**AACS1573 Introduction to Data Science**

MARKING CRITERIA

Your written assignment and presentation will be assessed against the following criteria.

**ASSIGNMENT RUBRICS (CLO3)**

**No.**

**Criteria**

**Evaluation/Marks**

**Mark Achieved**

**0**

**– Nothing**

**Presented/**

**Plagiarized**

**2**

**– Developing**

**4**

**–**

**Approaching**

**Expectation**

**6**

**– Complete**

**8**

**– Excellent**

**10**

**– Beyond**

**Expectation**

1

.

**Logic and**

**Organisation**

Does

not

develop

ideas

cogently,

uneven and

ineffective

overall

organisation,

unclear

introduction

and conclusion

Still

developingideas

cogently, uneven

and

ineffective

overall

organisation,

unclear

introduction and

conclusion

Develops unified

and

coherent

ideas

within

paragraphs with

generally

adequate

transitions; clear

overall

organisation

relating

most

ideas together

good

with

introduction and

conclusion

Develops

ideas

cogently organises

them logically with

paragraphs

and

them

connects

with

effective

transitions Clear

and

specific

introduction and

conclusion

Strongly Developed

cogently

ideas

them

organise

with

logically

paragraphs

and

connect them with

effective transitions

Clear and specific

and

introduction

conclusion n

High innovative

idea

cogently

organised and

logically

sequenced with

clear and

a

specific

conclusion.

2

.

**Conceptual**

**Understanding**

Does

not

respond

using course

content

using

Respond

appropriate and

sufficient course

content

Responds using

appropriate and

sufficient course

content

Respond

clearly

and

effectively

using appropriate

sufficient

and

Good Response in

using course content

and resources in

explaining

your

Highly effective

in

response

using

course

content

and

course

content

outside

and

sources

understanding of the

topic

to

sources

explain

the

understanding

3

.

**Evidence**

Does not

present data

Presents

accurately some

of the necessary

data

Presents clearly

and accurately

most of the

necessary data

Presents

clearly

and accurately all

of the necessary

data

Good Sequence and

Logical

thinking

behind Presenting

data

Most suitable

content

explaining the

end to end

points and with

clarity

4

.

**Relevance of**

**Content**

Irrelevant

content

Partially relevant

content

is

mentioned

Good relevant

content

is

mentioned

Right appropriate

Content is

mentioned

Right content with

clear logic content is

mentioned

Highly effective

content

explaining the

clear

logic

behind the topic

**AACS1573 Introduction To Data Science**

**No.**

**Criteria**

**Evaluation/Marks**

**Mark Achieved**

**0**

**– Nothing**

**2**

**– Developing**

**4**

**–**

**6**

**– Complete**

**8**

**– Excellent**

**10**

**– Beyond**



**PRESENTATION RUBRICS (CLO2)**

**AACS1573 Introduction To Data Science**

**No**

**Evaluation**

**Categories**

**0**

**- No**

**submission**

**1**

**– Weak**

**/Unsatisfactory**

**2**

**– Developing**

**/Needs Improvement**

**3**

**- Satisfactory**

**4**

**- Very Good**

**5**

**– Exceptional**

**/Excellent**

**Mark**

**Achieved**

**1**

**.**

**INTRODUCTION**

How well did the

speaker set the

scene

Did not

appear/atte

nd the

presentation

Lack of interest

and enthusiasm.

The presented

topic was not

introduced and

explained.

Individuals were

not introduced.

Shows a bit of

interest

and

enthusiasm.

The presented topic

was introduced but

vague and

irrelevant.

Introduced self but

not the members.

Seems interested,

but could be

more informed

on

the

introduction.

All members

were addressed,

including self.

and

Interested

enthusiastic

the

about

presentation. The

introduction has

been laid out

properly and to

the point.

All members

were addressed,

including self.

Eager about the

presentation.

Exceptional

introduction which

the

includes

summary.

All members were

addressed,

including self.

**2**

**.**

**CONTENTS**

Was the objective

identified? Was

the presentation

adapted to a wide

range of

audiences? Was it

well organized?

No clear

statement

offered.

