**Question 1**

Dataset: spenddata.csv and testdata.csv

You are given a set of survey data which captures spend amount among other data points. Some of the respondents have been tagged as belonging to group 1 – 6. However, due to a data calculation issue, some of the respondents have had their groups (**pov6**) missing.

Build a model that will classify these respondents back into one of the 6 groups.

1. Please explain the choice of metric / evaluation criterion used
2. What are the assumptions you made when building this model?
3. What were the approaches you considered? Please explain the reason for the technique / approach used as well as the pros and cons.
4. Please explain under what conditions will the model you choose **be not appropriate**
5. How confident are you of the model’s robustness and how would you explain the model’s performance?
6. Why is your model performing well / not well?

Was any feature engineering required? If yes, what were they. If no, why?

1a. Choice of Metrics used:

Weighted F1-score

F1 score weighted by the number of true instance of each class.

* Accuracy, precision, recall affected by imbalance multiclass dataset
* Macro F1 ignore class size
* Micro F1 gives more weight to larger classes
* Weighted F1 provided the best evaluation metrics by weighted the class size

1b. Assumptions

* Multiclass equal importance assumption
  + All classes are of primary interest
* Feature Independence Assumption.
  + Many ML models assume feature independence
* Null values assumption
  + Replacing null values with median – Quick assumption
  + Sweeping threshold (60%) Columns with more than 60% null are not useful

1c. Approach Considered

**Data Cleaning - Handling Null Values:**

Pros:

* Improves data quality and integrity.
* Reduces the potential for bias and errors in your analysis.

Cons:

* Data loss if you remove too many records with missing values.

**Outlier Detection using Z-Scores:**

Pros:

* Identifies potential data anomalies that can affect model performance.

Cons:

* Might incorrectly identify valid data points as outliers or miss subtle outliers.
* May require domain knowledge to set appropriate thresholds.

**Feature Engineering - Get Dummies for String Columns:**

Pros:

* Converts categorical data into a format suitable for machine learning algorithms.
* Allows the inclusion of categorical variables in models that require numerical inputs.

Cons:

* Increases the dimensionality of the dataset, potentially leading to the curse of dimensionality.

**2. Data Balancing:**

**ADASYN (Adaptive Synthetic Sampling):**

Pros:

* + - * Addresses class imbalance by generating synthetic samples for the minority class.
      * Helps prevent the model from being biased towards the majority class.

Cons:

* + - * Synthetic data may not accurately represent the true underlying data distribution.

**SMOTE (Synthetic Minority Over-sampling Technique):**

Pros:

* + - * Mitigates class imbalance by generating synthetic samples.
      * Improves the model's ability to learn from the minority class.

Cons:

* + - * Similar to ADASYN, synthetic data may not fully represent the true distribution.

**3. Feature Selection:**

Voting Through 5 Different Feature Selection Methods:

Pros:

* + - * Helps in identifying the most relevant features, reducing model complexity, and potentially improving model performance.
      * Robust to individual feature selection methods' limitations.

Cons:

* + - * May require more computational resources and time.
      * The choice of feature selection methods and the voting mechanism can introduce subjectivity.

1d. Not appropriate conditions

* Change in class percentage structure despite ADASYN
* Lack of domain knowledge:
  + There is a lack of context on the feature used and having the metadata on the columns before using the model is critical
* Features engineering is not done in the dataset as there is lack of understanding on the columns
* Change in Population Stability index over time. Meaning that the behaviour of the population change and new data is require to evaluate and predict ‘pov6’

1e. Model robustness and performance

* Model lacks robustness.
  + Data Cleaning was done without understand of the metadata nor domain based knowledge.
  + Outlier is not eliminated, employ a binning structure to the selected would be much better approach
* Performance rated based on f1\_weighted score

1f. Overall performance

A screenshot of a computer

Description automatically generated

* Test scores and validation scores for both f1\_weighted are approximately 0.93
* Overall average performance given the overall score
* Macro (ignoring class weights) only perform 0.74 which would not be sufficient to be use for strong prediction

**Future Improvements**

* Employ a binning structure would change the way the model predict the classes and reduce the impact of outliers
* Understanding meta data would give context and understanding on replacing null values or creating new columns

A chart with numbers and a yellow square

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated

.

Was Feature Engineering Required

Due to lack of information on metadata, feature engineering (get dummies) was only done on selected columns

‘Month’, ‘var9’