7390–Machine Learning and Data Sciences

Northeastern University, Fall 2017

Midterm, due on Saturday Dec 02, 2017

- The exam is take home.

Problem Description: When a customer applies for a loan, banks and other credit providers use statistical models to determine whether or not to grant the loan based on the likelihood of the loan being repaid. The factors involved in determining this likelihood are complex, and extensive statistical analysis and modeling are required to predict the outcome for each individual case. You should analyze a loan dataset using logistic regression, random forests, and boosting to predict loan repayment or default based on the data provided. The dataset consists of 94,440 loan records including the following fields:

- Loan ID: A unique Identifier for the loan information.
- Customer ID: A unique identifier for the customer. Customers may have more than one loan.
- Loan Status: A categorical variable indicating if the loan was paid back or defaulted. Target variable
- Current Loan Amount: This is the loan amount that was either completely paid off, or the amount that was defaulted.
- Term: A categorical variable indicating if it is a short term or long term loan.
- Credit Score: A value between 0 and 800 indicating the riskiness of the borrowers credit history.
- Years in current job: A categorical variable indicating how many years the customer has been in their current job.
- Home Ownership: Categorical variable indicating home ownership. Values are“Rent”,“Home Mortgage”, and “Own”. If the value is OWN, then the customer is a home owner with no mortgage
- Annual Income: The customer’s annual income
- Purpose: A description of the purpose of the loan.
- Monthly Debt: The customer’s monthly payment for their existing loans
- Years of Credit History: The years since the first entry in the customerss credit history
- Months since last delinquent: Months since the last loan delinquent payment
- Number of Open Accounts: The total number of open credit cards
- Number of Credit Problems: The number of credit problems in the customer records.
- Current Credit Balance: The current total debt for the customer
- Maximum Open Credit: The maximum credit limit for all credit sources.
- Bankruptcies: The number of bankruptcies
- Tax Liens: The number of tax liens.

Questions:

1. Analyze, process, cleanse the dataset and produce some numerical and graphical summaries of data. Answer the following questions:
   (a) Do there appear to be any patterns?
   (b) What important fields and information does the dataset have?
   (c) How do you clean the data and fill in the missing data?
2. Use logistic regression to analyze this data.
   (a) Use 100 bootstrap samples to estimate the the standard errors of the coefficients from the logistic regression fit AUC. Provide a 68% and 95% confidence interval for the model parameters of the logistic regression.
   (b) Calculate p-values for each feature. Do any of the features appear to be statistically significant? If so, which ones?
   (c) Use 5-fold cross-validation to estimate test AUC. What test AUC do you obtain? Provide a 68% and 95% confidence interval for test AUC.
   (d) Use 5-fold cross-validation and *Forward Stepwise Selection* approach for feature selection. In each iteration, choose the best model using the highest AUC. Report the features and AUC that appears to provide the best results.
   (e) Use 5-fold cross-validation and *Backward Stepwise Selection* approach for feature selection. In each iteration, choose the best model using the highest AUC. Report the features and AUC that appears to provide the best results.
3. Use random forests to analyze this data.
   (a) Use 100 trees for a range of values of max features consider. Describe the effect of the number of features considered at each split on the test AUC. What is the optimal value for max features?
   (b) What are most important features?
   (c) What test AUC do you obtain using 5-fold cross-validation?
4. Use boosting to analyze this data.
   (a) Perform boosting with 100 trees for a range of values of the shrinkage parameter λ. Produce a plot with different shrinkage values on the x-axis and the corresponding 5-fold cross-validation AUC on the y-axis. What is the optimal value for λ?
   (b) What are most important features?
   (c) What test AUC do you obtain using 5-fold cross-validation?
5. Compare the results between logistic regression, random forests, and boosting.
   (a) Plot ROC curves for the best logistic regression model, the best random forest model, and the best boosting model.
   (b) Which of these approaches yields the best performance?
6. Submit a Python or R script to show your work and evaluate the results on a test dataset.
   (a) For a given test dataset, it should provide the AUC and plot ROC curves for all three models
   (b) For a given test dataset and a given threshold, it should generate the confusion matrix and provide accuracy and FP and TP rates for all three models.