures

Part 1: Introduction

During the early years of 21st century, the housing price boom stimulated the development of the U.S. mortgage market while were also characterized by great uncertainties. Prior to 1980's, people evaluated mortgage default risk established on rule of thumb and their experience towards risk ratings. It was till the collapse of mortgage market in 2008 did people realize the importance to assess default risk quantitatively. Losses to mortgage lenders did not only cause a large economic downturn, but also stresses different financial systems. After the financial crisis, although mortgage default rates have been relatively low due to stricter regulation, it is still necessary to evaluate them due to the large size of market (15,419,529 millions of dollars at the end of 2018Q4) and high cost to borrowers with default loans.

A good measure of Probability of Default (PD) benefits financial institutions in assessing loan loss and some insights from modelling default risk can guide a competitive mortgage pricing and better underwriting practice. The features from PD prediction models can help financial institutions to understand the household incentives to default on mortgages and explore various relationship among different applicants as well. This paper applies statistical and machine learning techniques beyond traditional predictive models to enhance their performances and compare how these models effectively capture decisive features and classify default / non-default mortgages.

Part 2 Data Process

1. Data Sources

The information provided in this paper serves as a reference for understanding the Single-Family Loan-Level Dataset. The Dataset includes:

- Loan-level origination, monthly loan performance, and actual loss data on a portion of the fully amortizing 30-year fixed-rate1 Single Family mortgages that Freddie Mac acquired with origination dates from 2009Q4 to 2012Q3 (traced with 5 years performance data)
- The following types of mortgages were excluded from dataset:
    1) Adjustable Rate Mortgages (ARMs); mortgages with step rates
    2) Government-insured mortgages, including Federal Housing Administration/Veterans Affairs (FHA/VA), Guaranteed Rural Housing (GRH), HUD-Guaranteed Section 184 Native American mortgages
    3) Mortgages delivered to Freddie Mac under alternate agreements; documentation is not verified or waived;
    4) Mortgages associated with Mortgage Revenue Bonds purchased by Freddie Mac
- Loan performance data contains monthly performance, delinquency status and certain information including earliest event of Prepaid/Foreclosure/Repurchase/REO Disposition
- House Price Index downloaded from FRED Econ Data, which contains quarterly HPI from different area based on zip code from year 1975 to 2018.

2. Origination Data File

- 27 features are included in the Origination Data File, matched with each mortgage ID.
- Ten variables such as First Payment Date, Maturity Date, PPM Flag, Product Type, Seller Name

are deleted in that they not useful to predict default risk; 17 variables are kept after first scan

- First Time Home Buyer Flag: The empty entries is replaced by 'U', which means unknown
- MSA: If the MSA code could be found in our HPI file, then it would remain the same. Otherwise, it is replaced by the property state.
- Mortgage Insurance Payment: the data type for this column is modified from string to float
- Loan Size: The Original UPB column in the Freddie mac's origination data
- Super Conforming Flag: Replace empty entries to 'N'
- Loan ID: The original *LOAN SEQUENCE NUMBER* column, only loans which have records in the time data, and whose property state can be found in HPI data are picked.
- CLTV Highest: the max value of CLTV in the first 60 months
- The loan size variable is applied log transformation to have a more smooth range
- Delete mortgages whose number of units greater than 2 due to a low percentage
- Drop observations that have outliers (FICO score greater than 850; mortgage insurance payment greater than 180; LTV and CLTV greater than 200)

3. Monthly Performance Data

- 5 features are selected from 27 features in performance data
- Ind_Default_2: Equal to one when the loan ever reached D150 states or its zero-balance code was ever equal to 03 or 06 or 09; this is also the target variable for the following prediction

4. House Pricing Index Data

- 3 features are selected from original HPI data file
- The HPI file whose time from 1975 to 2008 is also deleted to match the origination data file; variable Year and Quarter are converted into one single variable QUARTER_DATE and has data

type of string.

• HPI_MAX/MIN: The loan's highest/lowest HPI in the first 60 months

• HPI_UP/DOWN_CHG: the max/min HPI percentage change in the first 60 months

5. Finalized Integral Data

3.5 million loan observations dataset is finalized by merging the Origination Data, Performance Data and HPI Data, a head file of 10 observations are displayed in Figure 1.

Part 3 Data Analysis and Feature Engineering

1. Single Variable Analysis

Single variables analysis can help understand the distribution for different features and check abnormalities from the original datasets.

• For categorical variables, mortgages are grouped by different categories and average default rate is computed for each level

• For continuous variables, if unique values of such variable are less than 15, then it is grouped by these different values; otherwise, cut the range of continuous variable values into 15 bins, plot them in a histogram and compute the mean value of default rates for each bin.

• The distribution of selected 17 features together with default rate are displayed from Figure 2-14. Seven important features are analyzed based on their importance in feature engineering process:

1) From the distribution of property type (Figure 3), single family has the highest portion up to

70% with average default rate of 0.65%; planned units are the second largest portion among different property types with the lowest average default rate. Condo as a type of living space similar to an apartment but independently sellable has highest default rate possibly due to the purpose of investment.

2) From the distribution of occupancy status (Figure 5), nearly 90% of mortgages are classified as primary residence; investment and second home have similar but relative low proportion; the default rate of investment is higher than other two groups in that investors focus more on the house return rather than a living space. The second home status has the lowest default rate in that they are more affluent than primary residence.

3) It is obvious that mortgages exceeding conforming loan (Figure 7) limits have a lower average default rate. The mortgages meet this flag are also called nonconforming or jumbo loan. Since jumbo loans have higher home values, applicants are faced much more rigorous credit requirement with no guarantee by Fannie Mae or Freddie Mac. Loans with super conforming flag will have lower default rates on average due to stringent requirements.

4) The distribution of FICO score (Figure 8) seems to skew left because of the negative direction on the score line. Most applicants have their credit scores ranging from 725-825. FICO score may be indicative of the likelihood that the borrower will timely repay future obligations. Thus, the higher FICO score, the lower default rate of mortgages on average.

5) The LTV ratio (Figure 10) is computed by dividing the original mortgage loan amount by the lesser of the mortgaged property's appraised value or its purchase price. Nearly half of loans have LTV ratio around 80%. The percentage of a mortgage turns to move with default rate in the same direction.

6) The HPI_DOWN_CHG (Figure 13) is calculated by *max (0, (-min HPI + original HPI)/original HPI)*, which reflects the maximum decrease of average housing price within a specific

area. The distribution may have a negative relationship with default rate: the faster the house price falls, people are more likely to default.

7) Number of borrowers (Figure 15) is number of people who are obligated to repay the mortgage secured by the mortgage property. 75% of mortgages have 2 borrowers and nearly 25% of mortgages only have one borrower. From the distribution of default rate, mortgages have 1 borrower are more than twice as likely to default, which can result from weak supervision between borrowers.

2. Correlation matrix analysis

It is necessary to conduct a correlation matrix between continuous variables in that this paper will apply logistic regression to predict default rate. The regression estimate will be unreliable if there is a high amount of correlation between different features. From the correlation matrix chart (Figure 16), original combined LTV and original LTV have a correlation up to 0.96. The column LTV is dropped to ensure assumptions of linear model.

