

2022.07.01

# Data Analysis Exercise

Chen Yi- Ju (Ernie)

# Table of contents



Overview



Pattern Discovery



Proposed Rule



Metrics and results

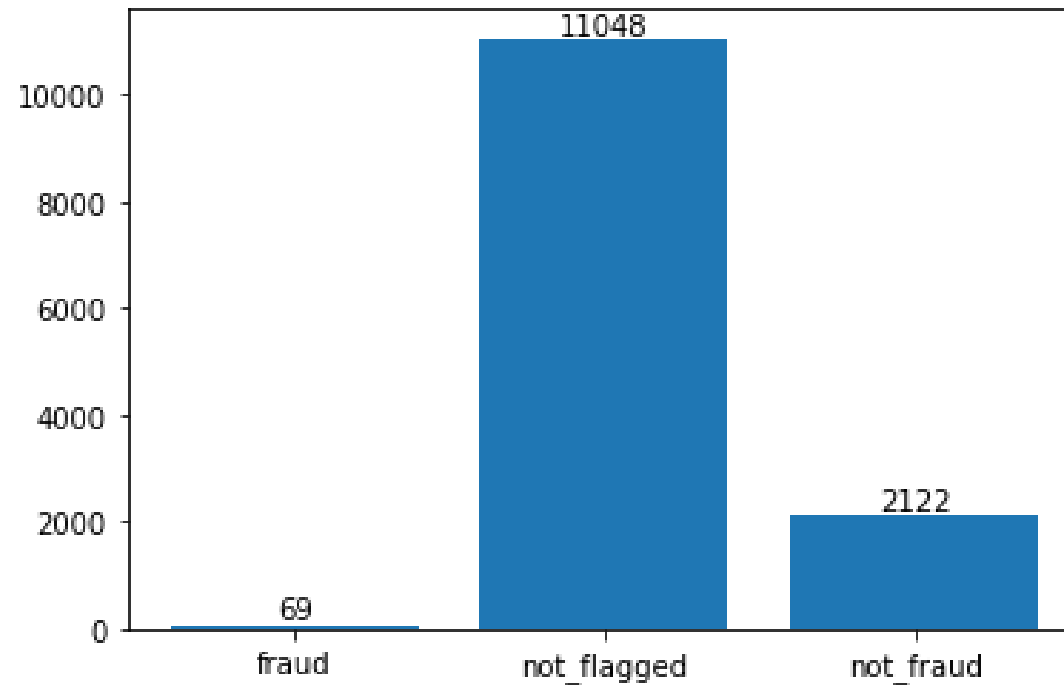


Conclusions and remarks

# Overview

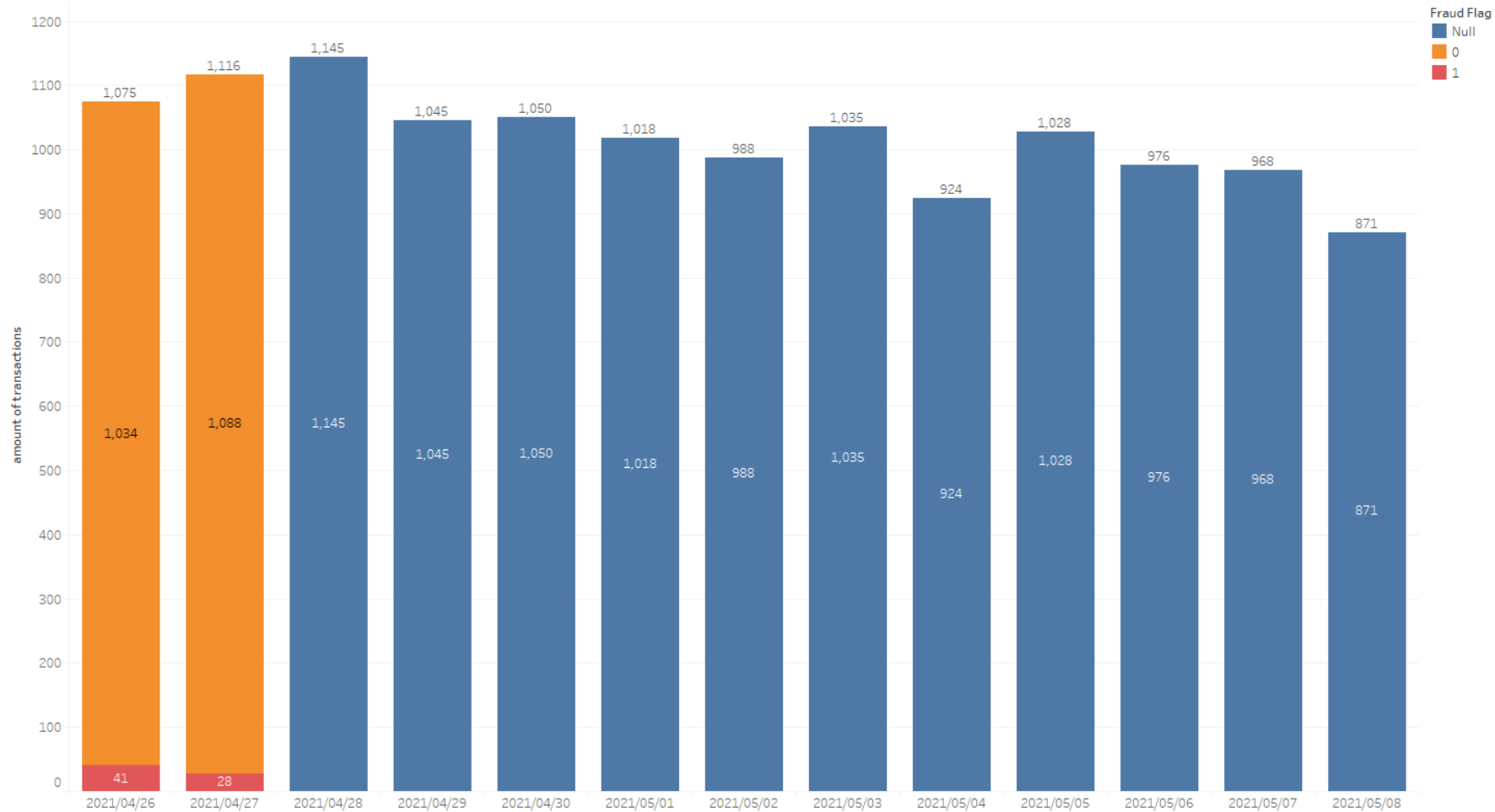
The data

> With a total of 13239 data points, there were 2191 flagged data points  $\sim 4/27$ , with the rest unflagged



# Overview

The data



# Overview

The Goal: answering the questions

**Discovering "why this happened" and "how to prevent it"**

# Overview

The Goal: answering the questions

- > What is the suspicious pattern?
- > Flag the payments that look suspicious from 4/28
- > Propose some rules to block the suspicious payments
- > What additional data points could help?

# Pattern Discovery

Key Findings from data flagged as fraud

- > All fraud happened in the blues store
- > All fraud had payment time happening just before or right after account creation(in less than 2 minutes) -> 0 Itv
- > There are three kinds of email domain used, of which includes “gomen-da.com”, a disposable temporary domain
- > All the transactions happened in new stores (0 days of age)
- > A relatively low percentage (less than 1/3 ) of the users had same phone numbers for Paidy and marketplace

