Credit Card Fraud Detection

Don't Let Fraud Crash Your Credit Party: Stay Sharp, Stay Secure!

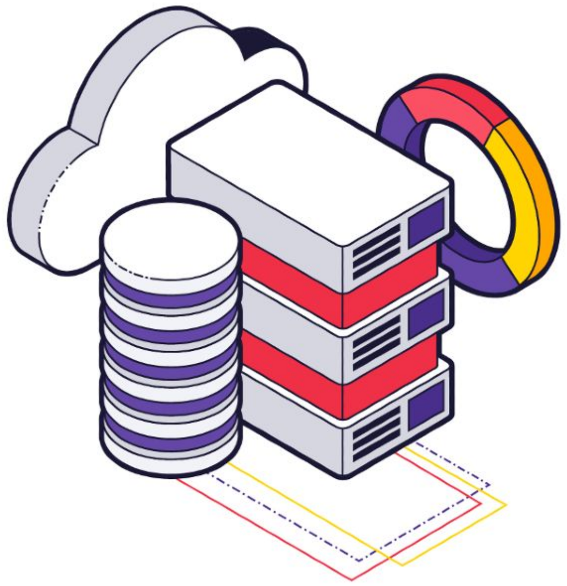
Credit Card Fraud Detection Project

Credit Card Fraud Detection Project

Problem Statement: The goal is to build a fraud detection system that quickly and reliably spots illegal transactions, while minimizing the inconvenience to customers by avoiding false positives.

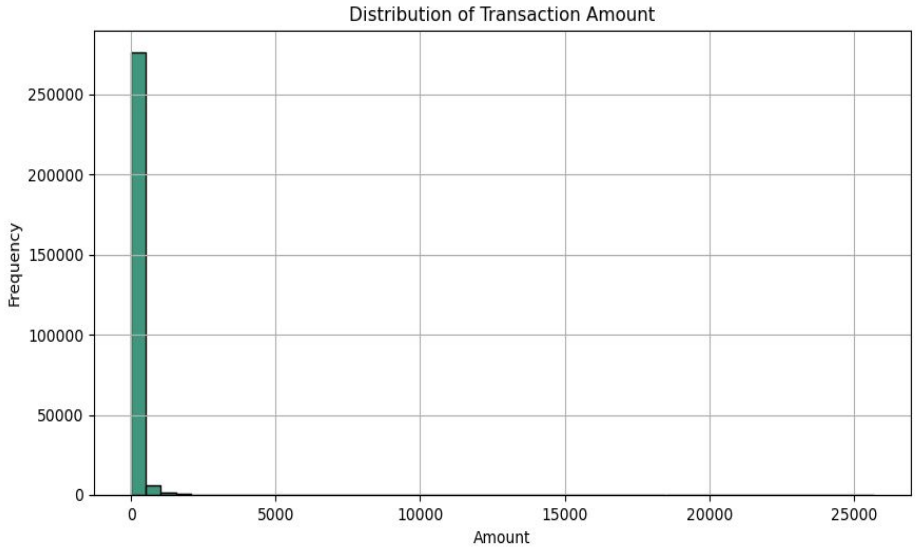
Opportunity: We want to build a model that can accurately anticipate fraud by using a comprehensive dataset from Kaggle. Our goal is to increase transaction security, reduce losses, and rebuild confidence in electronic payments so that credit card fraud will become less common in the future.