Scope too broad

or too narrow;

lacks depth; AND

uses too much

technical

language/ jargon.

clear

No

information

sequence; very

difficult to follow.

Incomplete or

unfocused.

Scope too broad or

too narrow OR

lacks depth OR

uses too much

technical

language/ jargon.

Evidence of some

organization

but

not in an optimal

order; difficult to

follow.

Reasonably clear.

Reasonable scope

and depth; lapse

into detail that

may

not be

accessible to the

audience. Ideas

presented

in

logical sequence;

reasonably easy

to follow.

and

Clear

Good

concise.

scope & depth

without losing

the audience in

technical detail; a

learning

good

experience.

Presented

in

logical

&interesting

ways; easy to

follow but not

oversimplified.

Clear, concise.

Engaging, and

thought-provoking.

Exceptional scope

& depth; a true

learning

experience;

exceeds

expectations.

Exceptional

organization

because the topic

is complex.

Assignment Page 6

**AACS1573 Introduction To Data Science**

The

opinion

clearly

expressed.

Opinion strongly

expressed

and

supported.

**4**

**.**

**CONCLUSIONS**

**CONFIDENCE**

**(**

**reading**

**materials)**

How

well did the

author conclude,

summarize and

recommend?

No

apparent

conclusions; no

discussion

of

implications.

Conclusions

are

restatements

of

previous

statements.

Brings

closure

some

with

synthesis

but

does not address

implications.

Synthesizes the

brings

work;

closure; allude to

broader

implications.

Synthesizes; brings

closure;

conveys

real implications;

suggest

new

perspectives.

**TOTAL MARKS (20%)**

**MARKING CALCULATION**

A. ASSIGNMENT RUBRICS (80%)

B. PRESENTATION RUBRICS (20%)

TOTAL A + B (100%)

CONVERT TO 50%

**Table Content**

Task

Pages

Introduction

1.0

1.1 Definition of Data Science

1.2 Significance of Data Science

10

2.0

Case Study

1

0

3.0

Data Science Process

3.1 Identify Data Source

3.2 Data / Feature Selection

3.3 Data Cleansing

3.5 Data Analysis

1

1

1

1

-1

3

1

3

-1

4

14-16

4.0

Result and Discussion

16

5.0

Conclusion

1

7

6.0

References

1

7

Appendix

1

8

# Introduction

Data science is the field of study which combines the scientific method, math and statistics, specialized programming, advanced analytics, and artificial intelligence to extract the meaningful business insight buried from the data. Banks utilize the data from customer previous history, trends, loyalty and communication by using data science. Nowadays, many banks are utilizing the power of data science to perform the important tasks like fraud detection and Customer Segmentation . Hence, investing in data science technology can add value to your business.

The using of data science in the banking sector is more than a tendency. Banks are using the data science to compete with their competition. Bank companies needed the data to develop insights and make data-driven decisions. Banks are able to make smarter decisions, focus their resources efficiently and improve performance by the helping of data science. Data science is playing some of the major role in banking industry. Here are some of the data science application which is risk modeling, fraud detection, customer lifetime value, customer segmentation, recommendation engines, real-time predictive analytics.

# Case study

In this study, fraud detection by using data science in the bank industry is discussed. Fraud transaction are increasing heavily every year and it has cause the company losing billions amount of money. In the past businesses are using rudimentary analytics and business rules to look for anomalies to create alerts from separate data sets. With the current fraud detection system and the advancements in machine learning, it has made the companies easier to detect frauds and irregularities in transactional patterns. Proactive fraud detection in banking is crucial for the bank companies for providing the security to customers and employees. The sooner the bank detect the fraud, the less money they will loses. By implementing the fraud detection able the banks to achieve necessary protection and avoid significant loses.

