Relevance of Personal Information in Credit Card Fraud Detection

Jason Lebov & Aarush Bedi

RESEARCH MOTIVATION

Binary Classification Study

- High Level of Concern Regarding Personal Information being used by Large Entities

- Current Credit Card Fraud Prediction Approach
  - Merchant, Transaction, & Card Holder Details

Research Question: Is a cardholder's personal information needed to accurately determine if a transaction is fraudulent? Which factors are the most influential in classification?

FEATURE ABLATION STUDY

‘A STUDY INVOLVING THE REMOVAL OF INDIVIDUAL FEATURES FROM THE MODEL AND MEASURING THE RESULTING CHANGES IN THE OUTPUT VARIABLES’

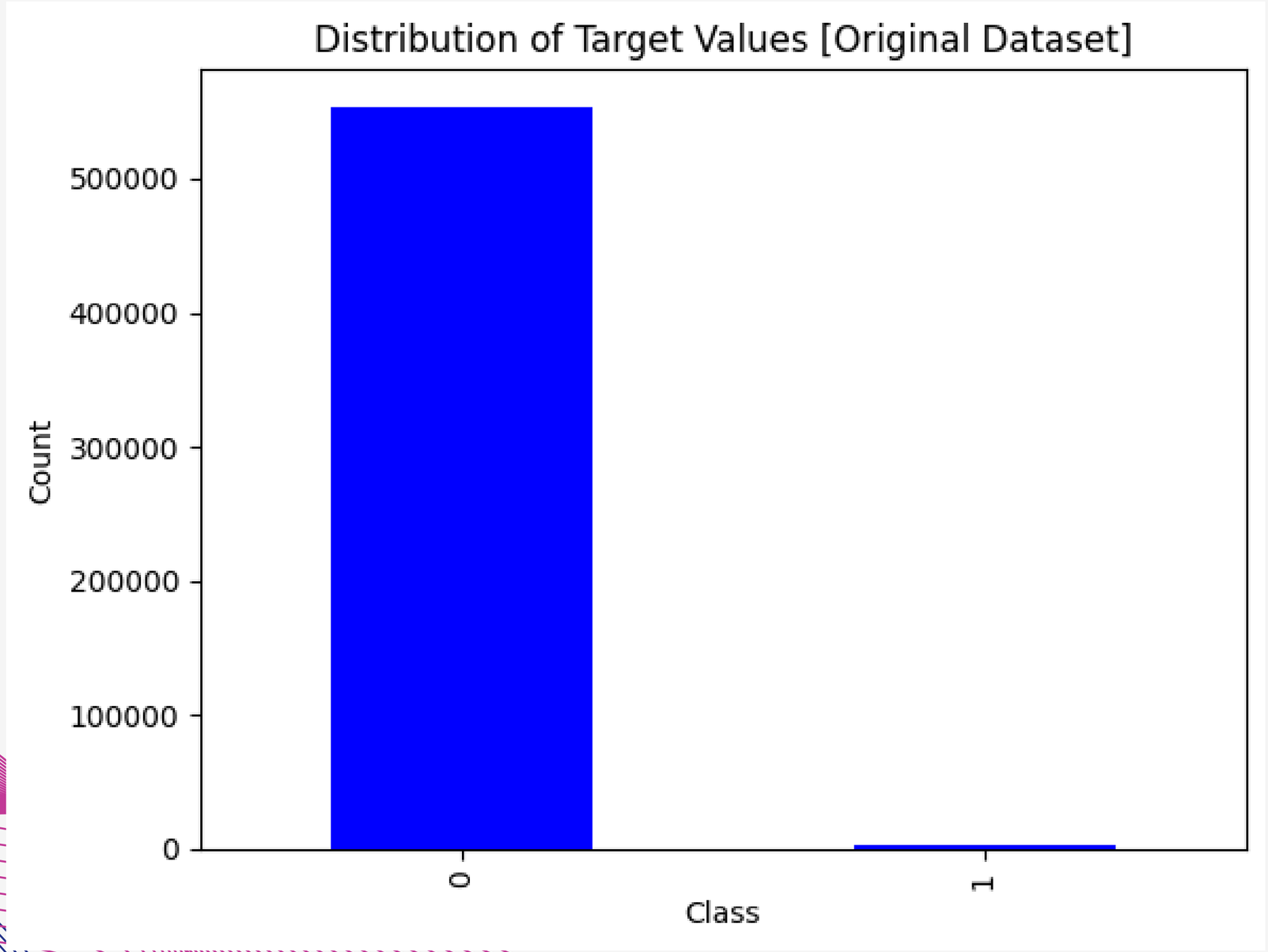
DATA EXPLORATION

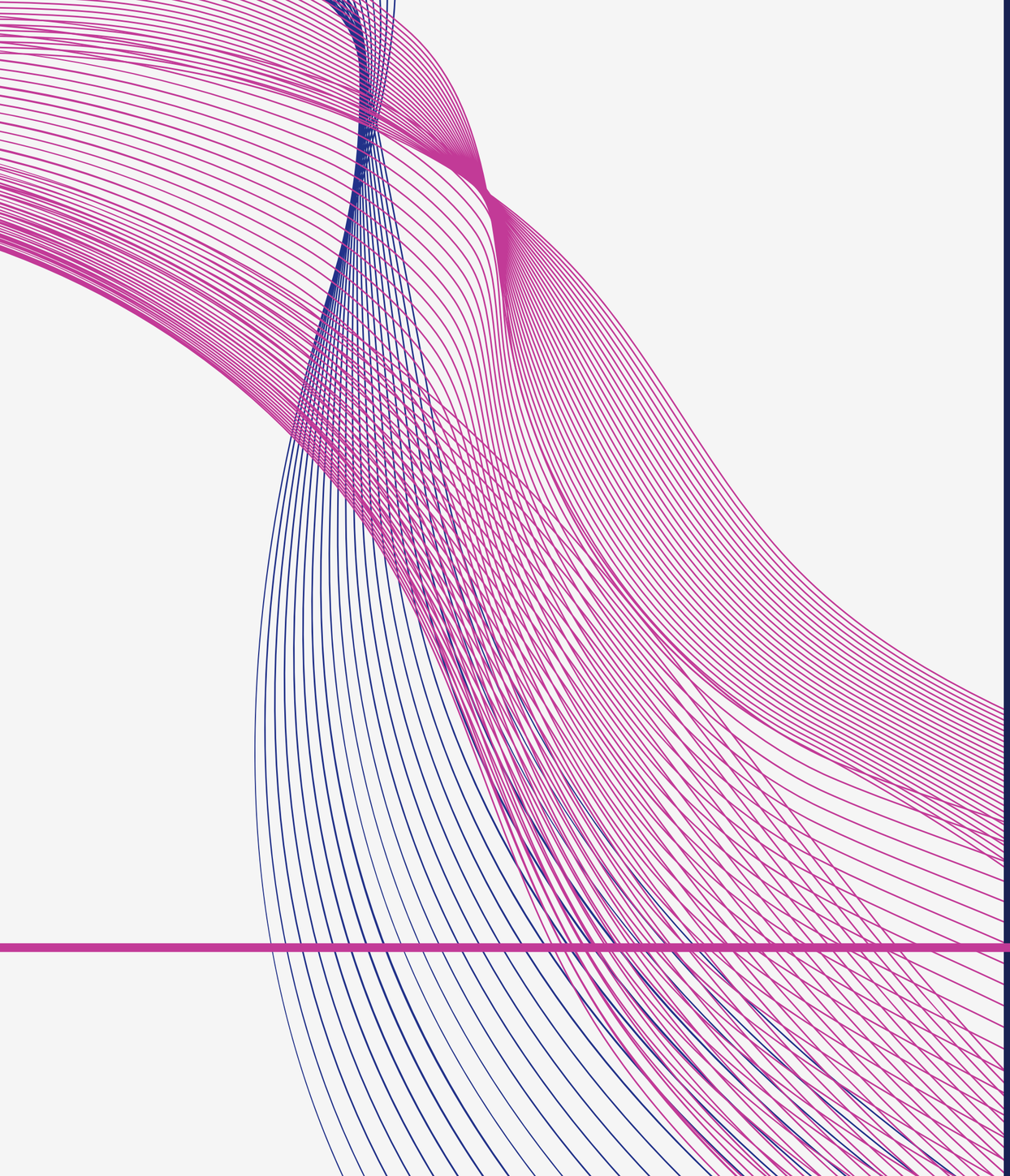
THE DATA

Kaggle Dataset - 555719 Samples, 23 Features

Information Regarding **Credit Card Transactions**...
| *Merchant Data, Transaction Data, Cardholder Data* |

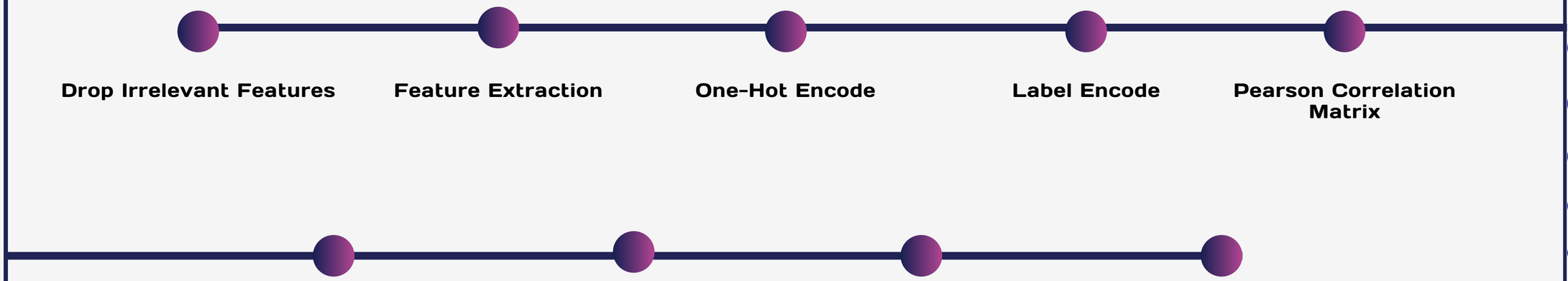
Target Variable *[Binary Classification]*
Fraudulent Transaction
Yes - 1, No - 0

