CREDIT CARD FRAUD DETECTION

BY: SHANMUKHI BALAGAMSETTY

INTRODUCTION

OBJECTIVE: DEPLOY DATA ANALYSIS FOR IN-DEPTH DETECTION AND UNDERSTANDING OF CREDIT CARD FRAUD.

METHODOLOGY: COMPREHENSIVE ANALYSIS USING MACHINE LEARNING MODELS TO IDENTIFY AND ANALYZE FRAUDULENT TRANSACTIONS.

CHALLENGES: ADDRESSING THE COMPLEXITIES OF TRANSACTIONAL DATA AND THE NUANCES OF FRAUD DETECTION.

FOCUS: PRIORITIZING RECALL TO CAPTURE MORE FRAUD CASES WITHOUT SIGNIFICANTLY AFFECTING PRECISION, THROUGH TARGETED MODEL TUNING AND DATA STRATEGY.

DATA OVERVIEW

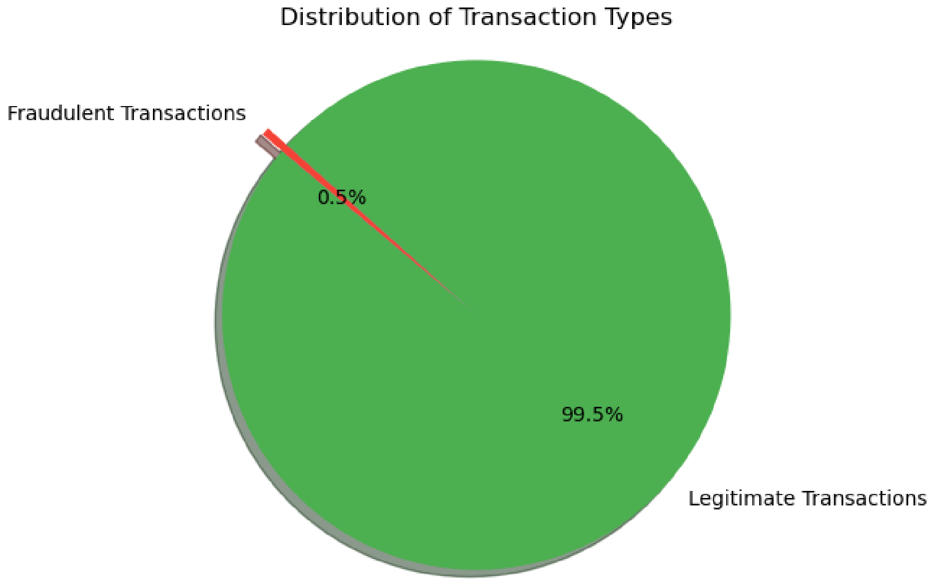
Dataset Source: Collected from Kaggle, covering Jan 2019 - Dec 2020.

Scope of Data: Encompasses transactions from 1,000 customers & 800 businesses.

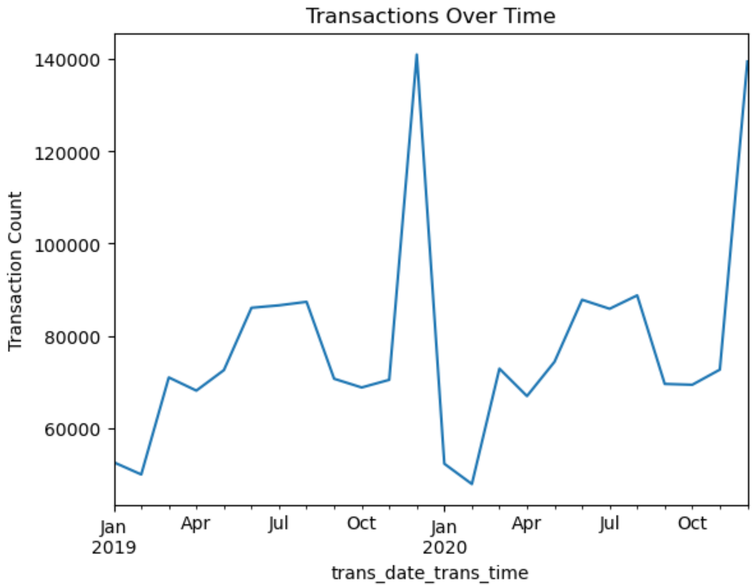
Data Specifics: Transactional details, merchant profiles, and geolocation.

