Fannie Mae Loan Performance Prediction

Kexin Wang[1], Ruoxi Liu[1], Tianqi Cui[2], Ziyan Lin[1]

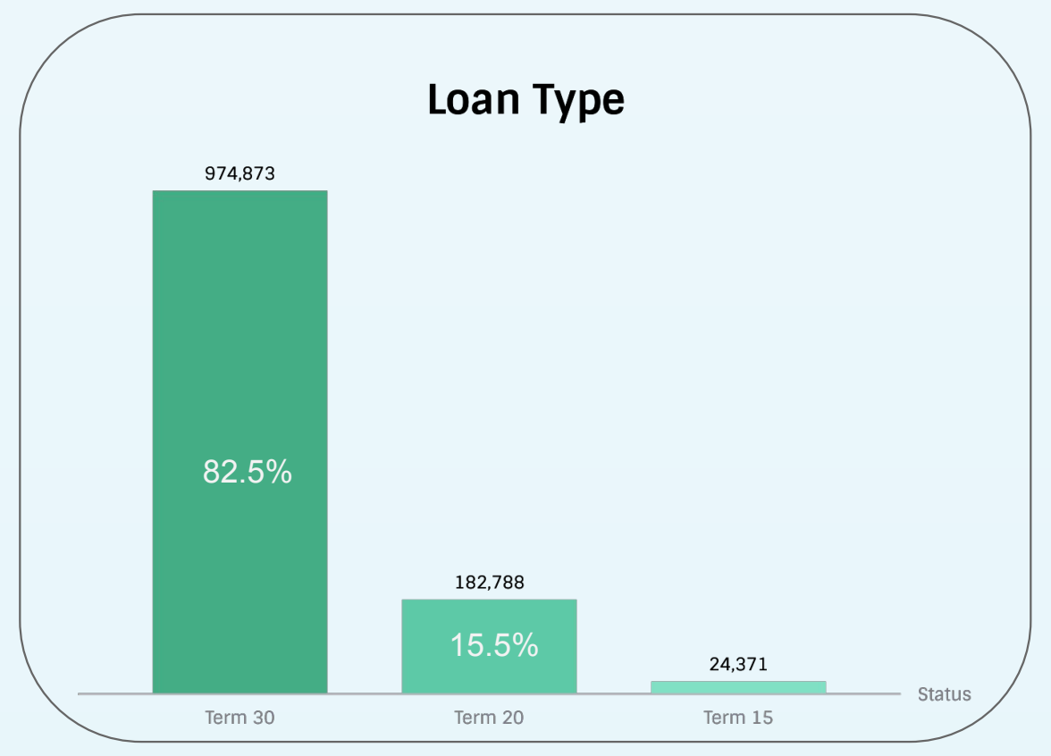
Applied Mathematics and Statistics[1], Chemical and Biomolecular Engineering[2], Johns Hopkins University

