Fraud Detection Case Study

Identifying Credit Card Fraudulent Transactions

Vinay Pavan GUNTUR

CREDIT CARD FRAUD DETECTION

Project Overview

- In this presentation, I have explored the process of building a machine learning model for credit card fraud detection.
- Covered the data preprocessing techniques, model evaluation, and the final model selection.
- Discussed the business implications and provided recommendations for future enhancements.

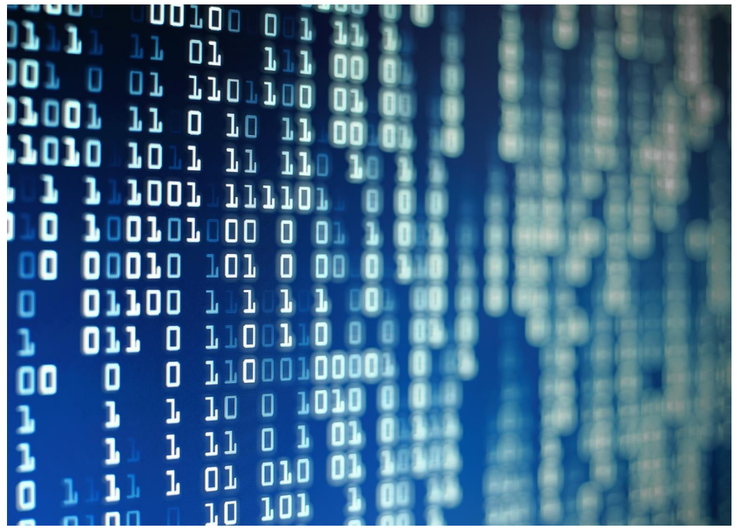
DATA ACQUISITION

Dataset Description

- The dataset contains information on a number of features, including the transaction amount, the time of the transaction, the merchant name, and the location of the transaction.

	step	customer	age	gender	zipcodeOri	merchant	zipMerchant	category	amount	fraud
0	0	'C1093826151'	'4'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	4.55	0
1	0	'C352968107'	'2'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	39.68	0
2	0	'C2054744914'	'4'	'F'	'28007'	'M1823072687'	'28007'	'es_transportation'	26.89	0
3	0	'C1760612790'	'3'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	17.25	0
4	0	'C757503768'	'5'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	35.72	0
...	...	...	...	...	...	...	...	...	...	...
594638	179	'C1753498738'	'3'	'F'	'28007'	'M1823072687'	'28007'	'es_transportation'	20.53	0
594639	179	'C650108285'	'4'	'F'	'28007'	'M1823072687'	'28007'	'es_transportation'	50.73	0
594640	179	'C123623130'	'2'	'F'	'28007'	'M349281107'	'28007'	'es_fashion'	22.44	0
594641	179	'C1499363341'	'5'	'M'	'28007'	'M1823072687'	'28007'	'es_transportation'	14.46	0
594642	179	'C616528518'	'4'	'F'	'28007'	'M1823072687'	'28007'	'es_transportation'	26.93	0

