Predict stock prices using long short-term memory networks

Suyash Tandon

1 Background

Understanding market behavior, and trading and investments are complex time-dependent problems. One could rely on the historic data available in the form of tabulated stock prices and company performance metric, and use tool from statics and mathematical modeling to predict future trends[1]. Such quantitative analysis is essential for many investment firms, and hedge funds. While company performance, and related metrics affect the stock trades, other factors like the brand value, policies, news and events, all influence the stock prices. Therefore, all these factors involved in the prediction – physical factors vs. physhological, rational and irrational behaviour, etc., make share prices volatile and very difficult to predict with a high degree of accuracy. In the recent years, machine learning has shown promising results in feature extraction, pattern recognition, and parameter tuning. Therefore, machine learning algorithms are a good candidate for making stock price predictions[2].

In this project, the historical data of stock prices of a publicly listed company will be used. A mix of machine learning algorithms will be used to predict the future stock price of this company, starting with simple algorithms like linear regression, extreme gradient boosting, and then moving on to advanced techniques like long short-term memory. The core idea behind this project is to showcase the learnings from this nanodegree program by implementing these algorithms, and to derive understanding that'll be useful for application in future research work.

2 Problem Description

Time Series forecasting and modeling plays an important role in data analysis. Time series analysis is a specialized branch of statistics used extensively in fields such as Econometrics and Operation Research. The aim of this project is to predict the future closing price of a given stock over a given period of time.

2.1 Goals:

- ○ Explore and pre-process stock price dataset,
- ○ Implement Linear Regression as the first benchmark,
- ○ Implement Extreme Gradient Boosting (XGBoost) as the second benchmark,
- ○ Implement Long Short-Term Memory using Keras,
- ○ Compile and compare the results to evaluate model performances.

[1] Brown, L. D. (1993). Earnings forecasting research: its implications for capital markets research. *International journal of forecasting*, 9(3), 295-320.

[2] Khan, Z. H., Alin, T. S., & Hussain, M. A. (2011). Price prediction of share market using artificial neural network (ANN). *International Journal of Computer Applications*, 22(2), 42-47.

3 Datasets and inputs

Stock behavior for different companies and industries is different. Hence, this project seeks to build an end-to-end framework, that learns the trends from a given dataset and then gives prediction for some future time. To demonstrate this framework, dataset of the technology giant, Apple Inc. will be pulled from the online repository hosted at Quandl[3]. For the purpose of this project, the acquired dataset will be split as 60% train, 20% validation, and 20% test. The model will be trained using the train set, model hyperparameters will be tuned using the validation set, and finally the performance of the model will be reported using the test set.

4 Solution Statement

The stock pice prediction using machine learning has become one of the standard, de-facto problems to test models on time-series data. There are many implementations for this problem[4]. A variant of the recurrent neural networks (RNN) is the Long Short-Term Memory (LSTM) network that specializes in learning long term dependencies [5]. Due to this property, the LSTMs appear to be good candidates for time-series data prediction[6]. The project will be implemented in a Jupyter Notebook using Amazon SageMaker. Using Keras implementation of the Tensor Flow library, the solution will use the LSTM neural net model. The measures of the performance will be based on the predicted stock price in comparison to both the actual price and the benchmark models predicted price.

5 Benchmark Model

This project implementation will use the Linear Regression and the Extreme Gradient Boosting (XGBoost) methods as benchmarks.

1. Linear Regression: The linear regression is a basic Machine Learning model that returns an equation which determines the relationship between the independent variables and the dependent variable. The equation for linear regression can be written as:

   $$Y = x_1\theta_1 + x_2\theta_2 + \cdots + x_n\theta_n.$$

   Here, x_i are the independent variables and θ_i are the respective coefficients for $i = 1, 2, \ldots, n$.

2. XGBoost: Gradient boosting is a process to convert weak learners to strong learners, in an iterative fashion. The name XGBoost refers to the engineering goal to push the limit of computational resources for boosted tree algorithms[7]. XGBoost has proven to be a very powerful machine learning technique and is usually the go-to algorithm in many Machine Learning competitions.

Both these models are extensively used for supervised learning, where given an input X, the method gives a prediction Y, and the model can be easily described as

$$Y = f(X).$$

It'd therefore, be necessary to adapt the time-series dataset for these regression models such that the previous time step becomes the input and the next time step is the output of the supervised learning[8]. In addition, the order between the observations must be preserved, and since there is no

[3] https://www.quandl.com

[4] https://www.analyticsvidhya.com/blog/2018/10/predicting-stock-price-machine-learningnd-deep-learning-techniques-python/

[5] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.

[6] Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long short term memory. *PloS one*, 12(7), e0180944.

[7] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SiGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM.

[8] https://machinelearningmastery.com/time-series-forecasting-supervised-learning/

previous value that can be used to predict the first value in the sequence, this row cannot be used.

6 Evaluation Metric

The model performance for this project will be evaluated by measuring the mean squared difference between the predicted and actual values of the target stock at adjusted close price, and the delta between the performance of the benchmark models and the LSTM implementation.

7 Project Design

A high-level overview of the project workflow is described below.

1. Setup the project resources
    - setup Notebook instance on Amazon SageMaker,
    - import necessary libraries,
    - download the dataset,
    - create a role, setup a session, create a S3 bucket.
2. Prepare dataset
    - load the dataset in pandas framework,
    - normalize the data,
    - split the dataset for training, validation, and testing.
3. Implement benchmark models
    - Setup LinearRegressor model,
    - Setup XGBoost model.
4. Implement LSTM model
    - Setup LSTM model with Keras utilizing parameters from benchmark model,
    - Tune hyperparameters.
5. Evaluate and compare model performance
6. Document and visualize results
    - Plot actual and predicted values,
    - Analyze and describe results for the project report.