Cornell University

Who get credits

ORIE 4741 Final Projects

Prepared by

Jie Zhu, Chuyang Zhang

December 2, 2021

Summary

One of the largest business sectors for any financial institution is mortgage. A mortgage lender is a financial institution or mortgage bank that offers and underwrites home loans. Lenders have specific borrowing guidelines to verify applicants' creditworthiness and ability to repay a loan. After the borrower submitted the application, the lender needs to access multiple conditions of the applicant to avoid the risk of any borrower fails to repay. Any decision to approve or deny an applicant needs to be evaluated very carefully, and these decisions need to be fair enough.

In our project, we aim to explore the feasibility of using models to help human reviewers predict applicants' reliability, so that further decisions based on model predictions will be more precise and fairer. First, we performed data analysis on a mortgage dataset to explore some basic features of data, then we conducted data processing and feature engineering to improve the data quality. Next, we used some linear models and ensembled tree models to fit data, we did some experiments in the process to discuss choice of regularization, SVM kernels and ensemble methods. Finally, we interpreted these models to diagnose unfairness, and talked about why it exists and how to solve it.

Dataset

The FFIEC Home Mortgage Disclosure Act (HMDA) provides mortgage data given year, geography and institution. This project is based on 2020 California mortgage dataset from 21ST MORTGAGE CORPORATION (institution code 549300XQVJ1XBNFA5536). The dataset has 11129 recordings of mortgage application within each 99 data fields are provided.

Data Filtering

We did some filtering operations on both axes to make sure only useful data is left. Some applications are withdrawn by users and some are closed for incompleteness, these applications are not normally finished and no decisions are made to them, so we will not use these data and they are removed.

Some data fields should be removed, too. There are four kinds of data fields we need to drop:

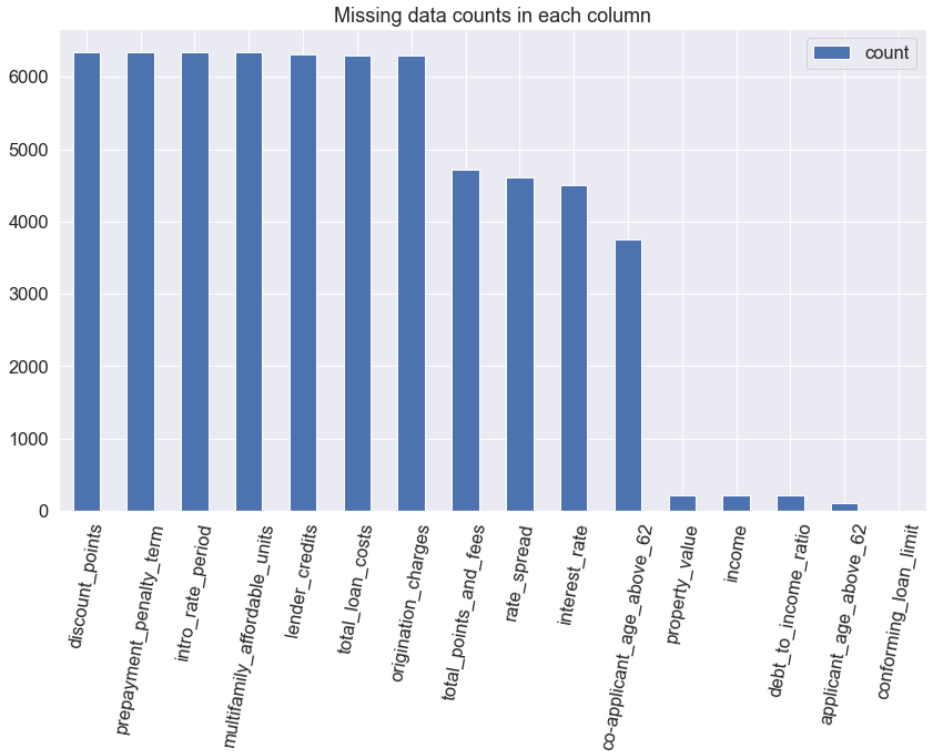
1. **Data fields that only have one distinct value.** For example, activity_year in this dataset has a distinct value of 2020, since there is no difference among samples on this field, it is not necessary to keep.
2. **Data fields that have too much distinct value.** For example, census_tract has 2760 distinct values, and one-hot encoding this feature will create 2760 additional dimensions, since huge

dimension here will leads to overfitting because the number of samples are not big enough to train such a complex model.

3. **Data Leakage.** The target variable is the action taken to the application, and denial_reason field in this dataset leaks the information of taken action. Keeping these fields are unnecessary because when the model is used for estimating before final decision is made, we cannot get these fields.
4. **Other unused data fields.** All data fields related to the census tract are not used, this is simply because they are too general and provide no information about the applicant himself.

Handle Missing Data

Sometimes the missing of data is informative and sometimes not. In our dataset, there are 16 columns that contain missing data, some columns have too much missing data like discount_points, these columns have more than half of the data missing, we can simply drop them because they provide very limit information. For others. we can either remove the rows, impute the value, or treat missing as value. In our case, we remove the rows because only 210 out of 6346 rows will be dropped, which is acceptable.

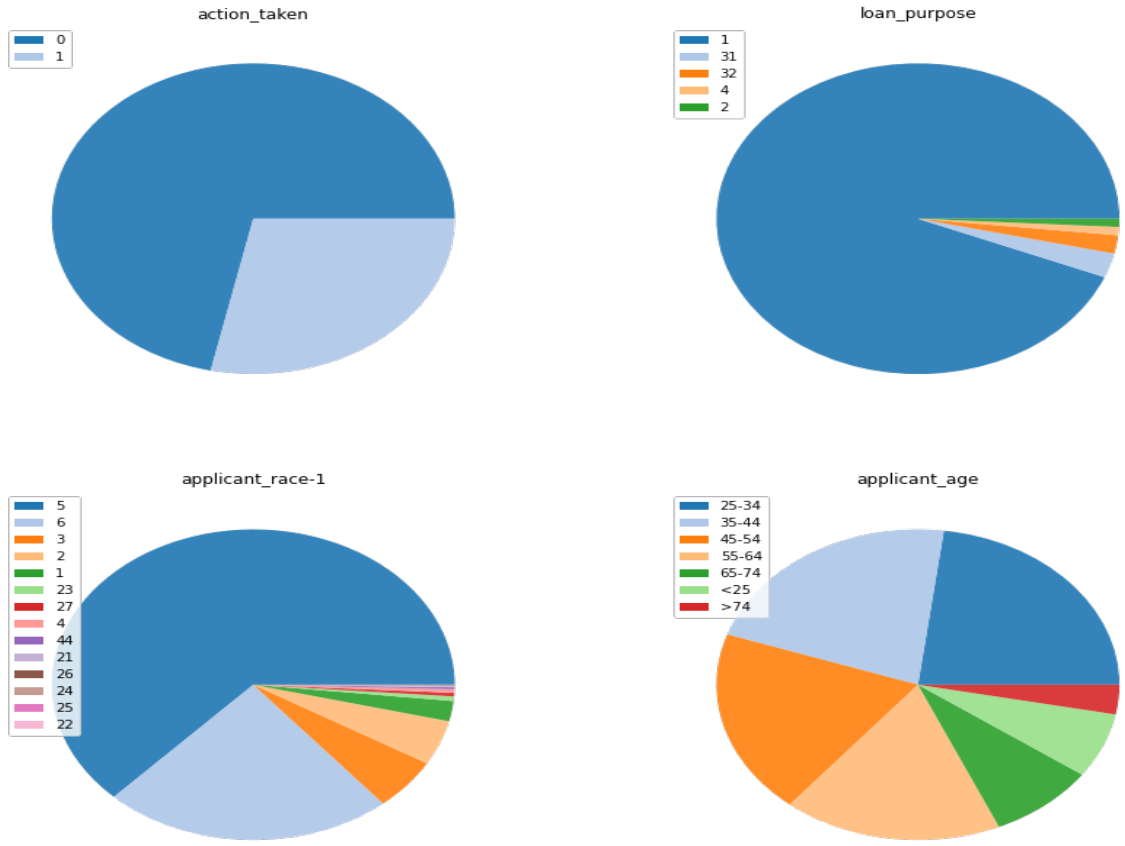


After preprocessing, the new dataset has 6136 rows and 39 columns, these data are purer and more valuable compared to the original dataset, and further results of data analyzing and modelling will be more convincing.

Primitive Analysis

Distribution of some properties

First step to understand the data is to look at distribution of its properties. Here we look at the distributions of action taken to application, the loan purpose, and applicant's race and age. The explanation of values can be found at here.

