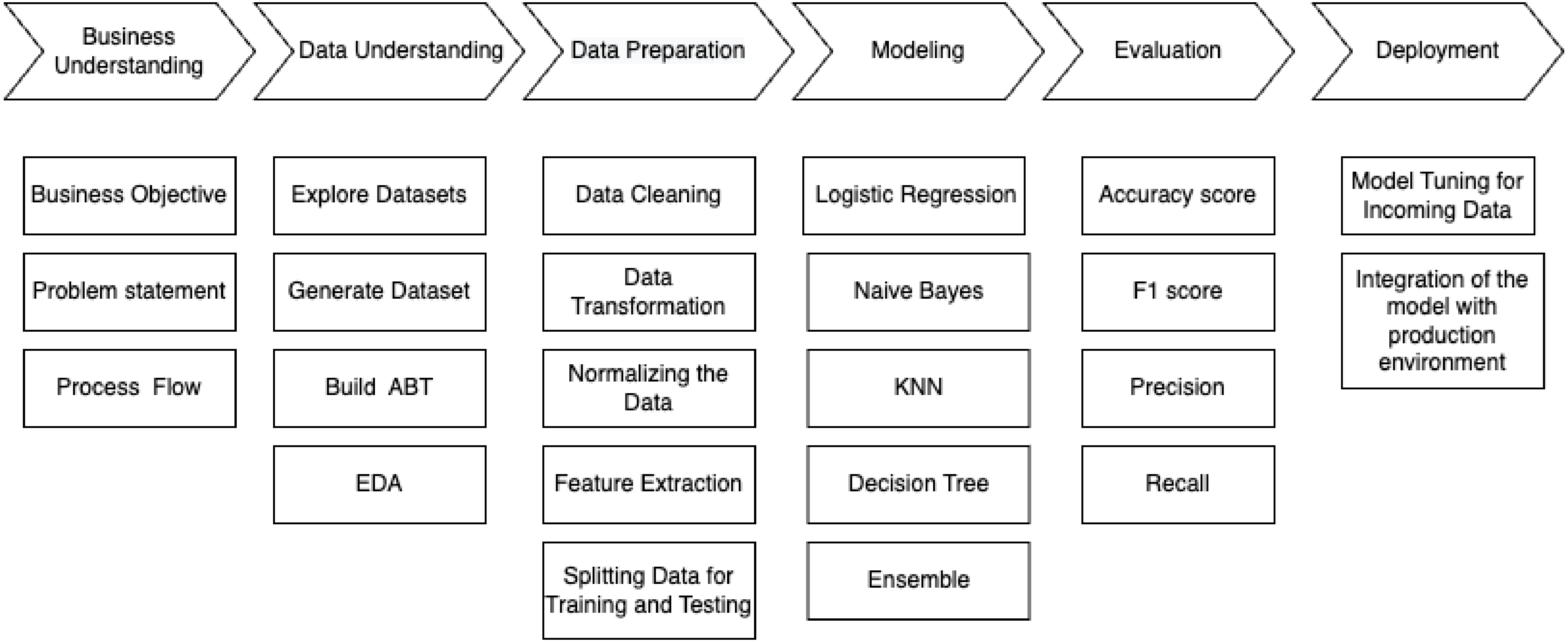
Credit Card Fraud Detection

Group 4 – Nupur, Revathi, Divya, Vani

Architecture



Business Understanding
- Business Objective
- Problem statement
- Process Flow

Data Understanding
- Explore Datasets
- Generate Dataset
- Build ABT
- EDA

Data Preparation
- Data Cleaning
- Data Transformation
- Normalizing the Data
- Feature Extraction
- Splitting Data for Training and Testing

Modeling
- Logistic Regression
- Naive Bayes
- KNN
- Decision Tree
- Ensemble

Evaluation
- Accuracy score
- F1 score
- Precision
- Recall

Deployment
- Model Tuning for Incoming Data
- Integration of the model with production environment

Introduction

- Advancement in Science and Technology has increased a lot in recent years. The rapid development in e-commerce, tap and pay systems, and e-payment methods resulted in a tremendous increase in financial frauds.
- Types of financial frauds : Unauthorized banking, Investment frauds, Identity thefts, Phishing, Advance fee fraud, Credit card transaction frauds and others
- In 2022, 46% of companies reported experiencing fraud, corruption and economic crimes compared to 47 % in 2020 and 49% in 2018.
- Digital fraud has impacted 38% of Americans in Q1 2022, 29% of people experienced phishing, followed by 26% for stolen credit cards.
- As fraudsters are increasing every day, it is important to identify such frauds and take necessary precautions to avoid them.

