A MINI-PROJECT ON MORTGAGE ANALYSIS

JIA(KEVIN) HE

Jia.he@tamu.edu

Part 1. Connect to database and get the data.

To read the data from an online MySQL database, two steps are required. The first step is to generate a connection. In this case, SQLAlchemy is selected, which is a Python SQL toolkit allowing users to communicate with the database very flexibly. The second step is to commit the commands after connecting to the database successfully. Module Pandas has the package which support to read SQL commands.

The data from the database has the following structure

Table 1. Data from database

Total Rows	20,000
Total Columns	19
Column Names	first_pmt_date', 'age', 'status', 'first_time_ho_flag', 'msa_code', 'mi_pct', 'credit_score', 'num_units', 'occupancy_status', 'orig_cltv', 'orig_dti', 'orig_upb', 'orig_ir', 'prop_type', 'state', 'zip', 'loan_seq_num', 'loan_purpose', 'num_borrowers'

Part 2. Mortgage status and origination year

This part will try to analyze how mortgage statuses respond to the year when the mortgage was originated.
In excel, the PivotTable allows users to group the data by different attributes. Similarly, in Python, DataFrame has a function called pivot_table which can achieve this goal.
To facilitate the group function, a new column named year is created, which originates from the date of the first monthly mortgage payment

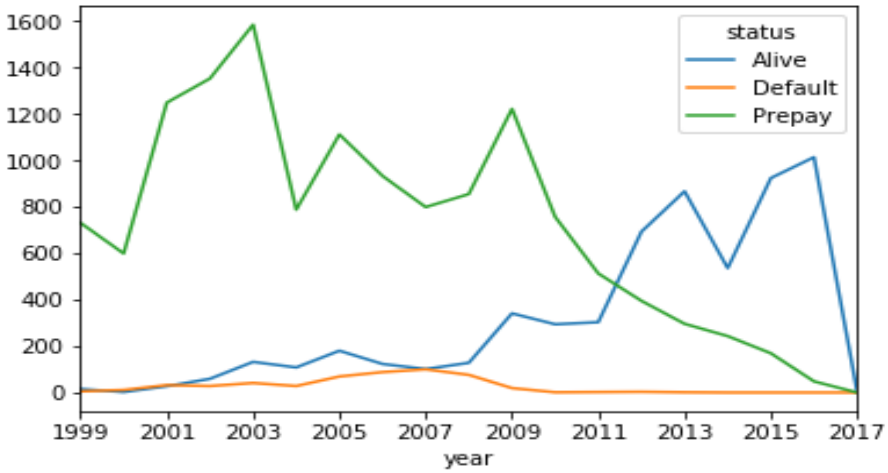

Figure 1 Status and Year

The figure shows the following characters:

1. Overall, the count of mortgage in prepay status is largest, next comes to alive status, and default status has a low rate of occupancies.

2. As time goes by, the number of mortgages in prepay status increases from 1999 to 2003 and is in decline after 2003. Similarly, the number of alive mortgages gradually rises. The number of mortgages in default is very small and stable all across the years. But it is obvious to observe that from 2005 to 2009, the default number is significantly higher, which may be caused by the subprime mortgage crisis.
3. There are zero records in any status in 2017, which means 2017 is the most recent year when the data was collected

Part 3. Mortgage status and age

This part will try to analyze the relationship between the mortgage statuses and its age

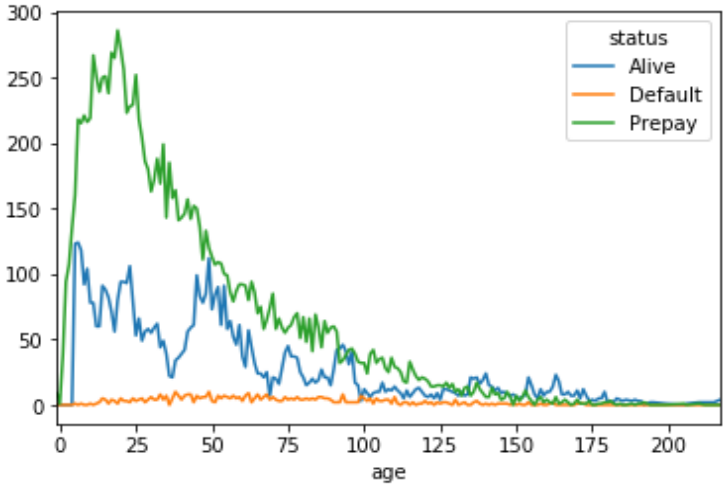

Figure 2 Status and Age

The figure shows these characters:

1. Among the three statuses, most mortgages are in prepay status, and a very small proportion are in default.
2. The homeowners are very likely to prepay the mortgage during the first 0 to 40 months, with the highest probability at around 70%(275/375) at age 25.
3. The total number of records drops as the age increases

Part 4. Determine the unconditional probabilities

Here, the time is limited to the first 5 years, which accordingly means the age is less than 60. Run the following two commands:

Table 2 Status in First 5 Years

Status of mortgage in the first 5 years				
Status	Alive	Default	Prepay	Total
Number	3868	231	10249	14348
Probability	26.96%	1.61%	71.43%	100%

The results in the table show that the default rate in the first 5 years is very small at 1.61%, while the prepaying rate during the same period is as high as 71.43%.

