**Instructions for the capstone project**

You have a total of 2 **questions** for this capstone project with timeline of up to 5 **weeks** to complete it.

You can return your answer back to the program manager via google class once you are done.

This capstone project will be used to evaluate the following essential merit criteria:

* Knowledge of analytics, machine learning, data science methodologies and toolkits, and techniques to evaluate machine learning models’ performance
* Proficiency in at least one analytics scripting language such as Python, R, SAS, or other ML/statistical languages
* Knowledge of how to apply multiple advanced analytical languages and platforms (i.e. SQL, R, Python, Scala, TensorFlow, MATLAB, and SAS, H20.ai, RapidMiner, DataRobot or related tools and technologies) to produce actionable outcomes.
* *Knowledge of current data management and data mining best practices and tools*
* *Knowledge of BI reporting platforms*
* *Knowledge of data engineering, preparation and transformation platforms*

Please note that your Excellent Written Communication Skills will be assessed throughout the project. Therefore, we recommend you use complete sentences where appropriate.

If you have any questions about the above instructions, please reach out to the program manager.

**GOOD LUCK**

**QUESTION 1**

|  |
| --- |
| * *Knowledge of current data management and data mining best practices and tools* * *Knowledge of BI reporting platforms* * *Knowledge of data engineering, preparation and transformation platforms* * *Excellent Written Communication Skills* |
| **Question 1**  As part of the selection process, we are asking you to complete a data exercise.  In the attached, you will find a data folder with .csv files that contain data extracts from a hotel system (**bookings, food orders, menu, requests and rooms**). Please be aware that, like any data, some files may have errors or items could have different names, which could make data classification and analysis difficult.  Using these files, please prepare a PowerPoint deck to show your findings and recommendations about the business. Please use full sentences and proper grammar as you will be evaluated on your written communication.  We expect to see: (minimum 4 slides, maximum 8 slides)   * Slide 1: Using the attached files, design a database structure and provide this as an ER diagram (Entity Relationship Diagram) with fields in each entity on a PowerPoint slide. You can use any tool to create the ER diagram and embed this as an image for this slide or create the diagram directly in PowerPoint. * Slide 2: Provide an explanation on your design. * Slide 3: Comment on any data quality issues if any and provide your recommendation on the any data quality processes that should be performed and what checks should be performed when ingesting this data into a database. * Slide 4: The data from this hotel system will need to be prepared in order to extract insights for senior executives. Provide some recommendations on design considerations and ETL transformations (such as creation of new features, joining of data) that will be helpful for data users to more easily consume the data, particularly when using a business intelligence tool. |
|  |

**QUESTION 2**

The Analytics team has been tasked to help the Collections team in the borrowers that are rolling

forward from an early due stage to later due stages. For that you are given a dataset that contains

customers that have been in an early due stage (1-30 days delinquent) and some of their main attributes.

Your task:

1. Predict whether the loan rolled or not rolled
2. What is your prediction accuracy

**Instructions:**

Use python or R for the process (data prep and exploration, modelling...) and create a presentation

sharing the process and final observations and recommendations (google slide, powerpoint or .pdf are fine):

In the presentation, share the process and findings (alongside with the

attributes used for the classification). Also include a summary section with the highlights.

**Dataset Description**

Field Name Description

Text

Description automatically generated

***See Data folder (dataset\_risk\_analytics.csv)***