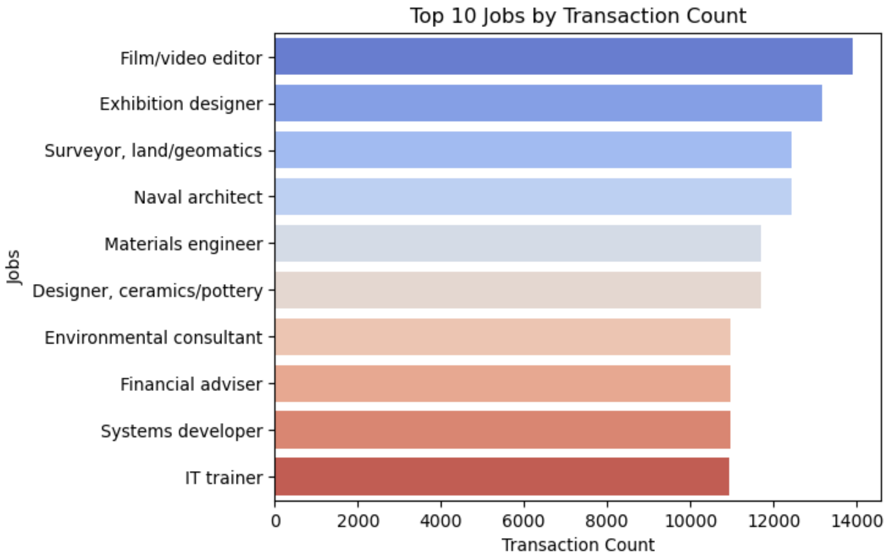
Transaction Types: Data on both legitimate and fraudulent credit card transactions.

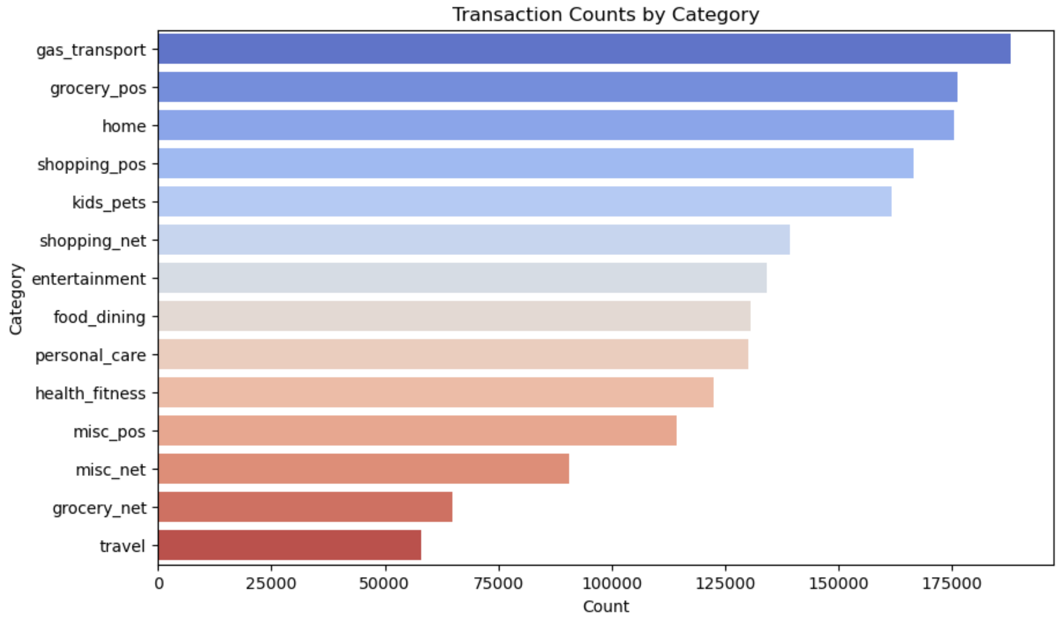
Data Integrity: 21 number of columns across 1852394 rows with diverse data types (no missing values or duplicates).



