PROJECT TITAN-X

SOW, Data Aqusition Acquisition & Understanding

AIDI-2005-01 - CAPSTONE TERM II
Course Facilitator: Marcos Bittencourt
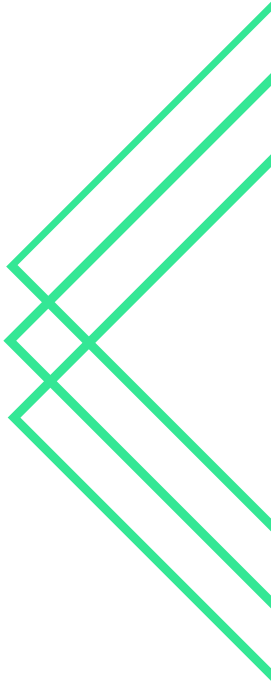Prepared by Team*Ace*

February 2020
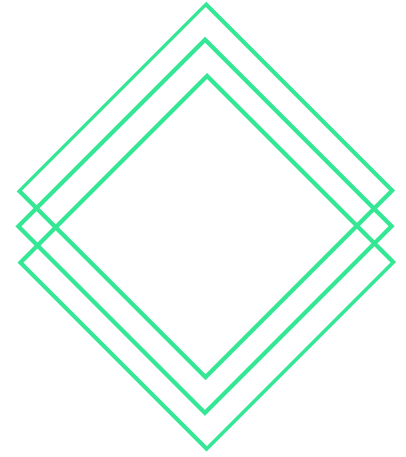
TABLE OF CONTENTS

Our Team

The people behind the project.
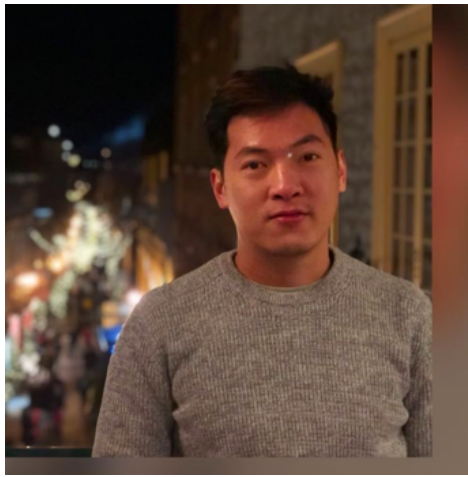
Team*Ace*

"We always aim to conquer any difficulties and challenges. 'Just do it, make it happen, and ACE it!' is our team's core spirit."

Keng Hin Cheong

100711777

AI Specialist

Au Quang Loc Nguyen

100741710

AI Specialist

Chutu Li

100765238

AI Specialist

Executive Summary/Intro

Stock trading has always been a task done by trading professionals in finance, since it can be very difficult, risky, time consuming and emotionally affected. Some people would agree that it is similar to gambling in a casino. Those who really want to invest in the stock market are always concerned for a plethora of reasons such as having too little time to understand the stock market and all of its options. And when the people actually go ahead and invest their hard-earned money into stock markets, it is most likely that they will lose it all, due to the lack of experience with the scenarios and even the most common situations of stock markets. In the end, more often than not, they become frustrated and decide for other types of investments.

Therefore, our project aims to address the problems mentioned previously and attempts to solve them in a way that brings everyone to invest in stocks a peace of mind, with a highly capable, machine-learning-driven engine doing most of the dirty work. The ML engine will not only decide and execute by itself, either to buy or to sell stocks, but will predict which kind of stocks is better suited the investment goals for each of the users, in order to maximize the returns. The ML engine will also aim to recommend stocks to the users based on correlation between stock descriptions and the inputs of the user's personal interests.