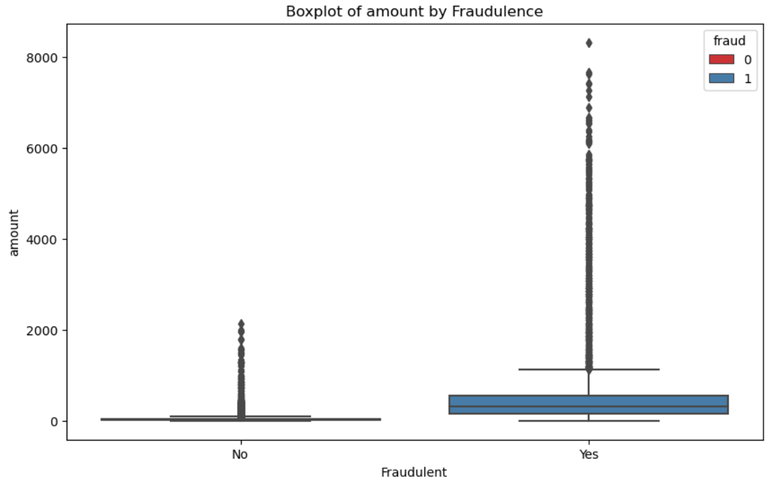
594643 rows × 10 columns

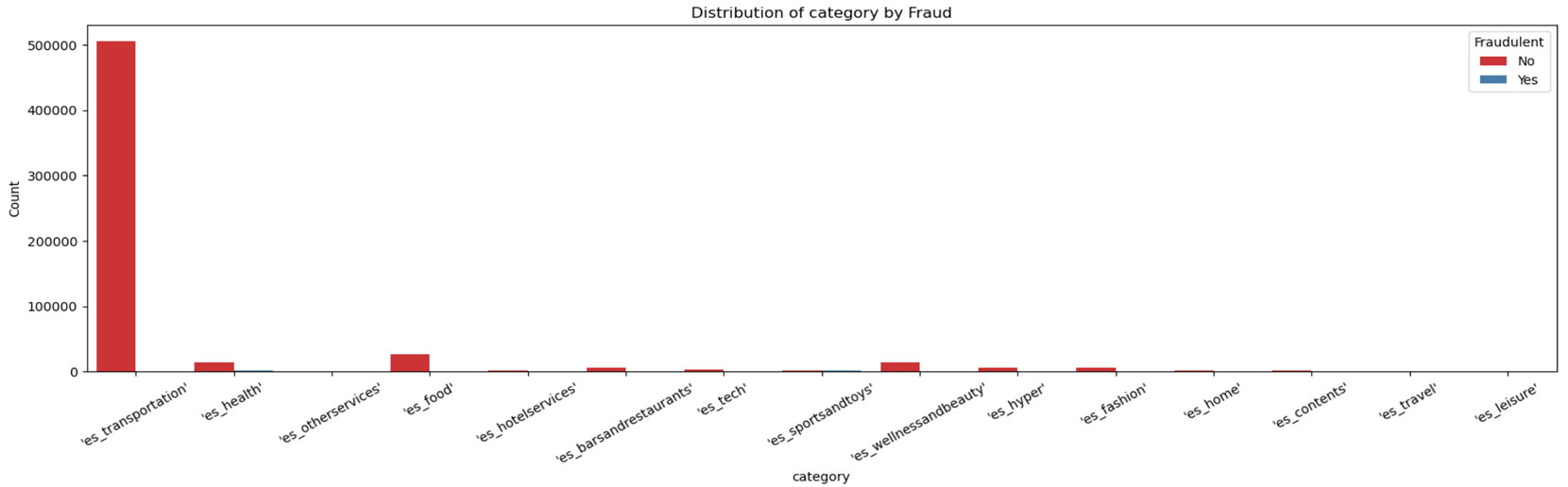


Methodology

Data Preprocessing

- The data was preprocessed to prepare it for modeling.
- This included handling missing values, encoding categorical variables, and scaling numerical variables.
- Also addressed the class imbalance in the dataset, as there were many more non-fraudulent transactions than fraudulent transactions.
- To address this imbalance, used a technique called SMOTE (Synthetic Minority Oversampling Technique) to oversample the minority class (fraudulent transactions).





Model Selection & Evaluation

Classification Algorithms

- A number of different machine learning models were evaluated for this project.
- These models included logistic regression, random forest, gradient boosting, and neural networks.
- Logistic Regression and Gradient Boosting Classifier have similar performance to each other, but fall short of the Random Forest in terms of accuracy and F1 score.
- Neural Network Classifier shows a good balance between precision and recall, but its overall accuracy is slightly lower than the Random Forest.

