DATA SCIENTIST SCREENING TEST QUESTIONS

As part of the recruitment process of a data scientist, the first level of screening involves a series of questions which revolves around the life of a data scientist, testing you from the fundamentals to the core elements of machine learning.

GENERAL QUESTIONS

This section involves basic programming and general theoretical tasks.

1. Write a simple program to implement selection sort.
   a. The solved program can be found under the attached file '**just_analytics__selection_sort.py'**

2. Given a function rand5() that returns a random int between 0 and 5, implement rand7()

   a. The program has been solved and will found in the file named as '**just_analytics__rand7.py'**

3. Write a function that sorts a stack (bonus: sort the stack in place without extra memory)

   a. The program has been solved and will found in the file named as **just_analytics__sort_stack.py'**

4. Implement a linked list (with insert and delete functions)
   a. The program has been solved and will found in the file named as **just_analytics__linked_list.py**'

5. Give examples of data cleaning techniques you have used in the past.

   Below are some of the data problems & cleaning techniques I tend to work with frequently:

   1. **Imbalanced Data [Oversampling / Under-sampling]:**
      This is a classification-related problem that typically refers to situations where the classes / categories are not represented equally in the data.
      E.g. Was recently working on a Census & Income-related project, whereby the target categorical field, related to an individual's Total Income, was highly imbalanced. There were only 2 unique categories present, but their proportion was highly skewed (84:16). In order to overcome this problem, I oversampled the minority class and trained the data on an equally balanced data set.
      While the overall accuracy decreased a bit (in test), the Precision increased dramatically.

2. **Highly Correlated, but not Causational, attributes [Keep the least impacting attributes]:**
   Herein, if there exist 2 or more attributes, in the data, which have a strong correlation amongst them, we must consider only the one which least impacts the other attributes – to reduce chances of one influencing the other(s) and, further, the eventual models. E.g. While working on the Census & Income-related project (mentioned above), the Gross Income was strongly correlated with the Age. Since we were unaware of the causational relation between the two, we kept Age – discarding Gross Income – as Age had a lower correlation with the other attributes.

3. **Dirty Data [Range of techniques (data imputation, clarification with stakeholders, deletion of data {if small quantities})]:**
   Majority of the times, while companies tend to store and keep digital data, nowadays, the data is often dirty & malformed. This can be due to a variety of reasons – ranging from data entry mistakes, spelling mistakes, system(s) breaking / malfunctioning, et al. E.g. In the Census project above, there were loads of spelling & typing errors – due to the manual entry of specific fields (such as address & pin code) which resulted in multiple rounds of clarifications with the stakeholders, followed by eventual dropping of those rows.

4. Could you please provide details about precision and recall in a machine learning context with an example?

   Both, precision & recall, are metrics used when evaluating Supervised + Classification-type Machine Learning problems.

   Precision is the ratio of the correctly predicted positive observations (*True Positives*) to the total predicted positive observations (*True Positives* + *False Positives*).

   Precision = TP/TP+FP = (True Positives/ True Positives + False Positives)

   Recall also called as Sensitivity is the ratio of correctly predicted positive observations (*True Positives*) to all observations in the actual class (*True Positives* + *False Negatives*).

   Recall = TP/TP+FN = (True Positives/True Positives + False Negatives)

   E.g. Let us consider the confusion matrix below, which can represent the classification of a tumour as malignant or benign.

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

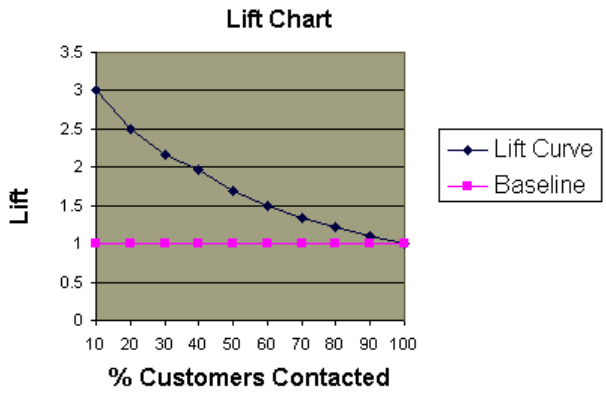
Hereby, we can see that the Precision score = 100/(100+10) = 100/110 = 90.91%
This implies that when we predict a tumour as being malignant, we are correct 90.91% of the times.

