Problem Statement

Situation

- **Our client is an IT company which provides backend systems to help retailers manage their inventories well**
- **They are using external assistance to detect fraudulent credit card transactions**
- **The model right now has many caveats:**
    - **In many instances, genuine payments are also declined**
    - **It classifies many good transactions as fraudulent**
    - **It does 'too good of a job'**
- **The client wants a new algorithm that can prevent good transactions being classified as fraudulent**
    - **They also want the new algorithm to be easier to explain**

Tasks

1. Calculate basic stats descriptive statistics (mean, median, min, max, standard deviation) for each field
2. Visualize distributions of data elements using histograms for key variables and predict which variables you expect to be most correlated with default/churn.
3. Insert (or use code provided) appropriate code to evaluate the credit card fraud detection performance of Decision Trees and Random Forest
4. Display the output visually using charts of your choosing and explain your choice. (ROC Curve, Confusion Matrix, Gains Table)
5. In addition to the spreadsheet/code or programming output you submit, include a separate written document of 250-500 words that summarizes:

    a) your interpretation of the data via the Exploratory data analysis

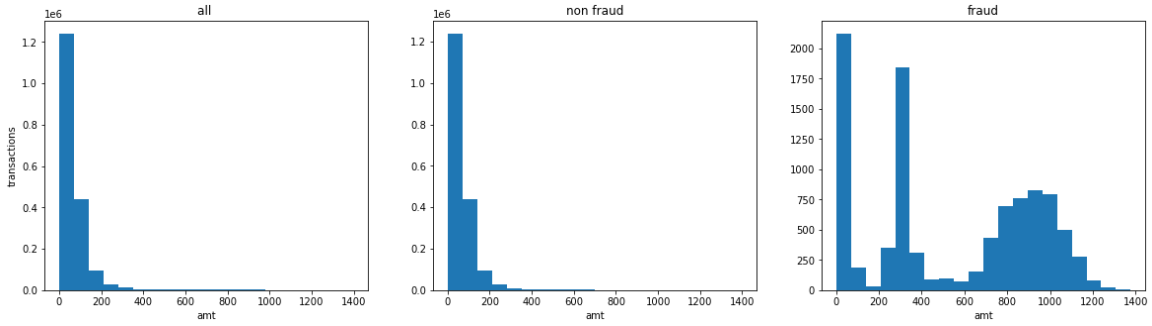
    b) your interpretation of modeling output of Logistic Regression, Decision Tree and Random Forest

    c) discuss which variables are most significant and any other unexpected insights.

