Are You A Victim Of Credit Card Fraud?

OR

Heard Of A Victim Of Credit Card Fraud?

Presented By: Carine Tan
Cohort 7, 26 November 2020

TOPIC : CREDIT CARD FRAUD DETECTION

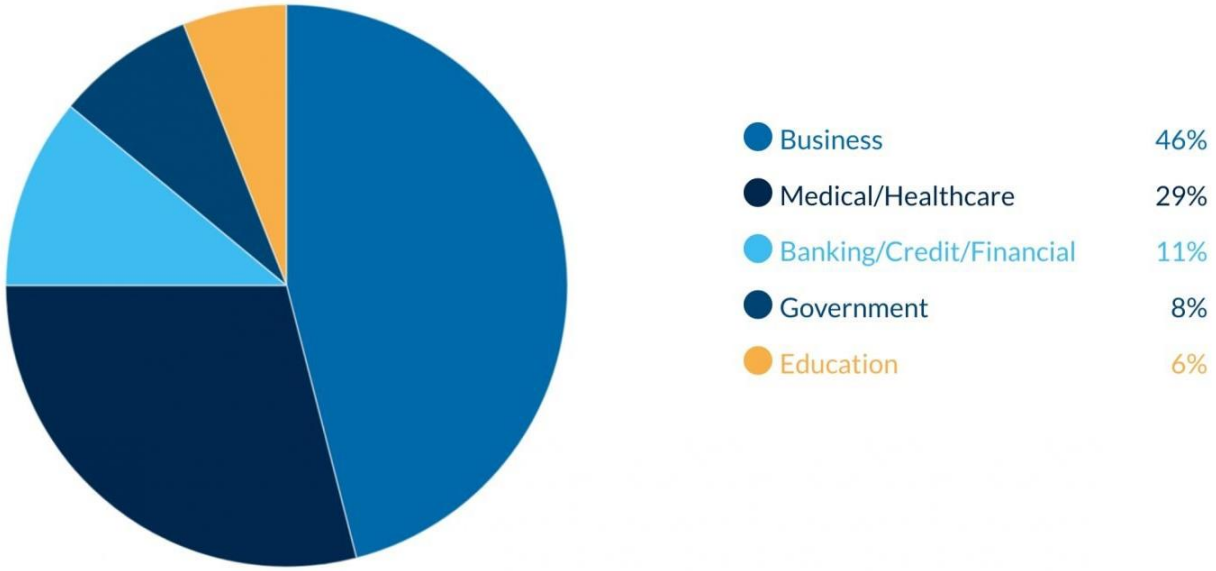
AGENDA

1. Backstory
2. Introduction
3. Case Study
   - EDA
   - Transformation
   - Outliers Detection & Removal
   - Visualization
   - Classification Algorithms
4. Pit Falls

BACKSTORY : WHY DID I PICK THIS TOPIC?

- The world is transitioning to cashless society.
- Retail shops adopt cashless payment methods e.g. contactless cards and smartphone payments.
- Sales transactions have gone online and increased after the COVID-19 pandemic outbreak. Fraud cases increased too.[#2#7]
- Data Breaches (PDPA) : Identity is stolen.
- Credit card details, personal information (name, telephone, address, email, contacts list). [#1]
- Fraud not limited in COO but transnational online credit card fraud.[#4]

Data Breaches By Business Sector in 2018



"In 2015, (credit cards) issuers bore 72% share of losses. Merchants & ATM acquirers assumed 28% of liability – (Nilson Report, October 2016). "[#1]

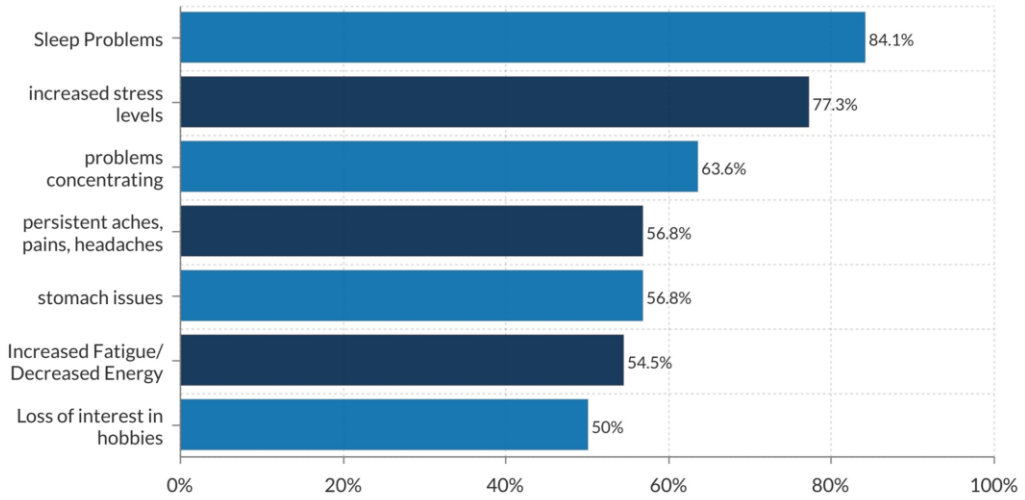
These are operational losses for credit card companies & merchants, lead them to look for credit card fraud detection softwares.

