Crypto Forecasting

CSE 523 Machine Learning

Prof Mehul Raval

END SEMESTER PRESENTATION

WINTER 2022

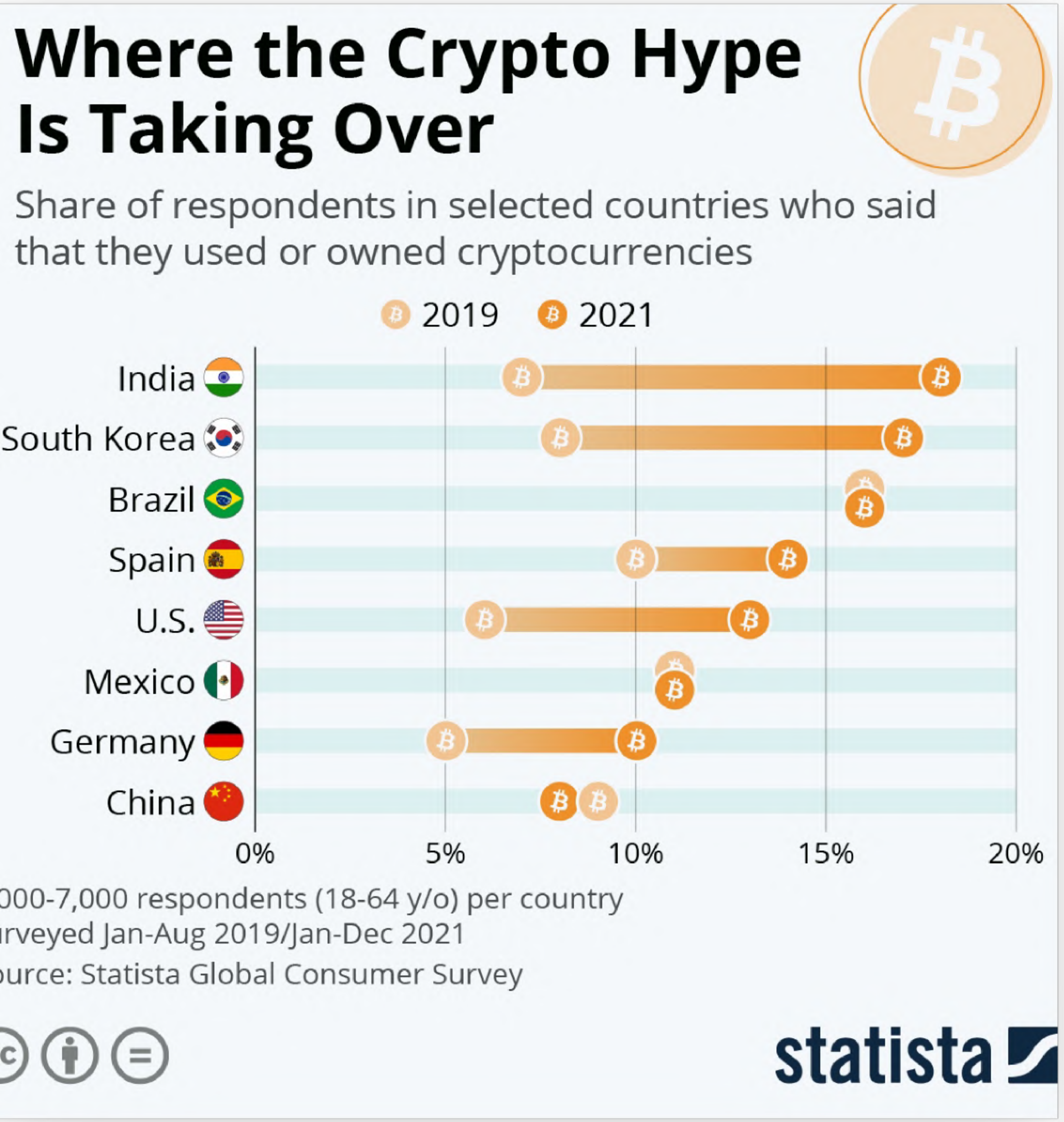
BY CRYPTOPOLICE

KHUSHI SHAH AU1920171

SAMEEP VANI AU1940049

KAVYA PATEL AU1940144

KASHVI GANDHI AU1940175



Introduction

"We have elected to put our money and faith in a mathematical framework that is free of politics and human error."
- Tyler Winklevoss

Cryptocurrencies are one of the most popular assets for speculation and investments but they are highly volatile. The fluctuating nature causes both hype and risk.

Crypto Forecasting is a process of utilizing the time-series data of given cypto currencies along with other features to predict the future value of these crypto currencies by analysing the past trends and data.

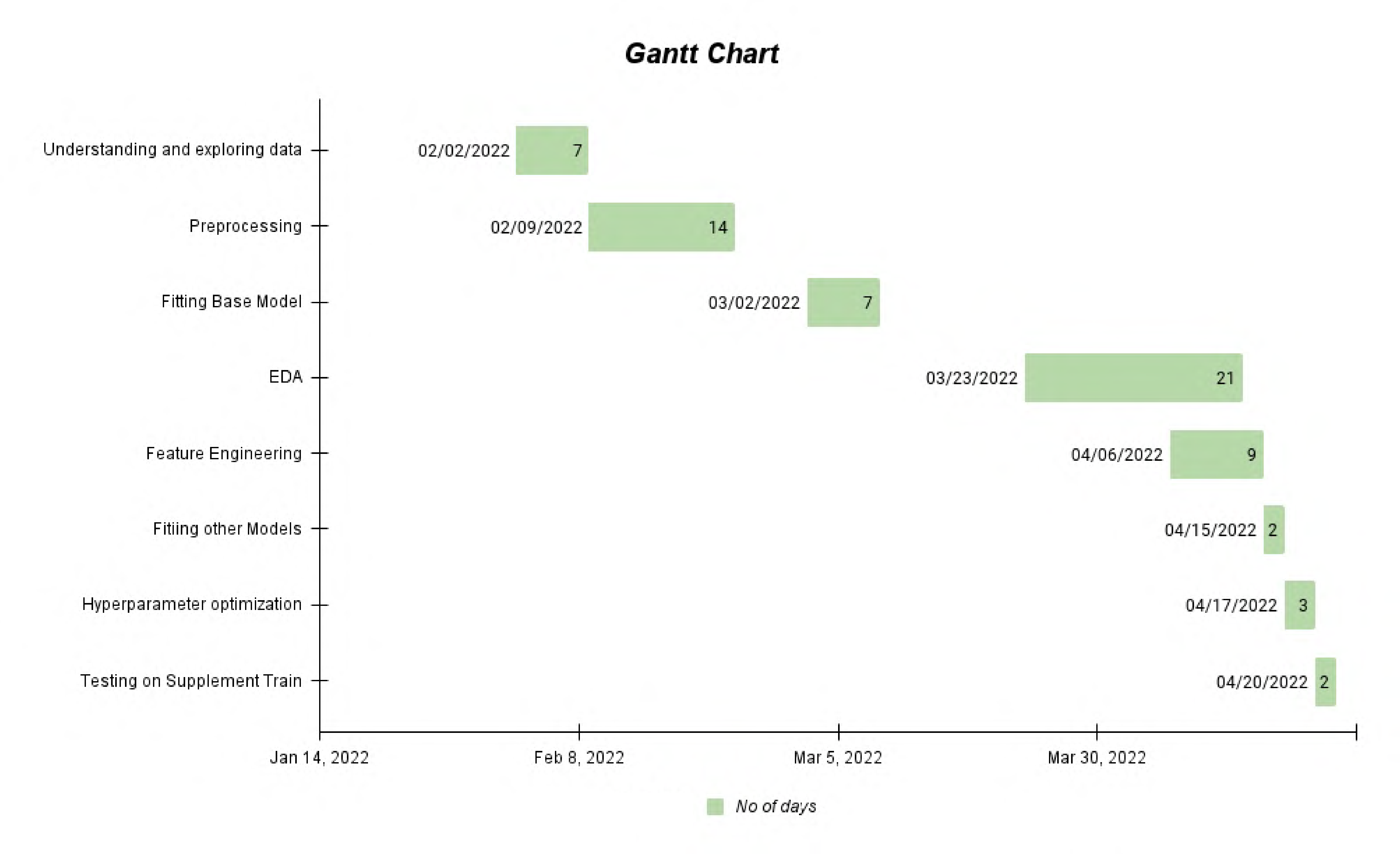
Through Crypto Forecasting, we intend to forecast short term return in 14 popular cryptocurencies using Machine Learning.

Reference: https://www.statista.com/chart/27070/cryptocurrency-use-selected-countries-over-time/

Problem Statement

Given the time series data, the challenge is to predict the future returns of 14 cryptocurrencies using various regression and time-series models.

GANTT CHART SHOWING PROJECT PROGRESS



Existing Body of work

Time-series forecasting of Bitcoin prices using high-dimensional features: a machine learning approach

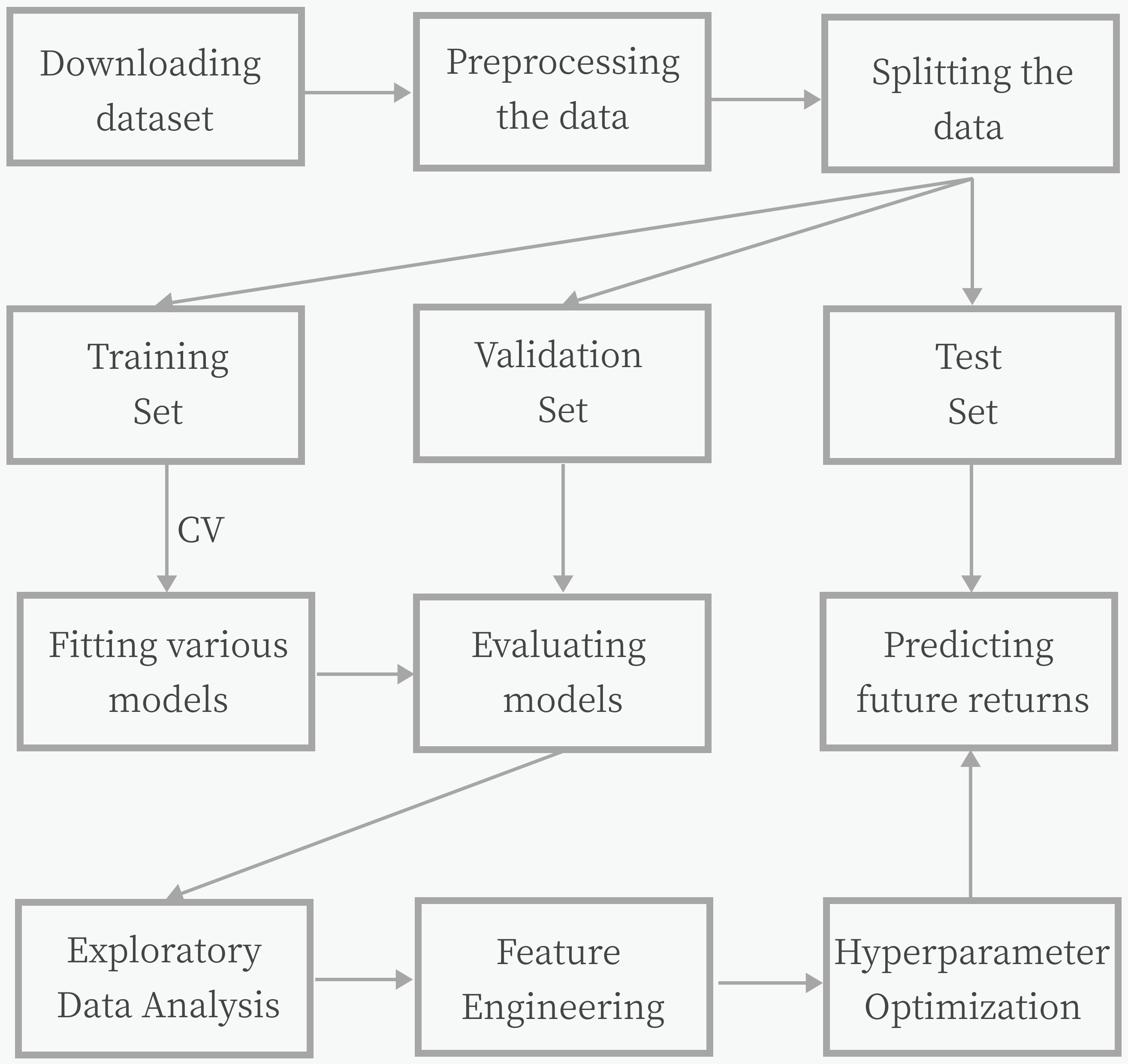
Forecasting and trading cryptocurrencies with machine learning under changing market conditions

Cryptocurrency price prediction using traditional statistical and machine - learning techniques: A survey

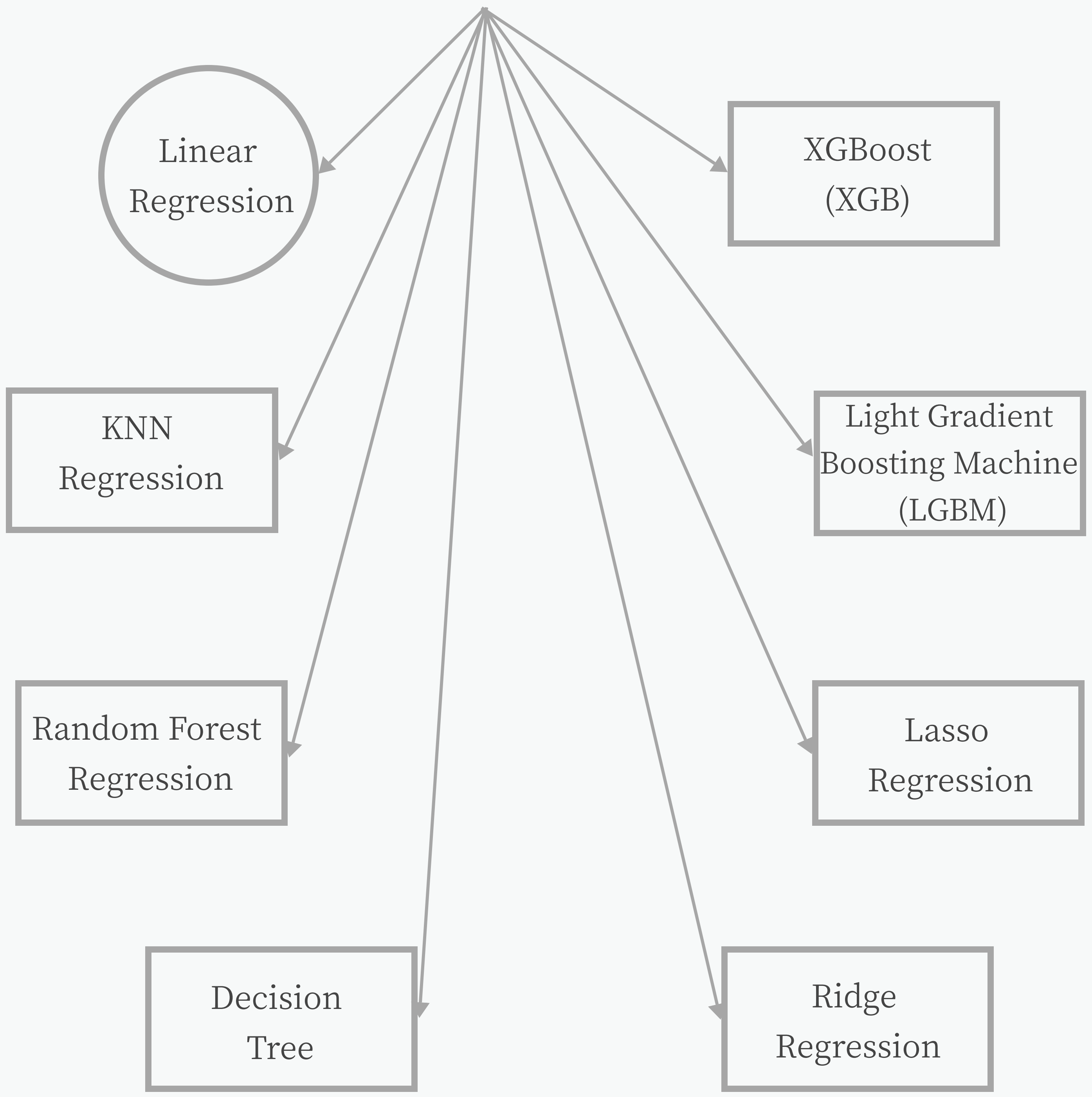
Machine Learning Strategies for Time Series Forecasting

APPROACH

OVERVIEW



MODEL SELECTION



APPROACH

DATA PREPROCESSING

1. Null value detection
2. Timestamp aggregation
3. Standardizing data

FEATURE ENGINEERING

1. Removing features with high correlation.
2. Creating new features.

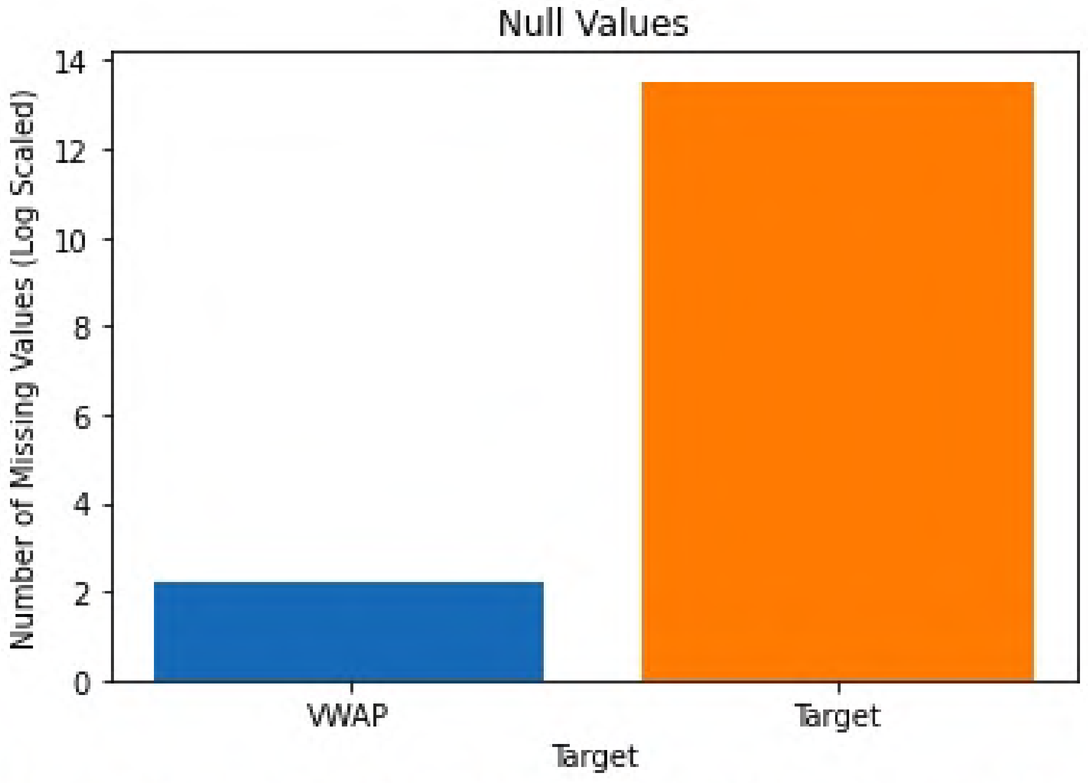
EXPLORATORY DATA ANALYSIS

1. Finding correlation between features
2. Detecting outliers
3. Plotting Feature vs Target graph
4. Seasonal trend
5. Lag features
6. Distribution of each feature

HYPERPARAMETER OPTIMIZATION

1. Grid Search
2. Optuna

NULL VALUES




```
Features with null values:  ['VWAP', 'Target']
```

TIMESTAMP AGGREGATION

timestamp
2018-01-01 00:01:00
2018-01-01 00:01:00
2018-01-01 00:01:00
2018-01-01 00:01:00
2018-01-01 00:01:00

→

new_date
2018-01-01
2018-01-02
2018-01-03
2018-01-04
2018-01-05

RESULTS

DATA PRE-PROCESSING

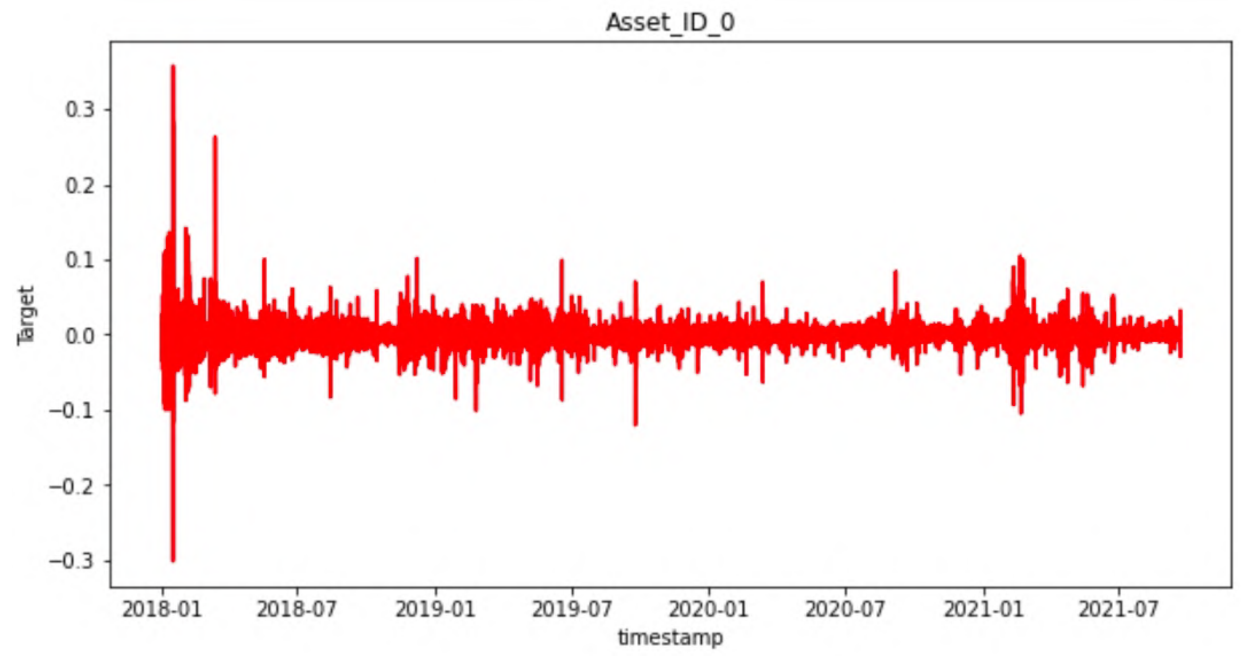
1. Null Values
2. Time stamp aggregartion

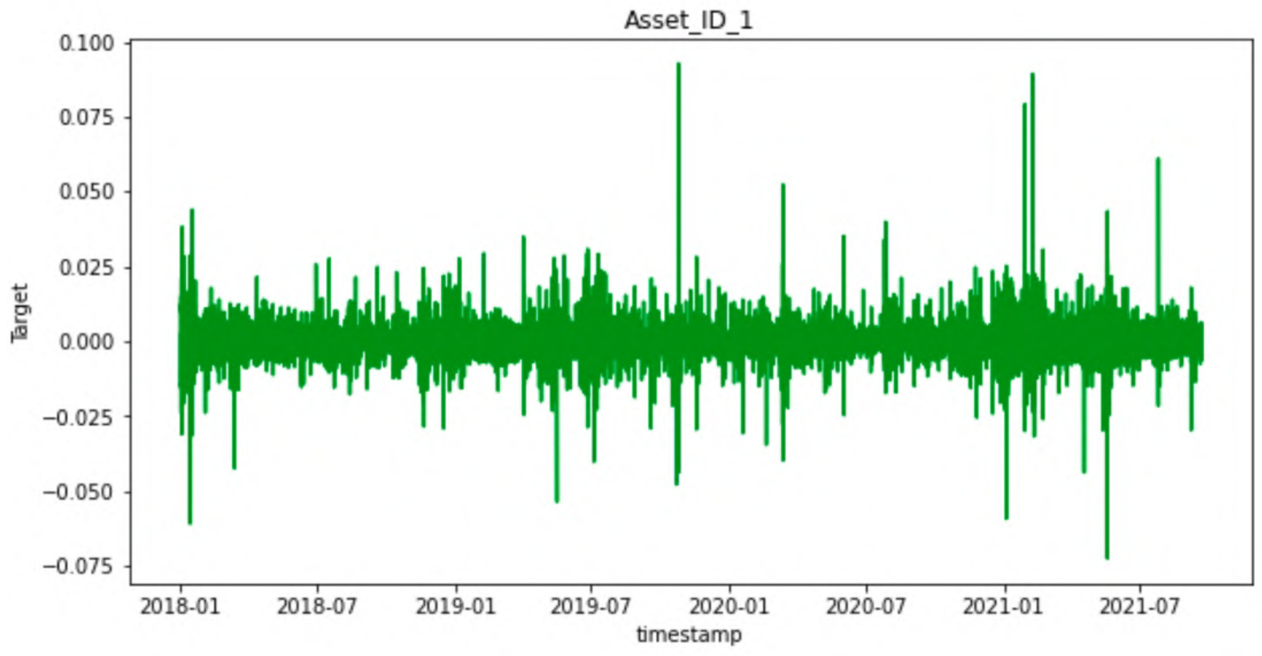
BASE MODEL - LINEAR REGRESSION

MSE = 3×10^{-5}

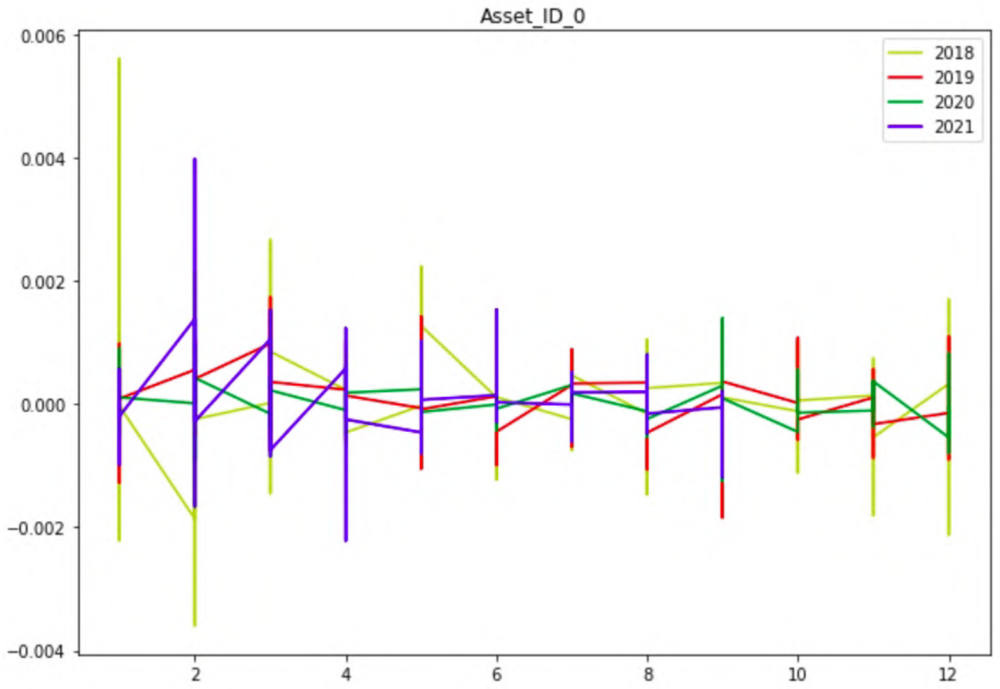
Conclusion - Can be a good fit.

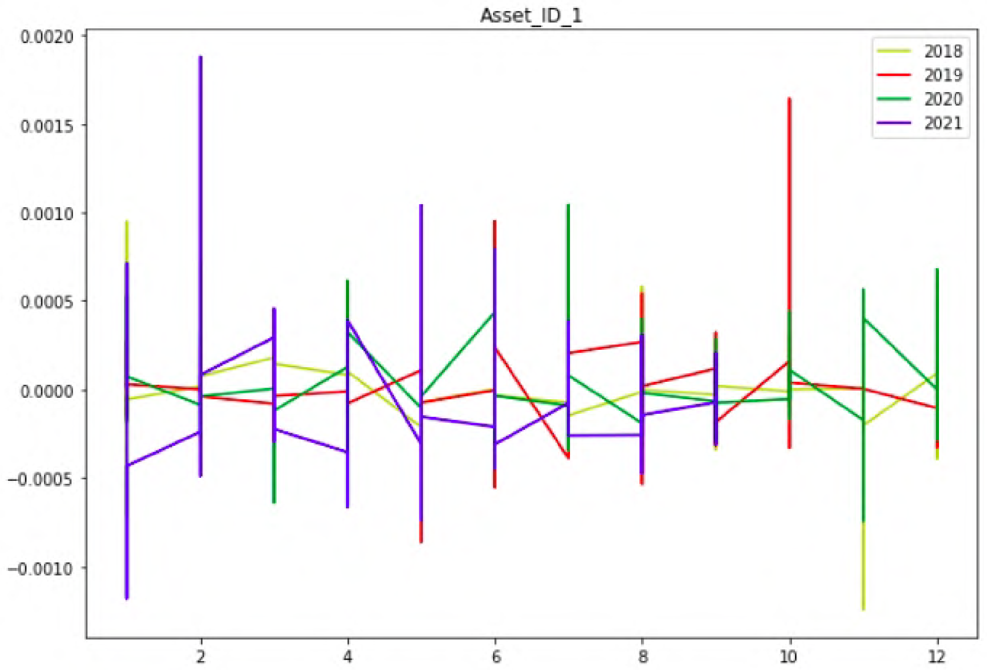
LINEAR REGRESSION TIMESTAMP VS TARGET





SEASONALITY GRAPH





RESULTS

EXPLORATORY DATA ANALYSIS (EDA)

Seasonal Trends

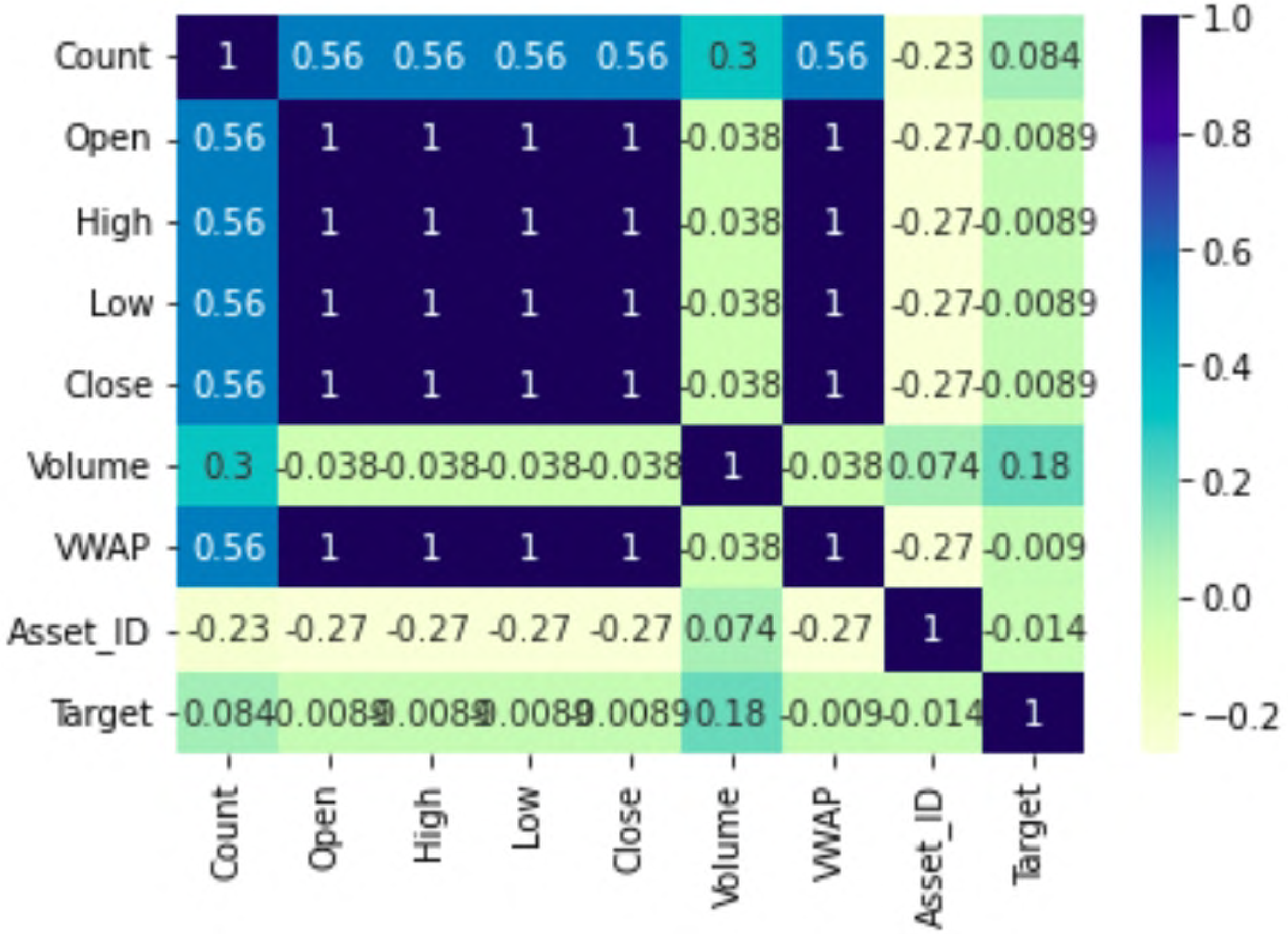
'Target' vs 'Month'

No strong seasonality

Correlation Graph

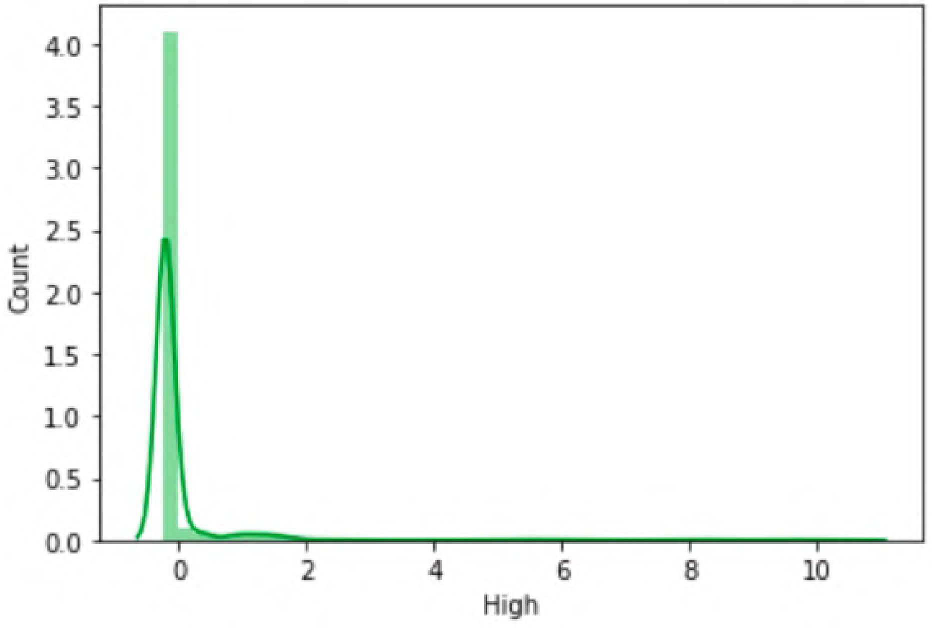
Conclusion - Removing features that are highly correlated

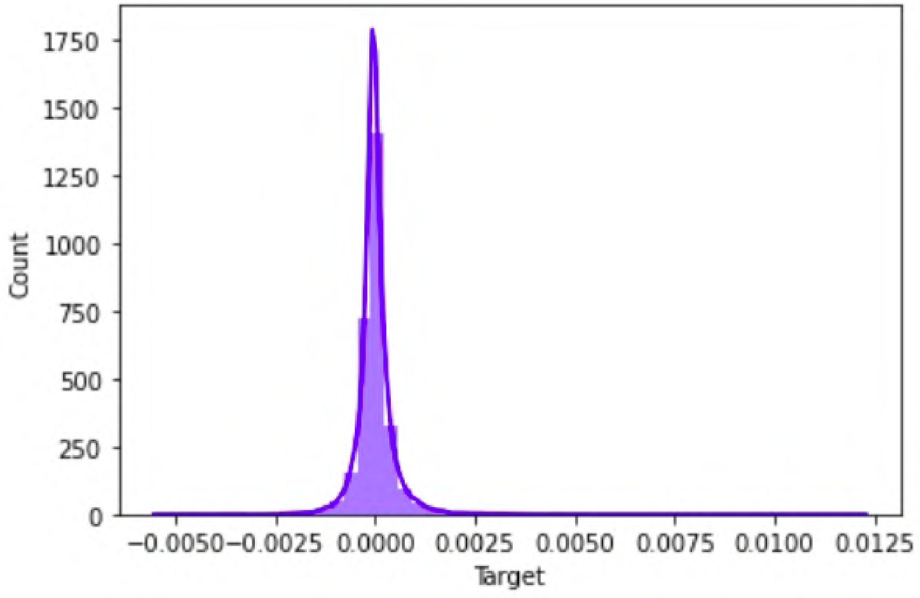
CORRELATION GRAPH FOR ALL ASSET ID



RESULTS

DISTRIBUTION OF FEATURES





EXPLORATORY DATA ANALYSIS (EDA)

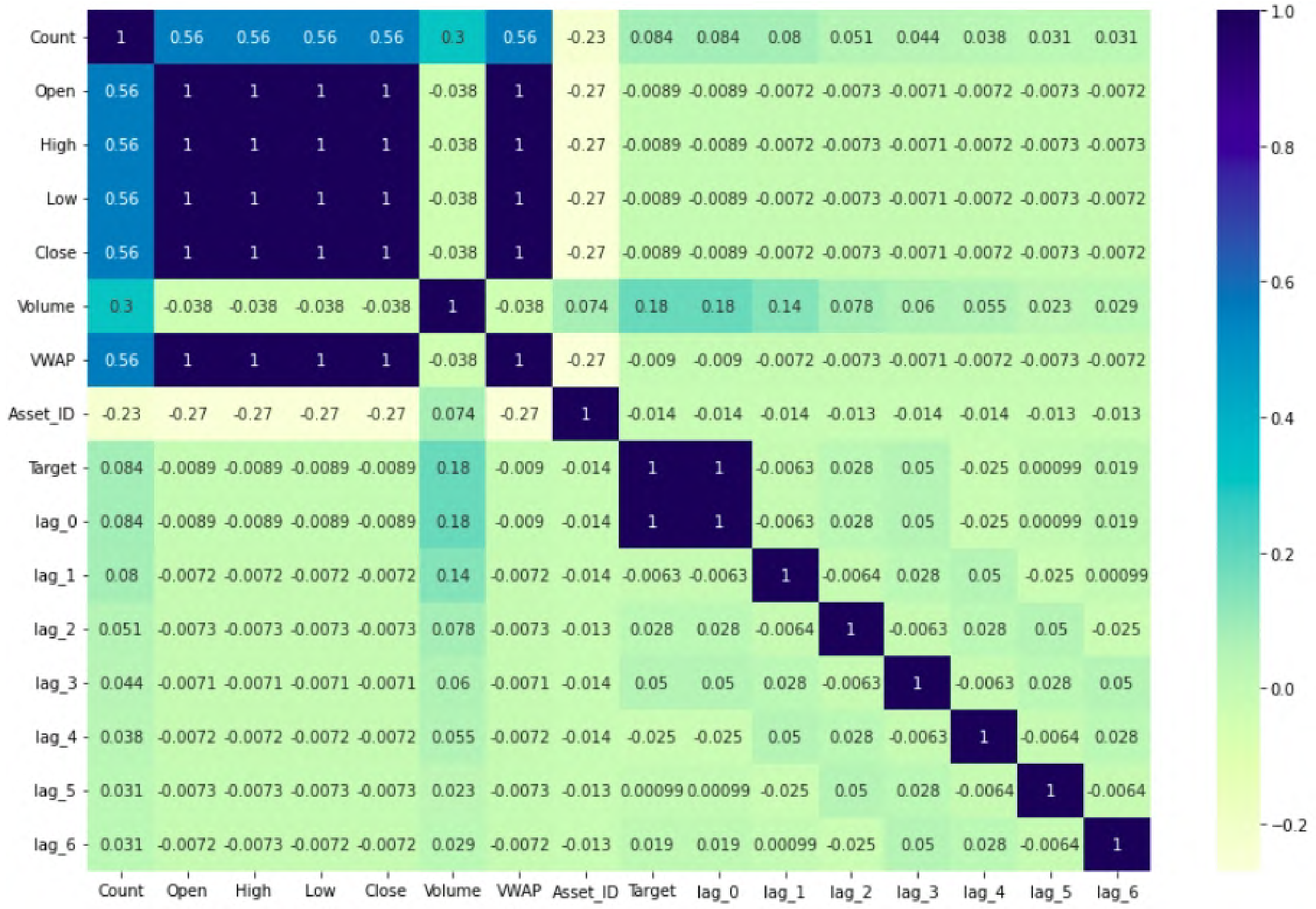
<u>Distribution of features</u>

1. Skewed Guassian distribution
2. Mean near to 0
3. Low variance
4. High kurtosis

<u>Lag features</u>

The correlation is not strong with 'Target'.

CORRELATION HEATMAP FOR LAG FEATURES



Pitch

FEATURES WITH **HIGH CORRELATION :**

1. Open
2. Close
3. High
4. Low
5. VWAP

NEW FEATURES

Range_Close_Open	Range_High_Low
0.009311	-0.106561
0.009823	-0.106239
0.009775	-0.106443
0.010801	-0.105861
0.011829	-0.102587

RESULTS

FEATURE ENGINEERING

Dropping columns with high correlation

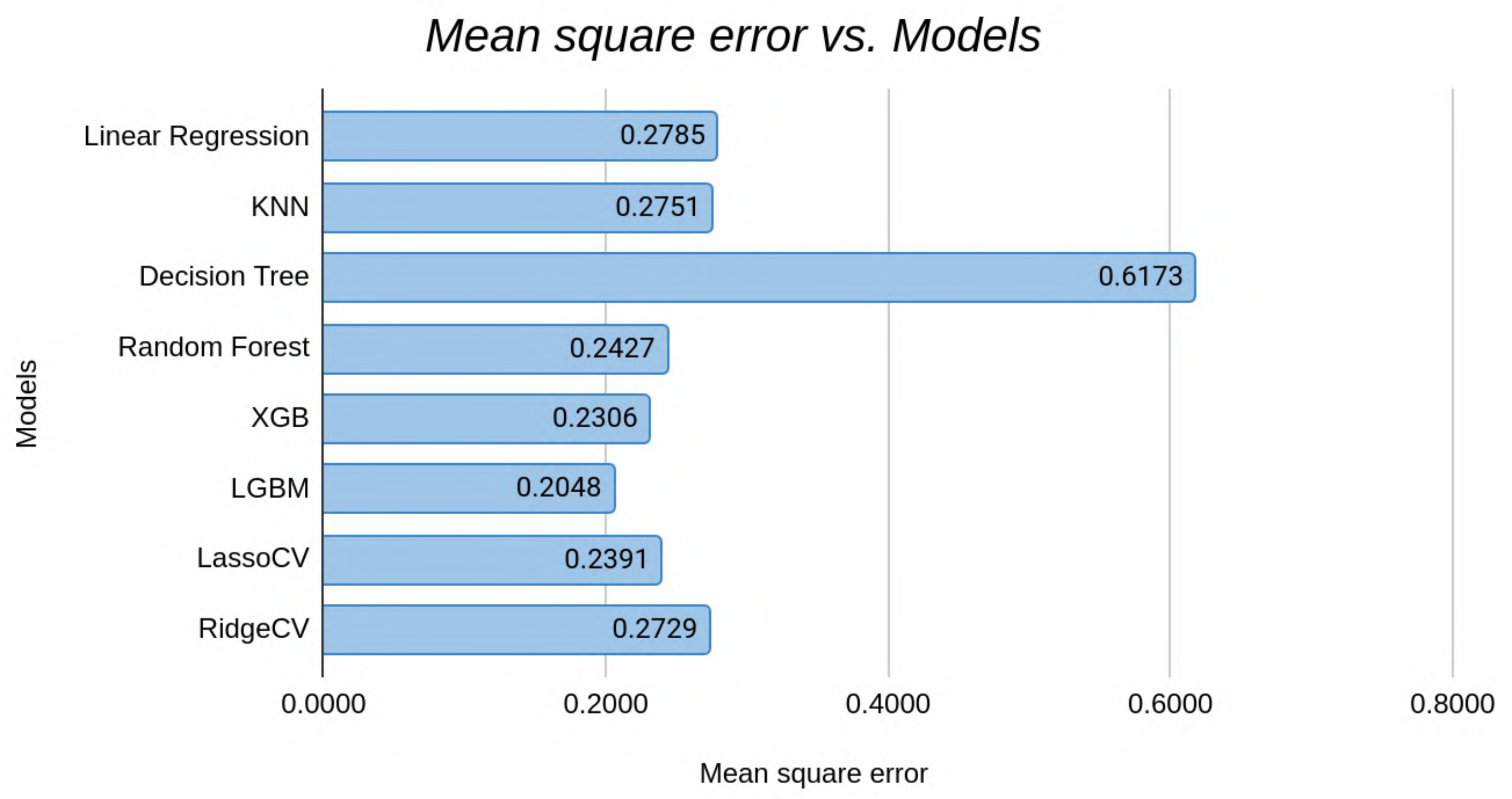
Adding new features
1) Open-Close range
2) High-Low range

MODEL FITTING

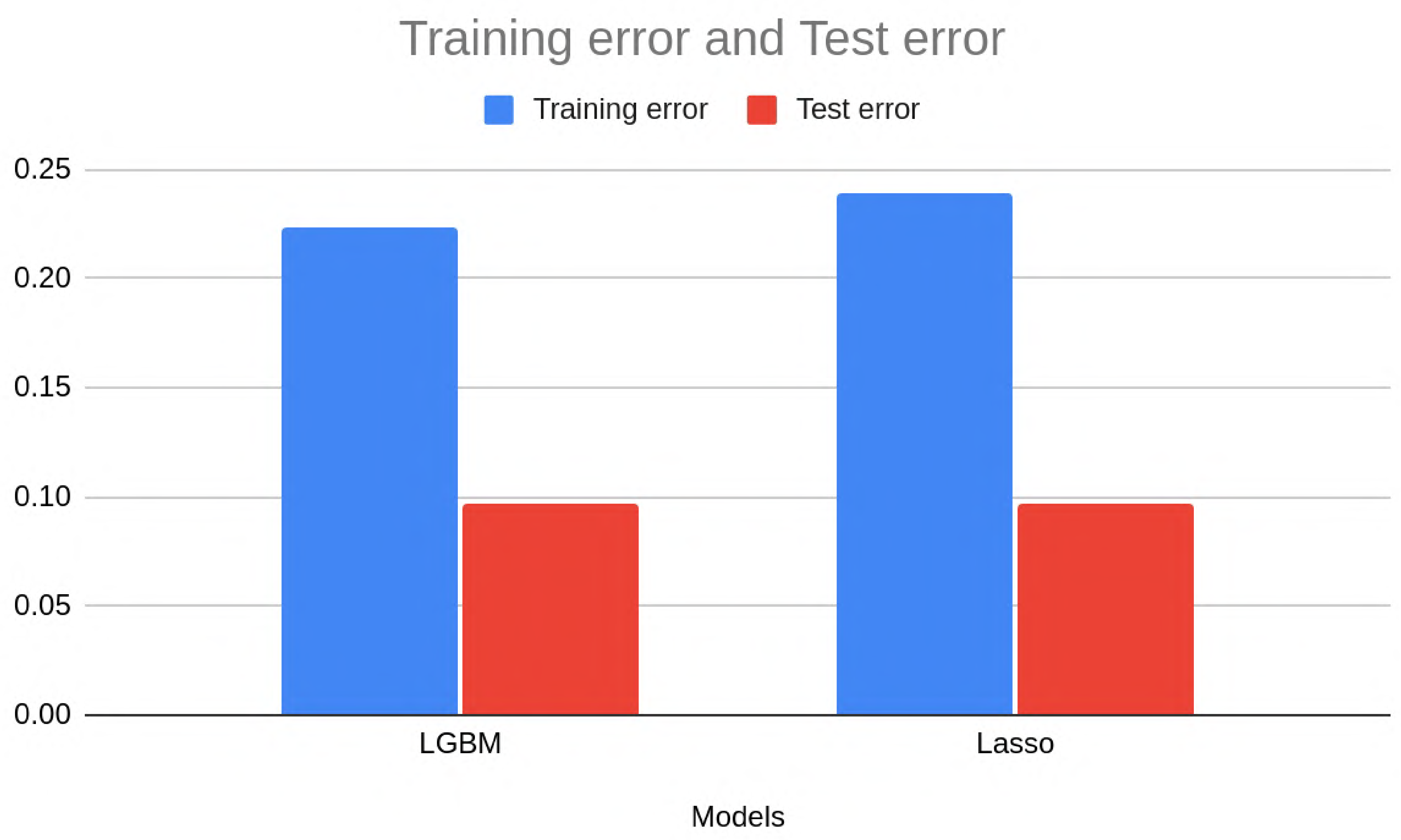
Performance metrics - Mean Sqaure Error

Top 3 Models :
1) LGBM
2) XGB
3) LassoCV

MODEL VS MSE



Pitch



RESULTS

HYPERPARAMETER TUNING

1) Optuna
2) Grid Search
The optimizer that gave better results was chosen.

<u>Error for XGB model</u>
Training error : 0.23056
Test error : 200
Conclusion - Overfitting

Generalization gap lower for LGBM

Conclusion - Final model is LGBM.

CONCLUSION

- Not all features of the dataset are needed to predict the returns
- Lasso and LGBM perform the best on Test Data Set
- There is a significant improvement from the base model.
- XGBoost overfits on Training Data Set
- The lowest Test Error achieved is 0.09 MSE

ROLE OF EACH MEMBER

Role	Name
Data Preprocessing	Khushi, Sameep, Kavya, Kashvi
Model Fitting	Khushi, Sameep -- Linear Regression, Lasso, KNN, XGBoost Kavya, Kashvi -- Random Forest, Decision Tree, LGBM, Ridge Regression
EDA , Feature Engineering	Khushi, Sameep, Kavya, Kashvi
Hyperparameter Optimization	Khushi, Kashvi -- Grid Search Sameep, Kavya -- Optuna

References:

1. A Comprehensive Guide to Time Series Analysis - Analytics Vidhya
2. Time-Series Forecasting with Spark ML: Part—1
3. G-Research Crypto Forecasting
4. The Complete Guide to Time Series Analysis and Forecasting
5. Learn Time Series Tutorials
6. How-To Guide on Exploratory Data Analysis for Time Series Data
7. Exploratory Data Analysis - Kaggle Source
8. Tuning Hyperparameter with Optuna
9. Guide to LightGBM Hyperparameter Tuning with Optuna
10. Grid Search for Model Tuning
11. Tune Hyperparameters with GridSearchCV