Objective

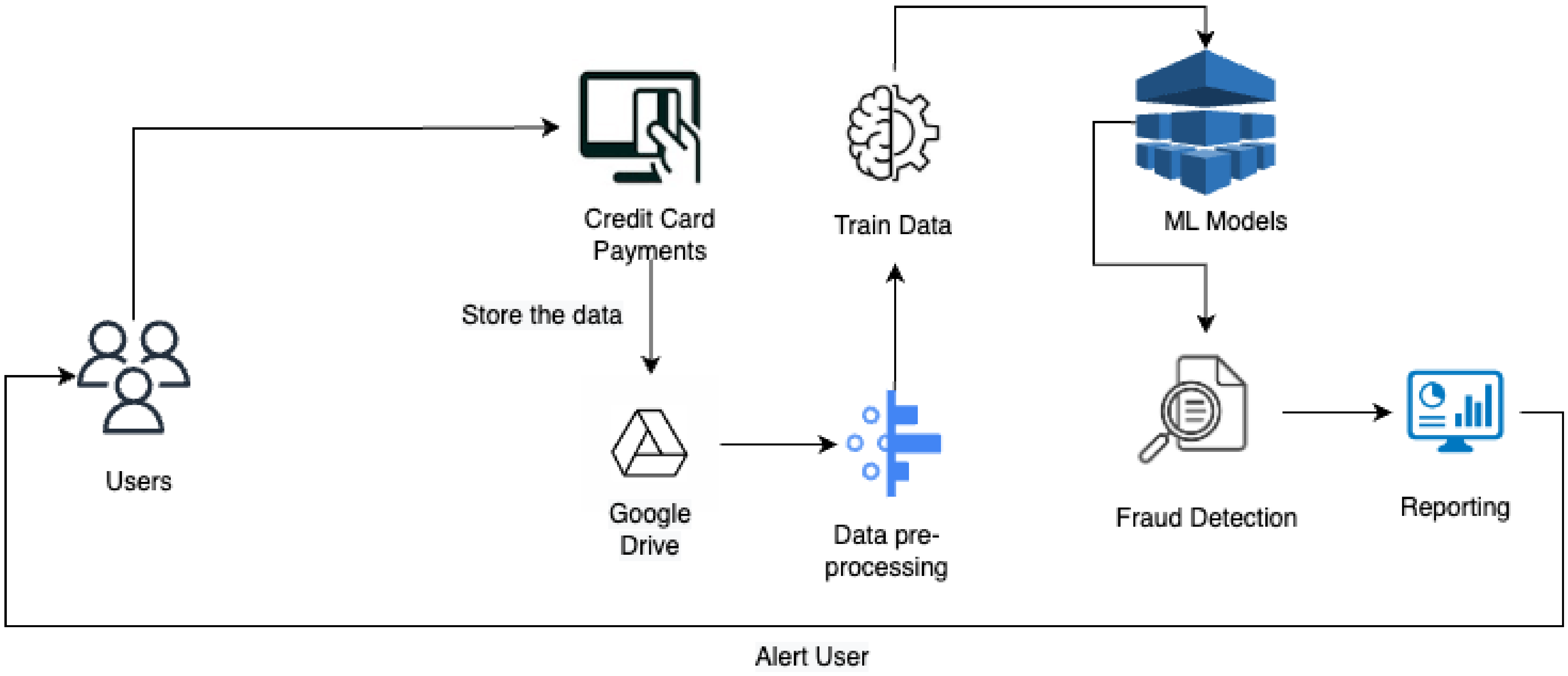
- In this project, our objective is to focus on credit card transaction frauds and discuss how such frauds can be handled by employing various Machine Learning algorithms.
- Credit card usage has increased tremendously over the decade which revolutionized the cashless payment methods, but it comes with its own set of risks.
- To tackle this issue, we plan to employ Machine Learning models like Logistic Regression, Naive Bayes, KNN, Decision Tree, Ensemble and compare the performance of each model to locate the best fitting model which helps in reducing the frauds and upgrading the system.

