Project Writeup - Predicting The S&P500 Market1

Objectives

Below is the list of questions that we are looking to answer as a final outcome of this project:

- Using historical data on the price of the S&P500 Index, make predictions about future stock market status.

Goal Significance

Why does the list of objectives above matter? What benefit we could derive by developing Stock Market Predicting Model. Below are the goals we can enlist:

- This information will help us forecast how the overall stock market as a whole will perform over the time.
- The stock market performance will, in turn, give us an insight about the overall economic forecast.
- The S&P500 Index movement can provide us a useful information about the business status of the top 500 corporate house in the nation.

Data

Data Source

The project uses the S&P500 Index data for the period of 03rd January 1950 thru 07th December 2015. The data is collected from the **yahoo finance** site.
S&P500 Index (01-03-1950 thru 12-07-2015)

Data Lists

The data is available in the .csv file format.
File name: `sphist.csv`

Data Extraction Details

The data contains following information about the S&P500 historical movement:

Date: Date of the market record
Open: The opening price on the date above when trading started.
High: The highest trading price on the date.
Low: The lowest trading price on the date.
Close: The closing price for the day when trading ended.
Volume: The number of shares traded on the date.

Adj Close: The daily closing price, adjusted retroactively to include any possible corporate actions.
(for more info on the adjusted closing price, please refer: http://www.investopedia.com/terms/a/adjusted_closing_price.asp)

Project Writeup - Predicting The S&P500 Market1

Model

How was the model created?

After reviewing the available info,

- Manipulate the available features and introduced some new features
- Extract the data pertaining to the new features
- Clean the data to facilitate the new features adjustments
- Verify the features correlations and predictive significance
- Carry out Linear Regression Analysis using the scikit-learn package.

Why only this model?

Linear Regression offers suitable supervised learning algorithm to provide better predictive model over continuous data. The reliability of the Linear Regression Model gets enhanced when its generalization over the unknown data can be better checked and confirmed by trying out over the cross-validation set.

Highlights of the code.

Software packages used:

- Python
- Pandas
- Numpy
- datetime
- Seaborn
- Matplotlib.pyplot
- Sklearn –

LinearRegression
mean_absolute_error

Overview:

- Read the data and form the dataframe
- Initialize new features
- Generating moving average and variance for different durations
- Update data frame with the data for the new features
- Data Cleaning, data slicing and development of training/test/CV sets
- Feature identification and feature scaling
- Development of predictive model using Linear Regression
- Feature significance assessment and model fit evaluation
- Learning algorithm predictive verification and error matrix analysis
- Check for model output generalization

Project Writeup - Predicting The S&P500 Market1

How does the data fit to the model?

The model fitting over the training set was assessed by checking the MAE and the same algorithm was verified over the test set.

Model Validation Details.

The model estimation over the cross-validation set was verified by getting the MAE. This error matrix was compared with the similar MAE output over the test set where the model is originally tested. The difference in errors were checked against the pre-determined tolerance to validate the model predictability.

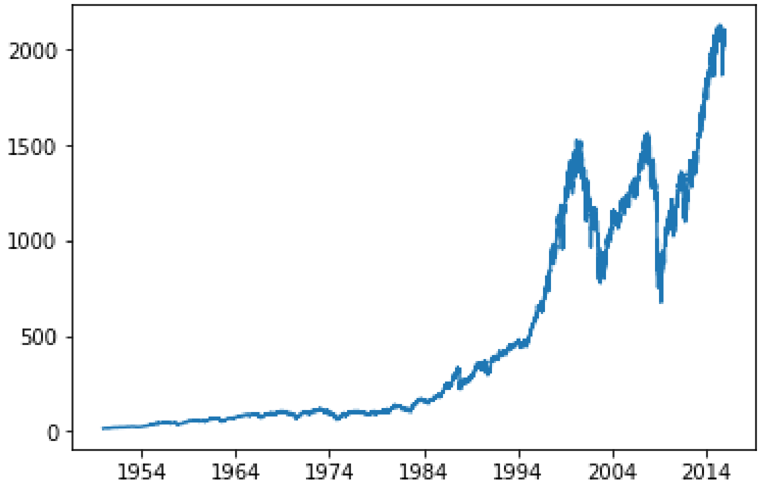
Justification for the meaningfulness of the model outcome.

The model predictive estimation error comes out to be **zero**, i.e. in other words, the model shows **100%** accuracy in predicting the future S&P500 index.