The ultimate goal of our project is not only to help new stock investors make money but to expand the community and get more people involved in this big industry that has no plans of stopping any time soon. However, due to the limits of time and resources, our project at this stage will only cover the SP500 IT Sector of the US stock market.
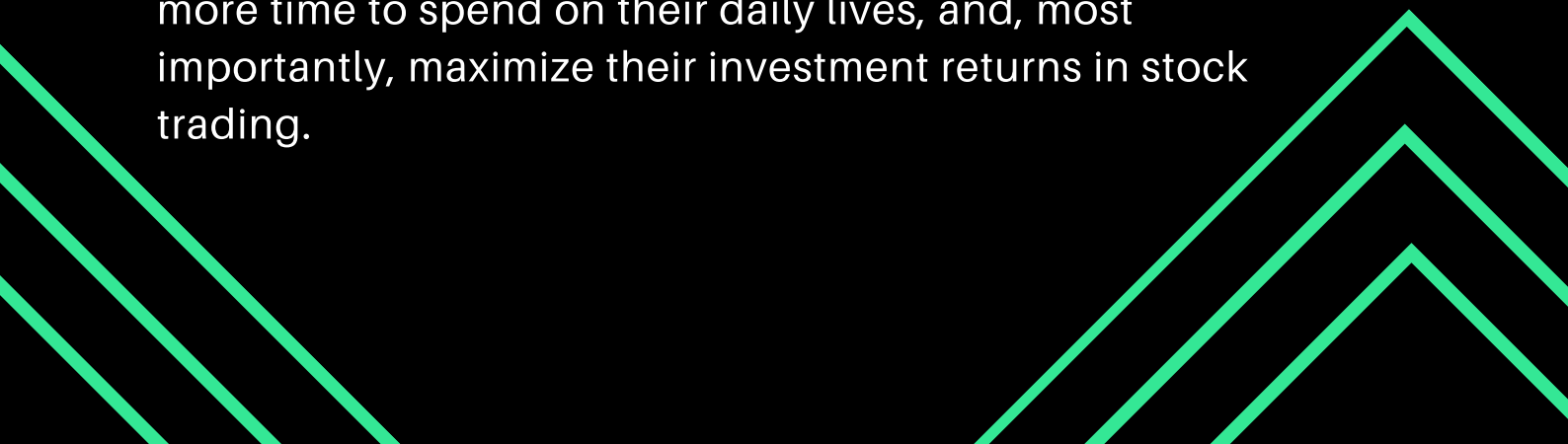
Rationale Statement

The main goal of our project will solve the difficulties of investing, which consists of removing the irrationality that leads to bad decisions because of human emotions and the large amount of time required to gain understanding of how stock market works and how to invest stocks with profits. The risks involved will be reduced since the ML engine will analyze and recommend the best stocks in terms of future returns. No previous understanding of the stock market is necessary for users with this ML engine developed by the project. All the information needed from users will only be their text or voice inputs for the kind of stocks they would like to invest and how long they would like to hold their investments, then the ML engine will come out with a list of stocks and pick the best ones for the users based on the future returns.

The real-time stock price will be compared to the model's prediction to verify its accuracy in predicting the stock price. Our goal is to make the ML engine's accuracy rate to be over 90% in prediction. We also expect user satisfaction to the engine to be higher than 95%.

Clients/users will continuously benefit from our project by being able to trade with ease and have less worry, more time to spend on their daily lives, and, most importantly, maximize their investment returns in stock trading.

Problem Statements

1. Investing in the stock market is hard and not for beginners. 53% of people think that they don't have enough money to invest, 21% don't know about stocks, 9% don't trust the stockbrokers and 7% think that stocks are too risky. The 10% left are afraid of higher fees and some other reasons*.
   a. Our project will make investing easy and care-free,since the user will only have to enter some inputs and the program will take care of the rest of the stock trading process to maximize the investment return.
2. Stock investors make irrational decisions based on emotions, accounting for almost 90% of losses in the stock market*.
   a. Our project will base its decisions on ML algorithms which learn from data and tendencies. It will improve the chances of winning and reduce that of loss.
3. Stock markets are different due to political and economic conditions. And because of the limit of resources and time, our project will focus on the IT Sector of SP500 in the US Stock market only.