Credit Card Fraud Statistics (for Year 2018)
Population: USA
Source: Shift Credit Card Processing [#1]

SHIFT Credit Card Processing

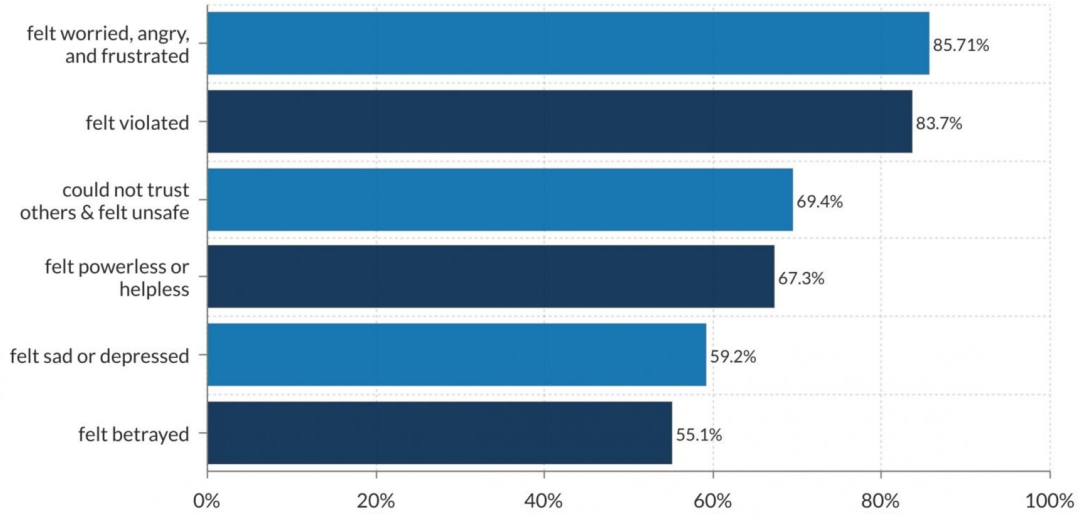
Physical Impact of Identity Crime Victims

Of the people that responded to this survey done by the Identity Theft Resource Center 2018 Aftermath Study:



Emotional Impact of Identity Crime Victims

Of the people that responded to this survey done by the Identity Theft Resource Center 2018 Aftermath Study:



INTRODUCTION:
What is Credit Card Fraud & How Can It Happen?

- **WHAT?[#1] :** It happens when someone uses your credit card or credit account to make transactions without your authorization.
- **HOW? [#7] :**
  - Counterfeit cards at the point of sale & ATMs
  - Card-not-present (CNP) transactions: online, mail, telephone, a social network, or mobile app. Authentication not required.
    - Cross-channel integrations allowing access to multiple accounts from 1 platform[#3].
  - Mobile banking
  - Fraudulent credit applications
  - Lost and stolen cards
  - Smaller categories

" Online shopping now creates the greatest fraud opportunity - the security of the EMV credit and debit cards (chip & PIN) are driving more fraudulent activity to the online or e-commerce industry. "

– Quoted from Reference #1

WALK THROUGH A CASE STUDY: DETECTING CREDIT CARD FRAUD USING MACHINE LEARNING MODEL[#5]

