Predicting Fraud and FICO Score

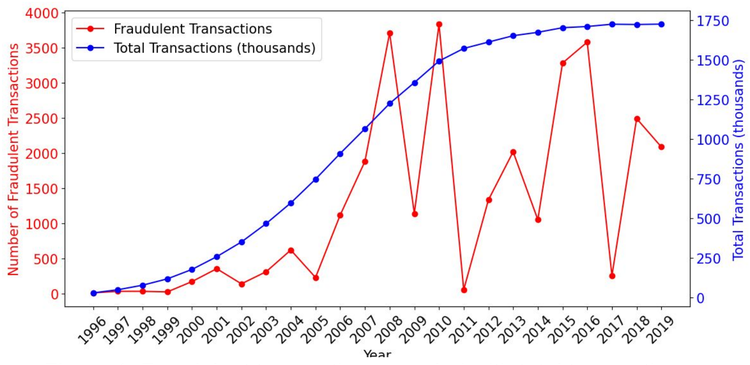
Matthew Keller, Computer Science '25
Moultrie Dangerfield, Computer Science '25
Emirhan Gencer, Computer Science ‘27
Jack Patterson, Computer Science ‘25

INTRODUCTION

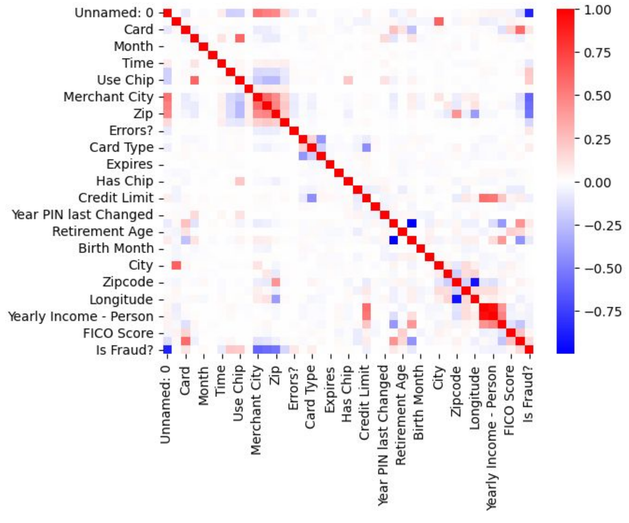
According to the Federal Trade Commision, fraud losses have increased by 14% between 2022 and 2023 and is projected to grow even further. Fraud detection and credit scoring are critical in financial services to mitigate risks and improve decision-making. This project explores predictive modeling for detecting fraudulent transactions and estimating FICO scores.

DATA

- Sourced from Kaggle
- Created by a multi-agent virtual world simulation performed by IBM
- 24 million credit card transactions ranging from 2009-2022
- The dataset is split into 3 main files
  1. **Transactions** - 24 million transactions
  2. **Credit cards** - Credit cards used to make purchases
  3. **Users** - Contains information on a total of ‘x’ users made purchases recorded in transactions.csv



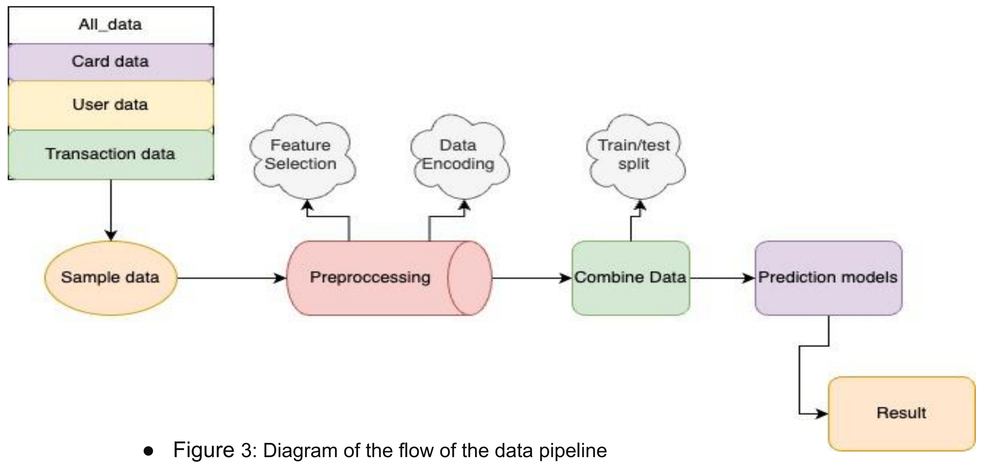
- Figure 1: Fraudulent transactions coupled with total transactions