Introduction to Dataset



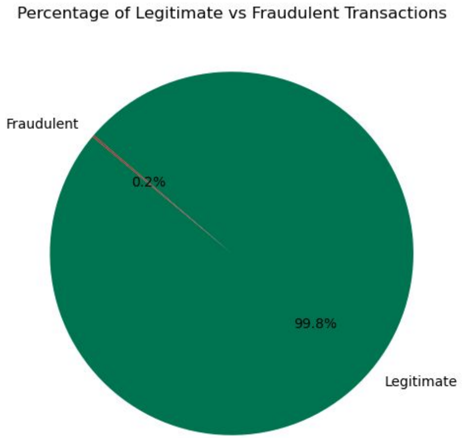
- Dataset sourced from Kaggle, containing 284,807 transactions (492 fraudulent).
- The dataset includes only numerical input variables derived from a PCA transformation.
- Due to confidentiality concerns, the original features and additional background information are not provided.
- Features V1, V2, ..., V28 are the principal components obtained through PCA.
- 'Class' is the response variable:
  - 1 indicates fraud.
  - 0 indicates a non-fraudulent transaction.

Distribution of Transaction Amount



- There's a peak around $0, indicating many small transactions.
- The average transaction amount is influenced by a small number of large transactions.

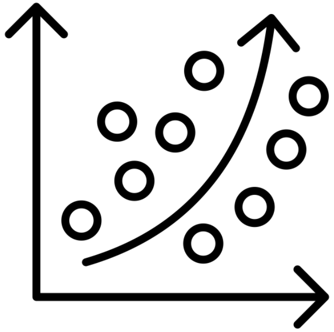
EDA for Imbalanced Datasets



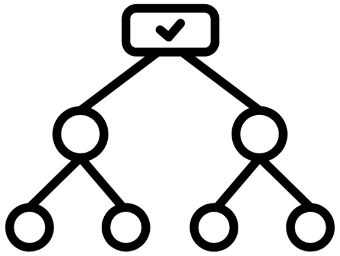
- The dataset contains:
    - 284,807 transactions in total.
- 99.83% of the transactions in the dataset are normal, while only 0.172% are fraudulent.
- This significant imbalance presents a challenge in accurately detecting fraudulent transactions.

Models

Logistic Regression



Decision Tree



Under Sampling: Techniques

Random UnderSampling Technique:

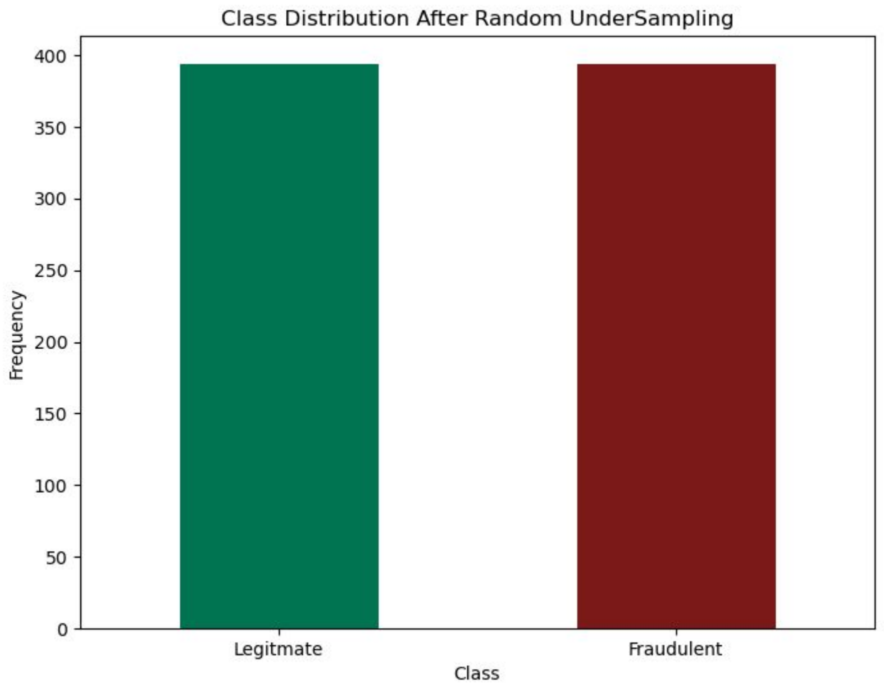
- Balances the dataset by reducing samples in the majority class (legitimate transactions).
- Helps machine learning model learn and predict both classes accurately.
- Potential drawback: Loss of information due to removal of many legitimate transactions.

