National Forensic Sciences University
Knowledge | Wisdom | Fulfilment
An Institution of National Importance
(Ministry of Home Affairs, Government of India)

INTERNSHIP REPORT
ON
“Transaction Fraud Detection”

Submitted To

School of Management Studies,

National Forensic Sciences University

For partial fulfilment for the award of degree

MASTER OF BUSINESS ADMINISTRATION

In

BUSINESS INTELLIGENCE

Submitted By

Vasu

(101MTMBBI2122024)

Under the Supervision of

Prof. Vipulkumar Joshi

National Forensic Sciences University,
Gandhinagar Campus, Gandhinagar – 382009, Gujarat, India.

DECLARATION

I certify that

A) The work contained in the Internship is original and has been done by myself under the supervision of my supervisor.

B) The work has not been submitted to any other Institute for any degree or diploma.

C) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

D) Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the internship/project work/dissertation and giving their details in the references.

E) Whenever I have quoted written materials from other sources and due credit is given to the sources by citing them.

F) From the plagiarism test, it is found that the similarity index of whole Project is less than 10 % as per the university guidelines.

Date: /07 /2023
Place: Gandhinagar

Student Name: Vasu
Enroll. No.: 101MTMBBI2122024

STUDENT DECLARATION

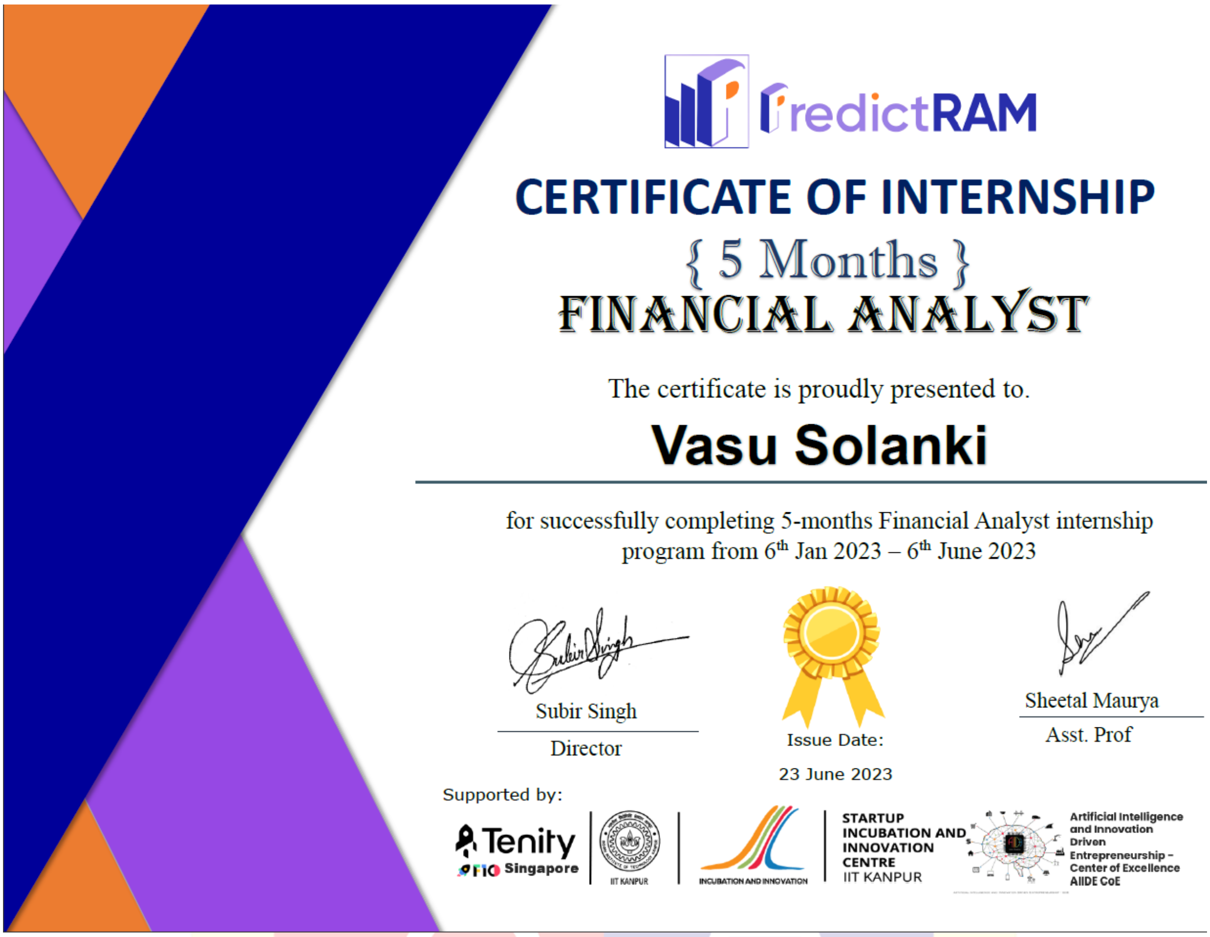
This is to certify that I, **Vasu 101mtmbbi2122024,** have completed the Project titled **"Transaction Pump and Dump Fraud detection model"** under the guidance of Prof. **Vipul Joshi** in partial fulfilment of the requirement for the award of Master of Business Administration in **Business Intelligence** at National Forensic Sciences University – School Management Studies.

We hereby declare that this project is an original piece of work and has not been submitted earlier elsewhere.

Date: 06/07/2023

Place: Gandhinagar

CERTIFICATE FROM THE COMPANY



PredictRAM

CERTIFICATE OF INTERNSHIP

{ 5 Months }

FINANCIAL ANALYST

The certificate is proudly presented to.

Vasu Solanki

for successfully completing 5-months Financial Analyst internship program from 6th Jan 2023 – 6th June 2023

Subir Singh
Director

Issue Date:
23 June 2023

Sheetal Maurya
Asst. Prof

Supported by:

Tenity
ario Singapore | IIT KANPUR | INCUBATION AND INNOVATION | STARTUP INCUBATION AND INNOVATION CENTRE IIT KANPUR | Artificial Intelligence and Innovation Driven Entrepreneurship - Center of Excellence AIIDE CoE

ACKNOWLEDGEMENTS

I'd like to express my profound appreciation to Prof Vipul Joshi my Guide for his important advice and support in finishing my project. He was there to help me every step of the way, and his motivation is what helped me to succeed. I completed my work successfully. I would also want to thank all the other supporting employees that aided me by providing the equipment that was required and crucial, without which I would not have been able to execute properly on this job.

I'd also want to thank the University for approving my project in my selected sector of expertise. I'd also want to thank the writers of the different blogs, videos, and scripts to which I referred during the process.

With Sincere Regards,
Vasu
MBA in Business Intelligence

ABSTRACT

Fraud detection for credit/debit card, loan defaulters and similar types is achievable with the assistance of Machine Learning (ML) algorithms as they are well capable of learning from previous fraud trends or historical data and spot them in current or future transactions.

A pump and dump (P&D) scam is a fraudulent practise that seeks to benefit by inflating the price of a stock by recommendations based on incorrect or misleading information.

or comments that are grossly overstated.

Recently this type of fraud has been increased as digitalization increases.

People are now aware of How fake news of stock market can result in money fluctuation and proper research has been going on for decades.

In this project we will talk about how transaction fraud occur and how pump and dump system is associated with it. We will examine mechanism, their motives and some mitigation measure.

ABOUT THE ORGANIZATION

PredictRam is a recognized **FinTech Startup by DPIIT (Govt of India) and a proud member of the FICCI Startup program,** they are the providers and developers of Collective-Intelligence platform for predicting Economic & financial markets events using blockchain, predictive analysis, big data and machine learning technologies.
By sourcing estimates from a diverse community of individuals.

They provide data that is not only more accurate but is also a more representative view of expectations when compared to sell-side-only data sets which suffer from demonstrable biases.
Goal at PredictRAM is to give the market a transparent data set of true expectations while providing analysts with a platform to build a verifiable track record.
PredictRam Providers and developers of open financial results & economic events collective intelligence predictive analysis platforms where hedge fund, independent, and sell-side analysts, along with investors, industry experts and students contribute their opinions and forecast estimates for companies & events.

Their diverse network consists of financial analysts, equity analysts, financial domain students, analysts, and data science students, data analysts, CA, CFA, and FRM students across the globe.

List of Figure

TABLE OF CONTENTS

Chapter – 1: Introduction

In this Major project, **we will focus on Transaction Fraud Detection using ML models, Pump and Dump system**, Transaction fraud refers to any unauthorized use of a legitimate user's payment information for making purchases, without their knowledge or consent. This type of fraud is quite broad and can include a wide range of activities. Typically, with credit cards, transaction fraud involves the unauthorized use of a victim's credit card to make purchases. This practice is illegal based on securities law and can lead to heavy fines. The burgeoning popularity of cryptocurrencies has resulted in the proliferation of pump-and-dump schemes within the industry.

Recent News About fraud-

1 First Republic was among the regional bank failures of early 2023-First Republic Bank customers included businesses and individuals with deposits of more than the $250,000. Federal Deposit Insurance Corp. (FDIC) insurance limit. Nearly two-thirds of its deposits were uninsured. First Republic’s failure was due to a run-on deposits following the collapses of Silicon Valley Bank and Signature Bank. JPMorgan Chase acquired First Republic Bank on May 1, 2023.

2. Adani Fraud - In month of Feb Gautam Adani, India’s top billionaire, had a good start to the year when he announced a follow-on public offer to raise INR 20,000 cr from the Indian stock market. Troubles started mounting soon after a New York-based investment research firm by the name of Hindenburg Research published a report accusing Adani of potential fraud. The scathing report alleged that the Adani Group indulged in accounting fraud and manipulation of stocks, a claim that has brought Adani down to his knees from the throne of the third richest in the world to slip off the ranking of even the top 20 as of today.

3. Silicon Valley Bankruptcy - On March 10, 2023, Silicon Valley Bank (SVB) failed after a bank run, marking the second-largest bank failure in United States history and the largest since the 2007–2008 financial crisis. It was one of three March 2023 United States bank failures. Seeking higher investment returns, in 2021 SVB began shifting its marketable securities portfolio from short-term to long-term Treasury bonds. The market value of these bonds decreased significantly through 2022 and into 2023 as the Federal Reserve raised interest rates to curb an inflation surge, causing unrealized losses on the portfolio.

4. Banks’ digital payments frauds nearly double in FY23- Even as the overall value of frauds reported by Indian banks halved from 59,819 crore in FY22 to30,252 crore in FY23, the value and volume of digital frauds committed using cards and internet-based payment methods nearly doubled in the previous financial year, data from the Reserve Bank of India's (RBI) FY23 annual report showed.

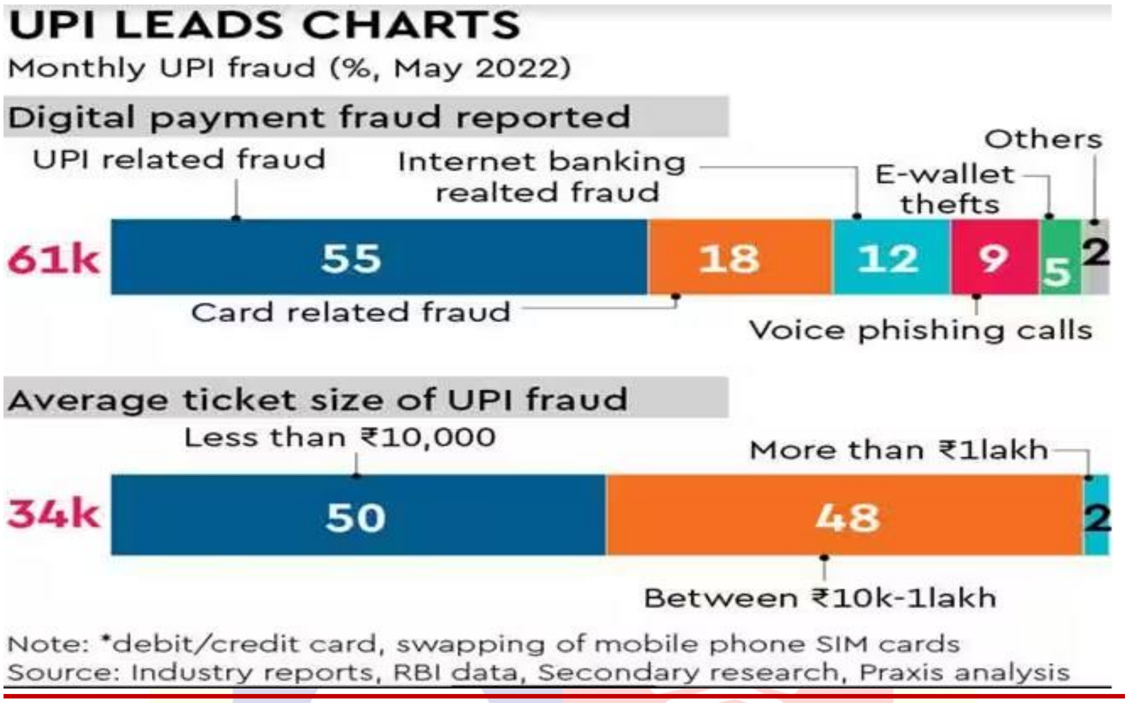


Fig 1 UPI lead Charts (May 2022)

5. Signature Bank was shut down on March 12, 2023- The bank's failure resulted from regulator concern about depositors withdrawing large amounts of money after the failure of Silicon Valley Bank (SVB) and the fear of continued contagion. Federal regulators said Signature Bank customers would get all deposits back, even amounts over $250,000 that are uninsured by the Federal Deposit Insurance Corp. (FDIC). An April 2023 FDIC report blamed Signature's failure on bank mismanagement, a lack of corporate governance, and failure to listen to and respond quickly to the FDIC's recommendations. Signature Bank's failure raised many policy questions around FDIC insurance, and bank and cryptocurrency oversight.