*See References

Data Requirement

1. All data needs to be timely, accurate, consistent with the US Stock market.
2. All data needs to be relevant to our project objectives.
3. For training, validating, and testing our machine learning algorithm properly, the SP500 IT sector data should at least have columns of "Date", "Close", where "Date" is the record date; "Close", is the price of the stock when the US stock market closes.
4. All the above stock dataset should have at least 3 years of data to be enough for effective machine learning.
5. The information dataset of each individual company should have company's description, sector, Industry.
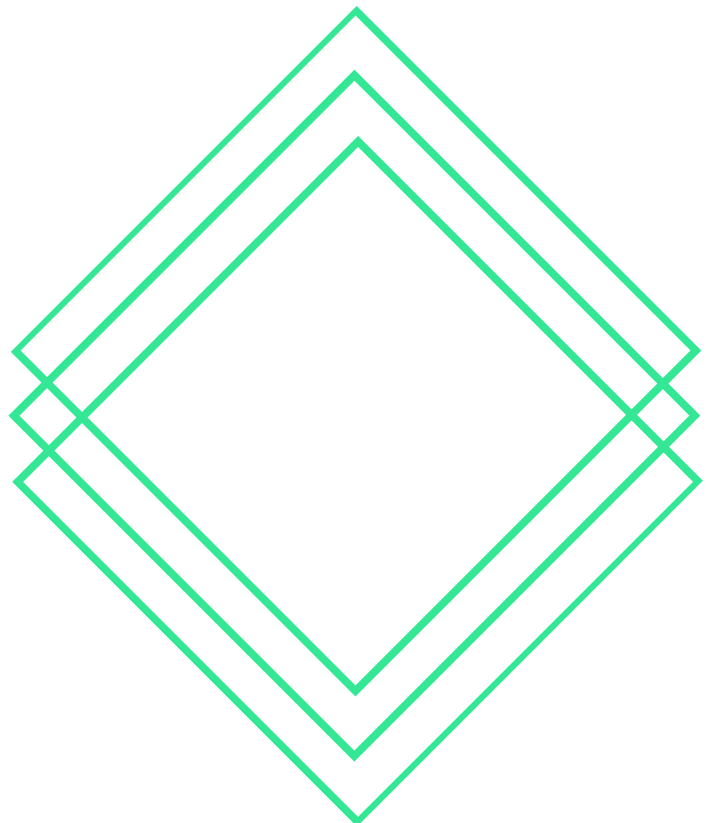
Data Requirement

Assumptions:
1. The stock/index data sources are trustable and accurate with the information provided.
2. The descriptions data of each company in the SP500 IT sector is accurate and up-to-date.

Constraints:
1. The dataset are all historical data, not real-time. Therefore, it may not reflect the most up-to-date info about the index/stock prices of our interest.
2. Due to the fact that all our stocks/index data only reflects daily changes, our project will produce better outcomes for medium to long term investors of the users, since it would be less dependent on the analyzing the immediate fluctuations on stock prices than short term investors.
3. More columns in the stock/index dataset are preferred, such as P/E ratio, intra-day high, and intra-low, for better predictive results.
4. The pre-market and after-market prices/index are not included in the datasets. The information could be helpful for price prediction for the machine learning algorithms.
5. Other factors could also influence stock prices, such as political, social factors. The data for those could be added to the project in the future if needed to provide better predictive results.

Data for the Project

For our project, we will mainly be using open financial data from the Internet. Below is the proposed list of data being used in training, validating, and testing our machine learning algorithms:

∗Because of lack of availability of the data files for the list of each individual company in SP500 IT sector, we will be using Python to scrap the information from a website below and make all the needed information into a csv file for our project:

1.The Website for the list of each individual company in SP500 IT sector:
    Source Link: https://www.barchart.com/stocks/indices/sp-sector/information-   technology
    Category of Data: Open data
    The Sample of the List:

