Machine Learning for Problem Solving
Case Study Phase 2: Predictive Modeling

Zirui Zheng, ziruizhe
Keyu Zhang, keyuzhan
David Lee, dlee3

3.1 Predictive Models of Default

1.

i) *How did you set up your model training and evaluation?*

We randomly split our data into 30000 samples to be trained and 20000 samples to be tested. See evaluation metrics and results below in question iii).

ii) *Which model hyper-parameters did you tune (for each model)?*

L1, L2 logistic regression: regularization strength
Decision Tree: Max Depth, Minimum samples to split, Minimum samples per leaf
Random Forest: Bootstrap, Max depth, Max features, Minimum samples to split, Minimum samples per leaf, number of estimators
Multi-layer perceptron: Activation, Learning Rate, Alpha, and Hidden Layer Sizes

iii) *Which performance measure(s) did you use? Report your evaluation results*.

	Accuracy	Precision (No Default)	Precision (Default)	Recall (No Default)	Recall (Default)	F1 (Default)	F1 (No Default)
Naive Bayes	0.7962	0.1429	0.7967	0.0005	0.9992	0.0010	0.8865
L1	0.79855	0.5939	0.8006	0.0288	0.9950	0.0549	0.8873
L2	0.79815	0.5438	0.8024	0.0443	0.9905	0.0819	0.8866
Decision Tree	0.7859	0.4173	0.8117	0.1340	0.9522	0.2029	0.8763
Random Forest	0.77345	0.4142	0.8298	0.2762	0.9003	0.3314	0.8636
Multi-layer Perceptron	0.79815	0.5257	0.8061	0.0730	0.9832	0.1283	0.8859

2. *What are some advantages and disadvantages of using these data splitting procedures?*

Random: Covers the entire range of data and provides all kinds of train/test sets using random seeding, but it is possible to randomly select certain parts of the data that is similar and then test on data that is different from the training. It would be less susceptible than temporal on macro events that would change a person's delinquency but isn't reflected in the data.

Temporal: Trains on the past and tests on the future. Very similar to how the model will be deployed. However, it is possible for the data's pattern to change during a certain time frame. Thus, the model you trained would be more susceptible to the change in attitude that occurred. For example, more people who aren't supposed to be delinquent went delinquent in 2008 and the next couple of years.

3.

i) Provide a list of aforementioned features that are derived by LendingClub and any other features that correlate/reflect those.

Derived by LendingClub /correlated features:

- Dti
- Grade
- Annual income
- Installment
- Loan amount

ii) What is the predictive power as compared to that for the models you trained in part 1?

The predictive power is lower than our models in part 1.

iii) Generate 100 independent train/test splits with different seeds and report average performance values along with the standard deviation for each model.

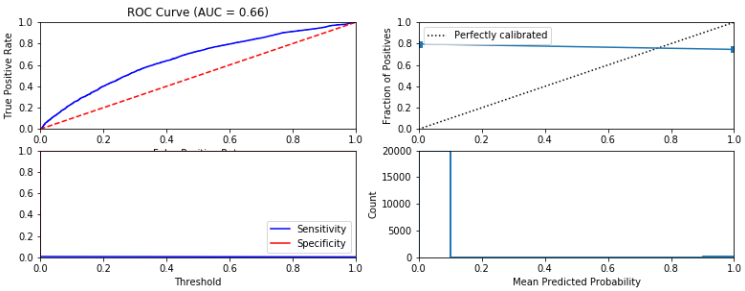
Without the derived features, the average overall performance decreases. Those derived features are calculated through rigorous modelling and have great explanatory power. That's why the scores are lower compared to previous model.

See next page for a detailed table of mean and standard deviation of different models.