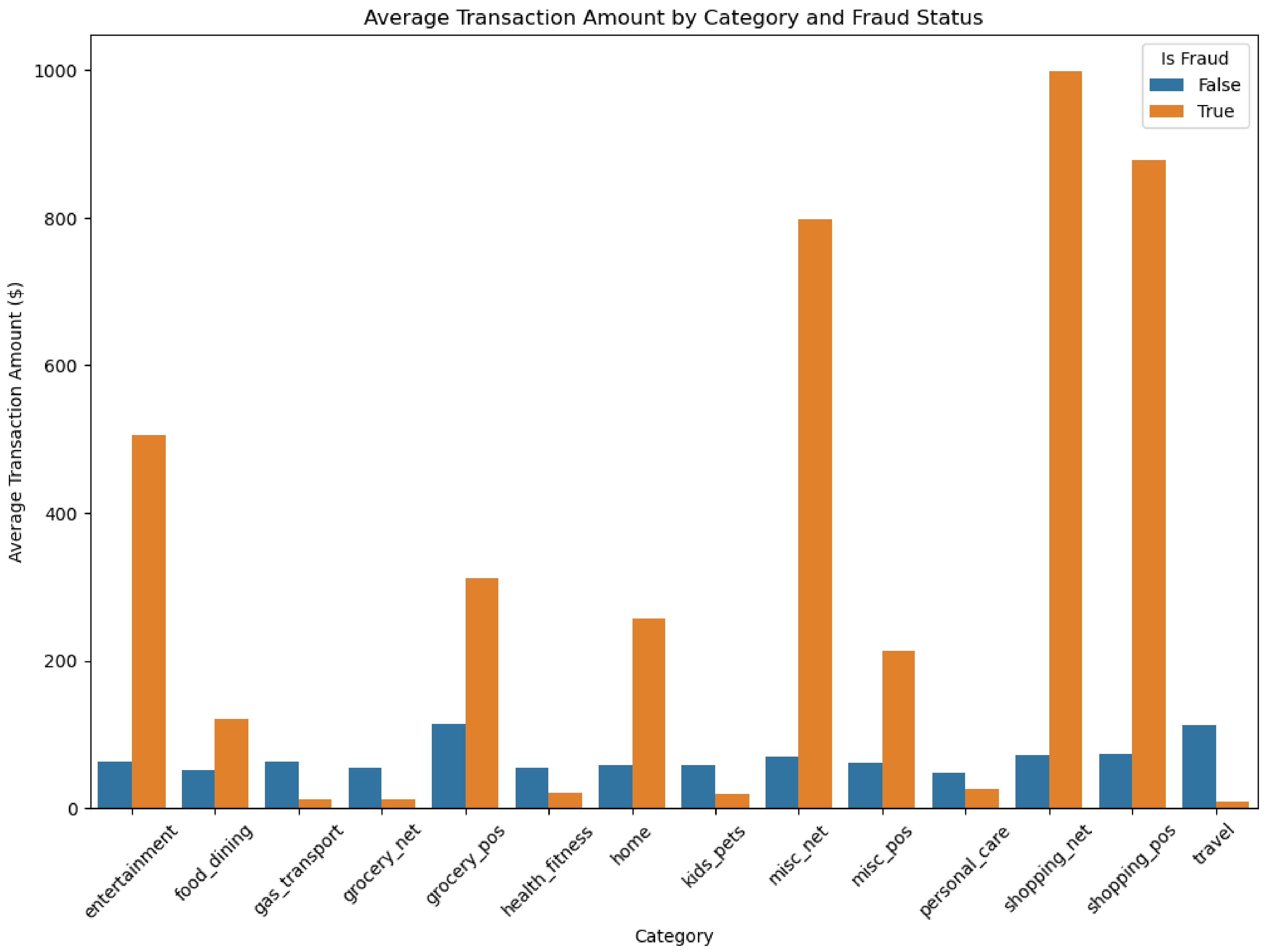
TRANSACTIONS OVERVIEW



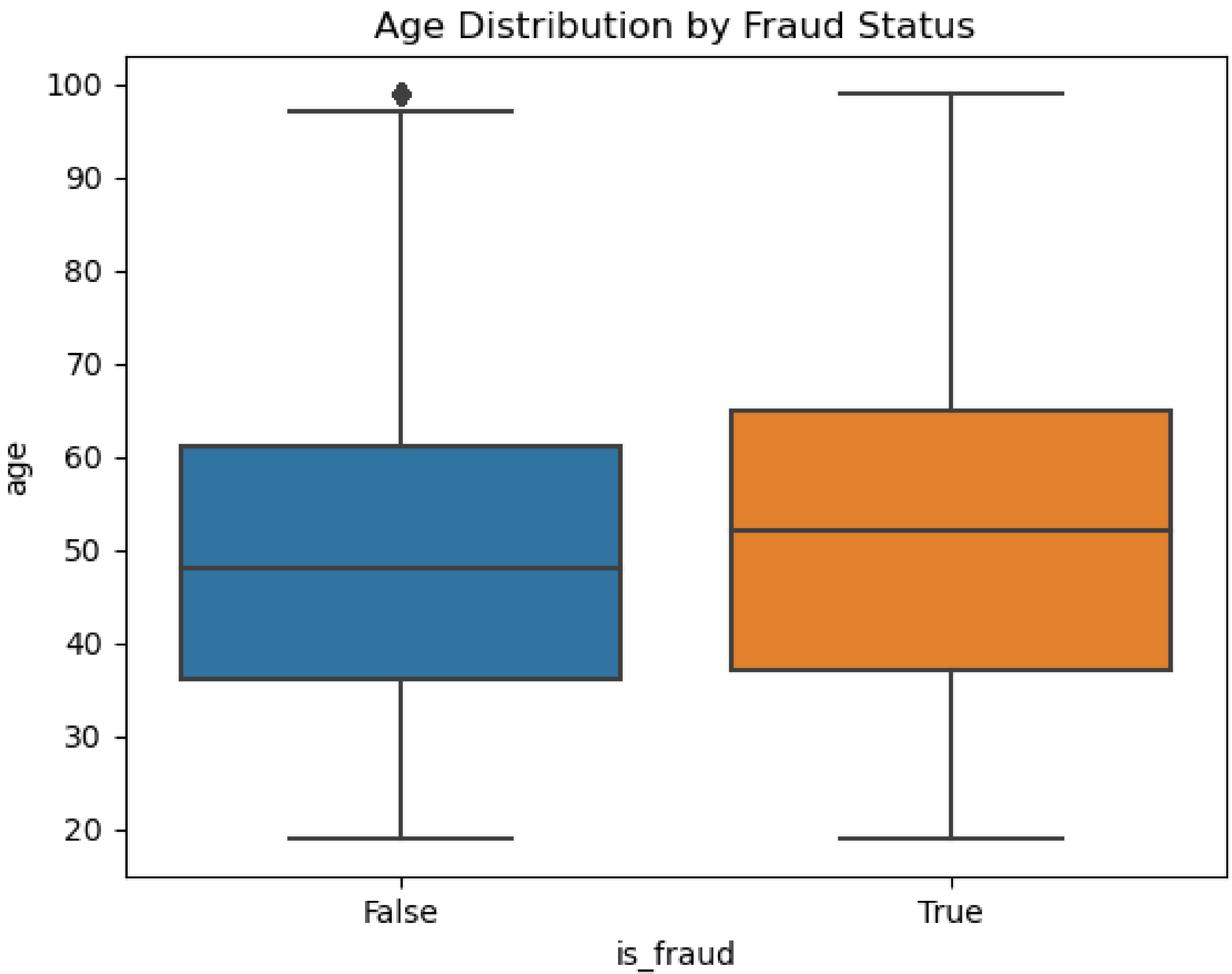


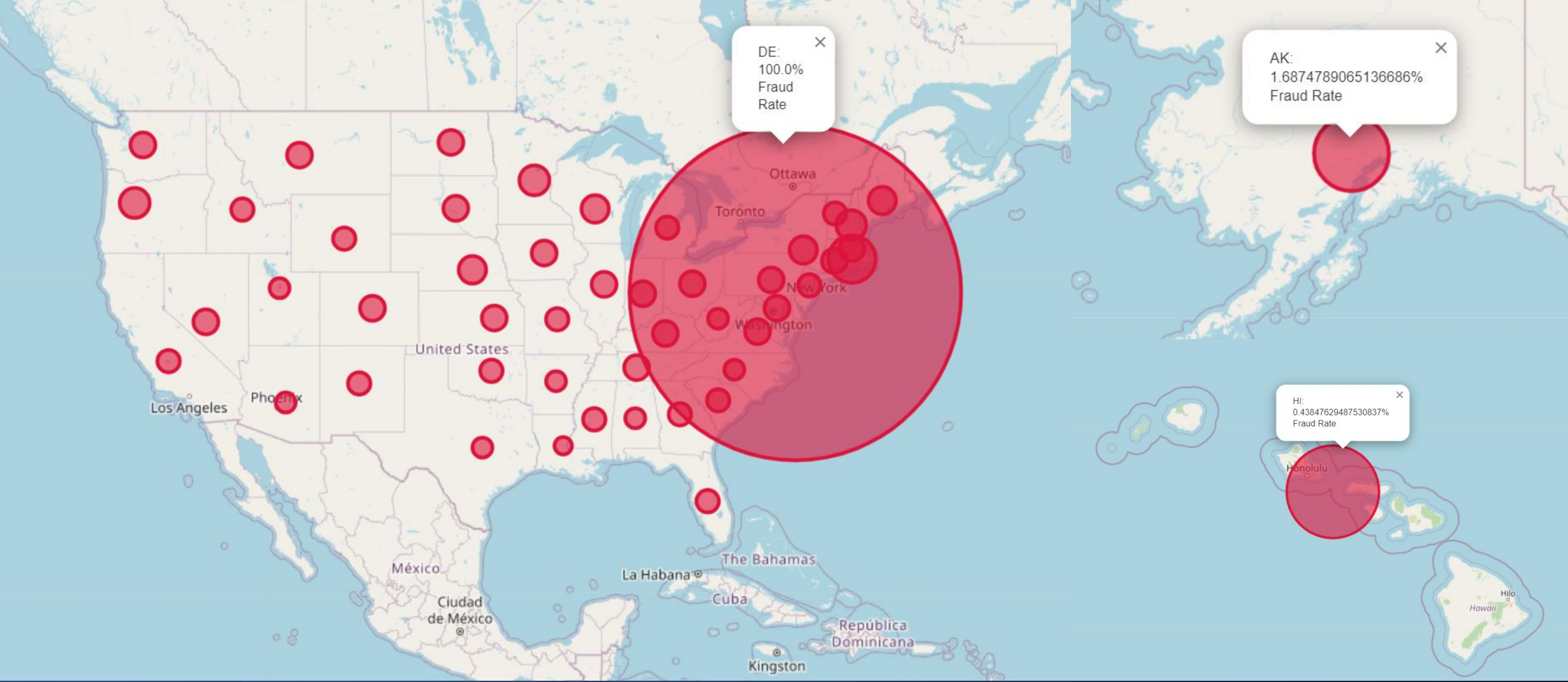


What types of purchases are most likely to be instances of fraud?



Are older customers significantly more likely to be victims of credit card fraud?





Fraud rates across different states.

FEATURE ENGINEERING

Distance Calculation: Combined customer and merchant coordinates to calculate transaction distances.

Time Features: Derived hour, day, month, year, and weekday from transaction timestamps.

Region Mapping: Segmented states into Northeast, Midwest, South, West to simplify regional analysis.

Job Categorization: Consolidated 497 job titles into broader career fields to reduce cardinality.

PREPROCESSING

Feature Removal: Eliminated non-predictive attributes including personal identifiers and redundant location details (e.g., 'cc_num', 'trans_num', 'first', 'last', etc.).

Dummy Variables: Transformed categorical variables into dummy/indicator variables for model compatibility.

Data Segregation: Separated features (X) and target variable (Y) to facilitate model training and evaluation.

Feature Scaling: Implemented Standard Scaler to normalize feature values, ensuring equal weight in distance-based algorithms.

Final Feature Set: Post-processing, the dataset contains 44 features engineered for optimal model performance.

BALANCING DATASET

Initial Challenge: Target class distribution was skewed in a large dataset of 1,852,394 rows.

Random Under Sampling (RUS): Implemented to balance the classes by reducing the size of the overrepresented class.

Complexity Consideration: Opted for RUS due to data complexity; other methods were less viable for handling such a vast dataset efficiently.

Post-RUS Dataset Size: Successfully reduced to 19,302 observations with 44 features.

Balanced Target Distribution: Achieved an equal split of the target variable with 9,651 instances in each class.

