MORTGAGE LOAN DEFAULT PREDICTION

Emmanuel Asante

Introduction

Real estate plays a fundamental role in the U.S. economy. Housing's combined contribution to the U.S. GDP generally averages between 15-18%. Real estate is commonly purchased using mortgages. A foreclosure is devastating to both the borrower and lender so being able to predict it is important.

This project analyzes publicly available information to predict loan default in the first 2 years of the mortgage lifespan.

Machine learning classifiers were employed to discriminate between default and active loans on balanced and imbalanced data.

Data Description

I used data from Fannie Mae loan data from 2013Q1 to 2018Q4 to train the models. Variables include loan features, customer features and external features. Some of the specific features are Debt-to-income ratio (dti) and Borrower credit score.

Machine learning classifiers

The cross-validation results for each machine learning classifier is shown Table 4 below. A 10- fold cross validation was implemented for each model on both the balanced class data and the imbalanced class data. Performance of models were assessed using accuracy, precision, recall and the F1 scores.. From the cross validation results all models performed well in terms of accuracy. However, transforming the variables by putting them on the same scale improved model performance. Also, the models fitted on the balanced data performed better than models fitted on imbalanced data. The imbalanced data models have higher accuracy levels but perform poorly on recall and precision due to the problem of class imbalance. The tree-based models such as the random forest, gradient boosting and Adaboost seemed to perform better than other models. The full results are summarized in the table below.

Table 5: Test results for select models for transformed models

	Transformed Variables							
	Balanced				Imbalanced			
Model	Acc	Pre	Rcl	F1	Acc	Pre	Rcl	F1
LR	0.83	0.82	0.85	0.84	0.91	0.63	0.33	0.43
LDA	0.83	0.81	0.86	0.83	0.90	0.53	0.44	0.48
QDA	0.5	1.00	0.1	0.1	0.81	0.33	0.78	0.46
KNN	0.82	0.8	0.86	0.83	0.91	0.58	0.41	0.48
CART	0.77	0.77	0.76	0.77	0.88	0.43	0.47	0.45
NB	0.8	0.75	0.88	0.81	0.79	0.31	0.82	0.45
AB	0.84	0.83	0.85	0.84	0.91	0.60	0.39	0.48
GBM	0.85	0.83	0.87	0.85	0.92	0.64	0.39	0.48
RF	0.84	0.83	0.86	0.84	0.92	0.67	0.38	0.48
ET	0.83	0.82	0.85	0.84	0.92	0.64	0.39	0.48

Conclusion

All the machine learning models performed well in terms of predictive accuracy but for loan mortgage prediction using recall to assess models is more appropriate. The Gaussian Naïve Bayes approach gave us the highest recall values of 88% and 82% for balanced and imbalanced class datasets respectively.