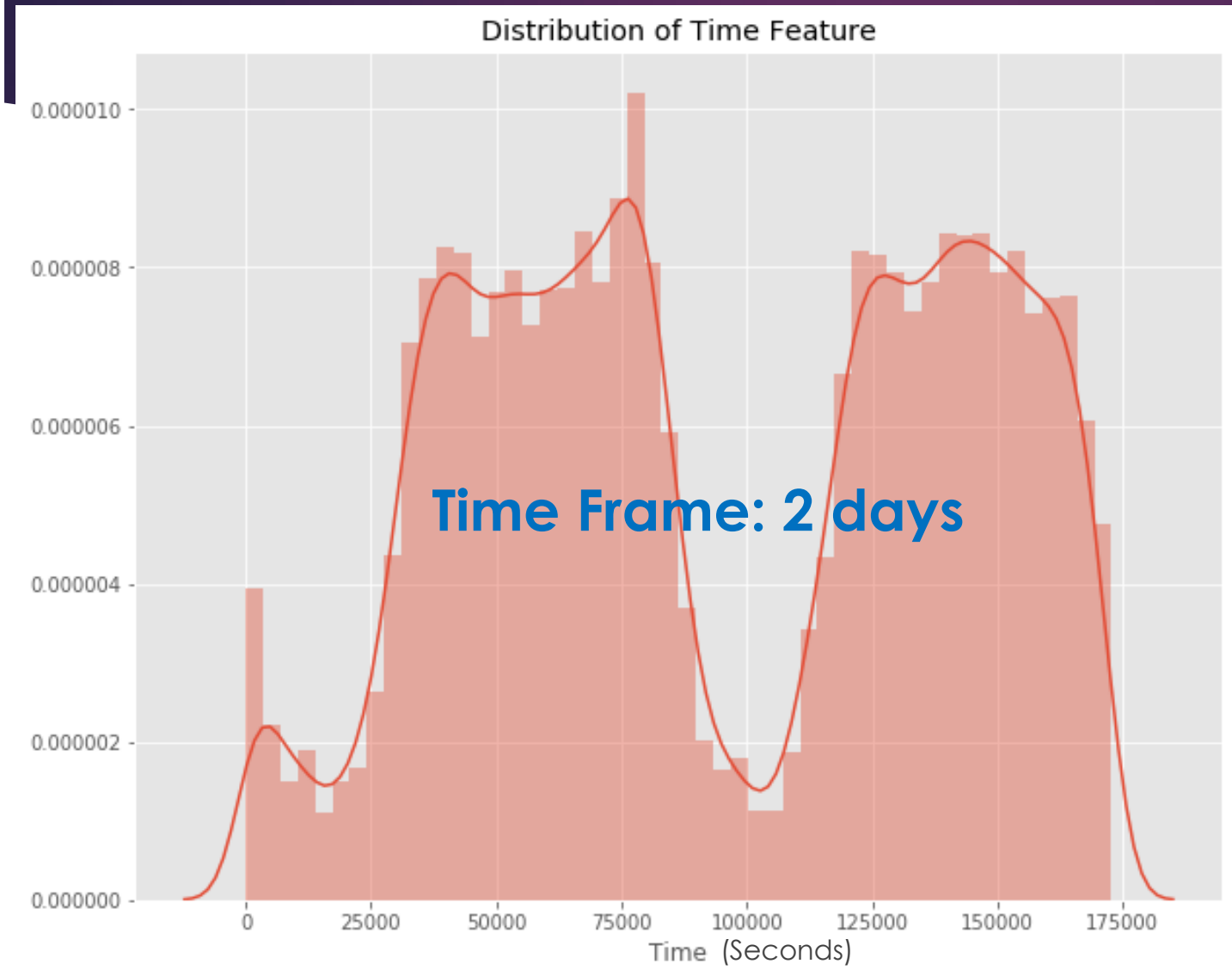
Background of Case Study

- **Raw Data Set :**
  - 31 variables – 28 anonymized, remaining 3 namely the Time, Transaction Amount and whether the transaction was fraudulent or not.
  - No missing values.
  - Size: 284,807 transactions
  - Time Frame: 2 days
  - Transaction Amount: Mean = $88.35, Max = $25,691.16 (Distribution is heavily skewed, majority transactions are small amount.)
  - 99.83% not fraudulent; 0.17% fraudulent

- **Data Preparation :**
  - Anonymized variables have been scaled and centered around zero.
  - Time and Amount are not scaled which may result M.L. algorithm to perform worse
    - To avoid this issue, Study Team standardized the Time and Amount columns.

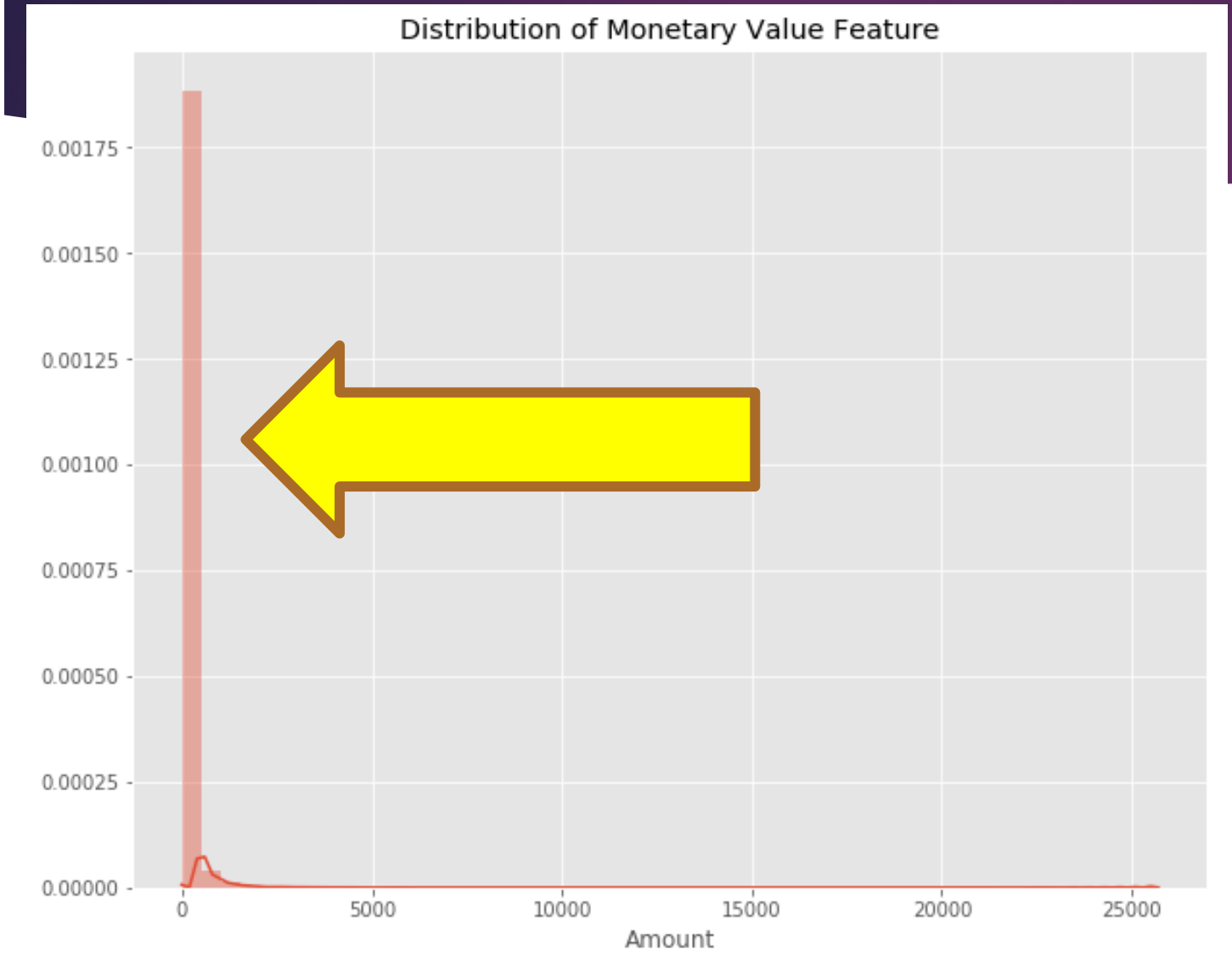
- **Creating Training Set :**
  - Heavily imbalanced: >99% not fraudulent
  - Want an algorithm that predicts fraudulent transaction with high performance instead of predicting non-fraudulent with 99% accuracy.