3. Variable Selection by Gradient Boosting Machine

1) Decision tree is a tree-based classification model. It splits the data into cases and repeat till make a decision. It is a way to show an algorithm that only contains conditional control statements. Leaf nodes, non-leaf nodes and stopping criteria are 3 elements that make up a decision tree.

2) Gradient Boosting Machine (GBM)

As one of the most powerful learning ideas in the past two decades, GBM is a statistical optimized method to minimize the loss function of the model by adding ‘weak’ learners to produce a ‘strong’ learner using a procedure that is very similar to gradient descent. GBM is built to continuously reduce the loss of model.

3) Feature importance in GBM

Gradient boosting algorithm is straightforward to give us a contribution of different variables after constructing boosted trees. The importance is measured by amount of entropy reduced for each feature and how much it has improved the performance through splits. The feature importance is then averaged across all trees.

The Figure 17 displays feature importance by GBM. From the figure, FICO score, number of borrowers, maximum HPI decrease, number of borrowers, CLTV, loan size and mortgage insurance payment are ranked by their contribution to the model in a descent order.

4. Eight variables are left after feature engineering process:

Continuous Variables: FICO; MORTGAGE_INSURANCE_PCT; ORGN_CLTV; LOG_LOAN_SIZE; HPI_DOWN_CHG; NUM_OF_BORROWERS.

Categorical Variables: PROP_TYPE; LOAN_PURPOSE.

Dummy variables are created for each categorical variable to represent different levels, finalized dataset is separated by 80% training data and 20% testing data as well.

Part 4 Models and Methods

1. The general loss function that different classification models aim to optimize is:

$$\mathrm{f(w,X,y)} = \sum_{i=1}^{n}\left(y_i \log\left(\frac{1}{1+e^{-w^T x_i}}\right) + (1-y_i)\log\left(1-\frac{1}{1+e^{-w^T x_i}}\right)\right);$$

where n is number of observations, wi is the weight for features.

1) Logistic Regression

Logistic regression is a form of binary regression model to predict a categorical dependent variable with input that at least contains one continuous dependent variable. In this paper, the dependent categorical variable will be *IND_DEFAULT2* (two levels: 1 represents default and 0 represents non-default).

To prevent overfitting of the model, it’s effective to augment our loss function with a term that serves to penalize large weights. The newly loss function after regularization for logistic regression is:

$$\mathrm{f(w, X, y)} = \sum_{i=1}^{n}\left(y_i \log\left(\frac{1}{1+e^{-w^T x_i}}\right) + (1-y_i)\log\left(1-\frac{1}{1+e^{-w^T x_i}}\right)\right) + C\sum_{i=1}^{n}|wi|;$$

where and C is a hyper parameter.

The regularization process can not only reduce the flexibility of fitting our training data, but also increase model performances in future. Thus, features are normalized due to the regularization process.

2) Cross Validation to select Hyper parameter for logistic regression

Often it is necessary to consider many models and ‘model’ simply means a particular setting of hyper parameters. The fitting process is trying to optimize the parameters and fit the training data as well as possible.

In the loss function of logistic regression, the constant C is a hyper parameter. To find the best constant C, a vector from 0.0001 to 10000 is created to search the best model hyper parameter. The best parameter C is 0.01 after cross validation.

3) Cross Validation to select Hyper parameters for Gradient Boosting Machine

Since GBM is a boosting method based on ensemble decision trees, here are several hyper parameters that needed to be optimized:

- Learning rate: it determines to what extent newly acquired information overrides old information. A large learning rate may be not even converging, while a small learning rate can take a long time to solve the problem.

The learning rates vector in GBM is [0.1,0.05,0.01,0.005].

• Minimum sample split: a stopping criteria in decision tree when number of samples in a node is less than pre-defined numbers. A small minimum sample split can cause overfitting problem and a large minimum sample split may yield a bad classifier.

Minimum sample split vector is [100,200,500,100].

• Maximum depth: another stopping criteria in decision tree to control the complexity of the model and overfitting problems.

Maximum depth vector is [2,4,6,8].

• Number of estimators: number of decision trees in the bagging process, which aims to control model complexity and overfitting problems.

Number of estimator vector is [100,300,500,1000]

After defining the range for hyper parameters in GBM, each combination (total 256 cases) is conducted to build a gradient boosting model and measure its performance by ROC/AUC score within training data. The best hyper parameters for GBM are: learning rate: 0.05; max depth: 2; min samples split: 500; number of estimators: 500.

4) The output of logistic regression with optimized hyper parameter is displayed below:

	Parameter
FICO	-0.67366
MORTGAGE_INSURANCE_PCT	0.08608
ORGN_CLTV	0.45311
LOG_LOAN_SIZE	-0.15314
HPI_DOWN_CHG	0.26521
NUM_OF_BOEEOWERS	-0.56572
PROP_TYPE_PU	-0.13693
LOAN_PURPOSE_N	-0.25358
LOAN_PURPOSE_P	-0.29953
INTERCEPT	-5.74926

From the summary, the magnitude (absolute value) of the parameter represent the weight assigned for this variable, while the sign of the parameter represents whether the variable will move with default risk in the same direction. From the above summary, FICO has a negative sign and the highest weight, which means it contributes most to capture default mortgages and moves in opposite direction against default probability. Similarly, number of borrowers and original CLTV are decisive to classify default and non-default mortgages due to relative high weights.

Part 5 Model Validation

Since the mortgage data is unbalanced (only 0.6% default mortgages). Applying accuracy rate to test model performance can cause lots of problems such as poor performance. The most used indices in practice are K-S test and ROC/AUC score.

1) Kolmogorov-Smirnov Test

KS test is a nonparametric test whether a sample follows a reference probability distribution. The statistics quantifies a distance based on cumulative distribution functions. The CDFs of default and non-default applicants are given by the relationships

$$F_{non-default} = \frac{1}{n}\sum_{i=1}^{n} I(S_i \leq a \wedge Dk = 0)$$

$$F_{default} = \frac{1}{m}\sum_{i=1}^{m} I(S_i \leq a \wedge Dk = 1)$$

- Si is the PD of the ith mortgage, n is number of non-default mortgages, m is number of default mortgages, I is indicator function where I (true) = 1 and I (false) = 0. A is a value within the probability domain [0, 1].

$$KS = max_{a\in[0,1]} \left|F_{m,default}(a) - F_{n,non\ default}(a)\right| ;$$

The KS statistics measures how well a model to separate default and non-default mortgages, a larger distance represents a higher KS score and a better performance.

The K-S curves for two models are plotted within the same picture (Figure 17). All four lines converge to 1 really quickly due to unbalanced data between default and non-default mortgages. The cumulative default line of GBM is slightly lower compared to the cumulative default line of logistic regression, which means GBM has a better performance to classify default mortgages.

2) Receiver operating characteristic

Receiver operating characteristic (ROC) is a graphic plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. In other words, ROC summarizes all of confusion matrix under different thresholds.