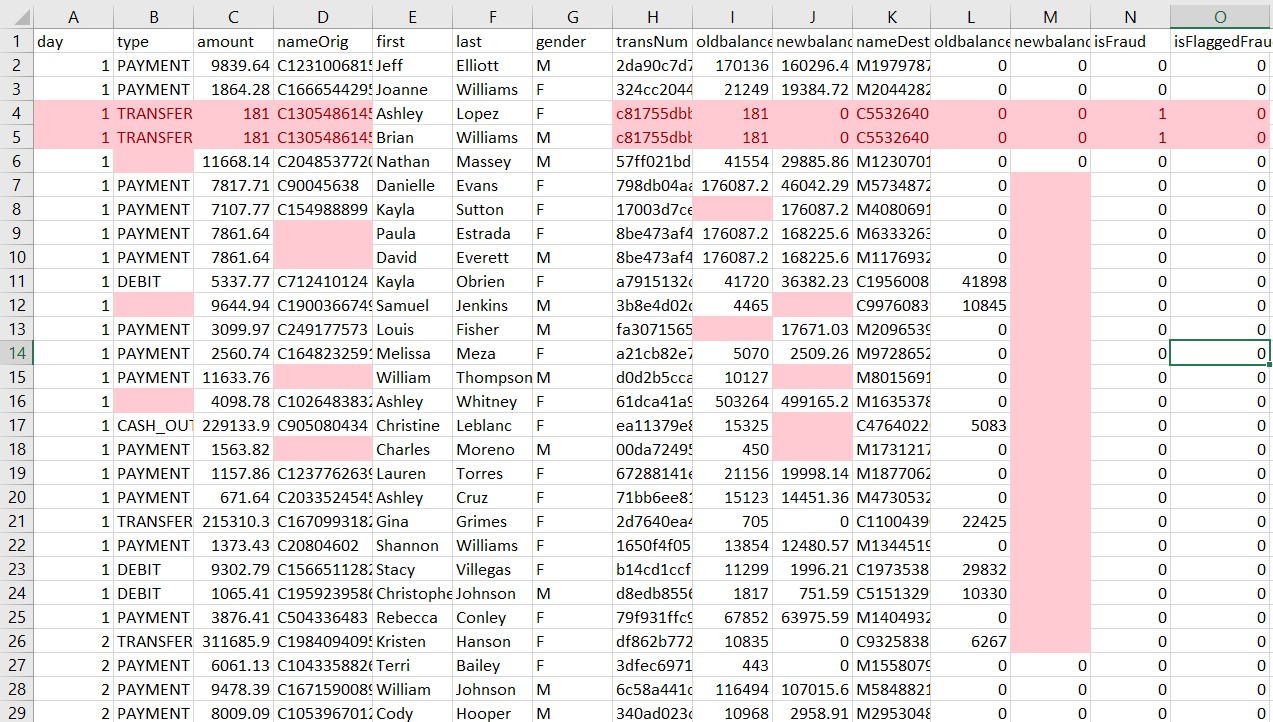
There are some key steps of fraud detection. The first step is Obtaining the data samplings for model estimation and preliminary testing. The second step is model estimation and the third step is testing stage and deployment. Since each of the dataset is different, each of them needs individual training and development by data scientist. Transforming in-depth theoretical knowledge into practical applications requires expertise in data mining techniques such as association, clustering, prediction and classification. As an example, when some unusually high transactions occur, the bank’s fraud prevention system will put the money on hold until the account holder confirms the deal.

## 3.1 Identifying Data Source

Nowadays, with the evolved of the technology, big data systems are helping all the bank to better understand the human behavior and predict for conclusions. Big data systems are able to help the bank industry to identify for the fraud transaction accurately. There are multiple places that people can get the datasets such as Kaggle, Google Public Datasets or UCI Machine Learning Repository. In this assignment, we are using the data sets from PaySim simulator n bank in order to produce a synthetic datasets that resembles the regular operation of transaction and injects the fraudulent behavior to test the performance of fraud detection method. PaySim simulator in bank will help aggragated data from private dataset to generate a synthetic dataset https://www.kaggle.com/ntnu-testimon/paysim1. Usually in the real-life application the source of the datasets is coming from the bank manager. Bank Manager will provide the data such as customer details and customer’s behaviour.