ML MODELS – LOGISTIC REGRESSION

Metrics	Accuracy	Precision	Recall	F1-score	CV Runtime
Solver					
‘lbfgs'	0.81 (+/- 0.02)	0.86 (+/- 0.03)	0.74 (+/- 0.06)	0.79 (+/- 0.03)	0.311936378
‘liblinear’	0.81 (+/- 0.02)	0.86 (+/- 0.03)	0.74 (+/- 0.06)	0.79 (+/- 0.03)	0.760750532
‘sag'	0.81 (+/- 0.02)	0.86 (+/- 0.03)	0.74 (+/- 0.06)	0.79 (+/- 0.03)	1.779671431
‘newton-cg'	0.81 (+/- 0.02)	0.86 (+/- 0.03)	0.74 (+/- 0.06)	0.79 (+/- 0.03)	0.430433273

ML MODELS – SUPPORT VECTOR MACHINE

Metrics	Accuracy	Precision	Recall	F1-score	CV Runtime
Kernel					
‘linear'	0.86 (+/- 0.01)	0.95 (+/- 0.01)	0.75 (+/- 0.03)	0.84 (+/- 0.02)	92.49470186
‘poly'	0.85 (+/- 0.04)	0.95 (+/- 0.02)	0.74 (+/- 0.08)	0.83 (+/- 0.05)	36.35176659
‘rbf'	0.87 (+/- 0.03)	0.94 (+/- 0.02)	0.78 (+/- 0.05)	0.86 (+/- 0.04)	44.87937307
‘sigmiod'	0.77 (+/- 0.01)	0.79 (+/- 0.02)	0.74 (+/- 0.04)	0.76 (+/- 0.01)	38.63424659

ML MODELS – K NEAREST NEIGHBORS

Metrics →	Accuracy	Precision	Recall	F1-score	CV Runtime
Neighbors ↓					
5	0.80 (+/- 0.03)	0.84 (+/- 0.02)	0.74 (+/- 0.06)	0.79 (+/- 0.03)	1.727298021
10	0.81 (+/- 0.03)	0.88 (+/- 0.01)	0.72 (+/- 0.06)	0.79 (+/- 0.04)	1.191655636
15	0.81 (+/- 0.03)	0.86 (+/- 0.02)	0.74 (+/- 0.06)	0.80 (+/- 0.04)	1.267908573
20	0.82 (+/- 0.03)	0.88 (+/- 0.03)	0.73 (+/- 0.06)	0.80 (+/- 0.03)	1.227131367
25	0.81 (+/- 0.02)	0.87 (+/- 0.02)	0.74 (+/- 0.05)	0.80 (+/- 0.03)	1.351754665

ML MODELS – DECISION TREES

Metrics / Criteria	Accuracy	Precision	Recall	F1-score	CV Runtime
'gini'	0.94 (+/- 0.07)	0.96 (+/- 0.01)	0.91 (+/- 0.14)	0.94 (+/- 0.08)	0.950036049
'entropy'	0.93 (+/- 0.07)	0.97 (+/- 0.01)	0.89 (+/- 0.14)	0.92 (+/- 0.08)	0.700759888

ML MODELS – RANDOM FOREST

criterion='gini'					
Metrics	**Accuracy**	**Precision**	**Recall**	**F1-score**	**CV Runtime**
Estimators					
10	0.96 (+/- 0.04)	0.98 (+/- 0.01)	0.95 (+/- 0.08)	0.96 (+/- 0.04)	4.81070471
20	0.96 (+/- 0.03)	0.98 (+/- 0.01)	0.95 (+/- 0.07)	0.96 (+/- 0.03)	8.82189441
50	0.96 (+/- 0.06)	0.98 (+/- 0.01)	0.95 (+/- 0.13)	0.96 (+/- 0.07)	22.3745515
100	0.96 (+/- 0.07)	0.98 (+/- 0.01)	0.94(+/- 0.13)	0.95 (+/- 0.07)	44.2629211
200	0.96 (+/- 0.06)	0.97 (+/- 0.01)	0.94(+/- 0.13)	0.96(+/- 0.07)	88.4445317
500	0.96 (+/- 0.07)	0.97 (+/- 0.01)	0.94(+/- 0.13)	0.95(+/- 0.08)	213.480791