		Actual	
		Default	Non-default
Predicted	Default	True Positive (+)	False Positive (+)
	Non-default	False Negative (-)	True Negative (-)

Figure 16: Confusion Matrix

From the ROC curve, the y axis is True Positive Rate (= true positive/ (true positive + false negative)), which measures the percentage of default applicants who are correctly classified. The x axis is False Positive Rate (= false positive/ (false positive + true negative)), which measures the percentage of sample classify as non-default that falsely predicted. Tradeoffs between true positive rate and false negative rate are plotted under different probabilities. A straight line with slope of 1 is added in ROC as well to represent a random binary distribution. The Area Under Curve (AUC) is a number that calculate the area under ROC curve, which measures the model's overall ability to identify default mortgages.

The ROC curves between logistic regression and gradient boosting machine (Figure 18), two models have similar area under the curve, while GBM curve slightly outperforms logistic regression near the thresholds of 0.4 and 0.5.

Part 6 Conclusion

After conducting K-S and ROC/AUC tests in model validation process, the test statistics of two models can be summarized as following:

	Logistic Regression	GBM
K-S statistics	0.481	0.485
ROC-AUC score	0.805	0.815

Comparing the performances between two models, gradient boosting machine has both higher K-S statistics and ROC-AUC score, which indicates it a stronger tool to capture default mortgage characteristics and separate them from non-default mortgages. Thus, gradient boosting machine is preferred to predict mortgage default risks.

From the feature importance and summary of logistic regression, FICO score, maximum decrease in HPI within 5 years and number of borrowers are the three strongest factors to determine whether a mortgage is likely to default. Once the risk tolerance interval is determined, some swap in/out strategies can be applied on the mortgage application side.

The results of two predictive models shows that machine learning techniques can greatly improve the performance of traditional predictive models and quantify mortgage risks, which does not only help firms understand credit market after financial crisis, but also improve the overall efficiency of financial systems.

Appendix A Feature Description among all datasets

1. Origination Data File

- Credit Score: summarize the borrower's creditworthiness
- First Payment Date: date of first scheduled payment under the terms of note
- First Time House Buyer Flag: indicates whether borrower is purchasing the property, will be primary residence, has on ownership interest during three year period preceding purchase
- Maturity Date: time of final payment
- Metropolitan Statistical Area (MSA): indicates the area in which the property locates
- Mortgage Insurance Percentage: % of loss coverage on the loan when default)
- Number of Units: whether the mortgage is a one/two or multiple unit property
- Occupancy Status: Whether the home is occupied, second home or investment property
- Original Combined Loan to Value (CLTV): (original loan amount + secondary mortgage loan amount) / purchase price
- Original Debt-To-Income Ratio: sum of borrower's monthly debt/total monthly income;
- UPB: Unpaid principal balance, the proportion that has not been remitted to the lender
- Original Loan-To-Value: original loan amount/purchase price
- Original Interest Rate: note rate indicated in the mortgage
- Channel: decide whether the mortgage origination involves retail/broker/correspondent
- Prepayment Penalty Flag: whether the mortgage is PPM
- Property Type: indicates the whether the property belongs to condo, planned development area, co-op or single family
- Loan Purpose: whether the loan is purchase, cash-out or non-cash-out refinance mortgage

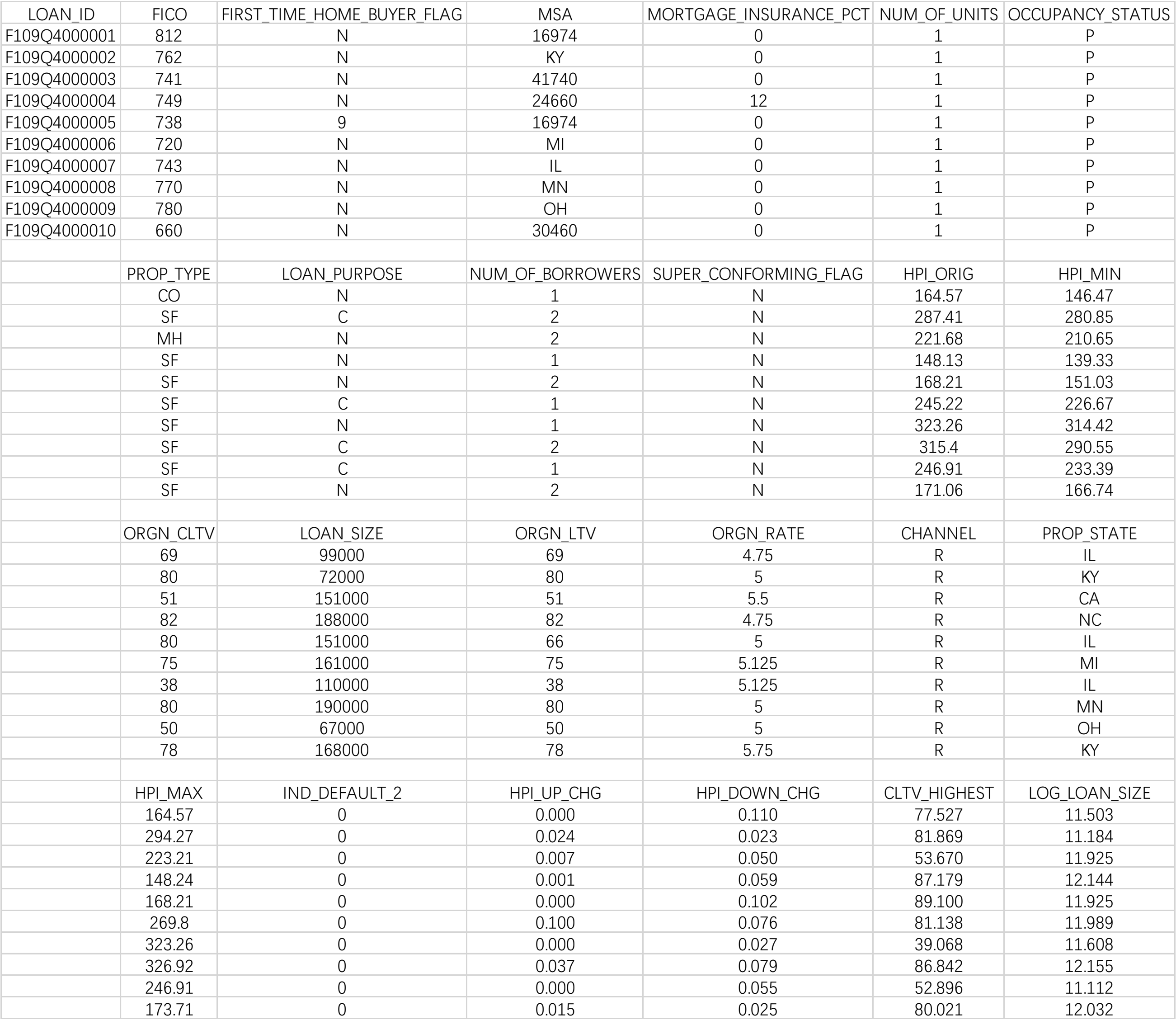
- Number of Borrowers: number of people who are obligated to repay mortgage note
- Super Conforming Flag: indicates whether mortgages exceeding conforming loan limits