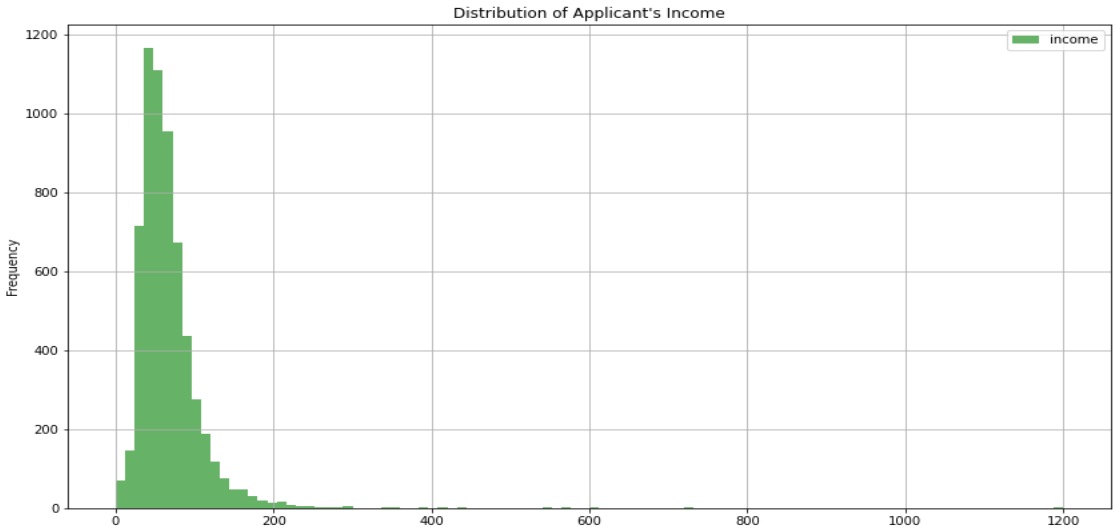


From the figure above, we noticed that

1. Most applications are denied, the percentage is about 70%
2. Most applicants loan for home purchase purpose
3. Most applicants are while and age from 25 to 54

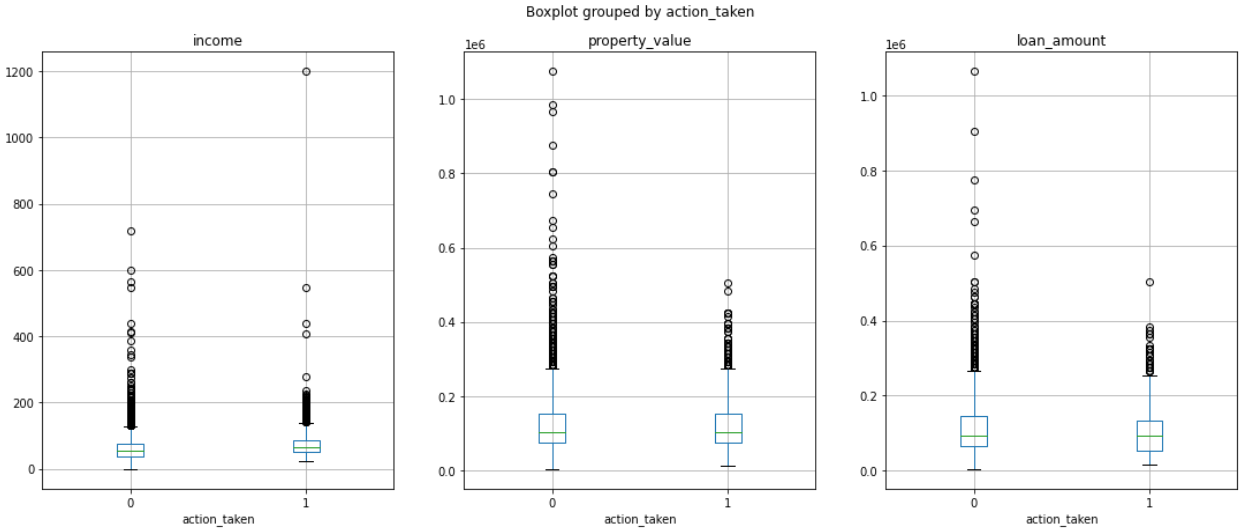
The applicant's income is important in mortgage application, and since it's a continuous variable, we choose to plot the histogram to observe its distribution.

We can conclude that the income nearly follows Poisson distribution, and most applicant's gross annual income is under 200,000 dollars, around 50,000 dollars. The poorest applicants have no income, and the richest have over 1,000,000 dollars each year, this is a very wide range and applicants with different income may be treated differently.