Pump and Dump System- A pump and dump (P&D) scheme is originally a fraudulent behavior aims to make a profit from boosting the price of a stock through recommendations based on false, misleading or greatly exaggerated statements. The essence of a pump and dump scheme is the ability to convince other investorsto buy a stock that is supposedly ready to take off. Because the perpetrators usually have an established position in the stock, they will make huge profits if the scheme succeed. As this practice disturbs the market, it is strictly regulated in the traditional financial market and can lead to heavy fines. However, as it is largely unregulated, the cryptocurrency market has become a paradise for perpetrators of P&D schemes.

Pump-and-Dump 3.0

The cryptocurrency market has become the newest arena for pump-and-dump schemes. The massive gains made by Bitcoin and Ethereum have kindled tremendous interest in cryptocurrencies of every stripe

Identifying Transaction Fraud- Basic distinction must be made

The term Securities Fraud covers a wide range of illegal activities, all of which involve the deception of investors or the manipulation of financial markets.

1 High Yield Investment Fraud- Characterized by promises of high rates of return with little or no risk. May involve various forms of investments (e.g. securities, commodities, real estate, precious metals, etc.) "Too good to be true" investment opportunities. Perpetrators may contact victims by telephone, e-mail, or in person. The offers are generally unsolicited.

2 Ponzi Schemes- Use money collected from new victims to pay the high rates of return promised to earlier investors. Payouts give the impression of a legitimate, money-making enterprise behind the fraudster's story. Investors are the only source of funding.

3 Pyramid Schemes- Victims advance relatively small sums of money in the hope of realizing much larger gains. Gains never materialize because there is no legitimate underlying investment. To participate a particular investment opportunity, victims must first send funds to cover "taxes" or "processing fees."After victims send the "fees," the perpetrators appropriate the funds and never deliver on the investment.

4 Advanced Fee Schemes- Victims advance relatively small sums of money in the hope of realizing much larger gains. Gains never materialize because there is no legitimate underlying investment. There are numerous types of advance fee fraud schemes, including lottery and sweepstakes scams, inheritance scams, loan scams, employment scams, romance scams, and business opportunity scams.

5 Foreign Currency Fraud- A number of investment schemes that involve trading in the foreign exchange market, also known as the forex market, used by traders, brokers, and financial institutions have been deemed by lawmakers to put the stability of the market at risk.

In some cases, these trading schemes are deemed fraudulent, attracting criminal liability for those that use and operate them.

When trading in the foreign exchange market, it is essential for traders to understand the potential risks and liabilities that can arise from trading activities. All traders should be aware of the legal implications of their activities, as well as the potential penalties that can result from any illegal activity**.**

6. Broker Embezzlement- Broker fraud can occur in variety of situations. Common types of broker fraud are: misrepresentation or omission, unauthorized trading, stock manipulation, embezzlement, breach of fiduciary duty, and overconcentration. Embezzlement is a type of fraud in which the embezzler, such as a broker, attains assets lawfully but then uses them for unintended purposes. Some warning signs of broker fraud include: promises of high returns with little or no risk, pressure to invest immediately, unregistered investments, unlicensed investment professionals and firms, complex or secretive strategies, and difficulty receiving payments.

7 Hedge Fund Related Fraud- Hedge fund fraud is a type of investment fraud and is any act of financial misconduct committed by or for the account of a hedge fund. any private, pooled-asset investment fund with a limited number of investors, a fee structure that pays management based on the percentage of funds-under-management plus a percentage of investment gains, and restrictive rules about when and how investors can withdraw their money. Hedge funds are lightly regulated and often use borrowed money (“leverage”) in pursuit of large returns.

8 Late Day Trading- Late-day trading is the illegal practice of recording trades executed after hours as having occurred prior to a mutual fund's calculation of its daily net asset value (NAV). It is normally associated with hedge funds placing orders to buy, or redeem, mutual fund shares after the current period's (usually daily) NAV is officially calculated, but receiving a price, usually more advantageous, based on the prior period's NAV that has already been documented. Late-day trading can dilute the value of a mutual fund's shares, harming long-term investors, and should not be confused with the completely legal and acceptable practice of after-hours trading.

Chapter – 2: Literature survey

2.1 Research paper no 1

- **Title** - **Detecting “Pump & Dump Schemes” on Cryptocurrency Market Using An Improved Apriori Algorithm**
- **Author:** Weili Chen∗†, YueJin Xu∗†, Zibin Zheng∗†, Yuren Zhou∗, Jianxun Eileen Yang‡ and Jing Bian
- **Publisher:** IEEE International Conference on Service-Oriented System Engineering (SOSE) (978-1-7281-1442-2/19/$31.00 ©2019 IEEE DOI 10.1109/SOSE.2019.00050)
- **Publishing Year**: 2019
- **Abstract:** With the popularity of Bitcoin, a cryptocurrency market emerged. However, because of insufficient supervision, the market attracts scams, for example, pump and dump (P&D) scheme, a famous fraudulent behaviour in stock markets, has been found rampant in the market**. To help deal with this issue, as a preliminary study, this paper proposes an improved apriori algorithm to detect user groups which may involve in P&D schemes**. The validity of the algorithm is verified by using the leaked transaction history of Mt. Gox Bitcoin exchange(Mt. Gox was a Tokyo-based cryptocurrency exchange that operated between 2010 and 2014. It was responsible for more than 70% of Bitcoin transactions at its peak.) Furthermore, by exploring some of the detected user groups, many abnormal trading behaviours in the exchange found. These findings provide new insights into the behavior of users in the cryptocurrency market, thus leading to meaningful implications for policymakers, investors, and managers dealing with the cryptocurrency market.
- **Summary and conclusion**

In this paper, they propose an improved **apriori algorithm** to detect user groups who may involve in “pump & dump” schemes. By using the leaked transaction history of the famous

Bitcoin exchange Mt. Gox, we found many user groups which buy or sell at the same time. To further analyse the detected groups, we found many abnormal trading records, i.e., abnormal trading behaviours and trading price. It is believed that some of these abnormal behaviours may result from the accounts controlled by the exchange.

2.2 Research paper 2

- **Title:** Pump-and-Dump Manipulation in Cryptocurrency Markets

- **Author:** Anirudh Dhawan and Talis J. Putnin¸
- **Publisher-** Indian Institute of Management Bangalore, India, 2University of Technology Sydney, Australia, Stockholm School of Economics in Riga, Latvia and 4Digital Finance Co-operative Research Centre, Australia
- **Publishing Year**: **:** 4 August 2022
- **Abstract:** The puzzle of widespread participation in cryptocurrency pump-and dump manipulation schemes. Unlike stock market manipulators, cryptocurrency manipulators openly declare their intentions to pump specific coins, rather than trying

  to deceive investors. Analysing a sample of 355 cases in 6 months, we find strong empirical support for both mechanisms. Pumps generate extreme price distortions of 65% on average, abnormal trading volumes in the millions of dollars, and large wealth transfers between participants.
- **Summary and conclusion-** Cryptocurrencies have given rise to a new form of pump-and-dump manipulation, which is similar in some respects to traditional pump-and-dump manipulation of stocks but completely different in other respects.

  Cryptocurrency pump-and-dump schemes affect welfare in three main ways. First, pumps cause wealth transfers. In aggregate, wealth is transferred from the least sophisticated players.

  (e.g., slow players, gamblers, and overconfident players) to manipulators and more sophisticated players (e.g., fast players).

  Second, cryptocurrency pump-and-dumps, like other forms of market manipulation, cause price distortions that harm price accuracy and informativeness. The price distortions could, in theory, degrade the efficiency of resource allocation.

  Third, widespread manipulation can damage the perceived integrity of cryptocurrency. markets and investor confidence in tokens and tokenization.

- **2.3 Research Paper 3**
- **Title:** Detecting cryptocurrency pump-and-dump frauds using market and social signals
- **Author:** Huy Nghiem *, Goran Muric , Fred Morstatter , Emilio Ferrara
- **Publisher:** University of Southern California, Information Sciences Institute, 4676 Admiralty Way, Marina del Ray, Los Angeles, California, USA
- **Publishing Year**: 2021

- **Abstract:** In this paper, they propose an approach to predict the target cryptocurrency for each pump before its announcement using market and social media signals using Neural Network-based architectures while offering interpretable insights into their black-box nature. Additionally, they construct models that are capable of forecasting the highest price induced by the pump after the cryptocurrency's identity is revealed within 6.1% error margin. They examine the optimal temporal windows and describe the limitations of social data to predict the manipulations in cryptocurrency trade. Their experimental results serve as proof of a feasible forecasting expert system for identifying cryptocurrency pump-and-dump frauds using publicly available data.
- **Summary and conclusion -** Baseline model- Logistic Regression model with adjusted class weights to accommodate the heavy imbalanced nature of data.

  CNN models: Convolution Neural Networks (CNN) architecture utilizes a set of kernels (filters of learnable weights), which convolve over the inputs and extract features in the process. CNNs have made significant contributions to the fields of Computer Vision.

  BLSTM models- Long Short Term Memory (LSTM), a special kind of Recurrent Neural Network. LSTM's building blocks are cells capable of identifying hidden dependencies from previous inputs.

  CLSTM models appears to be the most promising in terms of both performance and consistency. This architecture reliably produce MAPE comparable to baseline for Overall

  instances and lower MAPE for Pump instances across all configurations.

- **2.4 Research Paper 4**
- **Title:** Explainable AI-Driven Financial Transaction Fraud Detection using Machine Learning and Deep Neural Networks
- **Author:** Chaithanya Vamshi Saia , Debashish Dasa, *, Nouh Elmitwallya , Ogerta Elezaja Md Baharul Islamb
- **Publisher-** Department of Computing & Data Science, Faculty of Computing, Engineering & the Built Environment, Birmingham City University, United Kingdom College of Data Science and Engineering, American University of Malta, Malta Corresponding Author
- **Publishing Year: :** 18 May 2023
- **Abstract:** Artificial Intelligence (AI) is revolutionizing and transforming data-driven businesses like financial services, e-commerce, and banking. AI technology and the digitization of banking are rapidly increasing the volume of financial transaction frauds,

causing billions in revenue losses and unrecoverable funds. The primary business challenge faced by the finance and banking industry is understanding how their AI systems make predictions and explaining why a transaction is flagged as fraud. A solution to the “black box" challenge is known as Explainable Artificial Intelligence (XAI). XAI is a human-centric AI that accelerates business value through a greater understanding of predictions and makes complex black box decisions more transparent, trustable, explainable, and understandable to human users for better business decision-making without compromising prediction accuracy.

The outcome of the research can be applied to design Proof of Concept (POC) as a front-end web application and deployed in the cloud using Python and Streamlit to make it accessible to non-AI expert consumers like business managers, domain experts, business analysts, and internal stakeholders for better decision-making and generating business value through real-time predictions.

- **Summary and conclusion:** This research tackled and filled the gap of the lack of explain ability by implementing an Explainable AI (XAI)-driven Interface for Financial Transaction Fraud Detection using Machine Learning (ML) & Deep Neural Networks (DNN) The research contributed to both Explainable AI (XAI) research community and practical business value to solve fraud detection business problems in the financial services and banking industry by implementing five cutting-edge Explainable AI (XAI) methods for enhancing the model interpretability or explain ability of current black-box fraud detection systems used in the Finance and Banking industry. Therefore, XAI-driven financial transaction fraud detection system decisions can be operationalized into production with greater confidence and transparency.

- **2.5 Research paper 5 (Case Study-1)**
- **Title:** A Comparative Analysis - Credit card fraud detection using Machine Learning
- Techniques.
- **Author:** John O. Awoyemi, Adebayo O. Adetunmbi, Samuel A. Oluwadare
- **Publisher:** Department of Computer Science Federal University of Technology Akure Akure, Nigeria
- **Publishing Year**: 2017
- **Abstract:** Financial fraud is an ever growing menace with far consequences in the financial industry The performance of fraud detection in credit card transactions is

greatly affected by the sampling approach on dataset, selection of variables and detection technique(s) used.

This paper investigates the performance of naïve bayes, k-nearest neighbor and logistic regression on highly skewed credit card fraud data**. The results shows of optimal accuracy for naïve bayes, k-nearest neighbor and logistic regression classifiers are 97.92%, 97.69% and 54.86% respectively.**

- **Summary and conclusion**

  This paper investigates the comparative performance of Naïve Bayes, K-nearest neighbor and Logistic regression models in binary classification of imbalanced credit card fraud data. The rationale for investigating these three techniques is due to less comparison they have attracted in past literature. However, a subsequent study to compare other single and ensemble techniques using our approach is underway.

- **2.6 Research Paper 6 (Case study 2)**
- **Title:** Fraud Detection using Machine Learning and Deep Learning
- **Author:** Pradheepan Raghavan, Neamat El GayarSchool of Mathematical and Computer Sciences Heriot Watt University Dubai, UAE
- **Publisher: 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)**
- **Publishing Year: December 11-12-2019,**
- **Abstract:** Frauds are known to be dynamic and have no patterns, hence they are not easy to identify. This paper aims to benchmark multiple machine learning methods such as k-nearest neighbour (KNN), random forest and support vector machines (SVM), while the deep learning methods such as autoencoders, convolutional neural networks (CNN), restricted Boltzmann machine (RBM) and deep belief networks (DBN).
- **Summary and conclusion:** paper provides an empirical investigation comparing various machine learning and deep learning models on different data sets for the detection of fraudulent transaction. The main aim of this study is to find insights of which methods would best suitable for which type of datasets. As nowadays, many companies are investing in new techniques to improve their business this paper could potentially help practitioners and companies to better understand how different methods work on certain types of datasets**.**

Our study reveals that to detect fraud, the best methods with larger datasets would be using SVMs, potentially combined with CNNs to get a more reliable performance. For the smaller datasets, ensemble approaches of SVM, Random Forest and KNNs can provide good enhancements. Convolutional Neural Networks (CNN) usually, outperforms other deep learning methods such as Autoencoders, RBM and DBN. A limitation of this study is however that it only deals with detecting fraud in a supervised learning context.

