Big Data in Finance: Assignment 1

Prof. Tarun Ramadorai

January 2020

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 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.[1]
We created *three* datasets: please use "Loan_data_part_I.csv" as the main dataset for Part I of the problem set (which is based on 2013 data), and "Loan_data _part_II_a(b)" (based on (a) 2014 data and (b) 2016 data) only for Part II, which we will use to evaluate the models truly out-of-sample (i.e. using data the models have not been trained on).

The assignment will require you to use the data 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.).

Part I

Full Model

The cleaned dataset (FULL-MODEL) that you have downloaded has 35 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, 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 35 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] 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...

1. (10 points) Apply three different machine-learning techniques: linear discriminant analysis (LDA), Random Forest classification (with 100 trees), and 50-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 to predict default, compute the confusion matrix of each measure, using the method of 10-fold cross validation. Complete the following tables with the confusion matrices of each method for prediction (*note*: please normalize *all* confusion matrices by dividing each cell by its column-total, such that it reflects the *share* of predicted observations that fall in that outcome category (Current or Default) and both columns sum up to 1).

LDA

Predicted (Default or Current)	Outcome (Default or Current)	
	Current	Default
Current		
Default		

Random Forest

Predicted (Default or Current)	Outcome (Default or Current)	
	Current	Default
Current		
Default		

50-NN

Predicted (Default or Current)	Outcome (Default or Current)	
	Current	Default
Current		
Default		

In the previous question, you used the entire dataset but defaults are rare, which means that an approach in which we classify all observations as "Current" has relatively high accuracy (or low misclassification error), but is not a useful model. Undersampling or oversampling can be used to combat this.

2a. (10 points) Undersample the number "Current" to have 50% "Fail" and 50% "Current" in your training set. Rerun the three previous machine-learning techniques and compute the confusion matrices. Are the predictions better? Explain.

LDA

Predicted (Default or Current)	Outcome (Default or Current)	
	Current	Default
Current		
Default		

Random Forest

Predicted (Default or Current)	Outcome (Default or Current)	
	Current	Default
Current		
Default		

50-NN

Predicted (Default or Current)	Outcome (Default or Current)	
	Current	Default
Current		
Default		

2b. (5 points) In the previous questions, we gave you the hyperparameter for the k-nearest neighbors model (with k=50). This parameter should be found through cross-validation. Write a pseudo-code showing how you would find the optimal value for this hyperparameter.

Please note: for the following questions, always undersample the training set.

Reduced model

It is clear that not all of the 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.

3a. (5 points) Which loan attributes do you believe are the most informative? Please use your knowledge and intuition to choose 10 of the attributes, and provide a brief justification for why you chose these attributes. Let's call this the REDUCED-MODEL.

3b. (5 points) Apply the set of different machine-learning techniques with one change – logistic regression instead of LDA (i.e., logistic regression, Random Forest classification, and 50-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 confusion matrices using 10-fold cross validation. Please undersample again using the number "Current" to have 50% "Fail" and 50% "Current" in your subsample.

Logistic regression

Predicted (Default or Current)	Outcome (Default or Current)	
	Current	Default
Current		
Default		

Random Forest

Predicted (Default or Current)	Outcome (Default or Current)	
	Current	Default
Current		
Default		

50-NN

Predicted (Default or Current)	Outcome (Default or Current)	
	Current	Default
Current		
Default		

4. (10 points) Choosing the right model depends on the context in which it is used. Can you think of examples where one would worry more about false positives than false negatives, and vice versa? Another way to compare models is to compute the ROC (receiver-operating characteristic) curve that varies the classification threshold and shows the true positive rate (share of defaulters correctly classified) against the false positive rate (share of non-defaulters incorrectly classified). Plot the ROC curves and compute the area-under-the-curve (AUC) for a LDA and 50-NN model and explain which model should be preferred in your opinion.
Hint: Do that without 10-fold cross-validation.

AUC

LDA	**50-NN**

Lasso-reduced model

5. (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.

6. (5 points) First, let's use the Lasso approach, and compare it to previous methods. Apply the Logistic Regression model with Lasso penalization method to the to the default response variable (1 or 0), and the entire set of attributes. For now, set λ to 0.1. Compute the resulting confusion matrix using 10-fold cross-validation (fill in the table).

Penalized logistic regression - LASSO

Predicted (Default or Current)	Outcome (Default or Current)	
	Current	Default
Current		
Default		

7. (10 points) Next, apply the Logistic Regression model with Lasso penalization method to the default response variable (1 or 0), and the entire set of 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 (aim for 8 to 10). Let's call the resulting set of attributes the LASSO-MODEL. What is the value of λ now? 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.
Hint: Start with different values for λ, e.g. 1e-15, 1e-10, 1e-8, 1e-5,1e-4, 1e-3,1e-2, 1, 5, 10, check the number of coefficients that are zero, and iterate the value of λ until you get the desired number of coefficients that are shrunk to zero.

Understanding model performance

8. (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`

9. (5 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. (5 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 the confusion matrix. Please justify your choices.

Part II

Evaluating the models out-of-sample

11. (10 points) Lastly, we will do a proper out-of-sample test of your models, which is the challenge you would encounter in the real world. For that purpose, fit or train your Logistic, LDA and 50-NN models on the data from Part I (with undersampling), and test them on the datasets (a) and (b) for Part II (no undersampling - why?). Report the AUC for the three models below. Within each dataset, compare the relative performance of the three models with your previous results and discuss. Do you find overall differences in performance across datasets? Discuss.

AUC (OOS) - Dataset (a)

Logistic	LDA	50-NN

AUC (OOS) - Dataset (b)

Logistic	LDA	50-NN