Data Masters' Ultimate Fraud Detector

Ash Manoj, Ky Nguyen, Kiyonna Kapoor, Ethan Howard

Table of contents

01

Project Outline

“65 percent of people with credit or debit cards have experienced credit card fraud at least once”[1]

[1]https://www.security.org/digital-safety/credit-card-fraud-report/#:~:text=According%20to%20our%20research%2C%2065,had%20been%20victims%20of%20fraud.

Our Audience + Goals

- The primary audience is **Credit Card Companies**
- Our models can be used by creditors to flag fraudulent transactions.
- Minimize **false negatives**

Problem statement + Approach

- **Credit Card Fraud is a serious problem**

What we want to do:

- **Supervised Binary Classification to make predictions**
- **Suggest implementing the best model along with other best practices**

02

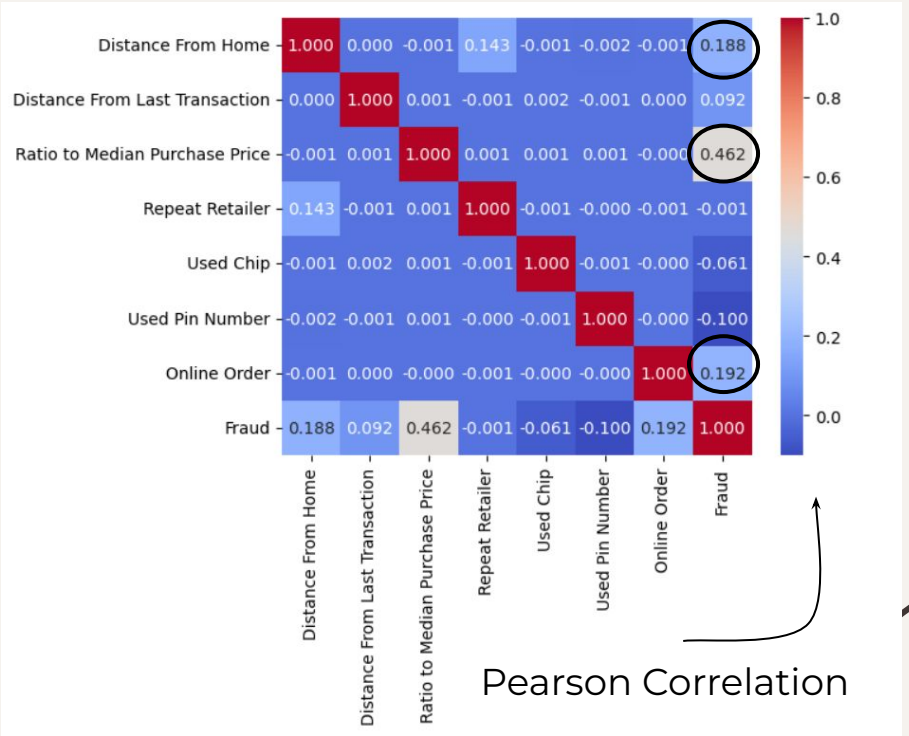
Dataset + Features

Feature Table

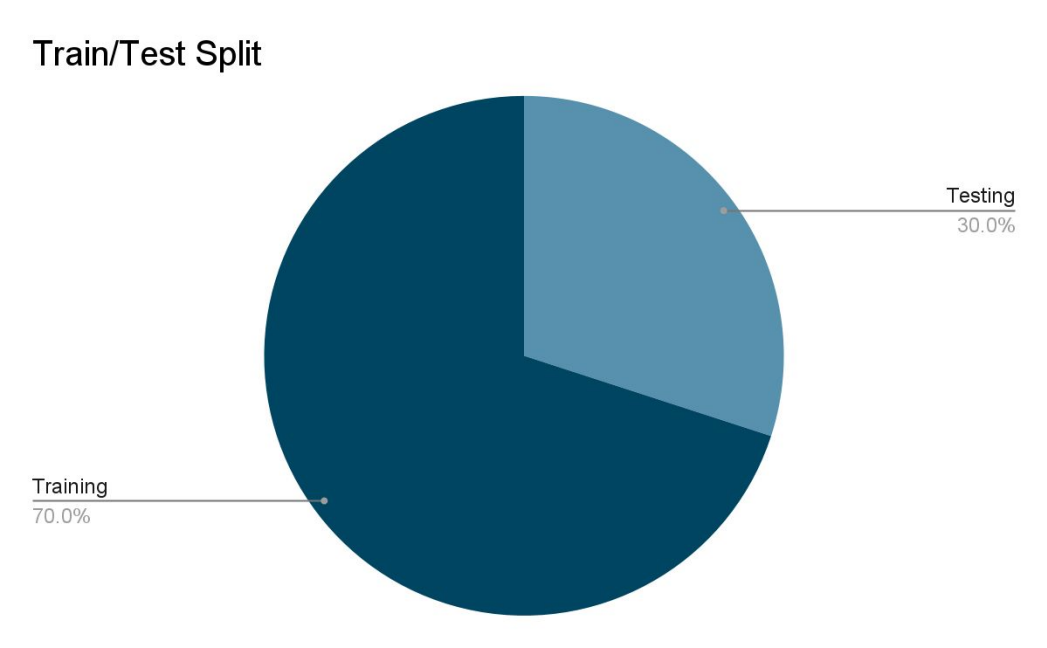
Feature	Distance from Home	Dist. from last transaction	Ratio to Median Purchase	Repeat Retailer	Used Chip	Used Pin Number	Online Order
Datatype	Float	Float	Float	Boolean	Boolean	Boolean	Boolean
Rationale	Longer distance, more likely fraud	Longer distance, more likely fraud	Higher ratio, more likely fraud	New retailer, more likely fraud	Chip is more secure	Using pin is less likely to be fraud	Easier to fraud an online purchase