## 3.2 Data/Feature Selection

Figure 1 shows the data of synthetic financial datasets for fraud detection

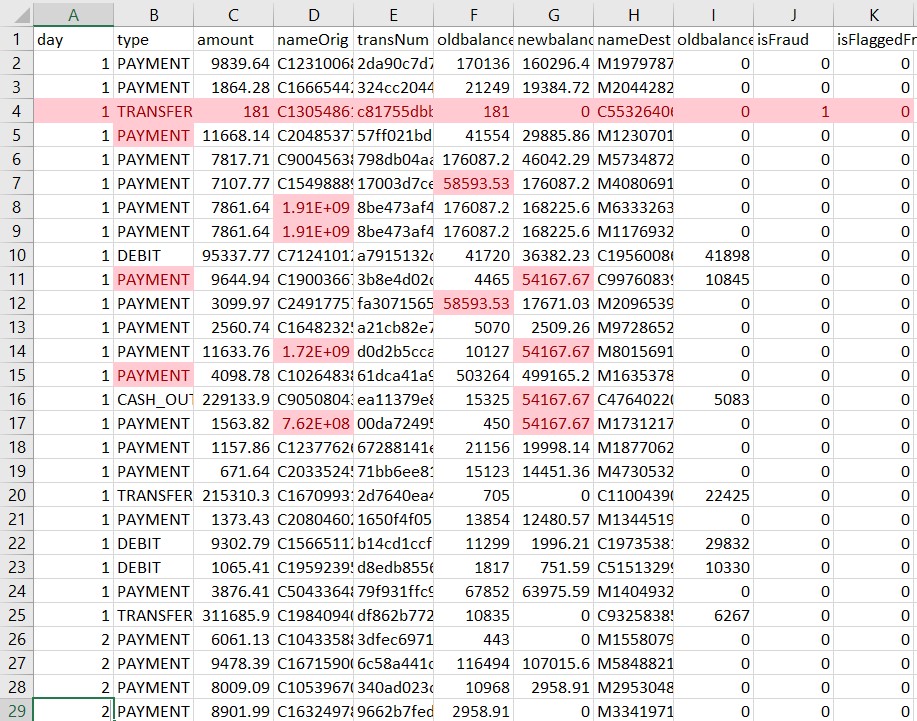


# Figure 1: Synthetic financial datasets for fraud detection before data selection

Figure 1 shows the Synthetic financial datasets for fraud detection before data selection. Figure 2 shows the Synthetic financial datasets for fraud detection after data selection. The dataset consists of 99 customer and 15 features. The feature are sort in to three categories .First is customer behavior feature such as day, and type of transaction. The second category is customer Information feature such as amount, nameOrig , first, last, gender. The third Category is the customer’s transaction information such as transNum, oldbalance, newbalance, nameDest, oldbalanceOrig, newbalanceOrig. In the datasets of the study which is shown in figure 1 contain 15 columns and 100 rows of data. There are multiple of data type which is day, type, amount, nameOrig,first, last, gender, transNum, oldbalance, newbalance, nameDest, oldbalanceOrig, newbalanceOrig, isFraud, isFlaggedFraud. This datasets help people to get know of customer behavior when processing the transaction.

Day is mean unit of time in real world and day 1 means 1 day of time in real world . Type is mean type of transaction such as debit, payment, cash-in, cash-out and transaction. Amount means that amount of the transaction. NameOrig means that customer who started the transaction. First,last and gender means the first name, last name and the gender of the customer. oldbalanceOrig means that initial balance of the bank account before the transaction. newbalanceOrig means that the new balance of the bank account after the transaction. nameDest means that customer who is the recipient of the transaction. oldbalanceDest means that initial balance recipient after the transaction. newbalanceDest means the new balance recipient after the transaction. Data type isFraud means the transactions made by the fraudulent agents inside the simulation. In this specific dataset the fraudulent behavior of the agents aims to profit by taking control or customers accounts and try to empty the funds by transferring to another account and then cashing out of the system. While isFlaggedFraud means the business model aims to control massive transfers from one account to another and flags illegal attempts. An illegal attempt in this dataset is an attempt to transfer more than 200.000 in a single transaction.

