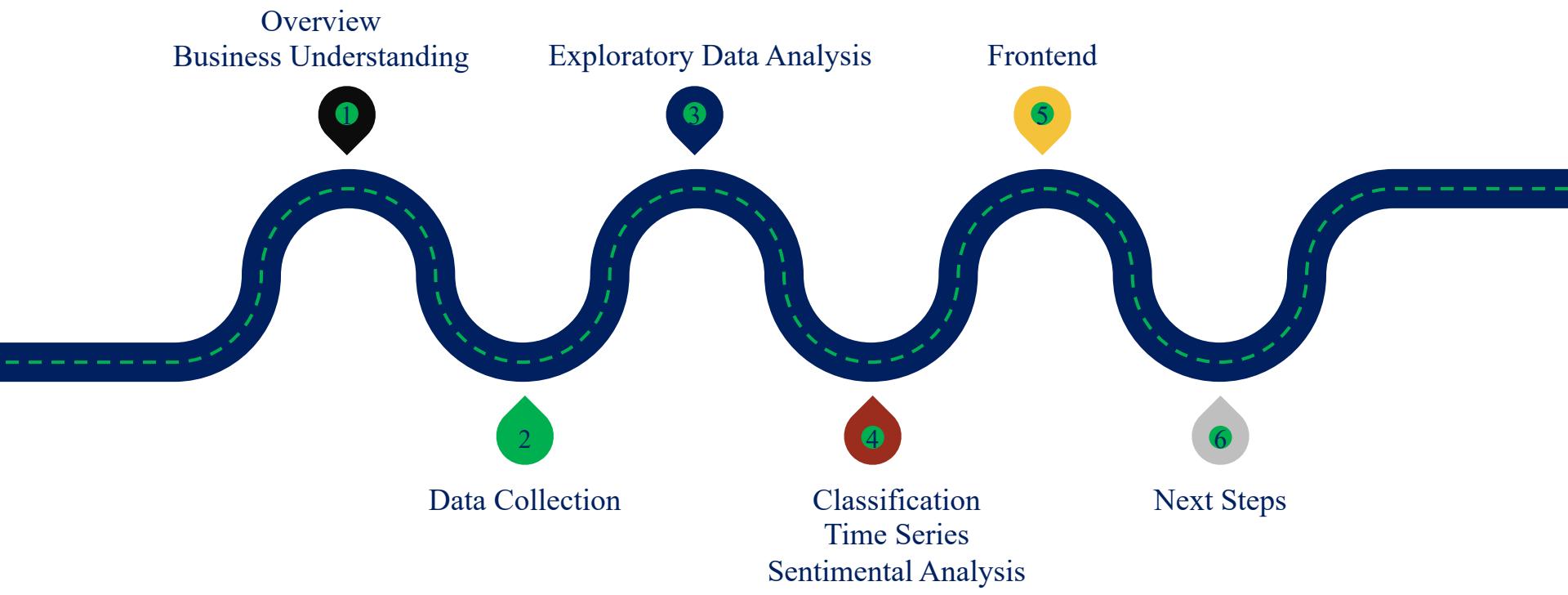


# Stock Market Prediction



Ning Chen

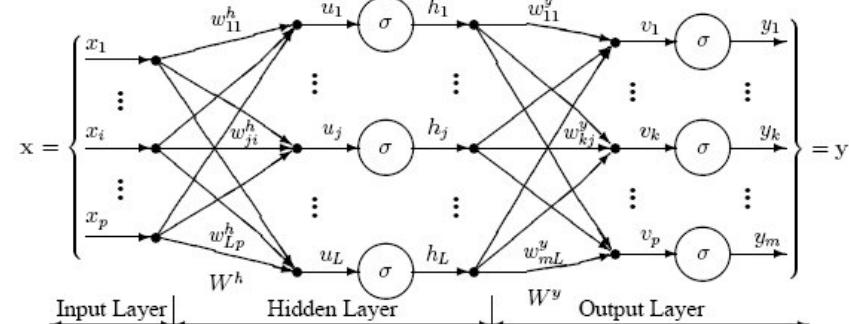
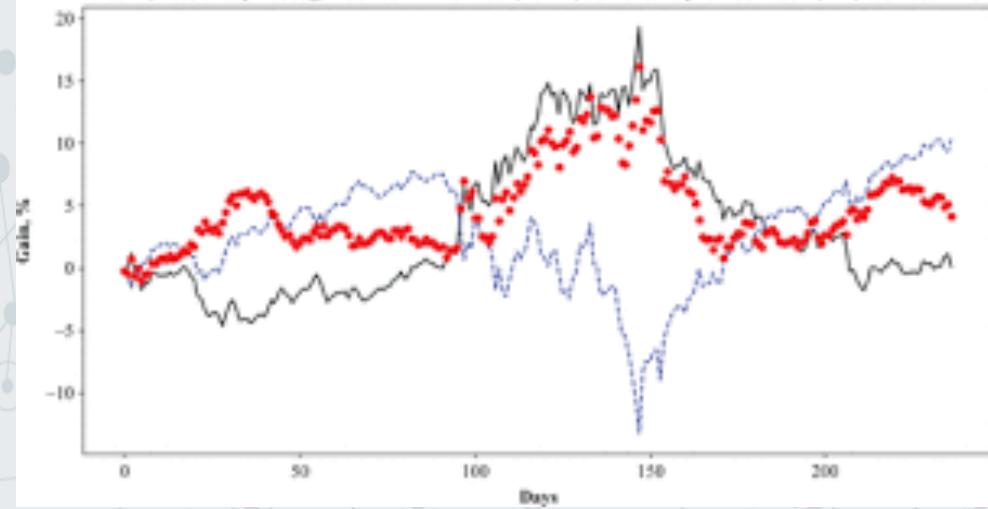
# Agenda



# Overview

Accurate prediction of stock market asset is a significant and challenging task due to complicated nature of the financial stock markets.

Table: Look-Ahead Performance of 1-Day Ahead NN Strategy  
(solid, 10-day training) for NYSE:SPY vs. SPY (dashed) vs. 10/5/8-Day MA Crossover (dots)



# Business Goals

Stock market prediction aims to determine the future movement of the stock value of a financial exchange.



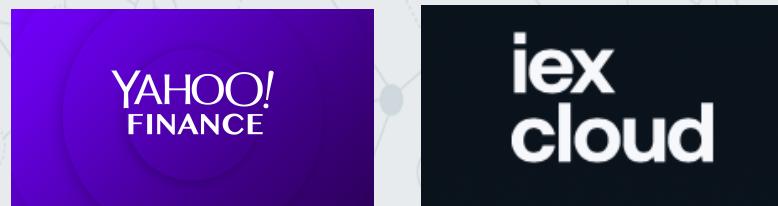
**Classification:** Trade Strategy (buy sell hold)

random forest, SVM, Xgboost, NN

**time series** regression model: SARIMAX, Facebook Prophet, LSTM

**sentimental analysis:** NLP

# Data Collection



A screenshot of a financial data page for Apple Inc. (AAPL) on Yahoo Finance. At the top, it shows major indices: S&amp;P 500 at 3,931.33 (-2.35 or -0.05%), Dow 30 at 31,613.02 (+90.27 or +0.29%), Nasdaq at 13,965.50 (+82.00 or +0.58%), Russell 2000 at 2,256.11 (-16.78 or -0.74%), Crude Oil at 60.94 (+0.85 or +1.40%), and Gold at 1,771.90 (+27.10 or +1.51%). Below the indices, the main focus is on Apple Inc. (AAPL), showing its price of 130.84, a change of -2.35 (-1.76%), and a volume of 130.87 with a 0.03 (0.02%) increase. The page includes tabs for Summary, Company Outlook, Chart, Conversations, Statistics, Historical Data, Profile, Financials (which is selected), Analysis, Options, Holders, and Sustainability. Under the Financials tab, there are links for Income Statement, Balance Sheet, and Cash Flow, with the Income Statement currently selected. It shows quarterly financial data from 2018 to 2020. To the right, there's a "People Also Watch" section listing other stocks like AMZN, FB, GOOG, TSLA, and NFLX with their latest prices and percentage changes. The overall layout is clean and organized, typical of a professional financial news website.

