Data Pre-Processing Test

Background

This test is designed for Data Scientist prospects to see your approach in handling raw data and develop the appropriate model for the specific client's use case. It is modelled in a case study format derived from the problems we have experienced and solved in our past projects.

We believe that this test mimics the position's responsibilities at ilmuOne Data. The data that our clients collect often do not align with their business goals or the analysis that they ask our team to perform, and thus, sometimes we have to do a bit of data wrangling before we are able to derive a model that is capable of serving the client's needs.

By completing this test, you will help us gain a better understanding of your potential with regards to this vacancy. Aside from your programming skills, you should consider this case study as a chance to show your problem-solving skills and your ability to think outside the box. In other words, you are free to infer things that are not directly stated in the case study. You are also welcome to state your opinions or any assumptions you might hold as long as it's still relevant to the task in hand.

Please note our evaluation will not be limited to your final submission. We will also consider intangibles including, but not limited to, your effort and professionalism. Our evaluation will also be subject to our judgement of your experience based on your CV and first interview. This case study will also highlight your ability to tackle new concepts. Even if you are unable to finish the test, please submit your best possible attempt. Good luck!

Instruction

Case Study

You had just recently completed the induction training in IlmuOne Data and you're immediately put in charge as the leading data scientist for a new client, a home appliances distributor company called PT. Denki Kobo.

PT. Denki Kobo is one of the oldest and largest distributors of Japanese home appliances in the country. They sell a myriad of home appliances from air conditioners, refrigerators, rice cookers, washing machines, vacuum cleaners, etc. The CEO of the company recognizes that in these modern times, businesses have a lot of insights to derive from analyzing their data. Regrettably enough, until this day, they have yet to employ proper data analysis techniques in their business. Therefore, they have requested our assistance in providing them with state-of-the-art data analytics solutions.

As the client is still in the middle of transitioning into the new age of data, they are well aware that the data they currently have is less than stellar. Although they have started keeping track of daily sales data since early 2010, the data was manually recorded by the company's sales managers, and as such, owing to human errors, there may be some days where the sales data is not available.

Since the client imports a large variety of home appliances from Japan, the daily sales data are consolidated by the set product category and maker. Interestingly, due to their lack of data keeping knowledge, the product category data is stored in 3 different columns, where a column contains only a single word and empty entries are marked as "null". That's why, for instance, the "Vacuum Cleaners" product category has "Vacuum" in the first column, "Cleaners" in the second column, and "null" in the third column.

You were also told beforehand that there is no sales data for weekends and public holidays as the warehouses only operate on weekdays. Moreover, for audit purposes, they broke down their sales data into separate files divided by quarters. Last but not least, the client has also supplied you with a small workbook containing their average margin of profit for every brand they have sold in the past 3 years.

The client had uploaded all their data in the following folder (attached in the email):
IOD - Data Scientist Test 2021 - Dataset.zip

Upon receiving the data from the client, your manager would like to see you submit a data description report before notifying them that the data has been well received. You were informed that this report will not only be used internally, but it will also be presented to the client to give them an idea of the current state of their data. Furthermore, it will be a crucial piece of documentation for the rest of the data science team to quickly refer to while they are working on the project.

1. **Write a data description report.**

After consulting with the client, you were notified that the data they have is actually mixed with the data belonging to their sister company, PT. D.K. Works, before their split last year. Luckily, you were also supplied with good news that these foreign data are easily identifiable, as they can be recognized by products having **one or more digits** in their identifier code. The identifier code is a string of 10 characters that consist of **only letters** in the data belonging to the client.

For example, a product with an identifier of “CTCNSTZEDP” belongs to PT. Denki Kobo (the client), while a product with an identifier of “PRN3TK7HWK” belongs to PT. D.K. Works (the sister company). Notice that the data belonging to PT D.K. Works has the number 3 and 7 in the identifier code.

Since the client and PT. D.K. Works had already parted ways, they requested that the latter’s data be taken out.

2. **Take out the data belonging to PT. D.K. Works.**

Distribution business is very risky in the sense that demand for the goods tends to fluctuate throughout time. This uncertainty has been an obstacle to the client’s ambition to expand their business by importing more goods from Japan to be sold in the country. Importing too few hurts their bottom line and importing too many risks their warehouse being flooded with unsold goods.

It is of the client’s best interest to be able to predict the demand of the goods and ultimately the future sales number. Hence, the client would like you to employ the art of data science to help them realize this aspiration.

3. **Demonstrate a method for predicting the client's future sales numbers.**

The client had set a target to sell a total of 30 million goods before the year 2021 ends. The client then asked you, at their current sales rate, when can this goal be accomplished?

4. **Predict when, if it happens, a total of 30 million goods is sold by the client in 2021.**

Due to the unexpected COVID-19 pandemic, the client is heavily considering shutting down several of their warehouses, and as they will have limited space to store their goods, they will be forced to cut one or a few of their product lines.

5. **Suggest to the client which product(s) you think they should cut.**

Moving forward, the client would like to improve their data collection methods, and so they asked you for guidance. Additionally, knowing the limited data that they currently have, they also asked you for advice on what other data they could collect in order to accommodate deeper data analysis in the future.

6. **Propose several ways the client could follow to improve their data collection and propose to them what data they could collect to accommodate future data analysis.**

Expected Output

You need to submit your work in the following formats:

1. Script
    - Python script (.py), or
    - Python notebook (.ipynb)
2. Analysis Report
    - PDF document (.pdf), or
    - PPT presentation (.ppt)