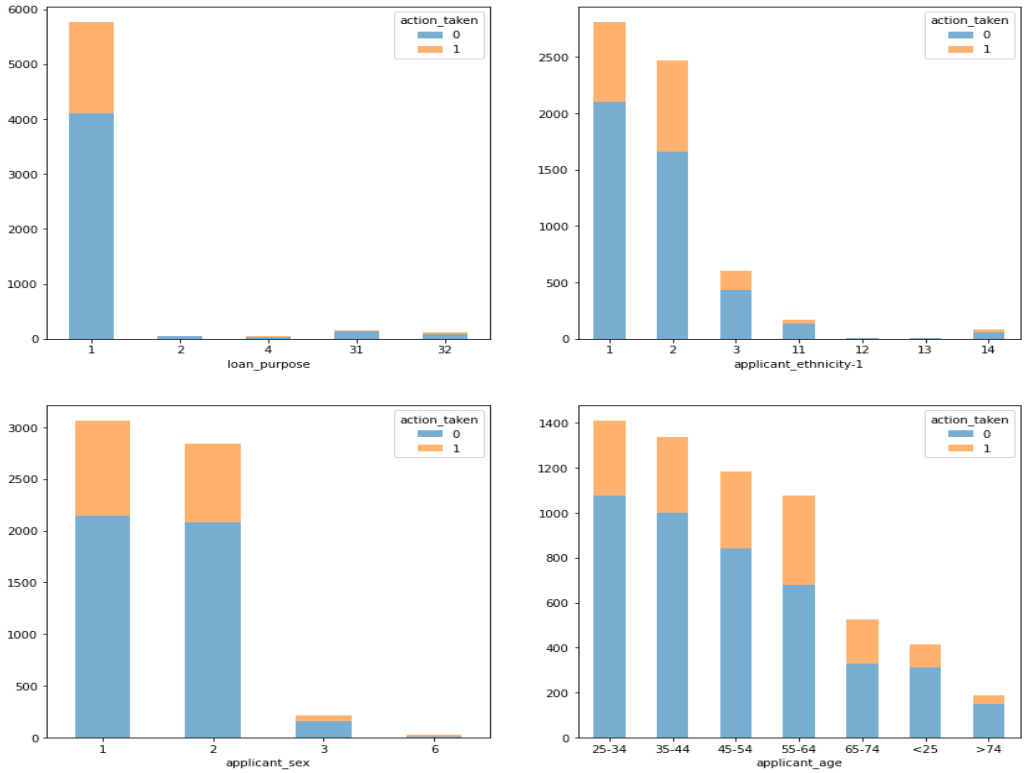
Bivariable analysis

To compare the income, property value and amount of loan of approved applicants and denied applicants, we generated these boxplots as follow.



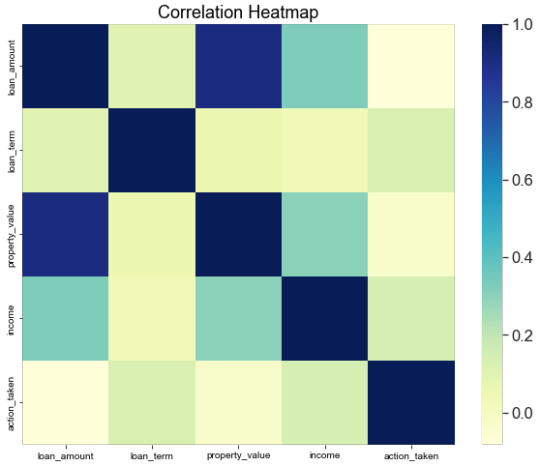
Accepted applicants usually have higher income, higher property value and less loan amount. This pattern reveals preferred feature of borrowers, because higher income and property value indicates the applicant can afford more loan repayment, and less loan amount have lower risk of unsuccessful repayment.

It is reasonable to evaluate applicant's ability to repay in the application, however, does other nonfinancial factors matters? Except loan purpose, we want to find out if ethnicity, sex, age affect the decision.



From the figure above, applicant that are not Hispanic or Latino, male applicant and applicant whose age is between 55-64, have higher probability of approved. The age property is probability related to stability or reliability, but how come sex and ethnicity matters so much? This may reveal unfairness in the application, that lender has bias on certain types of applicants.

