**Cognizant Virtual Experience Programme**

Task1: Exploratory Data Analysis

Question:

**Step 1: Prepare**

**Note:**This task assumes that you are able to access Google Colab. This will require you to have or create a free Google account.

First, you must download the CSV file provided within the additional resources (sample\_sales\_data.csv). This task will make use of coding notebooks, specifically using Python, which provide an interactive way of running and sharing code. There are 2 notebooks provided (eda. ipynb & eda\_walkthrough.ipynb), “eda.ipynb” provides some code snippets to get started but will then assume that the user is able to use their knowledge to perform analysis on the data in Python. “eda\_walkthrough. ipynb” provides the same code snippets, but also provides a walkthrough guide on how to complete this task. We recommend that if you’re comfortable with Python and data analysis, use “eda.ipynb”, otherwise use “eda\_walkthrough.ipynb”. Furthermore, you can run this notebook however you wish, but we recommend that you use “Google Colaboratory” (Google Colab). A video link is provided in the additional resources covering a getting started guide with Google Colab, as well as how to import a dataset into Google Colab. This should take 10-15 minutes.

**Step 2: Exploration**

Now that you’re ready to dive into the data, you should use the notebooks provided to conduct your exploratory data analysis. Try to use a rigorous and scientific approach when analyzing data. Always ask yourself this question: what do these results mean? You should aim to have a solid understanding of the statistical properties of the dataset when you’re complete (this may include the distributions of columns, descriptive statistics such as mean, median, max and min, etc.) By the end of the exploration, you should feel comfortable with the dataset and what it means. This should take you 25-30 minutes.

**Step 3: Communication**

Finally, you must summarize your findings in a concise and business-friendly manner within an email to the Data Science team leader. This email should also contain recommendations for what else we require in order to complete this task for the client. Please draft your email in a word document and submit this file for review. You can use the template provided in the additional resources section. This should take 5-10 minutes.

**Estimated time for task completion: 40-55 minutes, depending on your learning style.**

Here are some resources to help you

**Note:** The [eda\_walkthrough] and [eda.ipynb] resources below are a specific type of file that must be opened by a Python interpreter. So this task will give you great hands-on experience with Python interpreters! A Python interpreter is a program installed on your computer whose job it is to look at and run your Python code. For more information, see [this resource](https://cs.stanford.edu/people/nick/py/python-interpreter.html) by Stanford computer science. To open these files, you should use [Google Colab](https://colab.research.google.com/).

1. "C:\Users\gmpl\Downloads\sample\_sales\_data.csv" Sample\_sales\_data Dataset
2. Eda\_walkthrough
3. Eda.ipynb
4. Getting started with colab
5. AI from Data to Roi

https://cdn.theforage.com/vinternships/companyassets/e6nrxEAa6MHFh3Jmw/DCGoJxzfdJHirTYGe/1652743458777/ai-from-data-to-roi.pdf

1. How to upload a csv file into google Colab https://www.youtube.com/watch?v=woHxvbBLarQ
2. Seaborn documentation https://seaborn.pydata.org/
3. Matplotlib Documentation https://matplotlib.org/
4. Pandas Documentation https://pandas.pydata.org/
5. Email Template

### **Learn from the Example Answer**

Great work! Here's an example answer so you can see how your work compares.

Use the example answer to:

* Learn how a professional would complete this task.
* Ensure you’ve understood the key concepts of the task.
* Reflect on what you did well and where you can improve.

Dear [insert name of recipient],

I received the sample dataset from the Data Engineering team and I’ve been analysing the sample on behalf of the Data Science team.

I found the following insights as part of the analysis:

* Fruit & vegetables are the 2 most frequently bought product categories
* Non-members are the most frequent buyers within the store
* Cash is the most frequently used payment method
* 11am is the busiest hour with regards to number of transactions

As a reminder, the client indicated that they wanted to know the following: “How to better stock the items that they sell.”

With respect to this business question, my recommendations are as follows:

* This is a very broad statement and in order to tackle this with better accuracy, we need to identify a specific problem statement that the business would like to solve. For example, can we predict the demand of products on an hourly basis in order to procure products more intelligently?
* We need more data. The current sample only covers 7 days and 1 store.
* Based on the problem statement that we move forward with, we will need more datasets to help describe the outcome that we’re trying to model. For example, if we’re modelling demand for products, we may want to include information about stock levels or weather conditions.

Best regards,

[name of sender]

**TASK 2: Data Modelling**

Your work on the previous task was very helpful to propel this project forward with the client. Based on your recommendations, they want to focus on the following problem statement:

“Can we accurately predict the stock levels of products based on sales data and sensor data on an hourly basis in order to more intelligently procure products from our suppliers?”

The client has agreed to share more data in the form of sensor data. They use sensors to measure temperature storage facilities where products are stored in the warehouse, and they also use stock levels within the refrigerators and freezers in store.