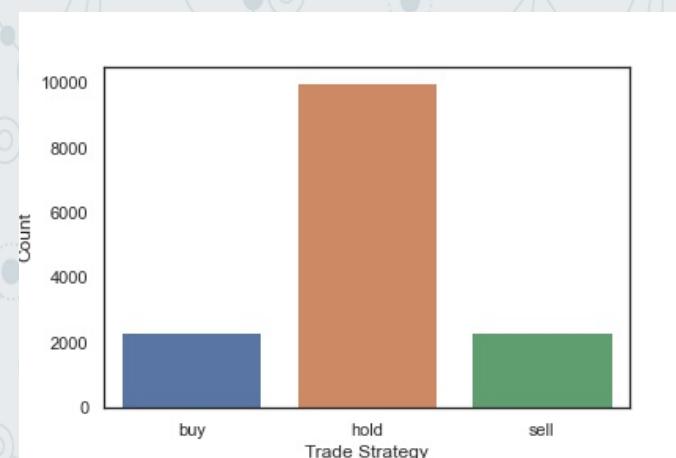
Data was collected from three different web sources by API calls or Web Scraping.

- Quarterly Report for Classification by Web Scrapping.
- Yahoo Finance and IEX API for Time Series by API calls.
- Twitter for sentimental data by VADER.

# Exploratory Data Analysis

## Trade Strategy

local minimum of the price to buy, local maximum of the price to sell, and all other time to hold.



# Exploratory Data Analysis

## Missing data

The missing data for weekends and holidays was filled by interpolation method.

The missing data of exogenous features was filled by propagating nearest valid observation backward/forward to next valid observation.