## Proposed Rule

**"if the user account is created within 3 minutes after the payment/transaction, identify it as fraud"**

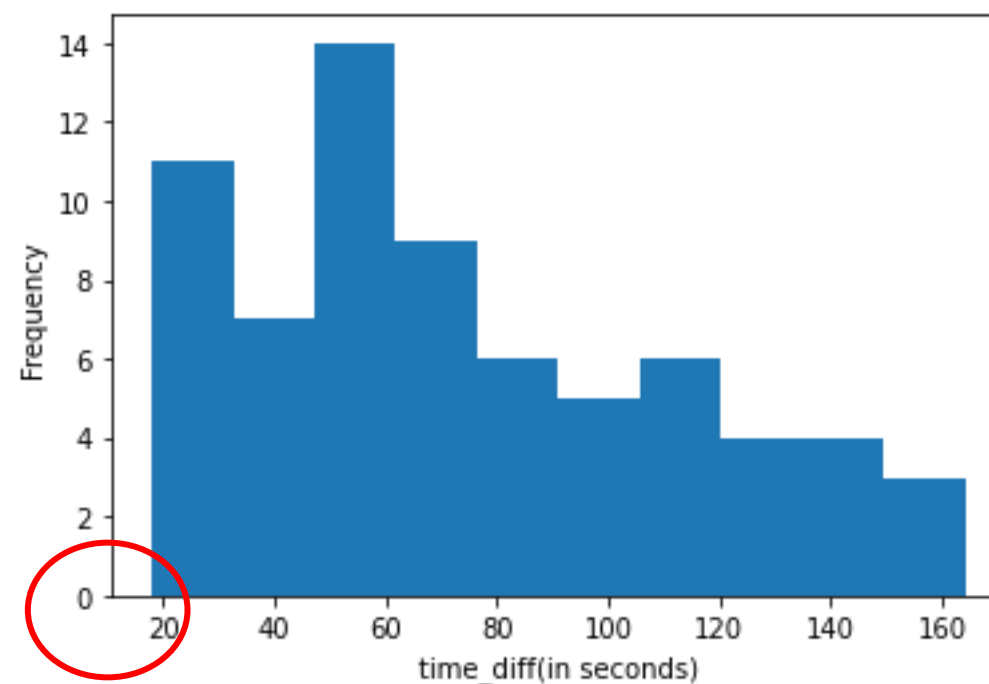
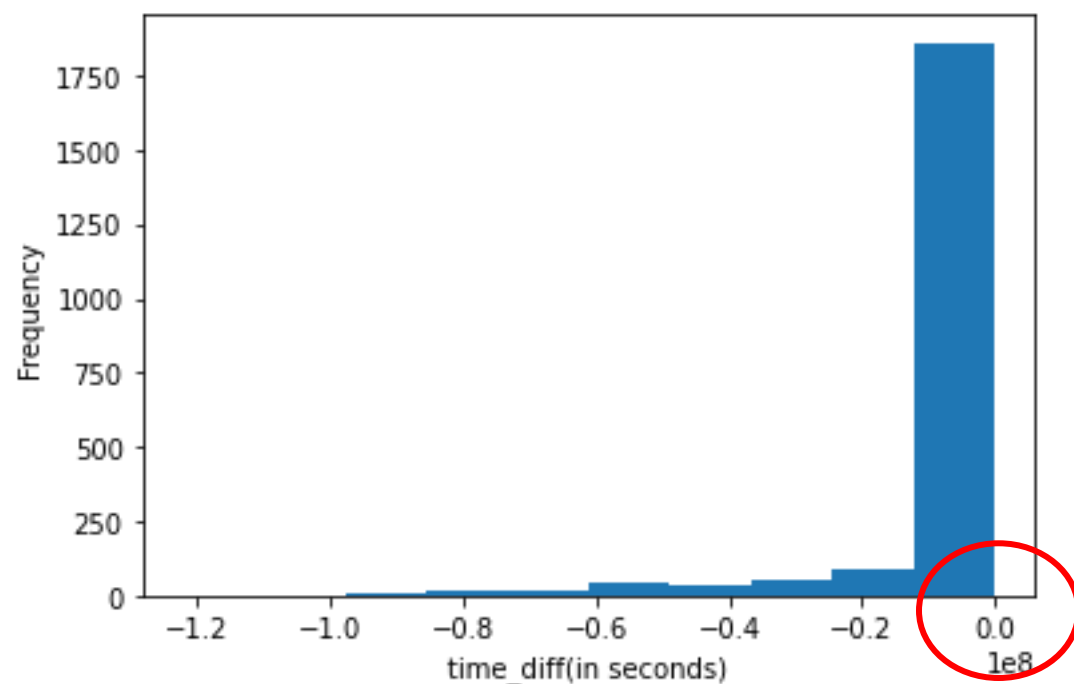


