PROJECT 602

PRINCIPLES OF DATA SCIENCE

Project part II

SUBMITTED TO: DR. Charles Forgy

SUBMITTED BY UID
Aneesh Krishna Rao Chepuri 121324382
Omkarnath Thakur 121335685
Ruthvick K Kandrala 121305206
Siddharth Pathania 121291592

1. The dataset contains features regarding customer profile at each statement date. Features are anonymized and the fall in these categories :-

Delinquency (D_*), Spend (S_*), Payment (P_*), Balance (B_*), and Risk (R_*).
We will be using these features for the final project. Along the process we will drop features which are less relevant, thus keeping only important features.

The target is a binary variable, and it tells us whether a customer has paid or has not paid his due amount in 120 days after their latest statement, therefore we will predict this target variable using the features.

2. Your initial plan for any feature engineering (e.g. adding a Boolean for weekend vs. weekday or taking the ratio of two other features).

For the American Express Default Prediction, the initial approach will involve:

- **Categorical columns**:
  - Apply one-hot encoding for nominal categories and for cardinal data.
  - Use frequency or target encoding with smoothing to capture relationships with the target variables.
- **Numerical columns**:
  - Handle missing data using KNN imputation or Missing Value imputation using ML models.
  - Scale features and create meaningful interactions (ratios, differences).
  - Apply PCA or LDA to reduce dimensionality.
- **Time-based columns (date columns)**:
  - Create lagged features and rolling windows to capture customer behavior over time.

- **Aggregation**:
    - Aggregate transaction-level data to customer-level using statistics like mean, sum, and standard deviation.
- **Feature selection**:
    - Apply techniques to retain the most important features for model refinement.

3. If you plan to use any sort of method for dealing with imbalanced data, detail your overall approach (e.g. "will oversample class XYZ", or "will create synthetic data using SMOTE"). If you do not intend to use any techniques to deal with imbalanced data, explain why it is not necessary (e.g. "all features are at worst imbalanced 7:3").

Based on the initial data analysis we saw that % data for people who default is less as compared to people who have paid. To handle this situation we are planning to use different methods, and will see which one works best for us. Also we know that credit defaults are typically rare events, often comprising less %age of the dataset.

- **Oversampling the minority class**: We'll use techniques like SMOTE (Synthetic Minority Over-sampling Technique) to create synthetic examples of the minority class. This will help balance the dataset without simply duplicating existing data points.

- **Undersampling the majority class:** We'll carefully undersample the majority class (non-defaulters) using methods like Random Under Sampling or Tomek links to remove some of the majority instances while maintaining the dataset's quality.

- **Ensemble methods:** We'll use ensemble techniques like BalancedRandomForestClassifier or EasyEnsembleClassifier, which are designed to handle imbalanced datasets effectively.

- **Class-weight adjustment**: Assign higher weights to the minority class in algorithms like Random Forest, XGBoost, or Logistic Regression making the model more sensitive to defaults.

Here are some initial techniques that we are planning to use. Along the process if we find more useful techniques we will use them as well, and then we will carefully monitor performance metrics to see which methods are more effective to give us more accurate results.