Interview Practice_Data Analyst

By Shaswati Ghosh on 10th May,2018

1. Describe a data project you worked on recently.

The latest project I worked on was data wrangling of OpenStreetMap data.

I extracted OSM file from openstreemap.org of my residential area in 'Milwaukee County'. The file size used for total project was around 150 MB. Aim of the project was to collect the data from open source, clean it, organize it, save the cleaned data in SQL database and use SQL query to find different information from those data about the area. Do correction in wrong data if you have proper and correct information.

As the file size was big enough, hence initial analysis was done on a sample of that file. After preparing the code with reference to that sample file, actual file was used for final execution and correction.

Different problems found in the file was like improper street name abbreviation, different format of phone number's, different format of postal codes, different spelling in city name, different format of city name(with and without state extension).

I retrived data from OSM file and converted XML data into specified format as a list of Python dictionaries. Then did the correction operation on those data and finally wrote into local CSV files.

These files data were then saved into MySQL data base. Later on using various SQL query I found many informations about my locality. I went through many new things like processing OSM file to CSV and then to data base. Also doing the audit part and finding the problems in data was really interesting.

2.You are given a ten piece box of chocolate truffles. You know based on the label that six of the pieces have an orange cream filling and four of the pieces have a coconut filling. If you were to eat four pieces in a row, what is the probability that the first two pieces you eat have an orange cream filling and the last two have a coconut filling?

Answer:

According to conditional probability:

Probability of 1st choice to be orange: 6/10

Probability of 2nd choice to be orange: (6-1)/(10-1) = 5/9

Probability of 3rd choice to be orange: 4/(10-2) = 4/8

Probability of 4th choice to be orange: (4-1)/(10-3) = 3/7

Total Probablity = (6/10) *(5/9)* (4/8) * (3/7)

```
In [1]: float(float((float(6)/float(10))) * float((float(5)/float(9))) * float((float(
        4)/float(8))) * float((float(3)/float(7))))
```

```
Out[1]: 0.07142857142857142
```

Hence the total probability is 0.0714

Follow-up question: If you were given an identical box of chocolates and again eat four pieces in a row, what is the probability that exactly two contain coconut filling?

Answer:

Total 4 chocolates out of 10 can be picked in 10C4 ways. Possible combination of 2 orange out of 6 is 6C2 Possible combination of 2 coconut out of 4 is 4C2

Hence the probality of exactly 2 being coconut out of picked 4 chocholet is

(6C2 * 4C2) / 10C2 = 6/14

3.Given the table users: construct a query to find the top 5 states with the highest number of active users. Include the number for each state in the query result.

Answer:

Selecting **State** and getting it's count, grouping by **State** name and displaying the top 5 name in `descending` order.

```
In [ ]: import sqlite3

        sqlite_file = 'test.db'    # name of the sqlite database file
        con = sqlite3.connect(sqlite_file) ## Connect to the database
        curs = con.cursor() # Get a cursor object

        curs. execute("""SELECT state, count(id) as count
        FROM users
        WHERE active EQ 1
        GROUP BY state
        ORDER BY count DESC LIMIT 5;""")
```

4.Define a function first_unique that takes a string as input and returns the first non-repeated (unique) character in the input string. If there are no unique characters return None. Note: Your code should be in Python.

Approach to Code logic:

Finding the count of each character of the given string in first loop and populating the dictionary **Counter** (holding the character and it's count) and putting each character of string in another list **Characters** in the same order of string. This order is needed to find the **first** single character. Then looping on character list and if count value 1 is found, returning the character and exiting from the loop. If nothing is found with count 1, returning **None** after completing the loop.

```
In [12]: def first_unique(string):

             Characters = []
             Counter = {} #Dictionary containing character and it's count in string

             for char in string:
                 if char in Counter:
                     Counter[char] +=1
                 else:
                     Counter[char] =1
                     Characters.append(char)

             for x in Characters:
                 if Counter[x] == 1:
                     return (x)
                     #return (x, Characters, Counter)
             return 'None'
```

```
In [13]: first_unique('aabbcdd123')

Out[13]: 'c'
```

```
In [14]: first_unique('a')

Out[14]: 'a'
```

```
In [15]: first_unique('112233')

Out[15]: 'None'
```

5.What are underfitting and overfitting in the context of Machine Learning? How might you balance them?

Answer:

Underfitting defines the situation when the model does not follow the trend in data. That means, even for training data set performance score is not good.

Overfitting happens when the model algorithm follows the training data set so well that it fails to generalize the trend for test data. Overfitted model shows very good result on training dataset, but poor results for test data set.

One of the main reason for overfitting/ under fitting is choice and count of features. Too many features or presence of noise in data can lead to overfitting of data. Similarly choosing features having very less effect on performance or less number of feature can lead to underfitting of data. Use of right number of correct features can balance underfitting and overfitting. Based on given primary features we can derive some features which can have better influence on test result. Few methods of selecting best features are `SelectKBest`, `feature_importances_ with noo-NUll importance` etc.

In general concept, we never get enough data to fit our model, removing a part for test or validation from train data set always causes a risk of underfitting. In k-fold cross validation, every time, one of the k subsets is used as the test data set and the other k-1 subsets are used as training set. This makes the model trained over all the data set and error estimation is averaged over all k trials to get the final performance matrics.

If different features value are of wide ranges, scaling or standardizaton is required to achive better result. Different scaling mechanism like MinMaxScaler or Normal Standardization can be used.

6.If you were to start your data analyst position today, what would be your goals a year from now?

My goals would be:

I want to do coding fluently in latest version of Python 3 and R-3.4.2 or later.

I would like to accquire the expertise in SQL as project work in Progrexion needs expert in this area.

I want to make myself expert on data collecting, cleaning and organizing of data to make the project work smooth.

Apart from personal carrier improvement, I would like to learn all the project specific rules and would attain zero delinquency in meeting deadlines.

I will work hard to become a senior data scientist.

In [ ]: