**Homework#0**

Due on Feb 9th, 2016

6:00pm

Submit Via NYUClass and bring a printed copy to class

- Questions from the readings and usecases:

**Hash Joins**

This is a fact: Distrusted architectures existed before Google and MapReduce. Hash Joins existed before Hadoop and MapReduce. Nowadays, most Relational Database Management Systems (RDBMS) offer Hash Joins instead of SQL Joins. Hash Joins is method for joining large data sets.

**Explain in your own words the concept of Hash Joins. Provide an example.**
**Describe the differences between Hash Joins and Normal SQL Join.**
**Explain this statement: *the computational cost of a Hash Join depends on the cost of building the hash table.***

**Twitter Predicts the Stock Market**

In the work mentioned in class for correlating sentiment analysis with stock market data, the authors relied on *a correlation/causality algorithm*. Briefly explain the algorithm in your own words. Provide an example.

**Flu Detection by Google**

Describe the overall architecture of the analytics behind Google trends and its prediction to Flu Outbreaks.
Explain in your own words why the Flu Outbreak detection analytics platforms by Google failed.

- Conceptual and High-level Architectural Questions

1) Explain and define the following concepts in ***your own words***. Provide examples/scenarios. The answer for each term should not exceed a page per concept (and a minimum of a paragraph per concept). You can also compare and summarize your comparison in a table. Please cite all your sources if you used any references including the required and recommended textbook.
    1. Big Data
    2. Predictive Analytics
    3. Analytic
    4. Data Science
    5. Data Mining
    6. Machine Learning
    7. Statistics
    8. Business Intelligence
    9. Cross-validation
    10. Confusion Matrix
    11. Unstructured Data
    12. Structured Data
    13. Semi-Structured Data
    14. Data Clustering
    15. Stream Mining
    16. Data Classification
    17. Supervised Learning
    18. Correlations
    19. Distributed Systems
    20. Unsupervised Learning
    21. Training Data
    22. Test Data
    23. Recommender Systems
    24. Trust Based Recommender Systems
    25. Biologically Inspired Data Mining
    26. Knowledge Discovery
    27. Class Label (in Data Classification)
    28. Standard Deviation
    29. Variance
    30. ETL Jobs
    31. SQL
    32. NoSQL
    33. RDBMS

- Why is Data Science is considered a Science? (please limit your answer to maximum one page – Please site all your sources)

2) (High Level Architectural Question) During the first lectures (*Analytics Project Lifecycle)*, we discussed high-level architecture of building data analytic solutions. We discussed two major use-cases. The first one is around predicting topics of textual data (email, project reports, news articles...), the second is on applying analytics to derive customers' segmentation for maximizing business profit. Consider the scenario where you are part of a data science team that will assist marketers to design a data-driven marketing strategy.

As part of the first phase in the Analytics Lifecycle, the data science team had several meetings with the marketers at your company -- which are the internal clients of this project. The following points summarizes meetings' minutes:

- Your business generally breaks down its customers into five types: loyal customer, discount customer, wandering customer, need-based customer, or a high-returned-items customer.
- Your business is eager to predict ahead of time its customers' type. And, depending on the customer's type, the marketer will decide on the marketing strategy he or she will adopt to target the customer.
- Loyal Customers are generally customers that shop regularly more than any other type of customers. They are the customers that make up at least 55% of the business' sales.
- Need-based customers are customers who have a clear intention on buying specific products. The items they will buy can be predicted from their previous purchase history.
- Discount Customers are customer who their purchase decision depends on the discount the business is offering.
- Wandering customers are generally customers who do not have a specific plan to purchase any specific items. The purchase patterns are unpredictable.
- High-returned-items customers are customers that has the tendency to buy multiple items and return the majority of the items they bought.
- The business has collected a *labeled Historical Data* of several customers. The data was mainly collected from online e-commerce site of the business. Customers start by creating accounts (customer registration) that has their personal information. All their transactions are being saved under their accounts.
- The business would like to predict the type of given customer after certain number of transactions made by the same customer.
- Depending on the customer predicted type, the marketers will assign their marketing strategy.

Provide an overview of the next steps in the predictive analytics lifecycle that you (the data scientist) would follow to predict the type of customer using the historical data collected by the business and the present data of the customer's behavior for a number of transactions. You do not have to go over the detailed algorithms for the model. Your overall grade for this question will be based on providing a high-level architecture that will solve this analytics problem and lead to build the predictive analytic solution.

Please find below the steps some of the issues that you might address in your answer to this problem. You do not have to necessarily follow those hints but they could be helpful to your analysis. Feel free to provide diagrams or utilize synthetic dataset of customers (Bonus +15pts for synthetic or real dataset) that you can create to explain how your approach works.

Data Understanding

What type of data would you need? (Data sources, data types – unstructured, structured…
What are the data sources that you will adopt to solve this problem?
What are the attributes that you might need to capture about each customer (e.g age, zip code, number of items purchases per a specific period...?)
What are some of the data problems that you might encounter? And how would you solve them (Data preparation Chapter)
Provide the overall architecture of your data matrix.
How would you reduce the data? (Refer to the Data Reduction Chapter)

Data Preparation

What are some of the techniques that you would adopt to prepare your data?
Think of feature selection, feature extraction...
How can *Singular Value Decomposition* be applied to your prepare your data?

Modeling

You can give a high-level overview of where data clustering, data classification can play a role in the core model that will predict the customer type.
You can provide an architectural diagram of the model. (Refer to the first lecture on supervised learning, unsupervised learning – Refer to the reading on Earthquake prediction using Twitter – Set#1 of the readings and how the classification model was used).

Evaluation

Provide an overview of this stage of your project. Refer to the cross-validation we used in the first practice and ***provide other methods (brief explanation)*** that you can also use to evaluate your model.
(Training Data, Test Data, ..)
How would you measure your results vis-à-vis the business objectives?

Model Deployment

Model creation does not end the predictive analytics lifecycle. Research and provide methods where you will deploy your model (refer to the CRISP-DM model deployment phase). How would you design a system that will leverage and make use of the model you created?