Credit Card Fraud Detection

Dataset

	Distance from Home	Distance from Last Transaction	Ratio to Median Purchase Price	Repeat Retailer	Used Chip	Used Pin Number	Online Order	fraud
0	57.877857	0.311140	1.945940	1.0	1.0	0.0	0.0	0.0
1	10.829943	0.175592	1.294219	1.0	0.0	0.0	0.0	0.0
2	5.091079	0.805153	0.427715	1.0	0.0	0.0	1.0	0.0
3	2.247564	5.600044	0.362663	1.0	1.0	0.0	1.0	0.0
4	44.190936	0.566486	2.222767	1.0	1.0	0.0	1.0	0.0

Problem Statement:

The difficulty in accurately detecting fraudulent transactions in financial systems. Fraudulent activities are rare compared to genuine ones, leading to a highly imbalanced dataset. This imbalance causes challenges for traditional detection methods, such as bias toward the majority class and poor identification of fraud cases. Additionally, fraud patterns constantly evolve, making it harder to rely on static or outdated techniques. These factors demand an advanced and adaptive machine-learning-based solution to improve fraud detection accuracy and maintain financial security.

Insights

Are there any outliers in the ratio, and could they indicate potential fraud?

What type of variation occurs in distance_from_home for fraudulent vs. non-fraudulent transactions? For instance, are frauds more likely when the distance is significantly larger?

Can distance_from_home be adjusted or used in combination with other features to improve model predictions?

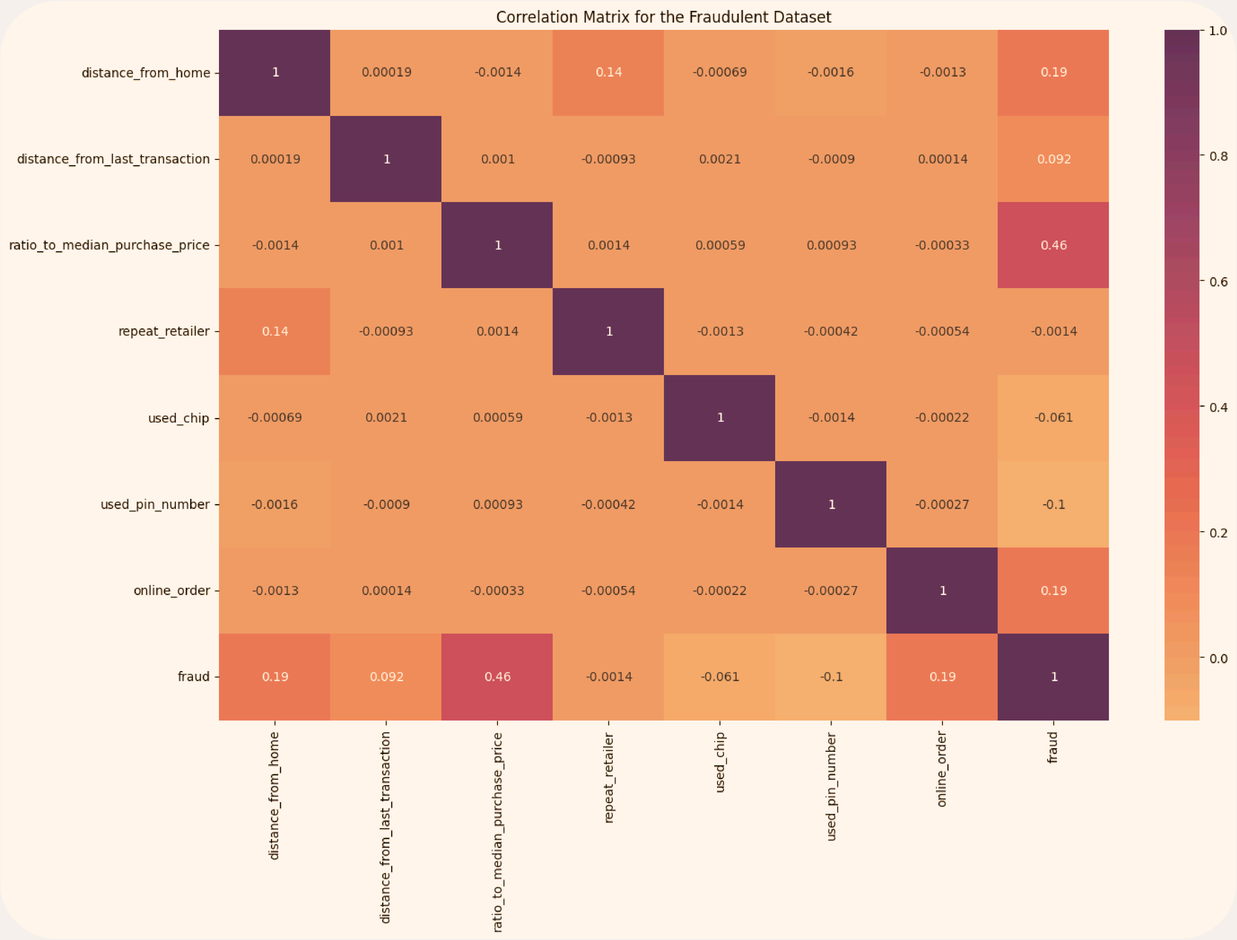
Machine Learning Models and Methods

The dataset was split into TrainData (80%) for training and TestData (20%) for evaluation. For testing, the following models were selected:

- Decision Tree Classifier
- Random Forest Classifier
- XGBoost
- Gaussian Naive Bayes (GaussianNB)

Each model was trained using six different normalization techniques applied to the original dataset, which are min-max scaler, z-score, decimal scaling, max absolute scaling, robust normalization, and unit vector normalization.

Performance was evaluated using Accuracy and RMSE to identify the best model.



Results

The top-performing model was the Random Forest with Min-Max and Robust Normalization, achieving 100% accuracy and RMSE of 0. The Random Forest with Z-Score Normalization delivered 99.9994% accuracy, showing consistency across techniques. The Decision Tree with Robust Normalization also achieved 100% accuracy but lacked consistency.

Conclusion

The Random Forest with Min-Max Normalization was selected as the final model,and with optimized hyperparameters, it achieved 100% accuracy and an RMSE of 0, significantly exceeding the 75% accuracy requirement.