Kexin Zhu

# MEMO

This report centered on two analyses:

- Descriptive comparisons of approved loans between Illinois and California in 2020
- Mortgage application status prediction model

Main findings summarized here (detailed how conclusions arrive see following pages)

**Part I**

Designed characteristics matrix is composed of 8 variables from three perspectives (state-wise; transaction-wise and customer-wise). Among approved loans in 2020, California (CA) and Illinois(IL) shared some commons:

- The two approval rates around 76% are statistically same
- A three-year period loan term as majority
- The most common purpose that people come to resort is Refinancing
- The rank in common loan type that customer is looking for is following Conventional, FHA, VA and others
- People aging from 35-54 is mainstay (45%)
- White people take the majority place in approved loan (around 85%)
- The median income of customer with approved loan is roughly the same around $110,000/yr

Also reveal variances:

- Average loan amount in California(mean=451,472.95) is an obvious high compared with Illinois'(mean=248,850.21), with a wider range(std)
- The average interest rate in CA is slightly lower than IL
- IL residents are more willing to take out a loan for home purchase than in California, where more people use it for cash-out refinancing.
- Middle-young people (25-34) have higher loan demand than middle-elders (55-64) in IL compared to CA's.
- Customer backgrounds are more varied in CA than IL, in terms of their race and income

**Part II**

After data preprocessing and feature engineering, Logistic Regression Classifier and AdaBoost Classifier were trained separately. Both shows a strength in Approval case, with high precision (avg 0.97) and recall (avg 0.96). While AdaBoost outperforms Logistic regression for Reject case after tunning learning rate to 0.6, which improved Recall from 0.62 to 0.91.

Data source: One Year National Loan Level Dataset

# Comparisons of CA and IL 2020 approved loans

In this report, comparative analysis is centered on comparing approved California and Illinois loans as comprehensively as possible.

- At the high abstraction level, how was loan itself.
- From the perspective of the transaction, what are some variations of main characteristics among the two states.
- From a customer perspective, who are they; what drives them come to mortgage; what are they looking for.

Indicators below were utilized to answer the above questions and followed with summaries.

Table 1: Variables used in comparative analysis

| State-wise | Transaction-wise | Customer-wise |
|---|---|---|
| loan_amount | loan_purpose | applicant_age |
| interest_rate | loan_type | income |
| loan_term | | derived_race |

**Note:** To curve the underlying pattern, no imputations were implemented at this point. The missing values would be ignored when producing summary tables or plots.

## State-wise comparisons

1. Approval rate

The proportions of approved loans in two states are roughly around 76%, fairly to say they are the same. Also, a Chi-Squared test result further supports the statement with a 99.96% p-value, which means that we do not reject the null hypothesis that getting loan approval is independent with being which state.

*Table 1: State-wise loan action taken contingency*

| action_taken | Approve | Reject | withdrawn |
|---|---|---|---|
| state_code | | | |
| CA | 0.755083 | 0.101167 | 0.143751 |
| IL | 0.768435 | 0.101337 | 0.130228 |

2. Loan amount and Interest rate

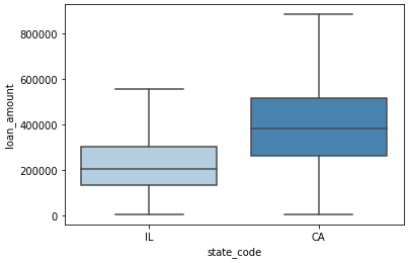


Fig 1. State-wise loan amount

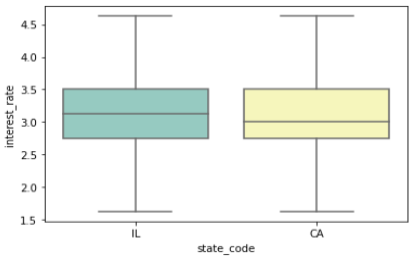


Fig 2. State-wise loan term

From boxplots above (Fig 1 and Fig 2), the average loan amount in California(mean=451,472.95) is an obvious high compared with Illinois'(mean=248,850.21), also a wider range(std) of the approved amount. The interest rate in California approved case is slightly lower than Illinois, around 3.0

3. Loan term

*Table 2: State-wise loan action taken contingency*

| state_code | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| CA | 2561416.0 | 332.541187 | 221.432701 | 1.0 | 360.0 | 360.0 | 360.0 | 339000.0 |
| IL | 634589.0 | 311.212733 | 82.813307 | 1.0 | 240.0 | 360.0 | 360.0 | 3660.0 |

The approved loan term in California and Illinois is comparatively same, with a three-year period majority, while there are some extreme long-term cases in California been seen.