- **2.7 Research Paper 7 (Case study 3)**
- **Title:** The advanced proprietary AI/ML solution as Anti-fraud- Tensorlink4cheque (AFTL4C) for Cheque fraud detection**.**
- **Author:** PRABHAKARA R UYYALA, Dr. DYAN CHANDRA YADAV,
- **Publisher: CSE DEPARTMENT, JS UNIVERSITY**
- **Publishing Year:** 2023
- **Abstract:** The Anti-fraud-Tensorlink4cheque (AFTL4C) is a new AI/ML-based solution developed to detect cheque fraud in real-time. Cheque fraud has become a major issue for financial institutions, resulting in millions of losses due to fraudulent activities. The AFTL4C solution uses Generative Adversarial Network (GAN) technique to compare various factors on scanned cheque images to identify potential counterfeits in real-time**.**
- **Summary and conclusion -** The AFTL4C solution was developed to detect cheque fraud in real-time using a Generative Adversarial Network (GAN) approach to compare various factors on scanned cheque images. The solution assigns a confidence point to each object of the scanned cheque image, which can be good, fraudulent, or requiring further review. The model can process up to 1,800 checks per second, with end-to-end response times of less than 62 milliseconds.

Facts and Figure

[1] First Republic was among the regional bank failures of early 2023-(https://www.investopedia.com/what-happened-to-first-republic-bank-7489214#:~:text=First%20Republic%20Bank%20was%20among,the%20broader%20interest%20rate%20environment.)

[2]Why Are Adani Shares Falling?- 2023 has been a tumultuous year for the Adani Group, which faced investor fury at the back of allegations of misgovernance and corporate fraud by U.S.-based short-seller group Hindenburg Research. (https://www.forbes.com/advisor/in/investing/why-adani-shares-are-falling/)

[3] Why SVB’s collapse is especially hard as a Black founder, says CEO (https://www.cnbc.com/2023/03/24/svb-collapse-squad-app-ceo-says-silicon-valley-bank-failure-was-agony.html)

[4] Banks’ digital payments frauds nearly double in FY23 While 3,596 frauds amounting to 155 crore using cards and internet banking services were reported in FY22, the volume nearly doubled to 6,659 digital frauds in FY23 amounting to276 crore. (https://www.financialexpress.com/industry/banking-finance/banks-digital-payments-frauds-nearly-double-in-fy23/3113522/#:~:text=While%203%2C596%20frauds%20amounting%20to,in%20FY23%20amounting%20to276%20crore.)

[5] Signature Bank becomes next casualty of banking turmoil after SVB (https://www.reuters.com/business/finance/new-york-state-regulators-close-signature-bank-2023-03-12/)

Chapter – 3: Problem statements

Pump-and-dump is a manipulative scheme that attempts to boost the price of a stock or security through fake recommendations. These recommendations are based on false, misleading, or greatly exaggerated statements. The perpetrators of a pump-and-dump scheme already have an established position in the company's stock and will sell their parts after the hype has led to a higher share price.

In this Project We will Focus on the pump and dump system (2.0 and 3.0) we will observe the financial model working and mechanism and possible mitigation measures to avoid these practices

Pump-and-Dump 2.0

The same scheme can be perpetrated by anyone with access to an online trading account and the ability to convince other investors to buy a stock that is supposedly "ready to take off.“

Pump-and-Dump 3.0

The cryptocurrency market has become the newest arena for pump-and-dump schemes. The massive gains made by Bitcoin and Ethereum have kindled tremendous interest in cryptocurrencies of every stripe.

Chapter – 4: Component / tools used

VS code complier- May 2023 (version 1.79)

Update 1.79.1: The update addresses this security issue. The update addresses these

Python- Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together**.**

Python 3.11 is between 10-60% faster than Python 3.10. On average, we measured a 1.25x speedup on the standard benchmark suite.

Libraries-

Pandas (2.0.3) – helps to load the data frame in a 2D array format and has multiple functions to perform analysis tasks in one go.

NumPy(1.25.0) – NumPy arrays are very fast and can perform large computations in a very short time. NumPy offers comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and more.

Matplotlib(3.7.1)/Seaborn (0.12.0) – Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics.

Sklearn(1.3.0) –is a Python module for machine learning built on top of SciPy and is distributed under the 3-Clause BSD license. This module contains multiple libraries having pre-implemented functions to perform tasks from data preprocessing to model development and evaluation.

XGBoost (1.7.6) – This contains the eXtreme Gradient Boosting machine learning algorithm which is one of the algorithms which helps us to achieve high accuracy on predictions.

Machine learning- Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, and to uncover key insights in data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics.

Financial Forecasting- Financial forecasting is the process of estimating or predicting how a business or project will perform financially in the future. Financial forecasting is a crucial element of financial planning as it estimates important financial metrics such as sales, income, and future revenue, which are necessary for finance-related operations such as budgeting and capital budgeting. A financial forecast is an estimate of future financial outcomes for a company or project, usually applied in budgeting, capital budgeting, and/or valuation. The most common type of financial forecast is an income statement; however, a complete financial model should include forecasts for all three financial statements: income statement, balance sheet, and cash flow statement.

Predictive analysis- a subset of advanced analytics that utilizes historical data, statistical modelling, data mining techniques, and machine learning to make predictions about future events or outcomes. It involves analyzing current and past data to identify patterns and trends, which are then used to build models that can predict future behaviors or events. Predictive analytics is widely used in business applications to identify opportunities and mitigate risks. Some common applications of predictive analytics include fraud detection, customer segmentation, and demand forecasting.

Chapter – 5: Flowchart

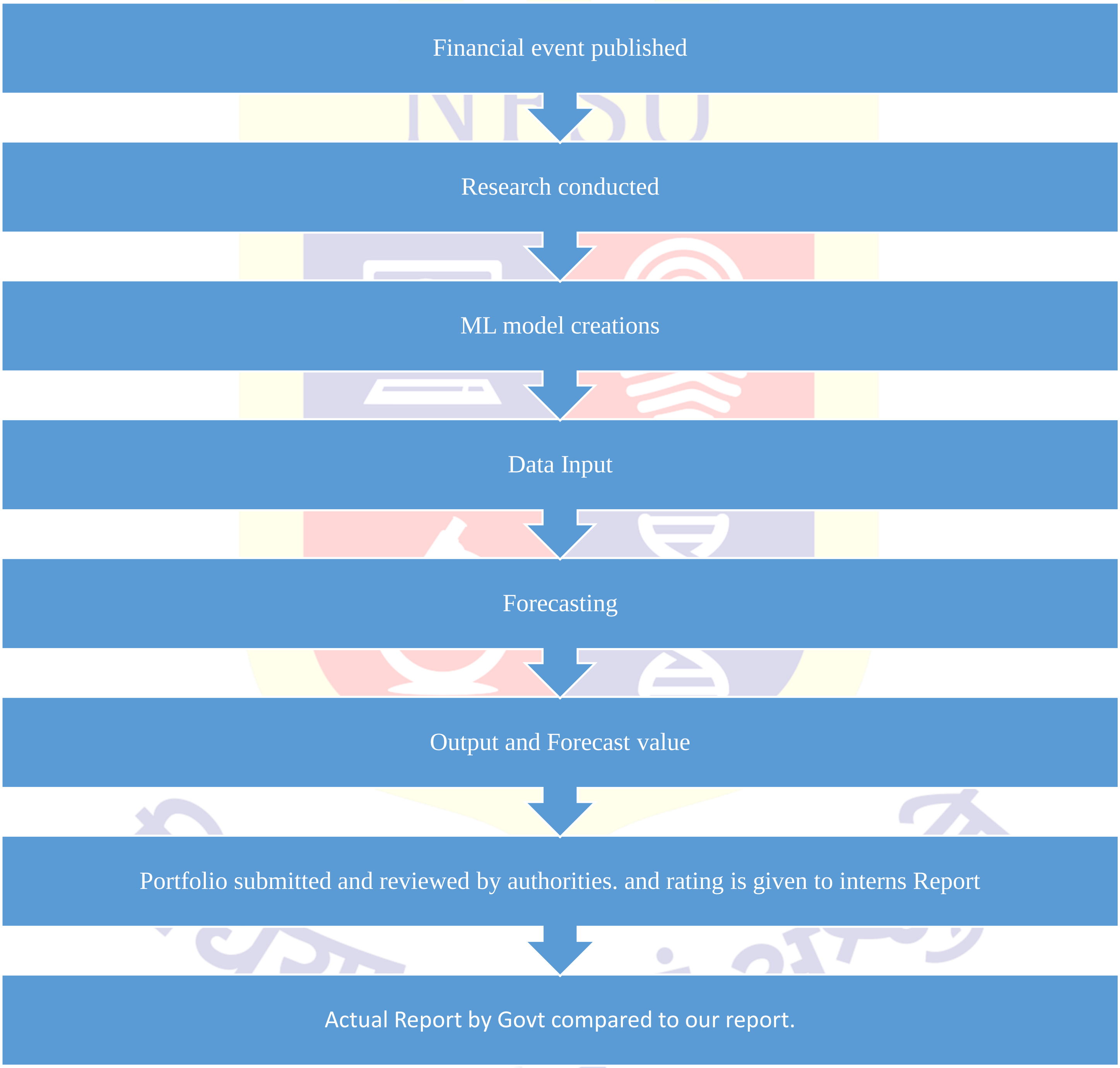
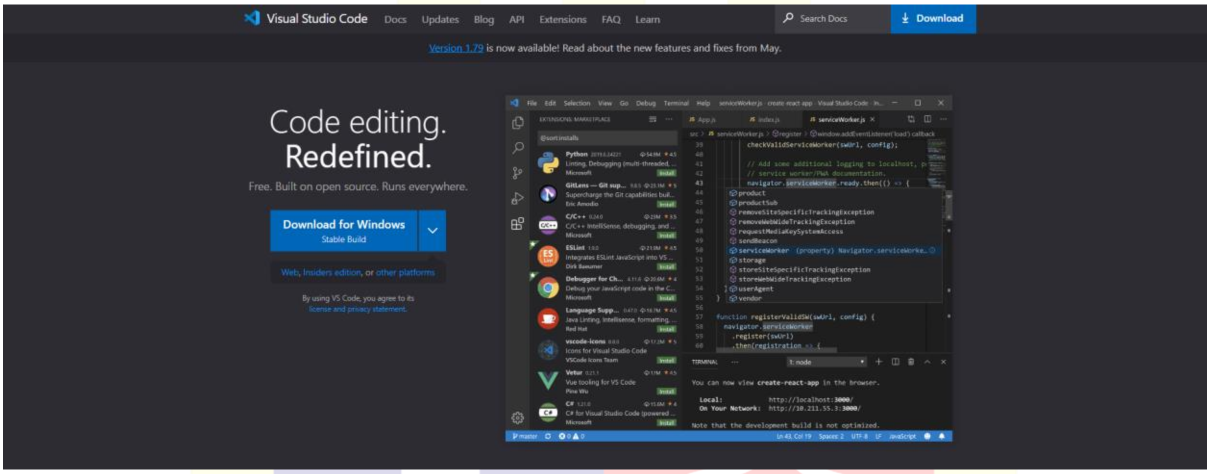

Financial event published
Research conducted
ML model creations
Data Input
Forecasting
Output and Forecast value
Portfolio submitted and reviewed by authorities. and rating is given to interns Report
Actual Report by Govt compared to our report.

Fig 2 Flow chart of complete Model.

Chapter – 6: Implementation & result (Snap Shot)

Stage 1

Install VS code (https://code.visualstudio.com/)



Version 1.79 is now available! Read about the new features and fixes from May.

Code editing.
Redefined.

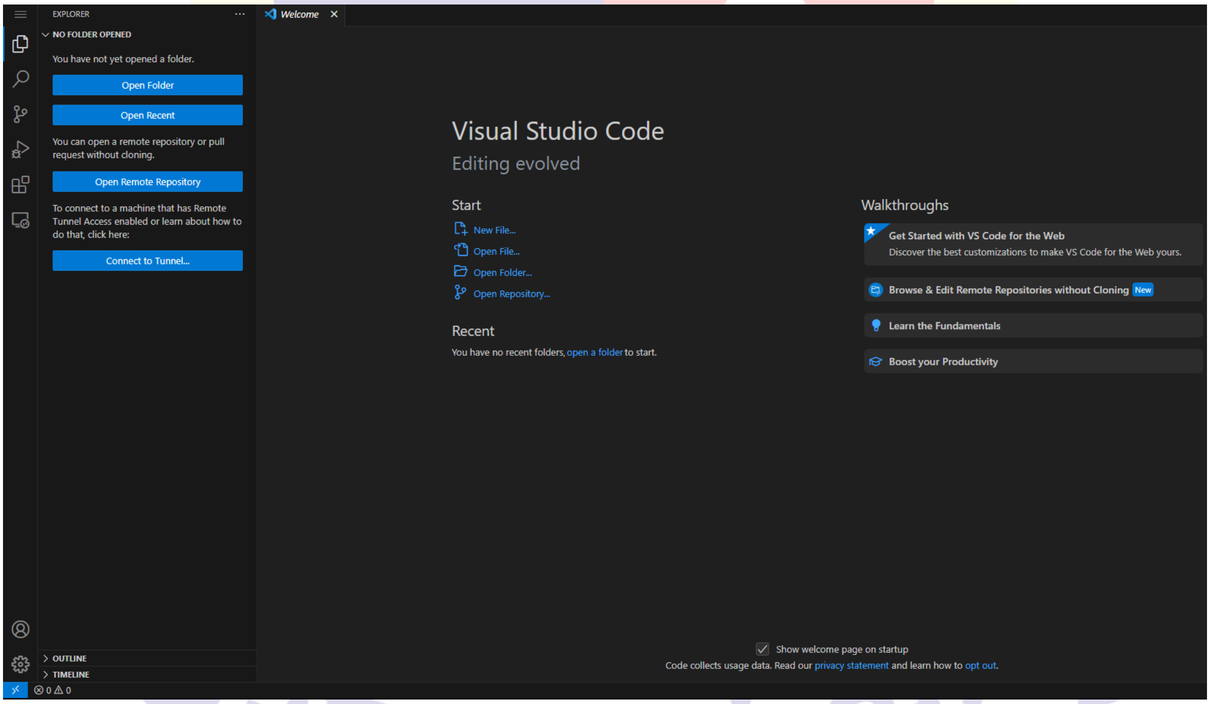
Free. Built on open source. Runs everywhere.