2. Monthly Performance Data

- Current Loan Delinquency Status: number of days the borrower is delinquent
- Loan Age: Number of months since loan origination
- Repurchase Flag: Indicates whether the mortgage has been repurchased
- Zero Balance Code: indicates the reasons the loan's balance reduced to zero (Prepaid, Foreclosure, Repurchased or REO Disposition)
- Other 22 variables are not used due to the target variable definition

3. HPI File

- Location: the MSA code for different areas
- Year/Quarter: HPI index of the indicated year and quarter
- HPI: the value of housing price



LOAN_ID	FICO	FIRST_TIME_HOME_BUYER_FLAG	MSA	MORTGAGE_INSURANCE_PCT	NUM_OF_UNITS	OCCUPANCY_STATUS
F109Q4000001	812	N	16974	0	1	P
F109Q4000002	762	N	KY	0	1	P
F109Q4000003	741	N	41740	0	1	P
F109Q4000004	749	N	24660	12	1	P
F109Q4000005	738	9	16974	0	1	P
F109Q4000006	720	N	MI	0	1	P
F109Q4000007	743	N	IL	0	1	P
F109Q4000008	770	N	MN	0	1	P
F109Q4000009	780	N	OH	0	1	P
F109Q4000010	660	N	30460	0	1	P
	PROP_TYPE	LOAN_PURPOSE	NUM_OF_BORROWERS	SUPER_CONFORMING_FLAG	HPI_ORIG	HPI_MIN
	CO	N	1	N	164.57	146.47
	SF	C	2	N	287.41	280.85
	MH	N	2	N	221.68	210.65
	SF	N	1	N	148.13	139.33
	SF	N	2	N	168.21	151.03
	SF	C	1	N	245.22	226.67
	SF	N	1	N	323.26	314.42
	SF	C	2	N	315.4	290.55
	SF	C	1	N	246.91	233.39
	SF	N	2	N	171.06	166.74
	ORGN_CLTV	LOAN_SIZE	ORGN_LTV	ORGN_RATE	CHANNEL	PROP_STATE
	69	99000	69	4.75	R	IL
	80	72000	80	5	R	KY
	51	151000	51	5.5	R	CA
	82	188000	82	4.75	R	NC
	80	151000	66	5	R	IL
	75	161000	75	5.125	R	MI
	38	110000	38	5.125	R	IL
	80	190000	80	5	R	MN
	50	67000	50	5	R	OH
	78	168000	78	5.75	R	KY
	HPI_MAX	IND_DEFAULT_2	HPI_UP_CHG	HPI_DOWN_CHG	CLTV_HIGHEST	LOG_LOAN_SIZE
	164.57	0	0.000	0.110	77.527	11.503
	294.27	0	0.024	0.023	81.869	11.184
	223.21	0	0.007	0.050	53.670	11.925
	148.24	0	0.001	0.059	87.179	12.144
	168.21	0	0.000	0.102	89.100	11.925
	269.8	0	0.100	0.076	81.138	11.989
	323.26	0	0.000	0.027	39.068	11.608
	326.92	0	0.037	0.079	86.842	12.155
	246.91	0	0.000	0.055	52.896	11.112
	173.71	0	0.015	0.025	80.021	12.032