S&P 500 Info Tech Components

Main View

Symbol	Name
AAPL	Apple Inc
ACN	Accenture Plc
ADBE	Adobe Systems Inc
ADI	Analog Devices
ADP	Automatic Data Procs
ADS	Alliance Data Systems Corp
ADSK	Autodesk Inc
AKAM	Akamai Technologies
AMAT	Applied Materials
AMD	Adv Micro Devices
ANET	Arista Networks Inc
ANSS	Ansys Inc
APH	Amphenol Corp
AVGO	Broadcom Ltd

Data for the Project

Furthermore, the stock prices data and description of each company in SP500 IT Sector will also be needed for our project. Therefore, we will be using an API called 'pandas_datareader.data' to get the stock data and scrapping the information of the description of each company in SP500 IT sector on from Yahoo Finance:

2. Category of Data: Open data
   For example, for the stock price of the Company of Apple:
   The sample of the Stock Dataset:

```python
import pandas_datareader.data as web
```

```python
In [91]:  #get the stock data for Apple Inc using pandas_datareader
          stock_symbol = 'AAPL'
          data = web.get_data_yahoo(stock_symbol, '1/1/2016', '12/6/2019',)
          data.reset_index(inplace=True,drop=False)
```

```python
In [96]:  data.head(10)
```

	Date	High	Low	Open	Close	Volume	Adj Close
0	2016-01-04	105.370003	102.000000	102.610001	105.349998	67649400.0	98.446655
1	2016-01-05	105.849998	102.410004	105.750000	102.709999	55791000.0	95.979675
2	2016-01-06	102.370003	99.870003	100.559998	100.699997	68457400.0	94.101387
3	2016-01-07	100.129997	96.430000	98.680000	96.449997	81094400.0	90.129868
4	2016-01-08	99.110001	96.760002	98.550003	96.959999	70798000.0	90.606438
5	2016-01-11	99.059998	97.339996	98.970001	98.529999	49739400.0	92.073563
6	2016-01-12	100.690002	98.839996	100.550003	99.959999	49154200.0	93.409874
7	2016-01-13	101.190002	97.300003	100.320000	97.389999	62439600.0	91.008270
8	2016-01-14	100.480003	95.739998	97.959999	99.519997	63170100.0	92.998695
9	2016-01-15	97.709999	95.360001	96.199997	97.129997	79833900.0	90.765305
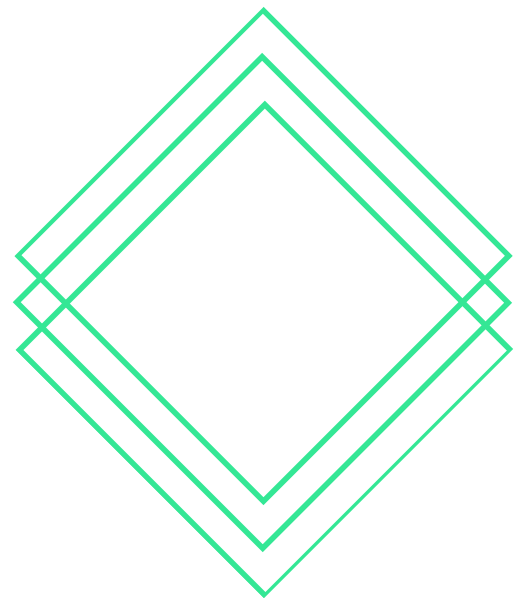
Data for the Project

3. Category of Data: Open data

Source Link: https://finance.yahoo.com/quote/AAPL/profile?p=AAPL

The Sample of the Description of the Company:

Apple Inc. (AAPL)
NasdaqGS - NasdaqGS Real Time Price. Currency in USD

218.82 -1.07 (-0.49%) **218.75** -0.07 (-0.03%)
At close: 4:00PM EDT   After hours: 7:59PM EDT

Summary Company Outlook Chart Conversations Statistics Historical Data **Profile** Financials NEW Analysis Optio

Apple Inc.