By using VS Code, you agree to its license and privacy statement.

```
checkValidServiceWorker(swUrl, config);

// Add some additional logging to localhost, p
// service worker/PWA documentation.
navigator.serviceWorker.ready.then(() => {
  product
  productSub
  removeSiteSpecificTrackingException
  removeWebWideTrackingException
  requestMediaKeySystemAccess
  sendBeacon
  serviceWorker  (property) Navigator.serviceWorke…
  storage
  storeSiteSpecificTrackingException
  storeWebWideTrackingException
  userAgent
  vendor
}

function registerValidSW(swUrl, config) {
  navigator.serviceWorker
    .register(swUrl)
    .then(registration => {
```

```
You can now view create-react-app in the browser.

  Local:            http://localhost:3000/
  On Your Network:  http://10.211.55.3:3000/

Note that the development build is not optimized.
```

Fig 3 (a) VS code compiler



You have not yet opened a folder.

You can open a remote repository or pull request without cloning.

To connect to a machine that has Remote Tunnel Access enabled or learn about how to do that, click here:

Visual Studio Code

Editing evolved

Start

Recent

You have no recent folders, open a folder to start.

Walkthroughs

Get Started with VS Code for the Web

Discover the best customizations to make VS Code for the Web yours.

Browse & Edit Remote Repositories without Cloning

Learn the Fundamentals

Boost your Productivity

☑ Show welcome page on startup

Code collects usage data. Read our privacy statement and learn how to opt out.

Fig 3 (b) VS Interface

Importing some libraries which will be used for various purpose.

Python libraries make it very easy for us to handle the data and perform typical and complex tasks with a single line of code.

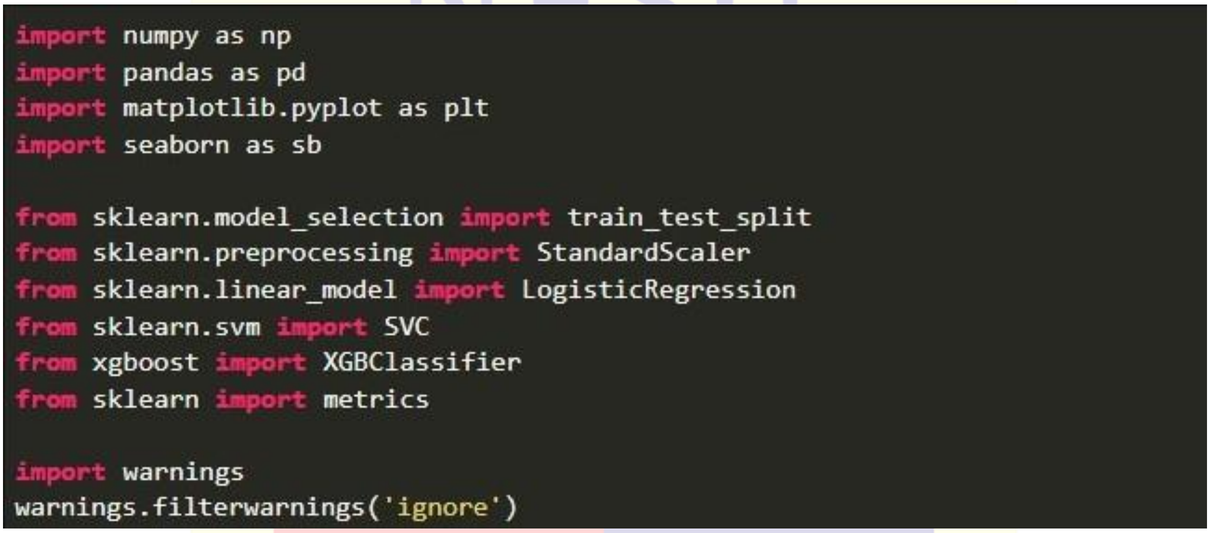
Pandas – This library helps to load the data frame in a 2D array format and has multiple functions to perform analysis tasks in one go.

Numpy – Numpy arrays are very fast and can perform large computations in a very short time.

Matplotlib/Seaborn – This library is used to draw visualizations.

Sklearn – This module contains multiple libraries having pre-implemented functions to perform tasks from data preprocessing to model development and evaluation.

XGBoost – This contains the eXtreme Gradient Boosting machine learning algorithm which is one of the algorithms which helps us to achieve high accuracy on predictions.



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sb

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from xgboost import XGBClassifier
from sklearn import metrics

import warnings
warnings.filterwarnings('ignore')
```

Fig 4 (a) Libraries in python

Importing Dataset

The dataset we will use here to perform the analysis and build a predictive model is Tesla Stock Price data.

OHLC(‘Open’, ‘High’, ‘Low’, ‘Close’) data from 1st January 2010 to 31st December 2017 which is for 8 years for the Tesla stocks.

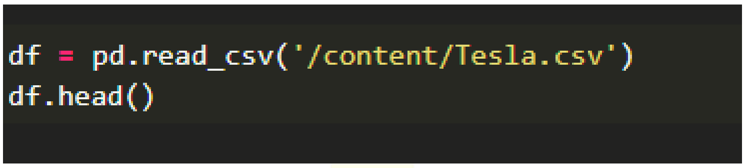
the CSV file from: https://www.kaggle.com/datasets/timoboz/tesla-stock-data-from-2010-to-2020

About Dataset

Context- TSLA has been on the rice recently, with a crazy +100% spike in the last few months.

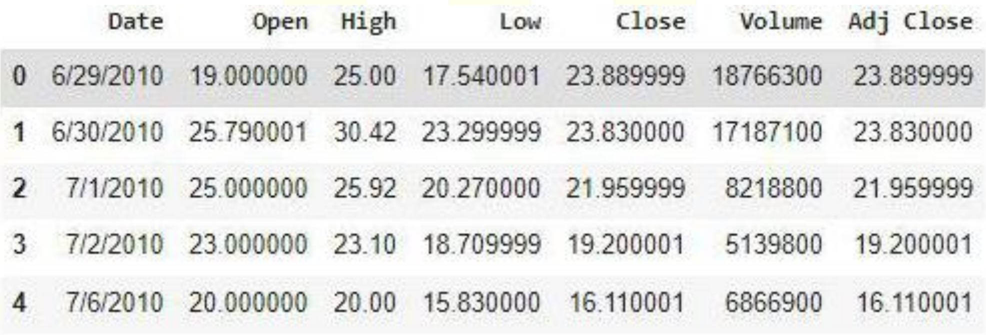
Content- EOD data for Tesla's stock from 2010 to 2020.

Data from https://finance.yahoo.com/quote/TSLA



```
df = pd.read_csv('/content/Tesla.csv')
df.head()
```

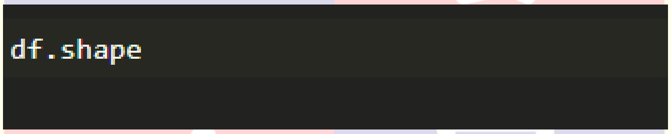
Fig 4 (b) input code for loading Dataset.



	Date	Open	High	Low	Close	Volume	Adj Close
0	6/29/2010	19.000000	25.00	17.540001	23.889999	18766300	23.889999
1	6/30/2010	25.790001	30.42	23.299999	23.830000	17187100	23.830000
2	7/1/2010	25.000000	25.92	20.270000	21.959999	8218800	21.959999
3	7/2/2010	23.000000	23.10	18.709999	19.200001	5139800	19.200001
4	7/6/2010	20.000000	20.00	15.830000	16.110001	6866900	16.110001

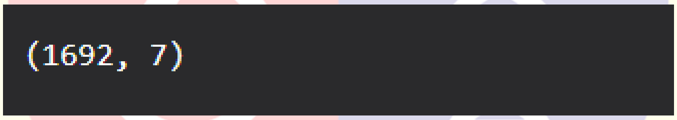
Fig 4 (c) code output code for loading Dataset.

From the first five rows, we can see that data for some of the dates is missing the reason for that is on weekends and holidays Stock Market remains closed hence no trading happens on those days.



```
df.shape
```

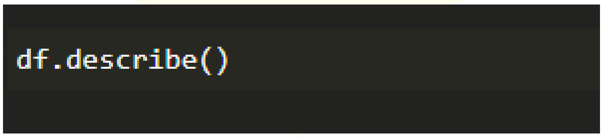
Fig 5 (a) shape of dataset.



```
(1692, 7)
```

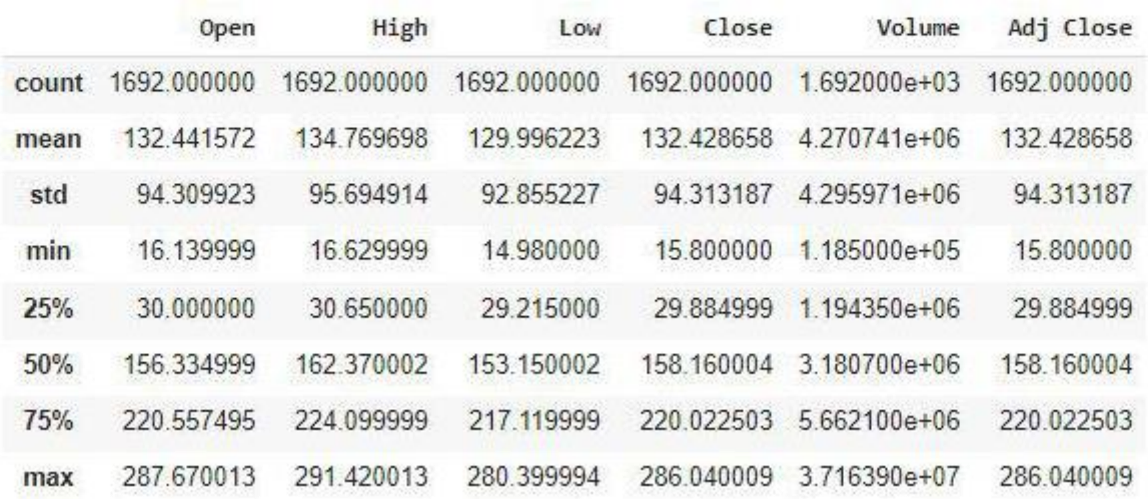
Fig 5 (b) Output.

we got to know that there are 1692 rows of data available and for each row, we have 7 different features or columns.



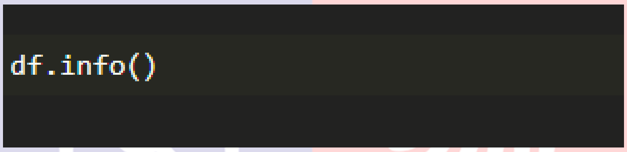
```
df.describe()
```

Fig 5 (c) Input code for data set description.



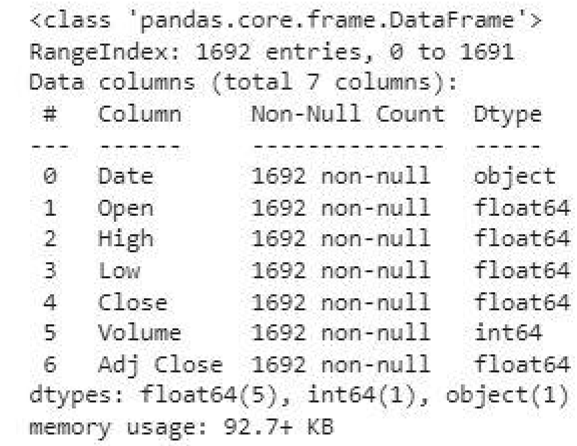
	Open	High	Low	Close	Volume	Adj Close
count	1692.000000	1692.000000	1692.000000	1692.000000	1.692000e+03	1692.000000
mean	132.441572	134.769698	129.996223	132.428658	4.270741e+06	132.428658
std	94.309923	95.694914	92.855227	94.313187	4.295971e+06	94.313187
min	16.139999	16.629999	14.980000	15.800000	1.185000e+05	15.800000
25%	30.000000	30.650000	29.215000	29.884999	1.194350e+06	29.884999
50%	156.334999	162.370002	153.150002	158.160004	3.180700e+06	158.160004
75%	220.557495	224.099999	217.119999	220.022503	5.662100e+06	220.022503
max	287.670013	291.420013	280.399994	286.040009	3.716390e+07	286.040009

Fig 5 (d) Output for description code.



```
df.info()
```

Fig 6 (a) input for information.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1692 entries, 0 to 1691
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   Date       1692 non-null   object 
 1   Open       1692 non-null   float64
 2   High       1692 non-null   float64
 3   Low        1692 non-null   float64
 4   Close      1692 non-null   float64
 5   Volume     1692 non-null   int64  
 6   Adj Close  1692 non-null   float64
dtypes: float64(5), int64(1), object(1)
memory usage: 92.7+ KB
```

Fig 6(b) Output for information.

Exploratory Data Analysis

Exploratory data analysis (EDA) is used by data scientists to analyse and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modelling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. **Originally developed by American mathematician John Tukey in the 1970s**, EDA techniques continue to be a widely used method in the data discovery process today.

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning.

There are four primary types of EDA:

Univariate non-graphical- This is simplest form of data analysis, where the data being analysed consists of just one variable. Since it’s a single variable, it doesn’t deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

Univariate graphical- Non-graphical methods don’t provide a full picture of the data. Graphical methods are therefore required. Common types of univariate graphics include:

Stem-and-leaf plots, which show all data values and the shape of the distribution.

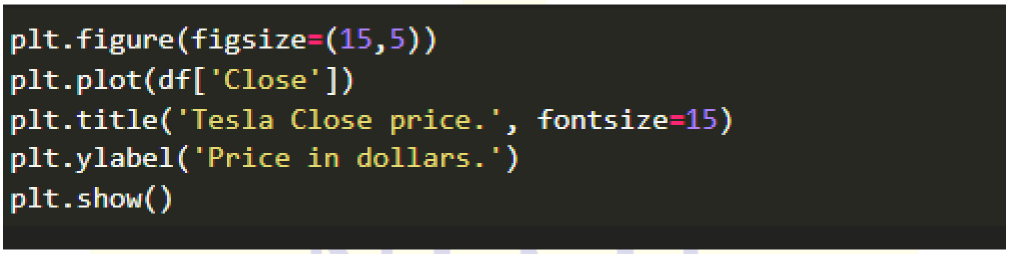
Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values.

