1 EFFICACY OF LOAN DEFAULT LIKELIHOOD PREDICTIVE MODELS FOR LENDING CLUB DATASET

In this section we compare the models that predict the likelihood of a given loan defaulting. We first describe the experimental setup that was used to compare the predictive models, followed by the comparison itself.

1.1 Experimental setup

All our experiments in this section are performed on the Lending Club[1] dataset obtained from [2]. This dataset contains complete loan data from the period 2007 to 2011. There are 42,538 loans with a total 114 features for each loan. Similar to the Prosper dataset, we first cleaned/analysed as described in Section 3.1. After clean the dataset, we end up with 38,331 loans with 18 features (include response variable). The 17 selected features and the response variable are described in Table 1.

Amongst these 38,331 loans, there are 5,517 defaulted loans. To avoid negative impact of such an unbalanced dataset on the prediction models, we apply the under-sampling technique [1]. Without loss of generality, we construct 5 sample datasets by randomly choosing 5,517 non-defaulting loans along with all 5,517 defaulting loans together as the sample dataset 5 times and show the average performance in this section. For training and testing the prediction algorithms, we split each sample dataset into a ratio of 80:20, indicating 80% of the loans will be used to train the predictive models and the remaining 20% used for testing. In order to avoid over-fitting and under-fitting of the model, we apply 5 Monte Carlo cross-validation (CV) on each sample dataset.

Table 1. Features and response variable description

Feature	Explanation	Type
loan_status	The current status of the loan.	Categorical
dti	Debt to income ratio.	Numerical
delinq_2yrs	The number of delinquency for the past two years.	Numerical
int_rate	The borrower's interest rate for this loan.	Numerical
installment	The monthly payment owed by the borrower.	Numerical
home_ownership	Specifies if the borrower is a homeowner or not.	Categorical
inq_last_6mths	The number of inquiries in past 6 months.	Numerical
emp_length	Employment length in years.	Numerical
revol_bal	Total credit revolving balance.	Numerical
annual_inc	The borrower's annual income.	Numerical
pub_rec_bankruptcies	Number of public record bankruptcies.	Numerical
loan_amnt	The listed amount of the loan applied for by the borrower.	Numerical
grade	LC assigned loan grade.	Numerical
open_acc	The number of open credit lines.	Numerical
addr_state	The state of the borrower.	Categorical
Term	The length of the loan expressed in months.	Numerical
revol_util	Revolving line utilization rate.	Numerical
total_acc	The total number of credit lines.	Numerical

[1] www.lendingclub.com

1.2 Comparison of models predicting the likelihood of loan default

n this section, we compare Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Multilayer Perceptron (MLP), Logistic Regression (LOGIT) and Random Forest (RF) models to predict the likelihood of a given loan defaulting, and thereby find the best predictive model. For comparison purposes, we select accuracy, True Positive Rate (TPR) and True Negative Rate (TNR) as the criterion to evaluate the goodness of fit of the model. The definition of these three measures are described in Section 3.2. In particular, the higher the accuracy, the better the model fits the dataset.

Table 2. Results of SVM, kNN, MLP, LOGIT and RF for five sample datasets with the measure: accuracy.

Sample dataset	1	2	3	4	5	Average
SVM	0.49	0.49	0.48	0.51	0.50	0.49
kNN	0.53	0.54	0.55	0.53	0.52	0.54
MLP	0.56	0.58	0.50	0.57	0.58	0.56
LOGIT	0.60	0.62	0.63	0.61	0.62	0.62
RF	0.63	0.66	0.65	0.62	0.64	0.64

Table 3. Results of SVM, kNN, MLP, LOGIT and RF for five sample datasets with the measure: True Positive Rate (TPR).

Sample dataset	1	2	3	4	5	Average
SVM	1.00	1.00	1.00	0	0	0.60
kNN	0.52	0.56	0.53	0.54	0.53	0.54
MLP	0.69	0.74	0	63	0.45	0.50
LOGIT	0.74	0.76	0.76	0.73	0.73	0.74
RF	0.56	0.56	0.57	0.54	0.54	0.55

Table 4. Results of SVM, kNN, MLP, LOGIT and RF for five sample datasets with the measure: True Negative Rate (TNR).

Sample dataset	1	2	3	4	5	Average
SVM	0.00	0.00	0.00	1.00	1.00	0.40
kNN	0.54	0.49	0.54	0.55	0.51	0.53
MLP	0.43	0.36	1.00	0.50	0.70	0.60
LOGIT	0.47	0.49	0.50	0.48	0.51	0.49
RF	0.69	0.71	0.71	0.69	0.72	0.70

Tables 2, 3 and 4 show the results of accuracy, TPR and TNR for five sample datasets based on SVM, kNN, MLP, LOGIT and RF models, respectively. From Table 2, it can be observed that RF achieves the highest average accuracy of 0.64 (64%). LOGIT also fits the sample datasets well with average accuracy of prediction of 0.62 (62%). In addition, the average accuracy of SVM, kNN and MLP are 0.49 (49%), 0.54 (54%) and 0.56 (56%), respectively. From Table 3, it can be observed that LOGIT achieves the highest TPR of 0.74 (74%) amongst these five machine learning models.

From Table 4, it can be observed that RF achieves the highest TNR of 0.70 (70%) amongst these five machine learning models. Hence, RF and LOGIT are the two best models to predict the likelihood of a given loan defaulting. However, compared to LOGIT, RF may take longer to execute with a large dataset, and one may select LOGIT as the predictive model with a trade-off of 0.02 (2%) reduced accuracy of prediction.

REFERENCES

[1] Chris Drummond, Robert C Holte, et al. 2003. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, Vol. 11. Citeseer, 1–8.

[2] LendingClub. 2016. Lending Club Statistics. https://www.lendingclub.com/info/download-data.action. Last accessed - 15/9/2016.