Final Report: Prediction of Loan Repayment

Problem Statement

Lending club, the fin tech market place bank, headquartered in San Francisco, California has been helping three Million members to reach their financial goals ever since 2007. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company and if the applicant is not likely to repay the loan, then approving the loan may lead to a financial loss for the company. The aim is to develop a predictive model that can be used as a reference tool that can help lending club to make decisions on issuing loans, so that the risk can be lowered, and the profit can be maximized. The given data set consists of 1048576 loan records with 136 columns. The status column shows the current state of each loan record.

By using the Lending Club Dataset, I developed a model as reference tool for helping the lending club management and the customers/members. Different exploratory data analysis methods and supervised learning classification methods were utilized to develop a model for this.

After cleaning my data and reducing my features to 6, my logistic regression model was able to classify 75% of the applicants that will default, compared to the other models.

Data Wrangling

The initial dataset consisted of 1048576 rows and 136 columns. In the initial data set, the loan status variable consisted of the categories fully paid, charged off, current, late, in grace period and default. This is a highly imbalanced data set since 46 percent are fully paid, 40 percent are current and only 2 percent are defaulters. The loans for which status is either fully paid or default are only considered for further analysis. After including loans for which status is either fully paid or default, now there are 482616 rows and 136 columns in the dataset. All columns which are theoretically not supposed to have a relationship with loan status and also which has more than 80 percent missing values were removed from the dataset.There were no duplicate values found in the dataset. Outliers were examined separately for each data point and examined if they were entered incorrectly or if the measurement was legitimate. In case of incorrect entries, the outliers

were removed and for legitimate measurement, they were kept as such. All the non numerical variables were converted into numerical variables by grouping them in different categories. The final data set consisted of 482616 rows and 104 columns after the data cleaning process.

Feature Engineering and Exploratory Data Analysis

The relationship of different factors that can possibly affect the home prices with the latest home price in Austin was examined using visualizations and correlations. To exploit maximum information from the feature annual income, it was converted into different categories and thus a new feature named incomegroup was created**.** The fico score variable was also grouped into three different categories to exploit maximum information Correlation is a technique for investigating the relationship between two quantitative, continuous variables in order to represent their inter-dependencies. The initial analysis shows no correlation between loan status variable and the numerical variables. In the final model, the income category and fico range variables are included instead of the income variable and fico range low since these two contain more useful information. Purpose and homeownership variables are also included in the final model. The visualization plots show association between income category, fico range, purpose, grade and homeownership variables with loan status. Hence in the final model, income category, fico range, purpose, grade and homeownership are included as the features.

Thus in total, there were 6 features in the final model and the dependent variable is loan status with “0” as fully paid and “1” as default . The remaining features were excluded in the final model based on preliminary trends.

Modeling

Three classifiers were tested for predicting loan default payment: Logistic Regression, Random forest classifier and decision tree classifiers. The metrics that were focused include precision, recall, f1 score and roc_auc_score. The data was divided into train and test data sets with a ratio of 80: 20. The models were fitted on the train data and the predictions were validated on the test data. Table 1 shows the evaluation metrics for the three initial models.

Table1: Evaluation Metrics for the initial models

Model	Roc_auc_score	precision	recall	F1
Random Forest	0.490613	0.0	0.0	0.0
Logistic Regression	0.843263	0.0	0.0	0.0
Decision trees	0.490624	0.0	0.0	0.0

The above results show that precision, recall and f1 score are all zero with all the three classifiers.roc_auc_score cannot be considered as a good evaluation method since the data is highly unbalanced. Hence the data is balanced using Synthetic Minority Oversampling technique

(SMOTE). Before balancing the data, only 0.004% is bad loan which will harm the performance of the model. After balancing the data using SMOTE, 50% are bad loan.

Table 2 shows the evaluation metrics for the three balanced models.

Table2: Evaluation Metrics for the balanced models

Model	Roc_auc_score	precision	recall	F1
Random Forest	0.576832	0.000129	0.25	0.000258
Logistic Regression	0.771044	0.000094	0.75	0.000188
Decision trees	0.576976	0.000129	0.25	0.000258

After oversampling the minority class by Synthetic Minority Over-Sampling Technique (SMOTE) in the training set, the recall score improves for every model, especially logistic regression. The recall score indicates that the logistic regression can classify 75% of the applicants that will default. The model was run with different random forest estimators but there were no differences in the performance with different random forest estimators. Thus, a random forest model that works better than logistic regression cannot be found out.

Based on the results above, logistic regression was selected as the best model for the analysis.

Conclusion

The objective of this model was to build a predictive model that can be used as a reference tool that can help lending club to make decisions on issuing loans, so that the risk can be lowered, and the profit can be maximized. The logistic regression can classify 75% of the applicants that will default when compared with 25 percent of default by random forest and decision tree classifiers. .The final model reached recall of 0.75 and roc_auc_score of 0.74 much higher than the initial unbalanced model. The final model shows the main factors affecting the loan status include purpose of loan, home ownership, grade, income category and fico score. The other factors like dti, employment length, revolving balance, hardship status etc were not included in the final model based on preliminary analyses.

Takeaways/Recommendations

Logistic regression was found to be outperforming all the other models in terms of scores and evaluation metrics compared to the decision trees and random forest classifiers. While the data was re sampled using SMOTE method, the performance of all models improved considerably. Many of the features were not found important in affecting the loan status and they were removed from the final model.

The lending club management and the customers/members might identify those loans that are likely to have chances for default, based on the model. The model based on logistic regression can correctly classify 75 percent of the bad loans that help to minimize risk and maximize the profitability of the institution.

Future Research

The analysis can be further extended by including boosting models like XG boosting regression models and also other machine learning models like support vector machine learning models and then examine how those models perform in comparison with the logistic regression model. Moreover, to find the optimized threshold for the models, the maximum profit needs to be located based on the revenue and cost.