Statistics + EDA Visualizations

	Distance From Home	Distance From Last Transaction	Ratio to Median Purchase Price
count	1000000.000000	1000000.000000	1000000.000000
mean	26.628792	5.036519	1.824182
std	65.390784	25.843093	2.799589
min	0.004874	0.000118	0.004399
25%	3.878008	0.296671	0.475673
50%	9.967760	0.998650	0.997717
75%	25.743985	3.355748	2.096370
max	10632.723672	11851.104565	267.802942



Our Sample



7:3 Ratio
Training to Testing

Resampling Methods

Undersampling		Oversampling	
Random	NearMiss	Random	SMOTE
Remove samples from the majority class, with or without replacement	Use k nearest-neighbors NearMiss3 1. For each negative sample, kept m nearest-neighbors 2. Select positive samples with largest average distance to the k nearest-neighbors		

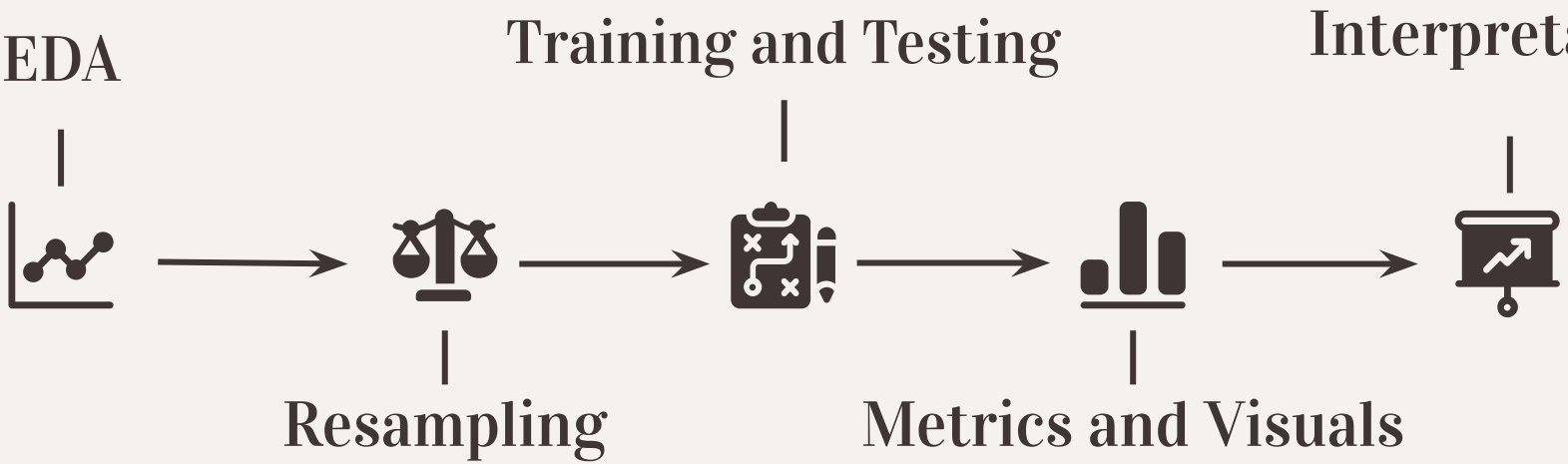
Resampling Methods

Undersampling		Oversampling	
Random	NearMiss	Random	SMOTE
		Randomly selecting examples from the minority class with replacement and adding them to the training dataset.	Use k nearest-neighbors. Select n of k instances to interpolate new synthetic instances by taking difference between a sample and its nearest neighbour and multiply the difference by a random value in (0, 1].