Literature Survey

Research Paper	Business Objective	Models Used	Performance Evaluation
Model for Credit Card Fraud Detection using Machine Learning Algorithm	Incorporated previous transactions details and identified the fraud transactions by analyzing inconsistent location calculations for every transaction	SVM, Logistic regression, KNN and Random Forest.	Evaluated the performance of the models by comparing **Accuracy score, F1 score** and **confusion matrix** for different models and identified the best fitting model.
Credit Card Fraud Detection: A Case Study	Collected information like registration details, login details, banking details, and others during the online transactions for detection fraud	Genetic Algorithm, Behavior Based Technique and Hidden Markov Models	Deployed all the models individually and then by taking the average of the values obtained from these three models, if the value is above the threshold value, they have defined that fraud has occurred.

https://ieeexplore.ieee.org/document/9673381
https://ieeexplore.ieee.org/document/7100189

Process Flow



Data Understanding

Data Collection

- Synthetic data for credit card transactions for 1,000 customers using Sparkov_Data_Generation-master
- Transaction duration: January 1, 2020, to December 31, 2021
- Csv file for each customer will all transactions for the given customer

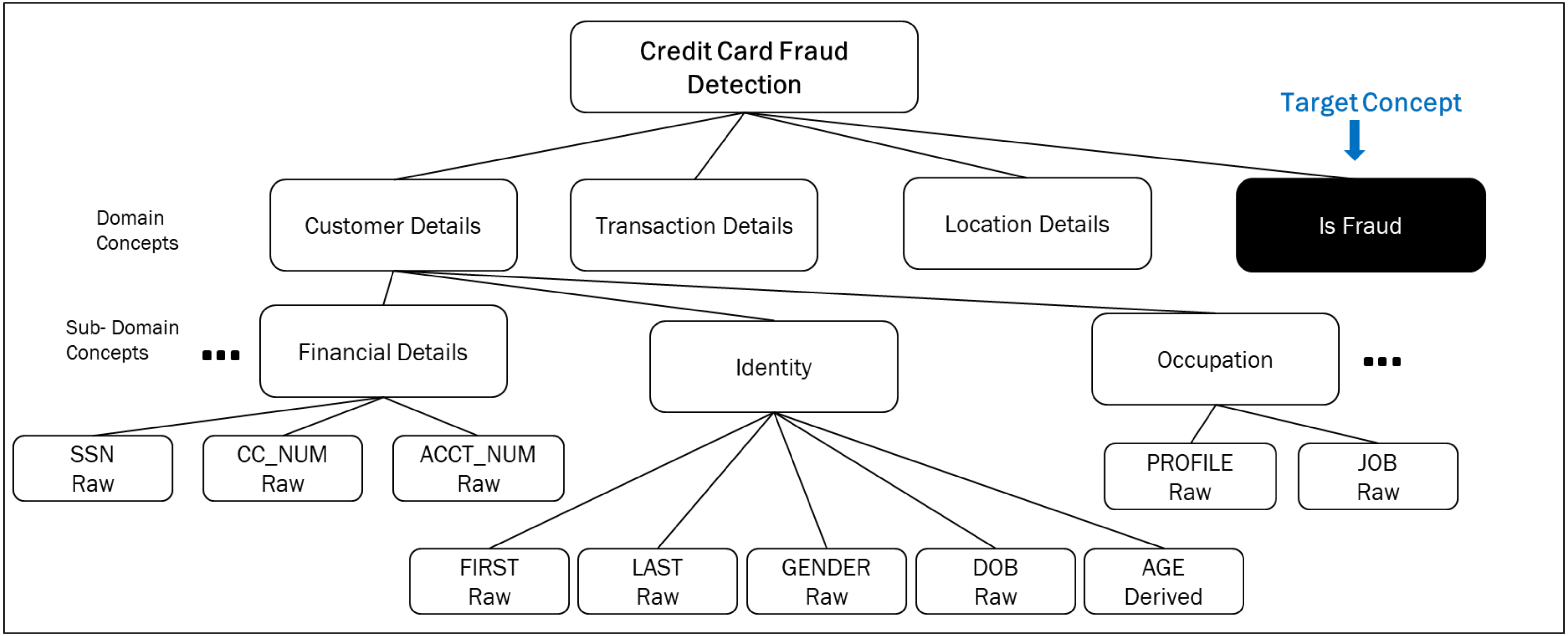
Data Import and ABT

- Union of all the transactions in the csv files and load in the dataframe for analysis
- Features:
  - Raw Features (25)
  - Derived Features (10)
- Data Exploration