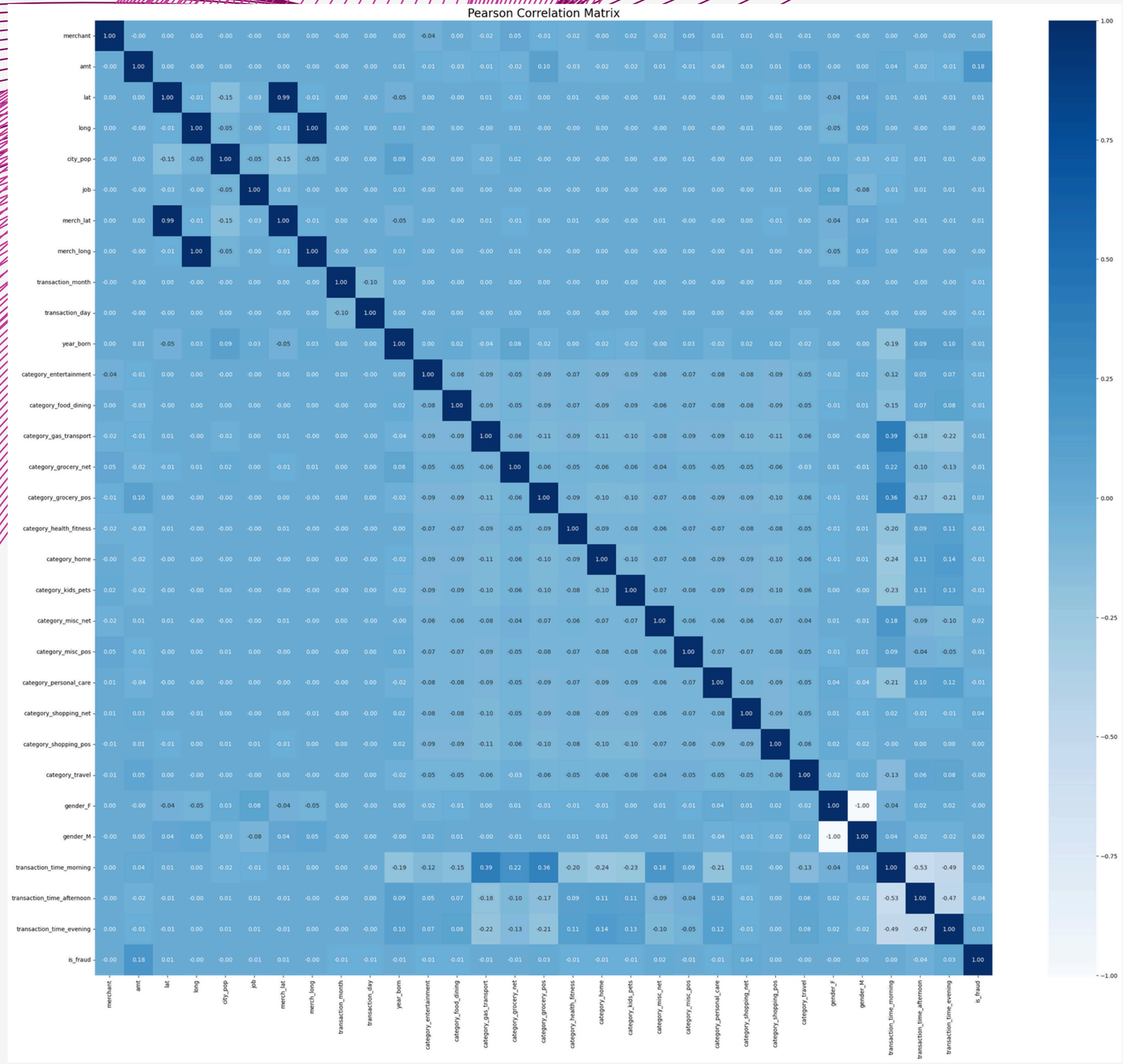

Distribution of Target Values [Original Dataset]
Count
500000
400000
300000
200000
100000
0
0
1
Class



DATA PREPROCESSING

PREPROCESSING STEPS





Drop:
‘merch_lat’ & ‘merch_long’

Post SMOTE Application

PI & No PI y_train Datasets

MODEL CREATION

Decision Tree

Fast

Class Imbalance

Easy to interpret

Simple to implement

GridSearchCV

Parameters Grids

Max Depth: range between (5, 100) w/ step size 5

Min Sample Split: [2, 10, 50, 100, 500]

Min Sample Leaf: [1, 5, 10, 50, 100]

Max Features: [’sqrt’, ‘log2’]

