Big Data in Finance: Assignment 1

Prof. Tarun Ramadorai

February 22, 2017

As demonstrated in Lecture 2, it is possible to build a reduced form model to predict loan default from a set of loan attributes. In this assignment you will build your own models using a peer-to-peer lending dataset, and test the predictive ability of a few different Machine Learning algorithms.

For this assignment, we will use data from *Lending Club*. They are the world's largest on-line marketplace, directly connecting borrowers and investors. A dataset containing complete loan data for all loans issued through the platform over the period from 2007 to 2015 is available at the Kaggle platform.[1] It includes the current loan status (Current, Late, Fully Paid, etc.) and a large set of attributes for each customer. We have selected a subset of these data and cleaned them for you (eliminating some variables, merging some others, filling missing data...), and the result can be downloaded from the Materials section of the Hub.

The assignment will require you to use the dataset to build models to predict loan default. To provide answers for the assignment, please use the space provided in the PDF template. Please restrict yourself to answering in the space provided - brevity and precision are highly valued. However, if you wish, at the end of your submission (after the PDF template) you can attach any other material that you consider relevant (e.g. code or images, and if you like, explanations or additional tests/text).

Full model

The cleaned dataset (**FULL-MODEL**) that you have downloaded has 123 different attributes of each loan, along with an indicator of the "loan status". The status of the loan is 1 if the loan was "charged off" (CO), delinquent, or late in payment, and 0 otherwise, and this is the target that we will be predicting.

Our first approach is to use all available information (all 123 attributes, we term this the **FULL-MODEL** set of attributes) to predict if a loan will default. Please take into account that at this stage, we are not interested in the exact probability of default, but rather whether there will or will not be a default (i.e., the domain of the predicted variable is binarized into 0 and 1, and in some cases we will use what is essentially a linear probability model).

1. (10 points) Apply three different machine learning techniques (logistic regression, tree classification, and 1-NN) to the **FULL-MODEL** and entire dataset in order to predict loan default. To obtain a measure of how good each technique is at using these attributes at predicting default, compute the *accuracy* of each measure,[2] using the method of 10-fold cross validation. Complete the following table with the accuracy of each method for prediction.

accuracy	logistic regression	tree classifier	1-NN
FULL-MODEL			

[1] https://www.kaggle.com/wendykan/lending-club-loan-data

[2] *Accuracy* is defined as the number of correct predictions divided by the total number of predictions, multiplied by 100 to turn it into a percentage.

Reduced model

It is clear that not all of the 123 attributes will be useful for predicting if a loan will default. The next step is to reduce this number of attributes to get a more tractable model, while still being able to predict default well.

2. (10 points) Which loan attributes do you *believe* are the most informative? Please use your knowledge and intuition to choose 10 of the 123 attributes, and provide a brief justification for why you chose these attributes. Let's call this the **REDUCED-MODEL**.

Now that we have reduced the dimensionality of the feature/attribute space, let's use these attributes to predict default, and check the performance of the resulting models.

3. (10 points) Apply the same three different machine learning techniques (logistic regression, tree classification, and 1-NN) to the **REDUCED-MODEL** dataset in order to predict loan default. Again, to obtain a measure of how good these techniques are at predicting default, compute their accuracy using 10-fold cross validation.

accuracy	logistic regression	tree classifier	1-NN
REDUCED-MODEL			

Lasso-reduced model

While using one's own knowledge and intuition for selecting relevant attributes might be a sensible thing to do, it may be that using a statistical approach to select variables for us might provide better results. Let's try that next.

4. (10 points) Please explain in your own words how you might use the LASSO estimation method to select a subset of attributes. In particular, please explain what the LASSO parameter λ is for, and how increases in λ change the estimates you obtain from the LASSO.

The LASSO estimation method selects only a subset of attributes which are estimated to have high predictive power on the current status of the loan.

LASSO performs this by adding a penalty term to the objective function (which we would like to minimise. The penalty term is the sum of absolute coefficient value, multiplied by parameter λ. A higher value of λ will decrease the number of selected attributes because the absolute magnitudes of regression coefficients are more heavily-weighted (and the algorithm will reduce some regression coefficients to zero).

5. (10 points) First, let-s calculate the accuracy of LASSO approach, and compare it to previous methods. For that, apply the LASSO to the default response variable (1 or 0), and the entire set of 123 attributes, and constrain the model to have at most 10 attributes in it (hint: in Matlab, use option *dfMax* in the lasso() command). Do this inside cross-validation, i.e., compute the method's accuracy using 10-fold cross-validation. Note that the set of ten attributes that the LASSO will select may be different across each of the 10 folds in the validation, but this is not a concern when we are estimating accuracy.

accuracy	logistic regression
LASSO-MODEL	82.04%

6. (10 points) Next, apply the LASSO to the default response variable (1 or 0), and the entire set of 123 attributes on the entire dataset (not inside 10-fold cross-validation) and once again, constrain the model to have at most 10 attributes in it. Let's call the resulting set of 10 attributes the **LASSO-MODEL**. Which attributes are those? Did your intuition about the correct set of attributes in the **REDUCED-MODEL** match the algorithmic results that you obtained from the **LASSO-MODEL**? Discuss.

The in-sample accuracy achieved after training the entire set of attributes on the entire dataset (not using cross-validation) is 82.05%, which is very close to the accuracy computed using 10-fold cross validation earlier.

The ten attributes selected by LASSO-MODEL is:
- loan_amnt
- int_rate
- installment
- annual_inc
- dti
- inq_last_6mths
- out_prncp
- total_rec_prncp
- issue_year
- GRADE_A

Two of the attributes obtained using LASSO-MODEL are the same as these selected in REDUCED-MODEL (i.e. int_rate, dti).

Understanding Model Performance

7. (10 points) Congratulations! You have now estimated a set of different models for default forecasting. Now assume that you are running a bank. Please pick one set of attributes and one algorithm to implement in the bank. Do you now have sufficient information to run a loan business?

8. (10 points) What metric (in addition to the previously used *accuracy*) might you use to distinguish between these models? Explain what this metric means and how it might be computed.

9. (10 points) How do the results using the full set of attributes **FULL-MODEL** compare to those that you personally selected **REDUCED-MODEL** in order to predict if a loan will default? What do you think is more important in this particular example, the set of attributes, or the classification technique? Discuss.

10. (10 points) Are these the best models you could possibly create? Name one other possible classification technique and any additional attributes that you might be able to add in order to improve accuracy. Please justify your choices.