Fannie Mae Loan Performance Analysis

Kexin Wang, Ruoxi Liu, Ziyan Lin, Tianqi Cui

Introduction

Fannie Mae is created to provide liquidity, stability and affordability to the mortgage market, and thus is important to the economic well-being of the U.S. However, as Fannie Mae acquires mortgages from the financial institutes, it is also exposed to the risks associated with the mortgages. This project is designed from Fannie Mae's prospective, with a goal to predict the performance of the mortgage loans in the future.

Our work mainly contains two parts. The first part is performing classification with the information that Fannie Mae has at the point of mortgage acquisition. The goal is to establish a quick estimation about the status of the loans within three years. The second part is Markov Chain transition analysis. We aim to derive transition probability matrices of the loan status, based on monthly mortgage payment information. The goal of this part is to provide relatively accurate transition probability of the loans, which could be used to compute the expected future losses related to certain loans.

Data Exploration

The Fannie Mae loan dataset contains loan acquisition data and performance data from 2000 Quarter 1 to 2017 Quarter 4. The acquisition dataset contains the basic information of the loan holders, and the performance dataset contains the monthly payment conditions or status changes of the loans. We further add in unemployment rate[1] and Housing Price Index (HPI)[2] as indicators for economic condition and geographical variation.

In order to understand the trends and get a big picture of the data, we explore the dataset by computing summary statistics. We choose loans originated in 2000 to do the descriptive statistics summary, because these loans cover the longest timespan. The total number of year 2000 loans is over 1.2 million, and we find 82% of them are 30-year term loans. Thus, we choose to focus on studying the behavior of the 30-year term loans.

We further summarize the current status of the loans that started in 2000, and find 99% of them have terminated. We also find prepaid is the main reason for termination (97.9% loans acquired in 2000 terminated because of prepaid). When a borrower prepays a loan, the lender misses out a lot of interest. As the statistics shows that the prepaid risk is threatening, so we add prepaid as a main target of our prediction.

Finally, we study the loan termination time, and find more than 80% loan terminated in three years. Because the behaviors of the loans vary a lot from the beginning years to the following years, we believe it is not practical to build one model that could be generally applied through all years. So we set three year as a cut off point. In the classification part, the goal is to predict a loan's status at the end of three year, while in the Markov Chain transition part, the goal is to model the transition probability in year 1, year 2, and year 3.

---

[1] From US Bureau of Labor Statistics, https://www.bls.gov/bls/unemployment.htm

[2] From Freddie Mac, http://www.freddiemac.com/research/indices/house-price-index.page

Classification

1. Methodology

Classification is one of the most useful approaches to identify categories of loan performance. For Fannie Mae dataset, we merge the acquisition and performance datasets by the Loan ID and treat the zero balance code at the end of year 3 as target (as our goal is to predict the loan performance at the end of year three). There are 3 categories of zero balance code: normal, default, and prepaid. Because the distribution of the 3 classes is highly unbalanced, using the whole dataset to perform classification would be inaccurate. As a result, we randomly select some data from the dataset.

In preprocessing step, it is necessary to deal with missing values, categorical variables, and datetime-format features. For missing values, we set 80% as threshold, which means one feature will not be considered if it has more than 20% null values. For those features whose missing values are less than 20% observations, mode values are used to replace null values. This helps to add variance to the dataset. For categorical variables and datetime-format features, we transfer them into dummy variables and numeric variables. One important categorical variable is the state. We try to use GDP, unemployment rate, and HPI as measurements to summarize relationship between states and loan performance.

For feature selection, we plot heatmap to show correlations between all features. Based on the heatmap, we delete one of the two features if they have high correlations with each other. We also use PCA to reduce dimensions of the features. After feature selection, we use 8 different classifiers to train our data. For each classifier, 10-fold cross validation with accuracy as the cv-score is used when training.

2. Results

The accuracy of the 8 classifiers, and the normalized confusion matrix of the random forest classifier (which gives the highest accuracy with lowest deviation) are reported in the following tables.

Classifiers	Accuracy
Perceptron	(68.77±8.18)%
Ridge Classifier	(72.36±6.89)%
Logistic Regression	(74.50±6.81)%
KNN Classifier	(79.13±6.95)%
Linear SVM Classifier	(75.39±7.69)%
RBF SVM Classifier	(76.06±8.35)%
Random Forest Classifier	(80.36±6.54)%
Naïve Bayes Classifier	(64.46±8.53)%

Table 1: Accuracy Summary

		Predicted Status		
		Normal	Default	Prepaid
Actual Status	Normal	88.44%	8.32%	3.24%
	Default	12.17%	78.91%	8.92%
	Prepaid	4.66%	17.65%	77.69%

Table2 : Normalized Confusion Matrix of Random Forest

Based on the confusion matrix, we could see that the accuracies are satisfactory, but not very high, especially for the Prepaid class. One possible reason behind this problem is the unbalanced dataset. Compare to Normal and Default classes, Prepaid data occupies a small part. This unbalance leads to low-accuracy results.

