CAPSTONE IDEA 1

**(1) What is the business problem?**

Can you identify the total sales for every listed product and store in the next month?

**(2) Who are the intended stakeholders, and why is this problem relevant to them?**

Retail stores (maybe). Predicting sales can help planning actions, budgets, and targets.

**(3) Where are the datasets available from?**

Kaggle

<https://www.kaggle.com/c/competitive-data-science-predict-future-sales/overview>

**(4) What data science approaches do you anticipate you will use to model the business problem as a data science problem? (\*)**

Time Series analysis and forecasting using Auto Regressive Integrated Moving Average (ARIMA)

**- How do you anticipate that you will evaluate the performance of each of the data science approaches that you currently envision?**

Mean Square Error &

Root Mean Square Error (to predict the sales with minimum error (i.e penalize high errors) so that inventory can be managed properly)

**(5) How do you anticipate that the intended clients will use the results of your CP2 to address the original business problem?**

Stores can use ensure the business as usual level they are going to achieve if nothing changes in their strategy. Moreover, they can calculate the incremental value of their new actions on top of this benchmark.

Second, it can be utilized for planning. They can plan their demand and supply actions by looking at the forecasts. It helps to see where to invest more.

Also, it is an excellent guide for planning budgets and targets.

Capstone Idea 2

(**1) What is the business problem?**

Which 5 movies can be recommended to a user that he will enjoy watching, based on the choice of a movie he liked previously?

**(2) Who are the intended stakeholders, and why is this problem relevant to them?**

**The intended stakeholders could be such companies who offer streaming services online like Youtube, Netflix, etc**

Recommendation systems are used not only for movies, but on multiple other products and services like Amazon (Books, Items), Pandora/Spotify (Music), Google (News, Search), YouTube (Videos) etc

**(3) Where are the datasets available from?**

a. TMDB Dataset(Kaggle) about 5000 Movies data(43.62MB)

(<https://www.kaggle.com/tmdb/tmdb-movie-metadata>)

b. 45,000 movies listed in the Full MovieLens Dataset available on

i. Kaggle for data until July 2017)

<https://www.kaggle.com/rounakbanik/the-movies-dataset>

ii. Latest data available on grouplens<https://grouplens.org/datasets/movielens/latest/>

**(4) What data science approaches do you anticipate you will use to model the business problem as a data science problem? (\*)**

COLLABORATIVE FILTERING (matrix factorization approach using Singular Value Decomposition algorithm)

**- How do you anticipate that you will evaluate the performance of each of the data science approaches that you currently envision?**

Need to understand the criteria more, but one of the possible ways is to be minimizing Root Mean Square Error between the predicted ratings and actual ratings.

Would also like to explore others I read about including Recall, Precision, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (DNCG)

**(5) How do you anticipate that the intended clients will use the results of your CP2 to address the original business problem?**

Recommendation systems have the potential to change the way websites communicate with users and to allow companies to maximize their ROI based on the information they can gather on each customer's preferences and purchases

CAPSTONE IDEA 3

**(1) What is the business problem?**

Given past credit card transactions with the knowledge of the ones that turned out to be fraud, which new transaction can be marked as fraudulent?.

**(2) Who are the intended stakeholders, and why is this problem relevant to them?**

Credit card issuing companies

Fraud is a major problem for the whole credit card industry that grows bigger with the increasing popularity of electronic money transfers. To effectively prevent the criminal actions that lead to the leakage of bank account information leak, skimming, counterfeit credit cards, the theft of billions of dollars annually, and the loss of reputation and customer loyalty, credit card issuers should consider the implementation of advanced Credit Card Fraud Prevention and Detection methods. Machine Learning-based methods can continuously improve the accuracy of fraud prevention based on information about each cardholder’s behaviour.

**(3) Where are the datasets available from?**

Kaggle (<https://www.kaggle.com/mlg-ulb/creditcardfraud>)

**(4) What data science approaches do you anticipate you will use to model the business problem as a data science problem? (\*)**

Supervised learning using Logistic Regression and Random Forests

**- How do you anticipate that you will evaluate the performance of each of the data science approaches that you currently envision?**

By evaluating the following metrics:

Accuracy: To check fraction of predictions the model gets right.

Recall: Out of the fraudulent transactions, what percentage of these are correctly identified by the model?

Precision: Out of all the transactions predicted to be fraudulent, what percentage were actually fraudulent?

F1 Score: Determine the performance of the model by combining Recall and Precision into one metric.

**(5) How do you anticipate that the intended clients will use the results of your CP2 to address the original business problem?**

They can train the Fraud Detection model continuously whenever new data arrives, so new fraud schemas/patterns can be learned, and fraudulent data detected as early as possible.