CISC 5352 Financial Data Analytics Quiz (4)

A). Nearest neighbor analytics (II) (60 points)

Finish the following analysis for the dataset: NBoption.csv

1. Partition data into 80% training data and 20% test data
2. Employ *k-NN* regression to predict volatility (you can choose your features) and evaluate its MSE
   (a) Find the samples whose values $error = |Volatility - predictedVolatility|$ are at the top 10% of all testing samples
   (b) Find the samples whose values $error = |Volatility - predictedVolatility|$ are at the bottom 10% of all testing samples
3. Employ *k-NN* regression to predict implied volatility without using volatility as a feature, that is, you exclude volatility in your features.
   (a) Find the samples whose values $error = |IV - predictedIV|$ are at the bottom/top 10% of all testing samples
4. Employ *k-NN* regression to predict implied volatility with using volatility as a feature, that is, you include volatility in your features.
   (a) Find the samples whose values $error = |IV - predictedIV|$ are at the bottom/top 10% of all testing samples
5. Compare their MSEs and draw your conclusion.
6. Try at least 3 different types of distances in k-NN regression and neighbors to find the best results for the previous problems.

B) Credit Risk Analytics (I) (20 points)

- Credit risk analytics is key in personal loan decision making for banks. Using credit risk analytics, banks are able to analyze previous lending data, along with associated default rates, to create an effective predictive model in loan decision making.
- The file *credit_data_risk_balanced.csv* has 16714 credit records with the following variables
  - 'Index': case number
  - 'Delinquency': this is a binary variable 1: means bad credit and 0: means good credit
  - 'Revolving Credit Percentage',
  - 'Capital Reserves',
  - 'Num Late 60',
  - 'Debt Ratio'
  - 'Monthly Income' ($)
  - 'Num Credit Lines' ($1000)
  - 'Num Late Past 90',
  - 'Num Real Estate',
  - 'Num Late 90',
  - 'Num Employees' (should not be more than 10 for a personal credit analytics)

1. Partition data as 80% for training and 20% for testing and use K-NN (you can try different distances) and support vector machines, to conduct credit risk analytics, i.e. do classification to determine good or bad credit records.
2. Draw your conclusion.

What should you turn in?

- 1. A folder that contains
  - A report to show details of your analytics (at least 15 pages)
  - your data
  - source files
  - corresponding related output.
- 2. Please name your folder last_name1_last-name2_CISC5352_Quiz_4. For example,
- 3. Send the zipped file (.zip instead of ,rar) of your folder to Blackboard before 11:59 pm Nov 7, 2018