Credit Card Anomaly Detection Using Supervised and Unsupervised Learning Models

Shruthi Sarode

12/22/2021

Image Source: https://miro.medium.com/max/3200/0*qj2HrX376Yv2oEcz

Contents

- Intro to Anomaly Detection
- Why ML algorithms for Anomaly detection
- Supervised and Unsupervised learning
- Problem Statement
- EDA
- Supervised Learning Modeling
- Unsupervised Learning Modeling
- Key Insights
- Next Steps

What is Anomaly?

- Rare events or observations which can raise suspicions by being statistically different from the rest of the observations.
- Help point out where an error is occurring
- Intrusion detection/cyber attack
- Fraud detection
- Systems health monitoring(failing machine)
- Event detection in sensor networks
- Ecosystem disturbances, etc

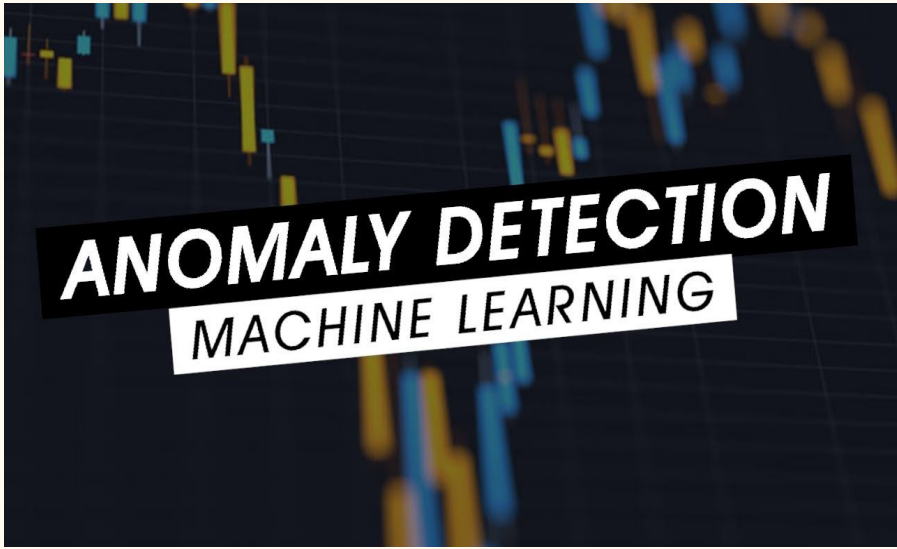
Image Source: https://unsplash.com/s/photos/different

Categories of Anomalies

- **Point anomalies** - single instance of data is anomalous.
  - Detecting credit card fraud based on “amount spent.”
- **Contextual anomalies** - abnormality is context specific.
  - Spending $100 on food every day during the holiday season is normal, but may be odd otherwise**.**
- **Collective anomalies** - set of data instances are anomalous.
  - a potential cyber attack.
- **Anomaly detection may also be similar to**
  - **Novelty detection**
  - **Outlier detection**

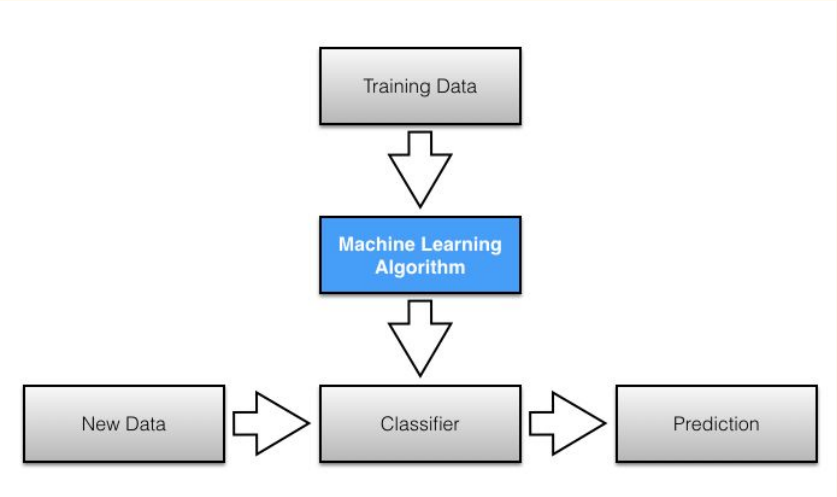
Why Use ML for Anomaly Detection

- Machine Learning (ML) is used in fraud detection because of two major reasons- Speed and Adaptability.
- ML is used for applications such as:
  - Predictions
  - Optimization
  - Detection
  - Classification