Figure 1: Finalized Integral Data

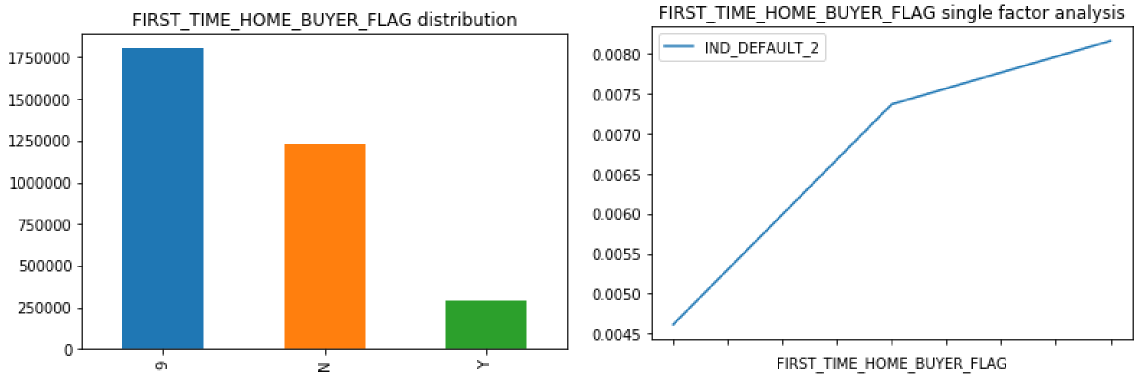


Figure 2: Distribution of First Time Home Buyer Flag

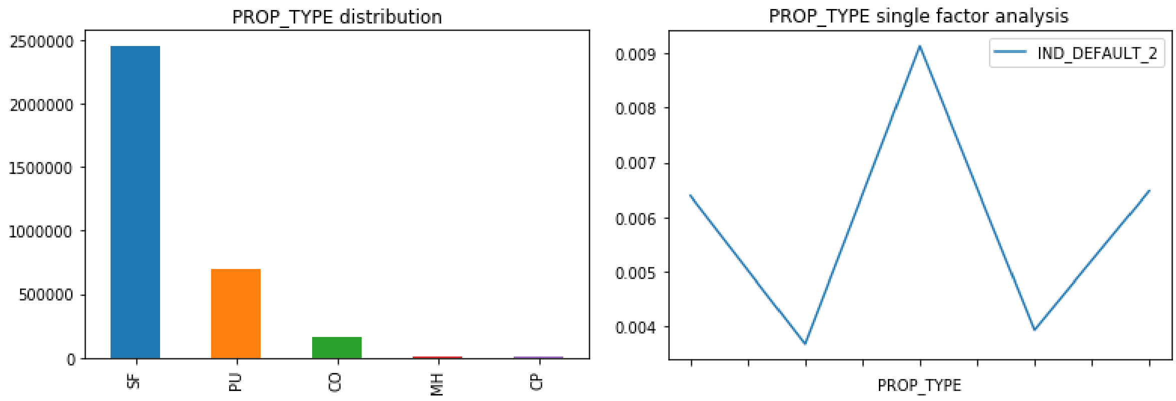


Figure 3: Distribution of Property Type

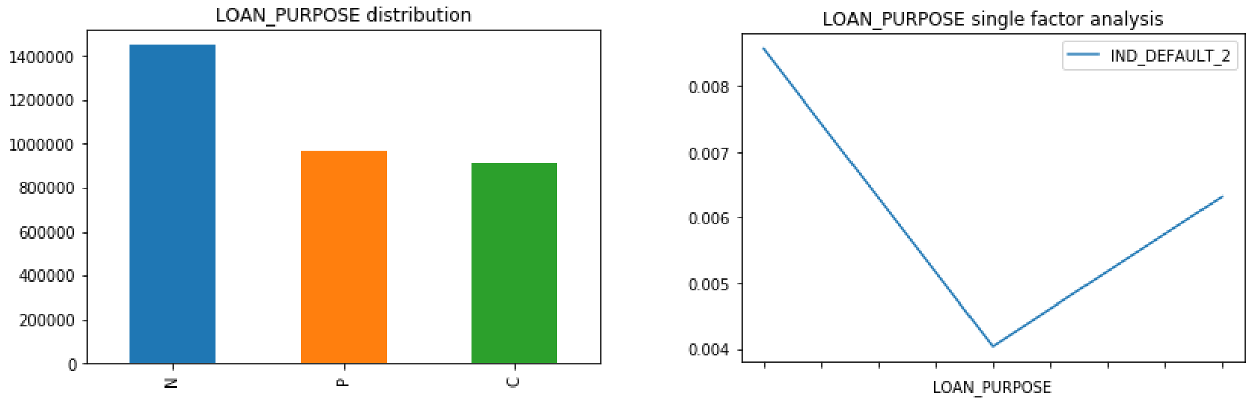


Figure 4: Distribution of Loan Purpose

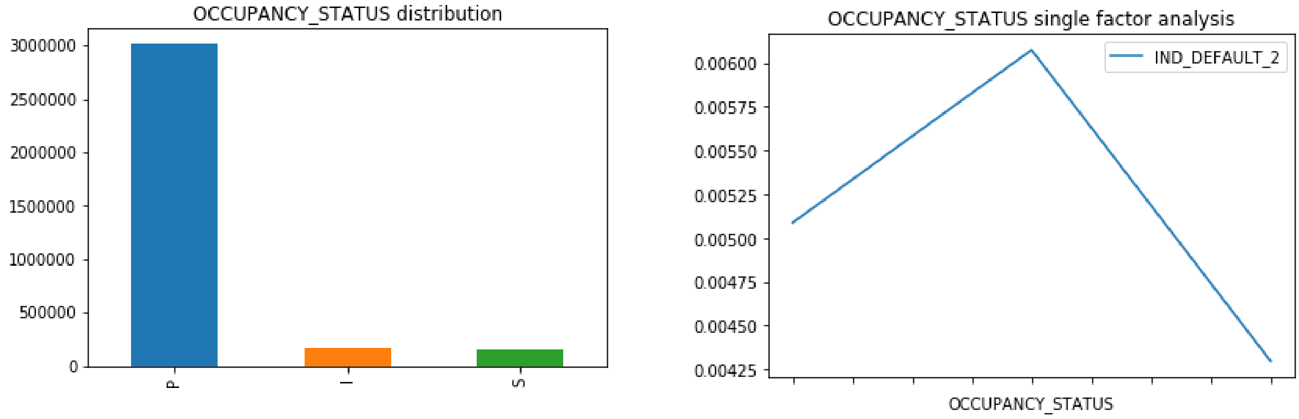


Figure 5: Distribution of Occupancy Status

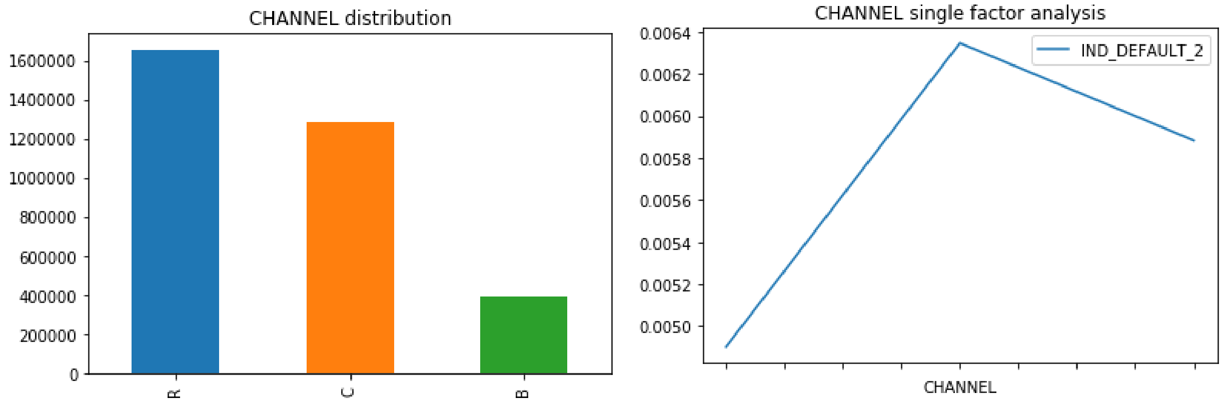