It is your task to look at the data model diagram that has been provided by the Data Engineering team and to decide on what data you’re going to use from the data available. In addition, we need you to create a strategic plan as to how you’ll use this data to complete the work to answer the problem statement.

You can summarize your choices and plan of work in a PowerPoint presentation. This PowerPoint will be sent to the Data Science team leader and the client for a review. Make sure to keep it concise (ideally 1 slide) and business-friendly.

### **Here is your task**

**Step 1: Data modeling**

Look at the data model provided in the additional resources. Look at all the data that is now available from the client and decide what you want to use for the modeling of the problem statement. This should take 10-15 minutes.

**Step 2: Strategic planning**

Come up with a plan as to how you’ll use this data to solve the problem statement that the client has positioned. This plan will be used to describe to the client how we are planning to complete the remaining work and to build trust with the client as a domain expert. If you need some guidance, use the provided resource video that describes the high-level overview of a data science project. This should take 10-15 minutes.

**Step 3: Communication**

Summarize the data that you want to make use of and the strategic plan of action in a single PowerPoint slide. This will be sent to the Data Science team leader and the client, so be sure to be concise and use business-friendly language. This should take 10-15 minutes.

**Estimated time for task completion: 30-45 minutes, depending on your learning style.**

**Necessary Document**

Data Model Diagram

Data Science Project Overview

<https://www.youtube.com/watch?v=s4uF8UOJz9k>

### **Learn from the Example Answer**



Task 3: Model Building And Interpretation:

It is now time to get started with some machine learning!

The client has provided 3 datasets, it is now your job to combine, transform and model these datasets in a suitable way to answer the problem statement that the business has requested.

Most importantly, once the modelling process is complete, we need you to communicate your work and analysis in the form of a single PowerPoint slide, so that we can present the results back to the business. The key here is to use business-friendly language and to explain your results in a way that the business will understand. For example, ensure that when you’re summarizing the

performance of the results you don’t use technical metrics, but rather convert it into numbers that they’ll understand.

### **Here is your task.**

You have previously outlined the strategic plan for completing the modeling work based on the problem statement.

It is now time for you to deliver on this strategic plan. Some additional resources have been provided to help you with this task. If you feel comfortable enough to complete this task on your own, use the “modeling.ipynb” Python notebook to get started. However, if you’re not sure where to start and would like to be guided through the modeling process, use the “modeling\_walkthrough.ipynb” notebook.

Feel free to use these notebooks in a similar way that we did before, using Google Colab. If you prefer to use them within your own development environment, that’s fine too.

The modeling process should take 35-45 minutes.

Once you’re done with the modeling, summarize your results in a business-friendly single PowerPoint slide. Be sure to explain whether this model can help them to tackle the problem statement.

This should take 10-15 minutes.

**Estimated time for task completion: 45-60 minutes, depending on your learning style.**

Here are some resources to help you

**Note:** The [modelling.ipynb] and [modelling\_walkthrough.ipynb] resources below are a specific type of file that must be opened by a Python interpreter. So this task will give you great hands-on experience with Python interpreters! A Python interpreter is a program installed on your computer whose job it is to look at and run your Python code. For more information, see [this resource](https://cs.stanford.edu/people/nick/py/python-interpreter.html) by Stanford computer science. To open these files, you should use [Google Colab](https://colab.research.google.com/).

1. Modelling .ipynb
2. Modelling\_walkthrough.ipynb
3. Intelligent Decisioning data <https://cdn.theforage.com/vinternships/companyassets/e6nrxEAa6MHFh3Jmw/DCGoJxzfdJHirTYGe/1652743643562/cognizant-intelligent-decisioning-data-driven-decisioning-for-modern-enterprises-brochure.pdf>
4. Scikit-learn Documentation <https://scikit-learn.org/stable/>
5. Sales.csv
6. Sensor\_stock\_levels

### **Learn from the Example Answer**

Chart, pie chart

Description automatically generated

Task 4: Machine Learning Production:

Gala Groceries saw the results of the machine learning model as promising and believe that with more data and time, it can add real value to the business.

To build the foundation for this machine learning use case, they want to implement a first version of the algorithm into production. In the current state, as a Python notebook, this is not suitable to productionize a machine learning model.

Therefore, as the Data Scientist that created this algorithm, it is your job to prepare a Python module that contains code to train a model and output the performance metrics when the file is run.

Move on to the next step to get started.

### **Here is your task**

Additional information about Python modules and running Python files is provided in the additional resources. You can assume for this task that the Python file does not need to process, clean or transform the dataset. The Python file should be able to load a CSV file into a data frame, then immediately start training on that data. Assume that the CSV file will contain the same columns as the dataset that you trained the model on in the previous task.

Be sure to write good quality code, this means following best practices and writing your code in a clear and uniform manner. More information about best practices are provided in the additional resources. Furthermore, make sure to document your code with comments, as this will help the ML engineering team to understand what you’ve written.

When you’re done, submit this file for review by the ML engineering team.