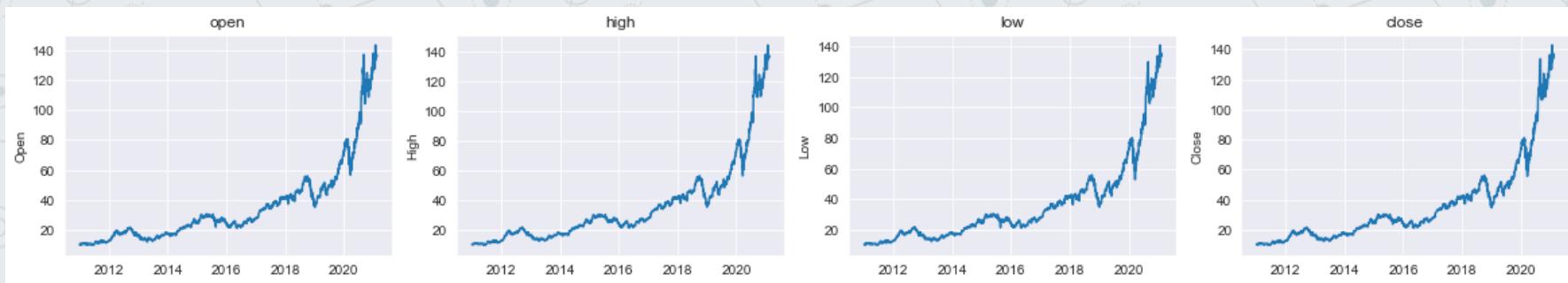
$$P_t = \sqrt[3]{P_{t-1}^2 * P_{t+2}}$$

$$P_{t+1} = \sqrt[3]{P_{t-1} * P_{t+2}^2}$$

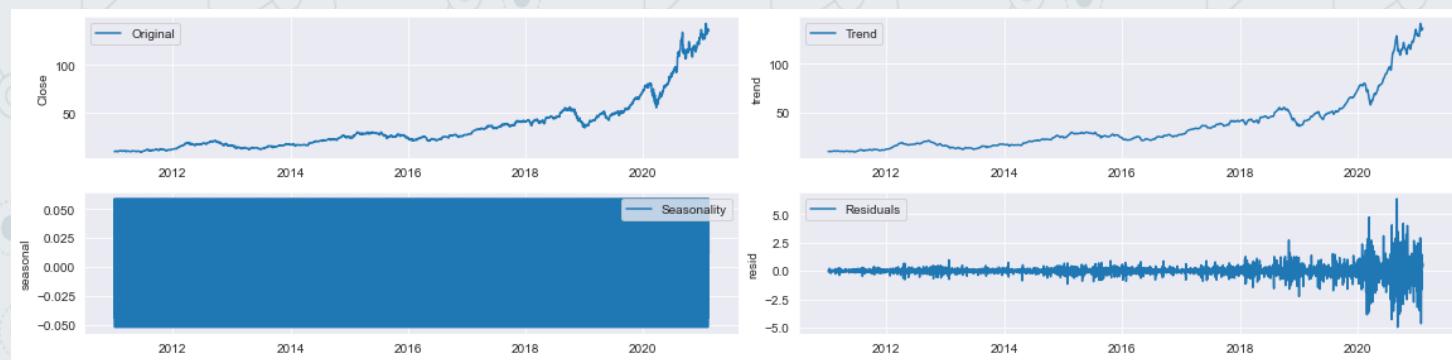
The fact that data are missing should not be neglected—quite often it is an indication of illiquidity. Using an average price results in an underestimate of volatility.

