Restatement and Summary

The given dataset consists of stock-quote historical data for different companies that are present in the Nifty 50 index. Some of the features are; the symbol code, the date of trading, the opening price, the closing price, the highest price during the trading period, the lowest price during the trading period, cumulative volume traded, and the total turnover. Also, it has standard and computed columns such as VWAP – for Volume Weighted Average Price, and all the date fields are split into year and month. Stock performance data available in the dataset covers several years, which helps to examine trends that exist in the stock market for the particular company under consideration over the years.

The focus is on if we have seasonality and trends, will it be preferable to identify these trends and adjust or model our forecast with those specifics in mind. Seasonality and forecast future closing price with each month and its affiliated and If companies are positively related could it be practically suitable that? forecast future position of stocks by the difference of the stock prices in the above mentioned. Finally, In the conclusion, can we establish a good model adding all the analysis that I have done now? which is able to forecast the future stock price with a significant level of accuracy (over 70%).

Analysis and Visualisation

It entailed tasks such as checking the data for missing value indicators in columns, removing the columns with missing values, converting the date Type to datetime type, extracting month and year from the date to determine seasonality, checking for outliers, and excluding unimportant columns. In preprocessing numberic Chinese columns, StandardScaler was used to transform numeric columns to the range of values from 0 to 1 so that all the processed feature scales are unified.

<u>Linear Regression</u>:

I used a basic Linear regression model to predict future stock prices. I used a very basic model as to know the actual complexity of my model and whether my calculation matrix give me good results. The results I got per company in my Nifty 50 were,

```
Symbol,MSE,RMSE,R-squared
ADANIPORTS,0.0049015032130839435,0.07001073641295272,0.14486726204123146
ASIANPAINT,0.09915058098690997,0.31488185242549305,0.4173322727616263
AXISBANK,0.014641602447177514,0.12100248942553832,0.4578282166119162
BAJAJ-AUTO,0.0164157450406568,0.12812394405674843,0.8204749454463022
BAJAJFINSV,0.23809588916825478,0.4879507036251252,0.8080923311453387
BAJFINANCE,0.24933595154270796,0.4993355099957422,0.5477975906482353
BHARTIARTL,0.004202896132508668,0.06482974728092551,0.3691646618487173
BPCL,0.003120179395319233,0.055858565997698444,0.4018463738806458
BRITANNIA,0.10289098881885436,0.3207662526184049,0.627300401438493
CIPLA,0.008248240592167425,0.09081982488513962,0.35358047568505613
COALINDIA,0.0002040922734682393,0.014286086709391039,0.7440011045089207
DRREDDY,0.052261585114106786,0.22860792880848815,0.7189503882825411
EICHERMOT,4.82521439477193,2.196637064872559,0.6523321914758629
GAIL,0.0013972132741365733,0.03737931612719223,0.4419527260053385
GRASIM,0.10471550895143421,0.3235977579518038,0.4792408696873909
HCLTECH,0.016471536672935747,0.12834148461403955,0.3993734579527346
HDFC,0.04994570204960407,0.22348535086131277,0.32703008687465407
HDFCBANK,0.02184396507626544,0.14779704014717426,0.6270311117973921
HEROMOTOCO,0.02440963045868083,0.15623581682405874,0.8624307497893899
HINDALCO,0.009368418935124155,0.096790593216098,0.49935834456282446
HINDUNILVR,0.03471988721624671,0.1863327325411365,0.504361095980592
ICICIBANK,0.009210310172663438,0.09597036090722717,0.538651199100803
INDUSINDBK,0.010242419271384563,0.10120483818170238,0.7857004636582221
INFY,0.2199483462667042,0.4689865096851979,0.5356778872053354
IOC,0.0020627944837095565,0.045417997354678205,0.2765220737604578
ITC,0.00937589779658131,0.09682921974580458,0.4147742119569313
JSWSTEEL,0.012202341746640191,0.11046421025219069,0.48616083880655736
KOTAKBANK,0.0056721943730728835,0.07531397196452251,0.8112072749160126
LT,0.023669576963789705,0.15384920202519642,0.5557200333895209
M&M,0.005692843982429415,0.07545093758482671,0.7264763511680103
MARUTI,0.20974183642238697,0.4579758033154011,0.8088423176806309
NESTLEIND,0.3578030270443925,0.5981663874244293,0.8612014944744705
NTPC,0.00010462496006138249,0.01022863432044486,0.463274641436184
ONGC,0.013183067505659493,0.11481754006100067,0.3911066869869425
POWERGRID,6.985607849544809e-05,0.008357994884866111,0.7330688447897754
RELIANCE,0.02140335090991124,0.14629884110925567,0.5617308403578392
SBIN,0.04850216996455373,0.22023208205108022,0.5421009752297278
SHREECEM,1.5080595125104208,1.2280307457512702,0.8379490207194148
SUNPHARMA,0.01727188672694773,0.13142255029844663,0.06962707419132008
TATAMOTORS,0.00838730277045276,0.0915822186368771,0.2027818175213183
TATASTEEL,0.0027380424508300105,0.05232630744501288,0.4758709369733056
TCS,0.02103304469739667,0.14502773768281937,0.7340027018418551
TECHM,0.02004983125169773,0.1415974267128387,0.4506870268797254
TITAN,0.0617214496720601,0.248438019779703,0.2673159389601958
ULTRACEMCO,0.03488254746920207,0.18676870045380214,0.9013009784413735
UPL,0.004790697755854114,0.06921486658120576,0.4285287162833603
VEDL,0.07170191577961561,0.2677721340610625,0.11220837427847619
WIPRO,0.06202316181318212,0.24904449765690892,0.39245641120945995
ZEEL,0.002142498122482811,0.046287126962934425,0.5452925749953512
```

And over all accuracy of LR was,

```
Overall MSE: 0.18523679630960369
Overall RMSE: 0.43039144544194147
Overall R-squared: 0.8148408711843901
```

<u>LSTM</u>: (Long Short – Term Memory)

The reason to select this model is because I have Date Time series Data set. And seasonality has clearly been seen in my DF. As a result, I used LSTM, which is an updated version of RNN (Recurrent Neural Network). By Hamad, R. (2023) LSTM is good to memories sequences and patterns hidden in the data set. RNN on other hand have the problem of vanishing gradients which means RNN has no ability to learn long-term dependencies.

The defined model is therefore composed of two LSTM layers, with a total of fifty units each. The properties of the first LSTM layer are as follows: it should be able to return sequences, which in turn will be used as an input for another LSTM layer. This second LSTM layer in turn feeds back the output at the final time step to the dense output layer of the architecture. Lastly, a dense layer is also implemented along with a single neuron that predicts the final output of the model. The Adam is used to compiles the model, which is the most efficient and fast optimizer used for many optimization problems. Secondly, the Mean Squared Error loss function is applied, which is suitable for regression tasks as the goal here is to minimize the square of the difference between the predicted and actual values of a continuous variable. I could have experimented with the LSTM layers but it took a much longer time than expected to run on the given data set.

	Symbol	MSE	RMSE	R-squared
1	ADANIPORTS	0.0013920876404242…	0.03731069069883685	0.20321835798615306
2	ASIANPAINT	0.011198604428009473	0.10582345877927764	0.6052615140546191
3	AXISBANK	0.0005619066645272…	0.023704570540873115	0.7105319876727632
4	BAJAJ-AUTO	0.0192581492520652…	0.13877373401355617	0.19570005937010537
5	BAJAJFINSV	0.47915218136539783	0.6922081922119947	-0.21724135918462784
6	BAJFINANCE	1.4264441728828878	1.1943383829061545	-5.8187740475326235
7	BHARTIARTL	0.0003877082484553…	0.0196903084906082…	0.6460817365931354
8	BPCL	0.01728587098835804	0.1314757429656058	-8.949049353306838
9	BRITANNIA	0.2504830361162329	0.5004828030174793	-0.5124695173820266
10	CIPLA	0.011220524644970283	0.1059269778902914	-5.274135033325812
11	COALINDIA	0.0018301312810847…	0.04278003367325385	-5.264557874283466
12	DRREDDY	0.26317483867976216	0.5130056906894525	-1.0127414089861762
13	EICHERMOT	11.166617589567998	3.3416489327228853	0.10901685446305465
14	GAIL	0.03633829555039274	0.19062606209643196	-11.582334262597747
15	GRASIM	0.0024227492550731…	0.04922143085154264	0.6580214307974006
16	HCLTECH	0.0012129395560357…	0.03482728177787901	0.7281890259868625
17	HDFC	0.0427914744748315…	0.20686100278890557	-1.940412618043624
18	HDFCBANK	0.011623130581565482	0.10781062369528098	0.5852357055669688
19	HEROMOTOCO	0.0089845144334568…	0.0947866785653812	0.787366196365652
20	HINDALCO	8.204671821410785e…	0.00905796435266268	0.7337971726029519
21	HINDUNILVR	0.10311834778270354	0.32112045681130863	-2.6620256188027622
22	ICICIBANK	0.0028999420059774…	0.053851109607671574	-1.14566398895137
23	INDUSINDBK	0.06152045734295085	0.2480331779076155	-0.9463498404367143
24	INFY	0.0034785368650055…	0.05897912228073214	0.537794576686415
25	IOC	0.00692423074986105	0.08321196278096708	-1.9025031026636534
26	ITC	0.0006090069919562…	0.02467806702228213	-1.35230906648719
27	JSWSTEEL	0.0001305716396550…	0.01142679481110402	0.8700018523473328
28	KOTAKBANK	0.0028519283780924…	0.05340344912168521	0.7552949812445975
29	LT	0.0006178845256208…	0.02485728315043344	0.8928935566053763
30	M&M	0.0038666906391284…	0.062182719779118885	0.7271938882118505
31	MARUTI	1.3206642997306899	1.1492015922938368	-5.408481783145292
32	NESTLEIND	0.3809514634232621	0.6172126565643823	0.5885667744757133
33	NTPC	0.0001520488127607…	0.0123308074658869…	-0.24231403578230948
34	ONGC	0.0009579214651490…	0.0309503063821519…	-2.8826186495235726
35	POWERGRID	0.0007109934346218…	0.0266644601412039…	-17.822603431539797
36	RELIANCE	0.0099771438586685…	0.09988565391821079	0.5571092935706645
37	SBIN	0.0026512812088201…	0.05149059340132085	-5.94270665931269
38	SHREECEM	1.0271218566683555	1.0134702051211746	0.364132255218548
39	SUNPHARMA	0.0203919643260170…	0.14280043531452208	-19.206649075016866
40	TATAMOTORS	0.0021351616879188…	0.04620780981521238	0.11863132261434117
41	TATASTEEL	0.0005582779860744…	0.0236279069338447…	0.7923108956154886
42	TCS	0.0368996329045348…	0.19209277160927962	-0.09301253971260248
43	TECHM	0.0003539229163761…	0.0188128391365084…	0.8438438087503721
44	TITAN	0.024110724602717442	0.15527628474019284	-0.860795334852062
45	ULTRACEMCO	0.25330007919721287	0.5032892599660883	-2.0162562310398497
46	UPL	0.04812014268958318	0.21936303856753803	-11.658138959461635
47	VEDL	0.0005290820318670…	0.0230017832323299…	0.3165078429851741
48	WIPRO	0.0220521587773781…	0.14849969285280742	-20.482843693669643
49	ZEEL	0.0048121469145903…	0.06936963971789387	-0.42797405697934376

```
Overall MSE: 0.3683391210029769
Overall RMSE: 0.6069094833687944
Overall R-squared: 0.8837588951567011
```