We believe our classification models serve their purposes of providing Fannie Mae a quick judgement of the loan performance in a three-year period since the loans' acquisition. This would help Fannie Mae to roughly determine the economic outcome of a loan portfolio at the acquisition time, when the loans' performance are unknown. Once the loans establish a performance history, more accurate prediction could be made based on the repayment history.

And this leads to our second part, which is to build a Markov Chain based on the loan payment information to forecast their performance in the future.

Markov Chain Transition Analysis

1. Methodology

Markov Chain is widely adopted in the industry to model the transition of financial events between different status. It is a stochastic process with an assumption that future events will only depend on the present rather than the past. However, this assumption would be too strong to be met by the real-world situations. To relax the assumption, we further assume that given a series of covariates, the time independence required by Markov Chain would be met. As a result, we design and include nine covariates in our model to build a conditional Markov Chain model.

We use multinomial logistic regression to derive the transition probabilities of the conditional Markov Chain model, which form the transition probability matrices. The matrices contain probabilities that a loan transfer between different status in one month. There are 9 covariates included in the model: unemployment rate, House Price Index, Unpaid Principal Balance, Loan Age, interest rate, credit score, Debt-to-Income Ratio, Loan Purpose and Loan to Value Ratio. The targets are 9 status depicting the performance of a loan: normal, prepaid, default, high-risk and delinquent for 1 month to 5 month. Besides the 2 absorbing status: prepaid and default, the rest of the status can transfer to any other status, so there are 7 models in total. For example, in model 'normal', we only consider the loans whose start status are normal, and model their behaviors transiting to all status. The other models are built with the same procedures. In addition, each model uses monthly data within a specific year, and thus, the transition probabilities are on monthly basis, and are valid for the whole year.

2. Results

To derive the Markov Chain transition matrices, we use the data from 2000 Quarter 1 to Quarter 4 to train multinomial logistic regression models. We then test the models' performance using the data of 2001, 2005 and 2015 to see if the model could be applied to other years. The model accuracy results are summarised in the table below, and we could see that the results are similar across different years. This proves that although the model is built with the data of 2000, it is still valid modeling the behaviors of the loans in other years (even for year 2015, which is far apart from 2000). We believe this generality verifies our assumption for the conditional markov chain: given the covariates, the loans could be regarded as being independent with each other and from time to time.

	2000	2001	2005	2015
Model "Normal"	97.42%	97.69%	98.60%	98.20%
Model 1	42.20%	41.26%	36.10%	47.81%
Model 2	29.88%	31.01%	38.34%	38.84%
Model 3	55.96%	50.50%	37.76%	53.40%
Model 4	63.08%	57.16%	39.68%	56.14%
Model 5	65.94%	62.91%	56.25%	56.15%
Model High Risk	88.90%	87.77%	89.79%	91.32%

However, we notice that there are several models do not perform very well. To be specific, model 1 and model 2 have low accuracy, and we believe this is mainly because of the density of our status setting. As we discussed before, we set up the loan status continuously from 0 to 5 to represent loans that being delinquent for 0 to 5 months. And we figure that our

model have difficulties capturing the transition behaviors of the intermediate status 1 to 4, because they tend to have equally transition probability to their adjacent status. For example, we find a loan that is currently delinquent for 2 times, is equally likely to transit to a higher delinquent status, which is 3, or a lower delinquent status, which is 1. And the covariates do not really help to identify the direction of the transition, which explains why model 2 has low accuracy.

We could easily improve the accuracy of the intermediate models by combing the status to form less dense status. For example, when we combine status 1,2,3 into one status, and combine status 4 and 5, the accuracy rises above 70%. And when we combine status 0 through 5, the accuracy is above 90%. But we still want to stick with the current model for building a complete transition matrix. The limitation of our current model is that the probability of transiting between the intermediate status is not very accurate, but it could accurately predict the transition to prepaid and default, which is what we care about most. In the future, we could design more informative covariates and include them in the model to help the model to decide the direction of transition, and improve the accuracy of the model.

As the result, the transition matrices for the first three year of a loan are shown on the right. The pattern of the transition probabilities within a year is similar across all three years. A loan is most likely to stay at its current status, and this is particular true for loans that are in status 0. As a loan's delinquency status increases, it is more likely to move to a higher delinquency status, or default. Also, we could notice that the loans in status 0,1,2,3 will not default, but the probability of prepaid is not associated with the loans' current status. All loans in all status carry the prepaid risk. And if we compare the numbers across the years, we could see that on average, loans are more likely to be prepaid in year 3.