Optimal Parameters

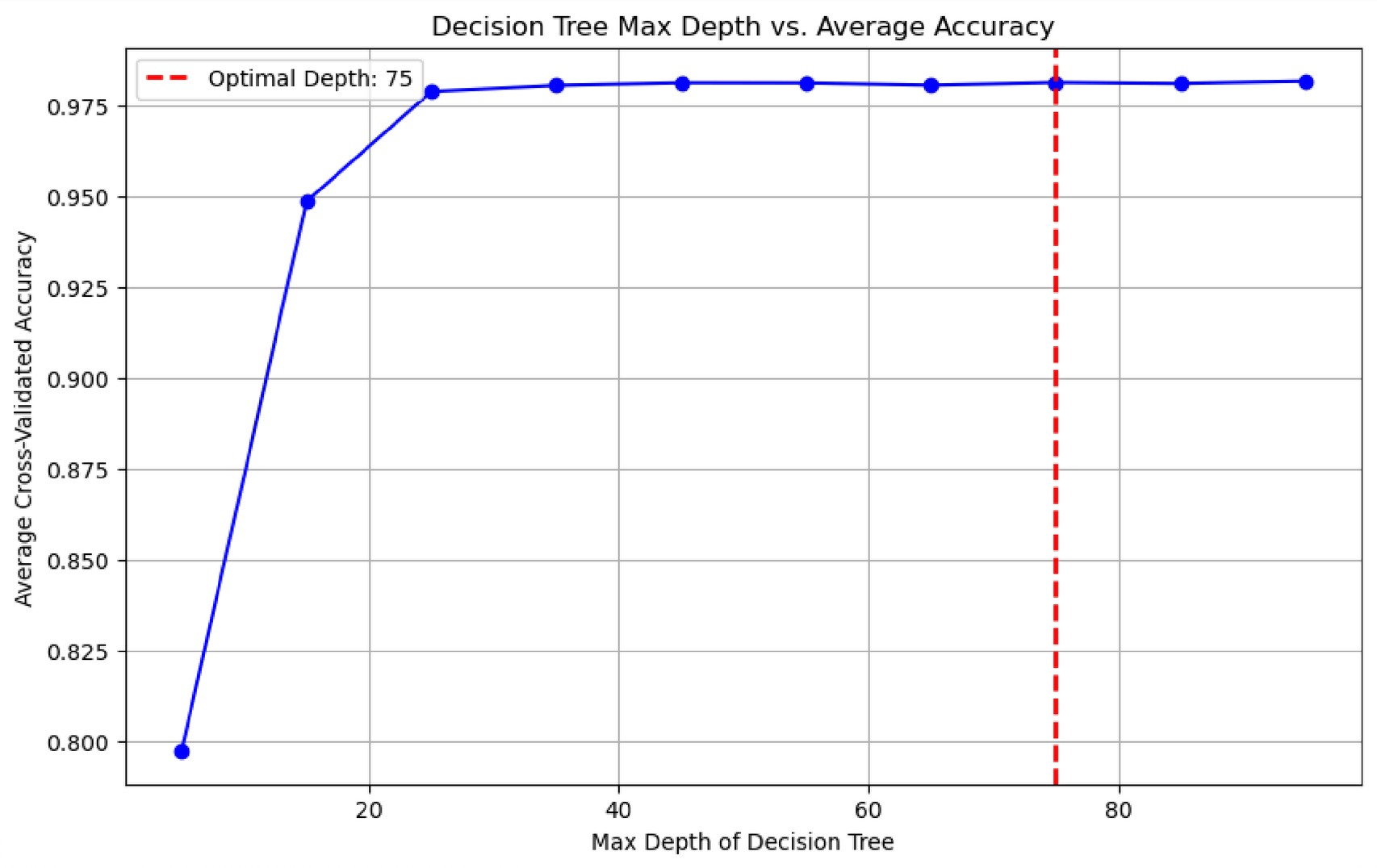
Max Depth: 75

Min Sample Split: 2

Min Sample Leaf: 1

Max Features: sqrt

Decision Tree - Max Depth



Decision Tree Models

01

Base Model - Full PI

02

No PI w/ Jobs

03

No PI w/ Lat & Long

04

No PI w/ Year Born

05

No PI w/ Gender

06

No PI

MODEL EVALUATION

CURRENT FINDINGS

Model Accuracy

- All Models have Similar Accuracies
  - **Conclusion:** PI is not Needed*
  - Year Born has the Highest Accuracy
- **Time** does decrease
- **Imbalanced Dataset** (testing data) results in Poor Metrics

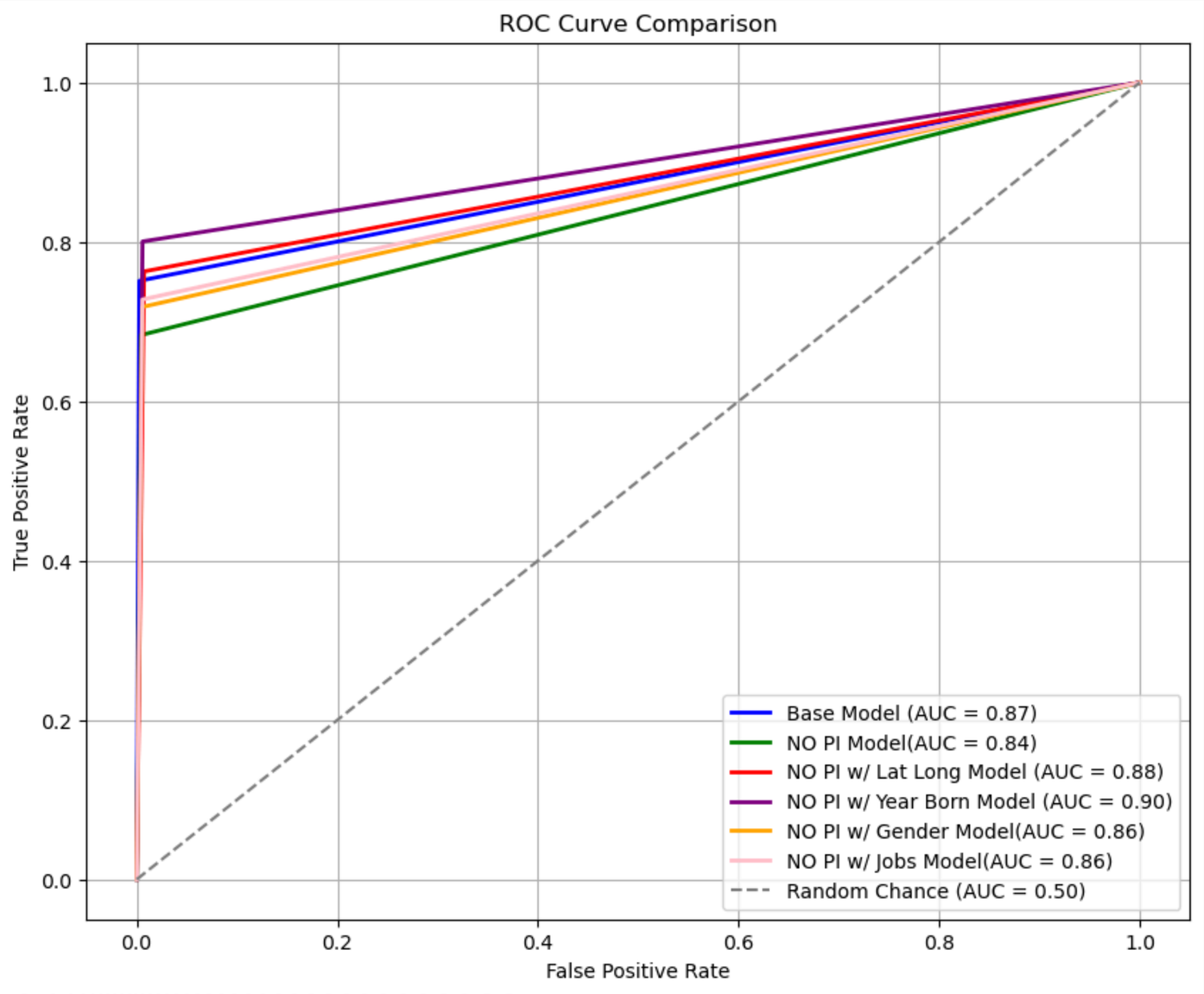
Feature Importance

- **‘amt’ feature -** large importance compared to other features
  - Others (including PI) are almost negligible.