One Apple Park Way
Cupertino, CA 95014
United States
408-996-1010
http://www.apple.com

Sector: **Technology**
Industry: **Consumer Electronics**
Full Time Employees: **100,000**

Key Executives

Name	Title	Pay	Exercised	Year Born
Mr. Timothy D. Cook	CEO & Director	15.68M	N/A	1961
Mr. Luca Maestri	CFO & Sr. VP	5.02M	N/A	1964
Mr. Jeffrey E. Williams	Chief Operating Officer	5.05M	N/A	1964
Ms. Katherine L. Adams	Sr. VP, Gen. Counsel & Sec.	5.31M	N/A	1964
Mr. Chris Kondo	Sr. Director of Corp. Accounting	N/A	N/A	N/A

Description

Apple Inc. designs, manufactures, and markets mobile communication and media devices, and personal computers. It also sells various related software, services, accessories, and third-party digital content and applications. The company offers iPhone, a line of smartphones; iPad, a line of multi-purpose tablets; and Mac, a line of desktop and portable personal computers, as well as iOS, macOS, watchOS, and tvOS operating systems. It also provides iTunes Store, an app store that allows customers to purchase and download, or stream music and TV shows; rent or purchase movies; and download free podcasts, as well as iCloud, a cloud service, which stores music, photos, contacts, calendars, mail, documents, and others. In addition, the company offers AppleCare support services; Apple Pay, a cashless payment service; Apple TV that connects to consumers' TVs and enables them to access digital content directly for streaming video, playing music and games, and viewing photos; and Apple Watch, a personal electronic device, as well as AirPods, Beats products, HomePod, iPod touch, and other Apple-branded and third-party accessories. The company serves consumers, and small and mid-sized businesses; and education, enterprise, and government customers worldwide. It sells and delivers digital content and applications through the iTunes Store, App Store, Mac App Store, TV App Store, Book Store, and Apple Music. The company also sells its products through its retail and online stores, and direct sales force; and third-party cellular network carriers, wholesalers, retailers, and resellers. Apple Inc. was founded in 1977 and is headquartered in Cupertino, California.
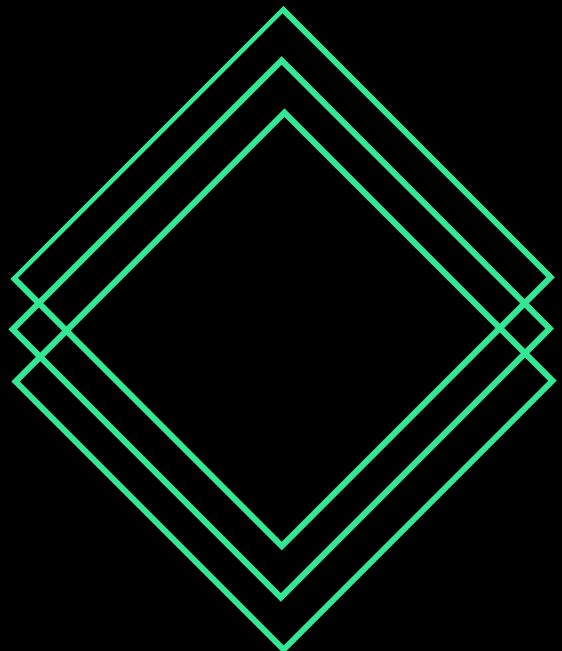
Proposed Model/Architecture Approach

The pipeline for the project consists of three main parts:

- the first part that recommends a list of stocks based on text/voice inputs from the users

- the second that is to analyze and predict a stock's future return

- the last that buy and/or sell stocks automatically for the user

The proposed main AI technologies involved in the project are as follows:

- **Natural Language Processing**
- **Computer Vision**
- **Regression**
- **Deep Learning**
- **Voice Recognition**
- **Text Analysis**

Key Performance Metrics

For the three main parts of the project, the metrics to be used for the performance evaluations are as follows:

- the first part that recommends a list of stocks based on text/voice inputs from the users:
  - Accuracy rate regarding comprehending the inputs from the user.
  - We aim to achieve the accuracy rate of over 90%.

- the second that is to analyze and predict a stock's future return
  - Back-testing and comparing the stock's historical return.
  - We aim to achieve the accuracy rate of over 90%.