EXPLORATORY RAW DATA ANALYSIS (EDA) :



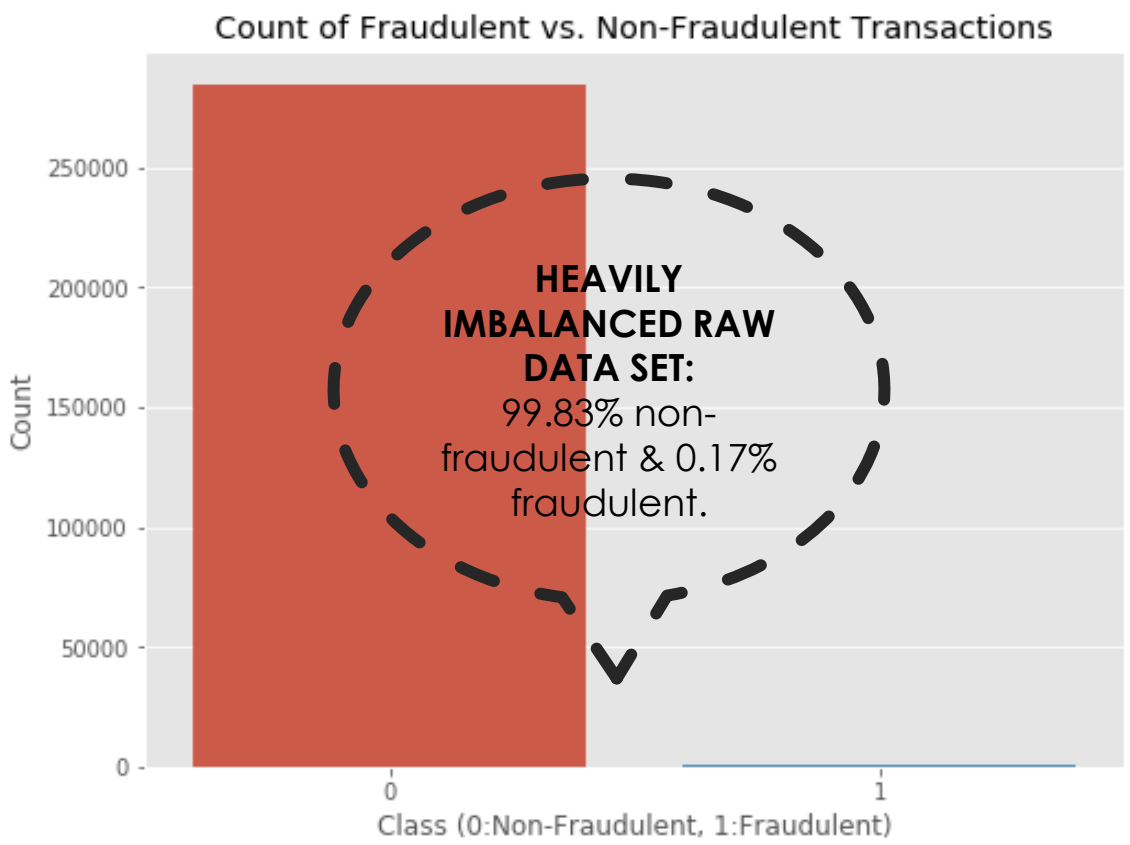
Approximately 28 hours after the first transaction, there is significant drop. It is assumed that most transactions happened during the day, and the drop occurred during the night!

EXPLORATORY RAW DATA ANALYSIS (EDA) :



Distribution is heavily left-skewed: Majority transactions are small amount & tiny fraction comes close to the maximum.