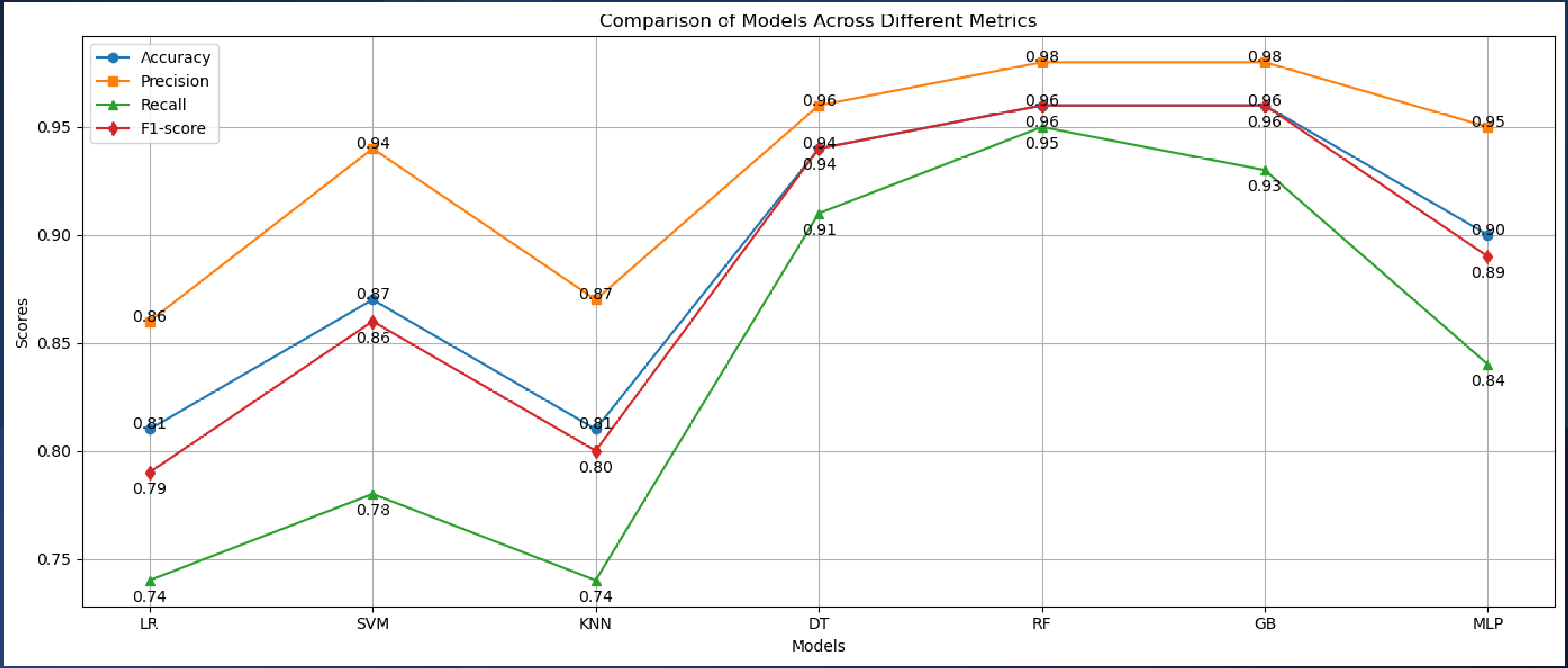
ML MODELS – GRADIENT BOOSTING

n_estimators=100					
Metrics	**Accuracy**	**Precision**	**Recall**	**F1-score**	**CV Runtime**
Maximum depth					
3	0.94 (+/- 0.04)	0.97 (+/- 0.00)	0.90 (+/- 0.09)	0.93 (+/- 0.05)	24.40419483
5	0.93 (+/- 0.04)	0.98 (+/- 0.01)	0.88 (+/- 0.08)	0.93 (+/- 0.05)	34.41374898
7	0.95 (+/- 0.03)	0.98 (+/- 0.00)	0.91 (+/- 0.07)	0.94 (+/- 0.04)	47.69124508
9	0.96 (+/- 0.03)	0.98 (+/- 0.01)	0.93 (+/- 0.07)	0.96 (+/- 0.04)	65.0622077
11	0.95 (+/- 0.07)	0.98 (+/- 0.01)	0.93 (+/- 0.14)	0.95 (+/- 0.08)	83.10816717