Figure 6: Distribution of Channel

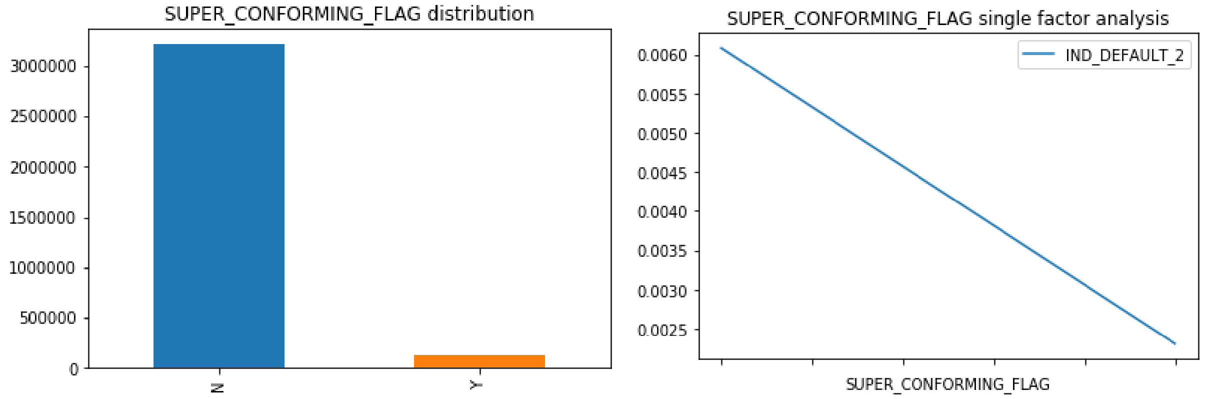


Figure 7: Distribution of Super Conforming Flag

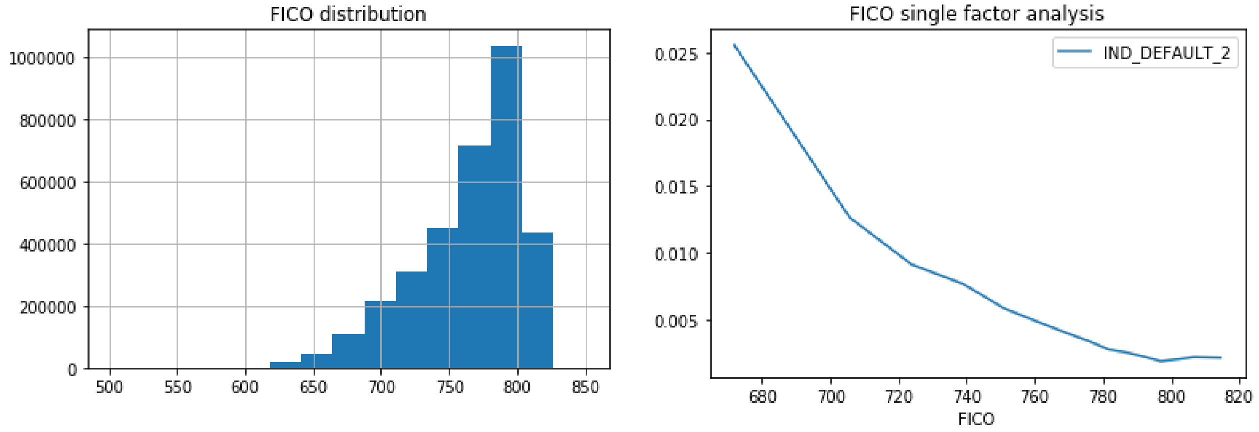


Figure 8: Distribution of FICO Score

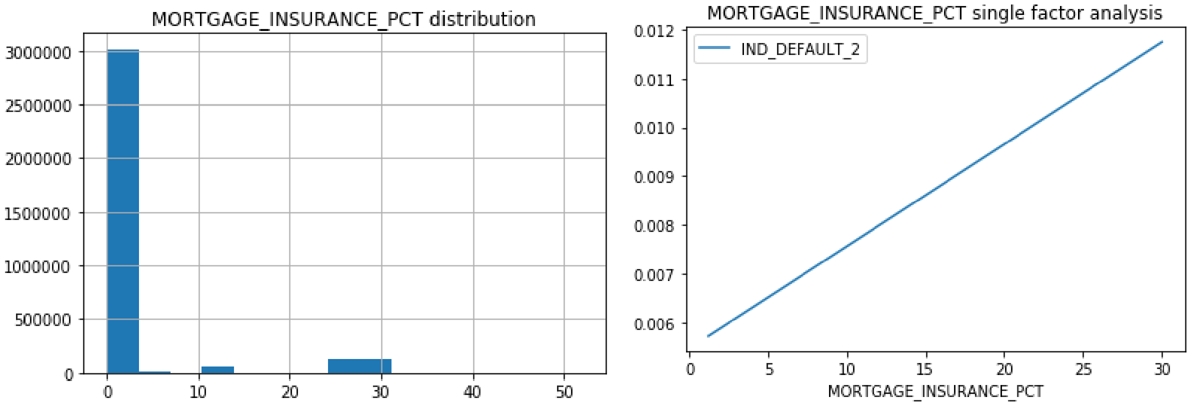


Figure 9: Distribution of Mortgage Insurance Payment

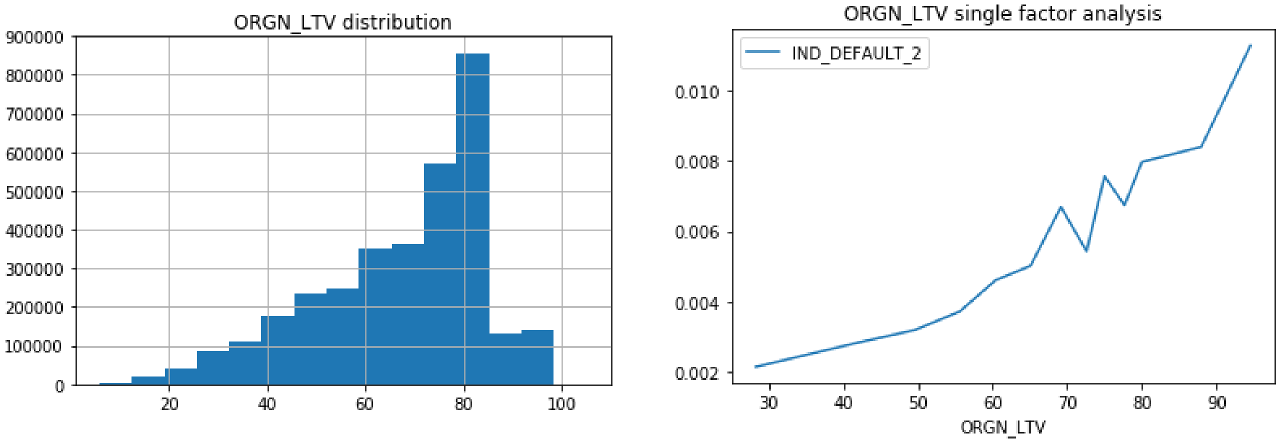


Figure 10: Distribution of Original LTV