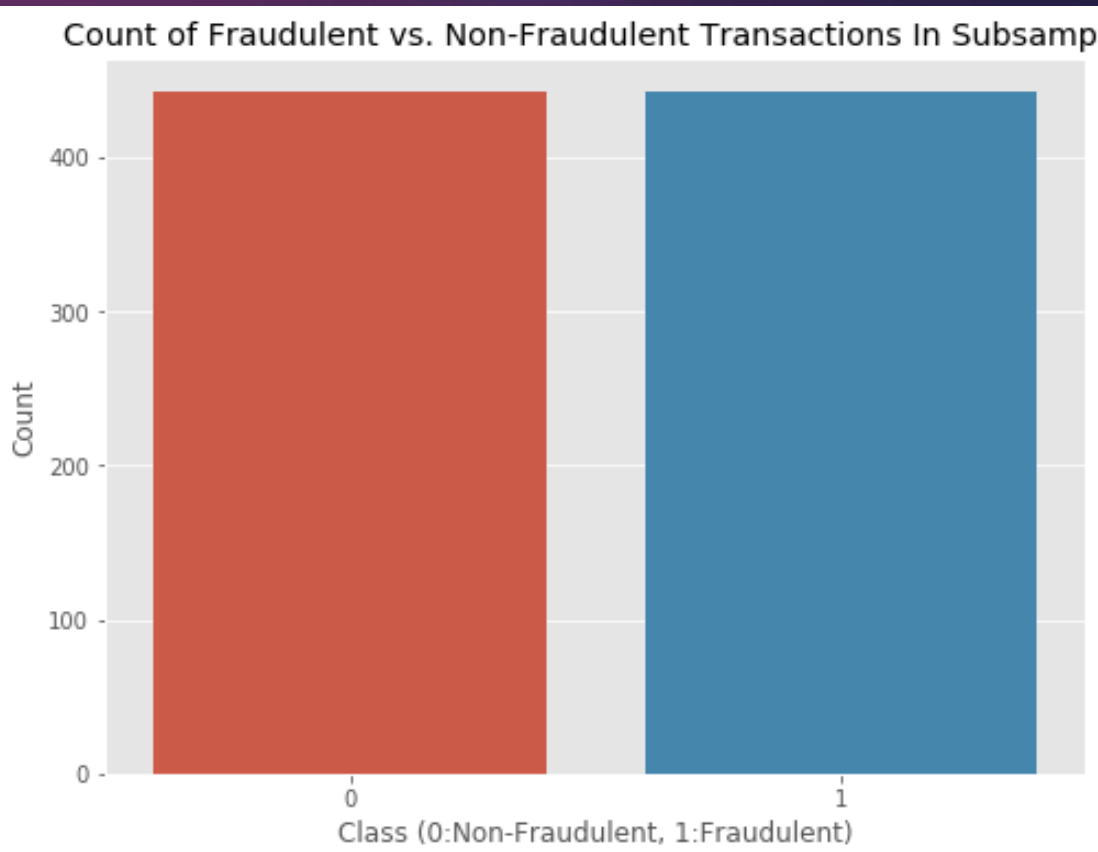
(1) TRANSFORMATION : Create A Balanced Data Set



Transform raw data set from this!

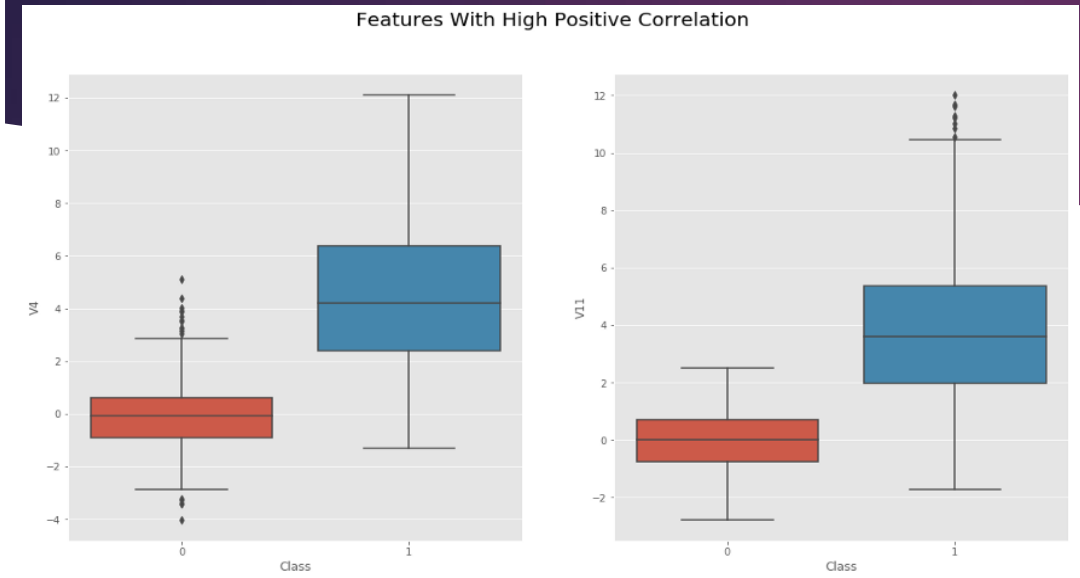
(1) TRANSFORMATION : Create A Balanced Data Set

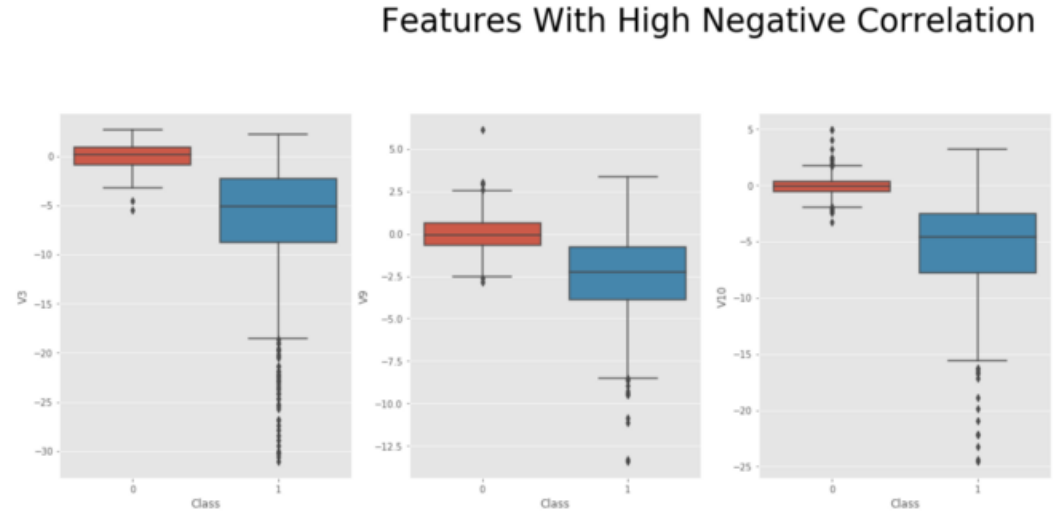
To this!