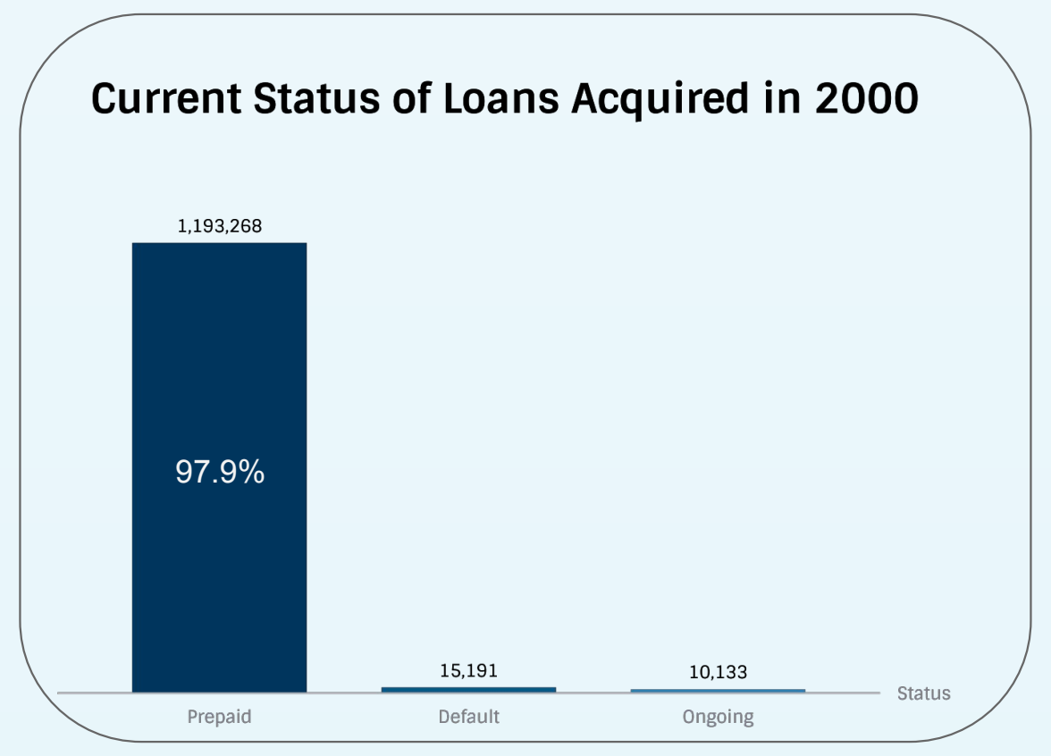
Introduction

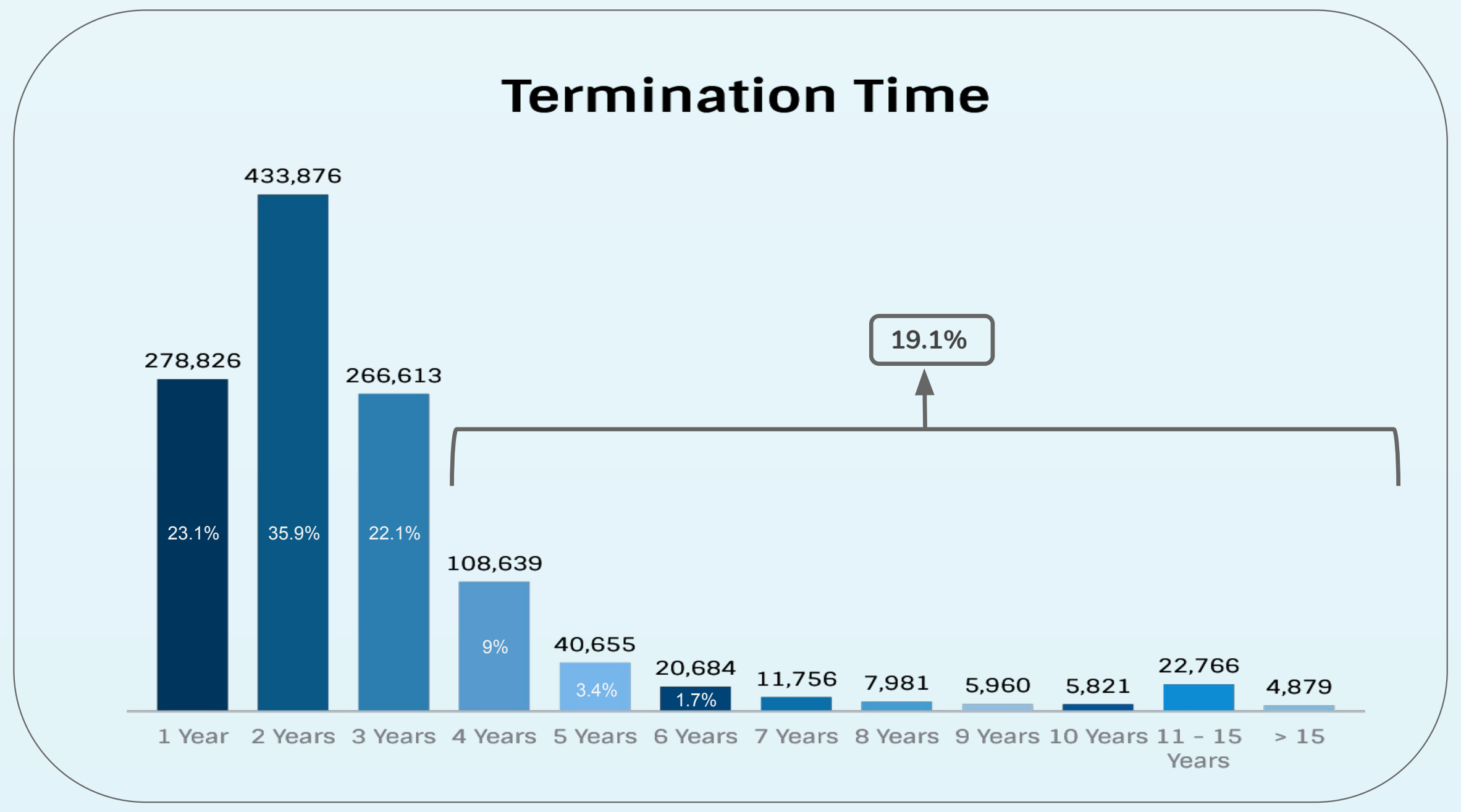
Fannie Mae is created to provide liquidity, stability and affordability to the mortgage market, and thus is important to the economic well-being of the U.S. However, as Fannie Mae acquires mortgages from the banks, it is also exposed to the risks associated with the mortgages. This project is designed 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 an estimation about the status of the loans within three years. The second part "Markov Chain Analysis" aims to derive transition probability matrix of the loan status, based on monthly mortgage payment information. The goal is to provide relatively accurate probability, which could be used to compute the expected future gain or loss related to certain loans.

