INFO 7390 Assignment Case studies – Spring 2017 – 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: Part 1 Midnight – 4/7/2017, Part 2 Midnight – 4/14/2017
- Each team will present their Part 1 on 4/8/2017- You will have 5 minutes
- You should include
  - One report summarizing all problems in WORD format
  - Share your data files and code through github with analyticsneu@gmail.com
  - Slide deck for all the problems
- On 15th, we will have a demo all the components of Part 2.
- Due to logistics and time constraints, we won't hand off the code to the 3 teams we discussed in last class. All teams are expected to come up with their workflow for deployment
- 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 working at a bank and you are considering investing in Lending club. Since there are no standard models, you are expected to build prediction models that will help you predict the interest rates based on various parameters users would input.

Part 1: Data wrangling and exploratory data analysis (100)

Data Download and pre-processing: (50 points)

Your first challenge is to programmatically download the data from https://www.lendingclub.com/info/download-data.action

Your goal is to download the data programmatically from the website and create one dataset for the entire database.

Things to think about:

- Data download: How will you download all loan data and create one dataset? How can you timestamp the data so you know when the data was recorded?
- Missing data analysis: How will you handle missing data?
- Feature engineering: What variables do you need to predict interest rates? Ensure users would be able to give you that information to help you predict rates
- Pipeline: Using Luigi/Pinball/Airflow automate the above 3 steps.
- You need to create one more pipeline to do this for the "Declined Loan data". Repeat above steps

NOTE: NO HUMAN INTERVERSION SHOULD BE NEEDED AT ALL. YOUR LUIGI/AIRFLOW script SHOULD DO EVERYTHING.

Dockerizing and Scheduling the pipeline: (30 points)

You should dockerize the whole project (excluding the Power BI dashboard) and write clear instructions on how to run the docker image. Research how you would schedule this on Amazon/Azure/Bluemix so that this image would run and execute the Luigi/Pinball/Airflow pipeline. After running this pipeline, the clean pre-processed data should be stored on S3/Object Storage/Blob storage(AWS/Bluemix/Azure)

Exploratory Data analysis: (20 points)

- Write a Jupyter notebook using R/Python to graphically represent different summaries of data. Summarize your findings in this notebook.
- Summarize your key insights about different user profiles, states, loan amounts etc.
- Create a Data scientist view of Power BI dashboards to illustrate your key insights

Note: If the data created by the pipeline is huge, extract summaries/portions that are interesting using Python/R and then run PoweBI on the summaries.

Part II: Building and evaluating models. (100)

Your next goal is to build a model to predict interest rates. You will get leads from people with different profiles and you must decide if you will give loans or not and if you will give a loan, how much interest you would charge for those loans.

Classification (25 points)

Use the "Loan Data" and the "Declined Loan Data" datasets to build classification models that will generate a flag whether to give a loan or not.

- Start with logistic regression using Jupyter and Python/R
- Compute ROC curve and Confusion matrices for training and testing datasets and comment on the results.
- Repeat this using Random Forest, Neural Network models and SVN algorithms.
- Choose one model you will deploy and implement this model on the Microsoft azure machine learning studio and create a REST API
- You should be able to a new record (You can define what features you will use) and the result will be a flag whether you would give a loan or not.

Clustering (25 points)

Once you have decided to give a loan, you should build models to decide what interest rate to give. You are debating whether to create one model for all customer prospects or segment data into clusters and then build prediction models specific to each cluster. You think of creating segments or clusters and build models one for each cluster. Your brainstorm with your team and come up with 3 possibilities

1. Segment data into clusters (you define how many) **manually** using categorical or numerical features. For example, you can segment by state, by ownership of home, by average dti or a combination of features.
2. You use a clustering algorithm (that can factor both numerical and categorical variables) and segment data into k clusters. You will then build prediction models for each cluster.
3. No clusters; Just use data as is

Once you do the clustering use t-sne to visualize your clusters for some sample test data. See http://distill.pub/2016/misread-tsne/ for guidance on using t-sne

Prediction (25 points)

- Write a prediction script in a Jupyter notebook in R/Python that builds a Regression model for the interest rate using data from the 3 clustering methodologies you worked
    - Try variable selection and build the best model for each segment/cluster (Note: You may have many segments and each model may have different coefficients based on the clusters

    used to train. You should automate it. Try parallel computing libraries to make things go faster)

    - Compute MAE, RMS, MAPE for training and testing datasets
    - Repeat this using Random Forest, Neural Network models and KNN algorithms.
    - Choose the best model amongst the 4 types of algorithms.
    - Deploy the best algorithm/algorithms on Azure ML studio
    - You will have a bunch of Rest APIs you should be able to choose from based on the cluster the record belongs to

Deployment (25 points)

Design the following workflow:

- Given a record, use a pre-trained clustering model to cluster the record to a segment.
- You will have 3 cluster assignments (1-manual, 2-based on your clustering algorithm, 3-default 1-cluster for all data)
- For each cluster, there should be a RestAPI which is linked to a chosen prediction model. Look up that API and use it to predict the 3 distinct interest rates.
- Select the highest interest rate and return it as your prescribed interest rate.

We will go through Part 2 in more detail next week during the lecture

Good luck!