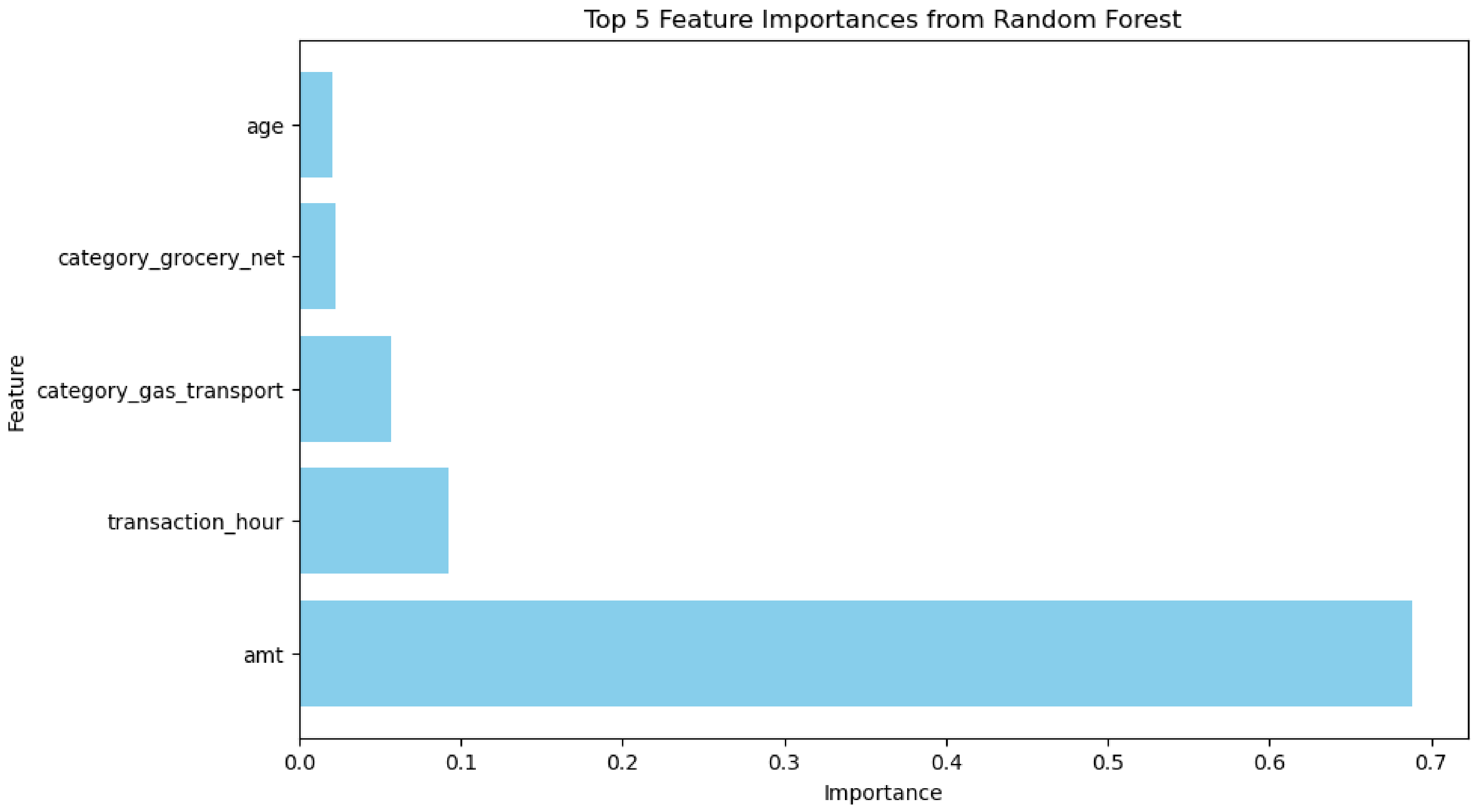
ML MODELS – MULTI LAYER PERCEPTON

Metrics	**Accuracy**	**Precision**	**Recall**	**F1-score**	**CV Runtime**
Solver					
‘lbfgs'	0.89 (+/- 0.05)	0.93 (+/- 0.01)	0.83 (+/- 0.10)	0.88 (+/- 0.06)	42.84511733
‘adam'	0.90 (+/- 0.05)	0.95 (+/- 0.01)	0.84 (+/- 0.10)	0.89 (+/- 0.05)	96.41093755
‘sgd’	0.87 (+/- 0.04)	0.92 (+/- 0.01)	0.81 (+/- 0.08)	0.86 (+/- 0.05)	115.5709627

COMPARISION BETWEEN ML MODELS



FEATURE IMPORTANCE



CONCLUSION

- Random Forest outperformed Gradient Boosting in recall, with both showing similar overall effectiveness.
- High-value internet and POS transactions were key indicators of fraud.
- Notable features: transaction amount, hour, and gas/transport categories.
- Challenges of high cardinality and large dataset management were mitigated by feature engineering and under sampling.
- Future efforts will concentrate on further feature reduction and investigating new feature selection methods.