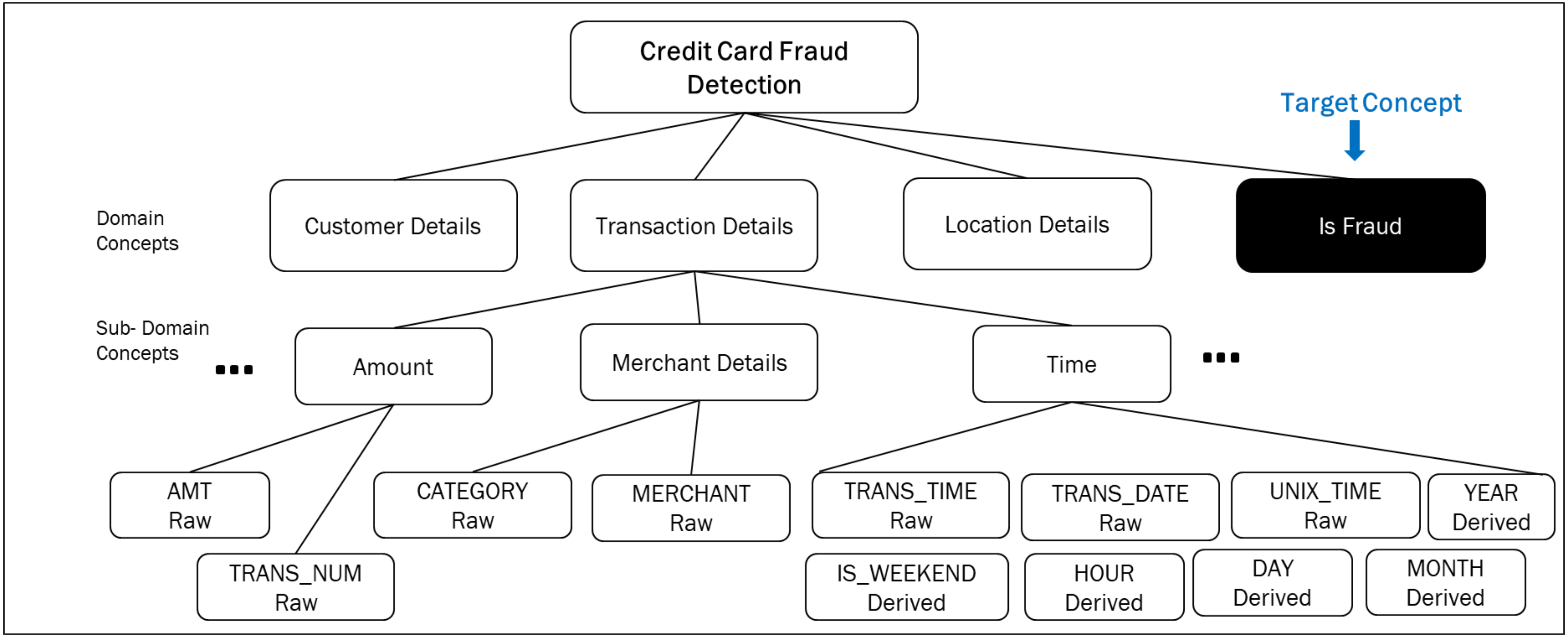
Exploratory Data Analysis

- Generate Data Quality Report for Categorical and Continuous features
- Histogram/ Bar plot for the feature distribution
- Correlation between continuous features