Supervised Learning

- Training data is labeled - nominal or anomaly
- Supervised learning algorithm analyzes the training data and produces an inferred function
- The algorithm will determine the class labels for unseen instances
- Popular ML algorithms for structured data:
    - Logistic Regression
    - Support vector machine learning
    - k-nearest neighbors (KNN)
    - Bayesian networks
    - Tree Based
    - Neural Networks



Unsupervised Learning

- Users do not need to supervise the model.
- Model works on its own to discover patterns and information that was previously undetected.
- Data is unlabeled - No ‘y’ variable to predict
- Unsupervised learning is classified into two categories of algorithms:
  - **Clustering**: discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
    - K-Means Clustering, DBSCAN, Hierarchical clustering
  - **Association**: discover rules that describe large portions of your data, new home buyers most likely buy new furniture
    - Apriori algorithm.

Source: https://www.guru99.com/unsupervised-machine-learning.html

Problem statement

Machine learning algorithms are being developed to detect fraudulent credit card transactions. The task is to apply Supervised and Unsupervised models on the credit card dataset and compare their performances.

Dataset (link: https://www.kaggle.com/mlg-ulb/creditcardfraud)

- Kaggle Credit Card Fraud Detection data for transactions in Sept 2013 by EU card holders.
- The dataset contains 284,807 transactions, 492 of which are fraudulent.
- The features of this dataset are already computed as a result of PCA(Except for Time and Amount). This helps us in 2 ways:
  - The confidentiality of the user data is maintained.
  - The features in the dataset are independent of each other due to PCA transformation.

Explore the Dataset

- Shape of the data - 284807 rows and 31 columns (features)
- Transaction distribution Fraud (1)= 473 and Non-Fraud (0)= 283253
- Dataset is highly imbalanced-Only 0.17% of transactions are fraudulent.
- Classic Imbalanced Dataset

CASE AMOUNT STATISTICS

NON FRAUD CASE AMOUNT STATS

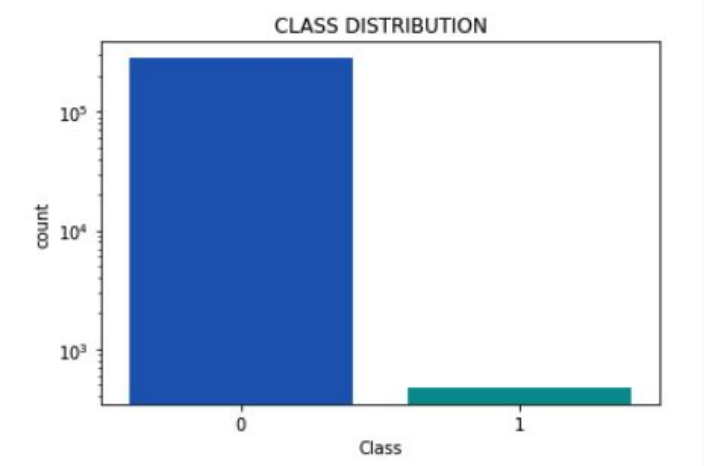
count	283253.000000
mean	88.413575
std	250.379023
min	0.000000
25%	5.670000
50%	22.000000
75%	77.460000
max	25691.160000