03

Our Process

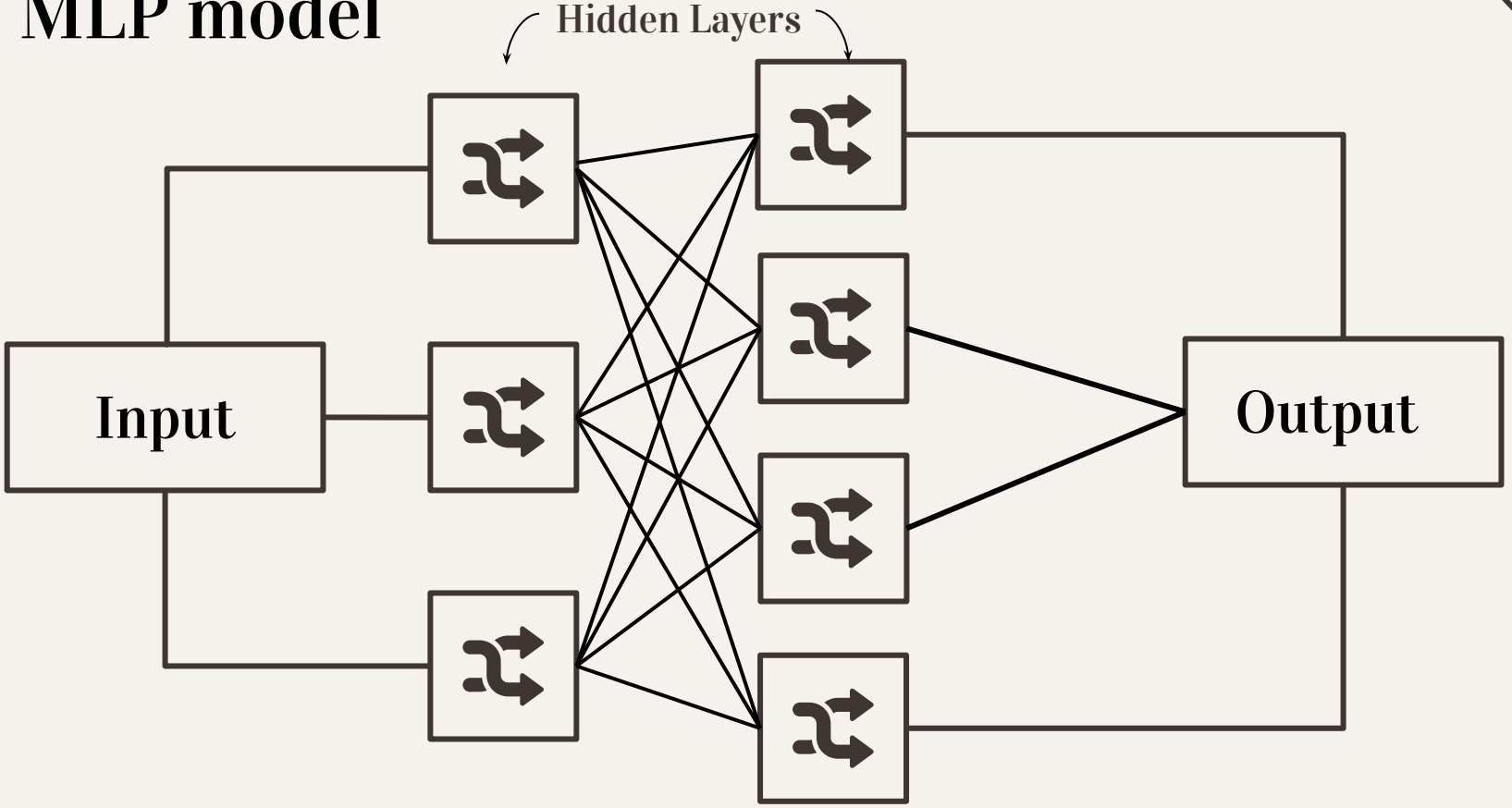
Our Process


EDA
Resampling
Training and Testing
Metrics and Visuals
Interpretation

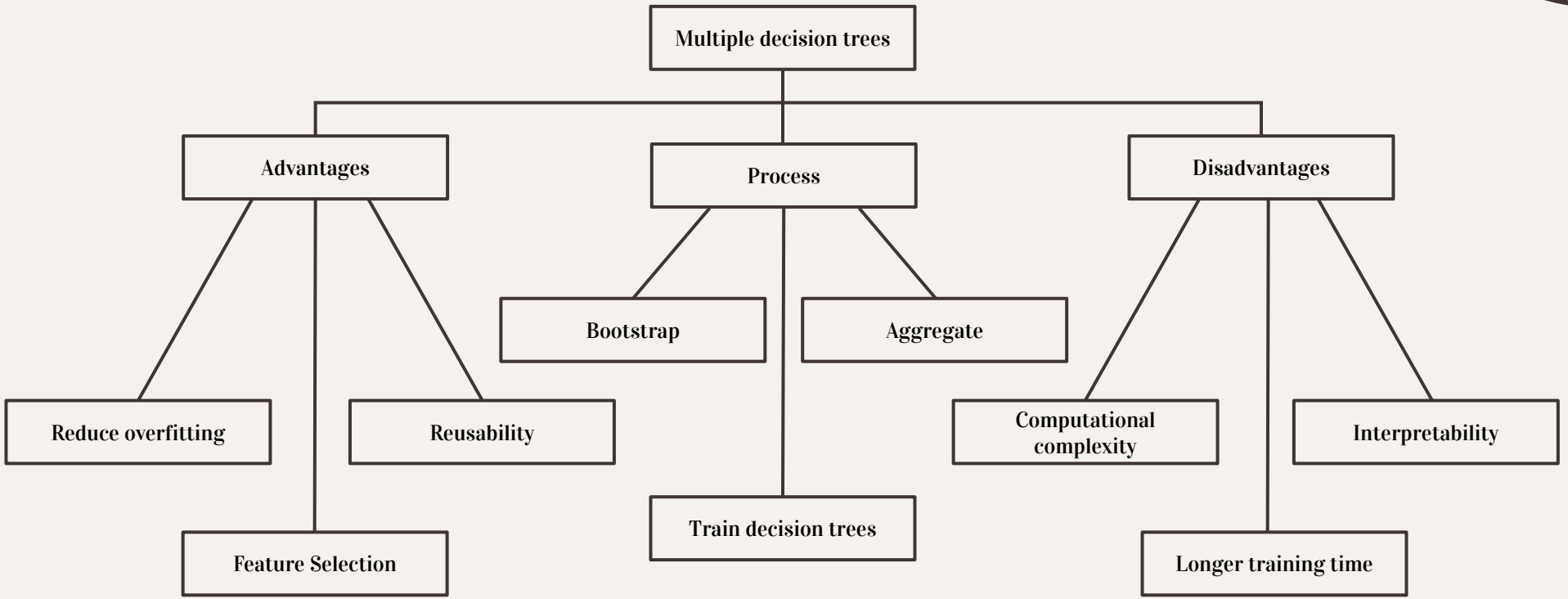
04

Our Models

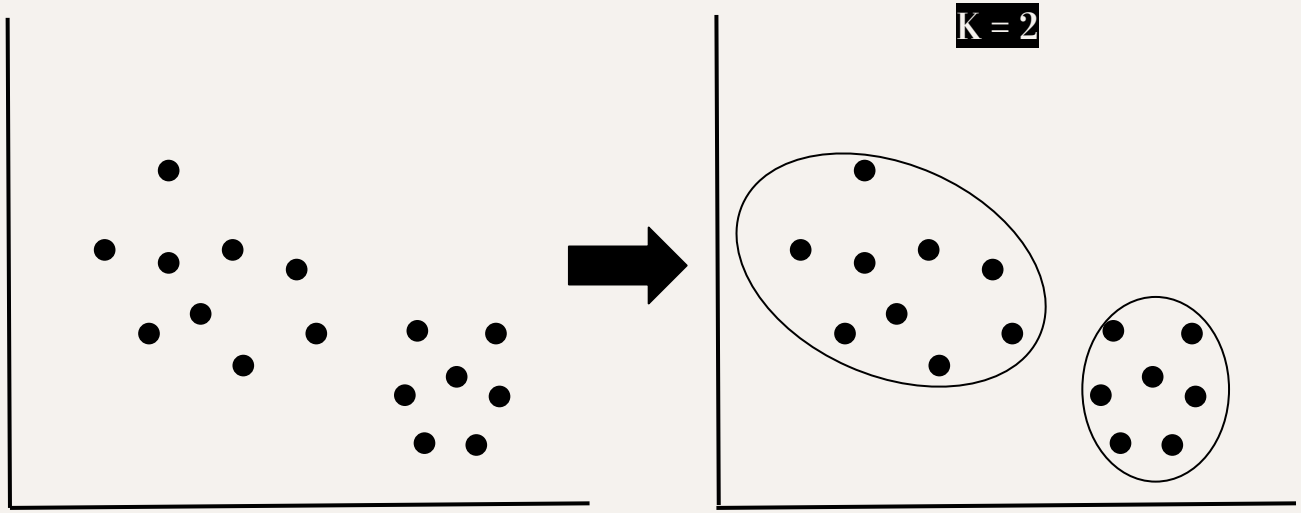
