A Thesis/Project/Dissertation Report

on

STOCK MARKET PREDICTION SYSTEM

Submitted in partial fulfillment of the

requirement for the award of the degree of

Bachelor of Technology

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of
Name of
Supervisor: Dr.
Nitin Mishra

Submitted By

Ujjwal Pant - 18SCSE1010388

Ashish Anand- 18SCSE1010358

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING /
DEPARTMENT OF COMPUTERAPPLICATION
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA

DECEMBER, 2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **"STOCK MARKET PREDICTION SYSTEM"** in partial fulfillment of the requirements for the award of Bachelor of Technology submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of July-2021 to December-2021 under the supervision of Dr. Nitin Mishra, Assistant Professor,, Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

Ujjwal Pant- 18SCSE1010388

Ashish Anand- 18SCSE1010358

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dr. Nitin Mishra

Assistant Professor

Table of Contents

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of Ujjwal Pant- 18SCSE1010388

Ashish Anand- 18SCSE1010358

has been held on _________________ and his/her work is recommended for the award of Bachelor of Technology in Computer Science and Engineering.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date:

Place: Greater Noida

Stock Market Prediction System

ABSTRACT

The prediction of a stock market direction may serve as an early recommendation system for short-term investors and as an early financial distress warning system for long-term shareholders. Forecasting accuracy is the most important factor in selecting any forecasting methods. Research efforts in improving the accuracy of forecasting models are increasing since the last decade. The appropriate stock selections those are suitable for investment is a very difficult task. The key factor for each investor is to earn maximum profits on their investments. In this project Regression a machine learning technique and long short term memory technique is used. Abstract-In Stock Market Prediction, the aim is to predict the future value of the financial stocks of a company. The recent trend in stock market prediction technologies is the use of machine learning which makes predictions based on the values of current stock market indices by training on their previous values. Machine learning itself employs different models to make prediction easier and authentic. The paper focuses on the use of Regression and LSTM based Machine learning to predict stock values.

Factors considered are open, close, low, high and volume. You would like to model stock prices correctly, so as a stock buyer you can reasonably decide when to buy stocks and when to sell them to make a profit. You need good machine learning models that can look at the history of a sequence of data and correctly predict what the future elements of the sequence are going to be. A correct prediction of stocks can lead to huge profits for the seller and the broker. Frequently, it is brought out that prediction is chaotic rather than random, which means it can be predicted by carefully analyzing the history of respective stock market. Machine learning is an efficient way to represent such

List of Figures

Acronyms

B.Tech.	Bachelor of Technology
M.Tech.	Master of Technology
BCA	Bachelor of Computer Applications
MCA	Master of Computer Applications
B.Sc. (CS)	Bachelor of Science in Computer Science
M.Sc. (CS)	Master of Science in Computer Science
SCSE	School of Computing Science and Engineering

Table of Contents

CHAPTER-1

Introduction

The prediction of a stock market direction may serve as an early recommendation system for short-term investors and as an early financial distress warning system for long-term shareholders. Forecasting accuracy is the most important factor in selecting any forecasting methods. Research efforts in improving the accuracy of forecasting models are increasing since the last decade. The appropriate stock selections those are suitable for investment is a very difficult task. The key factor for each investor is to earn maximum profits on their investments. In this project Regression a machine learning technique and long short term memory technique is used. Abstract-In Stock Market Prediction, the aim is to predict the future value of the financial stocks of a company. The recent trend in stock market prediction technologies is the use of machine learning which makes predictions based on the values of current stock market indices by training on their previous values. Machine learning itself employs different models to make prediction easier and authentic. The paper focuses on the use of Regression and LSTM based Machine learning to predict stock values.

Factors considered are open, close, low, high and volume. You would like to model stock prices correctly, so as a stock buyer you can reasonably decide when to buy stocks and when to sell them to make a profit. You need good machine learning models that can look at the history of a sequence of data and correctly predict what the future elements of the sequence are going to be. A correct prediction of stocks can lead to huge profits for the seller and the broker. Frequently, it is brought out that prediction is chaotic rather than random, which means it can be predicted by carefully analyzing the history of respective stock market. Machine learning is an efficient way to represent such processes. It predicts a market value close to the tangible value, thereby increasing the accuracy. Introduction of machine learning to the area of stock prediction has appealed to many researches because of its efficient and accurate measurements

LITERATURE REVIEW

Over the past two decades many important changes have taken place in the environment of financial markets. The development of powerful communication and trading facilities has enlarged the scope of selection for investors.

