PRESENTATION ON CAPSTONE PROJECT

Netflix Stock Prediction Using Machine Learning Algorithms

By

MANIKANTA JAGGAMPUDI

CONTENT

ABSTRACT

The objective of this research is to design a stock prediction linear model to predict the closing price of Netflix. This will be a comparative study of various machine learning models such as linear regression, k-nearest neighbour, and support vector machines. The attributes include in the model are Date, Open, Close, High, Low, Volume and Adj Close.

Data Collected from 2002 to 2022, I have divided dataset into two parts, training set is defined as 70% and testing set is 30 %. Final prediction is to be done in testing part only. Machine learning algorithms are used. With the help of RMSE (Root Mean Square Error) we will see how concentrated the data points are around the best fit line.

KEYWORDS: Stock Prediction, NFLX, Machine Learning

INTRODUCTION

Netflix is a combination of two words Net (Internet) and Flix (Flick used as an abbreviation for movie/film). In 1997, Netflix, in full Netflix, Inc., media streaming and video-rental company founded by American entrepreneurs Reed Hastings and Marc Randolph in Los Gatos California.

In 1999, Netflix started its business by offering an online subscription service through the Internet. Netflix started a DVD-by-mail rental service which provided an online catalogue of movies.

Subscribers chose movies and television shows from the Company's website, and the shows were then mailed to them in the form of DVDs, along with prepaid return envelopes, from one of the company's more than 100 distribution locations.

RESEARCH METHODOLOGY

This section explains the method to predict the closing price of Netflix for twenty years.

- Data Collection

The dataset used for this work was collected from Yahoo finance site for twenty years (from 11 Feb., 2002 to 11 Nov. 2022) Shown in Fig1.The data consist of 1009 instances and 7 features: date, the highest price of the day, the lowest price of the day, open price, close price, volume, and adjacent close price.

	Date	Open	High	Low	Close	Adj Close	Volume
0	2002-11-11	0.607143	0.607143	0.585714	0.600714	0.600714	371000
1	2002-11-12	0.600000	0.614286	0.572143	0.577857	0.577857	2191000
2	2002-11-13	0.579286	0.642857	0.555714	0.641429	0.641429	3231200
3	2002-11-14	0.642857	0.645000	0.618571	0.618571	0.618571	931000
4	2002-11-15	0.642143	0.653571	0.589286	0.650000	0.650000	2045400
...	...	...	...	...	...	...	...
5031	2022-11-04	272.019989	274.970001	255.320007	260.790009	260.790009	11119200
5032	2022-11-07	261.059998	261.149994	252.089996	258.600006	258.600006	7927000
5033	2022-11-08	259.920013	266.079987	256.880005	263.459991	263.459991	7683800
5034	2022-11-09	259.660004	260.899994	254.220001	254.660004	254.660004	7266000
5035	2022-11-10	265.989990	275.359985	260.089996	274.970001	274.970001	9667300

5036 rows × 7 columns

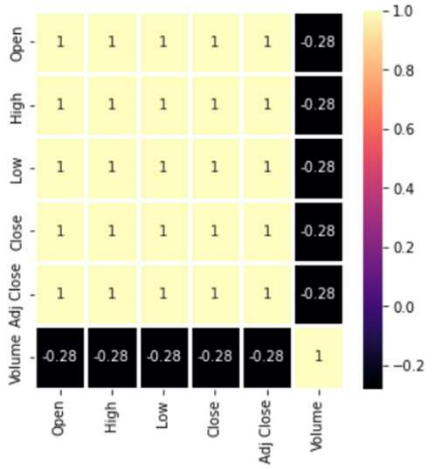
- Data Cleaning

After data collection, data cleaning is done which deals with missing data, duplicate data, and filtering out poor data. There is no nan values or duplicate values. Now we will check for the correlation coefficient by using heatmap shown in Fig2. For better understanding.