Data Exploration







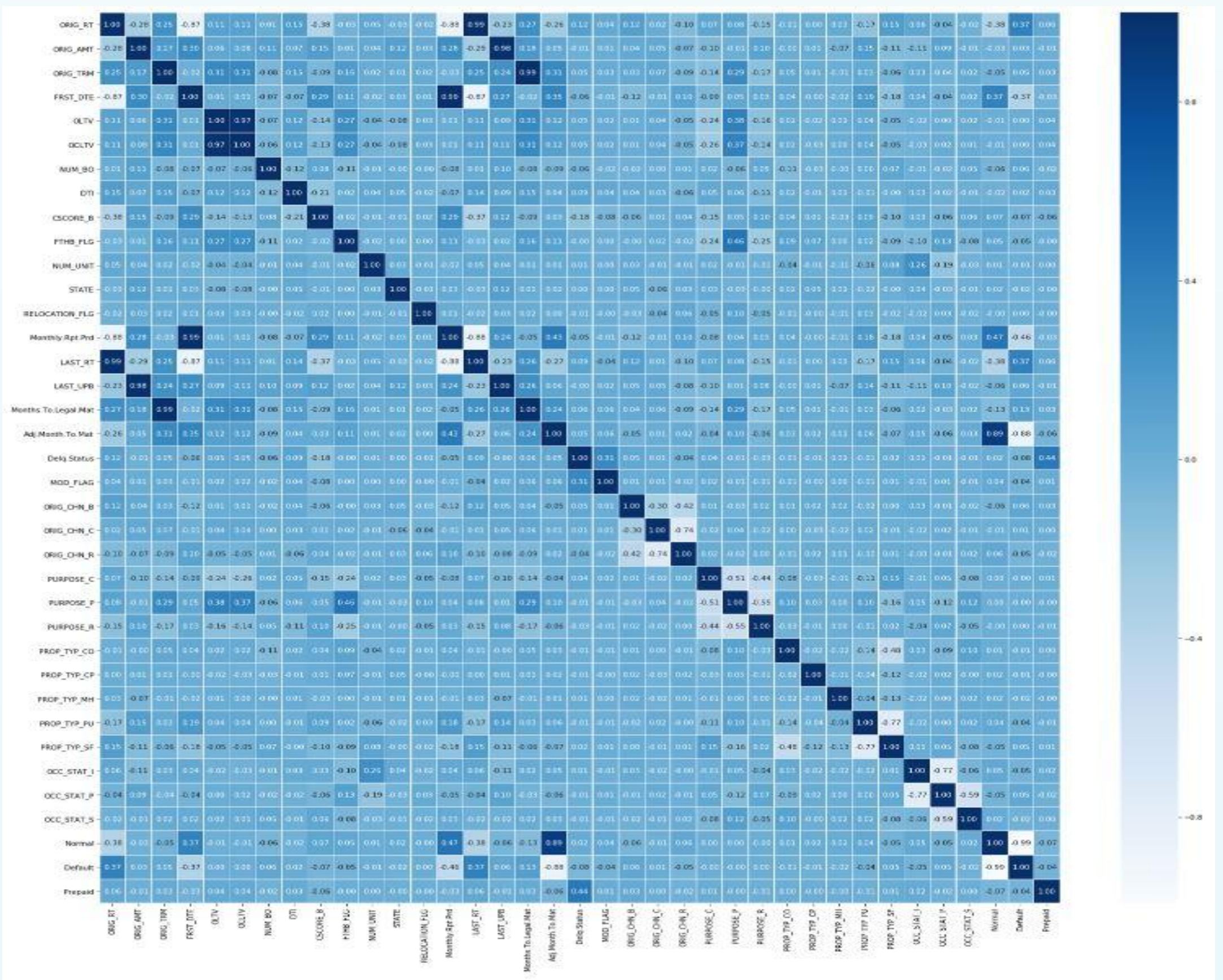
The Fannie Mae loan dataset includes loan acquisition data and performance data from 2000 Q1 to 2017 Q4. The acquisition data contains the basic information of each loan when Fannie Mae acquires them from banks, and the performance data contains the monthly payment or status changes of each loan. We further added in Unemployment Rate and Housing Price Index (HPI) as indicators for economic condition and geographical variation.