Year 1

	Normal	Deliquent-1	Deliquent-2	Deliquent-3	Deliquent-4	Deliquent-5	High Risk	Prepaid	Default
Normal	0.99	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00
Deliquent-1	0.40	0.40	0.10	0.04	0.01	0.00	0.00	0.05	0.00
Deliquent-2	0.20	0.16	0.30	0.27	0.01	0.00	0.00	0.04	0.00
Deliquent-3	0.12	0.07	0.07	0.24	0.47	0.00	0.00	0.03	0.00
Deliquent-4	0.11	0.05	0.02	0.03	0.22	0.54	0.01	0.03	0.01
Deliquent-5	0.09	0.03	0.01	0.01	0.02	0.22	0.54	0.04	0.04
High Risk	0.00	0.03	0.01	0.00	0.01	0.01	0.84	0.04	0.07
Prepaid	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Default	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Year 2

	Normal	Deliquent-1	Deliquent-2	Deliquent-3	Deliquent-4	Deliquent-5	High Risk	Prepaid	Default
Normal	0.95	0.01	0.00	0.00	0.00	0.00	0.00	0.04	0.00
Deliquent-1	0.40	0.40	0.10	0.01	0.00	0.00	0.00	0.06	0.00
Deliquent-2	0.20	0.19	0.26	0.29	0.01	0.00	0.00	0.05	0.00
Deliquent-3	0.12	0.08	0.09	0.16	0.50	0.00	0.00	0.04	0.00
Deliquent-4	0.12	0.04	0.02	0.05	0.14	0.56	0.00	0.04	0.01
Deliquent-5	0.10	0.03	0.01	0.02	0.03	0.15	0.59	0.04	0.04
High Risk	0.00	0.02	0.01	0.00	0.01	0.01	0.84	0.04	0.08
Prepaid	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Default	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Year 3

	Normal	Deliquent-1	Deliquent-2	Deliquent-3	Deliquent-4	Deliquent-5	High Risk	Prepaid	Default
Normal	0.93	0.02	0.00	0.00	0.00	0.00	0.00	0.05	0.00
Deliquent-1	0.35	0.44	0.15	0.00	0.00	0.00	0.00	0.06	0.00
Deliquent-2	0.20	0.20	0.28	0.28	0.01	0.00	0.00	0.05	0.00
Deliquent-3	0.10	0.07	0.09	0.18	0.50	0.00	0.00	0.04	0.00
Deliquent-4	0.10	0.04	0.03	0.05	0.16	0.55	0.00	0.04	0.02
Deliquent-5	0.10	0.03	0.01	0.02	0.05	0.17	0.50	0.04	0.04
High Risk	0.00	0.01	0.01	0.00	0.01	0.01	0.85	0.03	0.07
Prepaid	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Default	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

We also study the influence of a specific covariate to the transition probability. We study the influence of credit score, unemployment rate, HPI, interest rate, Debt to Income ratio, and loan purpose to the probability of defaulting and prepayment, by setting all other covariates the same and only changing the target covariate. Among all covariates stated above, we only find significant influence of credit score and loan purpose on the transition probabilities. Our result shows that for a loan holder who has been delinquent before, as his credit score increases from 500 to 800, the probability of defaulting decreases from 2.65% to 2.55%, and the probability of prepayment decreases from 2.94% to 2.8%. And for the loan purpose, we find that the probability of defaulting and prepayment for a loan holder who purchases the house for investment is 0.01% higher than a loan holder who purchases the house for primarily residential purpose. Please note that this difference is on monthly basis, and 0.01% higher probabilities of

defaulting and prepayment would cause 0.23% more loan terminations per year. Considering the fact that the number of loans is in millions, as well as the great dollar amount of the loans, 0.23% more loan termination could be economically significant.

Conclusion

Both parts of this project show valuable results and achieve our initial goals. In the first part, we provide satisfactory classification models, which would enable Fannie Mae to make initial judgements about the loan performance at the time of acquisition. Besides the first step judgement, more accurately predicted ongoing performance is another direction Fannie Mae would be interested in. So in the second part, we construct the conditional Markov Chain probability matrices which could successfully forecast default and prepayment. With the transition probability matrices, Fannie Mae will be able to calculate the expected gain or loss for any mortgage portfolios. All they need is the number of loans in each status, and the dollar amount of the loans. And then by multiplying the transition probabilities, they could forecast the expected gains or losses in 1 months, 6 months, or 1 year, 2 year, etc.

There are still some improvements that could be performed on our models. Firstly, some covariates used in our model have never been updated since acquisition, such as Debt-to-Income Ratio and credit score. It would be more reasonable to update them based on the economic conditions and the loan holders' payment history in order to provide more accurate information. Besides, when examining the influence of the covariates on the transition probabilities, some of them show minimum influence, which fail to reflect the real-world situation. There are several potential explanations. Firstly, our variables are not on the same numerical scale, which makes the logistic regression coefficients non-comparable. To solve this problem, normalization can be used in data preprocessing. Furthermore, there might exist cross effects between some covariates, such as HPI, unemployment rate and interest rate. Therefore, the individual influence of these covariates may not be significant, and introducing the cross effects will result in more accurate models.