Part 5. Use a model to predict mortgage status

1. Choose the response and factor

Estimating the default rate is a very import part of risk control, so status default is selected as the response. By intuition, it's reasonable to assume the default rate has a positive correlation with factors like Orig_cltv (loan to value) and negative correlation with factors like Credit_score (credit score). The solution in part 2 also shows that the number of default cases depends on the factor AGE, which will be used as the factor in this part

2. Machine learning model – Decision Tree

Identifying whether default or no default will happen is an issue of classification. Techniques used to do classification include logistic regression, K-nearest neighbors and decision tree. In this case, the decision tree is used to build the model, as it provides a very intuitive way to understand how the dependent variable reacts to independent variables.

The total data set is randomly divided into two groups, the training set which includes 80% of all data and test set which has 20%. The train set is used to build the decision tree model.

3. Evaluate the model

A model built based on training data may be underfitting or overfitting, so it is very important to test it with the test set. To evaluate the quality of the decision tree model, 2 methods are used: confusion matrix and decision tree visualization.

1) Confusion matrix

Table 3. Confusion Matrix

Confusion Matrix					**Accuracy Rate**	
		Predicted			True Default Rate	0%
	Total 4000	Not Default	Default		True Not Default Rate	100%
Actual	No Default	3887	0		Accuracy Rate	97.18%
	Default	113	0		Bench (null accuracy rate)	97.18%

The model has an accuracy rate of 97.18%, which is very high. However, it is no better than the null accuracy rate, which comes from the simple model by just predicting the most frequent status. As well, the result shows that the model can predict the true NOT DEFAULT 100% correctly while it cannot predict the DEFAULT status. The reason is that the default status only accounts for 2.4% proportion of the sample data. Using the majority vote, the model would vote all situations to NOT DEFAULT. (one way to solve this issue is using upsampling, but it may cause overfitting, so not used here)

2) Decision tree visualization

The rates above provide a general idea of how good the model is, the decision tree structure then will show how to apply it in a more practical way. The image below is a 3-level decision tree.

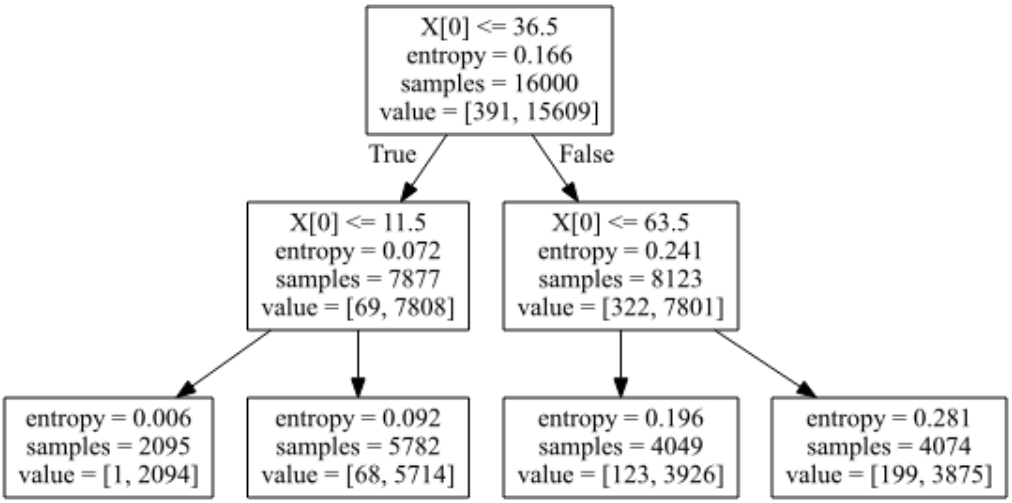

Figure 3 Decision Tree Structure

The decision tree above provides a distribution of default probabilities in different AGEs.

Table 4. Default Probability Distribution

Range of Age	Default	Not Default	Default Rate
0 - 11.5	1	2094	0.05%
11.5-36.5	68	5714	1.18%
36.5-63.5	123	3926	3.04%
63.5 or higher	199	3875	4.88%

So, the result shows that when the mortgages are within 11-months old, the owners are very unlikely to default as the default rate is only 0.05%. However, when the mortgages are more than 63-months, the default can be as high as close to 5%.
From the perspective of mortgage lenders, this information is really important for them to do risk control.

4. Mortgage price and investment strategy.

Financial investments always involve two issues: risk and return. A higher default rate increases the risk, and thus a higher return is expected. If the mortgage is packaged and traded on the market, the price will accordingly be lower.

Based on the result from the decision tree model, I think the following strategies should be taken into consideration, especially from the perspective of institutional participants like Unison.

1) Diversity investments in mortgages of different ages to reduce the idiosyncratic risk
2) Risk control should incline to mortgages that have a higher default rate. To be more specific in this case, more focus can be put on the mortgages with age larger than 36-months
3) Buy insurance to reduce the risk for those mortgages with a higher default rate if necessary.