INFO 7390 Midterm Case studies – Fall 2018 – version 1.0

Sri Krishnamurthy

- Answer all questions
- Submit an executive report (in MS Word) with your detailed analysis, explanation and interpretation of your analysis
- Deadline: Midnight – 11/29/2018
- You should include
  - One report summarizing all problems in PDF format
  - Share your data files and code through github with analyticsneu@gmail.com
  - Slide deck for all the problems
- On 1st ,3 teams will be asked to present their analysis. You will have 15 minutes. Focus on the demo. (10 minute demo + 5 minutes Q&A). You will be timed. Please rehearse to keep demo within 10 minutes.
- If you make assumptions, clearly state that in your report
- Include a slide (pie chart) with bullet points on contributions from each team member

You are hired by the US Housing and Urban Development (https://portal.hud.gov/hudportal/HUD) as a Data Scientist to understand the US housing market. You are given the Single-family loan data (http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.html) and asked to analyze the data and present your analysis. You will then be asked to build predictive analytics models using the datasets.

Part 1: Data wrangling (40)

Data Download and pre-processing: (10 points)

Your first challenge is to programmatically download the data from https://freddiemac.embs.com/FLoan/Data/download.php

You must figure how to programmatically download the data getting past the login page using a script in R/Python. (Hint: Look for ways to save and pass cookies). Once you login, your script will scrape the data page and download all the "sample" files from 2005 onwards. It will then summarize key information that you would like to present to your management team. You goal is to relay your understanding regarding summary measures for different attributes, trends over time for different variables. Counts, variability of numerical measures, location based statistics etc. You should do that for each file and generated one/more aggregate file for all data for 2005 onwards for Origination data and one/more aggregate file for PERFORMANCE data. Note: These summary files are large: You must preprocess data so that you can consolidate the "summaries" into one file keeping in mind the size of the file.

NOTE: NO HUMAN INTERVERSION SHOULD BE NEEDED AT ALL. ONE SCRIPT SHOULD DO EVERYTHING.

Exploratory Data analysis: (30 points)

- Write a Jupyter notebook using R/Python to graphically represent different summaries of data. Summarize your findings in this notebook.
- Analyse the quarterly data from 2007,2008, 2009 data including summary measures for different attributes, trends over time for different variables. Counts, variability of numerical measures etc. Leverage postal codes and states to indicate location specific information.
- Look at things like interest rate trends, delinquency trends over quarters, location specific insights etc. and other parameters. You have a lot of very interesting variables. Note: You should spend

some time analyzing which parameters matter and how things have changed. This is the kind of work of work Data scientists are expected to do. Please take this section seriously!

Sharing your work:

You present your findings to your teams in Boston. The teams there are intrigued and interested to review your work. They ask that they share your work. You should dockerize the whole project with required packages and write clear instructions on how the teams should run the docker image. Note: You should parameterize the cookies and provide instructions on how the teams can create their own logins and use their cookies to run the docker image. Put your image on Docker hub and instructions in Github to recreate the docker image.

Part II: Building and evaluating models. (60)

Your managers are impressed with the work you did! They shared it with economists and other experts who are intrigued by various parameters. They are interested in predicting interest rates in the next quarter based on information from the origination data **from the prior quarter**. In addition, they are interested in predicting whether a record is delinquent or not based on performance data. Fortunately, you have quarterly datasets readily available on the same data download page. You must do the following:

Prediction (30 points)

- Write a prediction script in a Jupyter notebook that given input (For example Q12005),
    - Programmatically downloads Q12005 and Q22005 origination data and pre-processes it.
    - Builds a Regression model for the interest rate using Q12005 data as training data (col 13)
    - Does variable selection to choose the best Regression model using Forward, Backward, Stepwise and Exhaustive search methods.
    - Validates against the Q22005 datasets
    - Computes MAE, RMS, MAPE for training and testing datasets
    - Repeat this using Random Forest & Neural Network algorithms.
    - Try TPOT, H20.Ai and AutoSKLearn Automl algorithms
    - Choose the best model amongst the different types of algorithms.
- You are asked to do what-if analysis had your algorithm used in various scenarios:
    - Financial crisis (https://www.stlouisfed.org/financial-crisis/full-timeline)
        - Run your algorithm for 4 rolling quarters and report your findings and discuss it in your report. (i.e Use Q12007, Q22007, Q32007, Q42007 for training and predict for Q22007,Q32007,Q42007,Q12008)
        - Run your algorithm 2 years later (i.e, 2009 for all 4 quarters)
    - Economic boom (1999, 2013) (https://www.thebalance.com/stock-market-returns-by-year-2388543)
        - Discuss your design and results in a report. Would you recommend using this model for the next quarter? Justify
    - Regime change (2016) from election

Classification (30 points)

- Write a new jupyter notebook that given input (For example Q12005),
    - Programmatically downloads Q12005 and Q22005 origination data and pre-processes it.
    - Builds a Logistic regression model for the CURRENT LOAN DELINQUENCY STATUS using Q12005 data as training data (col 4). Note anytime col 4 is > 0, add a new variable as Delinquent. Use this variable as your "Y" variable. IGNORE COL 4 AND DON'T USE IT IN YOUR MODEL.
    - Validates against Q22005 data and selects the best Classification model

    - Computes ROC curve and Confusion matrices for training and testing datasets
    - Repeat this using Random Forest & Neural Network algorithms.
    - Try TPOT, H20.Ai and AutoSKLearn Automl algorithms
    - Choose the best model amongst the different types of algorithms.
- Parameterize the input (example it should take Q12005) and modify the code so that it outputs the 5 parameters listed in the matrix below.
- Write another script that calls the above classification script from Q11999-Q42016 and computes the following matrix.
- Extra credit: Note that each invocation of is independent. There is scope to parallelize it. Parallelize this so that you can run two or more models concurrently (task parallel). Review R and Python packages that can be used to parallelize it to generate the below matrix as a dataframe and export it to a csv file.
    - Hints:
        - R- http://www.win-vector.com/blog/2016/01/parallel-computing-in-r/
        - Python-http://www.davekuhlman.org/python_multiprocessing_01.html

Quarter	Number of Actual Delinquents	Number of Predicted Delinquents	Number of records in the dataset	Number of Delinquents properly classified	Number of non-delinquents improperly classified as delinquents
Q21999					
through					
Q12016					

- Document your design and results in the report. Comment on the quality of the model and it's outputs. What can you do to do better? Would you recommend using this model to predict delinquents in the next quarter? Justify your answers
- Review https://en.wikipedia.org/wiki/Confusion_matrix . Which metrics do you recommend to use to evaluate the performance of your algorithm? Compute those metrics for Q12016.

Your Docker image should include Part 1 and Part II also so that you can share this with your team members!

Good luck!