- the last that buy and/or sell stocks automatically for the user
  - Comparing the results with the return performance of SP500 Index of the US stock market.
  - We aim to beat the performance of SP500 Index by at least 30%.
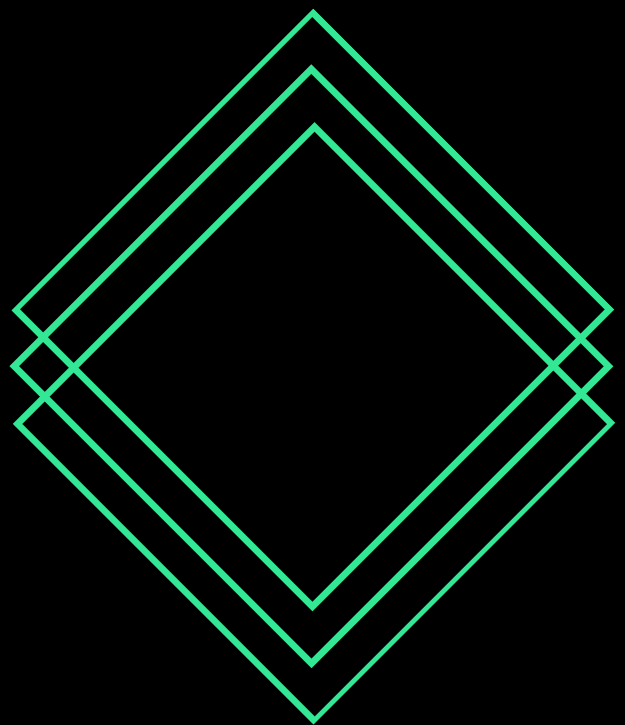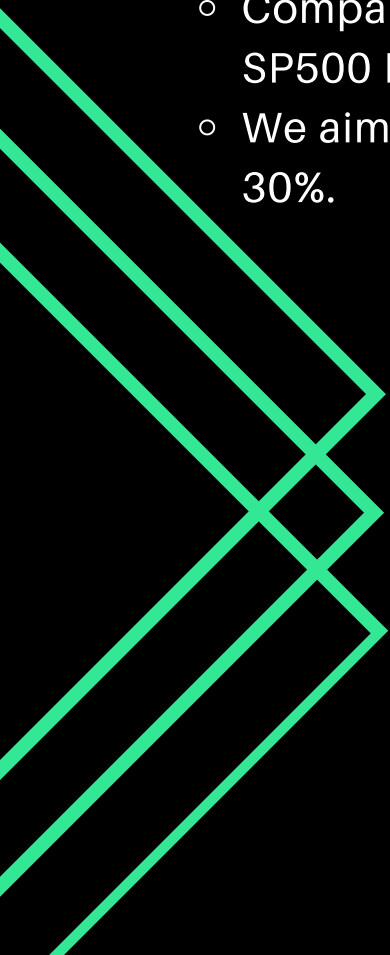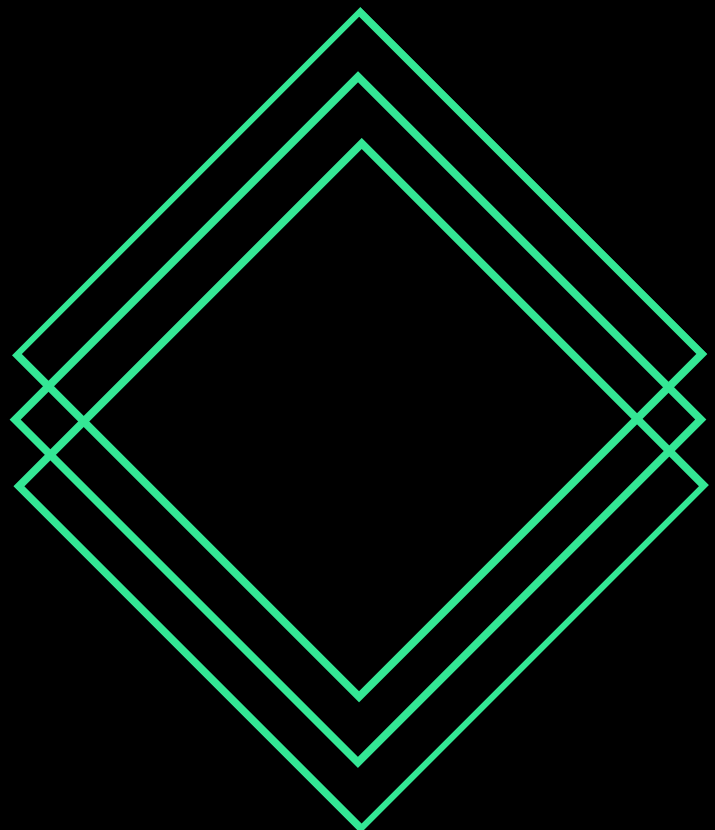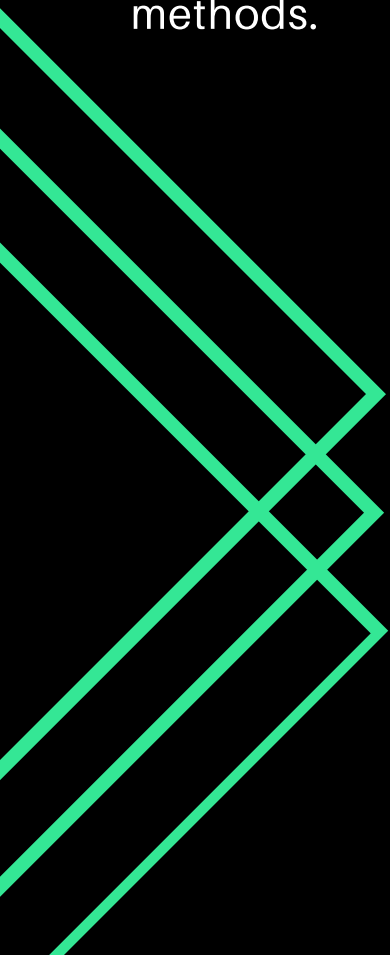
Exploratory Data Analysis