MLP model



Random Forest Model



K-Means Clustering



Elbow Point (k)

Measure the Distance

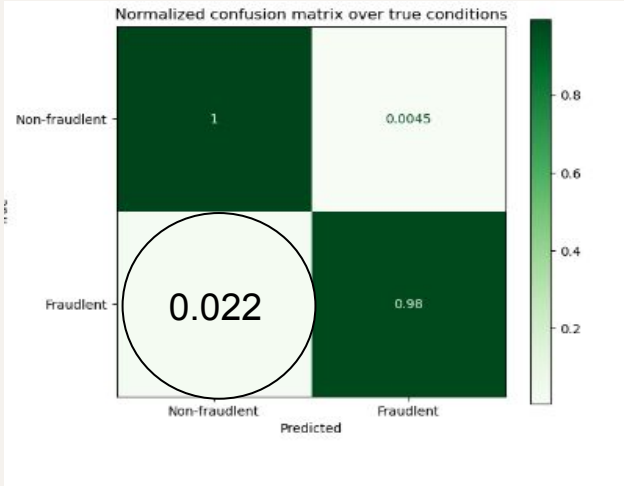
Grouping based on min Distance

Reposition the centroid

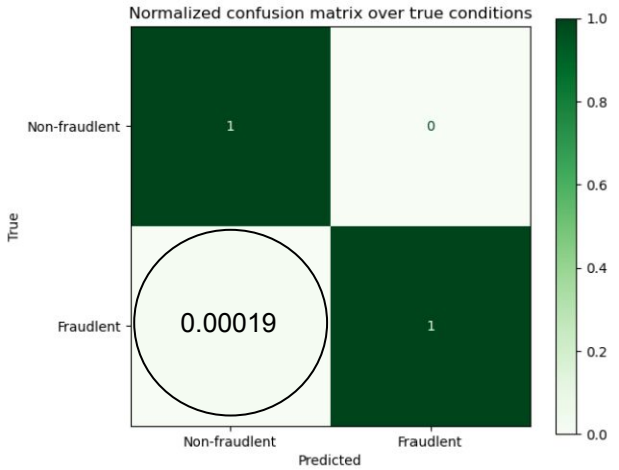
Convergence

Recall Confusion Matrices

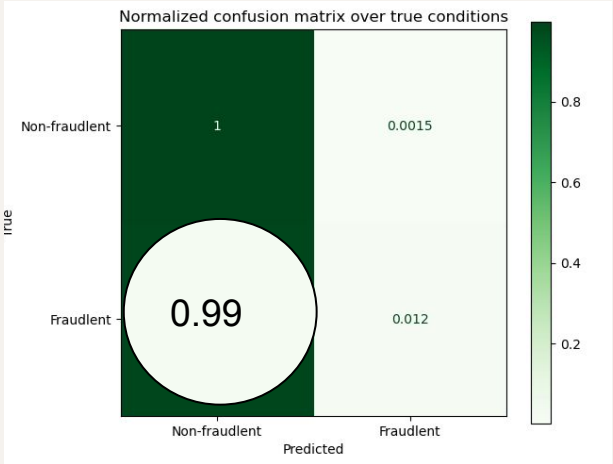
MLP



Random Forest



K-Means



The bottom left squares show the case we want to minimize (false negatives)
K-Means error due to imbalance of test data