# Data after data selection

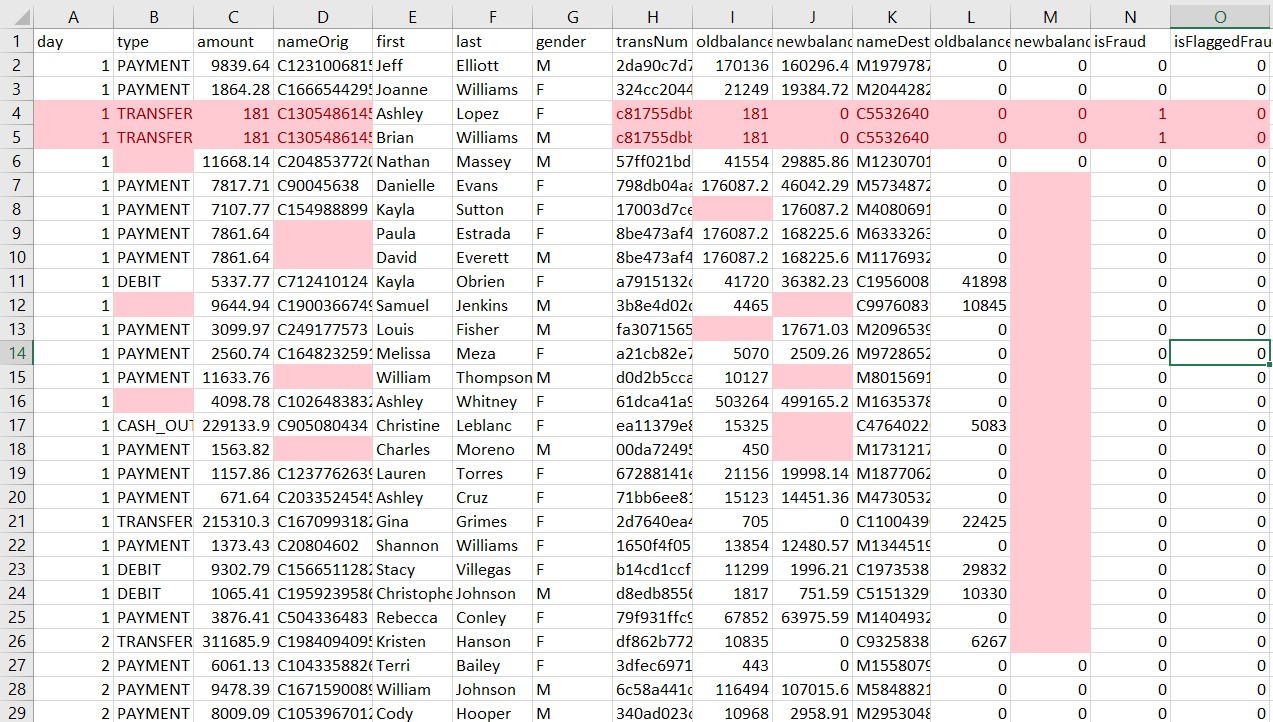


# Figure 2: Synthetic financial datasets for fraud detection after data selection

Data Before data selection contain few data such as day, type, amount, nameOrig,first, last, gender, transNum, oldbalance, newbalance, nameDest, oldbalanceOrig, newbalanceOrig, isFraud, isFlaggedFraud. After the data selection the data such as first name, last name and gender is being taken out. The reason these data is being taken out is because these data type are not suitable for this project. As a general rule, you would be expected to preserve anonymised data only. Personal data should be deleted as soon as possible because it will against the legal requirement for keeping the personal data.

## 3.3 Data cleansing

Figure 3 shows the Synthetic financial datasets for fraud detection before data cleansing while the Figure 4 shows the Synthetic financial datasets for fraud detection after data cleansing.



