Report: Credit Card Fraud Detection by Ruyue Xiao

1. Data Exploration

The dataset contains credit card transaction records, with is_fraud as the target variable. The key features include numerical and categorical variables such as:

- **Latitude/Longitude**: Transaction and merchant location.
- **Time Features**: Transaction date, time, and derived cyclic features like hour and weekday.
- **Amount** (amt): Transaction value.
- **Categorical Features**: Job, state, merchant, and gender.

No missing values were found in critical features after encoding and handling invalid values.

2. Feature Engineering and Selection

The following steps were applied to engineer useful features:

- **Distance Feature**: Calculated using the Haversine formula to determine geographic distance between transaction location (lat, long) and merchant location (merch_lat, merch_long).
- **Cyclic Time Features**: Created hour_sine and hour_cosine to represent the cyclical nature of hours in a day.
- **Label Encoding**: Applied to categorical columns like category, job, and state to convert them into numerical values.
- **Correlation Heatmap**: A correlation heatmap showed relationships between features and target, helping ensure no highly redundant features were included.

Features like distance_km and cyclic encodings showed strong relevance during feature importance analysis.

3. Model Selection

The following model was chosen:

- **LightGBM**: A gradient-boosting decision tree algorithm designed for high performance on large datasets.

Why LightGBM?

- Handles imbalanced data effectively (class weights).
- Fast training speed and efficient memory usage.
- Provides feature importance for interpretability.

4. Model Tuning

- **Hyperparameter Tuning**: Optimized using cross-validation with Stratified K-Folds to ensure class balance.
- Parameters: n_estimators=500, learning_rate=0.03, num_leaves=31.
- **Threshold Optimization**: After training, the decision threshold was tuned to maximize the F1-score using a grid search over thresholds ranging from 0.01 to 0.99.

5. Validation of Model

- **Local Validation**: A 5-Fold Stratified Cross-Validation (CV) approach was used. Out-of-fold (OOF) predictions ensured unbiased validation.
- **Evaluation Metric**: F1-score was chosen as the primary metric due to class imbalance, balancing precision and recall.
- **Results**:
    - **Best OOF F1-score** (local): **0.97236**

6. Comparison to Kaggle Results

The model was evaluated on Kaggle's leaderboard:

- **Public Score (30% Data Tested)**: **0.97480**
- **Private Score (Full Data Tested)**: **0.97536**

7. Pros and Cons of the Model

Pros:

- LightGBM is computationally efficient and scalable.
- Works well with imbalanced data using class weights.
- Features like distance_km and cyclic encodings improve predictive power.

Cons:

- Requires careful tuning of hyperparameters like learning_rate and num_leaves.
- Limited interpretability compared to simpler models (e.g., Logistic Regression).
- Sensitive to noisy or irrelevant features, requiring thorough preprocessing.

8. Decisions and Assumptions

- Assumed missing values were non-existent or irrelevant after encoding.
- Assumed that geographic distance is a strong predictor of fraud.
- Used a balanced class weight to address the class imbalance issue.
- Chose F1-score as the evaluation metric to ensure balanced performance on precision and recall.