In this study: they counted all of the fraudulent transactions in data set and randomly selected same number of non-fraudulent transactions; combine the two to create a new balanced subset.

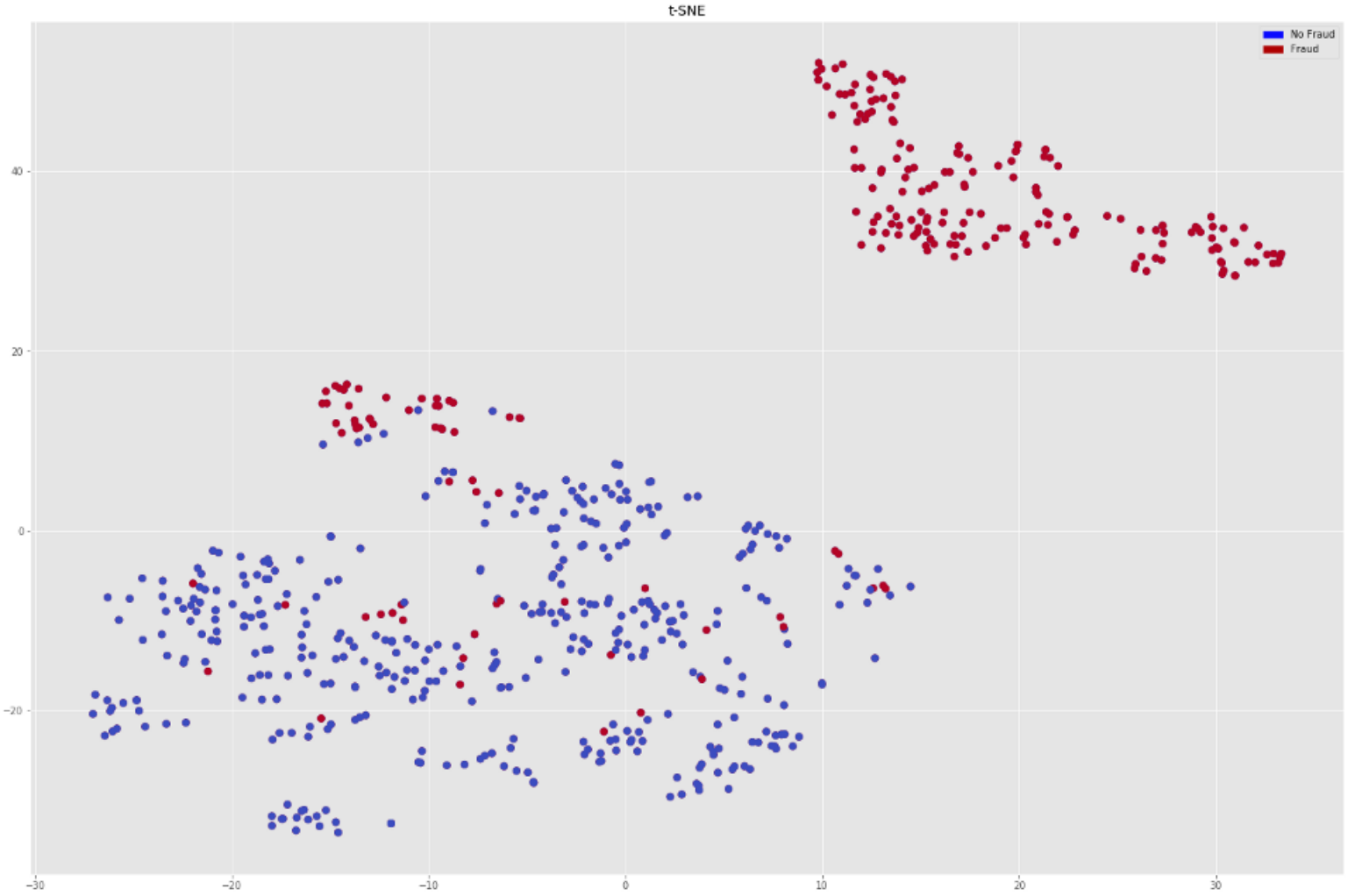
(2) OUTLIER DETECTION & REMOVAL





- Study used box plots to have good visualization about the outliers.
- Outside of 1.5 times the IQR (inter-Quartile Range), data are considered to be outliers.
- However, removing these outliers will significantly decrease the data size.
- Establish a trade off cut-off range of extreme outliers while keeping a good data size.
- In this study, outside of 2.5 times the IQR, data considered as extreme outlier and removed.