Before Random UnderSampling:

- Legitimate transactions : 227,451
- Fraudulent transactions : 394

After Random UnderSampling:

- Reduced number of legitimate transactions to match fraudulent transactions.
- Result: 394 legitimate and 394 fraudulent transactions.



Under Sampling Models

Logistic Regression

CLASS	PRECISION	RECALL	F1
0	1.00	0.96	0.98
1	0.04	0.92	0.07
ACCURACY			0.96

Decision Tree

CLASS	PRECISION	RECALL	F1
0	1.00	0.90	0.95
1	0.02	0.91	0.03
ACCURACY			0.90

Over Sampling: Techniques

Over Sampling Technique:

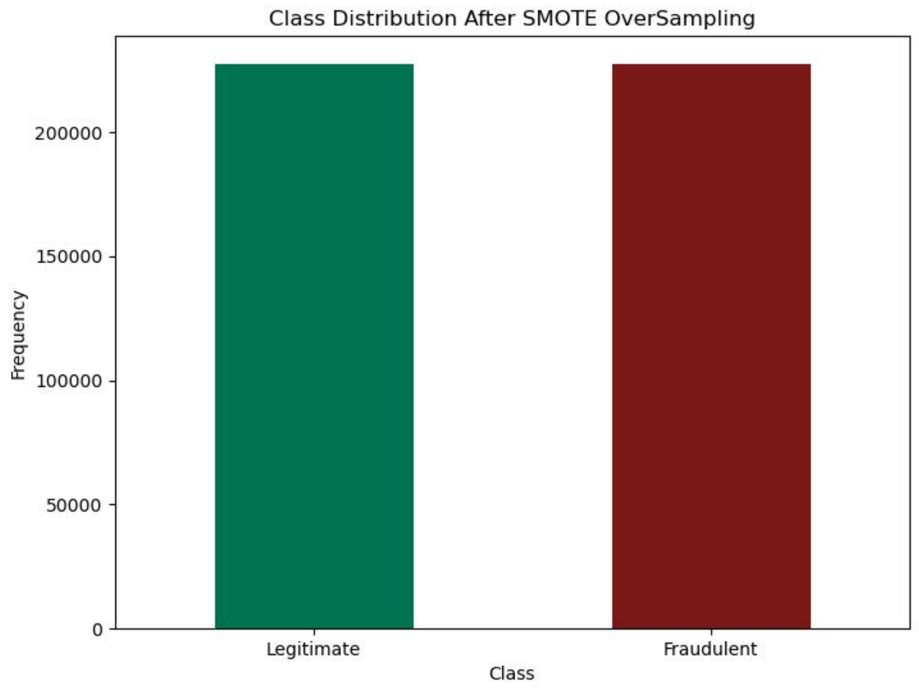
- **SMOTE :** Stands for Synthetic Minority Oversampling Technique.
- Unlike Random UnderSampling, SMOTE creates new synthetic points to balance the classes.

Before Over Sampling:

- Legitimate transactions : 227,451
- Fraudulent transactions : 394

After Over Sampling:

- Created synthetic fraudulent transactions to match the number of legitimate ones.
- After using SMOTE to balance the data, we now have an equal number of legitimate and fraudulent transactions (227,451 each).



