**Scenario**

You are a data scientist at a consulting firm that specializes in helping businesses leverage data for strategic decision-making. A client has approached your firm with a business challenge that requires predictive analytics to solve. Your manager has assigned you to lead this project, and you will be presenting your solution to the client's executive team. The success of this engagement could lead to a long-term partnership, so it's crucial that your approach is methodical, your solution is robust, and your communication is clear and compelling.

Finding and Selecting Data

When undertaking a data science project, you'll typically face a choice between problem-first or data-first approaches.

* In a **problem-first approach,** you begin with a business challenge and then seek appropriate data, ensuring strong alignment with business objectives but potentially struggling with data availability.
* A **data-first approach** starts with an interesting dataset and explores potential applications, guaranteeing data availability but risking solutions without clear business relevance.

For successful modeling, ensure your dataset has sufficient volume, quality, and variety. The data should be relevant to your problem domain and representative of the population you're studying. Consider striking a balance by starting with a business domain of interest, exploring available datasets within that domain, and refining your problem statement based on what the data can support.

**Remember**: Even the most sophisticated model can't overcome fundamentally unsuitable data, so thoughtful dataset selection is crucial to your project's success. You are allowed to also use more than one dataset if joining data would help create a more comprehensive project.

**Deliverables**

*Part 1-Project 2 Pitch*

The project pitch serves as a critical planning document that helps you clarify your approach and ensures alignment with project requirements before significant work begins. In particular for data science projects, project pitches are an important first step in determining if the data you have available is viable and relevant for your business problem.

You will need to include the following 3 parts of the pitch:

* Business Problem Scenario
* Problem Solving Process
* Timeline and Scope

*Part 2-Discussion MVP Project 2*

For this discussion, you'll share your project at approximately 90% completion (with your core analysis, modeling, and initial conclusions in place) to receive constructive feedback from peers.

This process mirrors the collaborative environment of industry data science teams and helps refine your final submission.

* **Initial Post** will include a Project Brief, Version 1 (Rough Draft) Project, and Written Reflection
* **Interaction Post** will review a peer's project

*Part 3-Project 2*

You'll demonstrate your ability to identify and solve business problems using machine learning, implement best practices for reproducible workflows, evaluate models in context of business objectives, and effectively communicate technical results to various stakeholders.

The project emphasizes both technical proficiency and the softer skills of problem-solving, critical thinking, and communication that distinguish exceptional data scientists.

You will be following a structured CRISP-DM approach to data science modeling, iteratively developing a machine learning model in order to solve specific problems.

Project GitHub repository containing:

* + Jupyter notebook/s with all code and analysis
  + README.md with project overview and instructions for running the code
  + Any relevant data or image files

*Part 4-Discussion Showcase*

This showcase provides an opportunity to present your project in a concise, compelling manner that highlights both your technical prowess and your ability to translate complex concepts into business value.

Additionally, you'll reflect on your journey through this project and provide constructive feedback to peers, mirroring the collaborative environment of professional data science teams.

**Instructions**

1. *Dataset Selection & Problem Definition*
   1. Choose a dataset suitable for classification or regression. Can be image classification as well.
   2. Define a specific business problem and identify key stakeholders.
   3. Formulate clear business objectives and success metrics.
   4. Determine how you will evaluate your models’ performance.
2. *Exploratory Data Analysis*
   1. Conduct comprehensive data profiling.
   2. Create visualizations to understand data distributions and relationships.
   3. Identify patterns, outliers, and potential features of interest.
   4. Document initial insights and hypotheses.
3. *Data Preprocessing & Feature Engineering*
   1. Implement data cleaning procedures.
   2. Handle missing values, outliers, and categorical features.
   3. Conduct new feature engineering if appropriate.\
   4. Build a reproducible preprocessing Pipeline using sci-kit learn.
4. *Model Deployment*
   1. Select appropriate algorithms based on the problem type.
   2. Implement at least three different modeling approaches.
   3. Perform extensive hyperparameter tuning
   4. Use cross-validation to assess model performance.
   5. Document the iterative improvement process.
5. *Model Evaluation & Selection*
   1. Evaluate models using appropriate metrics tied to business objectives.
   2. Select a final model based on performance and interpretability.
   3. Test the chosen model on held-out test data.
   4. Analyze feature importance and model behavior.
6. *Business Impact Analysis*
   1. Translate model metrics into business value.
   2. Create actionable recommendations based on model insights.
   3. Discuss implementation considerations and limitations.
7. *Documentation & Presentation*
   1. Finalize code documentation and notebook in Github
   2. Create a Minimum Viable Project (MVP) to share with your peers during the MVP Discussion to garner feedback before your final project is due.
   3. Implement appropriate feedback and submit your final code documentation and notebook in Github
   4. Develop a executive presentation targeting non-technical stakeholders and record presentation video. Outline the lessons learning in this process for the showcase discussion. Provide feedback to your peers.

**Possible Data Resources**

* [Kaggle DatasetsLinks to an external site.](https://www.kaggle.com/datasets)
* [UCI Machine Learning RepositoryLinks to an external site.](https://archive.ics.uci.edu/ml/index.php)
* [Google Dataset SearchLinks to an external site.](https://datasetsearch.research.google.com/)
* [AWS Open Data RegistryLinks to an external site.](https://registry.opendata.aws/)
* [Data.govLinks to an external site.](https://www.data.gov/)
* [World Bank Open DataLinks to an external site.](https://data.worldbank.org/)
* [Earth Data NASALinks to an external site.](https://earthdata.nasa.gov/)
* [FiveThirtyEight DataLinks to an external site.](https://data.fivethirtyeight.com/)
* [OpenMLLinks to an external site.](https://www.openml.org/)
* [TensorFlow DatasetsLinks to an external site.](https://www.tensorflow.org/datasets)