# Exploratory Data Analysis

stock price

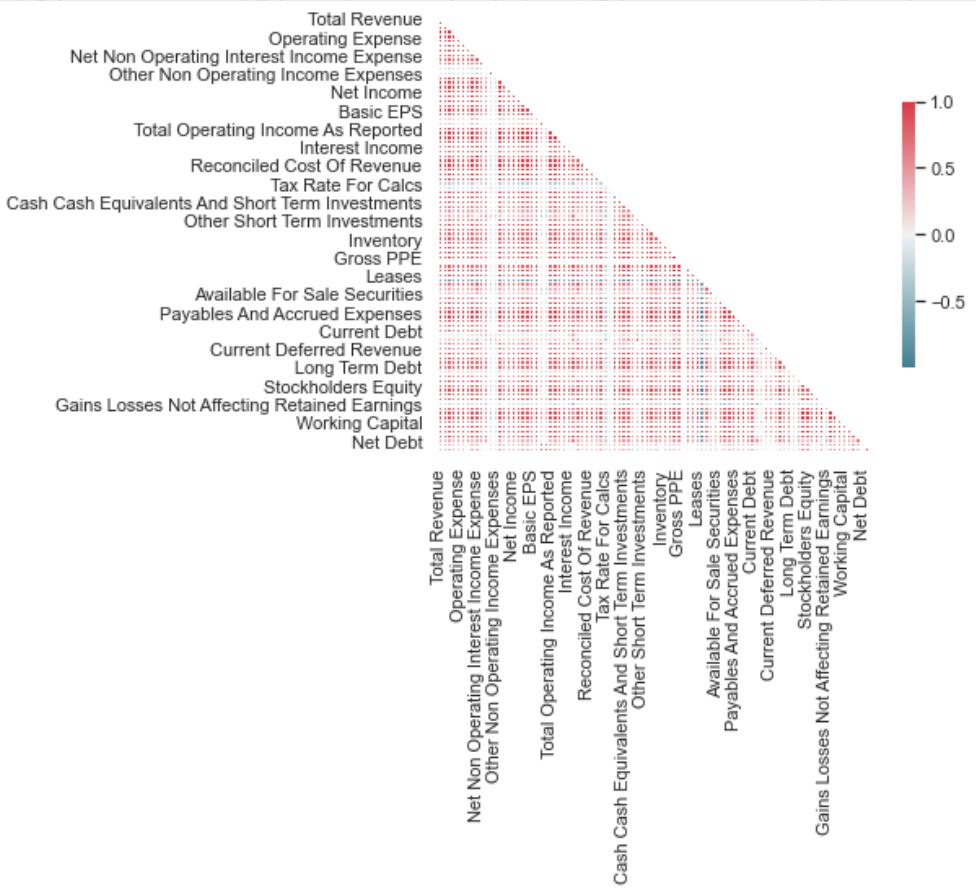


seasonal decompose

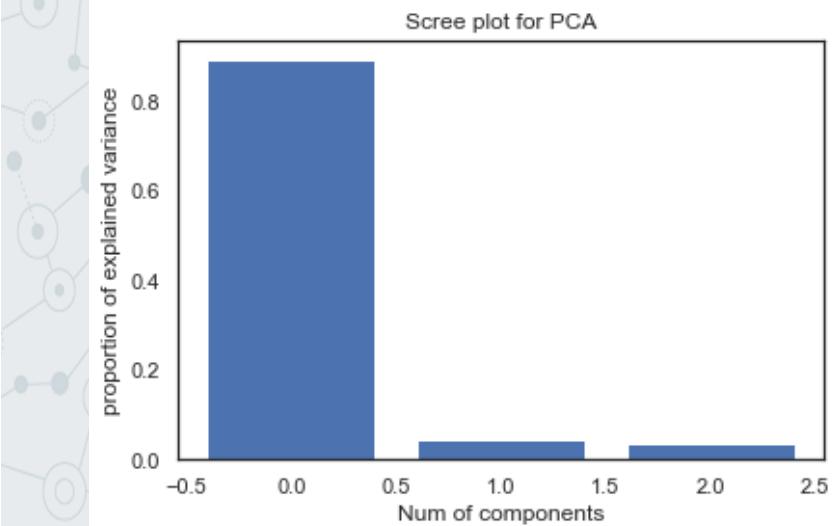


# Classification

## Correlation



## Principle Component Analysis (PCA)



# Classification

Gridsearch with classification algorithms

Training accuracy score:	0.7339201083276913			
Test accuracy score:	0.7307171853856563			
Training F1 score:	0.7339201083276912			
Test F1 score:	0.7307171853856562			
	precision	recall	f1-score	support
buy	0.00	0.00	0.00	105
hold	0.73	1.00	0.84	540
sell	0.00	0.00	0.00	94
accuracy			0.73	739
macro avg	0.24	0.33	0.28	739
weighted avg	0.53	0.73	0.62	739

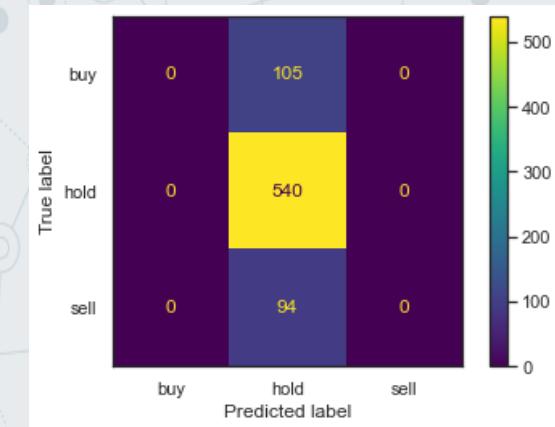
SVM

Training accuracy score:	0.6844660194174758			
Test accuracy score:	0.6740463215258855			
Training F1 score:	0.6844660194174758			
Test F1 score:	0.6740463215258855			
	precision	recall	f1-score	support
buy	0.00	0.00	0.00	462
hold	0.67	1.00	0.81	1979
sell	0.00	0.00	0.00	495
accuracy			0.67	2936
macro avg	0.22	0.33	0.27	2936
weighted avg	0.45	0.67	0.54	2936

random forest

Training accuracy score:	0.6844660194174758			
Test accuracy score:	0.6740463215258855			
Training F1 score:	0.6844660194174758			
Test F1 score:	0.6740463215258855			
	precision	recall	f1-score	support
buy	0.00	0.00	0.00	462
hold	0.67	1.00	0.81	1979
sell	0.00	0.00	0.00	495
accuracy			0.67	2936
macro avg	0.22	0.33	0.27	2936
weighted avg	0.45	0.67	0.54	2936