Fig 2 Correlation (Heatmap)

```
import seaborn as sns
plt.figure(figsize = [5,5])
sns.heatmap(df.corr(), cmap = "magma", linewidths = 3, annot =True)
```

```
<AxesSubplot:>
```



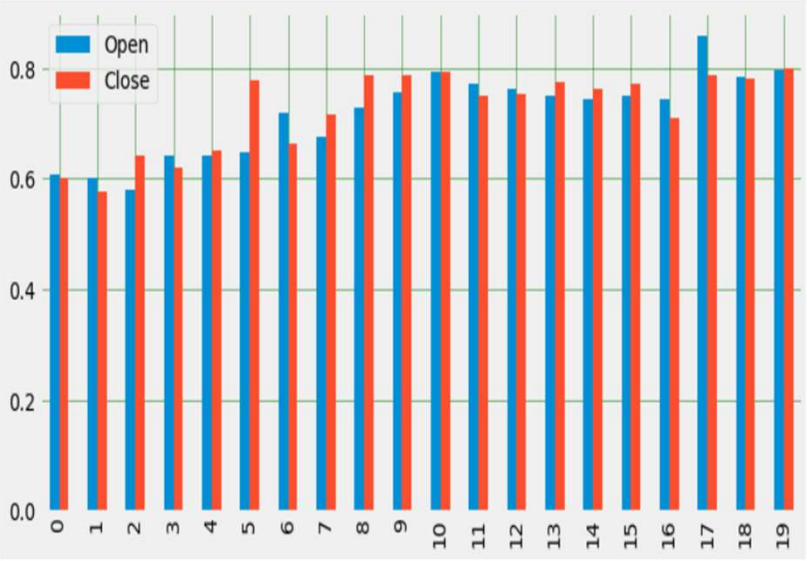
After visualizing the correlation coefficient, the light-yellow colour denotes the highly correlated features.

- Data Visualization

Visualize the Dependent Variable with Independent Features, by preparing the dataset which has high correlation will choose for the analysis.

fig 3 Date vs Closing Price

fig 4 Open vs Closing Price



Same like this we do for the High vs Close Price and Low vs Closing Price. After doing visualization of data, we will do the model training and testing.

- Data Transformation

Data consolidation is another term for it. The chosen data is translated into forms that can be used for data mining at this stage. The datasets were scaled to fit the model's tolerances and saved in the Commas Separated Value (.CVS) file format.

In this Dataset, record of each year is kept separately and is analysed using python libraries.

- Data Mining Stage

The data mining stage was divided into three phases. At each phase all the algorithms were used to analyse the stock datasets. The testing method adopted for this research was percentage split that train on a percentage of the dataset, cross validate on it and test on the remaining percentage. Thereafter interesting patterns representing knowledge were identified. . Fig 5 shows splitting data into training and testing.

Fig 5 Training and Testing data

```
from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test = train_test_split(X,Y,test_size=0.3,random_state = 1)
print(X_train.shape)
print(X_test.shape)
print(X_train.shape)
print(Y_test.shape)
```

```
(3525, 6)
(1511, 6)
(3525, 6)
(1511,)
```

Now you can see the size of dataset has been decreased as I have divided data into 80% training data and 20% Testing Data.

- Applying Machine Learning Algorithms

Now we apply the machine learning models such as linear regression, K-nearest neighbour, and support vector machines for the comparison shown in fig 6. After that we will apply cross validation for accuracy of the model shown in fig 7.

Fig 6 Linear Regression Model Training and Testing

```
from sklearn.linear_model import LinearRegression
lm = LinearRegression()
lm.fit(X_train,Y_train)
Y_pred = lm.predict(X_test)
```