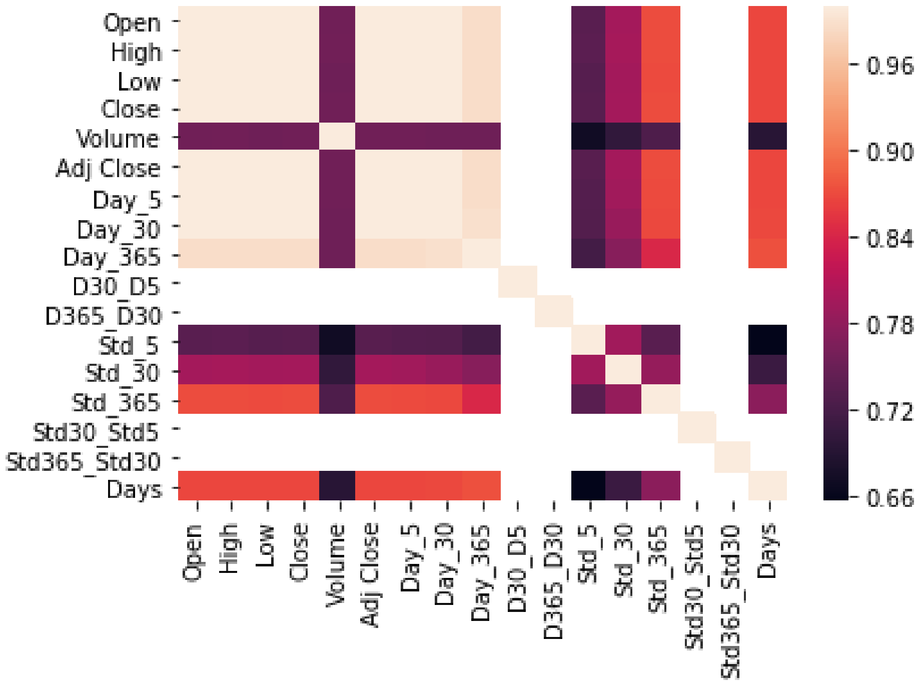
Results

Visualize the results.

