Credit Card Fraud Detection – Project Report

In this project, we addressed the problem of credit card fraud detection, a real-world machine learning task characterized by extreme class imbalance and high operational constraints. Fraudulent transactions represent less than 0.2% of the dataset, which makes standard classification approaches and naïve evaluation metrics unsuitable. Our objective was not only to build predictive models, but also to understand their limitations, compare their behaviors, and justify each methodological choice.

We began with an exploratory data analysis to understand the structure of the dataset. Most features were anonymized and already transformed using PCA, which reduced multicollinearity and simplified preprocessing. However, the Time and Amount variables were not scaled, so we standardized them to ensure consistency across features. We also carefully analyzed the class distribution and confirmed the extreme imbalance, which directly influenced our modeling and evaluation strategy.

As a first step, we implemented baseline models such as Logistic Regression and Decision Trees. These models were chosen deliberately: Logistic Regression provides a strong linear baseline with interpretable behavior, while Decision Trees allow us to capture non-linear relationships. We applied stratified train–test splits to preserve the fraud ratio and evaluated the models using metrics adapted to imbalanced data, such as recall, F1-score, ROC-AUC, and PR-AUC. These initial results clearly highlighted the core challenge of the problem: models either achieved high recall at the cost of many false positives, or high precision while missing a significant number of fraud cases.

To improve performance, we introduced hyperparameter tuning using GridSearchCV and explored ensemble methods. Bagging helped stabilize Decision Trees, while soft voting allowed us to combine complementary models. Although these approaches improved robustness, the results remained insufficient for a realistic fraud detection system. This led us to conclude that simpler models, even when tuned, are fundamentally limited on this type of dataset.

In the second part of the project, we focused on more robust and widely used models for fraud detection: Random Forest, XGBoost, Gradient Boosting, and LightGBM. These models are designed to handle complex non-linear patterns and benefit from ensemble learning principles. We paid particular attention to class imbalance handling, using techniques such as class weighting and scale_pos_weight, and selected recall-based scoring during cross-validation to reflect the cost of missing frauds.

Among these models, Random Forest, XGBoost, and LightGBM demonstrated strong discriminative power, with high ROC-AUC values and significantly improved PR-AUC compared to baseline methods. However, each model exhibited a different trade-off between precision and recall, reinforcing the idea that fraud detection is not about maximizing a single metric, but about selecting an appropriate operating point based on business constraints.

Finally, to better assess the intrinsic learning capacity of our best-performing model, we evaluated it on a balanced version of the dataset created through undersampling. This experiment showed a clear improvement in class-wise performance, with balanced precision and recall and high F1-scores. While these results confirm that the model can effectively learn fraud patterns, we explicitly acknowledge that this evaluation setting does not reflect real-world conditions. The original imbalanced dataset remains the most relevant benchmark for operational deployment.

In conclusion, this project demonstrates that credit card fraud detection requires more than applying standard machine learning models. It demands careful preprocessing, appropriate evaluation metrics, robust algorithms, and a clear understanding of the trade-offs induced by class imbalance. Through systematic experimentation and critical analysis, we showed that meaningful results can be achieved even in highly imbalanced settings, provided that methodological choices are well justified and aligned with the problem's constraints.