**Step 1: Plan**

Good quality code should be planned and should follow a uniform and clear structure. Before you start writing the Python module, take some time to think about how you want to structure this file and what needs to be included from your notebook. It is your choice whether you’d rather write it all in 1 block, separate it out into functions, or create a class with methods. Depending on your ability with coding, you may take a different approach. A starter file “module\_starter.py” is provided, which gives some hints as to how you may want to structure the file. If you’re a beginner, you may also want to make use of the file named “module\_helper.py”, which includes some functions that you may want to make use of. If you do use this helper file, you can simply copy and paste the functions that you would like to use into the module that you’re writing. This should take 5-10 minutes.

**Step 2: Write**

After planning the module that you’re going to write, you can start creating your file! Be sure to follow a consistent and clear structure, and use the additional resources for best practices. Also be sure to include plenty of comments and documentation, because the ML engineering team is not the team that wrote this code. Remember, you can assume that the Python file does not need to process, clean or transform the dataset. You can load the Python file as a CSV file directly into a data frame, and then immediately start training on that data. Assume that the CSV file will contain the same columns as the dataset that you trained the model on in the previous task. This should take you 30-40 minutes.

**Estimated time for task completion: 35-50 minutes, depending on your learning style.**

* Here are some Resource to help you
* Module \_starter <https://cdn.theforage.com/vinternships/companyassets/e6nrxEAa6MHFh3Jmw/DCGoJxzfdJHirTYGe/1652217970612/module_starter.py>
* Module\_helper <https://cdn.theforage.com/vinternships/companyassets/e6nrxEAa6MHFh3Jmw/DCGoJxzfdJHirTYGe/1652218007241/module_helper.py>
* Python modules <https://www.analyticsvidhya.com/blog/2021/07/working-with-modules-in-python-must-known-fundamentals-for-data-scientists/#:~:text=In%20Python%2C%20Modules%20are%20simply,to%20include%20in%20your%20application>
* Python best practices guide(PEP8)

<https://peps.python.org/pep-0008/>

### **Learn from the Example Answer**

<https://cdn.theforage.com/vinternships/companyassets/e6nrxEAa6MHFh3Jmw/DCGoJxzfdJHirTYGe/1652215017638/model_answer.py>

Task 5: Quality Assurance

The ML engineering team has taken your Python module and deployed the algorithm into production along with the DevOps, which is great!

Before it goes live, the DevOps team has been collecting some predictions from the algorithm and has provided these to the ML engineering team, who have performed some testing of the predictions against the actual results for ‘estimated\_stock\_pct’. The ML engineering team were testing the predictions vs actual results to see how well the algorithm is performing on “live” data.

After performing the tests, the ML engineering team wants to discuss with you some questions about the algorithm in order to further improve the model before the DevOps team integrates it with Gala Groceries’ live system.

Click next to get started.

### **Here is your task**

**Step 1: Prepare**

If you need to, remind yourself of the algorithm that you built and the code that you wrote. The ML engineering team will most likely be asking questions about how to improve the model, so you may want to do some research about how you can improve the performance of your model. If you need some support, take a look at the additional resources. This should take 10-15 minutes.

**Step 2: Answer questions**

To complete this task, answer the multiple-choice quiz. Please note there are 6 multiple choice questions to complete in this task. Please be patient as each question loads. This should take 10-20 minutes.

**Estimated time for task completion: 20-35 minutes, depending on your learning style.**

Here are some resource

1. How to improve machine learning models <https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/>
2. How to improve machine learning models #2
3. <https://www.dummies.com/article/technology/information-technology/ai/machine-learning/10-ways-improve-machine-learning-models-226836/>
4. CPG Digital Shelf Imroves Productivity
5. <https://cdn.theforage.com/vinternships/companyassets/e6nrxEAa6MHFh3Jmw/DCGoJxzfdJHirTYGe/1652743564124/cpg-digital-shelf-improves-productivity-codex4536.pdf>

Start The Quiz

**Q1/6: The accuracy of the predictions about the live data are not as good as the results that you showed during the training of the model. Why do you think this is?**

**Answer:** The model is underfit.

##### 2. Q2/6: Can you suggest a way to improve the predictions on the live data?

**Answer:** Collect more rows of data to train the model with.

##### Q3/6: The client has offered some additional datasets that may be useful. Can you let us know which two datasets you think would be best to include in the model?

##### Answer: Weather,delieveies.

##### Q4/6: We would like to explore the possibility of using a more complex machine learning algorithm to see how it compares to the current Random Forest. Can you suggest one to try?

**Answer:** Neural Network

##### Q5/6: What would be a disadvantage of using the more complex model from the previous question, against the current Random Forest?

##### Answer: More difficult to explain the algorithms ssresults.

##### Q6/6: Can you suggest a way that we can optimize the performance of the current Random Forest algorithm? In particular, we want to know how we can improve the MAE of the current algorithm.

**Answer:** Tune Hyperparameter