# Proposed Rule

## Reasoning



A clear distribution difference between fraud and non-fraud data

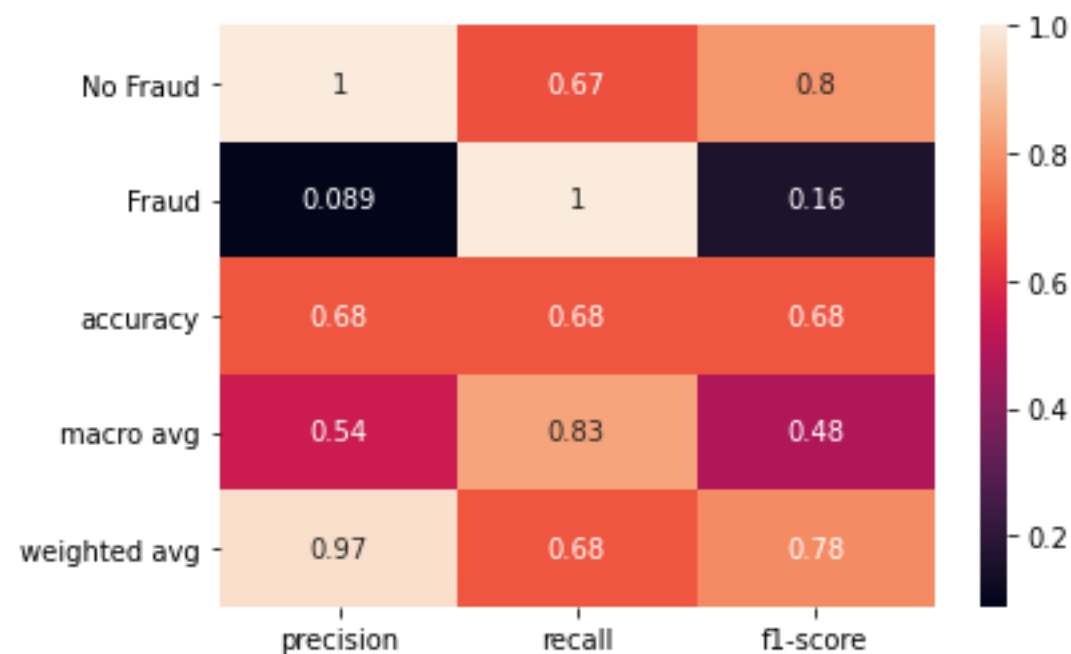
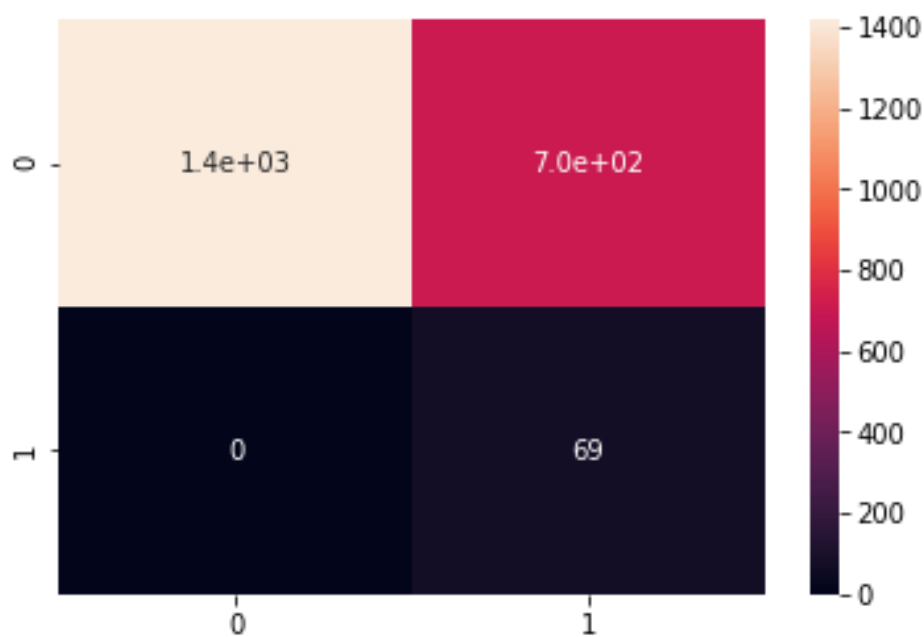