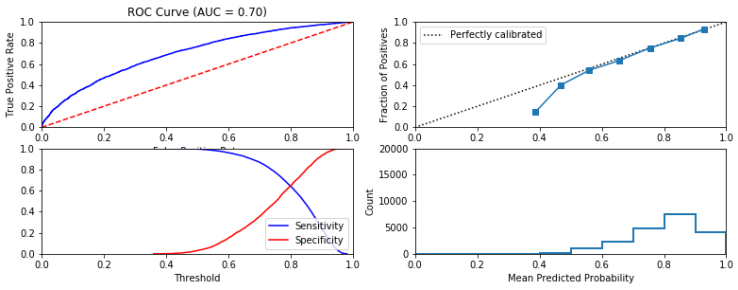
Weighted Average F1-score Across Splits

Model	Mean	Standard Deviation
Naive Bayes	0.7102	0.0040
L1 Logistic Regression	0.1833	0.2480
L2 Logistic Regression	0.7113	0.0045
Decision Tree	0.7240	0.0037
Random Forest	0.7399	0.0042
Multi-layer Perceptron	0.7085	0.0044

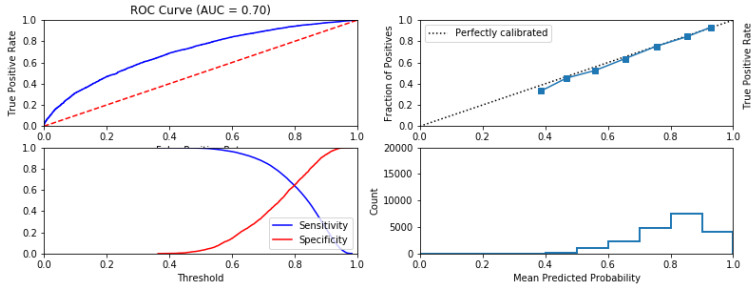
Naïve Bayes



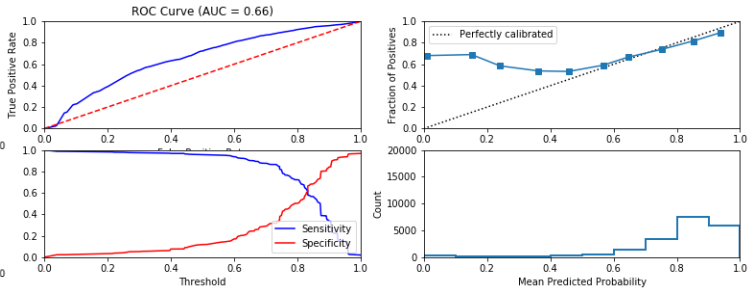
L1 Regression



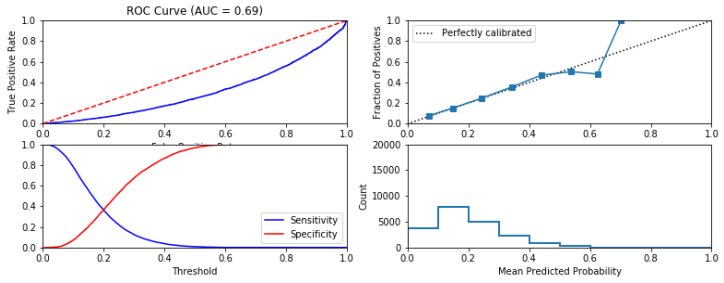
L2 Regression



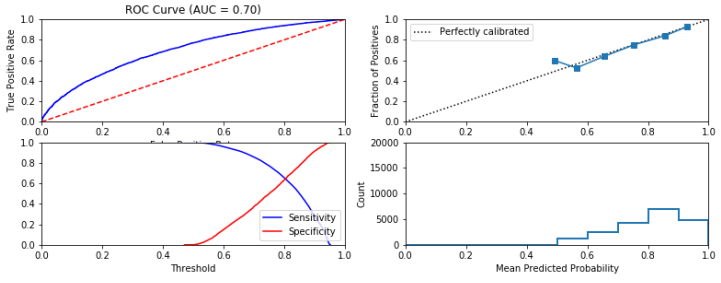
Decision Tree



Random Forest



MLP



4. *After modifying YourModel to ensure you did not include any features calculated by LendingClub, you want to assess the extent to which YourModel's scores agree with the grades assigned by LendingClub. How can you go about doing that? What is your observation?*

We used the table we got at phase 2 as our baseline. The table has the default percentage of loan for each grade. We use this percentage to help us estimate the prediction performance of our model. For example, 7.5% A-grade loans default, so if the model gives probability lower than 92.5%, we consider it doesn't underestimate the risk.