Forecasting stock return is an important financial subject that has attracted researchers' attention for many years. It involves an assumption thatfundamental information publicly available in the past has some predictive

relationships to the future stock returns. In order to be able to extract such relationships from the available data, data mining techniques are new techniques that can be used to extract the knowledge from this data. For that reason, several researchers have focused on technical analysis and using advanced math and science. Extensive attention has been dedicated to the fieldof artificial intelligence and data mining techniques. Some models have been proposed and implemented using the above mentioned techniques, the authorsof Tsang, P.M., Kwok, P., Choy, S.O., Kwan, R., Ng, S.C., Mak, J., Tsang, J., Koong,

K., and Wong, T. made an

empirical study on building a stock buying/selling alert system using back propagation neural networks (BPNN), their NN was codenamed NN5. The system was trained and tested with past price data from Hong Kong and Shanghai Banking Corporation Holdings over the period from January 2004 to December 2005. The empirical results showed that the implemented system was able to predict short-term price movement directions with accuracy about 74%.

The research by Wu, M.C., Lin, S.Y., and Lin, C.H., used decision tree technique to build on the work of Lin. where Lin tried to modify the filter rule that is to buywhen the stock price rises k% above its past local low and sell when it falls k% from its past local high. The proposed modification to the filter rule was by combining three decision variables associated with fundamental analysis. An empirical test, using the stocks of electronics companies in Taiwan, showed Lin'smethod outperformed the filter rule. According to Wu, M.C., Lin, S.Y., and Lin, C.H.,, in Lin's work, the criteria for clustering trading points involved only the past information; the future information was not considered at all. The researchby Wu, M.C., Lin, S.Y., and Lin, C.H., aimed to improve the filter rule and Lin's study by considering both the past and the future information inclustering the trading points. The researchers used the data of Taiwan stock market and that of NASDAQ to carry out empirical tests. Test results showed that the proposed method outperformed both Lin's method and the filter rule in the two stock markets.

The model of Wang, J.L., Chan, S.H. (2006) "Stock market trading rule discovery using two-layer bias decision tree", applied the concept of serial topology and designed a new decision system, namely the two layer

bias decision tree, for stock price prediction. The methodology developed by theauthors differs from other studies in two respects;

first, to reduce the classification error, the decision model was modified into abias decision model.

Second, a two-layer bias decision tree is used to improve purchasing accuracy.The empirical results indicated that the presented decision model produced excellent purchasing accuracy, and it significantly outperformed than random purchase.

The authors Enke, D., Thawornwong, S. presented an approach that used data mining methods and neural networks for forecasting stock market returns. An attempt has been made in this study to investigate the predictive power of financial and economic variables by adopting the variable relevance analysis technique in machine learning for data mining. The authors examined the effectiveness of the neural network models used for level estimation and classification. The results showed that the trading strategies guided by the neural network classification models generate higher profits under the same risk exposure than those suggested by other strategies.

PROBLEM FORMULATION

Investors are familiar with the saying, “buy low, sell high” but this does not provide enough context to make proper investment decisions. Before an investor invests in any stock, he needs to be aware how the stock market behaves. Investing in a good stock but at a bad time can have disastrous results, while investment in a mediocre stock at the right time can bear profits. Financial investors of today are facing this problem of trading as they do not properly understand as to which stocks to buy or which stocks to sell in order to get optimum profits. Predicting long term value of the stock is relatively easy than predicting on day-to-day basis as the stocks fluctuate rapidly every hour based on world events.

We aim to predict the daily adjusted closing prices of Vanguard Total Stock Market ETF (VTI), using data from the previous N days (ie. forecast horizon=1). We will use three years of historical prices for VTI from 2015–11–25 to 2018–11–23, which can be easily downloaded from yahoo finance.

We will split this dataset into 60% train, 20% validation, and 20% test. The modelwill be trained using the train set, model hyper parameters will be tuned using the validation set, and finally the performance of the model will be reported using the test set. Below plot shows the adjusted closing price split up into the respective train, validation and test sets.

To evaluate the effectiveness of our methods, we will use the root mean squareerror (RMSE) and mean absolute percentage error (MAPE) metrics. For both metrics, the lower the value, the better the prediction.

Last Value

In the Last Value method, we will simply set the prediction as the last observedvalue. In our context, this means we set the current adjusted closing price as theprevious day’s adjusted closing price. This is the most cost-effective forecasting model and is commonly used as a benchmark against which more sophisticatedmodels can be compared. There are no hyperparameters to be tuned here.

Moving Average

In the moving average method, the predicted value will be the mean of the previous N values. In our context, this means we set the current adjusted closing price as the mean of the adjusted closing price of the previous N days.The hyperparameter N needs to be tuned.

REQUIRED TOOLS

- JUPYTER NOTEBOOK
- DATASET FROM YAHOO FINANCE

PYTHON LIBRARIES

- Pandas
- Numpy
- Scikit Learn
- Matplotlib
- Pandas_datareaders
- Keras
- Math

These following Libraries can be installed by pip command in terminal andcan be used us the import function.

COMPLETE WORK PLAN LAYOUT

DATASET INFORMATION

Stock prices come in several different flavours. They are,

- Open: Opening stock price of the day
- Close: Closing stock price of the day
- High: Highest stock price of the data
- Low: Lowest stock price of the day

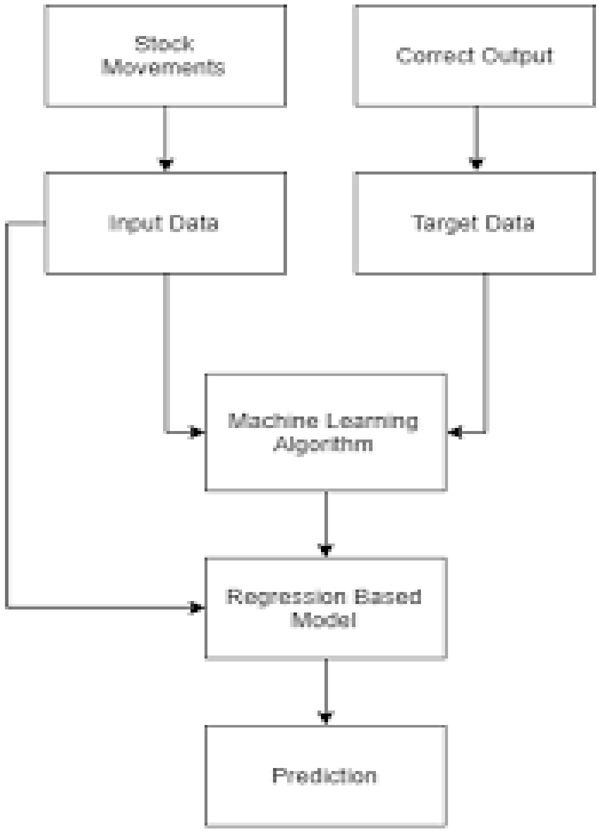
MODEL 1

Stock market prediction seems a complex problem because there are many factors that have yet to be addressed and it doesn't seem statistical at first. Butby proper use of machine learning techniques, one can relate previous data to the current data and train the machine to learn from it and make appropriate assumptions. Machine learning as such has many models but this paper focuses on two most important of them and made the predictions using them.

$$V = a + bk + error$$

Regression is used for predicting continuous values through some given independent values . The project is based upon the use of linear regression algorithm for predicting correct values by minimizing the error function as given in Figure1. This operation is called gradient descent. Regression uses a given linear function for predicting continuous values: Where, Vis a continuousvalue; K represents known independent values; and, a, b are coefficients. Work was carried out on csv format of data through panda library and calculated theparameter which is to be predicted, the price of the stocks with respect to time. The data is divided into different train sets for cross validation to avoid over fitting. The test set is generally kept 20% of the whole dataset. Linear regression as given by the above equation is performed on the data and thenpredictions are made, which are plotted to show the results of the stock market prices vs time

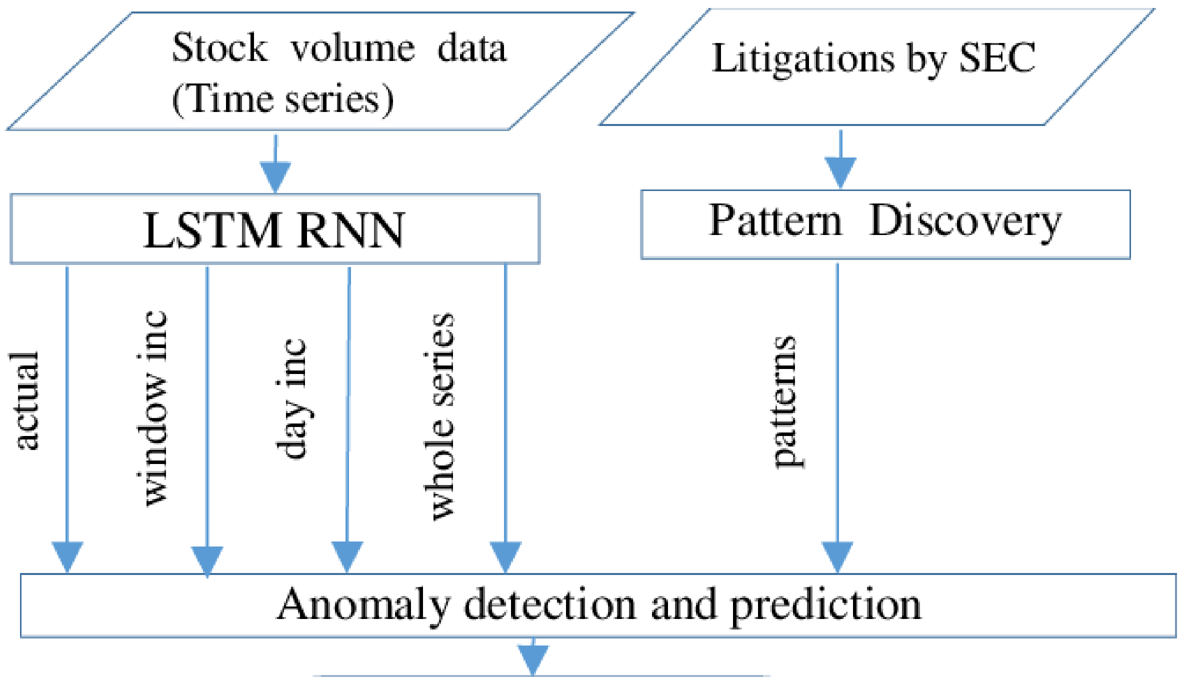
FLOW CHART


Stock Movements
Correct Output
Input Data
Target Data
Machine Learning Algorithm
Regression Based Model
Prediction

MODEL 2 based on LSMT

LSTM is the advanced version of Recurrent-NeuralNetworks (RNN) where the information belonging to previous state persists. These are different from RNNsas they involve long term dependencies and RNNs works on finding the relationship between the recent and the current information. This indicates thatthe interval of information is relatively smaller than that to LSTM. The main purpose behind using this model in stock market prediction is that the predictions depends on large amounts of data and are generally dependent on the long term history of the market . So LSTM regulates error by giving an aid tothe RNNs through retaining information for older stages making the prediction more accurate . Since stock market involves processing of huge data, the gradients with respect to the weight matrix may become very small and may degrade the learning rate of the system. This corresponds to the problem of Vanishing Gradient. LSTM prevents this from happening. The LSTM consists of aremembering cell, input gate, output gate and a forget gate. The cell remembersthe value for long term propagation and the gates regulate them . In this paper,a sequential model has been made which involves stacking two LSTM layers on top of each other with the output value of 256. The input to the layer is in the form of two layer and layer. A dropout value of 0.3 has been fixed which means that 0.3 out of total nodes will be frozen during the training process to avoid over-fitting of data and increase the speed of the training process. At last, the core dense layer where each neuron is connected to every other in the next layer is added providing input of 32 parameters to the next core layer which gives output as 1. The model is compiled with a mean square cost function to maintain the error throughout the process and accuracy is chosen as a metric for the prediction.

FLOW CHART



Stock volume data
(Time series)
Litigations by SEC
LSTM RNN
Pattern Discovery
actual
window inc
day inc
whole series
patterns
Anomaly detection and prediction

STEPS IN PREAPARATION OF MODELS

- Download data from yahoo finance
- Data exploration using pandas
- Splitingdataset into training and test sets
- Normalizing the data using MINMAXSCALER
- Predicting by Averaging
- Defining hyperparameters
- Defining inputs and oupts
- Parameters for LSTM and regression
- Calculating LSTM OUTPUT and feeding it to regression layer to get finalprediction
- Predict related calculations
- Visualising the prediction

IMPLEMENTATION

```
In [44]:
import math
import pandas_datareader as web
import numpy as np
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
import matplotlib.pyplot as plt
```

LOADING THE DATASET

```
In [14]:
df=web.DataReader('AAPL',data_source='yahoo',start='2012-01-01',end='2019-12-17')
df
```

	High	Low	Open	Close	Volume	Adj Close
Date						
2012-01-03	14.732142	14.607142	14.621428	14.686786	302220800.0	12.691425
2012-01-04	14.810000	14.617143	14.642858	14.765715	260022000.0	12.759631
2012-01-05	14.948215	14.738214	14.819643	14.929643	271269600.0	12.901293
2012-01-06	15.098214	14.972143	14.991786	15.085714	318292800.0	13.036158
2012-01-09	15.276786	15.048214	15.196428	15.061786	394024400.0	13.015480

```
In [15]:
df.shape

(2003, 6)
```

```
In [10]:
#VISUALIZE THE CLOSING PRICE HISTORY
```

```
In [19]:
plt.figure(figsize=(16,8))
plt.title(['Close Price history'])
plt.plot(df['Close'])
plt.xlabel('Date',fontsize=18)
plt.ylabel('Close Price USD($)')
plt.style.use('fivethirtyeight')
plt.show()
```


```
In [20]: #create a new dataframe with only the close column

In [21]: data=df.filter(['Close'])
         dataset=data.values

In [25]: training_data_len=math.ceil(len(dataset)* .8)
         training_data_len

         1603
```

```
In [26]: #SCALE THE DATA

In [27]: scaler=MinMaxScaler(feature_range=(0,1))
         scale_data=scaler.fit_transform(dataset)
         scale_data

         array([[0.01316509],
                [0.01457064],
                [0.01748985],
                ...,
                [0.97658263],
                [0.99755134],
                [1.        ]])
```

In [ ]:

```
#Create the training dataset
#Create the Scaled training data set

train_data=scale_data[0:training_data_len,:]

#spliting the data into X_train and Y_train sets
x_train=[]
y_train=[]

for i in range(60,len(train_data)):
    x_train.append(train_data[i-60:i,0])
    y_train.append(train_data[i,0])
    if i<=60:
        print(x_train)
        print(y_train)
```

In [33]:

```
#converting into the numpy array
x_train,y_train=np.array(x_train),np.array(y_train)
```

reshape data int 3 D dimnesion as lstm accepts 3 Dinmension values

In [38]:

```
x_train.shape
```

```
(1543, 60)
```

In [41]:

```
x_train=np.reshape(x_train,(x_train.shape[0],x_train.shape[1],1))
x_train.shape
```

```
(1543, 60, 1)
```

```
In [55]:
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout
from tensorflow.keras.layers import LSTM
```