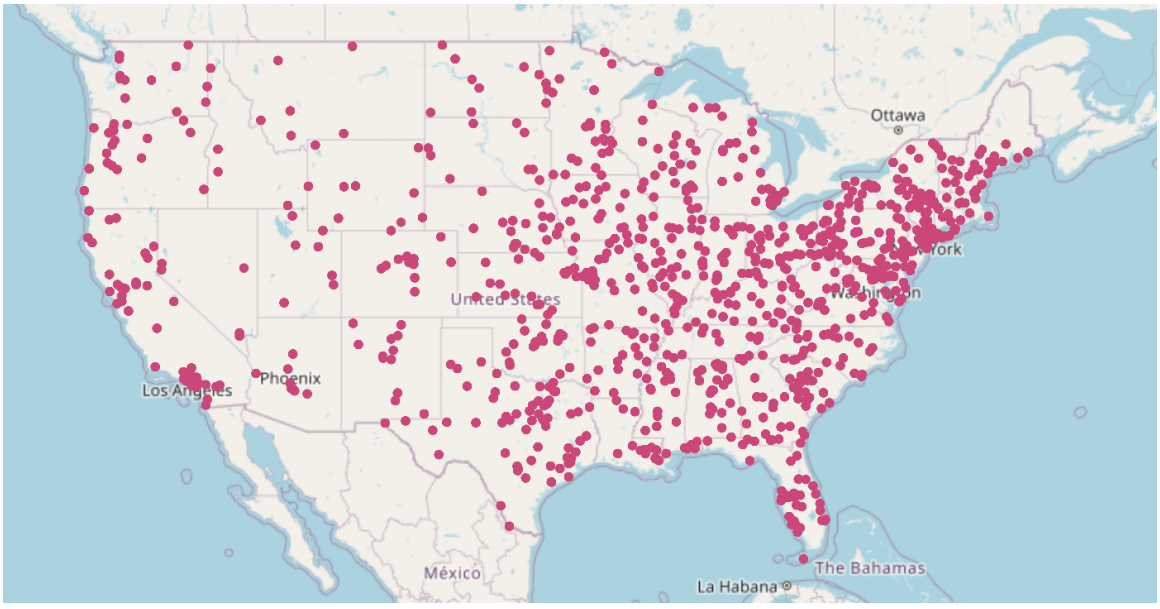
Exploratory Data Analysis

The compiled dataset comprises of 22 columns, and 1,852,394 observations. Each row is uniquely identified by a transaction date and time column. Each transaction is classified as fraudulent or non-fraudulent using the 'is_fraud' column (0 – normal, 1 – fraud).

- **Amount variable:**
  Most non-fraudulent transactions hover around 200-300 USD.
  However, we see a bigger spread for fraudulent transactions, with amounts going upto 1400 dollars as well.



- **Transaction Location**
  As visible from the transaction's geo map, there is a relatively higher concentration of fraudulent transactions in the eastern coast (especially near the New England belt)



- **Feature Engineering**
  60 days aggregated amounts and transactions were calculated for each particular credit card number, and these were added to the main data set to provide additional features

Modeling

- **Correlation with is_fraud variable**
  Amount variable has close to 20% correlation with the is_fraud classifier, followed by 60 days average amount

- **Oversampling**
  When comparing the instances for fraudulent observations, we can observe that only 0.5% transactions are fraudulent. This is a sign for imbalanced datasets, and it should be rectified through oversampling
  Post oversampling, the row count increases from 1,852,394 to 3,685,486 observations.

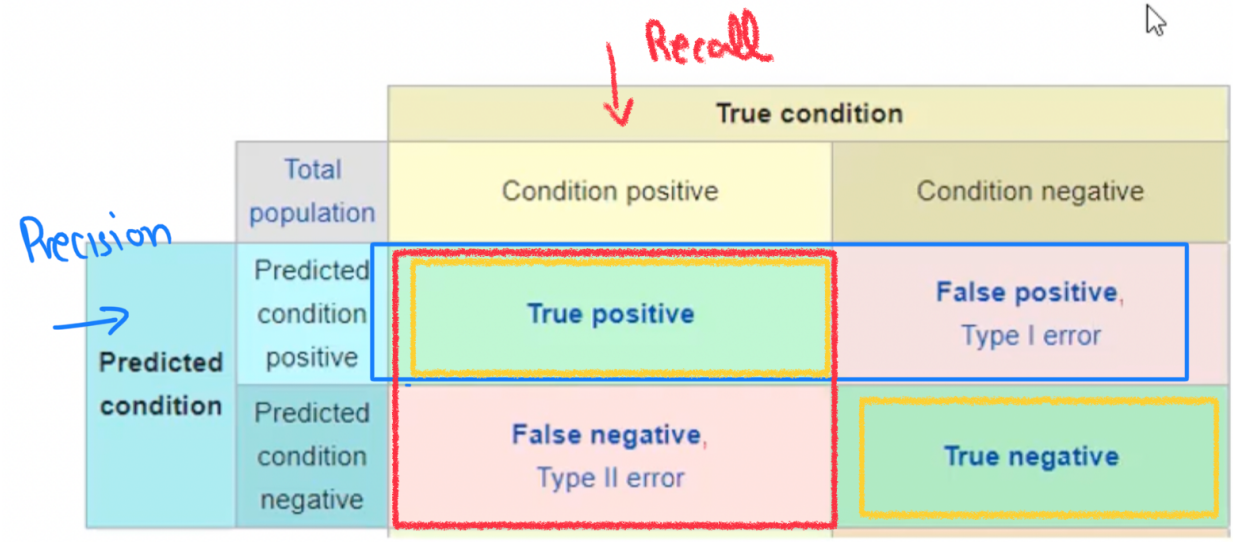
- **Train-Test Split**
  Predictors and Response variables were segregated into two different datasets.
  Both these datasets were divided into test and train datasets, with 67% train datasets and 33% test.

Model Comparisons

The current model is doing a good job for detecting fraudulent transactions, but it classifies many genuine transactions as fraudulent at the same time.

This indicates that the current model has high number of both true and false positives.
That would imply a **lower precision**, and we should try to improve this metric.

