

# Big Data in Finance: Assignment 1

Prof. Tarun Ramadorai

Due February 23, 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 peer-to-peer lending 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 the present is available on their website:

<https://www.lendingclub.com/info/download-data.action>

These data include 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 from 2017 Q3 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. Please note that this is NOT common in the real world - data cleaning is one of the most important and difficult tasks you will have to perform. However, in the interests of you learning the techniques, we have made the data as clean as possible.<sup>1</sup>

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 a small set of other material that you consider relevant (e.g. you MUST include your code, but if you wish to include additional figures, explanations, or additional tests/text, you can limit this to an additional 2-3 pages at the end, max.).

## Full model

The cleaned dataset (**FULL-MODEL**) that you have downloaded has 40 different attributes of each loan, along with an indicator of the “loan status”. The status of the loan is “Fail” if the loan was “charged off” (CO), delinquent, or late in payment, and “Current” otherwise, and this is the target that we will be predicting (you should obviously convert these to numerical values, 1 for Fail and 0 otherwise).

Our first approach is to use all available information (all 40 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 (quadratic discriminant analysis (QDA), tree classification, and 5-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,<sup>2</sup> using the method of 10-fold cross validation. Complete the following table with the accuracy of each method for prediction.

---

<sup>1</sup>Please note: any remaining data issues are very much your responsibility, to give you at least some of the feel of what is involved with real world data! Note: we have tested and trained algorithms using these data ourselves, with no problems...

<sup>2</sup>*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.

<i>accuracy</i>	QDA	Tree classifier	5-NN
<b>FULL-MODEL</b>			

## Reduced model

It is clear that not all of the 40 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 set of three different machine learning techniques with one change - logistic regression instead of QDA (i.e., logistic regression, tree classification, and 5-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.

<i>accuracy</i>	logistic regression	tree classifier	1-NN
<b>REDUCED-MODEL</b>			

## Ridge-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 Ridge estimation method to select a subset of attributes. In particular, please explain what the Ridge parameter  $\lambda$  is for, and how increases in  $\lambda$  change the estimates you obtain from the Ridge.

5. (10 points) First, let's calculate the accuracy of the Ridge approach, and compare it to previous methods. For that, apply the Ridge to the default response variable (1 or 0), and the entire set of 40 attributes, and constrain the model to have at most 10 attributes in it. 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 Ridge 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. Note also that you need to be very careful when applying the Ridge technique in terms of the relative magnitudes of the different variables!

<i>accuracy</i>		logistic regression
<b>Ridge-MODEL</b>		

6. (10 points) Next, apply the Ridge regression to the default response variable (1 or 0), and the entire set of 40 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 **Ridge-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 **Ridge-MODEL**? Discuss.

## 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? What else might you need to consider? Hint: see this paper:

[https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3072038](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3072038)

**8. (10 points)** What metric(s) (in addition to the previously used *accuracy*) might you use to distinguish between these models? Explain what these metrics mean and how they 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.