After calculation, the overall performance shows that the model agrees with the grades from LendingClub well.

5. *To this end, analyze whether YourModel trained (using the Random data splitting procedure in part 2. For cross validation) in 2009 performs worse in 2017 than YourModel trained on more recent data in 2016. What conclusion can you draw? Is your model stable?*

   The performances of the two settings don't differ much so that the model is stable.

   One thing we notice is the score in 2010's model is slightly higher than 2016's model, which may result from the that the model should be stricter in 2010 than in 2016, considering the aftermath of Subprime Crisis. That's why it gets higher score on default samples.

6. *Does anything surprise you about the performance of this model (averaged on out-of-sample test datasets) compared with the other models you have fit earlier?*

   It's quite surprising that the performance improves quite a bit but after carefully reviewing the features, we thought this makes sense. We introduced back variables that are not available at the time the loan was issued, which is not realistic and caused our model performance to improve. As we did before, we should remove these features.
7. *Report the performance results in corresponding entries in Table 3.1. Do they perform well? Can you tell?*

   The performance looks not so well based on low R-square scores. However, since our scenario have results mainly generated by human, and in addition, we have statistically significant predictors, low R-square scores here do not necessarily mean the regressors failed. In this case, we picked the Random Forest Regressor as our best regressor.

	Performance for each return calculation			
Model	M1	M2	M3(2.4%)	M3(6%)
L1 regressor	0.0110	0.0082	0.0092	0.0116
L2 regressor	0.0261	0.0159	0.0307	0.0315
Neural Network regressor	-0.0001	-0.0010	-0.0089	-0.0005
Random Forest regressor	0.0298	0.0173	0.0361	0.0364

8.

i) *Suppose you were to invest in 1000 loans using each of the four strategies, what would your returns be? Average your results over 100 independent train/test splits.*

See the table below.

ii) *Include the best possible solution(denoted Best) that corresponds to the top 1000 performing loans in hindsight, that is, the best 1000 loans you could have picked.*

	Return calculation			
Strategy	M1	M2	M3(2.4%)	M3(6%)
Rand	0.26%	4.49%	2.22%	5.59%
Def	0.06%	4.52%	2.07%	5.65%
Ret	1.62%	4.41%	2.18%	5.61%
DefRet	1.58%	4.54%	2.11%	5.59%
Best	8.59%	17.22%	10.08%	14.36%

iii) *Based on the above table, which data-driven investment strategy performs best? What can you tell about using the Random strategy? Does it cause you any loss? Why do think that is the case? How do the data-driven strategies compare to Random as well as BEST?*

Return-based perform the best but Default-Return-based is very close. I picked both for the next question.

The Random Strategy seems not bad. It has a similar return rate in M2, M3, and M4 with data-driven strategies, and it does not cause any loss. We think the reason is that LendingClub use interest rate to offset risks of loans. Also, the imbalanced data problem might cause most loans to have pretty good returns. Hence, randomly choosing loans can still result in good results.

Best Strategy outperformed all data-driven strategies by 4-5 times. This makes sense as we eyeballed all the returns and pick the highest 1000 loans. However, this won't work in reality as this method cannot measure returns when we only have features rather than the results themselves.

Theoretically, DefRet should have the highest return. However, in Phase-2 when we have taken into account the influence of default and non-default loans in calculating returns. Thus, the Default-return-based method here did not perform outstandingly.

9. *What trend do you observe? Why do you think that is the case?*

As portfolio size increases, the average return will decrease. This makes sense because in choosing loans, we sort the values of predicted return that we calculated by our best data-driven method. Hence, as we increase portfolio size, much lower-return loans are included hence generally decrease our investment return.