Xgboost

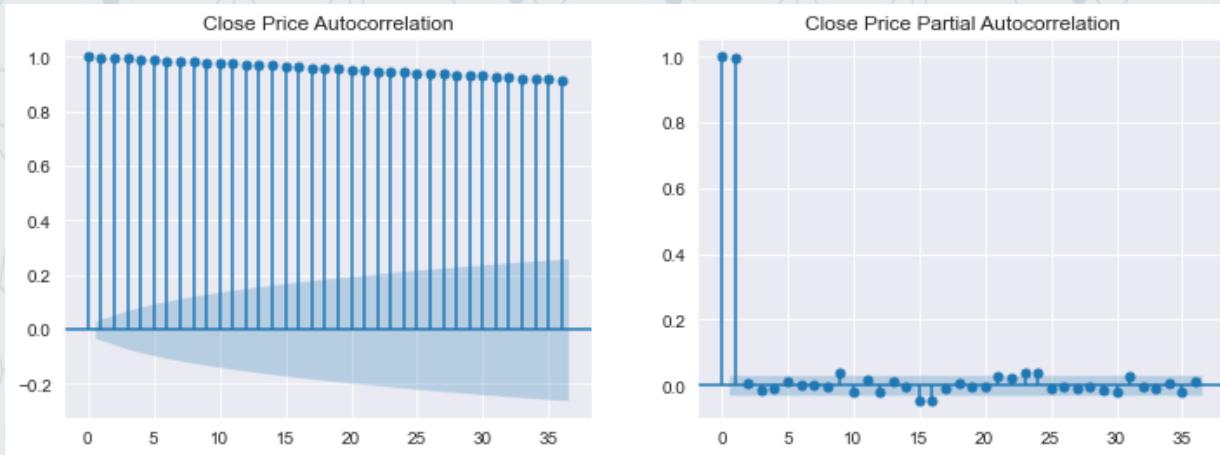


# Time Series

Dickey-Fuller Test

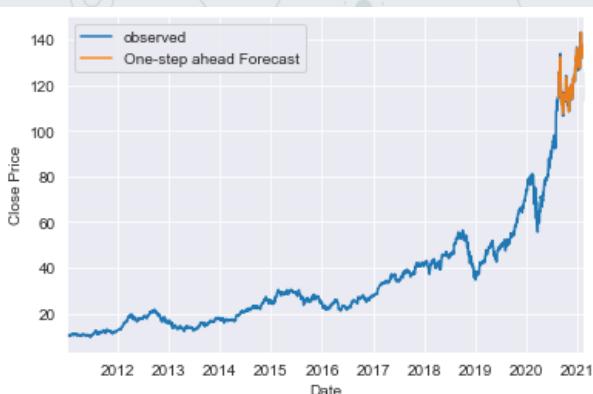


ACF & PACF



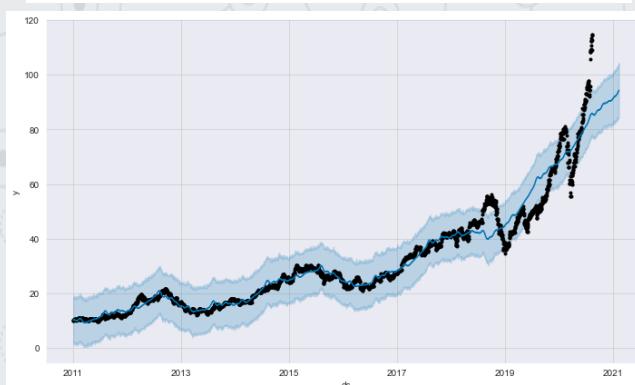
# Time Series

SARIMAX Model  
with  
exogenous features



SARIMAX RMSE of close price: 1.15  
SARIMAX MAE of close price: 0.82

Facebook Prophet



Facebook Prophet close price train RMSE: 4.34  
Facebook Prophet close price test RMSE: 33.89  
Facebook Prophet close price train MAE: 2.59  
Facebook Prophet close price test MAE: 33.05