Name: Amount, dtype: float64

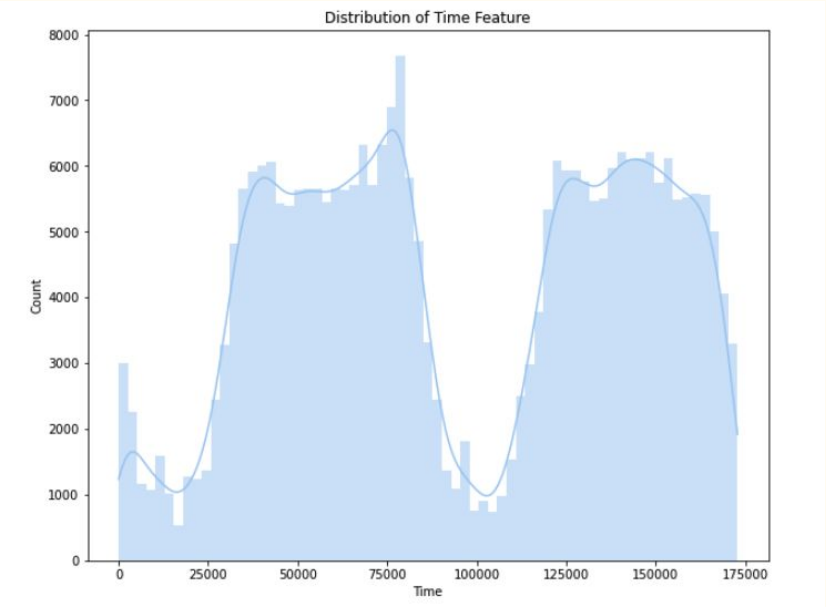
FRAUD CASE AMOUNT STATS

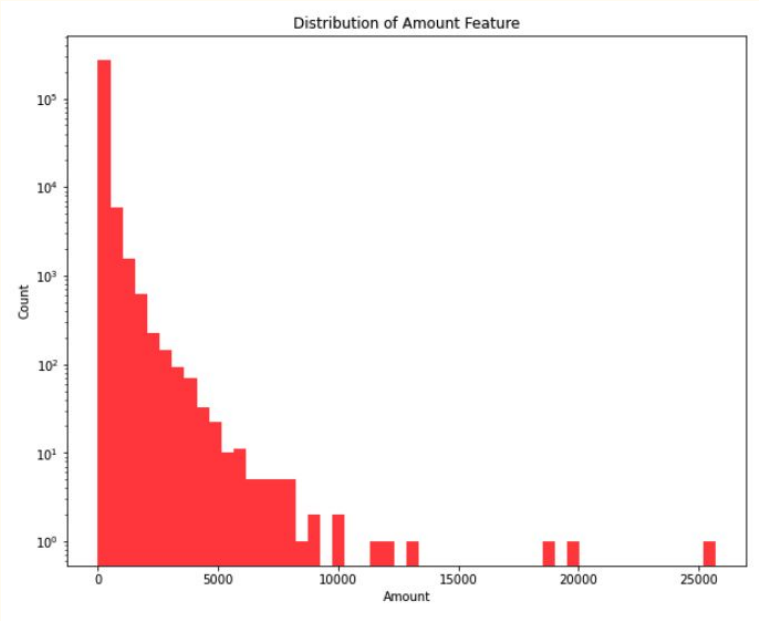
count	473.000000
mean	123.871860
std	260.211041
min	0.000000
25%	1.000000
50%	9.820000
75%	105.890000
max	2125.870000

