Data Science & Data Engineering Assessments

Instructions

- Complete the following questions and submit the solution in jupyter notebook (.ipynb) or python(.py) file depending on the questions you answered, you may compress multiple files into zip or tar.
- Submission shall be sent to the communication email between the candidate and Involve Asia when applying for the position. Should you receive no email communication from Involve Asia, you may submit your solution here.
- The submitted solution will be reviewed within a week.
- Copy the respective question as markdown/comment in the jupyter cell before writing down the solution as the questions might be revised from time to time.
- You may use external libraries whenever necessary.
- On average these assignments take 3 days starting from the date this question is opened, however you may submit earlier than the deadline.
- Any late submission **may not** be entertained.
- Complete as many questions as possible. **Completing all questions isn't expected** but is highly encouraged.

Survey

- Read a scientific literature titled “Bidding Machine: Learning to Bid for Directly Optimizing Profits in Display Advertising” from this link https://arxiv.org/pdf/1803.02194.pdf and answer the following survey.

    a. Have you published any research or scientific papers before? If yes, state the title or the link to the publication.
        i. No

    b. Based on your reading of the above scientific article, on a scale of 1 to 5, rate your level of understanding.
        i.

Scale	Description	
1	Not a clue	
2	Mostly don’t understand	

3	Understand non-mathematical parts	
4	Understand most parts	
5	Understand all parts	

c. If you were given a task to implement the techniques described by the paper into proof of concept how do you describe the task on a scale of 1 to 5.

i.

Scale	Description	
1	Unable to complete	
2	Able to code by myself but require mathematician help to implement	
3	Unable to code by myself but require engineer help to implement	
4	Able to implement most parts	
5	Able to implement all parts	

1) Machine Learning. (Ideally use a jupyter notebook). Tips: Beating the RMSE is one criteria but more points are given to describe the logic and process to obtain it.

   a) Download ***train.csv*** and ***test.csv*** from here https://goo.gl/do3PJY
   b) Perform simple Exploratory Data Analysis on ***train.csv*** and ***test.csv***
   c) Use a suitable algorithm to train a machine learning model from the ***train.csv***
   d) Interpret the model by showing its metrics.
   e) **Test.csv** contains extrapolated data. Use the trained model to infer/predict the y values in ***test.csv***. You may use y_hat as the predicted values column and retain the original y column to be used to complete the next step.
   f) Write down your train and test RMSE. You may attempt to improve your RMSE to as low as possible by experimenting with different ML algorithms or techniques, you may retain all the codes used to test different models and only pick the best result. For reference the base RMSE of **test data** is 25.36 and average is 15.99 and lowest recorded is 0.24

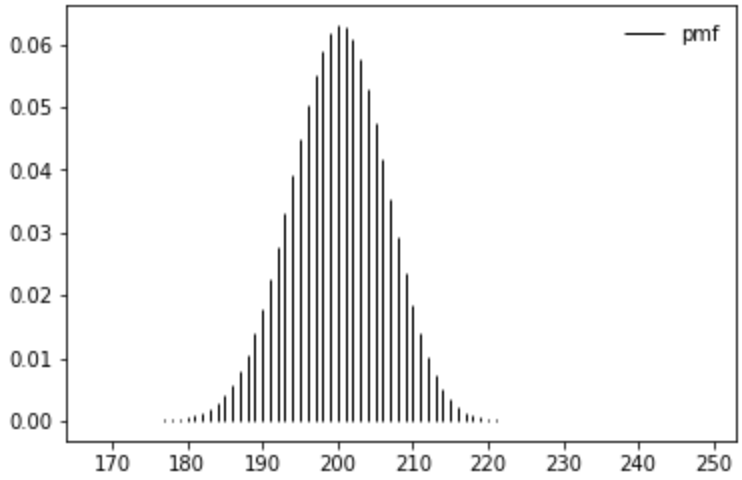
2) Web & data engineering. You **must** use any python web framework to accomplish this task i.e flask or django.