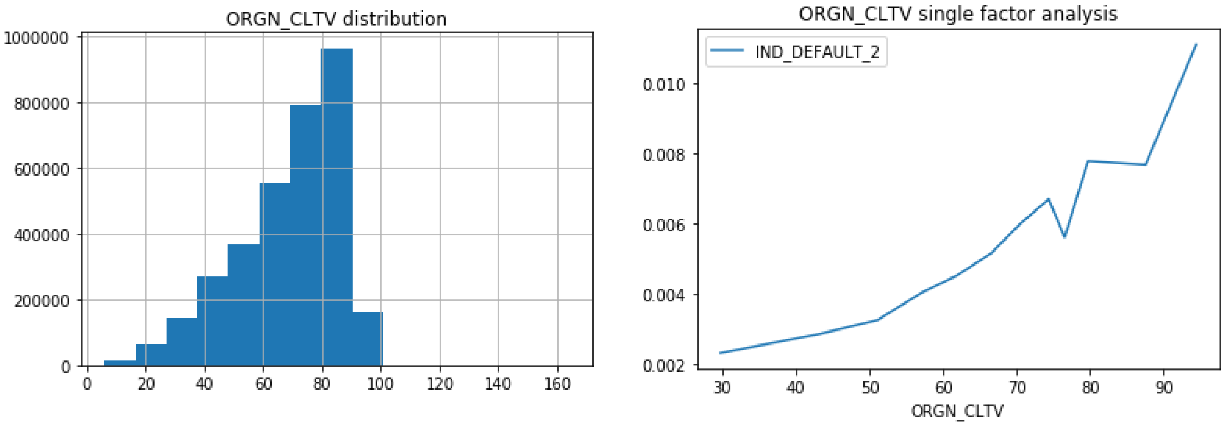


Figure 11: Distribution of Original CLTV

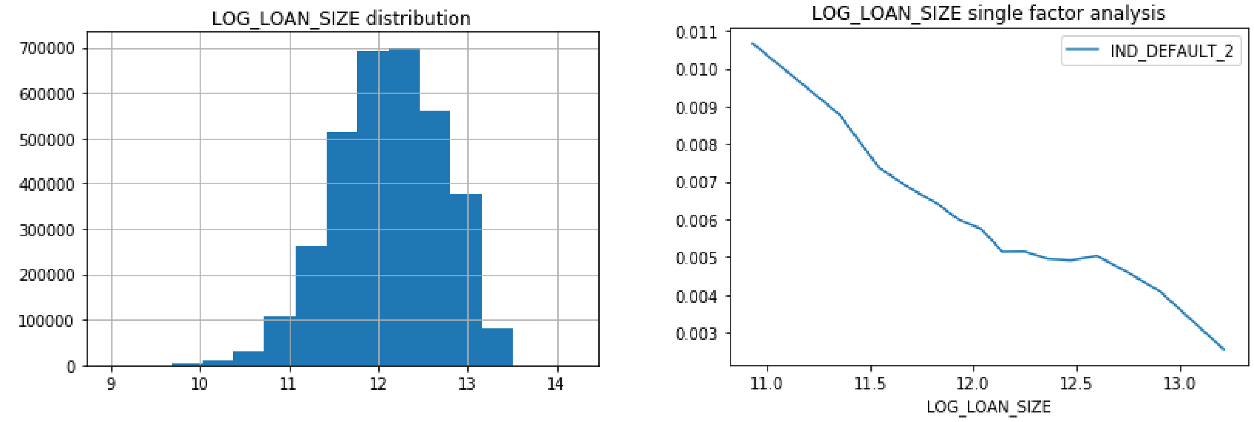


Figure 12: Distribution of Log Loan Size

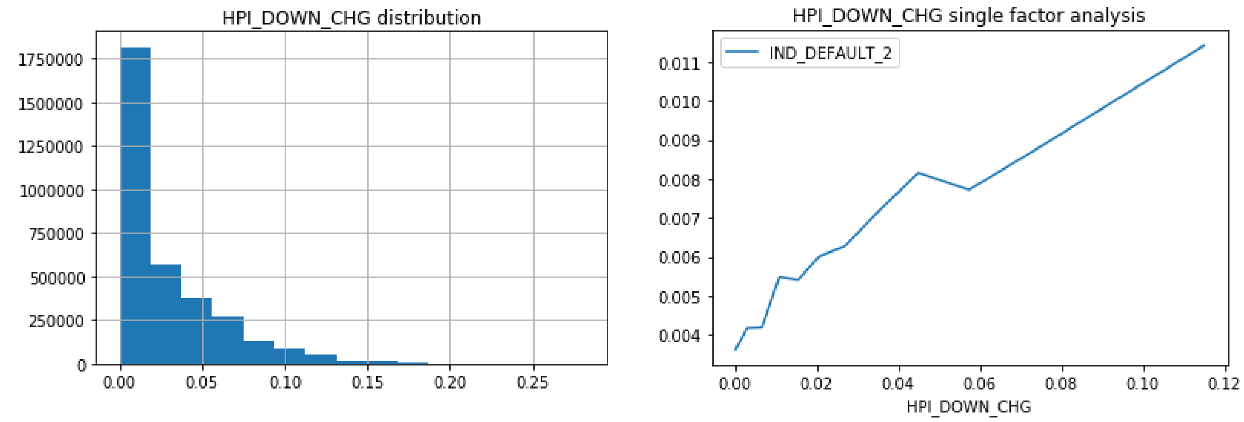


Figure 13: Distribution of HPI Maximum Decrease

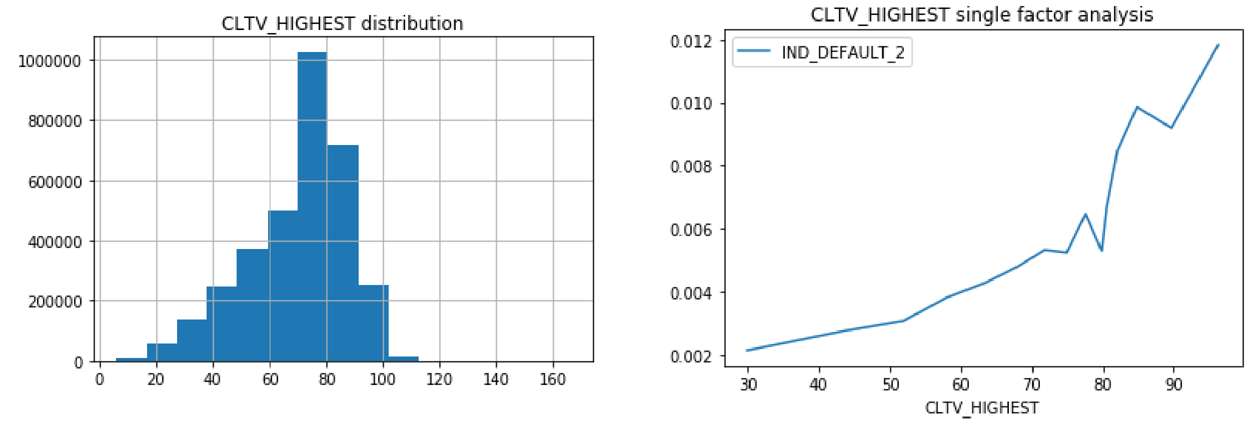


Figure 14: Distribution of Highest CLTV

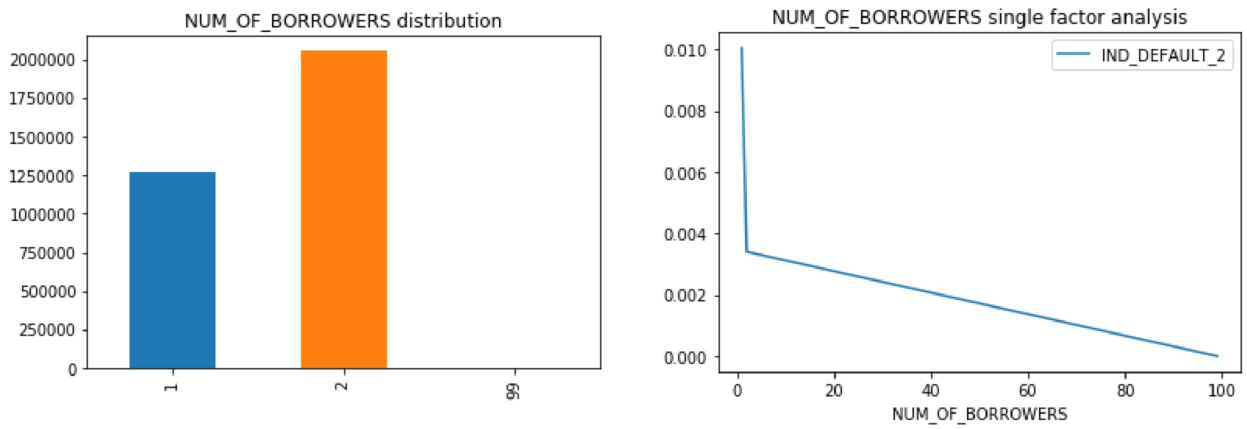
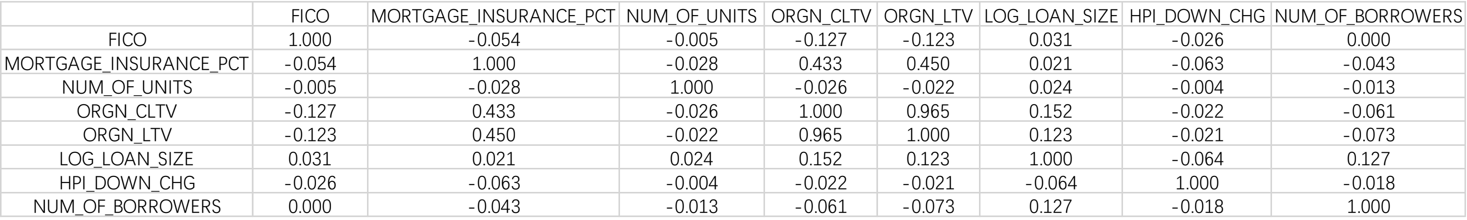