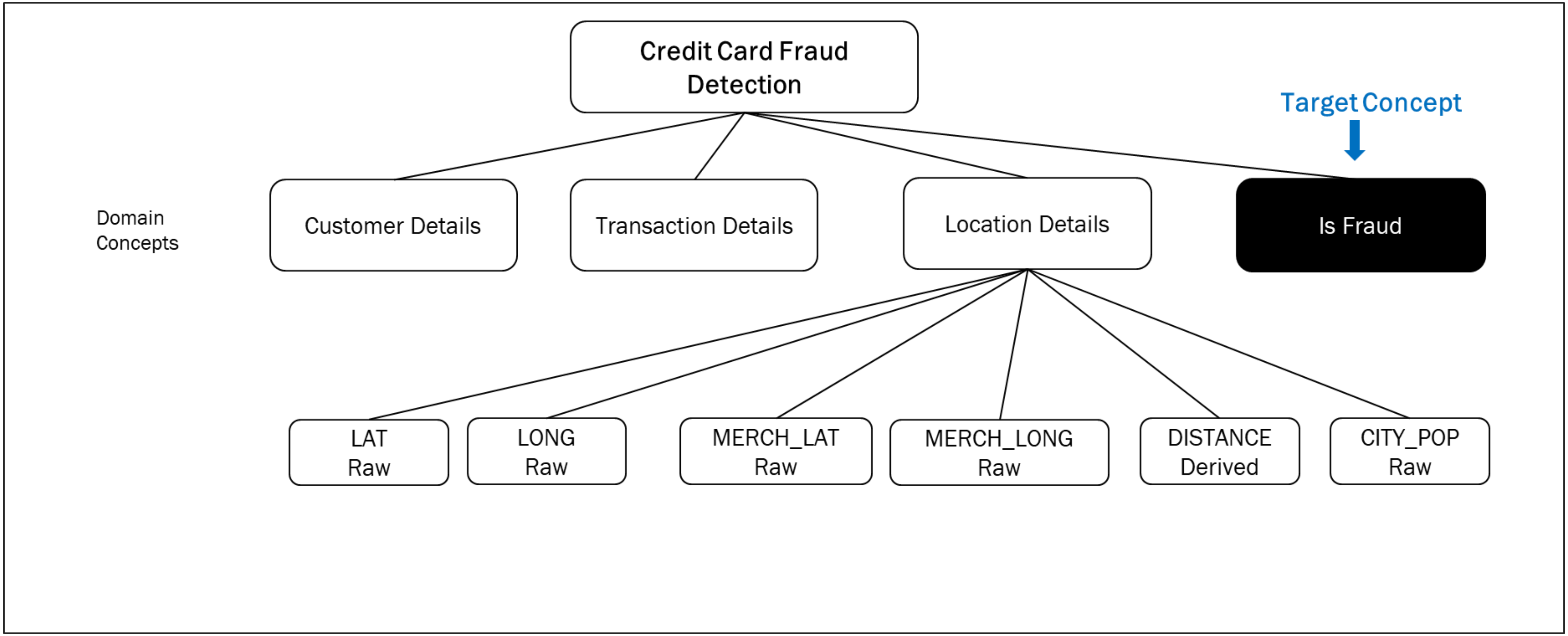
Domain Concepts



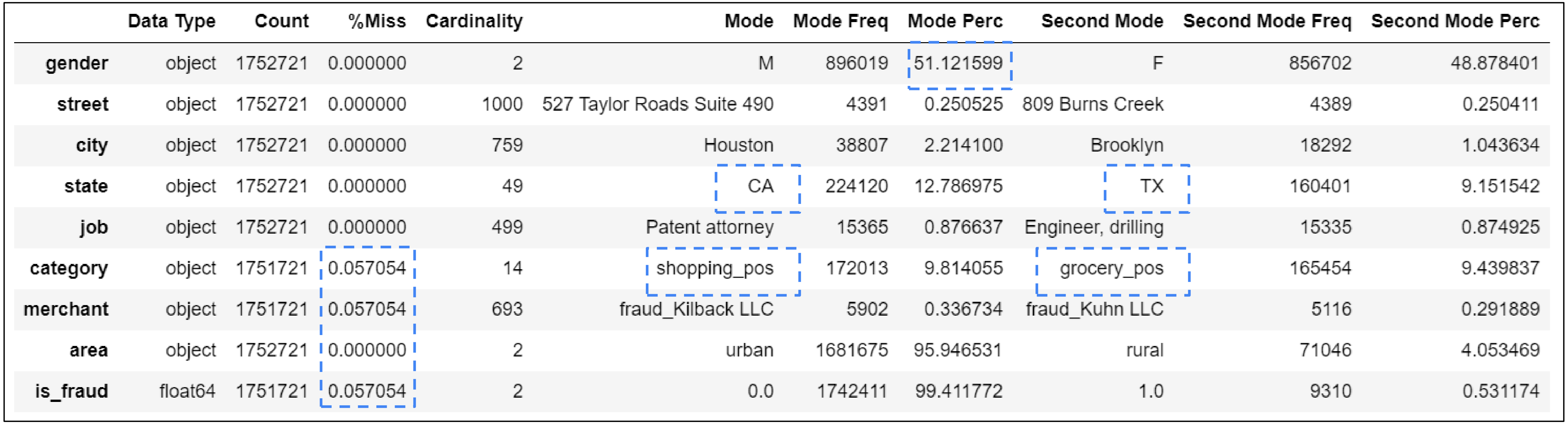
Domain Concepts



Domain Concepts



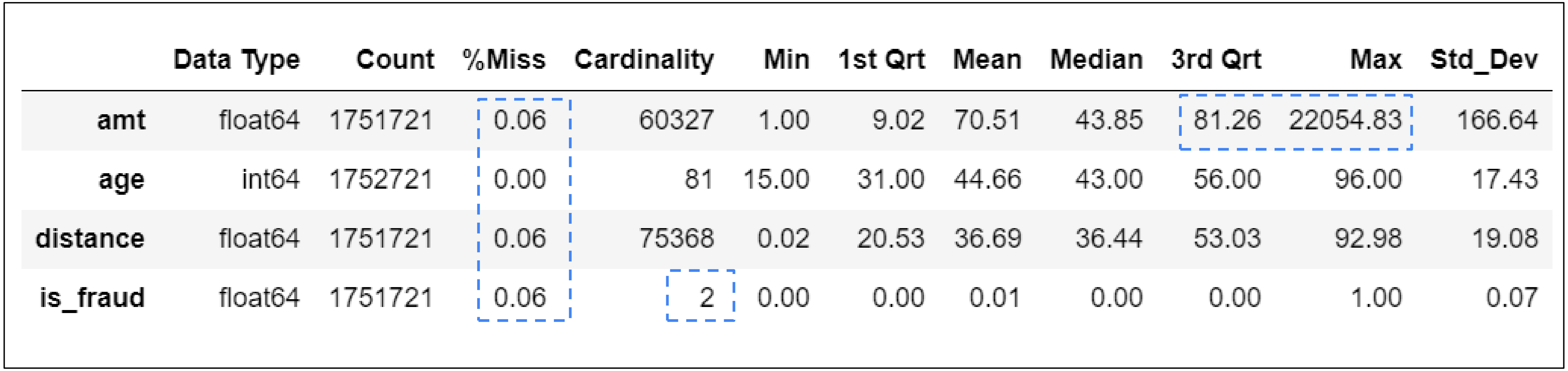
Data Quality Report – Categorical Features



	Data Type	Count	%Miss	Cardinality	Mode	Mode Freq	Mode Perc	Second Mode	Second Mode Freq	Second Mode Perc
gender	object	1752721	0.000000	2	M	896019	51.121599	F	856702	48.878401
street	object	1752721	0.000000	1000	527 Taylor Roads Suite 490	4391	0.250525	809 Burns Creek	4389	0.250411
city	object	1752721	0.000000	759	Houston	38807	2.214100	Brooklyn	18292	1.043634
state	object	1752721	0.000000	49	CA	224120	12.786975	TX	160401	9.151542
job	object	1752721	0.000000	499	Patent attorney	15365	0.876637	Engineer, drilling	15335	0.874925
category	object	1751721	0.057054	14	shopping_pos	172013	9.814055	grocery_pos	165454	9.439837
merchant	object	1751721	0.057054	693	fraud_Kilback LLC	5902	0.336734	fraud_Kuhn LLC	5116	0.291889
area	object	1752721	0.000000	2	urban	1681675	95.946531	rural	71046	4.053469
is_fraud	float64	1751721	0.057054	2	0.0	1742411	99.411772	1.0	9310	0.531174