Name: Amount, dtype: float64

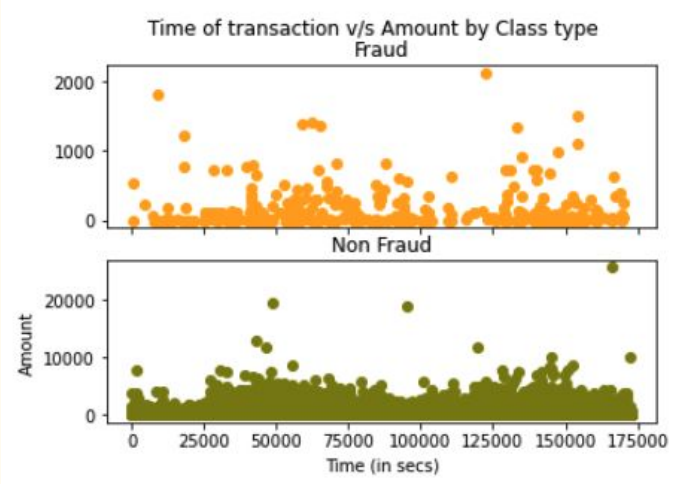


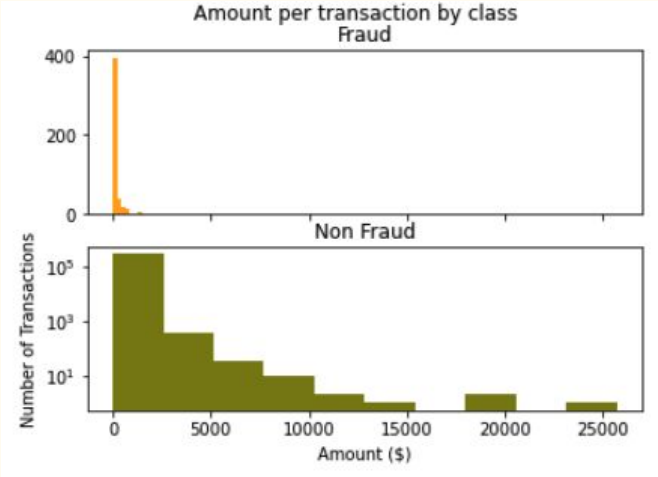
Visualizations of Time and Amount





Study of Fraudulent Behaviors

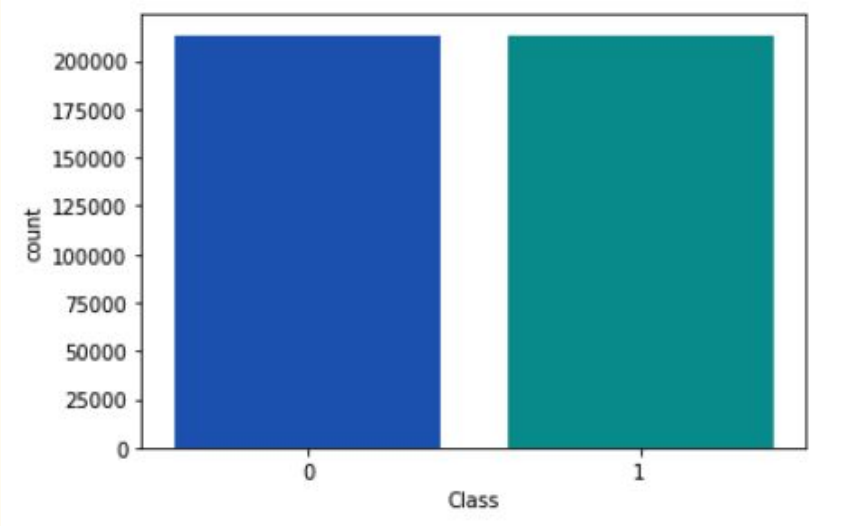




- Fraudulent transactions occur at the same time as normal transaction
- Most of the fraudulent transactions are small amount transactions
- Majority of normal transactions are also small amount transactions.

Supervised Learning Models (Data Preprocessing)

- Handle Imbalance in the dataset before modeling.
- Oversample the minority class - new examples can be synthesized from the existing examples
- Synthetic Minority Oversampling Technique - SMOTE technique
- Nearest neighbors to synthetically create fake datapoints to use for oversampling.
- Some of the other techniques:
  - Undersampling
  - Random Over Sampler
  - Class weight



Supervised models

Metrics chosen:

- Accuracy
- Specificity/Precision/True negative rate/Negative Predictive Value
- Sensitivity/Recall/True Positive rate/Positive Predictive value
- F1 Score

Modeling steps:

- Train-Test Split (Training Data - Used to model, Test data- set aside for validation)
- Standardize data (normalize i.e. $\mu = 0$ and $\sigma = 1$)
- Instantiate and fit to training data
- Predict on test data(unseen data)

Supervised Modeling Results

	model	train	test	precision	recall	f1_score	ROC AUC	Time
0	LogisticRegression	0.999246	0.999157	0.846154	0.626016	0.719626	0.042123	4.3 s
1	LogisticRegression-SMOTE	0.955376	0.975478	0.059207	0.886179	0.110998	0.027578	9.84 s
2	Random Forest-SMOTE	1.000000	0.999480	0.890909	0.796748	0.841202	0.964585	8 min 19s
3	Bagging Classifier-SMOTE	0.999991	0.998680	0.591195	0.764228	0.666667	0.912786	6 min
4	Support Vector-SMOTE	0.981783	0.984593	0.086587	0.829268	0.156802	0.907065	1hr 42 min
5	Neural Network-SMOTE	0.999848	0.998820	0.625806	0.788618	0.697842	0.893901	1 min 11s