Correlation



Correlation is interesting to observe, high correlation means the co-occurrence of certain combination of properties. Here we choose loan amount, loan term, property value, income and action taken as variables, the visualization of the correlation matrix is provided as follow. We noticed that the property value and income are highly related to loan amount, which means rich

people loan more money; and income is highly related to property value, because rich people have more expensive properties.

Predictive Model

Feature Engineering

There are 39 columns, some are numeric and others categorical, different processing are done to these two groups.

For each numeric feature, we first standardize it, then we add two additional features: its square transformed value and bucketized value. The standardization is needed because it can avoid the different scale of features affecting the training. Since we will use linear model later, the square transformation and bucketizing can transform the raw feature in non-linear form, which enables linear model to learn non-linear pattern from numeric features.

For each categorical feature, we simply apply one-hot transformation, so that each category becomes an independent dimension.

Train Strategy

We split the dataset randomly into two parts, 80% as trainset and 20% as test set. For each model, we use 5-fold cross validation to train 5 different instances, the prediction of instances will be mean aggregated into final prediction, which will be used to calculate different metrics with corresponding labels.

The performance of trained model will be tested by test set, and here we use accuracy, AUC and f1-score as metrics. The max iteration is set to 100,000 if the model is gradient-based optimized.

Model selection and performance

Our target variable is action taken to the application, this is either 1-approved or 0-denied, so this is a binary classification problem. We tried three types of models for given reason:

1. **Logistic Regression:** logistic regression is the simplest linear model in classification, its coefficients reveal how features are combined to compute the probability, therefore it is friendly for interpreting.
2. **Support Vector Machine:** similar with logistic regression but support vector machine tries to maximize the margin between the closest support vectors whereas logistic regression maximizes the posterior class probability.
3. **Ensembled Tree Model:** tree-based models are non-linear, and ensembled tree models includes random forest (bagging) and GBDT (boosting).

Moreover, for logistic regression model, we will try LR with no normalization and LR with L-2 normalization (also called Ridge Regression). For SVM, we will try linear kernel and non-linear RBF kernel to see which one is better for this task.

Here is the evaluated performance for each model on test set.

	LR	LR+L2	SVM+linear	SVM+rbf	Random Forest	GBDT
accuracy	0.971	0.980	0.980	0.971	0.970	**0.982**
auc	0.961	0.968	0.969	0.946	0.943	**0.970**
f1-score	0.946	0.963	0.963	0.943	0.940	**0.967**

1. LR with L-2 normalization is better than LR, this means here using normalization that add penalty to weights, LR has better generalization performance, normalization can avoid overfitting.
2. Linear kernel SVM is better than rbf kernel SVM, the pattern in our data cannot be well represented by rbf kernel.
3. GBDT is the best model here in this task. The boosting ensemble method enables GBDT to have smaller training residual error, and tree-based model is great at thinking like human!

Fairness and Improvement

One problem exists in machine learning is that, when data have some bad property, like discriminating, unfairness, fitting a model to such data, and model learned to act just as bad. We have found out that our data shows non-financial factors like sex, ethnicity can affect the probability of being approved, which is of course very unfair, and we can see if a LR model learned that way by looking at its coefficients.

Unfortunately, model is unfair, for example, if applicant age is above 62, the feature of applicant_age_above_62 will be Yes and its coefficients is -1.72, while No is 1.09. Such phenomenon is common among other biographical features. There two solutions that can solve this problem:

1. Train the model without discriminating features
2. (Recommend) Train the model with all features, then mask all applicants' non-financial features with default values, use masked features for predict.

Conclusion

In this report, we explore the HMDA mortgage dataset and reveals some hidden pattern, and we noticed that unfairness exists in our data. Then we tried different models, by comparing some optional configuration and apply useful training strategies, we successfully proved that GBDT will be the best model for our task. Finally, we talked about the unfairness in our model and recommend users to mask features that cause unfair or discriminating when predicting scores.

References

1. Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES. 96. 3-14. 10.1080/00220670209598786.
2. Evgeniou, Theodoros & Pontil, Massimiliano. (2001). Support Vector Machines: Theory and Applications. 2049. 249-257. 10.1007/3-540-44673-7_12.
3. Apostolidis-Afentoulis, Vasileios. (2015). SVM Classification with Linear and RBF kernels. 10.13140/RG.2.1.3351.4083.
4. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001). https://doi.org/10.1023/A:1010933404324
5. Mason, L.; Baxter, J.; Bartlett, P. L.; Frean, Marcus (1999). Müller (ed.). *Advances in Neural Information Processing Systems 12*. MIT Press. pp. 512–518.