# Figure 3: Synthetic financial datasets for fraud detection before data cleansing

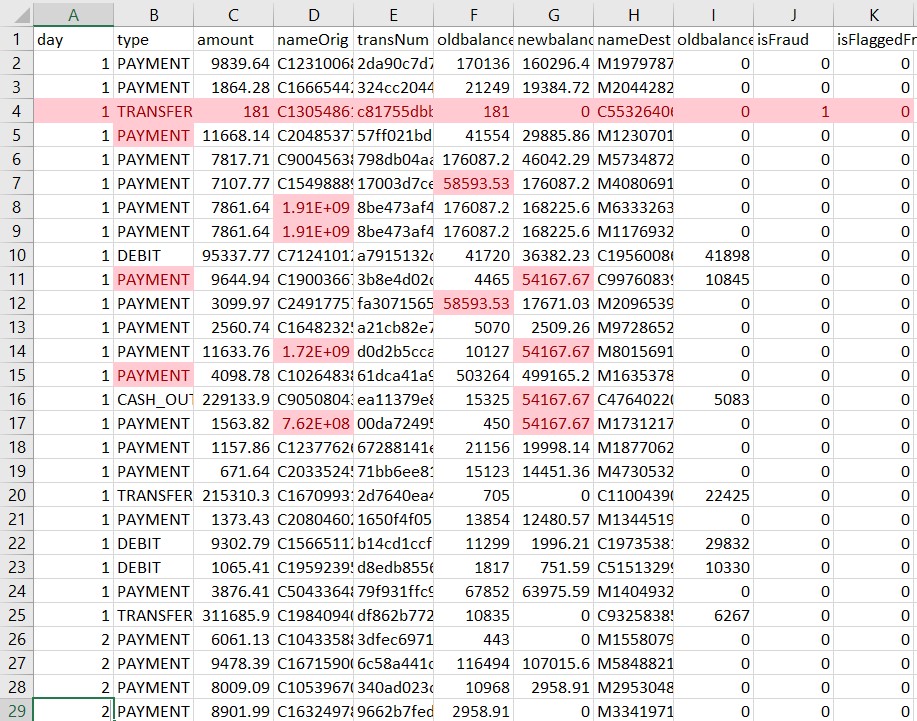


Figure 4: Synthetic financial datasets for fraud detection after data cleansing Data nameOrig before the data cleansing is using the wrong format, as an example of the wrong format 1912850431 and 1716932897. nameOrig data supposed to start with a alphabet C then with the number, as an example of the correct format,C1912850431 and C1716932897. Hence, we clean the data by fixing the error of the nameOrig data by adding a C infront of the number.

Next, there’s some missing value in the oldbalanceOrig and newbalanceOrig Data. This problem is being solve by calculate and filling the mean average value of the data in to the missing column. The reason of using mean value to fill these column is because there are only a few data are missing from the datasets. There’s also a lot of missing data in the newbalanceDest.

Hence, we clean the data by delete the whole column from the datasets.

Next, the there’s also some missing value in the type data. The missing value of the type data is fill by the most frequent data. In addition, there’s duplicate data in the datasets above. The duplicate data is being remove from the datasets and only one of the data is remain in the datasets.

## 3.5 DATA analysis

Data analysis is a necessity for making well-informed and efficient decisions. Data analysis is being used in this study to help the bank to find the fraud and predict the fraud transaction with the datasets provided in the figure 4. Figure 5 shows a pie chart is used to display frequency of the type of transaction of 99 transaction. Figure 6 shows a line chart is used to display the total amount of different of type of payment of 99 transaction. Figure 7 shows a histogram chart is used to display the total amount of transaction of 4 day.

PAYMENT;

57.14

%

TRANSFER;