Figure 15: Distribution of Number of Borrowers



	FICO	MORTGAGE_INSURANCE_PCT	NUM_OF_UNITS	ORGN_CLTV	ORGN_LTV	LOG_LOAN_SIZE	HPI_DOWN_CHG	NUM_OF_BORROWERS
FICO	1.000	-0.054	-0.005	-0.127	-0.123	0.031	-0.026	0.000
MORTGAGE_INSURANCE_PCT	-0.054	1.000	-0.028	0.433	0.450	0.021	-0.063	-0.043
NUM_OF_UNITS	-0.005	-0.028	1.000	-0.026	-0.022	0.024	-0.004	-0.013
ORGN_CLTV	-0.127	0.433	-0.026	1.000	0.965	0.152	-0.022	-0.061
ORGN_LTV	-0.123	0.450	-0.022	0.965	1.000	0.123	-0.021	-0.073
LOG_LOAN_SIZE	0.031	0.021	0.024	0.152	0.123	1.000	-0.064	0.127
HPI_DOWN_CHG	-0.026	-0.063	-0.004	-0.022	-0.021	-0.064	1.000	-0.018
NUM_OF_BORROWERS	0.000	-0.043	-0.013	-0.061	-0.073	0.127	-0.018	1.000

Figure 16: Correlation Matrix

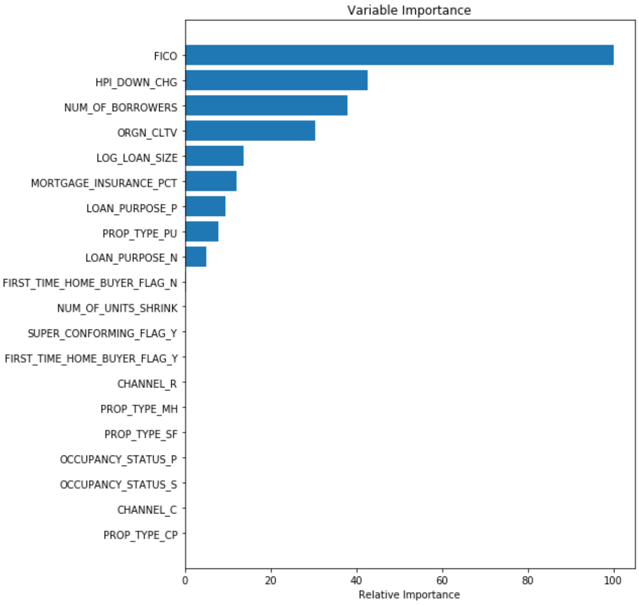


Figure 17: Feature Importance by GBM

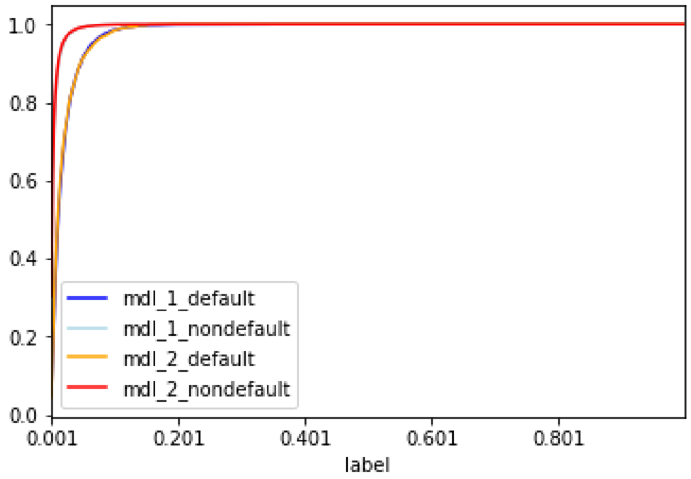


Figure 18: K-S test for two models

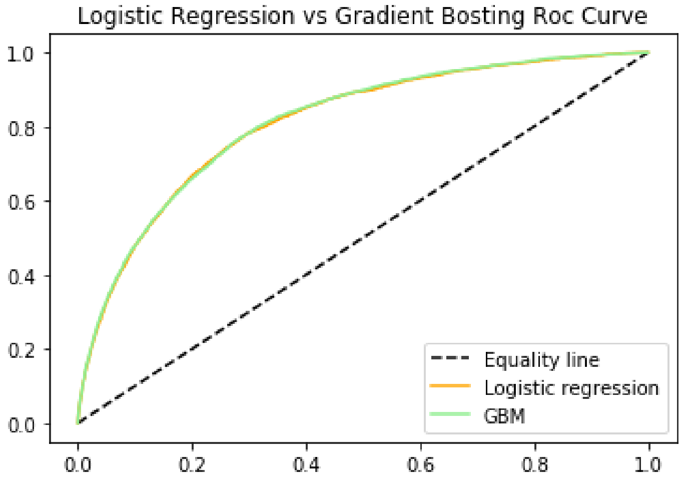


Figure 19: ROC curve for two models

References

[1] Li, Mingxin. "Residential Mortgage Probability of Default Models and Methods." The Journal of Financial Institutions Commission, British Columbia (2014)

[2] Single Family Loan-Level Dataset General User Guide issued by Freddie Mac in August 2018

[3] Federal Reserve Bank of St. Louis and US. Office of Management and Budget, Federal Housing Finance Agency: House Price Index, retrieved from FRED; December 25, 2018

[4] Condominium. (2019, February 28). Retrieved from https://en.wikipedia.org/wiki/Condominium

[5] Investopedia. Jumbo Loan. Retrieved from https://www.investopedia.com/terms/j/jumboloan.asp

[6] A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning. (2018, November 20). Retrieved from https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

[7] Hastie, T., Tibshirani, R., & Friedman, J. H. (2004). The elements of statistical learning: Data mining, inference, and prediction: With 200 full-color illustrations. New York: Springer.

[8] Loss function. Retrieved from https://en.wikipedia.org/wiki/Loss_function

[9] Zheng, Y., Zhang, B., & Cheng, Z. (2014). Hyper-heuristics with penalty parameter adaptation for constrained optimization. 2014 IEEE Congress on Evolutionary Computation (CEC). doi:10.1109/cec.2014.6900471

[10] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1). doi:10.18637/jss.v033.i01

[11] Romeijn, H. E. (n.d.). Random Search Methods. Encyclopedia of Optimization, 2175-2180. doi:10.1007/0-306-48332-7_424

[12] Rezac, Martin & Řezáč, František. (2011). How to Measure the Quality of Credit Scoring Models. Czech Journal of Economics and Finance (Finance a uver). 61. 486-507.

[13] Huang, J., & Ling, C. (2005). Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 17(3), 299-310. doi:10.1109/tkde.2005.50