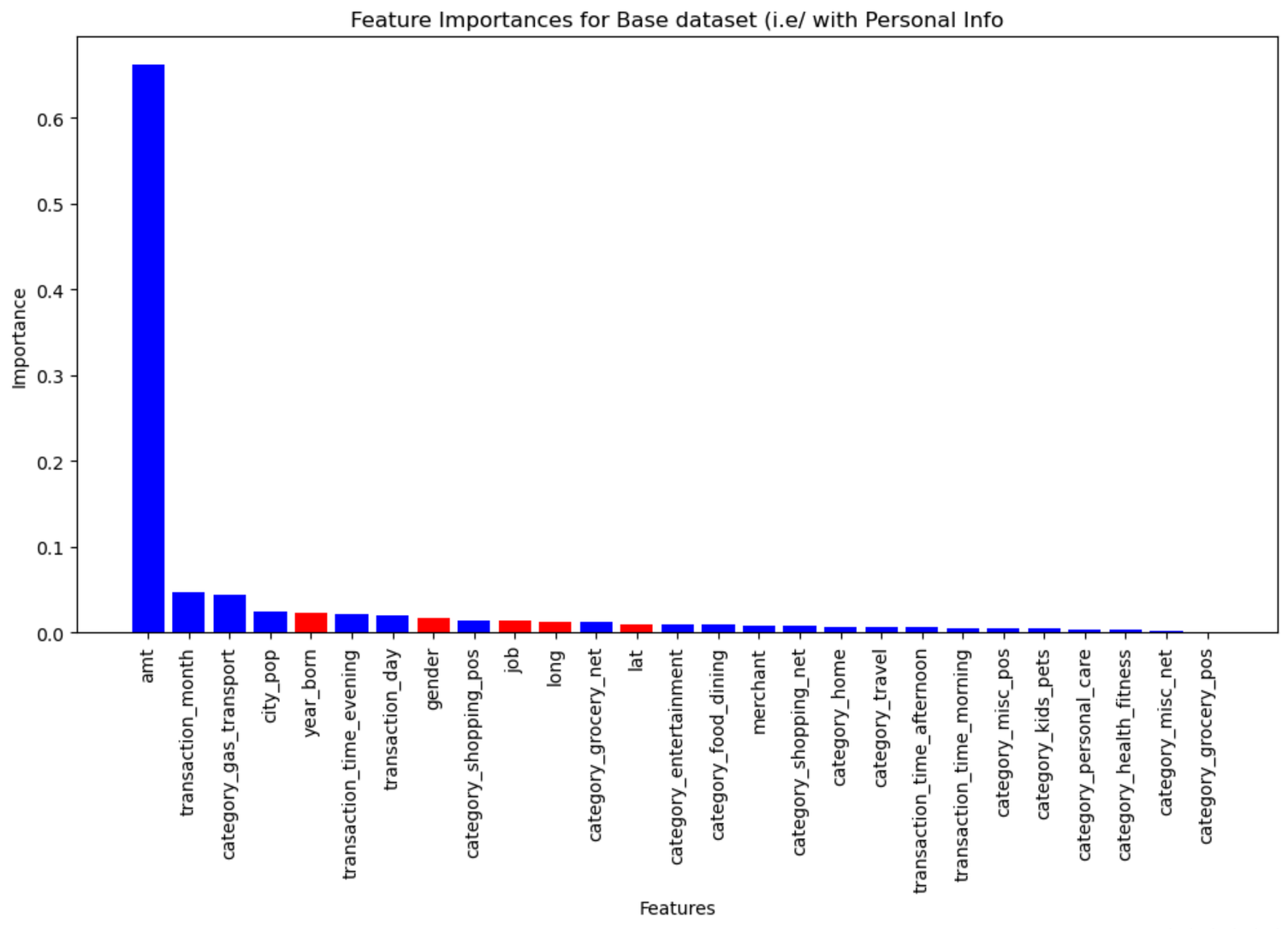
RESULTS

	Base - Full PI	No PI w/ Lat Long	No PI w/ Year Born	No PI w/ Gender	No PI w/ Jobs	No PI
AUC	0.87429	0.87780	0.89734	0.85645	0.86105	0.83934
AUPRC	0.39631	0.21956	0.27929	0.24369	0.24361	0.25606
F1	0.61896	0.41667	0.48506	0.45946	0.45723	0.48259
Accuracy	0.99646	0.99181	0.99349	0.99352	0.99338	0.99439
Time	2.64093	1.82571	1.49502	1.67914	1.50546	1.08064