Over Sampling Models

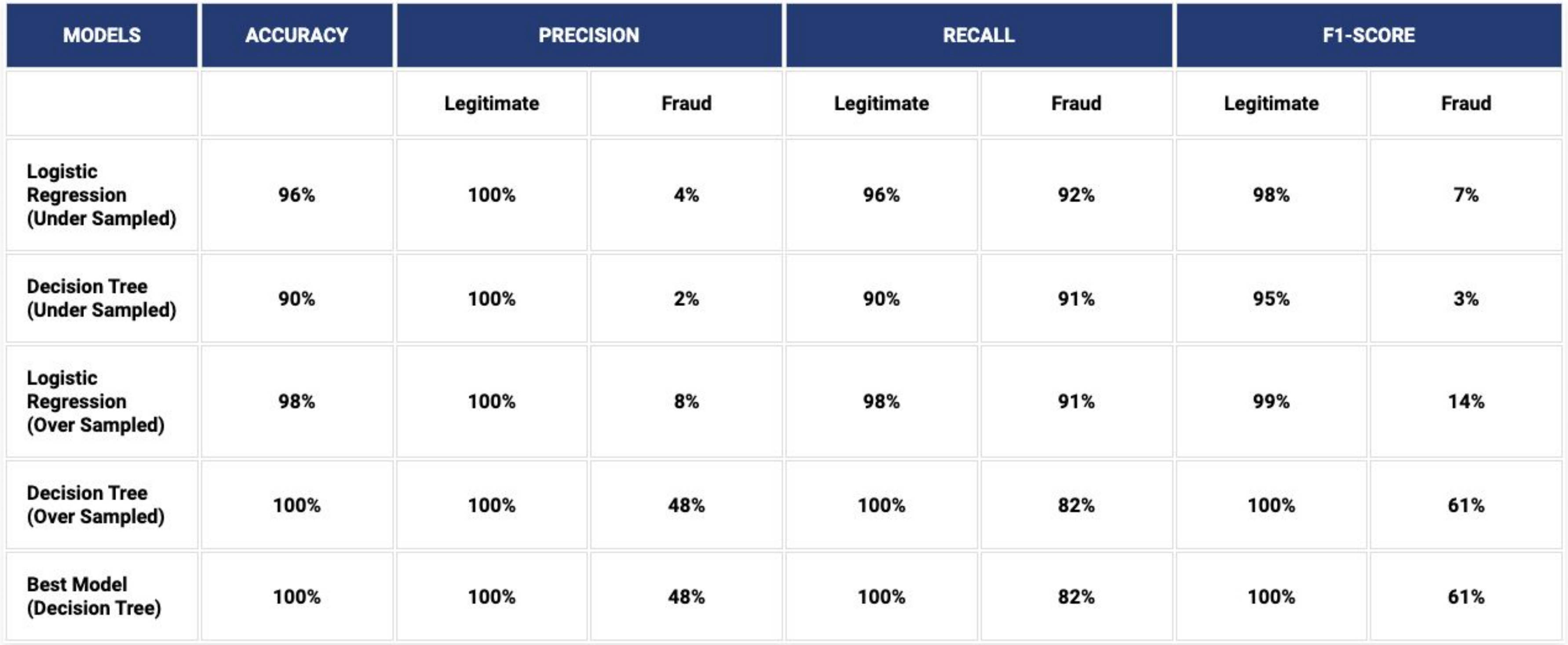
Logistic Regression

CLASS	PRECISION	RECALL	F1
0	1.00	0.98	0.99
1	0.08	0.91	0.14
ACCURACY			0.98

Decision Tree

CLASS	PRECISION	RECALL	F1
0	1.00	1.00	1.00
1	0.42	0.80	0.55
ACCURACY			1.00

Comparison of All Models



MODELS	ACCURACY	PRECISION		RECALL		F1-SCORE	
		Legitimate	Fraud	Legitimate	Fraud	Legitimate	Fraud
Logistic Regression (Under Sampled)	96%	100%	4%	96%	92%	98%	7%
Decision Tree (Under Sampled)	90%	100%	2%	90%	91%	95%	3%
Logistic Regression (Over Sampled)	98%	100%	8%	98%	91%	99%	14%
Decision Tree (Over Sampled)	100%	100%	48%	100%	82%	100%	61%
Best Model (Decision Tree)	100%	100%	48%	100%	82%	100%	61%

Thank You

By : Jashkirat Virdi