**Fraud Detection Assessment Summary**

**Approach Overview**

This assessment implements a comprehensive fraud detection system using the provided financial transaction dataset containing 6.3M records with features including transaction type, amounts, account balances, and fraud labels.

**Key Components**

**1. Data Analysis & Feature Engineering**

* **Exploratory Analysis**: Examined fraud patterns across transaction types, amounts, and time
* **Feature Creation**: Developed 8 new features including:
  + Balance difference calculations
  + Error detection in balance inconsistencies
  + Time-based features (hour of day)
  + Amount categorization and log transformation
  + Zero balance indicators

**2. Model Development Strategy**

Trained four different algorithm families to compare performance:

* **Logistic Regression**: Baseline linear model with feature scaling
* **Random Forest**: Ensemble method handling non-linear relationships
* **Gradient Boosting**: Advanced ensemble with sequential learning
* **Neural Network**: Deep learning approach with 3 hidden layers and dropout

**3. Data Preprocessing Decisions**

* **Class Imbalance**: Addressed severe imbalance (fraud ~0.13%) by balanced sampling
* **Feature Scaling**: Applied StandardScaler for linear models and neural networks
* **Train/Test Split**: 80/20 split with stratification to maintain fraud proportion

**Key Assumptions**

1. **Transaction Legitimacy**: Non-fraudulent transactions follow predictable patterns
2. **Feature Relevance**: Balance inconsistencies and transaction types are strong fraud indicators
3. **Temporal Stability**: Fraud patterns remain relatively consistent across time periods
4. **Data Quality**: Dataset accurately represents real-world transaction patterns

**Model Selection Rationale**

The assessment evaluates models using **AUC-ROC score** as the primary metric because:

* Handles class imbalance effectively
* Measures discrimination ability across all thresholds
* Industry standard for fraud detection systems
* Balances precision and recall considerations

**Expected Performance Hierarchy**:

1. **Gradient Boosting**: Likely best due to sequential error correction
2. **Random Forest**: Strong performance with feature interactions
3. **Neural Network**: Good for complex patterns but may overfit
4. **Logistic Regression**: Baseline performance with interpretability

**Trade-offs Made**

**Performance vs. Interpretability**

* **Chosen**: Ensemble methods for best performance
* **Trade-off**: Reduced interpretability compared to linear models

**Dataset Size vs. Training Speed**

* **Chosen**: Balanced sampling (100K majority + all minority samples)
* **Trade-off**: Potential loss of rare fraud patterns in full dataset
* **Justification**: Manageable training time while preserving fraud examples

**Model Complexity vs. Overfitting Risk**

* **Neural Network**: Added dropout layers and validation monitoring
* **Ensemble Methods**: Limited n-estimators to prevent overfitting
* **Cross-validation**: Could be implemented for more robust evaluation

**Expected Results**

Based on the feature engineering and modeling approach:

* **Best Model AUC**: Expected 0.85-0.95 range
* **Fraud Detection Rate**: 80-90% of fraudulent transactions identified
* **False Positive Rate**: <5% for practical business application