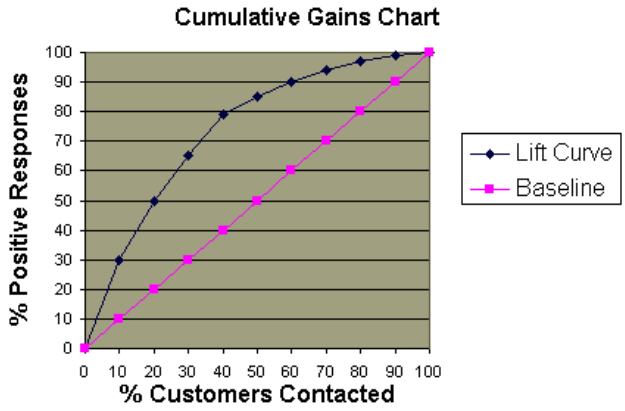
On the other hand, we have a Recall score = 100 / (100 + 5) = 100/105 = 95.24%
This implies that when a tumour is actually malignant, we correctly predict it 95.24% of the times.

5. If the CEO of a company asks you what model 'Lift' is, how would you explain it to him?

Lift is a **measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model**.

The Lift Curve in the Lift Chart indicates the True Positives the model has correctly predicted – as compared to the baseline approach (generally, non-usage of the model): hence, implying the advantage the model provides, as we consider more of the sample set (e.g. % of customers contacted).





Lift charts alongside Cumulative gain charts are good visual aids for measuring model performance.

DATA EXPLORATION CHALLENGE

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a dataset, which is exactly what this task demands.

The task is to download the dataset and do an exploratory analysis by using either R or Python. The dataset can be downloaded from **Beer Reviews**

Some of the questions you will need to answer based on the data set are:

1. Which brewery produces the strongest beers by ABV%?
   a. The Brewer ID that produces the strongest beer by ABV% is 6513.

2. If you had to pick 3 beers to recommend using only this data, which would you pick?

   a. The 3 beers to recommend using only this data would be as follows:

      1. Founders KBS (Kentucky Breakfast Stout)
      2. Pumpkin Ale
      3. Founders Breakfast Stout

3. Which of the factors (aroma, taste, appearance, palate) are most important in determining the overall quality of a beer?

   a. As per Recursive Feature Elimination, the factors that are most important in determining the overall quality of a beer are Appearance & Taste.

4. Lastly, if I typically enjoy a beer due to its aroma and appearance, which beer style should I try?

Bonus points if you can come up with your own set of questions and relevant insights.

a. The beer style that you should try due to its aroma and appearance is American Double/Imperial Stout.

DATA SCRAPING AND TEXT ANALYSIS

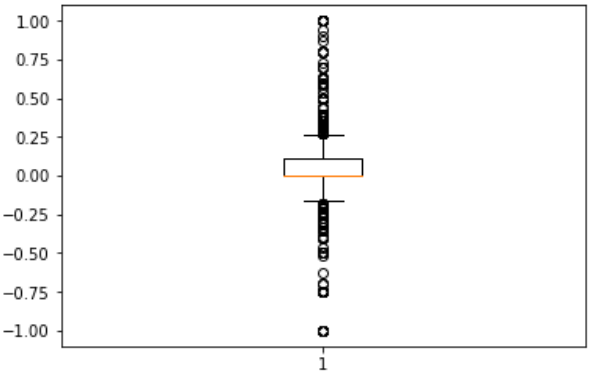
This task is to test your mettle in the areas of using APIs and then subsequently performing text analysis. You are required to scrape data from **Twitter API** for the following two airlines and store the data in a flat file.

- Air India
- Singapore Airlines

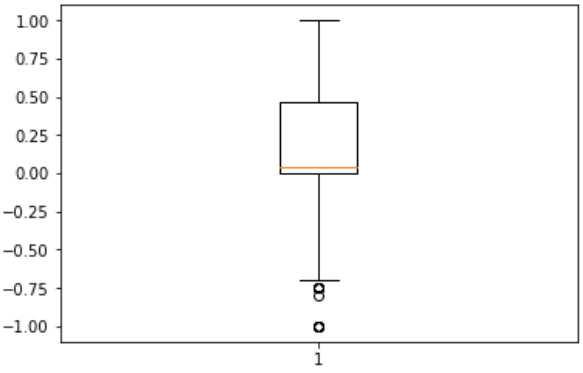
Further you are required to perform the following text analysis on the scraped data and answer the following questions (Applicable for both the airlines):

a. Are the customers satisfied with the services offered by the airlines?

1. Both airlines' customers' tweets indicate a nominally positive sentiment towards the companies.
2. SQ / Singapore Airlines received a mean sentiment score of +0.171 while AI / Air India received a mean sentiment score of +0.012 (almost neutral).
3. Constructed box & whisker plots (attached below) of the sentiment score for each airlines' tweets.



a. Above plot indicates that AI's tweets were equally positive & negative, with a relatively small IQR (Inter-Quartile Range) – implying that ~50% of the tweets had a sentiment score between 0.0 and +0.125.



b. Above plot indicates that SQ's tweets were far more positive than negative, with a relatively large IQR (Inter-Quartile Range) – implying that ~50% of the tweets had a sentiment score between 0.0 and +0.50.

