**Merchant Pricing Model - Methodology and Findings**

**1. Objective**

This project aims to assess whether a merchant's card processing fees are **competitive**, **neutral**, or **non-competitive** based on market data and merchant characteristics.

**2. Running the Model**

Run the .py file in the repository will train a random forest classifier on the given data. It will output an accuracy score for the model, as well as the prediction (classification) result of a sample line.

**3. Methodology**

**2.1 Data Preprocessing**

1. **Parsing Fees, calculating weighted average fee and give price competitiveness labels for each business (each line)**

* Fee strings (e.g., '0.1% + 2p') are separated into percentage and fixed fee components – 0.001, 0.02
* The percentage fees are then combined into a weighted average fee based on assumptions about card payments:
* 40% of card payments are from Mastercard cards, and 60% are from Visa cards
* 90% of card payments are made using Debit cards, 8% are made using credit cards, and 2% are made using business debit card.
* An additional assumption I made is that the fixed fee component is not taken into account. I made this assumption as the amount does not vary much between businesses and is relatively small comparing to the percentage component. Therefore, it does not affect the competitiveness assessment significantly.
* Price Competitiveness Labels - merchants are classified into 3 categories based on weighted fees:
  + Competitive: Fees below the 33rd percentile.
  + Neutral: Fees between the 33rd and 66th percentiles.
  + Non-Competitive: Fees above the 66th percentile.

1. **Binary Conversion:**

Columns like 'Is Registered' and 'Accepts Card' are mapped to 0 and 1 values to handle binary data effectively.

1. **Label Encoding Categorical Features:**

MCC Code (Merchant Category Code) and Current Provider are converted into numeric labels using LabelEncoder for compatibility with the model.

1. **Scaling Numerical Features:**

Numerical features are standardised using StandardScaler to ensure equal weighting during model training. Standardization is applied to both training and prediction data for consistency.

**2.2 Model Training and Evaluation**

1. **Classifier:** Random Forest Classifier with 100 estimators and a random state of 42 ensures consistency and robustness.
2. **Train-Test Split:** 75% training and 25% test data ensure a fair evaluation of the model’s performance.
3. **Performance Metric:** Evaluated using accuracy score to assess how well the model predicts pricing competitiveness.

**3. Assumptions**

1. **Fee Distribution:**
   * 40% Mastercard, 60% Visa split reflects card usage trends.
   * 90% Debit, 8% Credit, and 2% Business Debit usage breakdown aligns with industry averages.
   * The effect of fixed fees on competitiveness of a business is not significant
2. **Price Thresholds:**
   * Thresholds for competitiveness are based on the 33rd and 66th percentiles of weighted fees.

**4. Interesting Findings**

* + Businesses with higher turnover and larger transactions are generally more competitive. Low turnover businesses are non-competitive, possibly due to fewer economies of scale.
  + Turnover and transaction size significantly separate clusters.
  + The 33% percentile split results in a balanced distribution, ensuring no cluster is disproportionately large or small.