(3) VISUALIZATION : Dimensionality Reduction Technique t-SNE

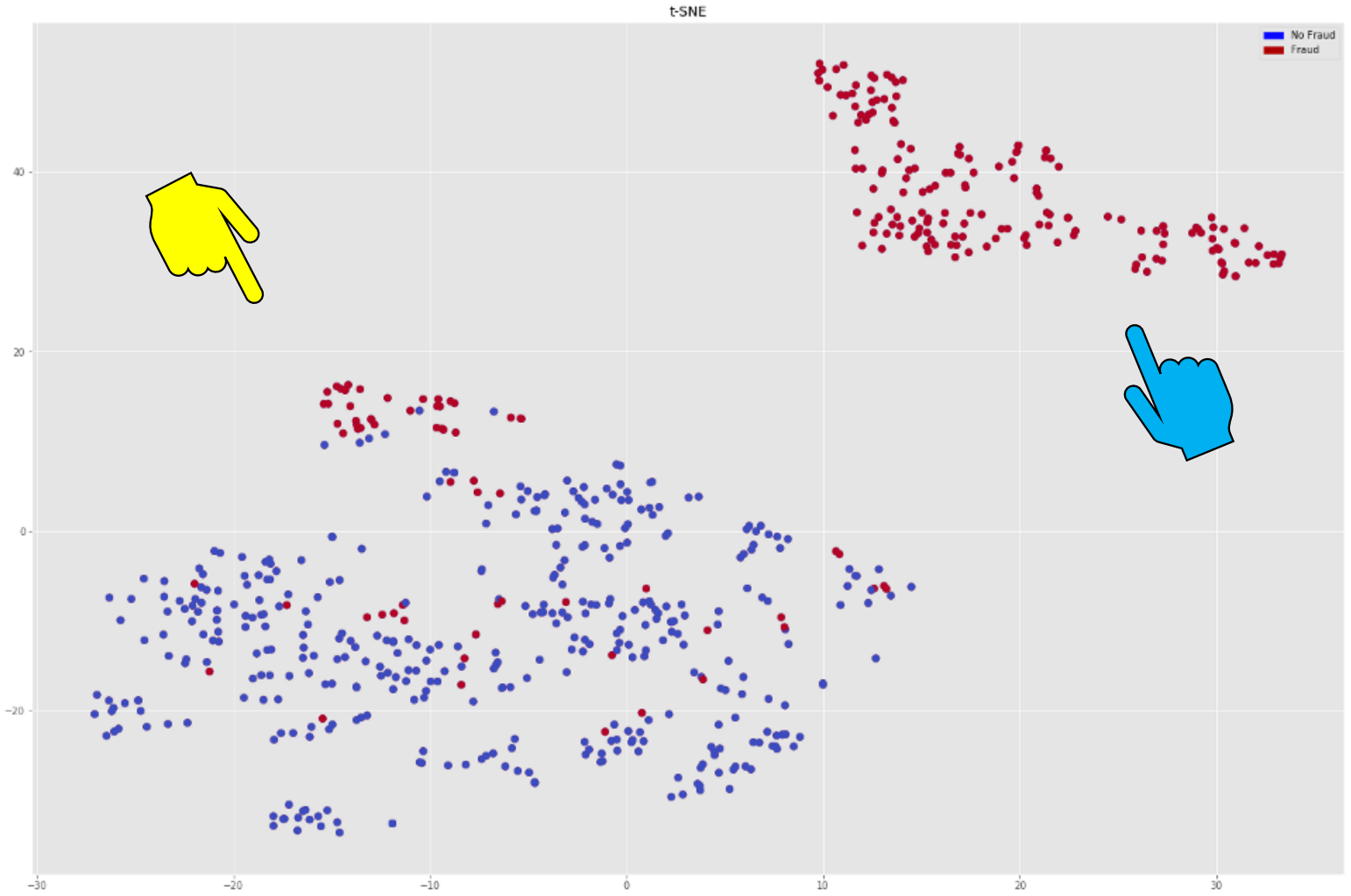


- t-distributed stochastic neighbor embedding (t-SNE) is a M.L. algorithm for visualization.[#9]
- Nonlinear dimensionality reduction technique.
- It models high-dimensional object by a 2- or 3-dimensional point in such a way that similar objects are modelled by nearby points and dissimilar objects are modelled by distant points with high probability.
- Algorithm: Wikipedia[#8]