4. Additionally, the above figures are all **sentiment scores (with a max & min value of +1.0 and -1.0 respectively)**; once we are provided a scale to classify these sentiment scores into categories (such as satisfaction / dissatisfaction), we can easily determine the

customers' satisfaction with the respective airlines, on the basis of these scraped tweets.

b. Can you extract the most relevant reasons as stated by the tweets related to why the customers might be satisfied or dissatisfied with the airlines?

1. Based on the tweets that each company attracted, we were able to zero in on the top words that helped determine the above sentiment scores:
    a. Singapore Airlines' frequently used & influencing words:
        i. Top 3
            1. airbus
            2. itsjoelpimental
            3. a350
        ii. Word cloud:

    b. Indian Airlines' frequently used & influencing words:
        i. Top 3:
            1. extremely
            2. praful_patel
            3. investigated

ii. Word Cloud:

For each of the questions above, please attach relevant codes written in R/Python 2/3, appropriate visualizations along with a short write-up about the approach taken.

(Twitter API Link : https://developer.twitter.com/en/docs)

TIME SERIES ANALYSIS & FORECASTING

The given dataset consists of cheese production volumes for three different types of cheese from the year 1995 to 2013. The data is clean but due to data collection issues has missing data for a few months. Your task is to perform a time series analysis on the historical data and forecast the cheese production volumes for the next year one. You can write your codes in Python 2/3 or R.

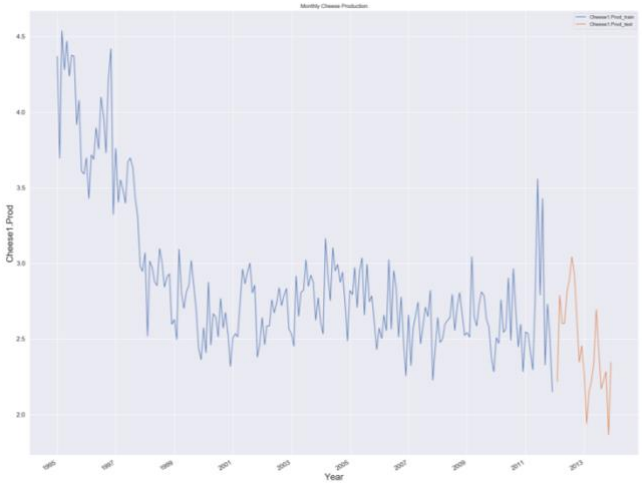
Also document the steps you took during data cleaning, imputation if any, analysis and forecasting and why you think the method you used is good for this scenario over other methods you might be aware of.

Dataset (Use the files provided if you cannot access the dataset below)

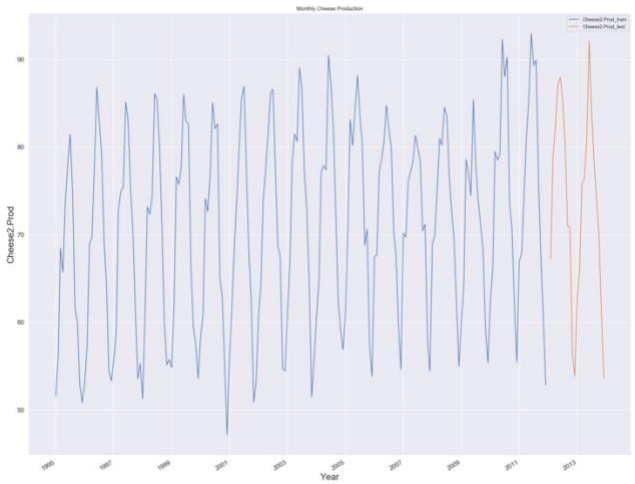
Cheese_Production_Data.txt

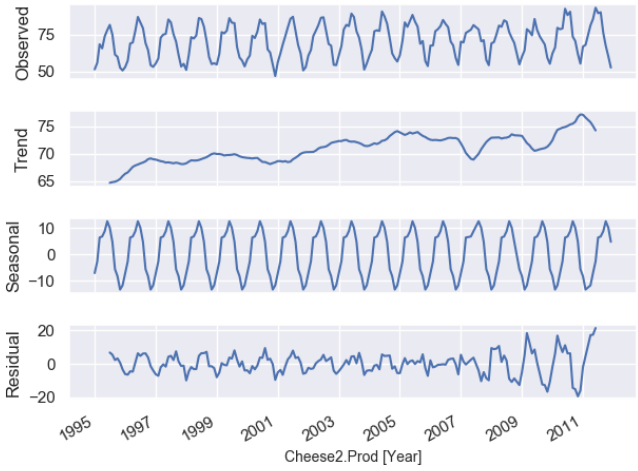
While working on this Time Series problem, the following are some of the key takeaways:

- **Insights**:
  - We are missing data for 6 months:
    - April 2007
    - July 2007
    - November 2008
    - March 2011
    - January 2012
    - September 2013
  - Due to time constraints, we decided not to impute these 6 data points (they formed <3% of the total possible data points)
  - Cheese1.Prod exhibits small variations annually, & stable / stationary, yet decreasing trends



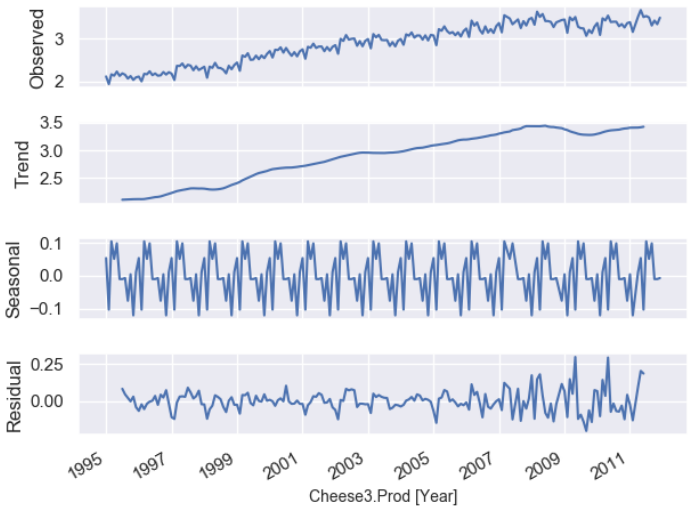
  - Cheese2.Prod --> exhibits large variations annually & almost constant / nominally increasing average (yearly) production trends. This is NOT a stationary time





series – even the DF test proved it.

- Cheese3.Prod --> exhibits small variations annually & increasing production trends. This is NOT a stationary time series – even the DF test proved it.



- **Techniques used:**
  - *Simple Exponential Smoothing*
    - Considers the entire history of all data points, while predicting – as opposed to Moving Average, Simple Average & Naïve approach
  - *ARMA*
    - Used in case the series is already stationary (e.g. while predicting for Cheese1.Pred)
  - *ARIMA*
    - Helps make a series stationary (via the I [Integration] component)
    - Aims to describe and consider the correlations in the data with each other, to make predictions

CODING CHALLENGE – FEATURE ENGINEERING & MACHINE LEARNING

The following dataset (csv) consists of credit card customers belonging to a financial institution in Taiwan. It consists of payment, history, credit and demographic information tagged to whether the card defaulted or not previously. The task is to build a prediction model around this data to predict the probability of default. From the perspective of risk management, the probability is considered more valuable than the binary classification itself. *You are free to use either Python 2/3 or R for coding.*

You will be tested on:

1. Feature Engineering –
    a. What features would you derive?
        1. The features that I would derive would be as follows:
            a. Pending Balance Amount, for each month
            b. Pending Average Balance, over the last 5 months
            c. Check who all have a current / latest Bill Amount > Maximum Credit Limit allocated by the bank.

    b. How you evaluate those features?
        1. Balance_amt1 = BILL_AMT2 – Pay_AMT1 [i.e. Past month bill – current payment amount]
        2. avg_bal = total balance for 5 months/5
        3. limit_crossed = LIMIT_BAL < BILL_AMT
    c. Techniques used.
        1. Used numerous Data Wrangling concepts baked into Pandas (e.g. conditional slicing and indexing), to derive the features
2. Machine Learning Models –
    a. Different algorithms and approach
        1. The different algorithms and approaches that were used for this project are as follows:

            a.Random Forest: It is one of the most popular & powerful **learning** algorithms currently – due to its ability to generate numerous decision trees against the same data and select the forest that best fits the current data set & target variable.
            It has proven itself to be a highly accurate classifier & can even run efficiently on larger datasets.

            b. Logistic Regression: The Swiss army knife of Classification problems, this approach does not assume a linear relationship between the Independent

& Dependent Variable. Additionally, the Dependent variable need not be normally distributed.

c. Decision Tree Classifier: Implicitly performs feature selection, requires relatively little effort from users for data preparation. Easy to interpret & helps fit the resulting rules to the provided data set quite well .

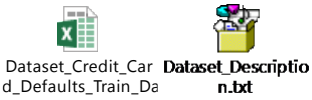
3. Optimizing the Models –
    a. What are the different approaches used?
        1. Approaches used were:
            a. Over sample the data → helps work with biased data and reducing prediction of only the majority class
            b. Then use the 3 above algorithms [RF, Logistic, DT] for classification → Predicting using all 3 helps prevent incorporating the bias / cons that come with using only 1 approach / technique

Evaluation

You will be evaluated based on your results on the test data.

Please write theoretical answers where you see fit in an accompanying document.

Attached Datasets

For training - 

For testing -