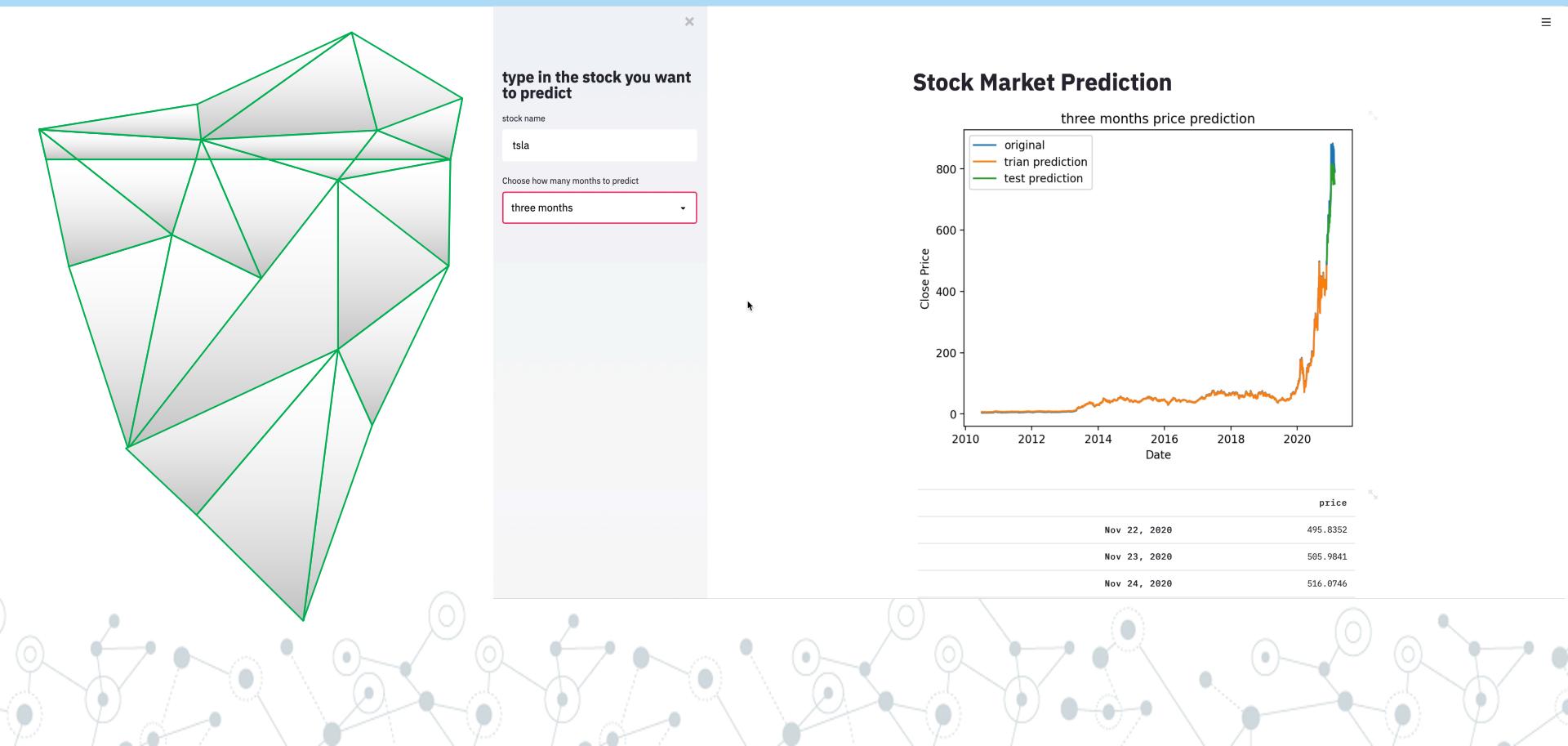
LSTM



LSTM Networks close price train RMSE: 0.70  
LSTM Networks close price test RMSE: 3.28  
LSTM Networks close price train MAE: 0.37  
LSTM Networks close price test MAE: 2.61

# Frontend

Streamlit was used to create a frontend to predict stock price with different time range.



# Next Step

- To access the updated quarterly reports timely and obtain more important features.
- To tune the hyperparameters (exogenous variables) in Time Series models. Technical indicators such as MACD, Stochastic, RSI, etc can be used.
- Besides Twitter, gathering more relevant sentimental data from other web sources.



# Thank you



GitHub: <https://github.com/ghcn345/Stock-Market-Prediction>

streamlit: <http://192.168.1.99:8501>

Blog: <https://kinder-chen.medium.com>

LinkedIn: <https://www.linkedin.com/in/ning-chen-bb1851145/>