Box plots, which graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.

Multivariate nongraphical: Multivariate data arises from more than one variable. Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation or statistics.

Multivariate graphical: Multivariate data uses graphics to display relationships between two or more sets of data. The most used graphic is a grouped bar plot or bar chart with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.

While performing the EDA of the Tesla Stock Price data we will analyze how prices of the stock have moved over the period of time and how the end of the quarters affects the prices of the stock.



```
plt.figure(figsize=(15,5))
plt.plot(df['Close'])
plt.title('Tesla Close price.', fontsize=15)
plt.ylabel('Price in dollars.')
plt.show()
```

Fig 7 (a) Input code of Graph.

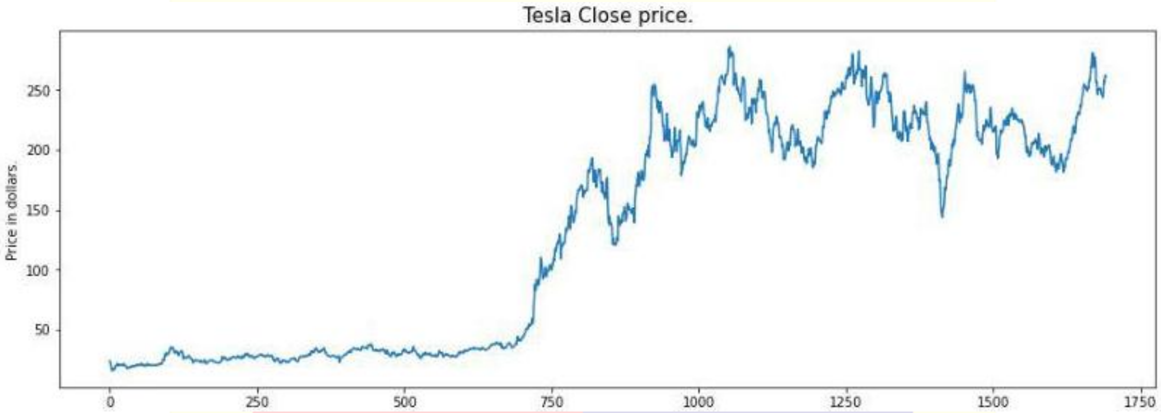
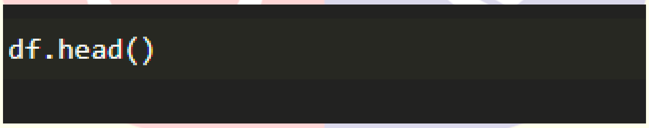


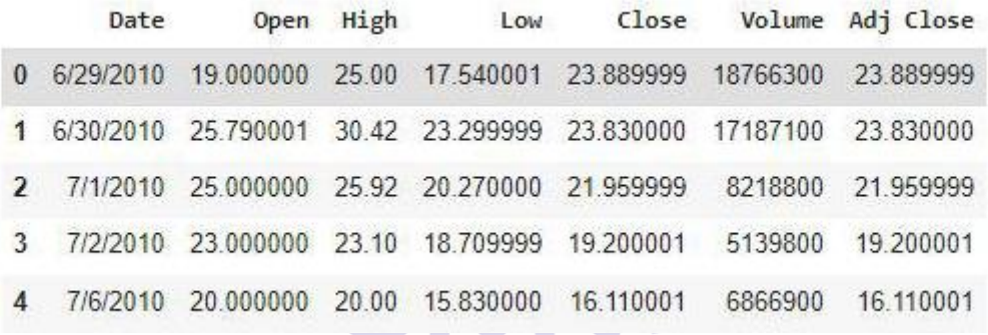
Fig 7 (b) Tesla close Price.

The prices of tesla stocks are showing an upward trend as depicted by the plot of the closing price of the stocks.



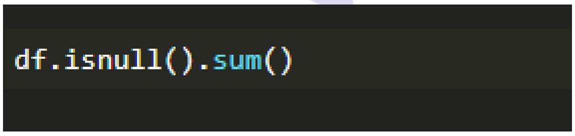
```
df.head()
```

Fig 8 (a) input code close/open values.



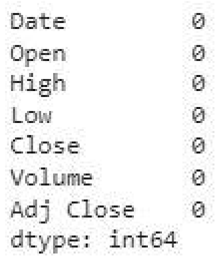
	Date	Open	High	Low	Close	Volume	Adj Close
0	6/29/2010	19.000000	25.00	17.540001	23.889999	18766300	23.889999
1	6/30/2010	25.790001	30.42	23.299999	23.830000	17187100	23.830000
2	7/1/2010	25.000000	25.92	20.270000	21.959999	8218800	21.959999
3	7/2/2010	23.000000	23.10	18.709999	19.200001	5139800	19.200001
4	7/6/2010	20.000000	20.00	15.830000	16.110001	6866900	16.110001

Fig 8 (b) open/close values.



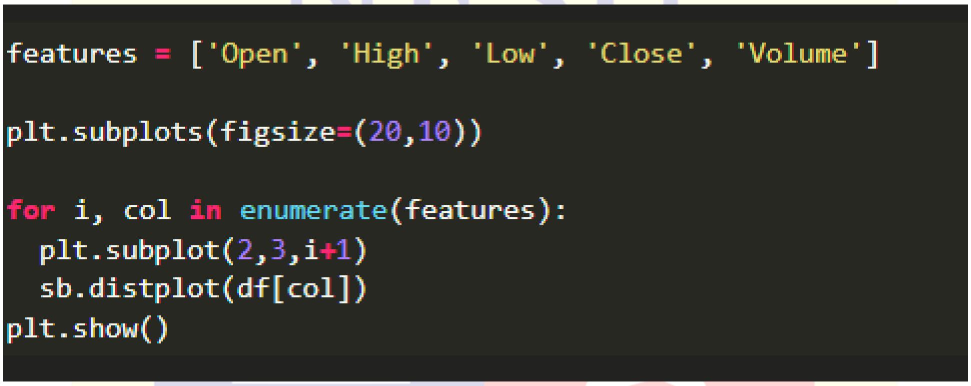
```
df.isnull().sum()
```

Fig 9 (a) check Null values.



```
Date         0
Open         0
High         0
Low          0
Close        0
Volume       0
Adj Close    0
dtype: int64
```

Fig 9 (b) no null values detected.



```
features = ['Open', 'High', 'Low', 'Close', 'Volume']

plt.subplots(figsize=(20,10))

for i, col in enumerate(features):
  plt.subplot(2,3,i+1)
  sb.distplot(df[col])
plt.show()
```

Fig 10 (a) Plotting of graphs.

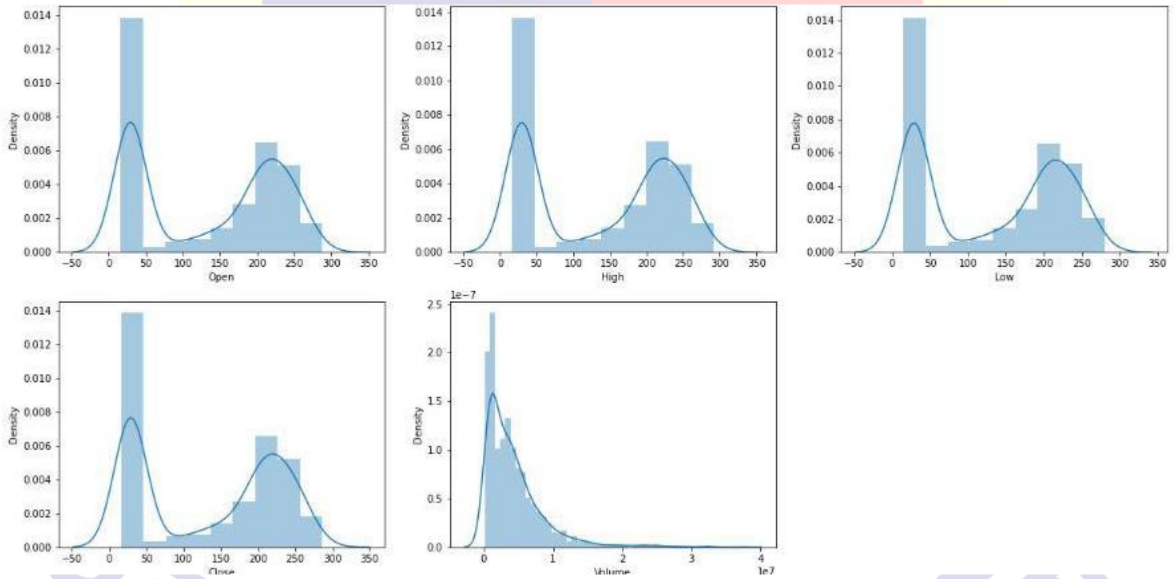
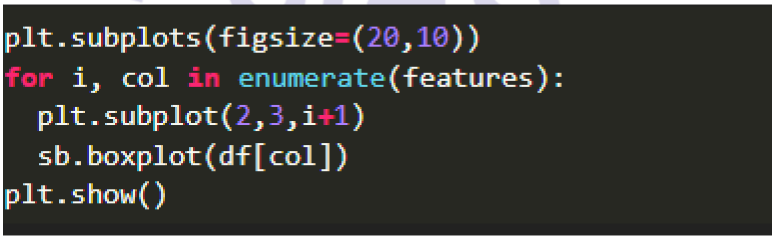


Fig 10 (b) Distribution Plot of the Continuous Variable.

In the distribution plot of OHLC data, we can see two peaks which means the data has varied significantly in two regions. And the Volume data is left-skewed.



```
plt.subplots(figsize=(20,10))
for i, col in enumerate(features):
  plt.subplot(2,3,i+1)
  sb.boxplot(df[col])
plt.show()
```

Fig 11 (a) Boxplot Code.

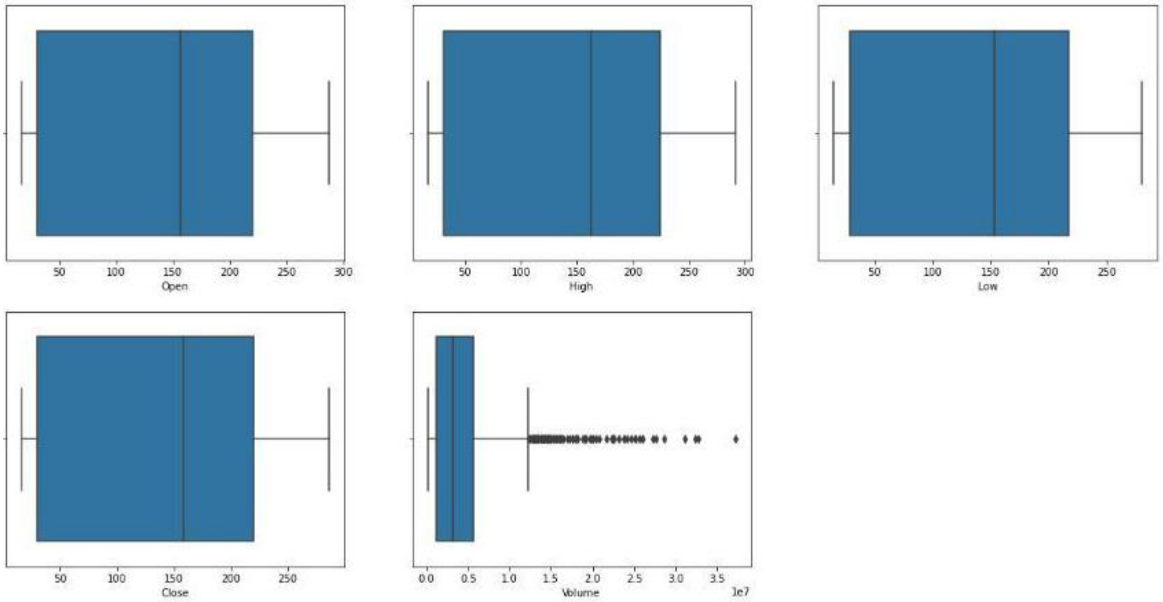
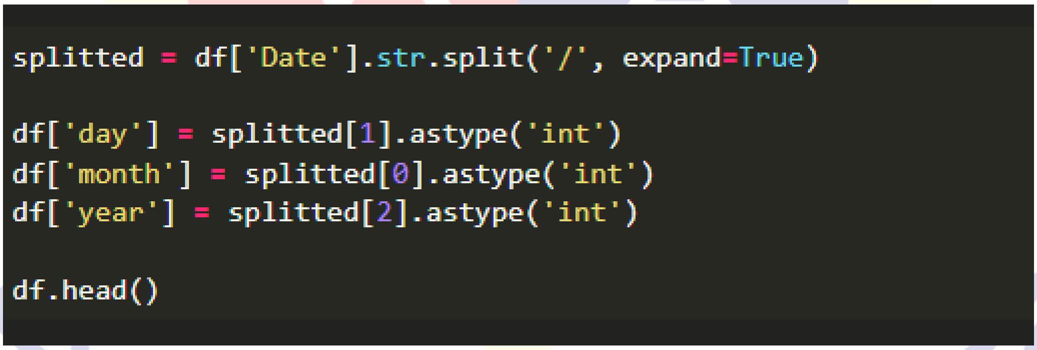


Fig 11 (b) Box Plot of the Continuous Variable.

From the above boxplots, we can conclude that only volume data contains outliers in it but the data in the rest of the columns are free from any outlier.