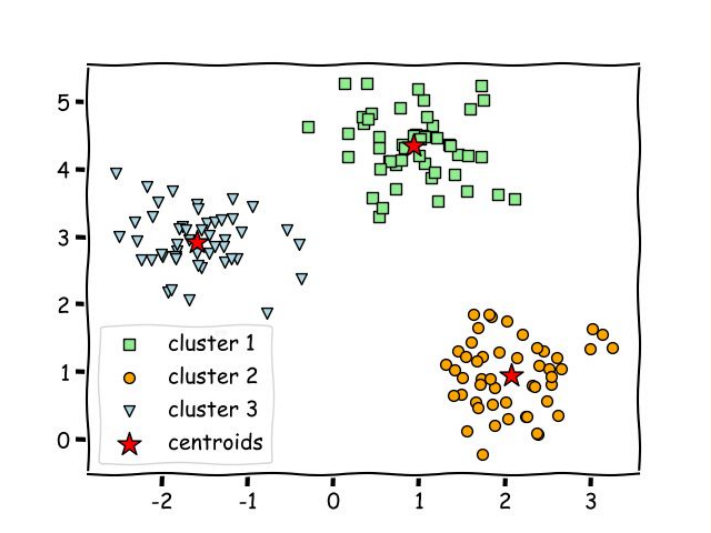
Imbalance data

- Our goal is to improve the recall on our test set
- TN - True Neg - nonfrauds class correctly predicted
- FP False positive- fraud outcome that model predicts as non fraud
- FN - false neg - nonfraud outcomes that model predicts as fraud
- TP - true posit - fraud class correctly predicted

```
TN:  71067
FP:  12
FN:  25
TP:  98
```

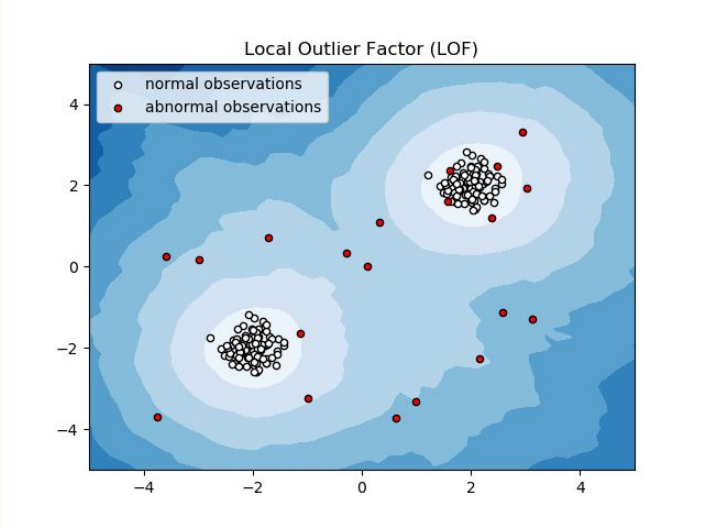
Unsupervised Models - K-Means Clustering

- Data points that are similar tend to belong to similar groups or clusters, as determined by their distance from local centroids.
- It aims to partition the observations into k-sets so as to minimize the within-cluster sum of squares.
- It starts with a group of randomly initialized centroids
- then performs iterative calculations to optimize the position of centroids until the centroids stabilize or the defined number of iterations is reached.



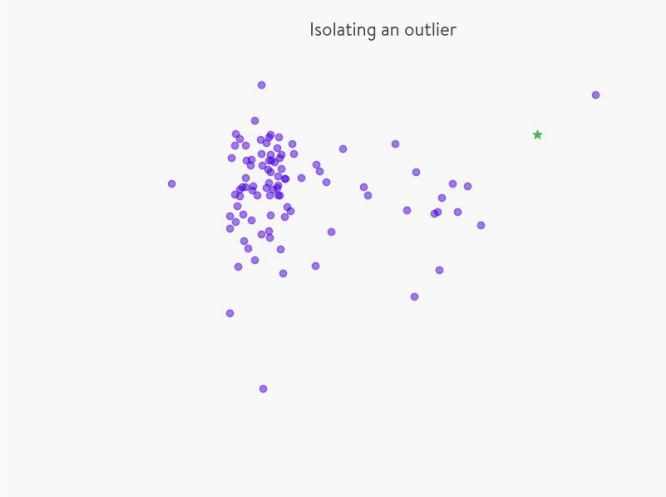
Unsupervised Models - Local Outlier Factor