## Transaction-wise comparisons

1. Loan_purpose


*Fig 3. Word cloud graphs of loan purposes in CA(left) and IL(right)*

The most common purpose that people come to resort is Refinancing, for both CA and IL. "Home purchase" takes the second largest proportion in Illinois and ranked higher in IL than in CA.

2. Loan type

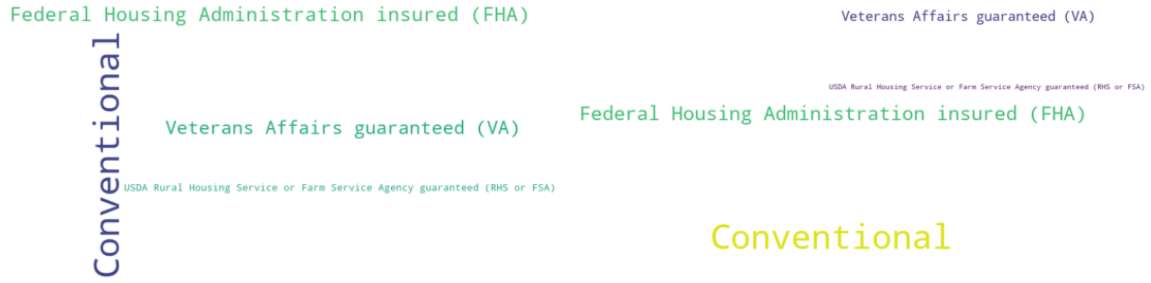


*Fig 4. Word cloud graphs of loan type in CA(left) and IL(right)*

The rank in common loan types customer are looking for is the same both in CA and IL, follows Conventional, FHA, VA and others.

## Customer-wise comparisons

1. Age

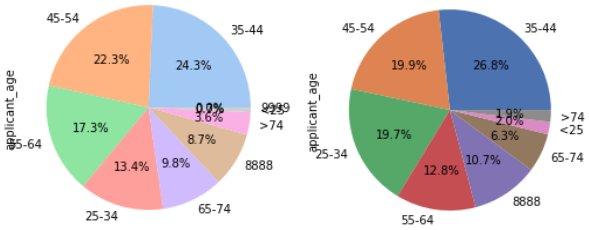


*Fig 5. Pie charts of customer age in CA(left) and IL(right)*

People aging from 35-54 is mainstay (45%) among customers with approved loan both in CA and IL. Middle-young people (25-34) have higher loan demand than middle-elders (55-64) in IL compared to CA's.

2. Race

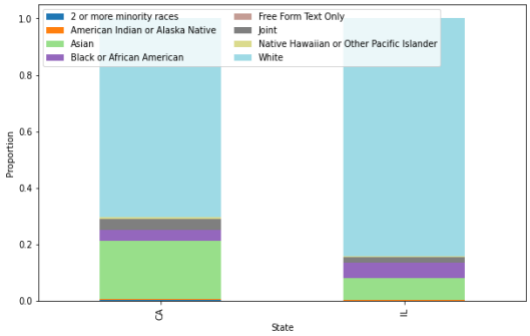


*Fig 6. Stacked bar charts of customer race in CA(left) and IL(right)*

White people takes the majority place in approved loan market both in IL (84.29%) and CA(70.40%). From another perspective, California's plate is more diverse with a higher proportion of other races.

3. Income

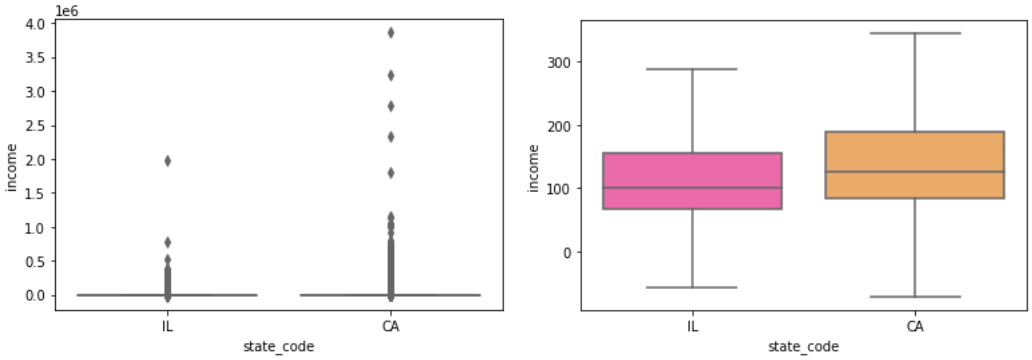


*Fig 7. Boxplots of customer income in CA and IL, original(left) and outliers_excluded(right)*