- Gender: Almost equal proportions of the gender for the credit card transactions
- State: CA and TX are the top two states with high number of credit card transactions
- Category: Most of the transactions are processed at shopping_pos and grocery_pos terminals
- % Miss: There are not many missing values for all the continuous features in the dataset

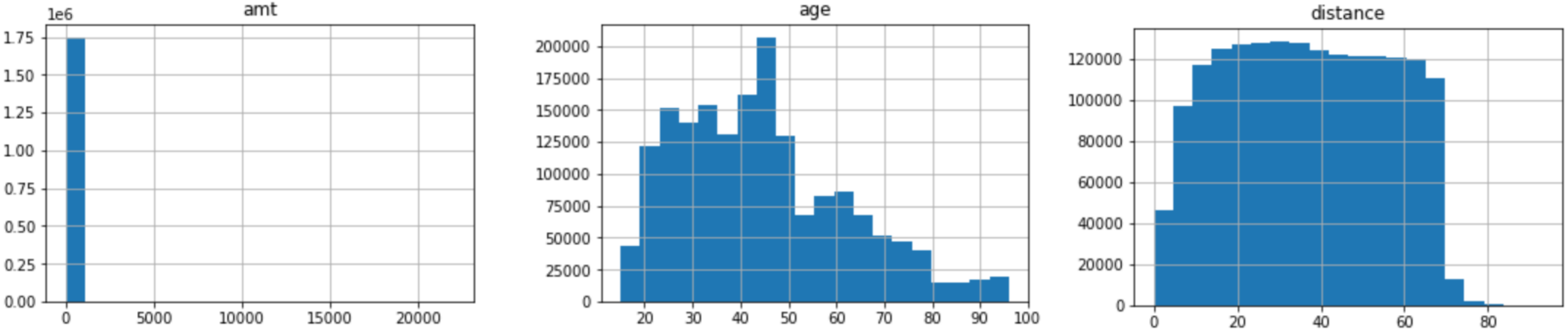
Data Quality Report – Continuous Features