a) From the question (1) above, create a Command Line Interface (CLI) script to automate the **steps** of downloading the training data, perform training tasks and finally save the trained model into file i.e pickle. Your CLI must accept an argument to skip the download step to use the existing downloaded training file.
b) Create a single page dashboard to display:
    i) A simple visualization from the training set
    ii) Input for user to manually enter x1 and x2
    iii) Display the inferred value (y) to the user
c) The visualization **must** be interactive e.g. by utilizing plotly
d) Using AJAX for input/output will be an added bonus.
e) Elegance solution i.e by showing correct way of using data structures and object oriented design will be an added bonus too.
f) You may use your imagination to accomplish this task as long as the above criterias are met.

3) Probability and uncertainty. (Ideally use a jupyter notebook). Note: dark **red bold keywords** are probabilistic or statistical term

a) Airlines are commonly known to utilize overbooking to earn extra profits or avoid losses over the fact that not all passengers will show up on boarding day. We will use a simple scenario here to demonstrate how airlines might do so.
    i) The passenger travel in this scenario is assumed to be alone. Factors such as weather, festival season, time of the flight and other external factors are assumed to not affect our model.
    ii) Historically passenger showed up rate is 80%
    iii) The number of tickets to be sold without overbooking is 200, and the cost of a ticket is $120.
    iv) The management of the airline is planning to overbook the seats by additional 50 tickets to achieve full capacity and avoid losses. Hence the total ticket to be sold is 250. (0.8x = 200, hence x = 250)
    v) However the risk is if more than 200 passengers showed on boarding day, the airline must compensate each bumped passenger by $300.

b) The revenue without overbook is $24,000 ($120*200), while for **best case** scenario where 250 tickets are sold and 50 no show passengers is $30,000 ($120*250). Show the revenue calculation for **worst case** scenario
c) Using the python library from matplotlib/seaborn and scipy.stats write codes to plot the **probability mass function (**pmf) for the above scenario. Hint: Your diagram should closely resemble the diagram below.



d) What is the probability of passengers showed up for:
   i) 200
   ii) 250

e) What is the **expected revenue** of 250 tickets sold?

f) Is 250 overbook tickets the most optimal number? Run a simulation of **expected revenue** from 201 to 260 tickets sold.

g) Plot the number of tickets sold against expected revenue and draw a vertical line to mark the optimal number of tickets sold.

h) By using the optimal number of tickets from **(f)** and with the shown up rate of 80%, generate 10,000 random numbers to simulate the possible number of passengers that show up . (hint : `.rvs(n, p, size=10000)`)

i) What is the min, max and 95% percentile of the number of passengers showing up from **(g)**?

j) What if the passenger isn't traveling alone, describe how it affects the model.

k) For the solutions, you **should** as much as possible minimize the usage of loop and instead be replaced with matrix computation (i.e numpy multiplication).

l) From your observation or research on Involve Asia, how do you think this scenario relates to the business model?

4) ABC is an online advertisement firm, the management found adblocker used by the users is affecting the revenue. The firm has invested significantly to distribute custom adblock whitelists. Below is the table showing the number of sample adblock detected before the distribution of the first 3 months (12 weeks) and the next 3 months after the whitelist has been distributed. (Ideally use jupyter notebook)

Week	1	2	3	4	5	6	7	8	9	10	11	12
Count	3100	2800	1900	2400	3200	2700	1600	4100	2300	3200	2900	3300

Week	13	14	15	16	17	18	19	20	21	22	23	24
Count	2300	2100	1900	2400	3500	1700	1800	2400	3300	2700	2100	2300

a) Convert the above information into a numpy **array** and compute the **average** difference between the first table and the second table.
b) Could the difference legit and not due to chance? Use a statistical approach to demonstrate your assumptions.
c) Because the firm already spent vast amounts of money, is the difference in the number of adblocker detected for the next 3 months statistically **significant** and worth the investment? You may provide your opinion here.
d) What's the **confidence intervals** of the average difference? You may use 90% or 95% percentile for confidence intervals, state this percentile in the solution.