Fig 7 Accuracy of Machine Models

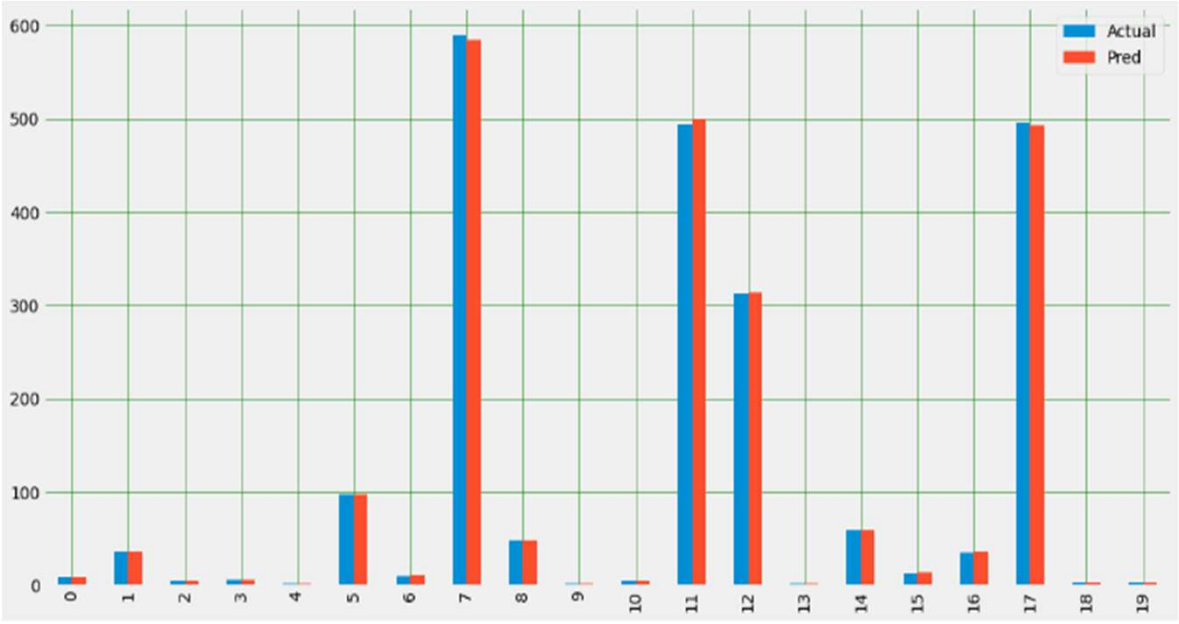
```
print("Accuracy of Linear Regression: ", val_scr.mean()*100)
print("Accuracy of KNN regression: ", val_knn_scr.mean()*100)
print("Accuracy of SVM Regression: ", val_svm_scr.mean()*100)
```

```
Accuracy of Linear Regression:  99.98612671426027
Accuracy of KNN regression:  99.63397890902742
Accuracy of SVM Regression:  97.11830488002003
```

RESULTS AND DISCUSSIONS

This section shows the results obtained from predicting the closing price of NFLX stocks of twenty years. The close price of Actual value and Predicted value of top twenty shown in fig 8.

Fig 8 Linear Actual vs Predicted



Now we will do same thing for the KNN and SVM after that we apply RMSE (Root Mean Square Error) for how concentrated the data points around the best fit line. R2 or r-squared will apply for knowing the score varies between 0 to 100% shown in fig 9.

Fig 9 R-squared error

```
print('Linear R2: ', r2_score(Y_test, Y_pred))
print('KNN R2: ', r2_score(Y_test, Y_knnpred))
print('SVM R2: ', r2_score(Y_test, Y_svmpred))
```

```
Linear R2:  0.9998691157823688
KNN R2:  0.9982238944187442
SVM R2:  0.9996039415030377
```

INFERENCES

In this the ML algorithms are used to predict the close price of NFLX by using technical indicators. The overall accuracy we obtained is good enough i.e., 99% since the stock trend could be affected by many random factors other than news and price information.

REFERENCES

1. Khan W, Malik U, Ghazanfar MA, Azam MA, Alyoubi KH, Alfakeeh AS (2019) Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis.

2. Khan, W., Ghazanfar, M.A., Azam, M.A. et al. Stock market prediction using machine learning classifiers and social media, news. J Ambient Intell Human Comput(2020).

3. Neha, & Rohan Lal. (2019). Prediction Analysis of Stock Price using Machine Learning Approach. Retrieved October 02, 2020.

4. Budak, C., Goel, S., & Rao, J. (2016). Fair and Balanced? Quantifying Media Bias Through Crowdsourced Content Analysis. SSRN Electronic Journal. doi:10.2139/ssrn.2526461

5. Kanade, P. A. (2020). Machine Learning Model for Stock Market Prediction. International Journal for Research in Applied Science and Engineering Technology.

Thank You