Improvement of Situation

Hyperparameter Optimization on LR:

Here I have used GridSearchCV to perform the fitting process of the LinearRegression model, and the fit_intercept parameter for the regression equation to decide whether to add an intercept term. The grid search is conducted for each country of the dataset with the model selection strategy being the 5-fold cross-validation (which changes the test set 5 times) to choose the model with the lowest mean squared error. This is useful in determining how the configuration of the linear regression model could be optimized when predicting the outcome of each of the groups.

After successful implementation of gridsearch and 5-fold-cross-validation I got the result where my accuracy increased by just 1%.

```
Overall MSE: 0.1852367963096037
Overall RMSE: 0.4303914454419415
Overall R-squared: 0.8148408711843901
```

Conclusion and Future work

Linear Regression and LSTM based models for stock price prediction. Linear Regression have a good average of the RMSE which is at 0. 430 and an R-squared value of 0. However, the accuracy and performance of LSTM model was superior, with the evaluated RMSE 0. 607 and a higher R-square equal to 0. 884, which shows that the proposed method enjoys higher predictive accuracy, as well as a better capability for temporal pattern analysis of the gathered data. This indicates that although Linear Regression is quite efficient, LSTM models are more appropriate when it comes to dealing with the imbedded seasonality and time features in the financial data sets.

It is clear to see that LSTM worked more efficiently on my data set as compared to the basic LR model even after hyperparameter tunning of LR. The accuracy I achieved with LSTM was 88% and the accuracy I achieved with LR was 80%, even after hyperparameter tuning of LR accuracy increased by just 1%.

I realised that hyperparameter tuning should have been used on other models like Random Forest where I could have introduced more hyperparameters and would come up with more good results with hyperparameter tuning. But my models as specially LSTM took a lot of time running on the datasets. Due to this complexity, I could not experiment with more models. My best model was LSTM.

To enhance accuracy, LSTM model's complexity has to be increased by adding more layers or neurons, testing with different time steps, and tuning hyperparameters. Additionally, increasing the training data size, inserting more relevant features, and using advanced techniques like attention mechanisms or ensemble methods could further enhance performance.

References

1) NIFTY-50 Stock Market Data (2000 - 2021) (2021). https://www.kaggle.com/datasets/rohanrao/nifty50-stock-market data?resource=download.

2) Agarrwal, K. (2022) 'What is NIFTY 50? How to invest in NIFTY 50,' Forbes Advisor INDIA, 22 April. https://www.forbes.com/advisor/in/investing/what-is-nifty 50-how-toinvest-in-nifty-50/.

3) Anushagali (2024) INFO6105_EDA. https://www.kaggle.com/code/anushagali/info6105eda.

4) India, N. (no date) Historical Index data. https://www.nseindia.com/reports-indices historical-index-data.

5) Hamad, R. (2023). *What is LSTM? Introduction to Long Short-Term Memory*. [online] Medium. Available at: https://medium.com/@rebeen.jaff/what-islstm-introduction-to-long-short-term-memory-66bd3855b9ce#:~:text=Long%20Short%2DTerm%20Memory%20(LSTM.