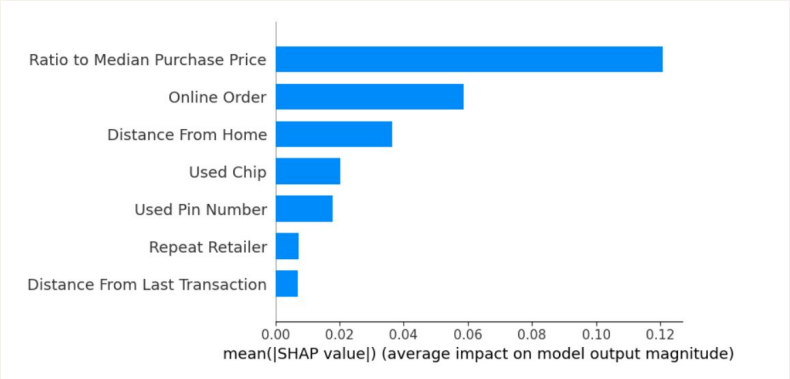
05

Results + Applications

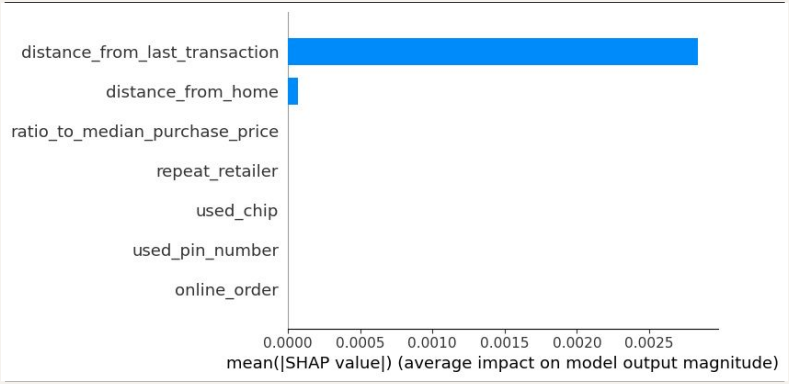
Overview + Accuracy Scores

Sampling Method	MLP	Random Forest	K-Means Clustering
Random Undersampling	92.36%	99.99+%	8.86%
NearMiss Undersampling	99.39%	99.99+%	91.28%
Random Oversampling	92.57%	99.99+%	91.15%
SMOTE Oversampling	92.92%	99.99+%	91.13%
None(control)	93.7%	97.72%	91.12%

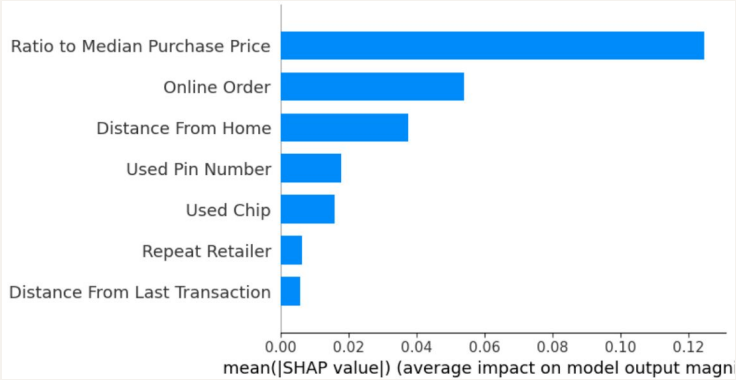
SHAP Analysis



MLP NearMiss



K Means



Random Forest ROS

06
Limitations

Limitations

Model Limitations	Sampling Limitations	Interpretability Limitations	Feature Limitations
K-Means model was unsuitable for this dataset Produced **8% accuracy for RUS** method, which raised concern	NearMiss undersampling **did not generate a balanced dataset** The dataset originally was **very imbalanced**, which is presumably the cause	We could only run **SHAP for 10%** of test data Interpretation might vary from truth	Dataset only had **7 features**, so there could be confounding variables that are not features

THANK YOU

GitHub: https://github.com/ashmanoj/CCFraudDetector

07

LIVE DEMO (surprise!!)

We will take questions while we set up our demo :)

Credit_Card_Fraud_Random_Forest_ros_model_live_demo_version2 - Jupyter Notebook

List of references

- https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud
- https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15
- https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
- https://scikit-learn.org/stable/modules/neural_networks_supervised.html
- http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
- https://www.security.org/digital-safety/credit-card-fraud-report/#:~:text=According%20to%20our%20research%2C%2065,had%20been%20victims%20of%20fraud.