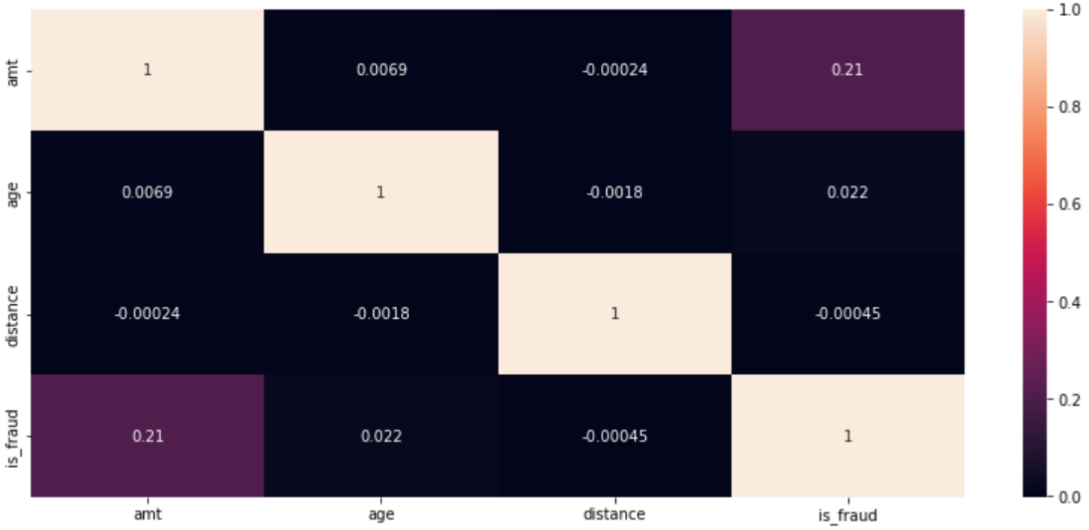


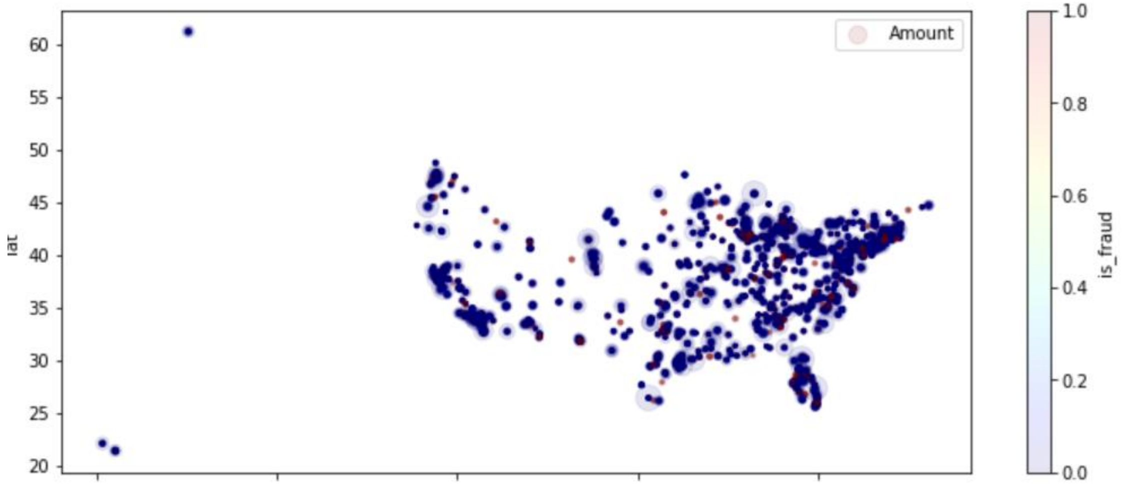
	Data Type	Count	%Miss	Cardinality	Min	1st Qrt	Mean	Median	3rd Qrt	Max	Std_Dev
amt	float64	1751721	0.06	60327	1.00	9.02	70.51	43.85	81.26	22054.83	166.64
age	int64	1752721	0.00	81	15.00	31.00	44.66	43.00	56.00	96.00	17.43
distance	float64	1751721	0.06	75368	0.02	20.53	36.69	36.44	53.03	92.98	19.08
is_fraud	float64	1751721	0.06	2	0.00	0.00	0.01	0.00	0.00	1.00	0.07

- Amt: There is a huge difference between the 3rd Qrt and Max amount value
- Is_fraud: The cardinality for the feature is 2, therefore this target feature can be considered categorical
- % Miss: There are not many missing values for all the continuous features in the dataset. As we are applying supervised machine learning algorithms, we will consider the instances where the target feature is populated

Histogram, Correlation Matrix, and Distribution of Transactions







Data Preparation

1. Data Cleaning

- Handle missing values
- Check for duplicates

Data Preparation

2. Data Transformation

- Data Selection
  - ➢ Drop the features that are not required in the modelling

- Data Attributes Decomposition / Composition
  - ➢ Transformation for the categorical fields:

    One-hot encoding of gender, area and category features

Data Preparation

3. Data Scaling

- Range normalizing the data

4. Data Splitting

- Split the dataset into test and train sets

Model Development

Models Shortlisted

Logistic Regression

Naïve Bayes

Decision Tree

KNN

Ensemble

Modulations

Resampling

Dimensionality Reduction

Hyperparameter Tuning

Model Comparison

S.Nr.	Model	Normalization	Collinearity Impact	Outlier Impact	Considerations
1	Logistic Regression	Yes	Yes	Yes	• Highly descriptive • Reasonable computational requirements
2	Naïve Bayes	N/A	Yes	Yes	• Suitable for small training data • Ignores interdependencies between attributes
3	Decision Tree	N/A	No	Yes	• Computationally heavy
4	KNN	Yes	Yes	Yes	• Lazy learner • Sensitive to curse of dimensionality
5	Ensemble	N/A	No	Yes	• Better prediction results • Limited explainability

References

https://www.analyticsvidhya.com/blog/2020/11/popular-classification-models-for-machine-learning/

https://www.researchgate.net/figure/A-comparison-between-the-various-classification-techniques_tbl1_292604633

https://www.templateswise.com/machine-learning-powerpoint-template/