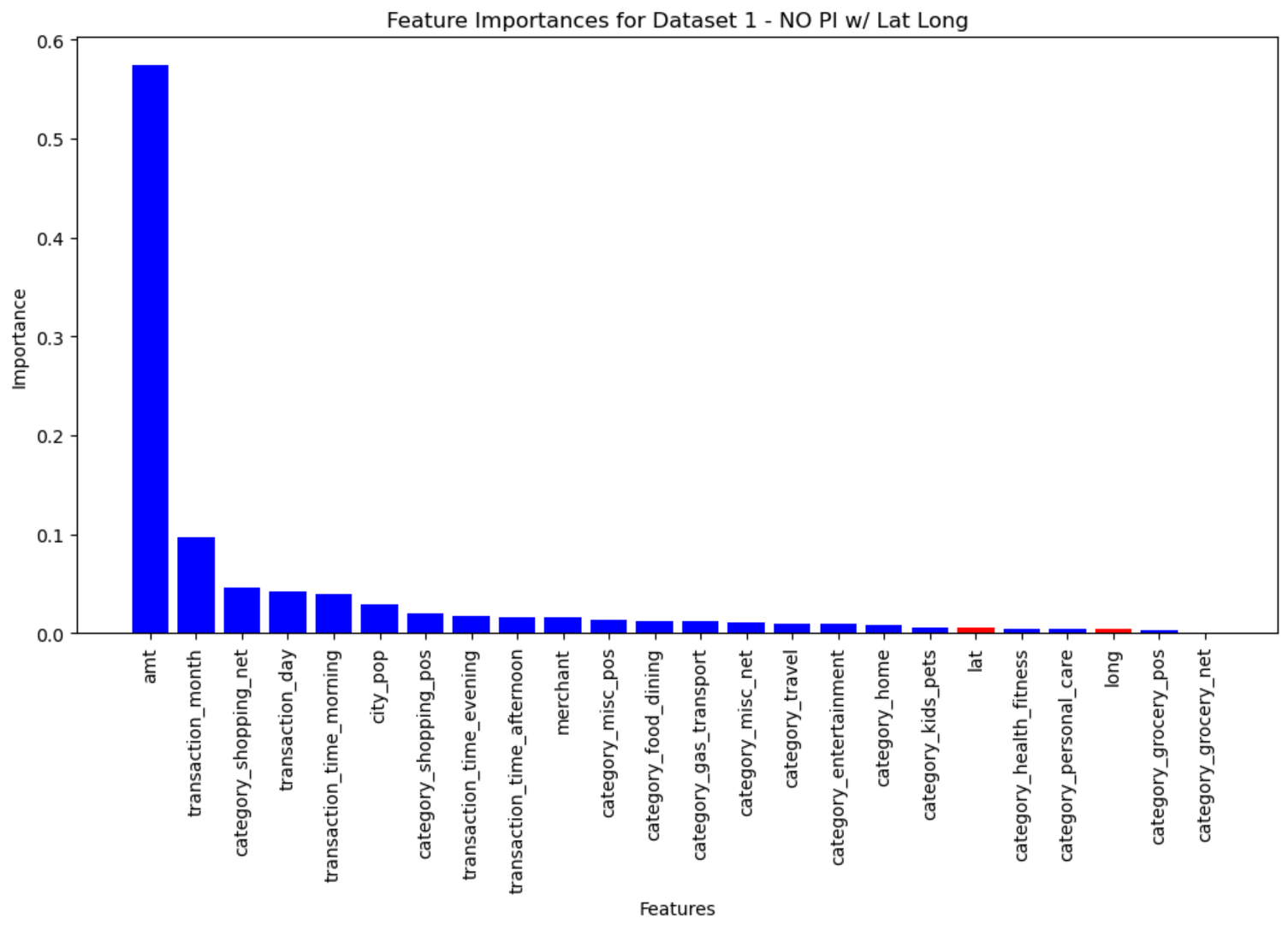
ROC Curve


ROC Curve Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Base Model (AUC = 0.87)
NO PI Model(AUC = 0.84)
NO PI w/ Lat Long Model (AUC = 0.88)
NO PI w/ Year Born Model (AUC = 0.90)
NO PI w/ Gender Model(AUC = 0.86)
NO PI w/ Jobs Model(AUC = 0.86)
Random Chance (AUC = 0.50)

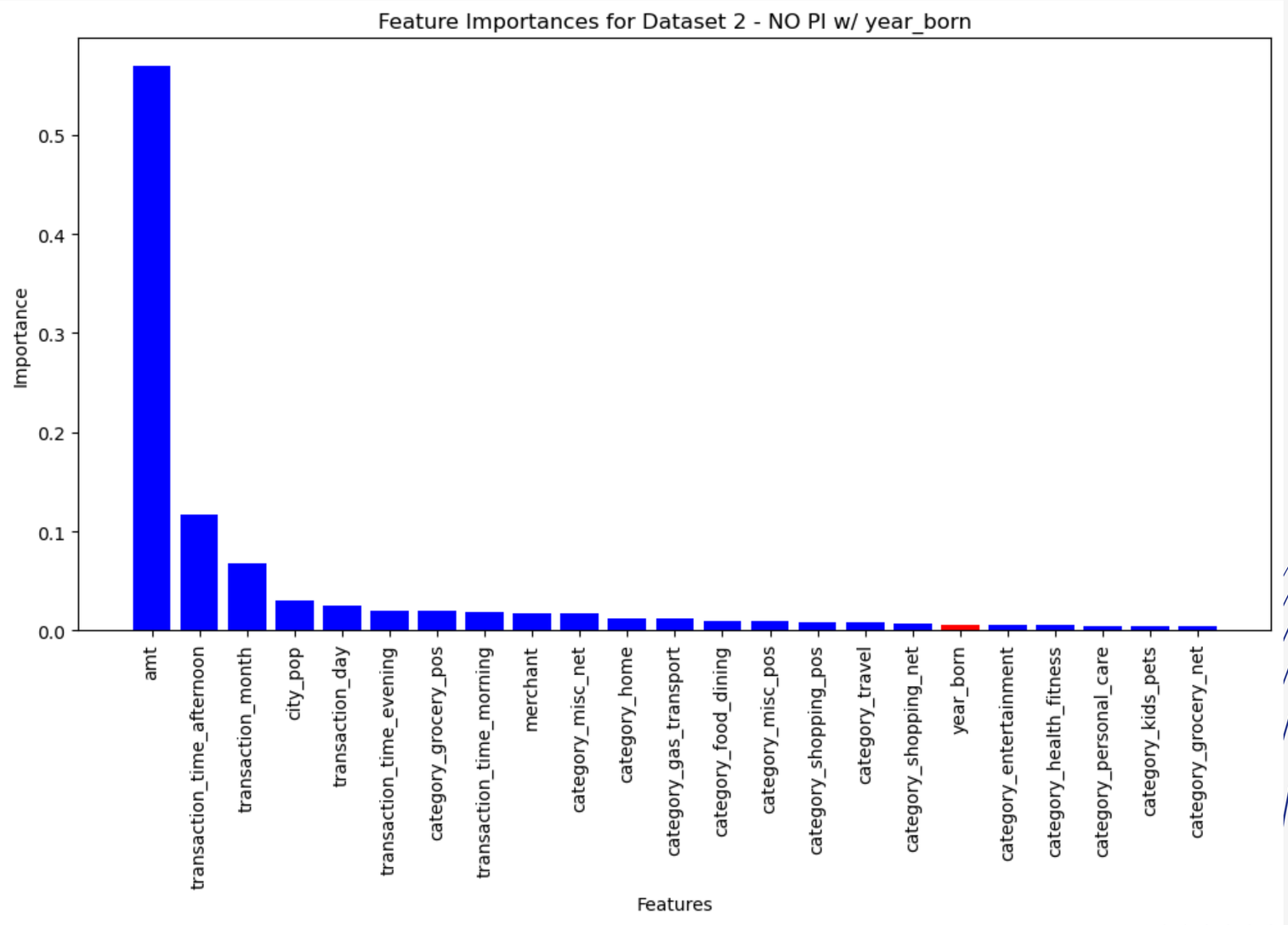
Feature Importance
Base Model (Full PI)