The median income of customer with approved loan is roughly the same around $110,000/yr. California income has a higher average($753.000/yr) and has a long-tailed distribution(larger variation and extremely large incomes exist)

# Mortgage application status prediction model

## Data wrangling

1. Missing value process

   The original one-year Loan-Level dataset for Illinois 2020 contains 99 related variables and 873,719 records in total. Before stepping into modeling phase, first check out the variables with missing values.

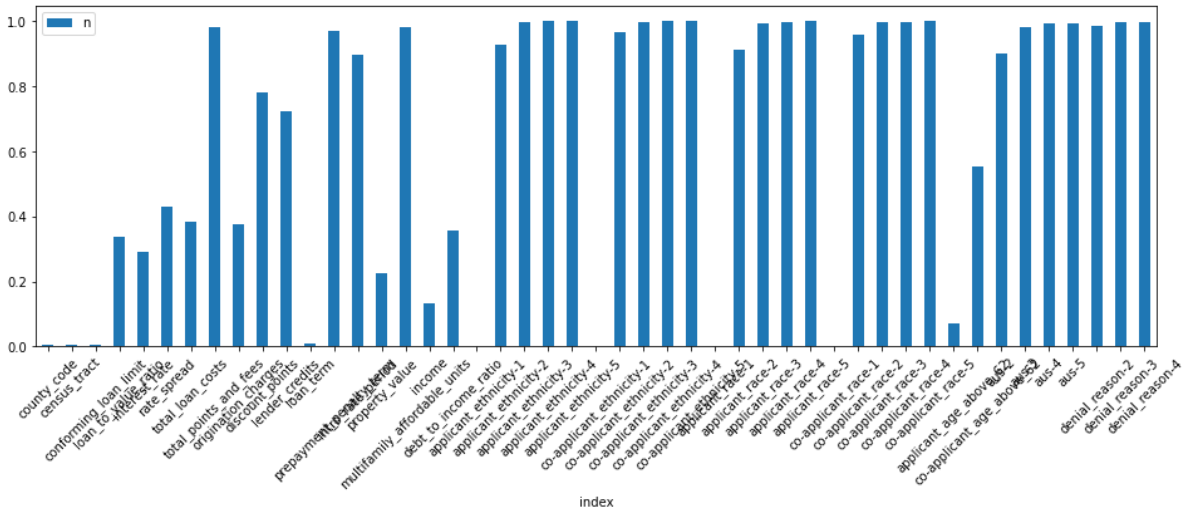


*Fig 8. Bar charts of missing value proportion in variables*

According to the modeling needs, practical meanings, and proportion of missing values, here list out the solutions, respectively.

*Table 3: Missing value solutions*

| Category | Action | Example | Count |
|---|---|---|---|
| At, or after loan that do not affect modeling | Delete | applicant_age | 10 |
| Additional Information / Already Integrated | Delete | income | 30 |
| Imputable variables of interest(low missing rate; Missing Completely at Random) | Imputate with median(numerical) or mode(categorical) | derived_race | 7 |

*Note: Considering the characteristics of loan and customer, hot-deck imputation may outperform mean/median replacement in some cases, while it would be a computational burden.*

2. Modeling data preparation

   Now there is a complete dataset with 59 variables. Next, targeting at make data usable for establishing models. The main actions summarized as below:

   - Utilize 3-sigma rule to leave out outliers
   - Factorize string columns categorical or ordinal
   - Format numerical variables

## Logistic Regression

1. Modeling Assumptions

   **What is the population of interest?**

   All mortgage application

   **Identify the features of interest**

According to Elul(2010)'is study, variables used in this stage are: 'action_taken', 'debt_to_income_ratio', 'loan_to_value_ratio', 'income', 'hoepa_status', and 'loan_amount'.