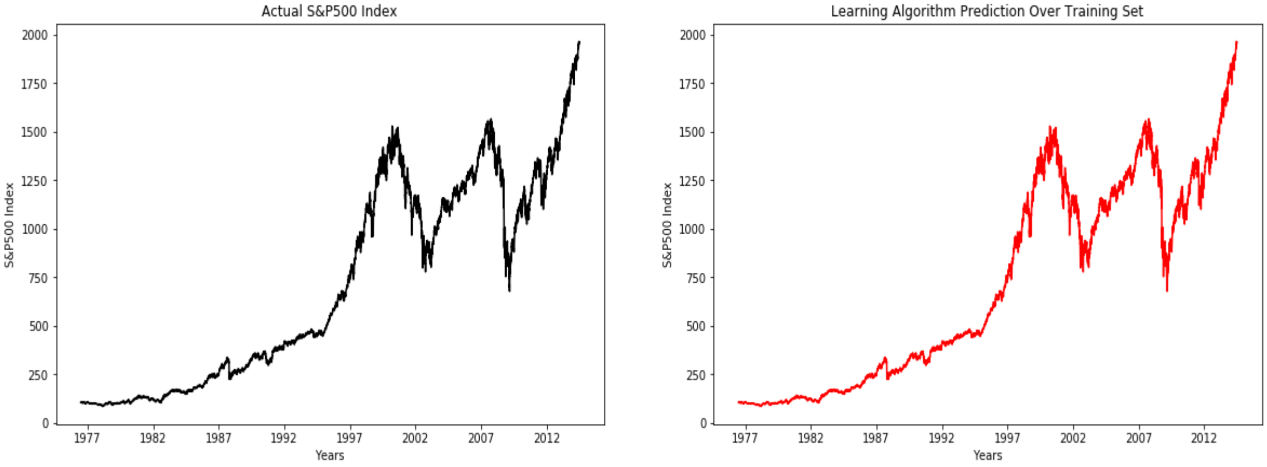
- **Closing S&P500 Index over Years**



- **Features Correlation Mapping**

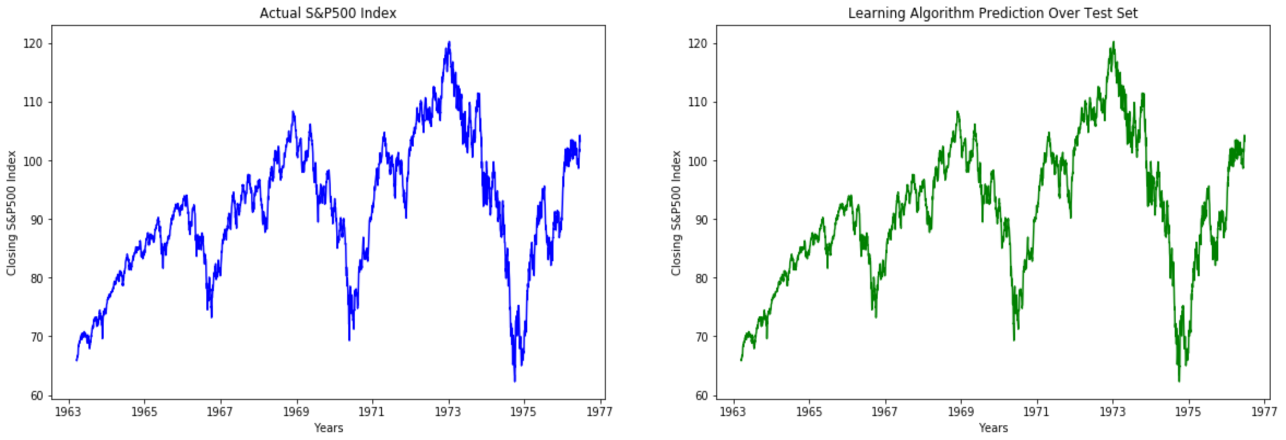


- **Comparing Model Fit over Training Set**

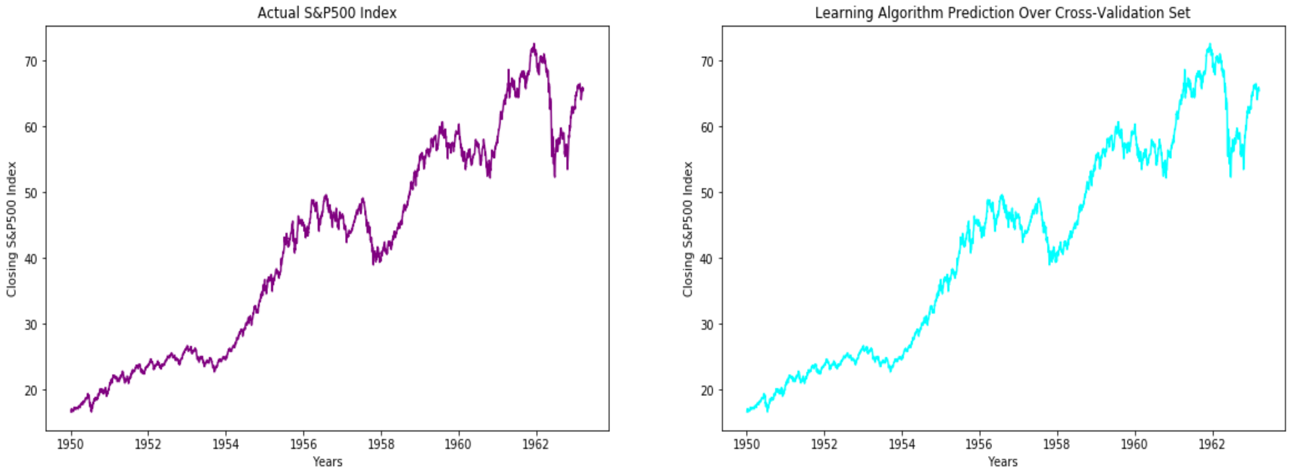


- **Comparing Model Trained Over the Test Set**

Project Writeup - Predicting The S&P500 Market1



- **Comparing Model Prediction over the Cross-Validation Set**



Explain the results in a simple, concise and easy way. (non-technical audience)

- The model shows highly reliable predictability for the future stock market status. In other words, we can get prediction about the likely S&P500 index with high confidence.

Explain the results in the most technical way. (technical, data-scientist audience)

- The Mean Absolute Error (MAE) for the Linear Regression turns out to be zero while model fit was tested over the training set. This may be caused by overfitting of the learning algorithm.
- The algorithm, while train over the "Test Set', gives zero MAE. This reduces our concern for the overfitting of the predictive outcome. However, to ascertain the model generalization, it needs to be verified for prediction over the cross-validation set, a completely independent and unknown data.
- Model prediction generates zero error over the "Cross-Validation Set'. This further enhances the learning algorithm's predictability confidence over the external data.

Conclusion

What we learn from this outcome. (non-technical audience)

The prediction about the S&P500 Index with reasonable confidence gives us better insight not only to manage our own investment portfolio but also about the overall micro economic growth trends.

Technical significance of the results. (technical, data-science audience)

- The model fit over the training set can be confirmed by the pattern similarity with the overall S&P500 index over the data for the last 64 years. This may pose a concern for model overfitting over the training set and its questionable predictions over the data set other than the training set.
- However, the predictive outcome of the learning algorithm shows zero Mean Absolute Error (MAE) over Test-Set as well as Cross-Validation Set. This can be easily visualized from the pattern similarity as shown in the comparative charts for both the cases.
- The results prove the model generalization for predictability over the unknown external data set.

Suggestion for Further Development

How differently you would have done this project, if get a chance to do it again. Why?

In case the model predictions over the cross-validation set would have shown some considerable errors, I might have tried to deduce some more features utilizing the available data and encourage the learning algorithm to develop more complex cost function for better predictability.

Your suggestions for someone if s/he wants to further continue this work.

Someone could pick one or more of the untouched data fields and continue this journey further to see the correlations between these factors and S&P500 Index pattern over the years.