CREDIT CARD FRAUD PREDICTION

With Random Forest and Feedforward Neural Networks

Yu Weng | June 10, 2025

DATASET

Source: Kaggle - Credit Card Fraud Prediction

555,719 instances and 22 attributes

	merchant	category	amt	gender	city	state	zip	lat	long	job	unix_time	merch_lat	merch_long	is_fraud	year	month	day	hour	minute	age
0	fraud_Kirlin and Sons	personal_care	2.86	M	Columbia	SC	29209	33.9659	-80.9355	Mechanical engineer	1371816865	33.986391	-81.200714	0	2020	6	21	12	14	57
1	fraud_Sporer-Keebler	personal_care	29.84	F	Altonah	UT	84002	40.3207	-110.4360	Sales professional, IT	1371816873	39.450498	-109.960431	0	2020	6	21	12	14	35
2	fraud_Swaniawski, Nitzsche and Welch	health_fitness	41.28	F	Bellmore	NY	11710	40.6729	-73.5365	Librarian, public	1371816893	40.495810	-74.196111	0	2020	6	21	12	14	54

PREPARING DATA

Data Exploration and Understanding

- Highly imbalanced data (**99.62% not fraud** vs. **0.38% fraud**) - SMOTE

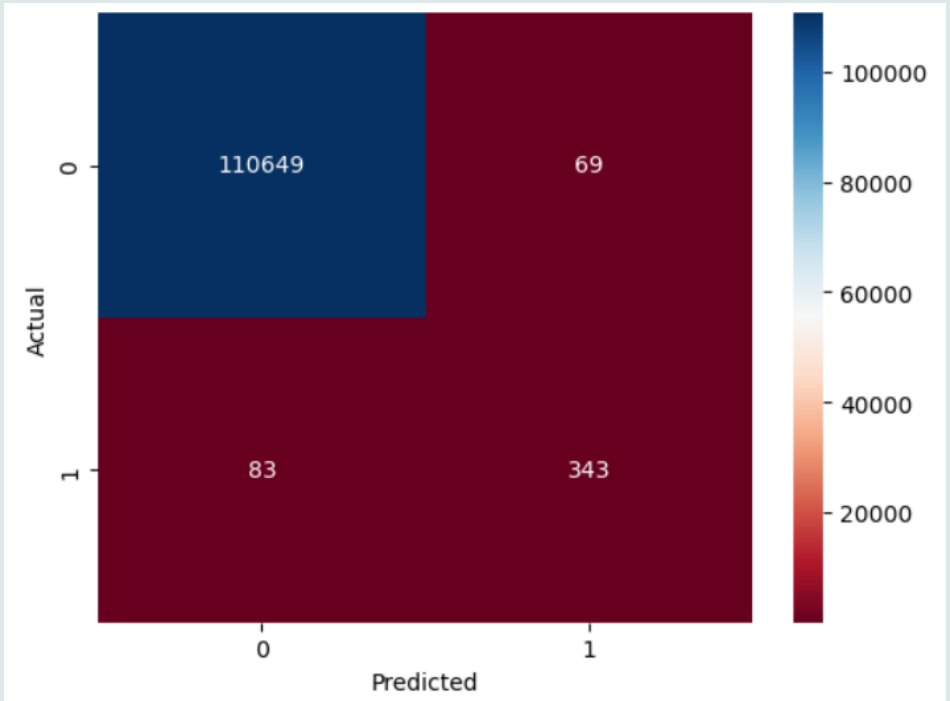
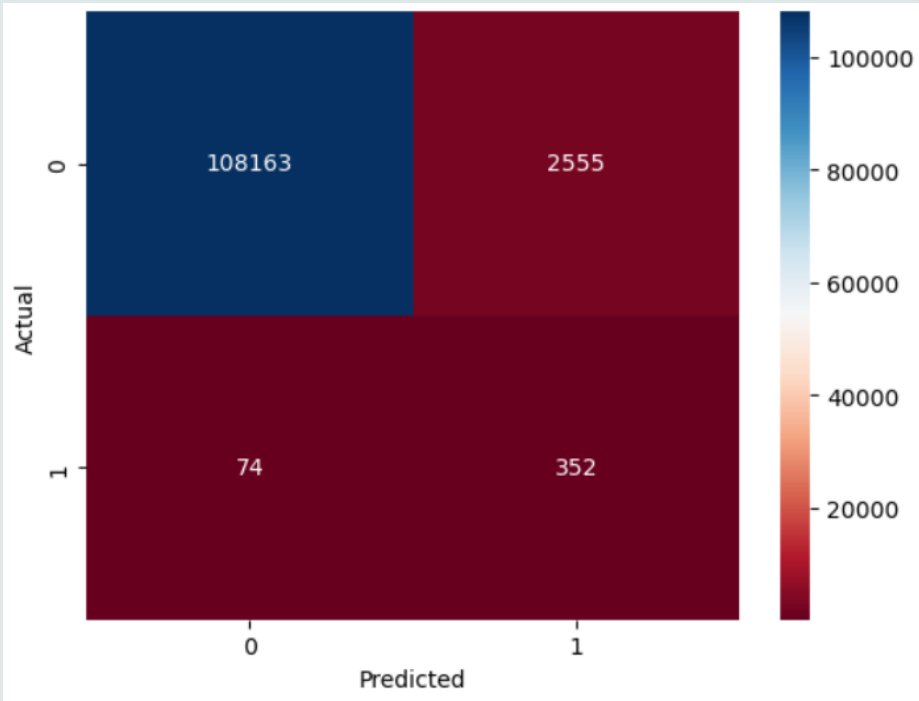
Feature Engineering