Build the LSTM model

```
In [57]:
model=Sequential()
model.add(LSTM(50, return_sequences=True, input_shape= (x_train.shape[1],1)))
model.add(LSTM(50, return_sequences=False))
model.add(Dense(25))
model.add(Dense(1))
```

```
In [58]:
#Complie the model
model.compile(optimizer='adam',loss='mean_squared_error')
```

```
In [59]:
model.fit(x_train,y_train,batch_size=1,epochs=1)

1543/1543 [==============================] - 18s 12ms/step - loss: 7.3976e-04
```

```
In [60]:
#creating the testing the dataset
#Create a new array containing scaled values from inex 1543
```

```
In [96]:
test_data=scale_data[training_data_len - 60: , :]
```

```
In [97]:
#Create the data sets x_test and y_test
x_test=[]
y_test=dataset[training_data_len:,:]

for i in range(60,len(test_data)):
    x_test.append(test_data[i-60:i, 0])
```

```
In [98]:
x_test=np.array(x_test)
```

```
In [99]:
x_test= np.reshape(x_test, (x_test.shape[0], x_test.shape[1], 1))
```

get the model prediction

```
In [100]:
predictions=model.predict(x_test)
predictions=scaler.inverse_transform(predictions)
```

get the root mean squared error(RMSE)

```
In [101]:
rmse=np.sqrt(np.mean(predictions-y_test)**2)
rmse

2.0913076114654543
```

plot the data

In [102]:

```
train=data[:training_data_len]
valid=data[training_data_len:]
valid['Predictions']=predictions
```

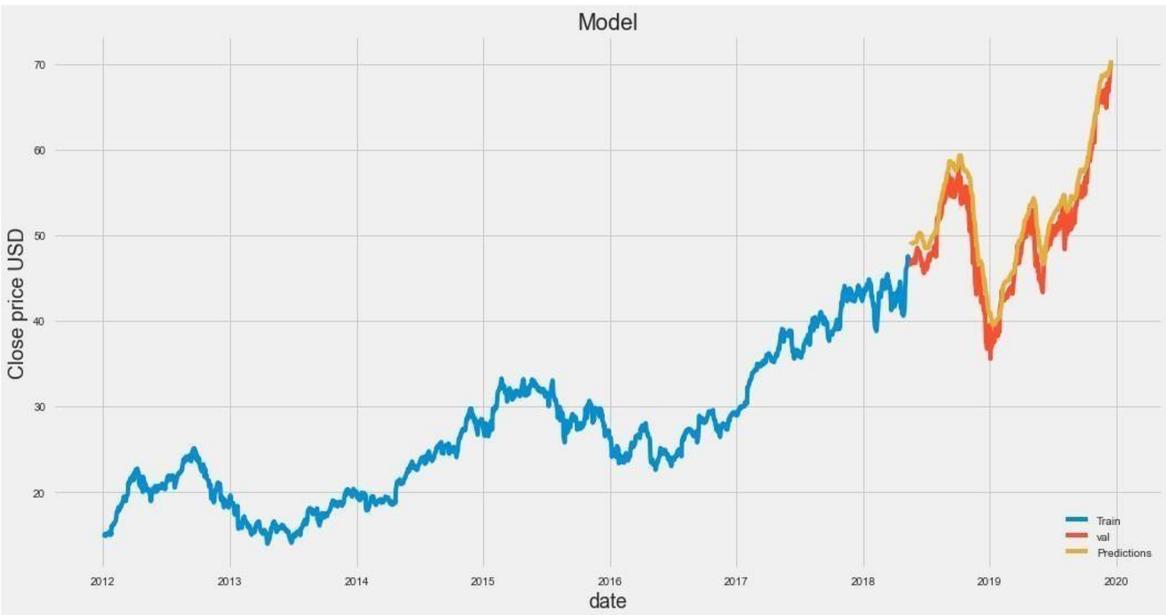
```
c:\python37\lib\site-packages\ipykernel_launcher.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  This is separate from the ipykernel package so we can avoid doing imports until
```

visualize the data

In [104]:

```
plt.figure(figsize=(16,8))
plt.title('Model')
plt.xlabel('date',fontsize=18)
plt.ylabel('Close price USD',fontsize=18)
plt.plot(train['Close'])
plt.plot(valid[['Close','Predictions']])
plt.legend(['Train','val','Predictions'],loc='lower right')
```

PLOTING THE GRAPH FOR PREDICTED PRICE WITH ACTUAL VALUE



In [105]:

```
valid
```

Date	Close	Predictions
2018-05-17	46.747501	48.874680
2018-05-18	46.577499	48.959053
2018-05-21	46.907501	48.975101
2018-05-22	46.790001	48.999039
2018-05-23	47.090000	49.007614
...	...	...
2019-12-11	67.692497	69.118729
2019-12-12	67.864998	69.347015
2019-12-13	68.787498	69.590561
2019-12-16	69.964996	69.928429
2019-12-17	70.102501	70.411484