In this study,

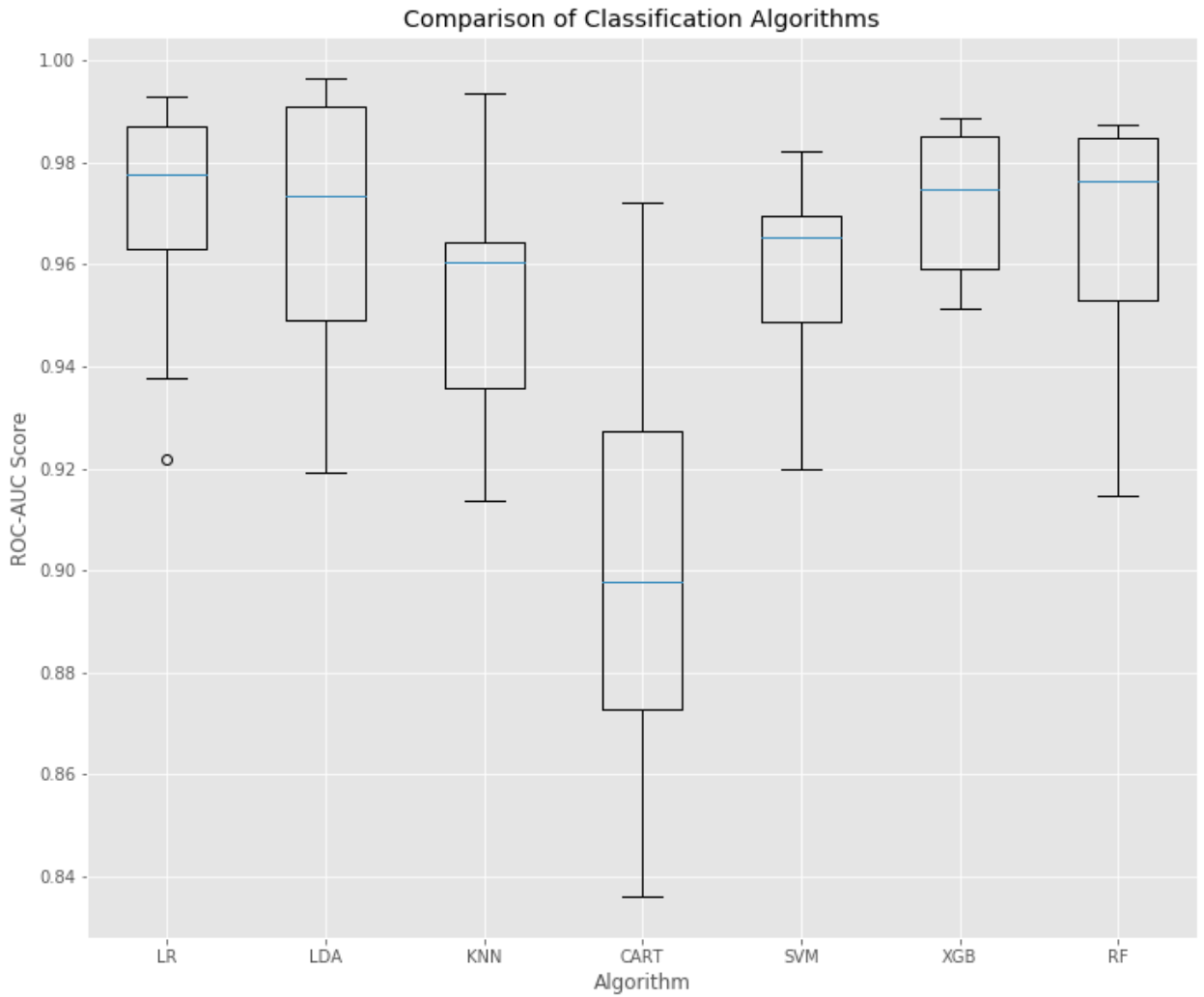
- *Project 30-dimensional plot into a 2-dimensional space.*
- *Result: A scatter plot showing clusters of fraudulent and non-fraudulent transactions.*

(3) VISUALIZATION : Dimensionality Reduction Technique t-SNE



- Majority of the fraud transactions are clustered on the high probability scale.
- All non-fraudulent transactions are clustered on the low probability scale.

(4) CLASSIFICATIONS ALGORITHMS : Testing the Performance of Algorithms



- Study team has a set of algorithms.
- Objective : Test and pick out the algorithm that best perform on the data set.
- **HOW? :** Split the balanced data set into 80/20 train-test and applied k-fold cross-validation resampling technique to test fit each algorithm.
- **PERFORMANCE MEASURE :** ROC-AUC. Outputs a value between 0 (worst) and 1 (perfect score). If an algorithm has a ROC-AUG score of above 0.5, its is achieving a higher performance than random guessing.

RESULT :

- Random Forest (RF) chosen as the best fitting model.
- Higher degree of comprehensiveness in accuracy & also create business value.
- Slightly decreasing performance.
- **Limitation :** not able to see the specific factors for detecting fraudulent transactions as most features in data set are anonymized.
- **Future Work :** Fine tuning of the RF algorithm with better quality data set.

It's a long-haul battle against hackers & frauds. Technology security is advancing so are the fraudsters!

Positive: Replacement of human work in running through records to find fraudulent transactions or proactive reporting from consumers on suspecting credit card frauds.

It is a detection method but not corrective and preventive method.

Credit cards would be completely phased out in the near future. May be replaced with facial or fingerprint recognition or alternative payment acknowledgement methods, but still these can be manipulated and we would be dealing with new security challenges. Use data science to counter-prevent and fight against these potential hackings & frauds.

REFERENCES

1. Shift Processing, Credit Card Fraud Statistics [Updated September 2020] https://shiftprocessing.com/credit-card-fraud-statistics/
2. Forbes, This Week In Credit Card News: Pandemic Brings Huge Increases In Card Fraud And Mobile Banking https://www.forbes.com/sites/billhardekopf/2020/05/29/this-week-in-credit-card-news-pandemic-brings-huge-increases-in-card-fraud-and-mobile-banking/?sh=711a2c8d56c8
3. Forbes, Social Media Platforms Increasingly Popular With Cybercriminals https://www.forbes.com/sites/jessicabaron/2019/04/30/social-media-platforms-increasingly-popular-with-cybercriminals/?sh=3da0b3d17324
4. The Straits Times, At least 21 people in Singapore fall prey to transnational online credit card fraud syndicate https://www.straitstimes.com/singapore/courts-crime/at-least-21-people-in-singapore-fall-prey-to-transnational-online-credit-card
5. Towards Data Science, Detecting Credit Card Fraud Using Machine Learning https://towardsdatascience.com/detecting-credit-card-fraud-using-machine-learning-a3d83423d3b8
6. Towards Data Science, Credit Card Fraud Detection. Staying Vigilant in the Virtual World https://www.datasciencecentral.com/profiles/blogs/credit-card-fraud-detection-case-study-improving-safety-and
7. Forbes, U.S. Card Fraud Losses Could Exceed $12B By 2020 https://www.forbes.com/sites/rogeraitken/2016/10/26/us-card-fraud-losses-could-exceed-12bn-by-2020/?sh=122bfab5d243
8. Wikipedia, t-distributed stochastic neighbor embedding https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding
9. Journal of Machine Learning Research, Visualizing High-Dimensional Data Using t-SNE (2014) https://lvdmaaten.github.io/publications/papers/JMLR_2014.pdf

THE END

THANK YOU!