- Feature Transformation
- Variable Encoding
  - Categorical features: gender, merchant, category, state, job…
- Feature Scaling
  - Numerical features: transaction amount, age

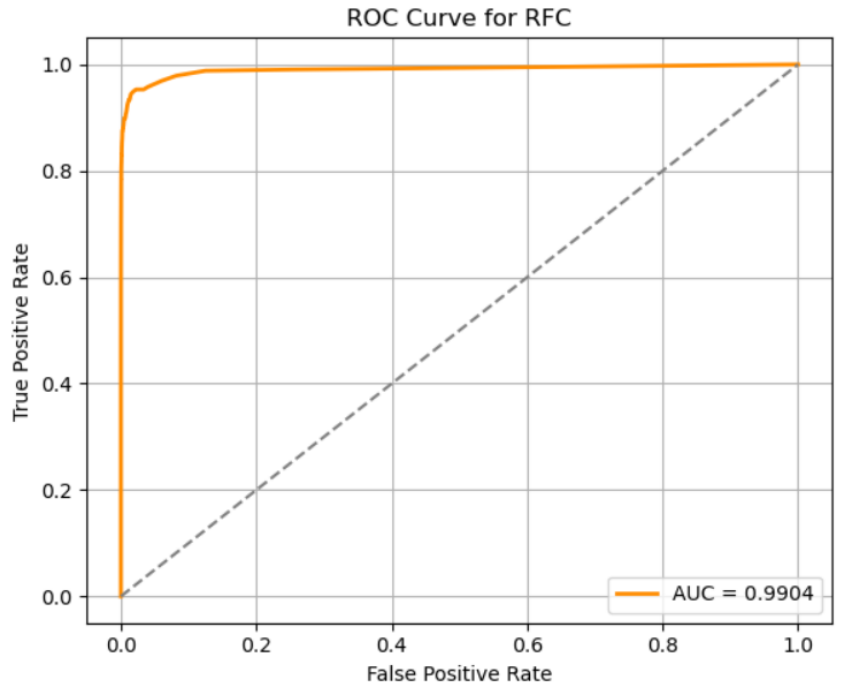
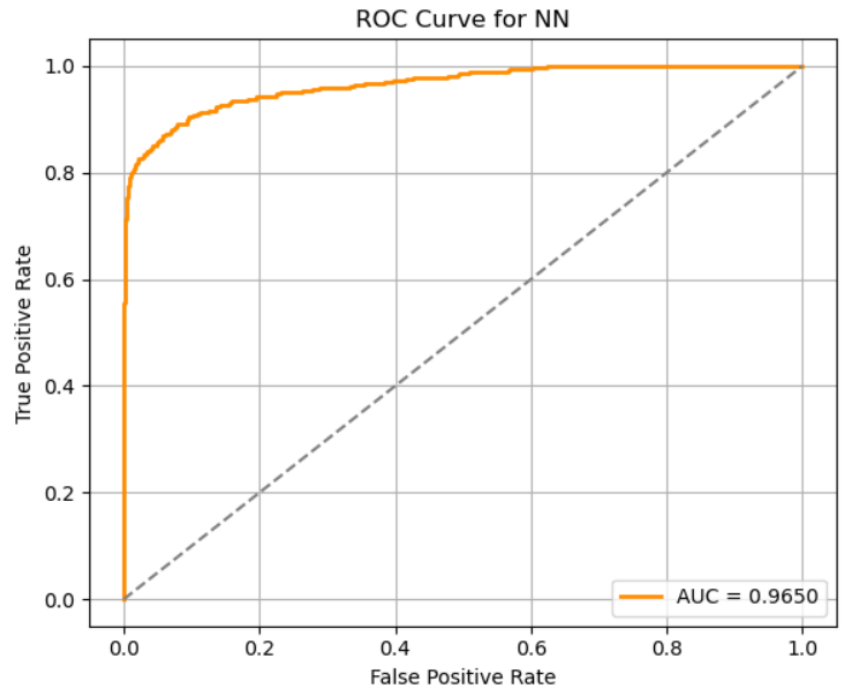
MODEL SELECTION

APPROACH	MODEL	STRENGTHS
Traditional Machine Learning	Random Forest Classifier	Robust, interpretable, handles imbalanced/tabular data well
Deep Learning	Feedforward Neural Network	Captures complex, non-linear relationships in the data

PERFORMANCE – CONFUSION MATRIX

Random Forest Classifier	Feedforward Neural Network
	

PERFORMANCE – ROC CURVE

Random Forest Classifier	Feedforward Neural Network
	

PERFORMANCE

Metrics	Random Forest Classifier	Feedforward Neural Network
Precision	0.83	0.12 *too many false positives!
Recall (Sensitivity)	0.81	0.83
F1-Score	0.82	0.21
AUC	0.99	0.96

CONCLUSION

In fraud detection:

- Prioritize recall to catch more frauds
- Low precision can overload analysts and slow down approvals
- Accuracy is misleading due to imbalance

Random Forest is a strong baseline model for fraud detection

Neural Network could outperform Random Forest if:

- Time-based behavioral patterns are included
- Embedded features are utilized
- Scaled to handle tens of millions of transactions

GitHub Repository