400 rows × 2 columns

PREDICTING THE CLOSSING PRICE USING LAST 60 DAYS DATA

```
In [106]:
#get the qoute
apple_quote=web.DataReader('AAPL',data_source='yahoo',start='2012-01-01',end='2019-12-17')
#create a new dataframe
new_df=apple_quote.filter(['Close'])
#get the last 60 day closing price values and convert the dataframe to an array
last_60_days=new_df[-60:].values
last_60_days_scaled=scaler.transform(last_60_days)
```

```
In [107]:
#create an empty list
X_test=[]
#append the last 60 lasts
X_test.append(last_60_days_scaled)
#convert the X_test set to numpy array
X_test=np.array(X_test)
#reshape the data
X_test=np.reshape(X_test,(X_test.shape[0],X_test.shape[1],1))
#get the predicted price
pred_price=model.predict(X_test)
#undo scaling
pred_price=scaler.inverse_transform(pred_price)
print(pred_price)

[[70.909615]]
```

Limitations and Future Scope of the Project

1) Machine-learning methods HAVE been successfully used by various individuals and institutional 'in-house' groups, but most 'public' individuals, such as yourself, will NOT learn of 'THE' SPECIFIC methodologies that have yielded 'lucrative' returns and results.When 'huge' money is involved, and this IS the case when 'dealing with' the financial markets, NO ONE is going to publicly 'share' their 'edge' derived from applying THEIR successful methods to trading hence, you're not likely to hear of, nor see, detailed
studies and reports of such successes.

2) MOST 'academic' researchers who publish papers attempting to apply computer-processing algorithms to trading markets simply do NOT truly UNDERSTAND the underlying 'dynamics' of market price behaviors, so 'naive' applications of methodologies are attempted and 'researched', with the result that 'less than stellar' outcomes are generated frequently. To be 'effective' in developing 'successful' trading methods requires a rather 'deep' understanding of 'general underlying dynamic behaviors' of what makes the markets 'tick'. In particular,....

3) Markets (stocks, futures, forex, options, etc) generate data that form (statistically) NON-STATIONARY, time-series of numbers over ANY period of 'time window' that one may want to examine, 'forecast' upon, and trade. 'Prediction' (which is highly 'precise') isessentially impossible, but to a greater or lesser degree, 'forecastability' (less 'precise', but more 'probabilistic') IS applicable to market time-series data, with the exception of what are called 'event shocks', such as USA's 9/11, October of 1987, 'flash crashes', and similar types of 'events'. (From a 'risk-management' standpoint, any 'good' and 'effective' trading strategy/system MUST make provision for such occurrences in order to protect trading capital and prevent financial 'disaster'!)

4) From an engineering (and computer science) perspective, a 'trading system' can be 'thought of' as a 'combined' mathematical/logical TRANSFORM that uses 'appropriately-conditioned' time-series 'market' data as input and then attempts to 'functionally' convert this input into a monotonically-increasing 'capital-capture' output time-series. Before attempting to EFFECTIVELY design such a 'transform', one MUST have a relatively 'decent' understanding of the characteristics AND 'character' of the time-series 'input' data to which the 'transform' is to be applied........MOST researchers don't have an adequate, NOR realistic, market-dynamics UNDERSTANDING hence, their market MODELS are 'inadequate' and THIS is another reason why you rarely see public information of 'successful' machine-learning methods as applied to trading the markets.

CONCLUSION

The scope of Machine Learning is not limited to the investment sector. Rather, it is expanding across all fields such as banking and finance, information technology, media & entertainment, gaming, and the automotive industry. As the Machine Learning scope is very high, there are some of the areas where researchers are working toward revolutionizing the world for the future. Lastly here, but not finally, patterns do frequently recur in market-oriented time- series data that can be exploited when designing a transform such as mentioned in the previous paragraph. These patterns and features can certainly,and effectively be discerned by means of machine-learning methods.