To perform the necessary exploratory data analysis for model trainings, we have chosen to use Jupyter Notebook in Python to get a more interactive environment so that the job can be done more easily. Please visit the notebook for more details.

The hyperlink to the Jupyter Notebook is:
https://github.com/teamAceAIDIDurhamC/docs

The Benefits of Feature Engineering
It is very often said that "data is the fuel of machine learning." The fact is before-processed data is like the crude oil of machine learning. This "crude oil", without being refined, can hardly be useful for training a model. Therefore, feature engineering is the process of extracting relevant features from a raw dataset, so that an accurate model can be produced using machine learning methods.

References

https://www.quora.com/Why-do-people-not-invest-in-stock-market-when-we-have-seen-tremendous-results-in-long-term-investments

https://www.investing-arena.com/why-do-90-of-people-lose-money-on-the-stock-market/

https://www.investopedia.com/articles/basics/03/062003.asphttps://chartyourtrade.com/7-common-trading-problems-and-their-solutions/

https://www.desiretotrade.com/5-difficulties-in-trading-and-how-to-overcome-them/

https://medium.com/datadriveninvestor/machine-learning-for-stock-market-investing-f90ad3478b64

https://towardsdatascience.com/machine-learning-techniques-applied-to-stock-price-prediction-6c1994da8001

https://medium.com/recombee-blog/machine-learning-for-recommender-systems-part-1-algorithms-evaluation-and-cold-start-6f696683d0ed

https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

https://www.toptal.com/machine-learning/supervised-machine-learning-algorithmshttps://medium.com/python-pandemonium/regression-for-understanding-machine-learning-ef89a62906a9

https://finance.yahoo.com

https://www.investing.com

https://www.ft.com

https://money.cnn.com/data/us_markets/

https://www.wsj.com