24.49

%

DEBIT; 9.18%

CASH\_OUT; 9.18%

frequency of type of transaction

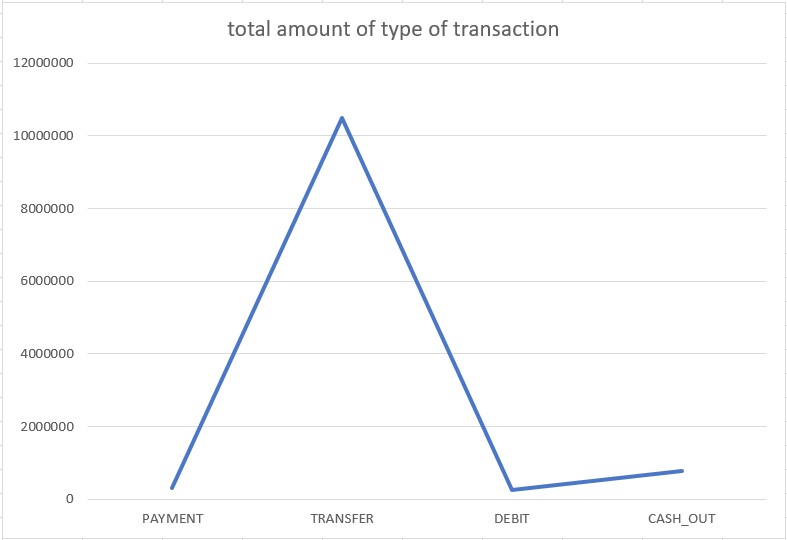
PAYMENT

TRANSFER

DEBIT

CASH\_OUT

# Figure 5: frequency of the type of transaction of 99 transaction



# Figure 6: the total amount of different of type of payment of 99 transaction

Day 1

Day 2

Day 3

Day 4

0

2000000

4000000

6000000

8000000

10000000

12000000

Total amount of transaction of 4 day

Figure 7: the total amount of transaction of 4 day

# 4 Result and Discussion

The result of the pie chart shows that PAYMENT is the type of payment method that is frequently used by the customer following by TRANSFER transaction method. Next, number of customer using the DEBIT and CASH\_OUT payment method are the same. This result of pie chart shows that customers favorite transaction method is using PAYMENT transaction method following by TRANSFER, DEBIT, CASH\_OUT transaction method.

The result of the line graph shows that TRANSFER transaction method involves the most of the money compare to other transaction method following by CASH\_OUT, PAYMENT, DEBIT transaction method. The result of the line graph shows that TRANSFER transaction method is the transaction method that the customer using when transfer large amount of money and it also shows that customer more willing to use this transaction method when transfer large amount of money following by CASH\_OUT, TRANSFER, DEBIT transaction method.

The result of the histogram graph shows that Day 4 involve the most transaction amount compare to the other day such as Day 1, Day 2 and Day 3. It shows that there are a lot of customers are transferring big amount of money in Day 4 then followed by Day 1, Day 3 and Day 2.

# CONCLUCISON

As a conclusion, large amount of customer transaction data are being generated and recorded by using data science. Data science able to help the company to extract the data and provide more insight to the management. With the insight provided it make it easier for the bank to identify the new trends and the behavior of the customer and these data are able to help. Even though it have a lot of good advantage but it also have a lot of disadvantage which is high costs. To understand the pattern of the fraud and detect fraud, a team of data scientists are needed to design and continually update the system. With the involvement of massive amounts of data, businesses are also required to invest in data storage and management as well. Next, it is difficult to collect good data. A great deal of good data are needed in order to build a machine learning model with long-term functionality.

However, from the output of the chart we can understand, we are not able to achieve the objective which is Fraud detection. The reason of this is because fraud detection aren’t able to achieve without the machine learning. Fraud detection and Machine Learning becomes possible due to the ability of Machine Learning algorithms to learn from the historical fraud patterns and recognize them in the future transaction. So the fraud detection objective is not achieve.

# 6 Refrences

Medium,Guru99 2021,What is Data Analysis? Research | Types | Methods | Techniques,viewed

10 February 2021, Available at: <https://www.guru99.com/what-is-data-

analysis.html#:~:text=Data%20analysis%20is%20defined%20as,based%20upon%20the%20data %20analysis.>

2021 Kaggle Inc,Students&#39; Academic Performance Dataset,viewed 12 February

2021,Available at:< https://www.kaggle.com/ntnu-testimon/paysim1>

Craig Stedman,2005 - 2021,DEFINITION data analytics (DA),TechTarget, viewed 12 February

2021, Available at: <https://searchdatamanagement.techtarget.com/definition/dataanalytics?

\_gl=1\*t6qvxf\*\_ga\*MTIxMzQ5NTM0OC4xNjA3NDQxNTI2\*\_ga\_RRBYR9CGB9\*MTYxMz

MwMjIxNS4xLjAuMTYxMzMwMjIxNS4w&amp;\_ga=2.229203418.751049824.16133022161213495348.1607441526>

Available at:<https://wso2.com/whitepapers/fraud-detection-and-prevention-a-data-analyticsapproach/>