**Identify the population parameter**

Suppose the relationship between the logit function of conditional probability (mortgage being approved) could be represented as a linear model:

$$logit\,\mathbb{P}(Mortgage\ application = Approve \mid debt\_to\_income\_ratio = x_1, loan\_to\_value\_ratio = x_2,\ income = x_3, hoepa\_status = x_4, loan\_amount = x_5)$$

2. Model fit and results

In general, the variance between loan type(product) is taking different factors into decision-making. Here pick up the most common type Commercial as example. The entire dataset is 576,869 rows, with one dependent variable(action_taken) and other six independent variables.

- Randomly form a 7:3 training and test set from the dataset.
- To avoid predictions dominated by variables with too much dimensionality, normalize predictors before training the model.
- The train test split, model fit and evaluation are done within Scikit-learn.

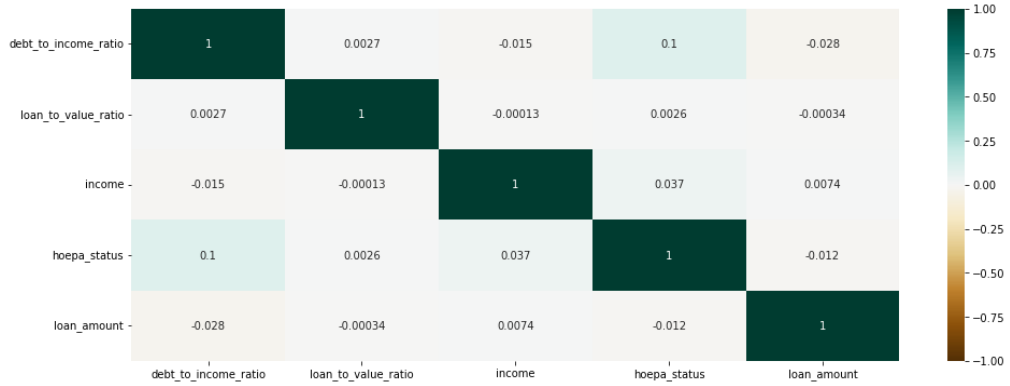


*Fig 9: Heatmap of correlation between independent variables*

From the heatmap, there is no evidence showing the model would suffer from multicollinearity.

*Table 4: Confusion Matrix of logistic regression*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.95 | 0.97 | 0.96 | 151984 |
| 1 | 0.72 | 0.62 | 0.67 | 21077 |
| accuracy | | | 0.92 | 173061 |
| macro avg | 0.84 | 0.79 | 0.81 | 173061 |
| weighted avg | 0.92 | 0.92 | 0.92 | 173061 |

From the table above, the logistic regression shows a satisfying overall accuracy 92%. The model performs well for Approval case(label=0) prediction with 95% precision and 97% recall. While in terms of Reject case(label=1), still leave room to be advanced. (Recall=0.62)

Out of the risk control's need, an improvement on Reject case (label 1) should be on the plate. The R-square of fitted logistic regression is only 0.2953. This signal warns that logistic regression may not be good for linear non-separable case.

## AdaBoost Classifier

The boosting algorithms are tree-based and are suitable for the concept of information gain or Gini impurity. They perform well on datasets with no linear relationship. In this sense, here take AdaBoost classifier as enhancement, with Decision Tree as base estimator. Testing learning rate ranges from 0.1 to 1, 0.1 as on step, and take 0.6 which gives a highest recall on Reject case (label 1).

*Table 5: Confusion Matrix of logistic regression*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 0.94 | 0.96 | 151984 |
| 1 | 0.68 | 0.91 | 0.78 | 21077 |
| accuracy | | | 0.94 | 173061 |
| macro avg | 0.83 | 0.92 | 0.87 | 173061 |
| weighted avg | 0.95 | 0.94 | 0.94 | 173061 |

Compared with Logistic regression, AdaBoost shows an obvious improvement in f1-socre of Reject case (label 1) and continue with a satisfying performances on Approve case (label 0). There is no one-fit-all model but the optimal choice under a trade-off. Here put higher priority to the Reject case (label 1), AdaBoost will be the final model to go.

[i] Elul, R., Souleles, N. S., Chomsisengphet, S., Glennon, D., & Hunt, R. (2010). What" triggers" mortgage default?. *American Economic Review*, *100*(2), 490-94.