# Metrics and results

Using the original data to evaluate



Below is the classification results using the original labeled data.

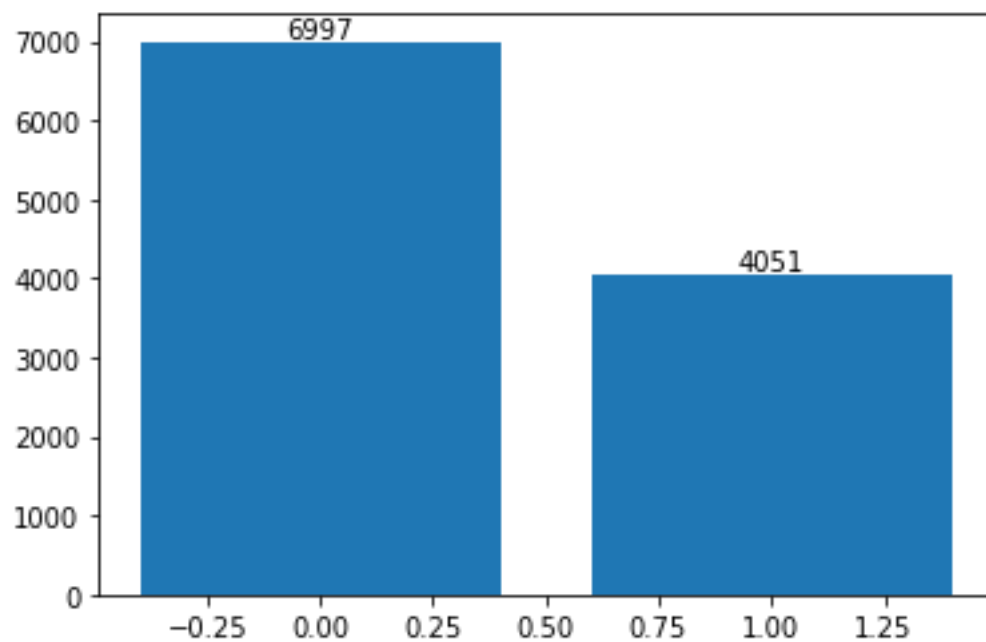


# Metrics and results

The classification on the unlabeled data



Below is the classification results by imposing the rule.



# Conclusions and remarks

The conflict between “making rules” and “correct classification”

- > The results of using the proposed rules were not satisfactory.
- > The problem of “imbalanced data”
- > Relatively complex models(logistic/trees/k-nearest) imposed were able to produce better classification results
- > However, they could not impose “rules” to execute that makes business sense
- > Choosing “interpretability” over “precision” in this case

# Conclusions and remarks

Additional suggestions & personal thoughts

- > email domains whitelists could be provided to weed out domains that fall prey to easy abuse
- > Needs more data and research to create the whitelist
- > There could be potential security risks in the blue market and particular versions of OS systems
- > “Analytics is art”
- > Focus on solving “the actual question” over using particular methods



**Thank you!** **paidy**