Logistic Regression

F1 Score before imbalance: 0.7258
F1 Score after balance: 0.9610

Random Forest

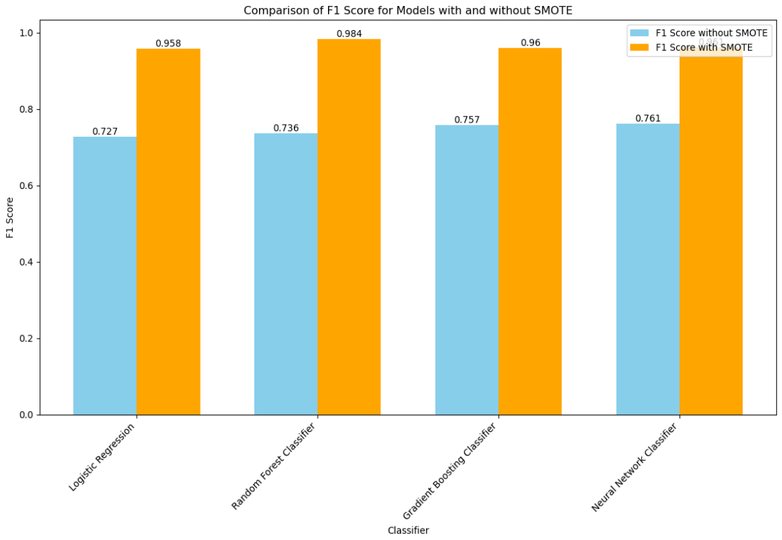
F1 Score before imbalance: 0.7461
F1 Score after balance: 0.9842

Gradient Boosting

F1 Score before imbalance: 0.7565
F1 Score after balance: 0.9636

Neural Networks

F1 Score before imbalance: 0.7595
F1 Score after balance: 0.9712



Testing with test dataset

Evaluating model performance on unseen data

- Utilized a separate test dataset to evaluate the model's performance on unseen data.
- Applied the trained random forest classifier SMOTE model to make predictions on the test dataset.
- Assessed the model's ability to correctly classify fraudulent and non-fraudulent transactions.

Evaluation Metrics after applying SMOTE:

	Classifier	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.956736	0.928724	0.989297	0.958054
1	Random Forest Classifier	0.983701	0.975967	0.991785	0.983813
2	Gradient Boosting Classifier	0.958298	0.929227	0.992058	0.959615
3	Neural Network Classifier	0.959656	0.936835	0.985676	0.960635

	Value	Count
0	0	584655
1	1	9988

Fine-tuning and optimization

Hyperparameter Optimization

- Fine-tuned model hyperparameters using Grid Search Cross-Validation, optimizing n_estimators, max_depth, min_samples_split & min_samples_leaf.

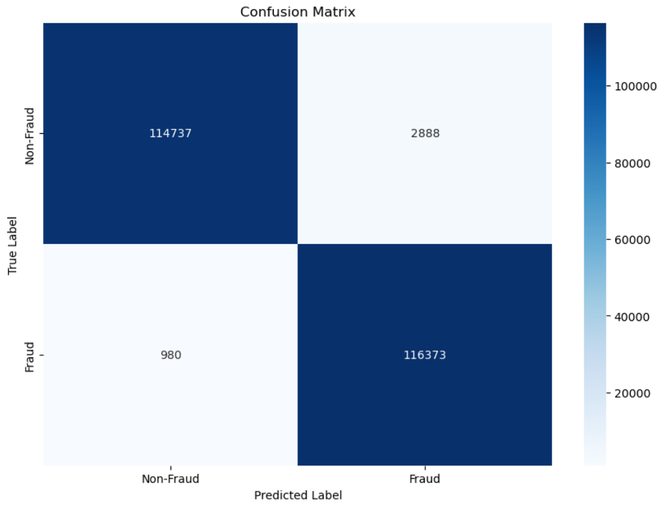
- Grid Search helped identify the best combination of hyperparameters to maximize model performance