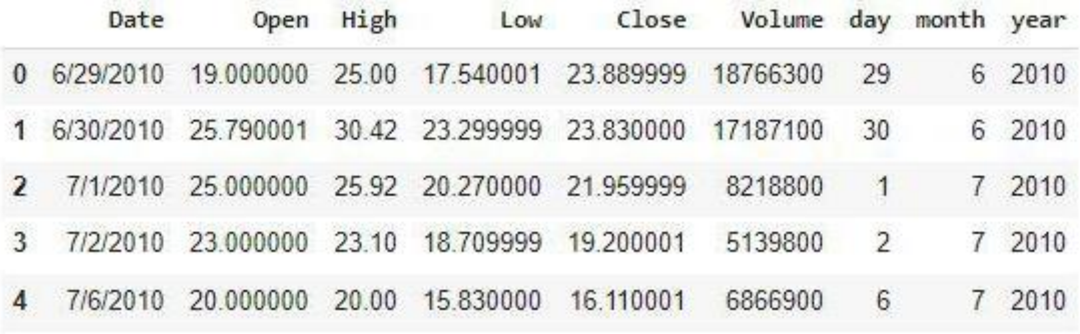
Feature Engineering

Feature Engineering helps to derive some valuable features from the existing ones. These extra features sometimes help in increasing the performance of the model significantly and certainly help to gain deeper insights into the data.



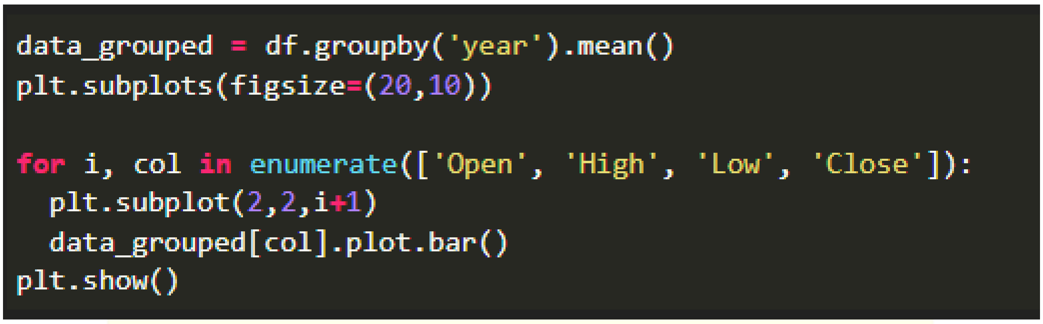
```
splitted = df['Date'].str.split('/', expand=True)

df['day'] = splitted[1].astype('int')
df['month'] = splitted[0].astype('int')
df['year'] = splitted[2].astype('int')

df.head()
```

Fig 12 (a) expanding date format.



	Date	Open	High	Low	Close	Volume	day	month	year
0	6/29/2010	19.000000	25.00	17.540001	23.889999	18766300	29	6	2010
1	6/30/2010	25.790001	30.42	23.299999	23.830000	17187100	30	6	2010
2	7/1/2010	25.000000	25.92	20.270000	21.959999	8218800	1	7	2010
3	7/2/2010	23.000000	23.10	18.709999	19.200001	5139800	2	7	2010
4	7/6/2010	20.000000	20.00	15.830000	16.110001	6866900	6	7	2010

Fig 12 (b) expanding date format output.



```
data_grouped = df.groupby('year').mean()
plt.subplots(figsize=(20,10))

for i, col in enumerate(['Open', 'High', 'Low', 'Close']):
  plt.subplot(2,2,i+1)
  data_grouped[col].plot.bar()
plt.show()
```

Fig 13 (a) Input for Bar graph.

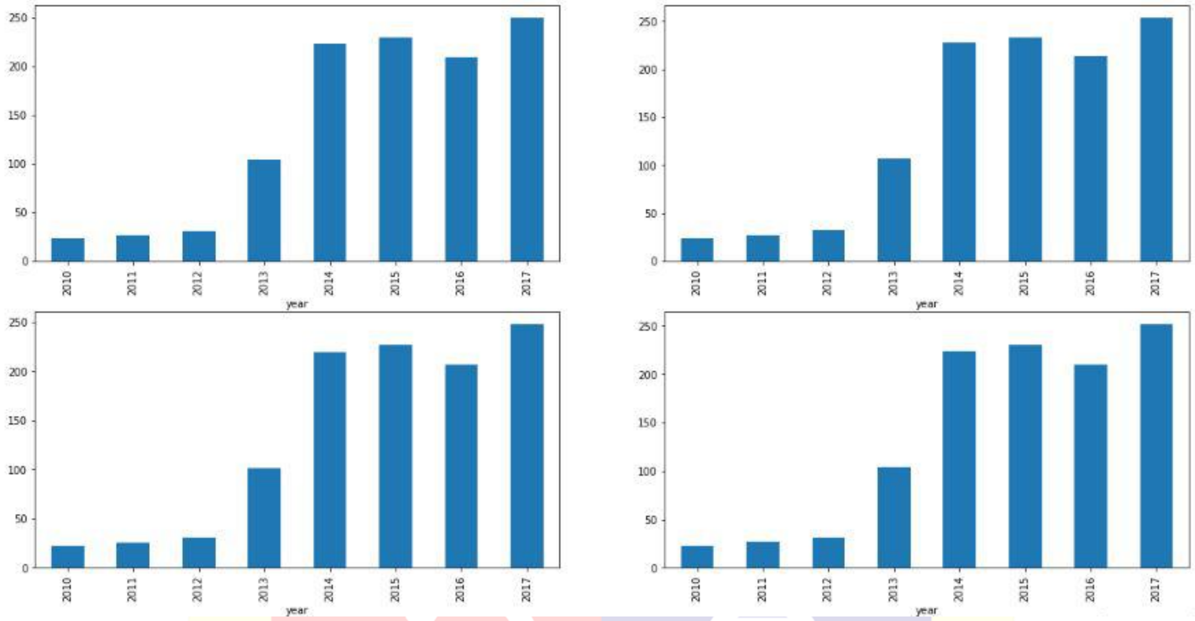
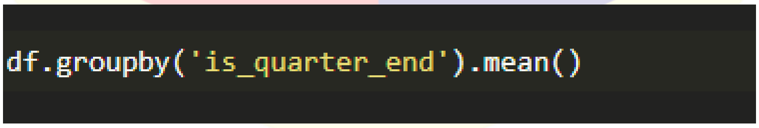


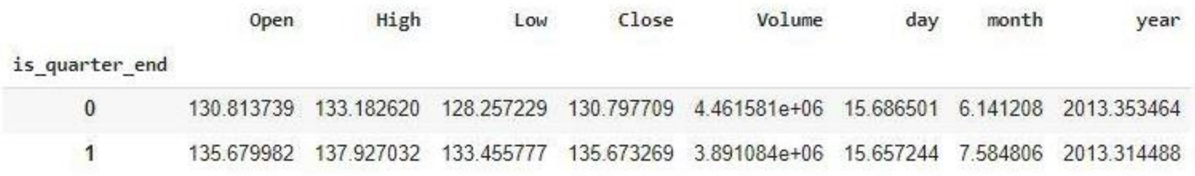
Fig 13 (b) code output for bar graph.

From the above bar graph, we can conclude that the stock prices have doubled from the year 2013 to that in 2014.



```
df.groupby('is_quarter_end').mean()
```

Fig 14 (a) mean of quarter end.

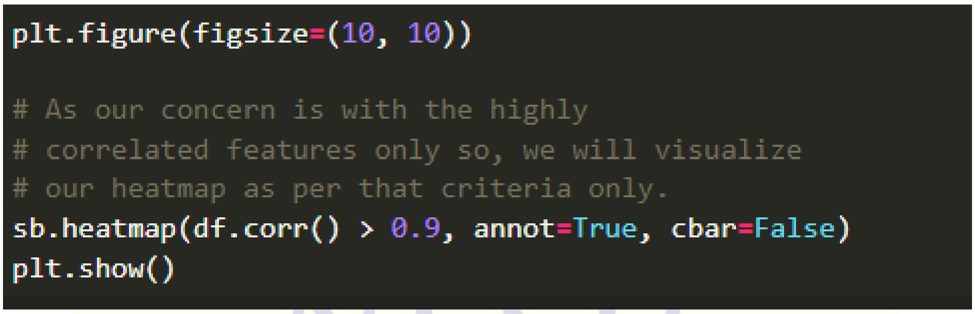


```
                      Open        High         Low       Close        Volume        day     month          year
is_quarter_end
             0  130.813739  133.182620  128.257229  130.797709  4.461581e+06  15.686501  6.141208  2013.353464
             1  135.679982  137.927032  133.455777  135.673269  3.891084e+06  15.657244  7.584806  2013.314488
```

Fig 14 (b) Output mean of quarter end.

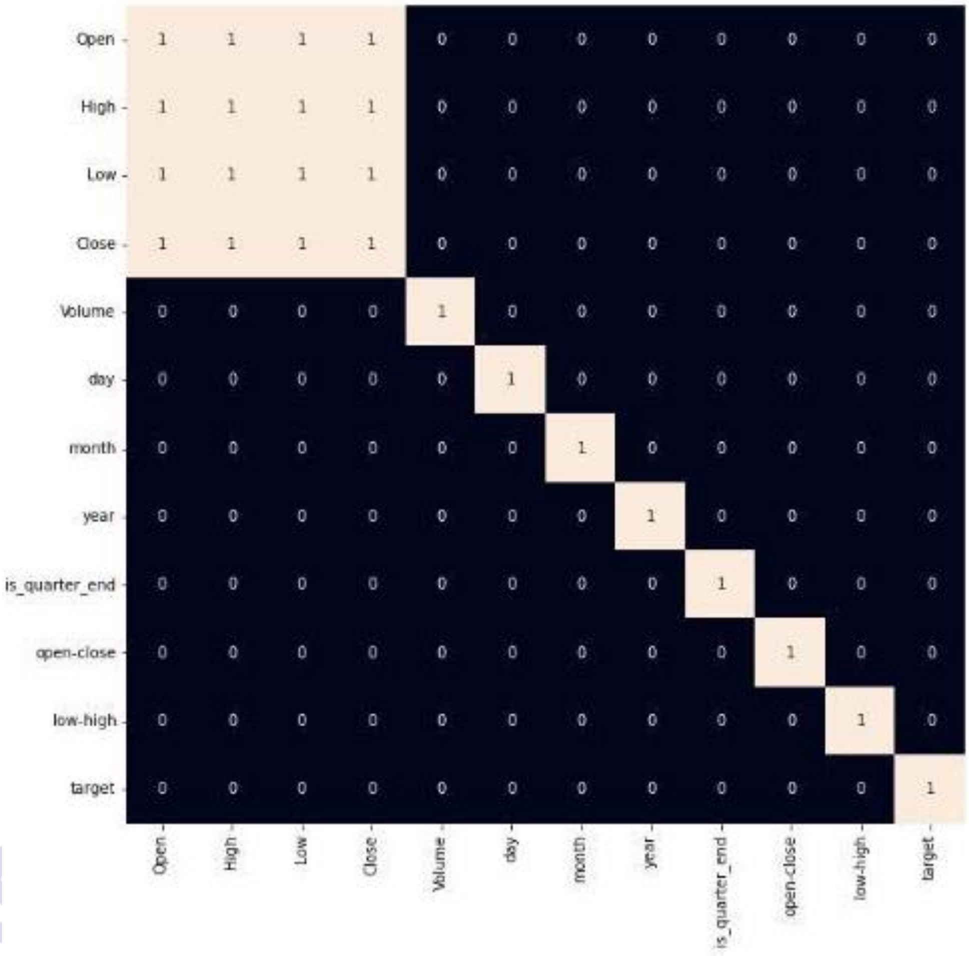
Prices are higher in the months which are quarter end as compared to that of the non-quarter end months. The volume of trades is lower in the months which are quarter end.

When we add features to our dataset we have to ensure that there are no highly correlated features as they do not help in the learning process of the algorithm.



```
plt.figure(figsize=(10, 10))

# As our concern is with the highly
# correlated features only so, we will visualize
# our heatmap as per that criteria only.
sb.heatmap(df.corr() > 0.9, annot=True, cbar=False)
plt.show()
```

Fig 15 (a) Heat map code.



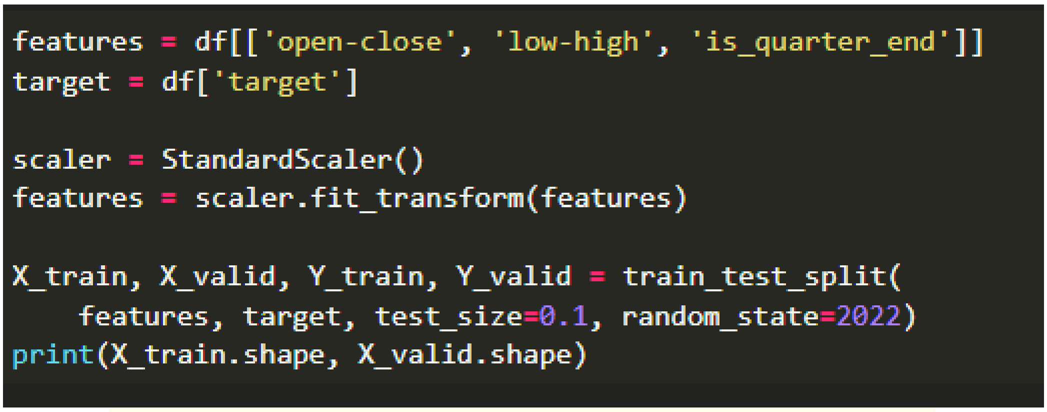
	Open	High	Low	Close	Volume	day	month	year	is_quarter_end	open-close	low-high	target
Open	1	1	1	1	0	0	0	0	0	0	0	0
High	1	1	1	1	0	0	0	0	0	0	0	0
Low	1	1	1	1	0	0	0	0	0	0	0	0
Close	1	1	1	1	0	0	0	0	0	0	0	0
Volume	0	0	0	0	1	0	0	0	0	0	0	0
day	0	0	0	0	0	1	0	0	0	0	0	0
month	0	0	0	0	0	0	1	0	0	0	0	0
year	0	0	0	0	0	0	0	1	0	0	0	0
is_quarter_end	0	0	0	0	0	0	0	0	1	0	0	0
open-close	0	0	0	0	0	0	0	0	0	1	0	0
low-high	0	0	0	0	0	0	0	0	0	0	1	0
target	0	0	0	0	0	0	0	0	0	0	0	1

Fig 15 (b) Heatmap of the correlation between the features.

From the above heatmap, we can say that there is a high correlation between OHLC that is pretty obvious, and the added features are not highly correlated with each other or previously provided features which means that we are good to go and build our model.