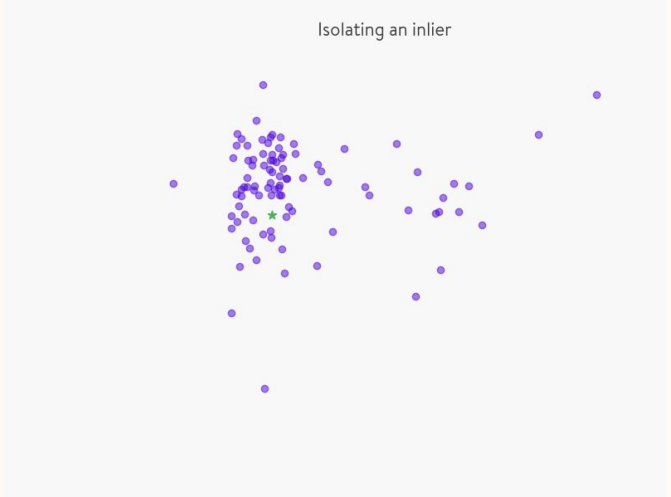
- Density of data points as key factor to detect outliers
- Anomalies appear in low density areas
- LOF compares the local density of its points with respect to its neighbors
- It produces an anomaly score that represents data points which are outliers in the data set.
- Normal points has low score
- Anomalous points have higher scores.



Unsupervised Models - Isolation Forest

- The algorithm isolates each point in the data and splits them into outliers or inliers. Uses Random decision trees.
- This split depends on how long it takes to separate the points.
- Randomly select a feature and then randomly select a split value between the max and min values of the selected feature
- Isolating anomaly observations is easier - fewer conditions to separate
- Isolating normal observations require more conditions.
- An anomaly score can be calculated as the number of conditions required to separate a given observation.

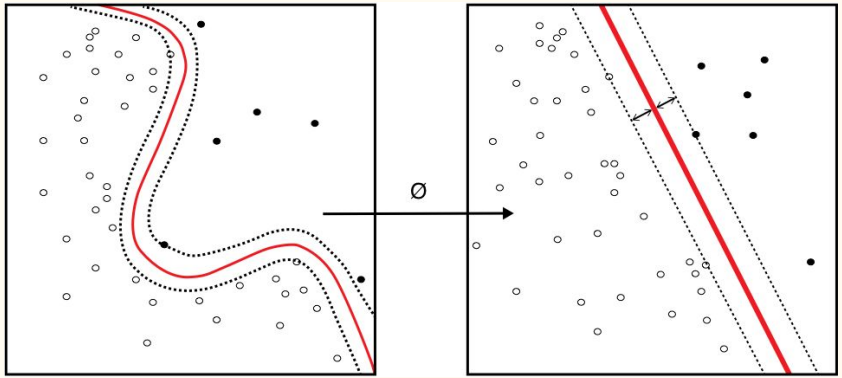




Source: https://quantdare.com/isolation-forest-algorithm/

Unsupervised Models - One-Class SVM

- One-class SVM is a variation of the SVM
- used in one class problem, where all data belong to a single class.
- SVM model divides the training sample points into separate categories based on mac-margin hyperplane
- SVM model then makes predictions by assigning points to one side of the gap or the other.
- One-Class SVM nfers the properties of normal cases and from these properties can predict which examples are unlike the normal examples.



Source: https://commons.wikimedia.org/wiki/User:Zirguezi

Unsupervised Modeling Results

Model Name	Accuracy	Precision	Recall	F1 score	Errors	Time
K Means	0.54	0.001	0.39	0.002	130255	9.26 s
Isolation Forest	0.997	0.28	0.28	0.28	643	44.2 s
Local Outlier Factor	0.996	0.00	0.00	0.00	985	36min
One Class SVM	0.951	0.85	0.03	0.06	13951	37 min 50s

- Isolation Forest detected 643 errors
- IF has a 99.7% accuracy thn LOF, Kmeans or SVM
- Error precision and recall: IF performed much better than all the others
- Detections of fraud cases is around 28% versus LOF detection rate 0% and SVM of 3%
- Finally the model computation time was also better than the others.

Key Insights

- Supervised and Unsupervised models
- These analytical models are run on credit card dataset and
- Models are evaluated with help of confusion matrix.
- Among the all models, random forest and Isolation Forest performed best in terms of accuracy, precision and recall and computation times.

Next Steps

- Conduct gridsearch to tune the hyper parameters for better scores
- Conduct Anomaly Detection with Time series approach
- The spending behavior of the customer can be studied over time
- This can may be help detect the transactions as legitimate or fraud in real time

Thank You!

Image Source: https://unsplash.com/s/photos/different