INFO 7390 Projects
Project selection guidelines and evaluation criteria

Project theme/topic selection deadline: **Jan 31, 2024**
Project change deadline: **Feb 4, 2024**
Project submission deadline: **April 19, 2024**
Maximum marks: **40**

1. <u>Team size</u> can range from two to a max of four, with each team making the contributions of each team member transparent. **Teaming up is highly recommended and reflects a real-world, corporate team. Exceptions :** Reasons and background skills for doing a one-person project need to be discussed.

2. <u>Suggested topics</u>. These are suggested routes
   (a) Use a Kaggle dataset. You could either come up with your own prediction goals or use Kaggle's problem definition. Advantages of picking a Kaggle dataset include well curated, large datasets, problem is pre-defined and most importantly, you can see where your model performance stands against the leaderboard!
   (b) You can gather and work on datasets from other sources like Google bigquery.
   (c) Webscraping, twitterbots, yelp reviews, and a host of other domain specific datasets exist out there. Youtube videos, for example, has been the source of many a project.

3. <u>Project submission details</u>. Please submit:
   (a) code inside of a Jupyter notebook, that goes all the way from reading in the data to printing out model performance visualizations. Please make sure to submit fully commented code,
   (b) a short write-up explaining the goals, and importance of the project,
   (c) Slide deck with not more than 6-7 slides summarizing the project, packages used, goal, etc.

Point breakdown

The project is worth 40 points and involves developing a substantial data science solution that addresses complex real-world challenges. This write-up provides detailed rubrics to guide your progress and evaluation.

1. **Problem Definition and Understanding (5 points):**
   - Clearly define the data science problem your project aims to solve.
   - Demonstrate a deep understanding of the problem domain and the significance of your solution.
2. **Data Collection and Preprocessing (5 points):**
   - Collect and preprocess relevant data for analysis.
   - Handle missing values, outliers, and ensure data quality.
   - Clearly document your data cleaning and preprocessing steps.
3. **Exploratory Data Analysis (EDA) and Visualization (5 points):**
   - Conduct a thorough EDA to gain insights into the data.
   - Utilize appropriate visualization techniques to communicate key findings.
   - Demonstrate creativity in visualizing complex relationships within the dataset.
4. **Feature Engineering and Model Development (5 points):**
   - Implement advanced feature engineering techniques to enhance model performance.
   - Develop a robust data science model using machine learning or statistical methods.
   - Experiment with multiple models and clearly justify your choices.
5. **Model Evaluation and Interpretability (5 points):**
   - Evaluate your model using appropriate metrics, considering the problem's context.
   - Interpret model results and provide actionable insights.
   - Address any limitations or challenges encountered during the modeling process.
6. **Code Quality and Documentation (5 points):**
   - Write well-structured, documented, and modular Python code.
   - Include comments and docstrings to explain key components of your code.
   - Provide a README file with clear instructions on running your data science project.
7. **Presentation and Communication (5 points):**
   - Develop a clear and engaging presentation to effectively communicate your findings.
   - Clearly articulate the problem, your approach, and the impact of your data science solution.

Brownie Points (5 points):

In order to earn brownie points for creativity in your data science project, consider implementing the following suggestions:

1. **Novel Feature Engineering:**
    - o   Explore unconventional features or create new variables to gain unique insights into the problem.
2. **Innovative Visualization Techniques:**
    - o   Utilize novel visualization methods or combine visual elements creatively to convey complex relationships within the data.
3. **Algorithmic Innovation:**
    - o   Explore advanced algorithms, customize existing ones, or combine different models in creative ways to address specific nuances.
4. **Interactive Dashboards:**
    - o   Build interactive dashboards using tools like Plotly or Bokeh to allow dynamic exploration of the data.
5. **Integration of External Data Sources:**
    - o   Integrate external data sources to enhance analysis and provide a broader perspective on the problem.
6. **Creative Problem Framing:**
    - o   Think creatively about problem framing, considering alternative perspectives or redefining the problem for unique solutions.
7. **Ethical Considerations:**
    - o   Address potential biases in the data creatively and ensure fairness and transparency in models.
8. **Open-Ended Exploration:**
    - o   Pursue aspects of the project that genuinely interest you, allowing for personal creativity to shine through.

Note: Evaluation will be based on the overall quality of your work, adherence to the rubrics, and the extent to which you meet or exceed expectations. The brownie points are awarded for outstanding efforts that demonstrate creativity and innovation in the field of data science.