Assignment: Feature Engineering & Data Preparation Plan

**Week:**

**Week 4: Feature Engineering & Data Preparation**

**Due Date: Week 4**

**Objective:**

In this assignment, you will advance your capstone project by performing feature engineering and data preparation, two critical steps for enhancing model performance. You will create new features, apply transformation techniques, and prepare your dataset for machine learning by splitting it into training, validation, and test sets. The deliverable for this assignment will be a comprehensive Feature Engineering and Data Preparation Plan.

**Assignment Instructions:**

**1. Feature Engineering:**

* **Objective:** Improve your dataset by creating new features that may improve the predictive power of your model.
* **Tasks:**
  + **Identify Potential Features:** Examine your dataset and identify any features that can be created based on existing data. For example:
    - Creating ratios or differences between numerical features.
    - Extracting date-related features (e.g., year, month, day) from datetime variables.
  + **Transformation Techniques:** Apply transformations to numerical features to improve their distribution:
    - **Scaling:** Normalize or standardize numerical features to bring them to a similar scale (e.g., using MinMaxScaler or StandardScaler in Python).
    - **Log Transformation:** Apply logarithmic transformations to skewed data to reduce the impact of outliers.
    - **Polynomial Features:** If applicable, consider adding polynomial terms to increase the model's ability to capture non-linear relationships.
  + **Encoding Categorical Variables:** Use appropriate encoding methods to convert categorical features into numerical values. Techniques include:
    - **One-hot encoding** for nominal variables (e.g., creating binary columns for each category).
    - **Ordinal encoding** for ordered categories.

**Deliverable:**  
Provide a list of new features you created or transformed, including explanations for each feature and why it was engineered. Include any code used for feature transformations and encoding.

**2. Data Preparation for Modeling:**

* **Objective:** Prepare your dataset for modeling by splitting it into training, validation, and test sets, ensuring your data is ready for machine learning.
* **Tasks:**
  + **Data Splitting:** Split your dataset into training, validation, and test sets. Explain how you split the data and why (e.g., 70% training, 15% validation, 15% test).
    - **Training Set:** Used for fitting the model.
    - **Validation Set:** Used for tuning hyperparameters and preventing overfitting.
    - **Test Set:** Used for evaluating the final model performance.
  + **Handling Imbalanced Data:** If applicable, discuss how you handled class imbalances (e.g., using SMOTE or class weighting).
  + **Shuffle and Random Seed:** Ensure data is shuffled before splitting, and set a random seed for reproducibility.

**Deliverable:**  
Provide a summary of your data preparation process, including:

* The code used for splitting the data.
* The rationale behind the split percentages.
* Any steps taken to handle imbalanced classes, if applicable.

**Submission Requirements:**

* **Report Format:**  
  Submit your Feature Engineering and Data Preparation Plan as a Jupyter Notebook (or PDF if using another platform). Ensure all code is well-commented, and each section is clearly labeled.
* **Structure of Submission:**
  + Feature Engineering: List of created/transformed features and their rationale.
  + Data Preparation: Description of how the data was split, why the chosen split percentages were used, and any techniques for handling imbalances.
  + Include all code used for feature engineering, encoding, transformation, and splitting the data.
* **Length:**  
  The report should be concise but thorough, with an expected length of no more than 5–7 pages, including code and visualizations.