REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

REVA Academy for Corporate Excellence (RACE)

Topic: Trading Analytics for Day Trading in Stock Market

Name of the Presenter(s)

Anand Mohan

Batch:MBA06

Trimester: THIRD TRIMESTER

SRN: R19MBA53

Mentor: J B SIMHA

www.race.reva.edu.in

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Agenda

- Introduction
- Problem Statement
- Technical Solution Design
- Framework – CRISP-DM
- Analysis
- Future Scope

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Introduction

- **Exchange-high volatility-New field for Researchers!**
- **Numerous Stock Prediction Techniques today**
- **consistency of prediction Performance remains debatable.**
- **Short Prediction-High Frequency Trading via Broker -Eat up the potential profit due to high commission rate**
- **Exploring Machine Learning ways for Stock Prediction.**
- **Information extracted from returns data of HDFC Stocks between 2000 to 2021 being utilized**
- **Various Machine Leaning Algorithms using Regression Models explored**
- **Machine Learning Algorithms ranked based on their relative expected outcome**

Problem Statement

- **Exchange is technically accessible to traditional voters.**
- **Still, these markets are dominated by big investors, perpetuating the wealth divide.**
- **With algorithmic Trading, unsuccessful, long-term predictions on company filings, any derived and associated formula may go fine on back testing in controlled environments, but the main challenge is live testing.**
- **Algorithmic Trading gives many benefits but retail investors do not have the desired technology to create such systems.**
- **Hence, we will introspect more on simple and easy to apply Modelling Mechanisms using Statistical and Machine Learning approaches.**
- **investors should be able to utilize the paper's findings to assist them to guide their quality allocation and create buy-sell selections that best meet their needed returns expectations with High Accuracy and Minimizing risks.**

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956, Approved by AICTE, New Delhi

Technical Solution Design

Data Extraction:

Daily Data of HDFC company from the year 2000 to 2021

Data Preprocessing:

→ Handling Missing values

→ Features Addition

→ **MinMax Scaler**

Modeling / Classifiers:

→SMA EMA T Test Metrics

→SMA EMA Z Test Metrics

→Auto Keras Classification

→KNN Classification

→Logistic Regression

→ARIMA

→ OLS Regression

→ Lasso Regression-CV

→ k-Nearest Neighbours Model

→Decision Tree

→GridSearchCV

→Random Forest

→XGBoost Model

Validation:

→Accuracy

→Precision

→Recall

→MAE

→MSE

→RMSE

→MAPE

Phase -1

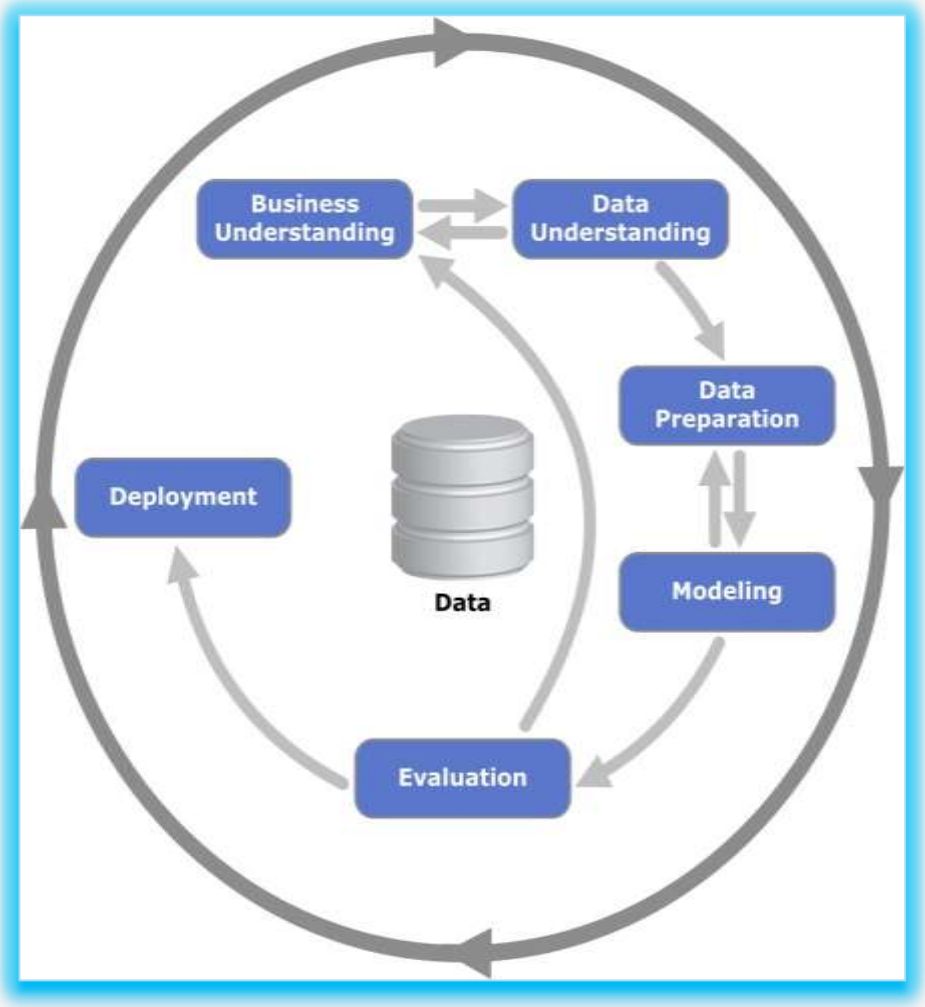
Deployment:

→Dashboard

→State of the art API

Phase - 2

Framework-Crisp DM



Business Understanding

Data Understanding

Data Preparation

Modelling

Evaluation

Deployment

Business Understanding

There exist 2 main conventional approaches to the analysis of the stock markets:

(1) Fundamental analysis

(2) Technical analysis.

Modern Approaches for stock exchange Prediction:

1. Hypothesis Testing:

The null Hypothesis (H0) is assumed to be true. Hypothesis testing starts by stating and assuming a null hypothesis then the method determines whether or not the belief is probably going to be true or false.

The vital purpose to notice is that there is a component of doubt concerning validity of Null Hypothesis. whatever data that is against the declared null hypothesis is captured within the alternative Hypothesis (H1).

2. ARIMA Model:

Business Understanding

3. Machine Learning Approach

3. Prediction with Deep Learning

Long Short-Term Memory (LSTM

Auto Keras

5. Sentiment Analysis Approach:

The most-used platforms on the net are social media. it is calculable that social media users everywhere around the globe can number around 3.07 billion. there is a high association between stock costs and events associated with stocks on the net.

The Stock Twits, Twitter, and Yahoo Finance are the typical medium for sentiments research.

Data Understanding

Daily Data of HDFC company from the year 2000 to 2021 which is traded on the stock exchange in India, is being used for this study. To properly scan stocks, we should tend to first perceive what every column within the stock chart means:

Date	Symbol	Series	Prev Close	Open	High	Low	Last	Close	VWAP	Volume	Turnover
2000-01-03	HDFCBANK	EQ	157.40	166.00	170.00	166.00	170.00	170.00	169.52	33259	5.638122e+11
2000-01-04	HDFCBANK	EQ	170.00	182.00	183.45	171.00	174.00	173.80	174.99	168710	2.952261e+12
2000-01-05	HDFCBANK	EQ	173.80	170.00	173.90	165.00	168.00	166.95	169.20	159820	2.704094e+12
2000-01-06	HDFCBANK	EQ	166.95	168.00	170.00	165.30	168.95	168.30	168.44	85026	1.432166e+12
2000-01-07	HDFCBANK	EQ	168.30	162.15	171.00	162.15	170.75	168.35	166.79	85144	1.420158e+12

- **The previous close nearly always refers to the previous day's final price of security once the market formally closes for the day.**
- **The opening price is the first trade price that was recorded. This is usually employed in relevance to the present price or the closing price from the previous day session to quantify the stock's movement.**
- **The last price is the one at which the foremost recent transaction happens.**
- **The close is the last price once the market is closed on the day.**

Data Understanding

- **If the closing price is up or down over five-percent more than the previous day's shut, the whole listing for that stock is listed in bold. It is to be noted that you are not almost certain to get this price if you get the stock consecutive day the reason being that the share price is consistently dynamic (even when the exchange is closed for the day).**

- **The volume-weighted average worth (VWAP) may be a technical analysis indicator used on intraday charts that resets at the beginning of each new commerce session. it is a business benchmark that represents the typical price which the security listed throughout the day, based on both volume and price.**

- **Trading Volume shows the number of shares listed for the day, listed in lots of 100 quantities of shares.**

- **Share turnover may be an estimation of stock liquidity, calculated by dividing the whole number of shares traded throughout some period by the average number of shares outstanding for the same duration of time.**

Data Preparation

Handling Missing values: Three of the features—Trades', 'Deliverable Volume',' % Deliverable had quite one hundred periods missing values therefore those columns will need to be dropped as they are having several missing values.

It is needed to refrain from implementing the mean, media, and mode imputation methodology because those might render values that may introduce bias into our dataset. Second, the strategy solely looks at the variable itself and therefore might come up with values that don't seem to be representative of trends within the dataset.

Features Addition: In the dataset, computed variables are added that for sure would influence stock returns. These are moving averages for rolling periods of seven days,13 days,20 days,100 days, and two hundred days. Exponential moving averages is being conjointly enclosed for seven days,13 days,20 days,100 days, and two hundred days. These derived features were useful in evaluating the securities market returns.

Data Preparation

```
scaler = MinMaxScaler()
feature_transform = scaler.fit_transform(df[features])
feature_transform= pd.DataFrame(columns=features, data=feature_transform, index=df.index)
feature_transform.head()
```

Date	Prev Close	Open	High	Low	Last	VWAP	Volume_lag_1d
2000-01-04	0.002747	0.008258	0.006438	0.005841	0.004583	0.005641	0.000320
2000-01-05	0.004329	0.003266	0.002484	0.003338	0.002083	0.003237	0.001667
2000-01-06	0.001478	0.002434	0.000869	0.003463	0.002479	0.002922	0.001579
2000-01-07	0.002040	0.000000	0.001283	0.002149	0.003229	0.002237	0.000835
2000-01-10	0.002060	0.004472	0.004803	0.003338	0.001375	0.002652	0.000836

MinMax Scaler is one of the approaches to data scaling that is being used.

Here, the minimum of features is created up to zero, and the most of features are up to one.

MinMax Scaler shrinks the data inside the given range, sometimes from zero to one.

It transforms data by scaling variables to a given range.

It scales the price to a selected value range while not varying the form of the initial distribution.

Modelling

HDFC excel data is put in Tabular form in step 1.

Step 2: The time series data is plotted for the HDFC stock that is provided as a dataset for the project for all ten years.

The 7-day moving average time series data is added in step 3.

Step 4: The data for a 7-day moving average time series is being plotted.

Step 5: The data from a rolling 7-day moving average is included in the Data frame.

Step 6: It is determined whether the closing price value on a certain prior day was lower or higher than the current 7-day moving average.

Modelling

Step 7: The same step is performed for the moving averages of 13 days, 20 days, 100 days, and 200 days.

Step 8: Exponential Moving Average is used to recreate the five different models created using Simple Moving Average.

Step 9: ARIMA Time series modelling is used to create an additional five different models.

The construction of all 15 models, as seen above, will be used to forecast day trading in the stock market.

When the majority of the 15 various models or all of them move in the same direction, a choice on whether to purchase or sell the stock must be made.

Modelling

various Classification models namely AutoKeras Classification Model (Structured Data Classifier), K-neighbours Classifier Model, and Logistic Regression Classification Model deployed and their prediction accuracy is being compared with Simple Moving Average Models, Exponential Moving Average Models, and ARIMA Models.

further ahead various Regression Models including both Machine Learning and Deep learning techniques are deployed and Metrics namely Mean Absolute error and Mean Absolute percentage errors are deployed to estimate the quality of the predictions on the close price of the HDFC share.

These Regression Models are Ordinary Least Squares(OLS)-Linear Regression Model, Lasso Regression Model, Lasso regression Model Using Cross Validation, The k-Nearest Neighbours(KNN) Algorithm, Decision Tree Algorithm, GridSearchCV Algorithm with Hyperparameter Tuning, Random Forest Regression Model, XGBoost ML Model, Using Principal Component Analysis (PCA) with LSTM, Using Principal Component Analysis (PCA) with LSTM with Moving Average variables(Feature Engineering), Long Short-Term Memory(LSTM) Neural Network Model, Regression Model using AutoKeras.

Data Evaluation

Initially, A rule-based model is being developed to try to do hypothesis testing to work out whether or not the chosen stock's price is crossing any of the moving averages. Then prediction based on the Hypothesis Testing Rule is decided that is employed as a Metric to work out the accuracy for predicting the upward Trend or Downward trend of the HDFC shares.

Data Evaluation

A few Classifications Based Models will be conjointly built. Metrics being employed for classification Models would be accuracy score and confusion matrix which can facilitate in determining the accuracy of predicting the upward Trend or Downward trend of the HDFC shares.

Auto Keras Accuracy score

```
#PRECISION AND RECALL
print("ACCUIRACY SCORE")
print(metrics.accuracy_score(test_y,y_predict))
```

```
ACCUIRACY SCORE
0.8491988689915174
```

```
#PRECISION RECALL MATRIX
print("precision/recall Metrics")
print(metrics.classification_report(test_y,y_predict))
```

```
precision/recall Metrics
              precision    recall  f1-score   support

           0       0.76      0.98      0.86       488
           1       0.97      0.74      0.84       573

    accuracy                           0.85      1061
   macro avg       0.87      0.86      0.85      1061
weighted avg       0.88      0.85      0.85      1061
```

K neighbors Classifier Accuracy score

```
#PRECISION AND RECALL
print("ACCUIRACY SCORE")
print(metrics.accuracy_score(test_y,predicted_values))
```

```
ACCUIRACY SCORE
0.7408105560791706
```

```
#PRECISION RECALL MATRIX
print("precision/recall Metrics")
print(metrics.classification_report(test_y,predicted_values))
```

```
precision/recall Metrics
              precision    recall  f1-score   support

           0       0.69      0.81      0.74       488
           1       0.80      0.69      0.74       565
           2       0.00      0.00      0.00         8

    accuracy                           0.74      1061
   macro avg       0.50      0.50      0.50      1061
weighted avg       0.74      0.74      0.74      1061
```

Logistic Regression Accuracy score

```
#PRECISION AND RECALL
print("ACCUIRACY SCORE")
print(metrics.accuracy_score(test_y,classes))
```

```
ACCUIRACY SCORE
0.9010367577756834
```

```
#PRECISION RECALL MATRIX
print("precision/recall Metrics")
print(metrics.classification_report(test_y,classes))
```

```
precision/recall Metrics
              precision    recall  f1-score   support

           0       0.97      0.83      0.89       488
           1       0.86      0.98      0.91       565
           2       0.00      0.00      0.00         8

    accuracy                           0.90      1061
   macro avg       0.61      0.60      0.60      1061
weighted avg       0.90      0.90      0.90      1061
```

Data Evaluation

ARIMA Modelling:

Five ARIMA models are created using Moving Average as the Target variable because it would smoothen the curve for the close price of the HDFC stock price. When a model for prediction functions in statistic Time series analysis is created, a stationary statistic Time series for a higher prediction is required.

ADF (Augmented Dickey-Fuller) test is a statistical significance test which means the test will end up in hypothesis tests with null and alternative hypotheses. As a result, we will have a p-value from that we will have to be compelled to create inferences regarding the Time series, whether or not it is stationary or not.

To perform the ADF test in any statistic package, the stats model provides the implementation operation adfuller (). Function adfuller () provides the subsequent data particularly p-value, Value of the test statistic, Number of lags for testing consideration, and critical values.

if the results of the ADF test are bigger than 0.05 then we were required to fail to reject Null Hypothesis H0 and are available to reasoning that point Series is not Stationary. If the results of the ADF test would are lesser than 0.05 then we were required to reject Null Hypothesis H0 and are available to reasoning that Time Series data is Stationary.

Data Evaluation

Different Regression Models are built using each of the Machine Learning and Deep Learning algorithms to work out the Accuracy in predicting the expected close price of the HDFC stock that is that the Target or dependent variable for the Modelling Algorithms.

The metrics that is required to be verified for the accuracy of predictions in the case of regression Modelling are Mean Absolute Error (MAE), Mean square Error (MSE), Root Mean square Error (RMSE), Median Absolute Error (MAE), Mean Absolute Percentage Error (MAPE).

$$MAE = \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\text{test set}} \underbrace{|y_i}_{\text{predicted vaue}} - \underbrace{\hat{y}_i|}_{\text{actual value}}$$

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)^2}_{\substack{\text{The square of the difference}\\ \text{between actual and}\\ \text{predicted}}}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$$MAPE = \underbrace{\frac{100\%}{n}}_{\substack{\text{Multiplying by 100\%}\\ \text{converts to percentage}}}\Sigma\left|\frac{\overbrace{y - \hat{y}}^{\text{The residual}}}{\underbrace{y}_{\substack{\text{Each residual is scaled}\\ \text{against the actual value}}}}\right|$$

$$MedAE(y, \hat{y}) = median(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|)$$

Data Evaluation

24)Area under the curve (AUC)–

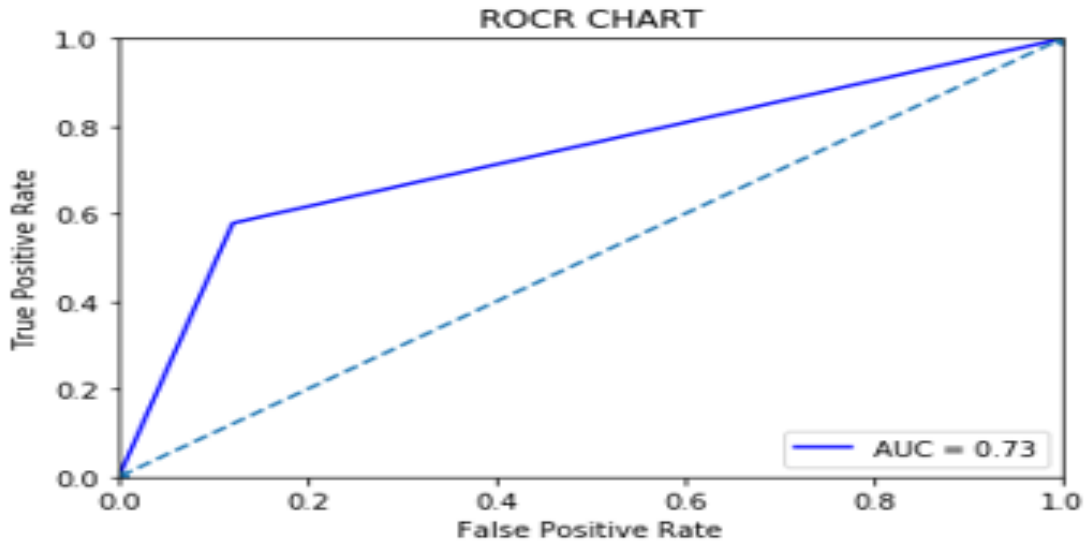
Typically, this is area under the curve of the ROC curve.

Higher the value of AUC better is the binary classification.

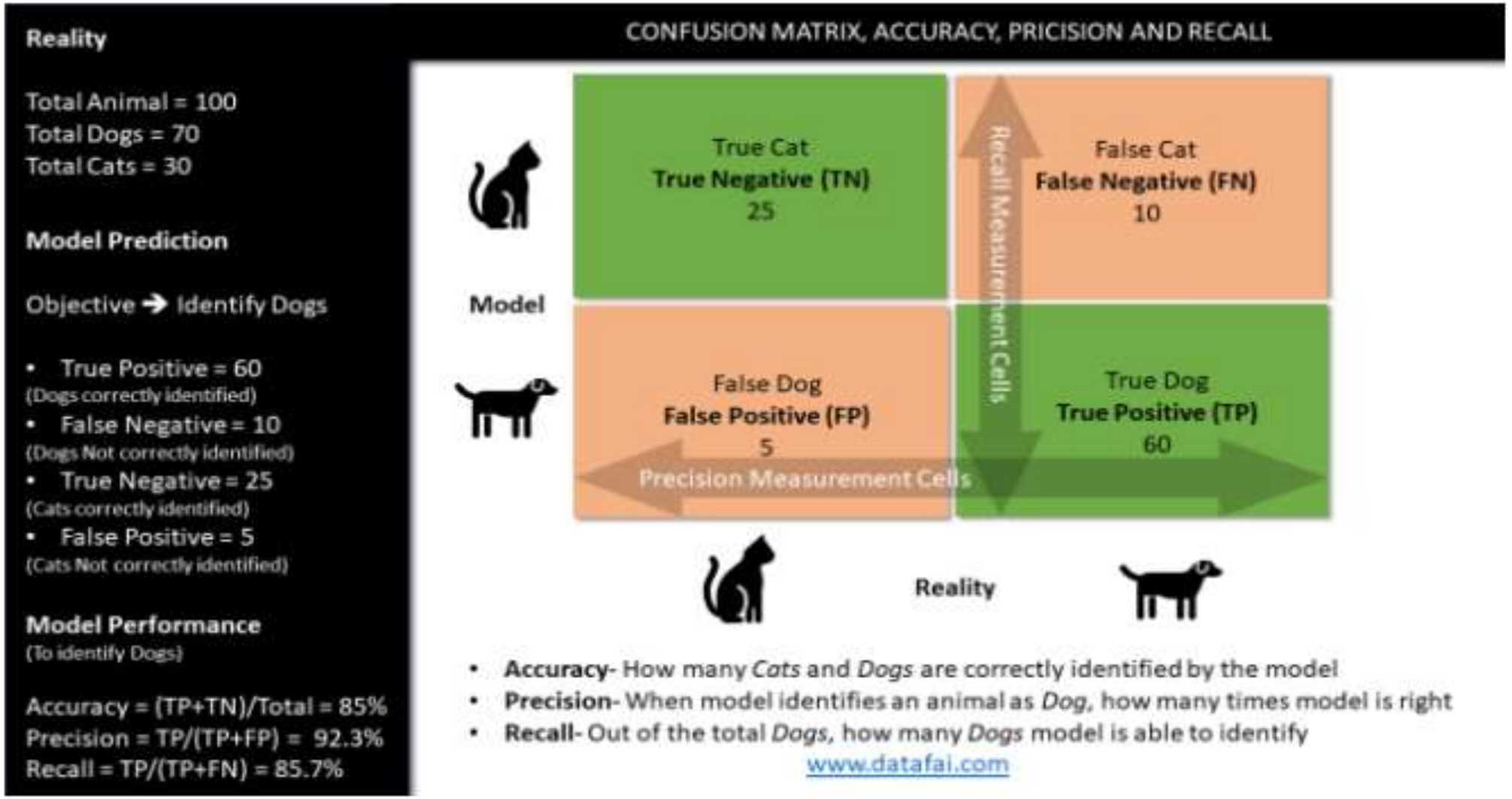
The Best possible value of AUC is 1 (or 100%) and the worst possible value of AUC is 0 (0%).

There are two main limitations of AUC- first, it is not applicable for multiclass classification and second, it is not a right metric for unbalanced data, i.e., for the data where one class is represented much higher than the other class.

For example, in fraud classification where fraud incidence rate in the typically less than 1%.



Data Evaluation



Deployment

SMA EMA T Test Metrics:

SERIAL NUMBERS	DESCRIPTIONS	TOTAL	TRUE COUNT	FALSE COUNT	EFFICIENCY
SMA7	Simple moving average-7 samples	5297	4114	1183	77.67
SMA13	Simple moving average-13 samples	5291	3474	1817	65.66
SMA20	Simple moving average-20 samples	5284	3217	2067	60.88
EMA7	Exponential moving average-7 samples	5297	4077	1220	76.97
EMA13	Exponential moving average-13 samples	5291	3486	1805	65.89
EMA20	Exponential moving average-20 samples	5284	3236	2048	61.24

It can be observed that T-test Hypothesis testing done for a rolling 7-day moving average data has given the highest efficiency in correctly predicting the upward or downward trend closely followed by exponential moving averages with a span of 7-days. however, prediction efficiency is least for 20 days of Simple moving average data and 20-days exponential moving average data.

Deployment

SMA EMA Z Test Metrics:

SERIAL NUMBERS	DESCRIPTIONS	TOTAL	TRUE COUNT	FALSE COUNT	EFFICIENCY
SMA100	Simple moving average-100 samples	5204	2798	2406	53.77
SMA200	Simple moving average-200 samples	5104	2754	2350	53.96
EMA100	Exponential moving average-100 samples	5204	2829	2375	54.36
EMA200	Exponential moving average-200 samples	5104	2779	2325	54.45

It can be observed that Z-test Hypothesis testing done for a rolling 100-day moving average And 200-day moving average has given lesser efficiency in correctly predicting the upward or downward trend compared to the prediction done with Hypothesis testing done on smaller samples using T-test Hypothesis testing.

Similar inferences can be drawn for Exponential moving average with 100 days and 200 days span as well.

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956, Approved by AICTE, New Delhi

Deployment

Classification Model Metrics:

SERIAL NUMBERS	DESCRIPTIONS	TOTAL	TRUE COUNT	FALSE COUNT	EFFICIENCY
Structured Data Classifier	Auto Keras Classification Model	1061	901	160	84.92
K Neighbors Classifier	KNN Classification Model	1061	786	267	74.08
Logistic Regression	Logistic Regression Classification Model	1061	956	97	90.10

It can be observed that Logistic Regression Classification Model and Auto Keras classification Model have given the accuracy of near about 85 to 90% in able to correctly predict the direction of the close price.

The highest Accuracy in predicting the direction by Hypothesis Testing using simple moving Average and Exponential Moving averages was near about 77%.

Hence, it can be safely concluded that Deep Learning models and Machine Learning Models were able to provide better outputs compared to Statistical methods of Hypothesis Testing.

Deployment

ARIMA Models Metrics:

SERIAL NUMBERS	DESCRIPTIONS	MEAN ABSOLUTE ERROR (MAE) FOR TEST DATA	MEAN SQUARE ERROR (MSE) FOR TEST DATA	ROOT MEAN SQUARE ERROR (RMSE) FOR TEST DATA	Median Absolute Error FOR TEST DATA	MAPE FOR TEST DATA
EMA_200ARIMA	Auto Arima model using Exponential moving average-200 samples	84.21	9662.99	98.30	96.06	Nan
SMA_100ARIMA	Auto Arima model using Simple moving average-100 samples	112.25	19404.28	139.30	95.51	9.42
SMA_20ARIMA	Auto Arima model using Simple moving average-20 samples	183.76	45227.79	212.67	181.82	16.29
SMA_13ARIMA	Auto Arima model using Simple moving average-13 samples	184.73	44482.52	210.91	172.64	16.171
SMA_7ARIMA	Auto Arima model using Simple moving average-7 samples	185.64	47486.11	217.91	173.93	15.09

Deployment

ARIMA Models Metrics:

In all results of the ADF test for ARIMA Modelling on our dataset for HDFC stock, it was observed that the p-value obtained is bigger than 0.05 thus we the null hypothesis is not rejected and it is concluded that the statistic for Dataset under consideration is non-stationary.

It can be observed that mean Absolute Error, Mean Square Error, Root Mean Square Error, Median Absolute Error, and Mean Absolute Percentage Error are far too high in the case of all Auto ARIMA Modelling. Hence, it can be concluded that the dataset under consideration was not suitable for Time series Modelling using the ARIMA Modelling algorithm.

Deployment

Regression Models-Part1 Metrics:It can be observed that mean Absolute Error and Mean Absolute Percentage Error was satisfactory for the Ordinary Least Squares (OLS)-Linear Regression Model. However, other Regression Models were not able to provide MAPE within the acceptable range.

SERIAL NUMBERS	DESCRIPTIONS	MEAN ABSOLUTE ERROR (MAE) FOR TEST DATA	MEAN SQUARE ERROR (MSE) FOR TEST DATA	ROOT MEAN SQUARE ERROR (RMSE) FOR TEST DATA	Median Absolute Error FOR TEST DATA	MAPE FOR TEST DATA
OLS Model	Ordinary Least Squares (OLS)-Linear Regression Model	2.03	11.83	3.44	1.14	0.227
LASSO Model	Lasso Regression Model	7.56	132.63	11.52	4.67	0.85
CVLASSO Model	Lasso regression Model Using Cross-Validation	7.55	132.59	11.51	4.66	0.85
KNN Model	The k-Nearest Neighbors (KNN) Algorithm	5.42	132.08	11.49	3.16	0.59

Deployment

Regression Models-Part2 Metrics:It can be observed that mean Absolute Error and Mean Absolute Percentage Error were satisfactory for Random Forest Regression Model. However, other Regression Models were able to provide fairly acceptable MAPE but still lower MAPE would have been better.

SERIAL NUMBERS	DESCRIPTIONS	MEAN ABSOLUTE ERROR (MAE) FOR TEST DATA	MEAN SQUARE ERROR (MSE) FOR TEST DATA	ROOT MEAN SQUARE ERROR (RMSE) FOR TEST DATA	Median Absolute Error FOR TEST DATA	MAPE FOR TEST DATA
DT Model	Decision Tree Algorithm	3.26	23.95	4.89	2.10	0.383
GRIDSEARCHCV Model	GridSearchCV Algorithm with Hyper-parameter Tuning	3.22	23.16	4.81	2.10	0.38
RF Model	Random Forest Regression Model	2.45	15.25	3.90	1.49	0.29
XGBOOST Model	XGBoost ML Model	3.25	22.78	4.77	2.12	0.37

REVA UNIVERSITY
Bengaluru, India

Deployment

Regression Models-Part3 Metrics:It can be observed that mean Absolute Error and Mean Absolute Percentage Error was satisfactory for both Using Principal Component Analysis (PCA) with LSTM and Regression Model using AutoKeras.However, other Regression Models were able to provide fairly acceptable MAPE but still, their MAE would have been better.

SERIAL NUMBERS	DESCRIPTIONS	MEAN ABSOLUTE ERROR (MAE) FOR TEST DATA	MEAN SQUARE ERROR (MSE) FOR TEST DATA	ROOT MEAN SQUARE ERROR (RMSE) FOR TEST DATA	Median Absolute Error FOR TEST DATA	MAPE FOR TEST DATA
PCA LSTM Model	Using Principal Component Analysis (PCA) with LSTM	4.37	34.70	5.89	3.60	33.44
PCA LSTM Moving Averages Model	Using Principal Component Analysis (PCA) with LSTM with Moving Average variables (Feature Engineering)	7.75	135.03	11.62	5.99	33.47
LSTM Model	Long Short-Term Memory-LSTM Neural Network Model	9.71	159.01	12.61	8.20	33.40
Auto Keras Model	Regression Model using AutoKeras	2.59	242.51	15.57	1.10	0.27

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Analysis and Results

Classification Metrics Comparison:

SERIAL NUMBERS	DESCRIPTIONS	EFFICIENCY>67%
SMA7	Simple moving average-7 samples	YES-77.67
SMA13	Simple moving average-13 samples	NO-65.66
SMA20	Simple moving average-20 samples	NO-60.88
EMA7	Exponential moving average-7 samples	YES-76.97
EMA13	Exponential moving average-13 samples	NO-65.89
EMA20	Exponential moving average-20 samples	NO-61.24
SMA100	Simple moving average-100 samples	NO-53.77
SMA200	Simple moving average-200 samples	NO-53.96
EMA100	Exponential moving average-100 samples	NO-54.36
EMA200	Exponential moving average-200 samples	NO-54.45
Structured Data Classifier	Auto Keras Classification Model	yes-84.92
K Neighbours Classifier	KNN Classification Model	yes-74.08
Logistic Regression	Logistic Regression Classification Model	yes-90.10

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Analysis and Results

Classification Metrics Comparison:

SERIAL NUMBERS	DESCRIPTIONS	EFFICIENCY>67%
SMA7	Simple moving average-7 samples	YES-77.67
SMA13	Simple moving average-13 samples	NO-65.66
SMA20	Simple moving average-20 samples	NO-60.88
EMA7	Exponential moving average-7 samples	YES-76.97
EMA13	Exponential moving average-13 samples	NO-65.89
EMA20	Exponential moving average-20 samples	NO-61.24
SMA100	Simple moving average-100 samples	NO-53.77
SMA200	Simple moving average-200 samples	NO-53.96
EMA100	Exponential moving average-100 samples	NO-54.36
EMA200	Exponential moving average-200 samples	NO-54.45
Structured Data Classifier	Auto Keras Classification Model	yes-84.92
K Neighbours Classifier	KNN Classification Model	yes-74.08
Logistic Regression	Logistic Regression Classification Model	yes-90.10

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956, Approved by AICTE, New Delhi

Analysis and Results

Regression Metrics Comparison:

SERIAL NUMBERS	DESCRIPTIONS	MAE<=5	MAPE<=0.33
OLS Model	Ordinary Least Squares (OLS)-Linear Regression Model	YES-2.034	YES-0.23
LASSO Model	Lasso Regression Model	NO-7.555	NO-0.85
CVLASSO Model	Lasso regression Model Using Cross-Validation	NO-7.55	NO-0.85
KNN Model	The k-Nearest Neighbours (KNN) Algorithm	NO-5.423	NO-0.59
DT Model	Decision Tree Algorithm	YES-3.26	NO-0.38
GRIDSEARCHCV Model	GridSearchCV Algorithm with Hyper-parameter Tuning	YES-3.218	NO-0.38
RF Model	Random Forest Regression Model	YES-2.45	YES-0.29
XG Boost Model	XGBoost ML Model	YES-3.25	NO-0.37
PCA LSTM Model	Using Principal Component Analysis (PCA) with LSTM	YES-4.366	YES-33.44
PCA LSTM Moving Averages Model	Using Principal Component Analysis (PCA) with LSTM with Moving Average variables (Feature Engineering)	NO-7.75	YES-33.47
LSTM Model	Long Short-Term Memory-LSTM Neural Network Model	NO-9.71	YES-33.40
Auto Keras Model	Regression Model using AutoKeras	YES-2.59	YES-0.27

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Analysis and Results

It can be observed that Logistic Regression Classification Model and Auto Keras classification Model have given the accuracy of near about 85 to 90% in able to correctly predict the direction of the close price.

The highest Accuracy in predicting the direction by Hypothesis Testing using simple moving Average and Exponential Moving averages was near about 77%. other Hypothesis testing using T-test and Z-test statistical algorithms were not satisfactory in able to predict the direction of the close price of the HDFC share.

It can be observed that Ordinary Least Squares (OLS)-Linear Regression Model, Random Forest Regression Model, Using Principal Component Analysis (PCA) with LSTM and Regression Model using AutoKeras provide MAE<=5 and MAPE<=0.33. Hence these Regression Models were most successful in predicting the close value of the stock price.

Recommendations for Future Work

It is assumed that returns are more or less constant over time. However, Returns are highly dependent on time.

In future projects, it can be shown how to define Bullish and Bearish regimes using modern machine learning techniques. Then the techniques already discussed can be used in this project to estimate the direction of close price for each of the Normal and Crash regimes.

The Sentiment Analysis Approach may also be explored using Text Analytics for predicting stock market returns.

REVA UNIVERSITY

Bengaluru, India

Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Thank You!