- Figure 2: Correlation map of all features. Note features related to is_fraud?

METHODOLOGY

- Sampled data to get 50% fraudulent transactions and 50% legitimate transactions
- Categorized raw FICO score into 4 classes (Very Low,Low,Medium,High) based on industry standards
- Selected relevant features for each model
- Encoded categorical features to numerical values



- Figure 3: Diagram of the flow of the data pipeline

- Fraudulent Detection features:

  In US, in state, ZIP code, online transaction, swipe transaction, is suspicious amount, amount, number of credit cards

- FICO Prediction features:

  Current Age, Birth Year, Gender, Yearly Income - Person, Total Debt, Number of Credit Cards

RESULTS

Target Variable	Model	K	Accuracy	Precision	Recall	F1
FICO	NN (cosine)	25	38.50%	43.22%	81.70%	56.38%
	NN (euclid)	25	39.25%	43.40%	81.66%	56.67%
	NB	-	42.25%	42.25%	100.00%	59.40%
	Random	-	42.25%	41.18%	28.99%	34.00%
	Majority	-	42.25%	42.25%	100.00%	59.40%
Fraud	NN (cosine)	101	75.85%	86.85%	69.37%	70.35%
	NN (euclid)	5	75.6%	87.1%	64.1%	72.4%
	NB	-	36.95%	24.64%	16.50%	19.77%
	Random	-	49.90%	49.84%	37.29%	41.55%
	Majority	-	36.95%	24.64%	16.50%	19.77%

- Figure 4: Performance Metrics for Fraud Detection and FICO score prediction models

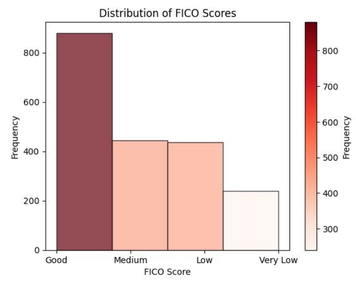
- Fraud detection best model: Cosine and Euclidean Performed 30% better than baseline random model for the F1 metric.

Measure / model	K	Accuracy	Precision	Recall	F1
Cosine	1	70.7%	81.1%	64.2%	60.4%
Cosine	50	74.00%	84.6%	67.9%	67.3%
Cosine	101	75%	85.9%	68.9%	69.2%
Euclidean	1	74.7%	83.1%	66%	67.6%
Euclidean	3	76.3%	86.8%	66.6%	72.6%
Euclidean	5	75.1%	87.1%	64.1%	72.4%

- Euclidean 3 KNN yielded the best performance metrics for fraud detection.
- FICO score prediction best model: Naive Bayes. Performed 22.38% better than baseline random model for the F1 metric.
- 25 KNN yielded the best performance metrics for FICO score prediction.

ERROR ANALYSIS

- One fraudulent charge for every ~800 legitimate charges.
- Skewed dataset- Large majority of FICO scores fell into the “good” category. This caused errors for Naive Bayes and large KNN.



- Figure 5: Uneven distribution of FICO scores

- Majority classes heavily favored in both models

DISCUSSION

- Challenges: Large data size, encoding nominal features, skewed data.
- Our classification of the FICO scores into four categories may have oversimplified the variability of credit score.
- Further work on this subject could benefit from additional data such as regional economic data.

https://colab.research.google.com/drive/1f0ZdyU4-x5zj-0CD3qSza46WXbIhPvIV?usp=sharing