Data Splitting and Normalization

After selecting the features to train the model on we should normalize the data because normalized data leads to stable and fast training of the model. After that whole data has been split into two parts with a 90/10 ratio so, that we can evaluate the performance of our model on unseen data. Result of the below is (1522, 3) (170, 3).



```
features = df[['open-close', 'low-high', 'is_quarter_end']]
target = df['target']

scaler = StandardScaler()
features = scaler.fit_transform(features)

X_train, X_valid, Y_train, Y_valid = train_test_split(
    features, target, test_size=0.1, random_state=2022)
print(X_train.shape, X_valid.shape)
```

Fig 16 (a) data Split and normalization.

Model Development and Evaluation

Machine learning models**(Logistic Regression, Support Vector Machine, XGBClassifier),** and then based on their performance on the training and validation data we will choose which ML model is serving the purpose at hand better.

For the evaluation metric, we will use **the ROC-AUC** curve but why this is because instead of predicting the hard probability that is 0 or 1 we would like it to predict soft probabilities that are continuous values between 0 to 1. And with soft probabilities, the ROC-AUC curve is generally used to measure the accuracy of the predictions.

Logistic Regression- Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class. It is used for classification algorithms its name is logistic regression. it's referred to as regression because it takes the output of the linear regression function as input and uses a sigmoid function to estimate the probability for the given class.

Support Vector Machine- (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well it's best suited for classification. The main objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space. The hyperplane tries that the margin between the closest points of different classes should be as maximum as possible. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

XGBClassifier- XGBoost is an implementation of Gradient Boosted decision trees. This library was written in C++. It is a type of Software library that was designed basically to improve speed and model performance. It has recently been dominating in applied machine learning. XGBoost models majorly dominate in many Kaggle Competitions. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and the variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

ROC Curve

ROC stands for Receiver Operating Characteristics, and the ROC curve is the graphical representation of the effectiveness of the binary classification model. It plots the true positive rate (TPR) vs the false positive rate (FPR) at different classification thresholds.

AUC Curve:

AUC stands for Area Under the Curve, and the AUC curve represents the area under the ROC curve. It measures the overall performance of the binary classification model. As both **TPR and FPR range between 0 to 1,** So, the area will always lie between 0 and 1, and A greater value of AUC denotes better model performance. Our main goal is to maximize this area in order to have the highest TPR and lowest FPR at the given threshold. The AUC measures the probability that the model will assign a randomly chosen positive instance a higher predicted probability compared to a randomly chosen negative instance.

It represents the probability with with our model is able to distinguish between the two classes which are present in our target.

Basically, the ROC curve is a graph that shows the performance of a classification model at all possible thresholds(threshold is a particular value beyond which you say a point belongs to a particular class). The curve is plotted between two parameters.

TPR – True Positive Rate.

FPR – False Positive Rate.

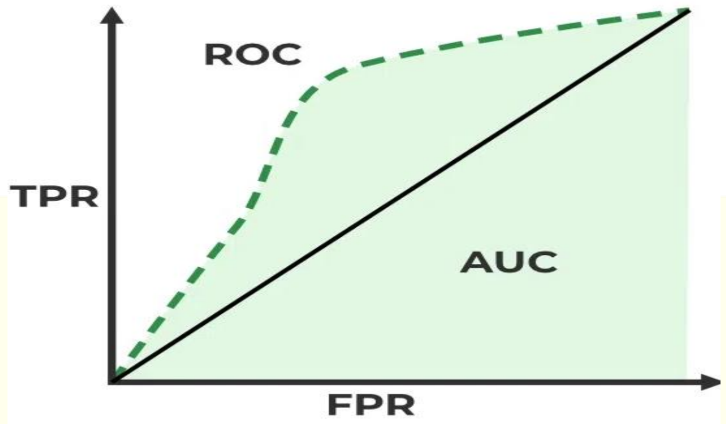
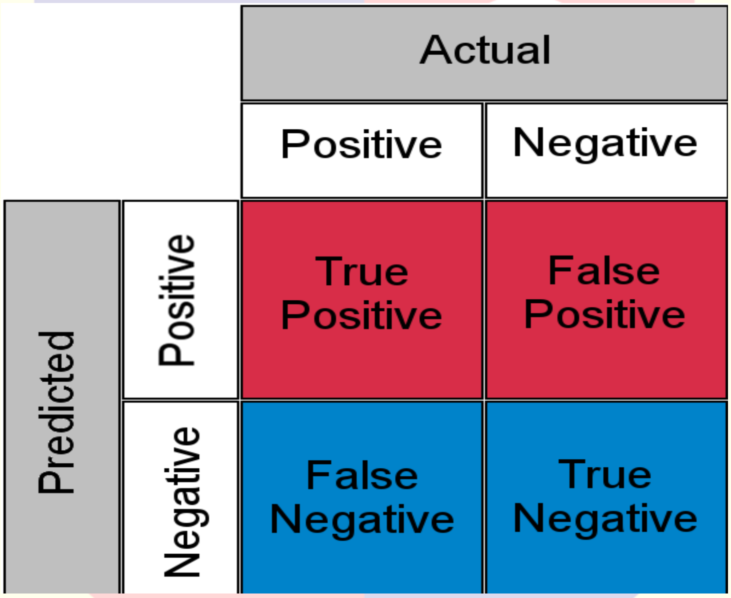


Fig 16 (b) TRP/FRP Relationship.



		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Fig 16 (c) Confusion Matrix for a Classification Task.

- **True Positive:** Actual Positive and Predicted as Positive
- **True Negative:** Actual Negative and Predicted as Negative
- **False Positive(Type I Error):** Actual Negative but predicted as Positive.
- **False Negative(Type II Error):** Actual Positive but predicted as Negative.

Sensitivity / True Positive Rate / Recall -TPR/Recall/Sensitivity is the ratio of positive examples that are correctly identified. It represents the ability of the model to correctly identify positive instances.

Formula For Calculation – TRP = TP/TP+FN

False Positive Rate- False Positive Rate is the ratio of negative examples that are incorrectly classified.

Formula for Calculation- FRP = FP/TN+FP = 1-specificity

Specificity measures the proportion of actual negative instances that are correctly identified by the model as negative. It represents the ability of the model to correctly identify negative instances.

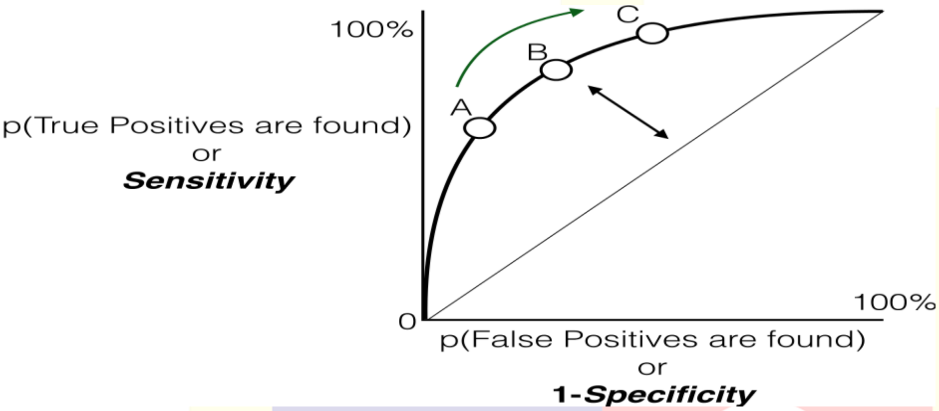
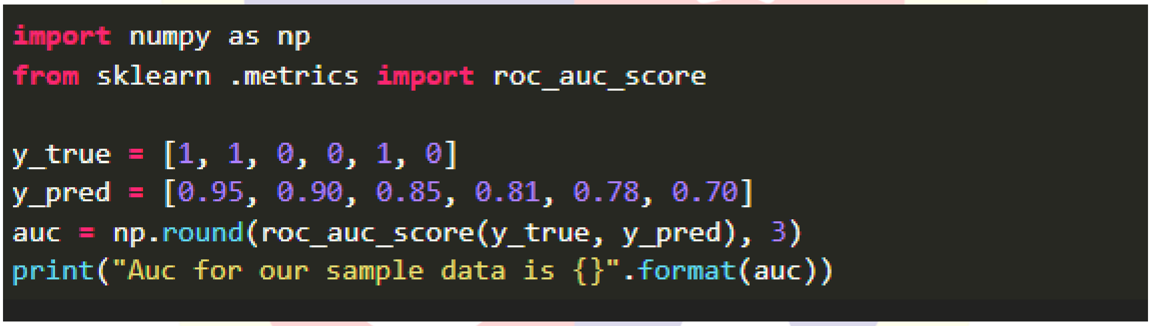


Fig 16 (d) Sensitivity versus False Positive Rate plot.

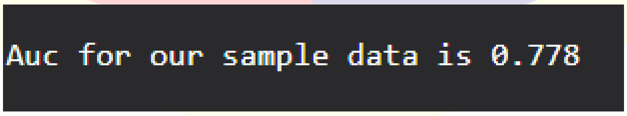
How does AUC-ROC work?

AUC measures how well a model is able to distinguish between classes.



```
import numpy as np
from sklearn .metrics import roc_auc_score

y_true = [1, 1, 0, 0, 1, 0]
y_pred = [0.95, 0.90, 0.85, 0.81, 0.78, 0.70]
auc = np.round(roc_auc_score(y_true, y_pred), 3)
print("Auc for our sample data is {}".format(auc))
```

Fig 17 (a) Roc_Auc_score input.

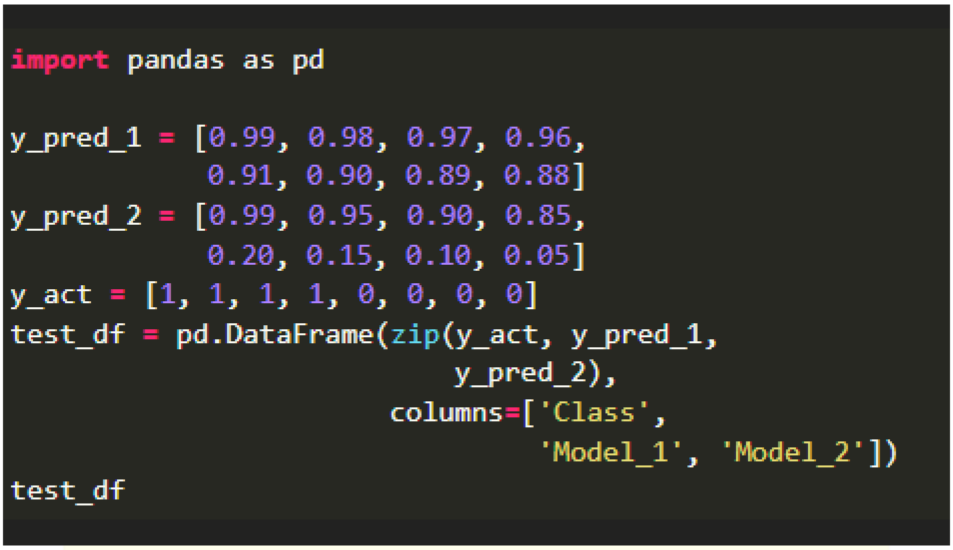


```
Auc for our sample data is 0.778
```

Fig 17 (b) Auc output.

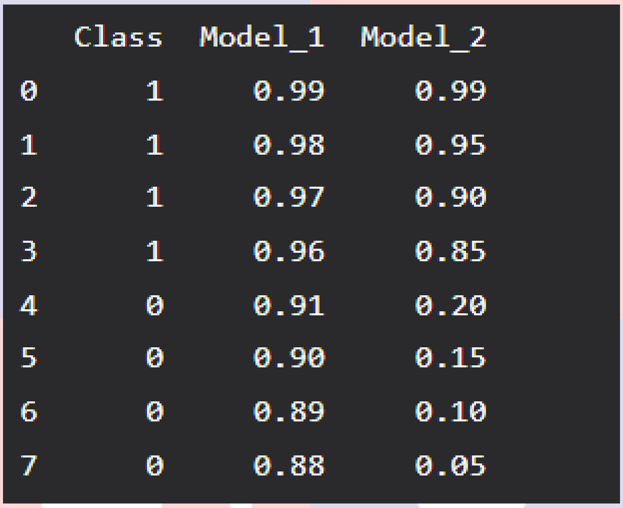
When to Use AUC-ROC curve?

ROC-AUC does not work well under severe imbalance in the dataset, to give some intuition for this let us look back at the geometric interpretation here. Basically, ROC is the plot between TPR and FPR(assuming the minority class is a positive class). ROC-AUC tries to measure if the rank ordering of **classifications** is correct it does not take into account actually predicted probabilities.



```
import pandas as pd

y_pred_1 = [0.99, 0.98, 0.97, 0.96,
            0.91, 0.90, 0.89, 0.88]
y_pred_2 = [0.99, 0.95, 0.90, 0.85,
            0.20, 0.15, 0.10, 0.05]
y_act = [1, 1, 1, 1, 0, 0, 0, 0]
test_df = pd.DataFrame(zip(y_act, y_pred_1,
                           y_pred_2),
                       columns=['Class',
                                'Model_1', 'Model_2'])
test_df
```

Fig 18 (a) Classification of 2 test Model.



```
   Class  Model_1  Model_2
0      1     0.99     0.99
1      1     0.98     0.95
2      1     0.97     0.90
3      1     0.96     0.85
4      0     0.91     0.20
5      0     0.90     0.15
6      0     0.89     0.10
7      0     0.88     0.05
```

Fig 18 (b) code output.

What is Classification?

Classification is a process of categorizing data or objects into predefined classes or categories based on their features or attributes. In machine learning, classification is a type of supervised learning technique where an algorithm is trained on a labelled dataset to predict the class or category of new, unseen data.

The main objective of classification is to build a model that can accurately assign a label or category to a new observation based on its features.

Types of Classification

Binary Classification: In binary classification, the goal is to classify the input into one of two classes or categories.

Example – On the basis of the given health conditions of a person, we have to determine whether the person has a certain disease or not.