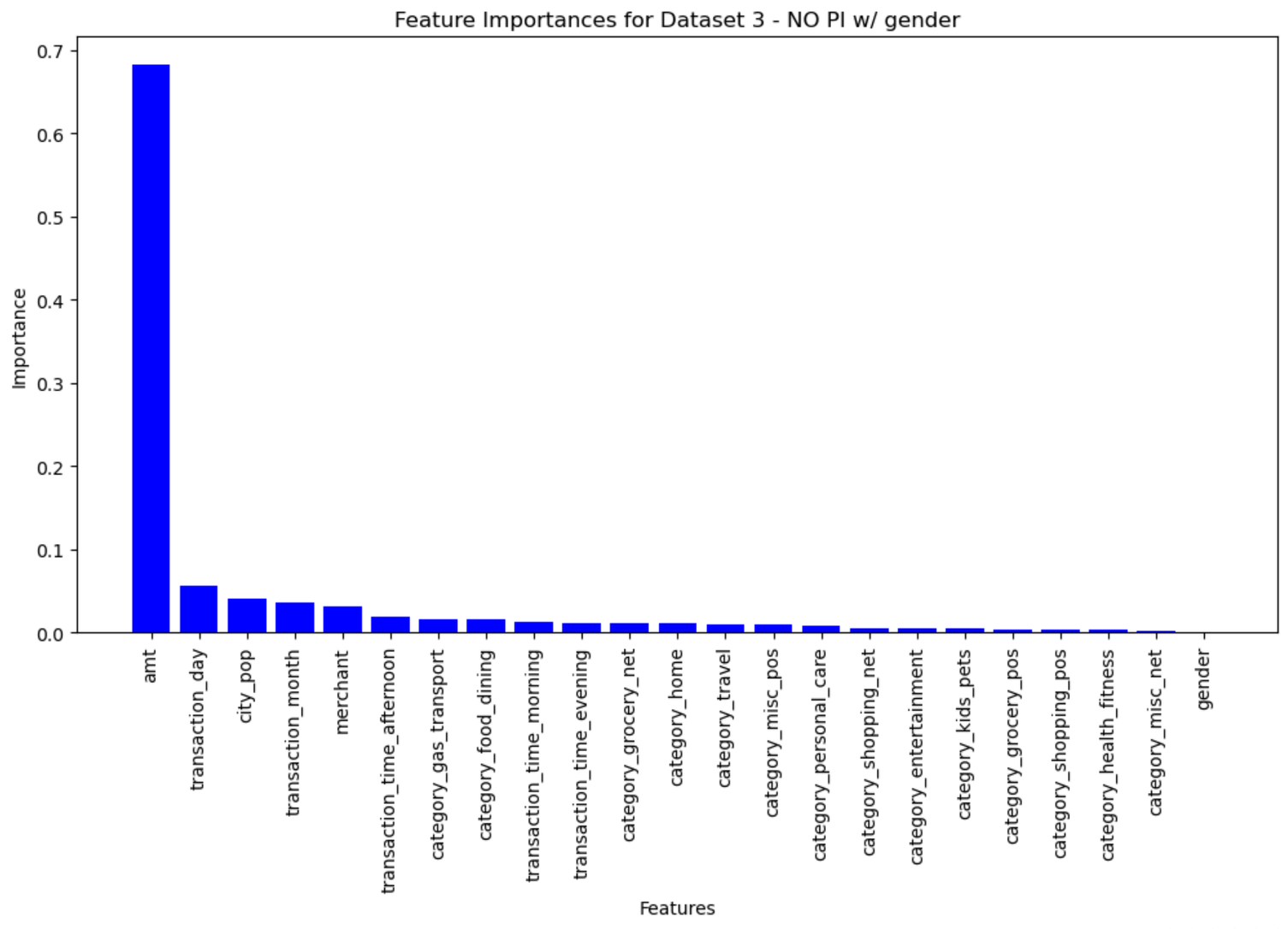
Feature Importance
No PI w/ Lat & Long



Feature Importance
No PI w/ Year Born

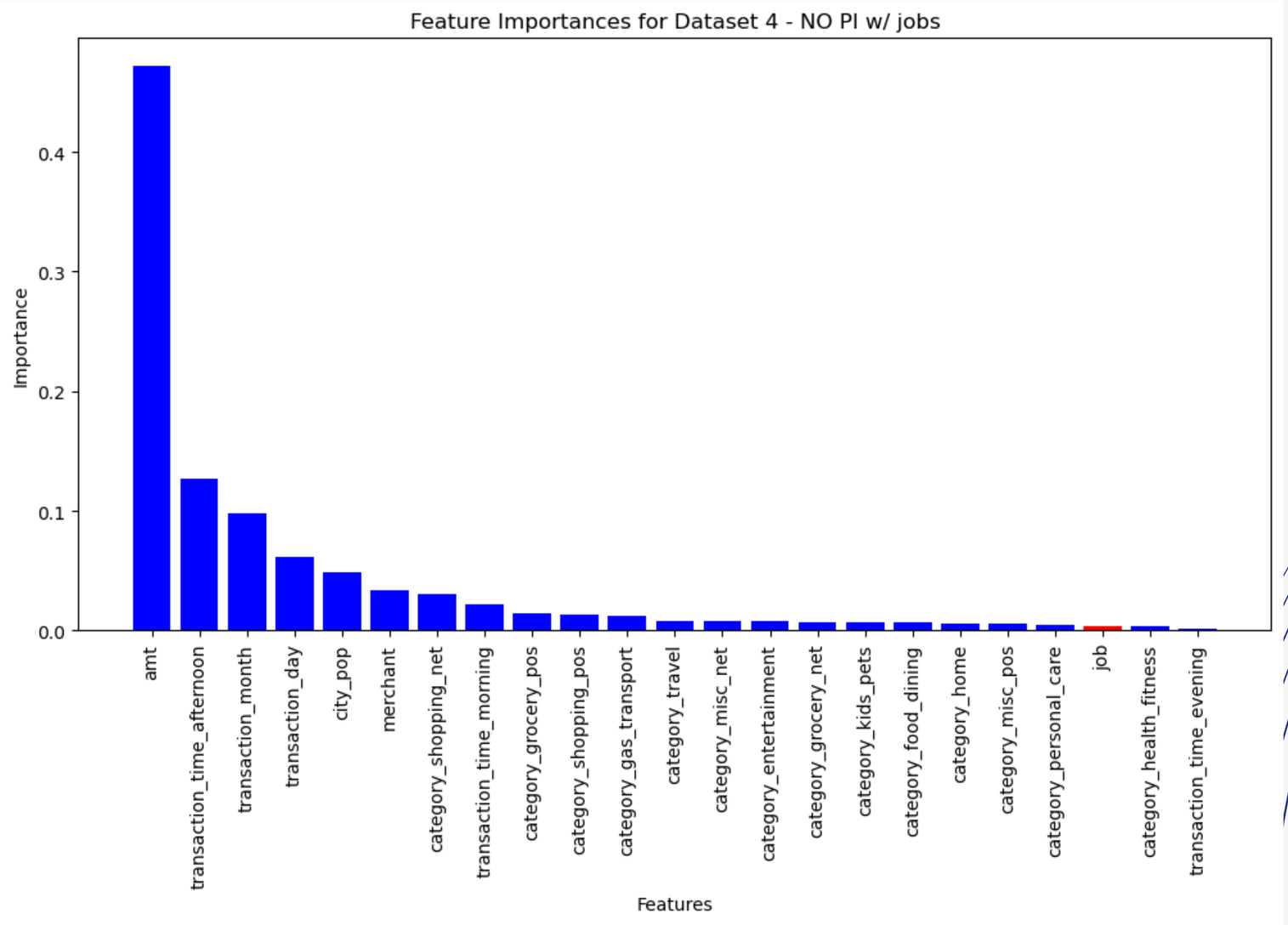


Feature Importance
No PI w/ Gender



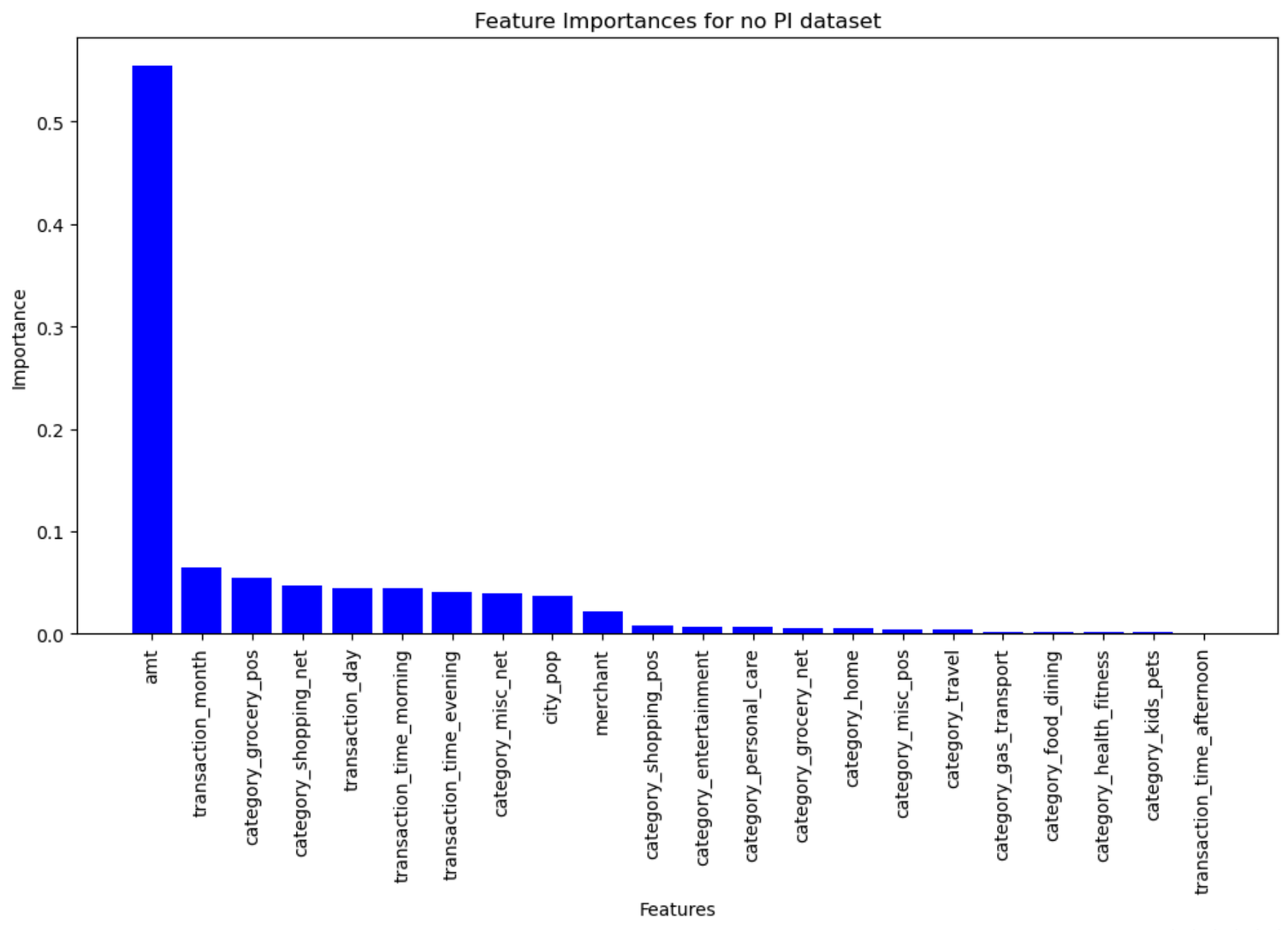
Feature Importance

No PI w/ Jobs



Feature Importance

No PI



FUTURE STEPS

More Models

- Try using different models like KNN and Neural Networks to see if our hypothesis still stands and if we get improved accuracy

More Parameters

- Use More Values for the ‘param_grid’ in GridSearchCV
- Test Model Specific Parameters

More PI Features

- Use Word Vectorization to extract more personal information
  - Ex: **‘first’** and **‘last’** names can be vectorized to give more context to the cardholder