In order to understand the trends and get a big picture of the data, we explored the dataset by computing summary statistics. The result charts are shown as above.

Classification

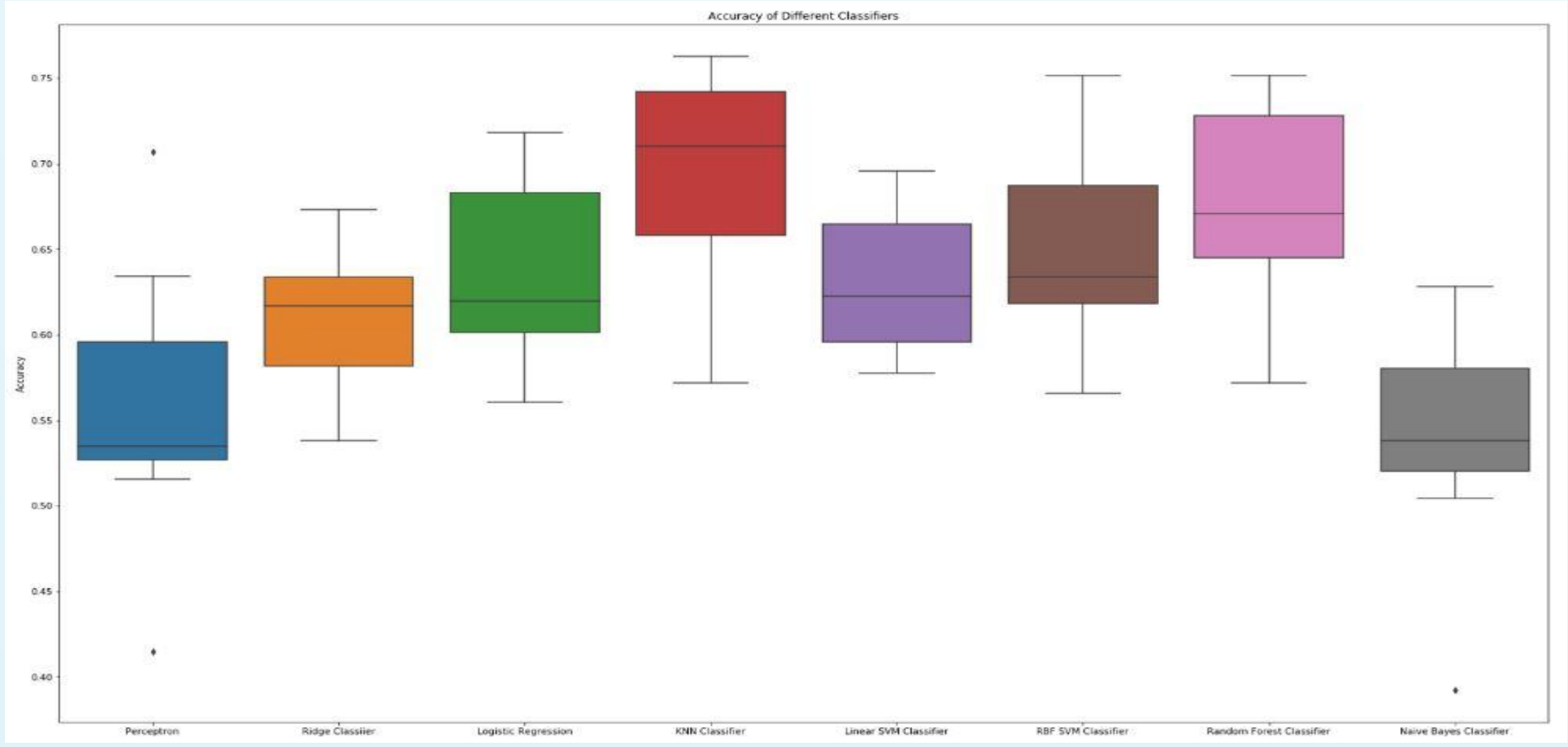
Classification methods are applied to identify categories of loan performance. Zero Balance Code contains information about loan status, and it is treated as target. All other variables are considered as predicted features. In preprocessing step, we delete feature if it miss more than 20% observation and use mode to replace missing values.