Precision of a model = True Positives / True Positives + False Positives



Our client also wants a model that can be easily explained and interpreted with various stakeholders. So very advanced models might not be a good proposition.

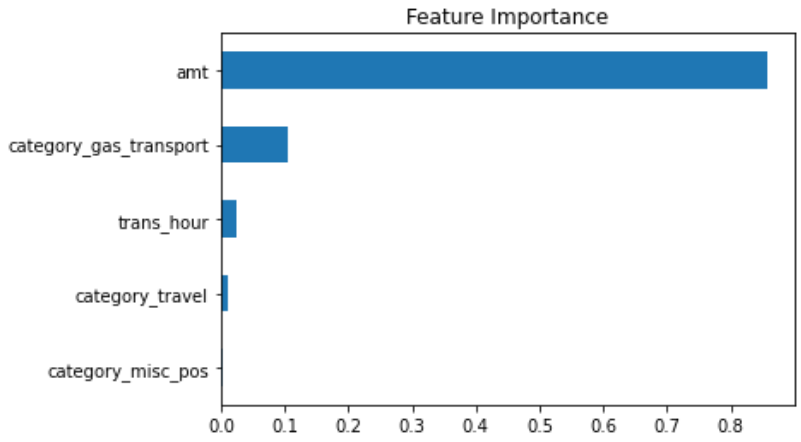
Comparing the 4 widely used models on test evaluation metrics:

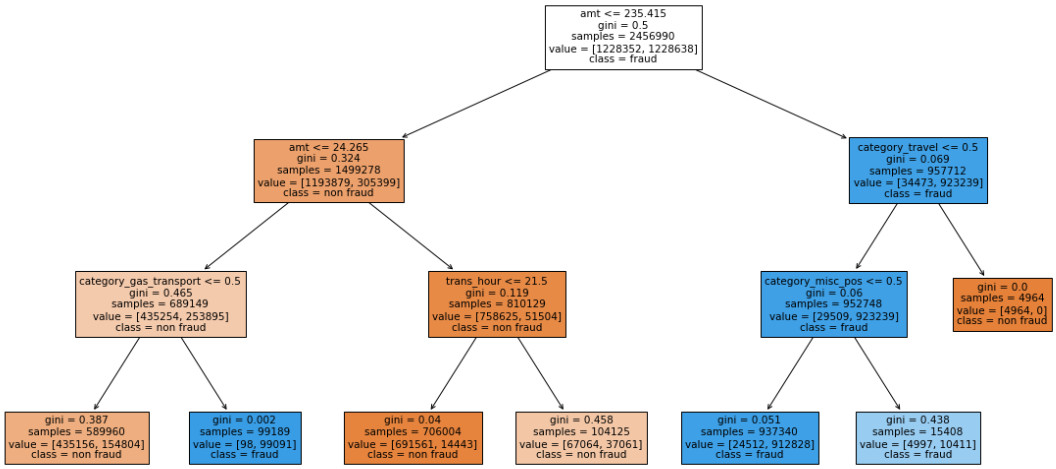
	Logistic Regression	**Decision Tree**	**Random Forest**	**XGBoost**
Test Accuracy	0.85	0.91	0.89	0.95
Test Precision (for 1)	0.89	0.97	0.94	0.95
Test Recall (for 1)	0.78	0.93	0.82	0.94
Interpretability	Medium	High	Medium	Low

From the evaluation table above, we can concur:

- All models are providing good test accuracies.
- Both precision and recall have a good range of figures for all four models
- Decision trees provide the best precision figures among all the models. This is extremely critical for us, given that our client wants to reduce miss-classifying non-fraudulent transactions
- From a communication point of view, a **decision tree** is the most interpretable by non-technical team members as well

Diving deeper into the Decision Tree Model:





- While running the model, we used a hyperparameter to fine tune our decision tree. We limited the max_depth criterion to 3 (only 3 levels of branching will be done by the model)

- The model is classifying the data based on 5 variables:
    - Amount (amt)
    - Transaction hour
    - 3 Categories of payments (included as dummy variables in the model)
        - Travel
        - Gas Transport
        - Miscellaneous
    - Transaction hour and category misc can be removed from the model without having a significant impact on the predictions

- Given the high interpretability and high precision for this model, we can strongly suggest a new prediction tool based on decision trees to our client