Multiclass Classification: In multi-class classification, the goal is to classify the input into one of several classes or categories. For Example – On the basis of data about different species of flowers, we have to determine which specie our observation belongs to.

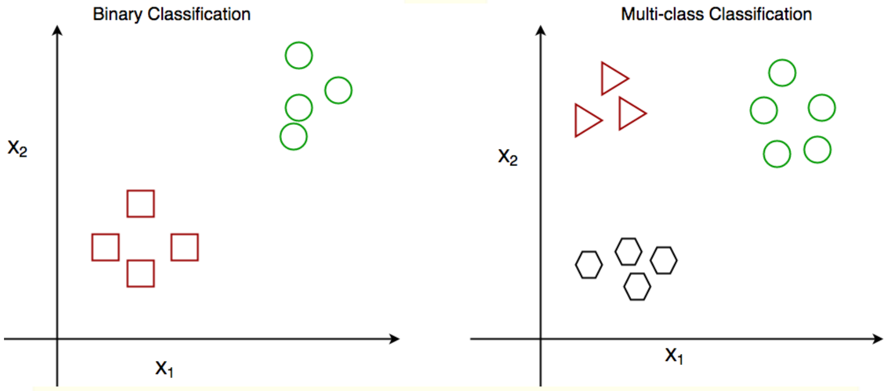
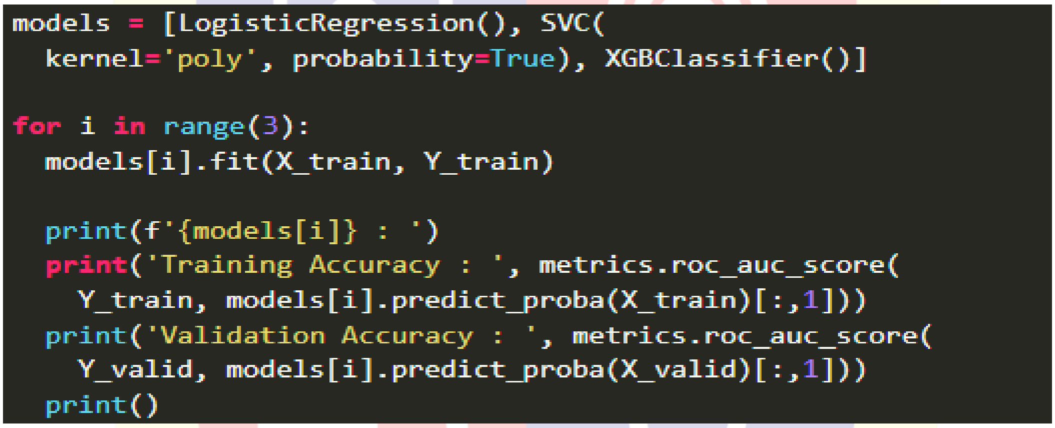
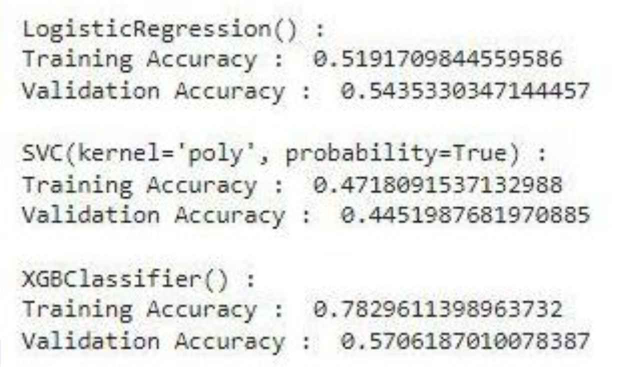


Fig 19 Binary vs Multi class classification.



```
models = [LogisticRegression(), SVC(
  kernel='poly', probability=True), XGBClassifier()]

for i in range(3):
  models[i].fit(X_train, Y_train)

  print(f'{models[i]} : ')
  print('Training Accuracy : ', metrics.roc_auc_score(
    Y_train, models[i].predict_proba(X_train)[:,1]))
  print('Validation Accuracy : ', metrics.roc_auc_score(
    Y_valid, models[i].predict_proba(X_valid)[:,1]))
  print()
```

Fig 20 (a) Model Training.



```
LogisticRegression() :
Training Accuracy :  0.5191709844559586
Validation Accuracy :  0.5435330347144457

SVC(kernel='poly', probability=True) :
Training Accuracy :  0.4718091537132988
Validation Accuracy :  0.4451987681970885

XGBClassifier() :
Training Accuracy :  0.7829611398963732
Validation Accuracy :  0.5706187010078387
```

Fig 20 (b)Model Accuracy.

Classification model Evaluations

Evaluating a classification model is an important step in machine learning, as it helps to assess the performance and generalization ability of the model on new, unseen data. There are several metrics and techniques that can be used to evaluate a classification model, depending on the specific problem and requirements. Here are some commonly used evaluation metrics:

Classification Accuracy: The proportion of correctly classified instances over the total number of instances in the test set. It is a simple and intuitive metric but can be misleading in imbalanced datasets where the majority class dominates the accuracy score.

Confusion matrix: A table that shows the number of true positives, true negatives, false positives, and false negatives for each class, which can be used to calculate various evaluation metrics.

Precision and Recall: Precision measures the proportion of true positives over the total number of predicted positives, while recall measures the proportion of true positives over the total number of actual positives. These metrics are useful in scenarios where one class is more important than the other, or when there is a trade-off between false positives and false negatives.

F1-Score: The harmonic mean of precision and recall, calculated as 2 x (precision x recall) / (precision + recall). It is a useful metric for imbalanced datasets where both precision and recall are important.

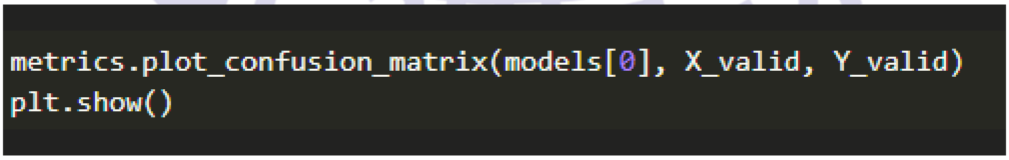
ROC curve and AUC: The Receiver Operating Characteristic (ROC) curve is a plot of the true positive rate (recall) against the false positive rate (1-specificity) for different threshold values of the classifier's decision function. The Area Under the Curve (AUC) measures the overall performance of the classifier, with values ranging from 0.5 (random guessing) to 1 (perfect classification).

Cross-validation: A technique that divides the data into multiple folds and trains the model on each fold while testing on the others, to obtain a more robust estimate of the model's performance.

In the project we used confusion Matrix

Validation Through - Confusion Matrix

A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data. It is often used to measure the performance of classification models, which aim to predict a categorical label for each input instance. The matrix displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) produced by the model on the test data.



```
metrics.plot_confusion_matrix(models[0], X_valid, Y_valid)
plt.show()
```

Fig 21 (a) confusion matrix code.

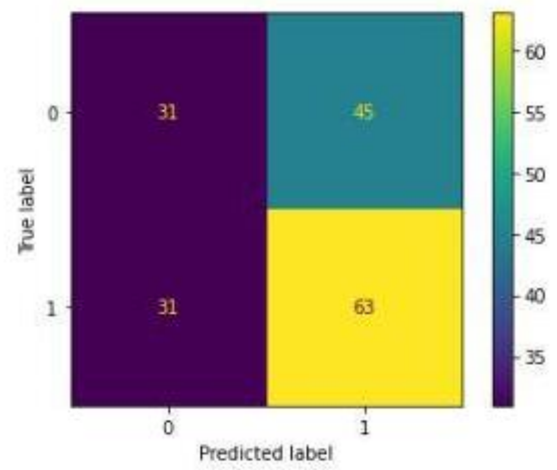


Fig 21 (b) confusion matrix.

Learnings in the Project

Learning are as follows-

- How-to pick-up stocks.
- How we can perform analysis on the data provided.
- How we can build a stock portfolio.
- How to read the stock graphs.
- Learning Financial Industry concepts like (WPI, CPI, FX reserves, and Dump and Pump concepts, Industrial and Manufacturing production, India Bank Loan Growth.
- ML model creations by Python and in-depth analysis performed.

Mitigation Measures

Protect Yourself—How You Can Avoid Becoming a Victim

Identify the warning signs

- Does the offer sound too good to be true?
- Is the seller using high pressure sales tactics?
- Did the investment offer unsolicited?
- Did the seller ask for information that is usually considered personal (e.g. social security number, credit card information, etc.) over the phone or Internet?
- If you answer "yes" to any of the above questions when considering an investment opportunity, you may be the target of a scam artist.

Take action to avoid fraud.

Don't believe everything you are told by the seller. Take the time to do your own research on the investment's potential.

- Don't assume the solicitor is who he or she claims to be.
- Check with federal and state securities regulators to find out if there have been any complaints against the company.
- Ask the promoter whether—and how much—he or she has been paid to tout the opportunity.
- Ask where the company is incorporated and then call that state to ensure that the company has a current annual report on file.
- Request written financial information, such as a prospectus, annual report, offering circular, or financial statements, then compare the written information to what you were told.
- Get offers in writing and save a copy for your records.
- Check with a trusted financial advisor, your broker, or an attorney about any investments you are considering.

Chapter – 7: Future work & conclusion

For the next stage, the focus would be on the practical part of the project.

- More Research on fraud detection and prevention and how we can Associate AI with it. And Dump and Pump system 3.0
- Application of ML and security purpose.
- As this project was the application of the internship with predictRam more focus will on Financial Model Further improvement, algorithm improvements, Article studies and Blogs and Case studies.

Through this project I want to make awareness about the Pump and Dump system in the stock market as we have already seen current example like (SV Bank Failed, and various other bank failures).

References

List of papers/books/websites etc. referred for project

[1] Detecting “Pump & Dump Schemes” on Cryptocurrency Market Using An Improved AprioriAlgorithm Weili Chen∗†, YueJin Xu∗†, Zibin Zheng∗†, Yuren Zhou∗, Jianxun Eileen Yang‡ and Jing Bian∗ School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China National Engineering Research Center of Digital Life, Sun Yatsen University, 510006, Guangzhou, China corresponding author. Shenzhen.

[2] Sai, Chaithanya Vamshi, et al. "Explainable Ai-Driven Financial Transaction Fraud Detection Using Machine Learning and Deep Neural Networks." Available at SSRN 4439980.

[3]. A NewWolf in Town? Pump-and-Dump Manipulation in Cryptocurrency Markets* Anirudh Dhawan1,2 and Talis J. Putnins2,3,4 1Indian Institute of Management Bangalore, India, 2University of Technology Sydney, Australia, 3Stockholm School of Economics in Riga, Latvia and 4Digital Finance Co-operative Research Centre, Australia
(Review of Finance, 2023, 935–975 https://doi.org/10.1093/rof/rfac051 Publication Date: 4 August 2022).

[4] Detecting cryptocurrency pump-and-dump frauds using market and social signalsHuy Nghiem *, Goran Muric , Fred Morstatter , Emilio Ferrara University of Southern California, Information Sciences Institute, 4676 Admiralty Way, Marina del Ray, Los Angeles, California, USA(journalhomepage:www.elsevier.com/locate/eswa)
https://doi.org/10.1016/j.eswa.2021.115284 Received 18 November 2020; Received in revised form 17 May 2021; Accepted 23 May 2021.

[5] The economics of stock touting during Internet based pump and dump campaigns Michael Siering Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt, Germany (Info Systems J. 2018;1–28. wileyonlinelibrary.com/journal/isj) © 2018 John Wiley & Sons Ltd (DOI: 10.1111/isj.12216).

[6] Explainable AI-Driven Financial Transaction Fraud Detection using Machine Learning and Deep Neural Networks Chaithanya Vamshi Saia , Debashish Dasa, *, Nouh Elmitwallya , Ogerta Elezaja , Md Baharul Islamb (https://ssrn.com/abstract=4439980).

[7] A Comparative Analysis Credit card fraud detection using Machine Learning Techniques: Samuel A. Oluwadare, Adebayo O., Adetunmbi, John O. Awoyemi (978-1-5090-4642-3/17/$31.00 ©2017 IEEE).

[8] Deep Learning Detecting Fraud in Credit Card Transactions Abhimanyu Roy, Jingyi Sun, Robert Mahoney, Loreto Alonzi, Stephen Adams, Peter Beling 978-1-5386-6343-1/18/$31.00 ©2018 IEEE.

[9] Fraud Detection using Machine Learning and Deep Learning Pradheepan Raghavan, Neamat El GayarSchool of Mathematical and Computer Sciences Heriot Watt University Dubai, UAE2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE 978-1-7281-3778-0/19/$31.00 ©2019 IEEE).

[10] Predictive-Analysis-based Machine Learning Model for Fraud Detection with Boosting Classifiers M. Valavan and S. Rita* Department of Statistics, Periyar University, Salem, Tamilnadu, India (Computer Systems Science & Engineering DOI: 10.32604/csse.2023.026508).

[11] The advanced proprietary AI/ML solution as Anti-fraud- Tensorlink4cheque (AFTL4C) for Cheque fraud detection. Dr. DYAN CHANDRA YADAV, PRABHAKARA R UYYALA COMPUTER AND SCIENCE ENGINEERING Research Scholar, JS UNIVERSITY, The International journal of analytical and experimental modal analysis Volume XV, Issue IV, April/2023 ISSN NO: 0886-9367.

[12] Abu-Mostafa, Y.S., Atiya, A.F. Introduction to financial forecasting. Appl Intell 6, 205–213 (1996). https://doi.org/10.1007/BF00126626.

[13] PREDICTIVE ANALYTICS Extending the Value of Your Data Warehousing Investment (Book) By Wayne W. Eckerson.

WEBSITE REFERENCES:-

[1] https://www.forbes.com/advisor/in/investing/why-adani-shares-are-falling/

[2] https://seon.io/resources/transaction-fraud-detection/

[3]https://www.cnbc.com/2023/03/24/svb-collapse-squad-app-ceo-says-silicon-valley-bank-failure-was-agony.html

[4] https://www.investopedia.com/terms/p/pumpanddump.asp