The dataset includes many categorical variables and features in datetime format. We set them as dummies and transfer them into numeric variables.



We plot heatmap to show correlation between them. Based on output, we delete features if they have very high or low correlation with target variable, and then we use PCA to double check.

We implement Multi-class classification method with Perceptron, Ridge, Logistic Regression, KNN, Linear SVM, RBF SVM, Random Forest, QDA, and Naive Bayes. The models are evaluated by cross validation, and results are compared in confusion matrix.

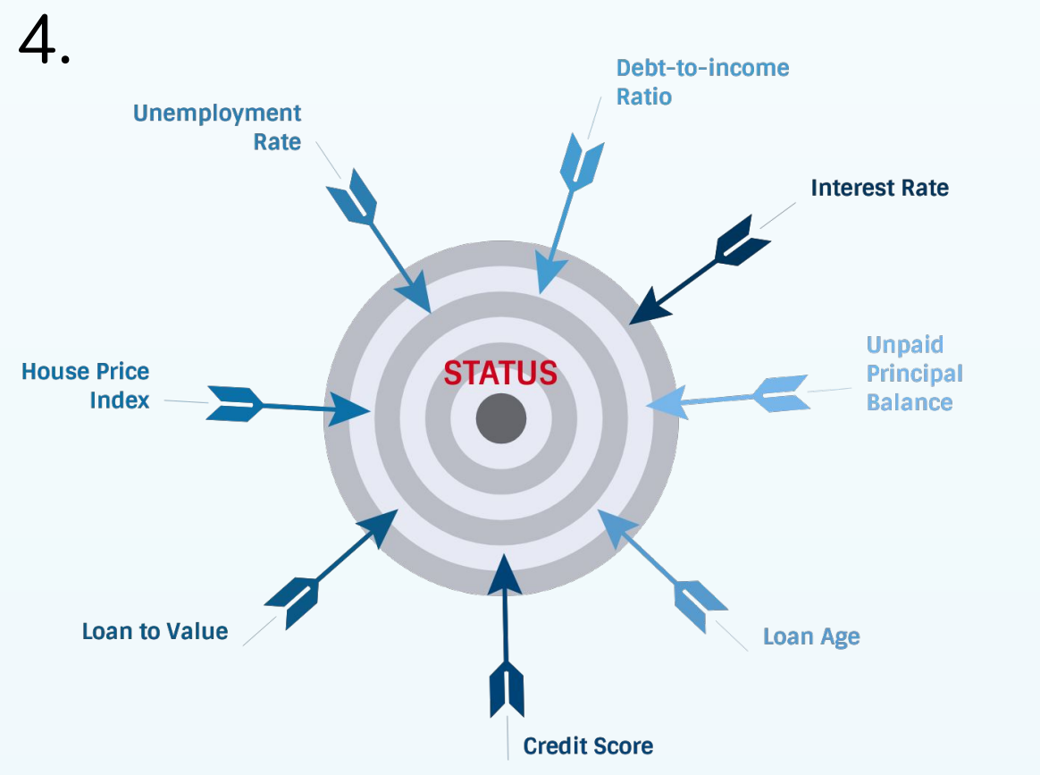


The confusion matrix of KNN classifier

		Predicted Status		
		Normal	Default	Prepaid
Actual Status	Normal	0.821501	0.146045	0.032454
	Default	0.165010	0.643611	0.191379
	Prepaid	0.117647	0.255578	0.626775

Markov Chain Transition - Methodology

Multinomial Logistic Regression is implemented to predict the probability of a loan transferring between the status. Based on "Zero Balance Code" and "Delinquency Status" , we define 9 status : Prepaid, Default, Normal, High Risk and Delinquent 1 to 4.



Since we train the model on 1-year dataset, the transition probability will be valid for the corresponding year. Thus, given the features of a loan and its present status, our model will be able to predict the following status.

Markov Chain Transition Matrix

We used Year 2000 data to fit multinomial logistic regression models, then tested the models' performance with data of 2001, 2005 and 2015, in order to see if the models have generality.

The in-sample testing and out-of-sample testing accuracy (for the first year models) are shown as the followings:

	2000	2001	2005	2015
Model 0	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%

Across the years, the performance of the models is pretty consistent, showing that the models trained with 2000 data are applicable to all other years. However, there are certain models have low accuracy (Model 1 - 3), and that is mainly caused by the complexity of the intermediate status. The remedy approaches will be discussed in later section.

The transition matrices based on the models are then calculated:

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

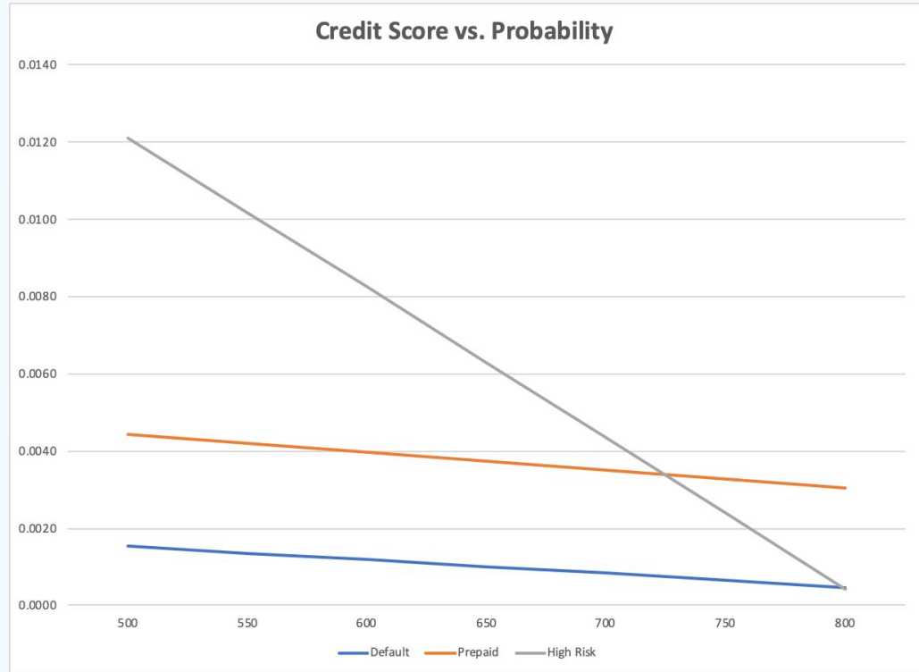
Year 3

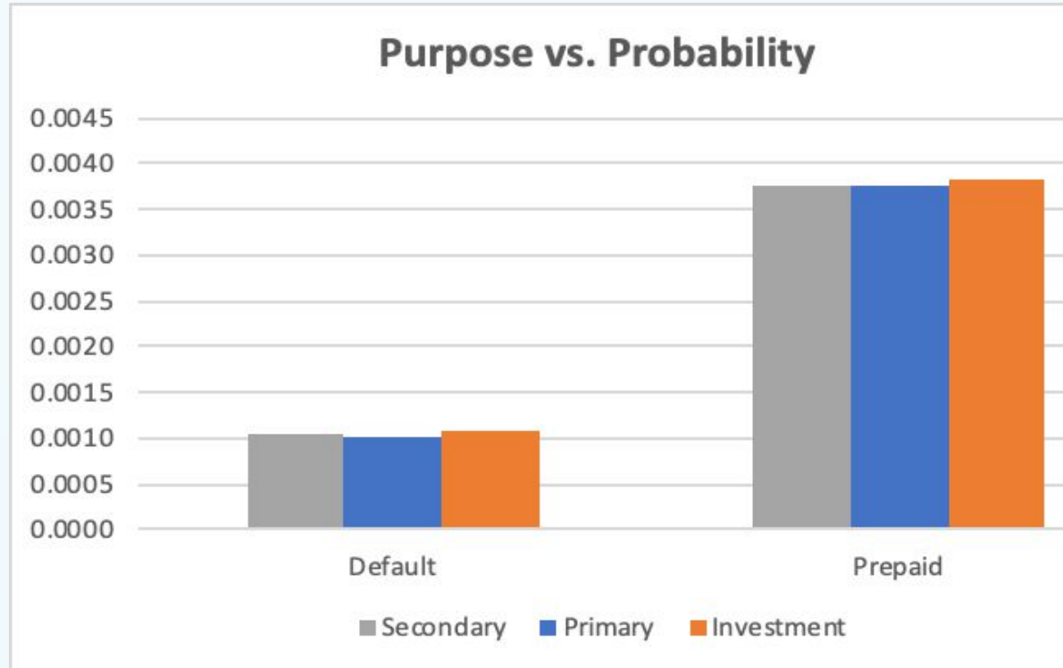
	Normal	Deliquent-1	Deliquent-2	Deliquent-3	Deliquent-4	Deliquent-5	High Risk	Prepaid	Default
Normal	0.93	0.02	0.00	0.00	0.00	0.00	0.00	0.05	0.00
Deliquent-1	0.35	0.44	0.15	0.00	0.00	0.00	0.00	0.06	0.00
Deliquent-2	0.20	0.20	0.28	0.28	0.01	0.00	0.00	0.05	0.00
Deliquent-3	0.10	0.07	0.09	0.18	0.50	0.00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

** Probabilities are for monthly transition*

Probability Change

In order to figure out the influence of a specific covariate, we set other covariates the same, and only change the target covariate. We study the influence of credit score, unemployment rate, HPI, interest rate, debt to income (DTI), and loan purpose on the transition probability, and only find two covariates have significant effect.





The model is fitted by 2000 Year 1 data. The trends prove that our model is a strong prediction tool. Credit score is perfectly negative on the probability of moving to high risk status. Besides, the higher the credit score is, the less likely the loan will be default or prepaid. This is an accurate reflection of the the real-world financial market.

The right barplot explores the relationship between purpose and probability. Purpose consists of investment, primary and secondary. Even though there is not a big difference between their monthly transition probability, mortgage for investment purpose do show higher probability to be default and prepaid than the other two kinds. And such a difference will have significant impact if accumulated in the long-run.

We also plot other covariates' influence but it seems that the trend is not obvious. Firstly, our variables are not on the same numerical scale. To solve this problem, normalization can be used in data preprocessing. Additionally, the original monthly probability is small, so the change is also insignificant over our prediction time period. 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.

Future Work

In this project, we develop models to identify and predict prognostic factors that are strongly associated with the performance of a loan.

There are lot of things we could try to improve our models. Firstly, some variables, such as debt-to-income ratio and credit score, are collected at the acquisition time, and have not been updated. We could update them according to a loan holder's payment history, interest rate and HPI to bring these information up to date. Besides, unbalanced data is a common problem in both parts, which means we need to consider a better weighting procedure. In addition, we believe that the variables in model 1, 2 and 3 do not provide enough information to the model to make accurate judgement. Adding in more variables, such as dummy variables to indicate the borrowers' previous credit or past delinquency records will be definitely meaningful.