# Credit Card Fraud Detection: Business Summary Report

## 1. Introduction

Credit card fraud poses a major financial and operational risk to payment networks, merchants, and banks. Fraudulent activity represents a very small percentage of total transactions, but each false negative has a high cost and each false positive wastes operational time.

This project develops a machine learning solution to detect fraudulent transactions using the *Credit Card Fraud Detection Dataset*, which contains **284,807 transactions**, with only **0.17% fraud cases**. The goal is to improve fraud detection accuracy while minimizing false alerts.

---

## 2. Key Insights From the Data

### 2.1 Transaction Amount Patterns

- Fraudulent transactions are typically **low-value**, with a median amount of **~9 units**, suggesting intentional avoidance of detection thresholds.
- Fraud still includes occasional medium-value spikes (~1,000–2,000 units).
- Legitimate transactions include very large outliers (up to ~25,000), creating a skewed distribution.

### 2.2 Transaction Timing Patterns

- Fraud does not occur randomly; it clusters in a specific time range.
- Fraud tends to occur between **40,000–130,000 seconds** into the dataset period.
- This suggests exploitation of specific time windows, such as low monitoring periods.

### 2.3 PCA Feature Analysis

The dataset's features (V1–V28) are PCA-transformed components for privacy.

Even without direct meaning, they show clear fraud behavior patterns:

- **V14 and V17** exhibit strong separation between fraud and non-fraud.
- **V12** shows moderate separation.
- These components provide strong predictive power for machine learning models.

---

# 3. Handling Class Imbalance

Fraud = **0.17%** → severe imbalance problem.

Three methods were tested:

1. **Class-weighted Logistic Regression**
    - High recall (0.92) but extremely low precision (0.06)
    - Too many false alarms → impractical for production.
2. **SMOTE Oversampling**
    - Improves balance
    - Better F1-score (0.23)
    - Still insufficient precision for business use.
3. **Random Forest with Balanced Class Weights**
    - Best balance between precision and recall
    - Most effective model overall

---

# 4. Final Model Performance

## Random Forest (Selected Model)

- **Precision (Fraud): 0.96**
  → When the model says "fraud," it is correct 96% of the time.
- **Recall (Fraud): 0.76**
  → The model captures 76% of all fraud cases.
- **F1-score (Fraud): 0.85**
  → Strong overall fraud detection performance.

## Why This Model?

- High precision keeps the number of false alerts low.
- Strong recall ensures the system still catches the majority of fraud.
- Robust to noisy and non-linear patterns.
- Handles imbalance effectively when class weights are applied.

---

# 5. Business Impact

Implementing this Random Forest model can offer:

### Reduced Financial Loss

Detects fraudulent transactions early with fewer missed cases.

### Lower Operational Cost

High precision significantly reduces unnecessary investigations.

### Improved Customer Trust

Minimizes the chance of fraud slipping through undetected.

### Scalable Solution

The model can operate in real-time with minimal computational overhead.

---

## 6. Limitations

- PCA-transformed features reduce explainability.
- Dataset covers only two days of transactions.
- Lacks contextual data (merchant category, user history, geo-location).
- XGBoost could not be used due to macOS library limitations (libomp dependency).

---

## 7. Recommendations & Future Work

- Add contextual features (merchant category, region, device).
- Deploy anomaly detection techniques (Isolation Forest, Autoencoders).
- Use real-time model monitoring with adaptive retraining.
- Integrate cost-sensitive learning to handle financial trade-offs more directly.

---

## 8. Conclusion

This project demonstrates a practical and effective machine learning pipeline for credit card fraud detection. Despite extreme class imbalance, the final Random Forest model achieves a strong balance between precision and recall, making it suitable for business use cases that require both efficiency and accuracy.

This system can significantly reduce fraud-related losses, improve operational workflows, and provide actionable insights for risk management teams.