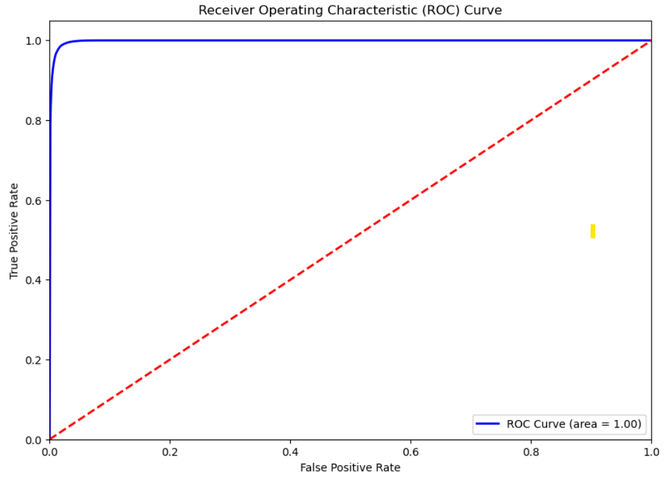
- Achieved an exceptional ROC AUC score of 0.998, signifying excellent discrimination between fraudulent and non-fraudulent transactions..

Classification Report:

	precision	recall	f1-score	support
0	0.99	0.98	0.98	117625
1	0.98	0.99	0.98	117353
accuracy			0.98	234978
macro avg	0.98	0.98	0.98	234978
weighted avg	0.98	0.98	0.98	234978

Confusion Matrix:
[[114737 2888]
[980 116373]]





Rule based vs ML based fraud detection systems

Key differences

Rule-based fraud detection	ML-based fraud detection
Catching obvious fraudulent scenarios	Finding hidden and implicit correlations in data
Requires much manual work to enumerate all possible detection rules	Automatic detection of possible fraud scenarios
Multiple verification steps that harm user experience	The reduced number of verification measures
Long-term processing	Real-time processing

Key Features

- step int64
- category object (Used One Hot Encoder here)
- amount float64
- fraud int64
- age_'0' int64
- age_'1' int64
- age_'2' int64
- age_'3' int64
- age_'4' int64
- age_'5' int64
- age_'6' int64
- age_'U' int64
- gender_'E' int64
- gender_'F' int64
- gender_'M' int64
- gender_'U' int64

Model Selection Features

These are the key comments used for the model building that gave us the accuracy and F1 score needed.

Key outcome

Business Value

- Test the model’s effectiveness on completely unseen future data to ensure its reliability in real-world scenarios

- Credit card fraud is a significant problem for financial institutions.
- In 2020, the estimated global cost of credit card fraud was $30.5 billion.
- By implementing a machine learning model for credit card fraud detection, financial institutions can reduce their losses from fraud.
- Extract additional features such as transaction time, location, and device used to enhance model performance and fraud detection accuracy.

Recommendations

Regular model improvement

- Implementing the model on a dataset of 370,604 transactions occurring over 118 hours, with a fraud rate of 1.21%, can significantly improve fraud detection and financial security.

- Continuously monitor the model's performance to adapt to evolving fraud patterns and ensure its effectiveness over time.

- Update the model with new data, including newly identified fraudulent transactions, and consider incorporating additional features such as device information to improve its accuracy over time.

- By monitoring and updating the model, we can help ensure that it remains effective in detecting credit card fraud.

Target

Progressive Chart

Audience

Adult

Woman

Teenager

Money Loosed before building model — **$ 30B**

Money saved after implementing model — **$ 25B**

Conclusion

- In conclusion, machine learning can be a powerful tool for credit card fraud detection.
- The model that developed in this project was able to achieve a high level of accuracy on the test set (98.4%), while also maintaining a good balance between precision (97.5%) and recall (99.3%).
- By implementing this model, financial institutions can reduce their losses from fraud.
- However, it is important to remember that the model is not perfect and needs to be monitored and updated on an ongoing basis.

THANK YOU

- Vinay Pavan GUNTUR