University of Wisconsin-Madison

Temporal Analysis of Spotify Stock Prices

Final Report

Josie Beres, Neeraj Deshingkar, AJ Koenig, Angad Vir Singh

ISyE 603

Executive Summary

Making up nearly one third of the market for audio streaming services, Spotify is an application that millions of people use on a daily basis. Seeing how Spotify has been growing lately and its apparent success, our team decided to analyze the prices of Spotify stocks since the company went public four years ago and develop temporal models to check if one can accurately forecast future values.The goal of this project was to determine if Spotify stock is a worthwhile investment, what the best model is for predicting stock prices, and if it is possible to predict an increase or decrease in Spotify stock prices on a daily basis.

We began our analysis by partitioning the data so that 90% of it could be used as a training set and the remaining 10% as a testing or validation set. The rolling forward method was also used when training the models so that only the most recent data would be taken into account while forecasting and the results would be more accurate. Several predictive models were developed, but our analysis showed that the linear regression model was the most accurate due to its ability to model the continuous downward trend found in the validation dataset. The accuracy of the deep learning model was quite similar to that of the linear regression model, but it was computationally expensive to train this model and hypertune the parameters. This deep learning model with three hidden layers was quite flexible in its forecast and therefore, could be used as a baseline for future analysis and research - especially with a larger planning horizon.

Ultimately, using these models, it was determined that Spotify stock prices will likely continue to decrease or simply level out in the near future, so stock will be fairly cheap to purchase. With this in mind, an investor must be looking for long-term growth rather than short-term profitability. Additionally, through the creation of a logistic regression model, we determined that predictions for daily increases or decreases in stock prices are basically the equivalence of guessing when it comes to accuracy. In conclusion, this analysis revealed that it is very difficult to predict stock prices and that such prediction models should be used only as a guide rather than relying on them fully.

Introduction

Since its launch in 2008, Spotify has become an international streaming giant, ending 2021 with 406 million monthly active users. According to a study by Midia Research, as of the end of the second quarter of 2021, Spotify was the largest digital service provider, amassing a 31% share of the audio streaming market. Due to the company's growth and apparent success in recent years, our team was interested in analyzing how its market value has fluctuated and determining if it would be possible to accurately predict future values.

Description of Data

Spotify went public in April 2008 and has since been trading stock on the New York Stock Exchange (NYSE). The dataset selected for this project contains the stock prices for Spotify from April 3rd, 2018 to March 4th, 2022, and was obtained from the NASDAQ website. The closing, opening, high, and low stock price for each date within the timeframe, along with the volume of stocks, are included in this dataset and were considered for analysis.

Description of Problem

In order to determine if Spotify stock is a worthwhile investment, an in-depth, temporal analysis was conducted on its stock prices. Historical data was examined and then used to develop predictive models to forecast stock prices for a validation set. In addition to determining whether Spotify stock is a worthwhile investment, we wanted to identify which type of predictive model provides the most accurate forecasts for stock prices. We also wanted to determine whether or not it is possible to predict if Spotify stock prices will increase or decrease on a daily basis. With these goals established, it was possible to shift into the analysis stage of the project.

Approaches

Throughout this project, several analytical approaches that were learned in the classroom were put to use. Smoothing methods and decomposition were used on the data once it was cleaned in order to better understand its four components: trend, seasonality, level, and noise. Following these basic explorations, the data was modeled in a linear manner with time series components. A linear regression model and an ARIMA model were developed, along with a simple neural network and a deep learning model. Upon completion of these models, a logistic regression model with a binary predictor was developed. Each of these models were used with training and validation sets to forecast future stock prices. Lag variables and rolling windows were also used to improve the accuracy of these forecasts.

Data Preprocessing

In order to best handle the data of interest, both tsibble and timeseries objects were used for modeling. It was critical to identify that there were missing dates within the series due to federal holidays and weekends. For the purposes of this analysis, it was assumed that missing days due to holidays do not have an impact on the next open stock price for the next day. This meant that we could not assign a frequency, such as weekly or monthly, to our time series objects. Rolling windows were then used to iterate through the data and choose the best training and validation periods to model it. Feature engineering was also performed to create lag variables for the open price and volume. These variables, consisting of seven for the open price and one for the volume of stocks, were then used as predictors in our regression models. The use of Lag variables was essential since stock price is inherently volatile and the dynamics of the stock market change everyday. By incorporating these features into our models, it was possible to produce better forecasts.

Summary of Preliminary Results

To begin the analysis, a timeplot was created for the data; the opening stock price for each day was used in this plot as it will be the most relevant value for an individual looking to purchase or sell stock. It can be noted in Figure 1 that there was a dramatic increase in stock prices between 2020 and 2021, with a peak visible around the first quarter of 2021. Since then, the open stock price appears to be decreasing and there has been a sharp decline at the beginning of 2022.

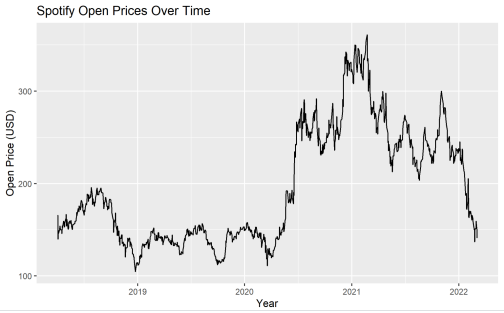

Figure 1: Spotify opening stock price time plot.

Since the stock market is not open on weekends or holidays, we thought it could be interesting to assign another factor to each datapoint that identifies which day of the week the date corresponded with; this could reveal if the dataset contained any seasonality based on the days of the week. Figure 2 contains a bar plot of the number of datapoints associated with each weekday. Tuesdays appear to be the most prevalent in the dataset, while Mondays were missing the most often due to the celebration of federal holidays. We were then able to look further into how stock prices tend to change based on the day of the week. Figure 2 also depicts the number of days in which opening stock prices went up and the weekday that they correspond with. It appears that stock prices increased most often on Thursdays and Fridays.

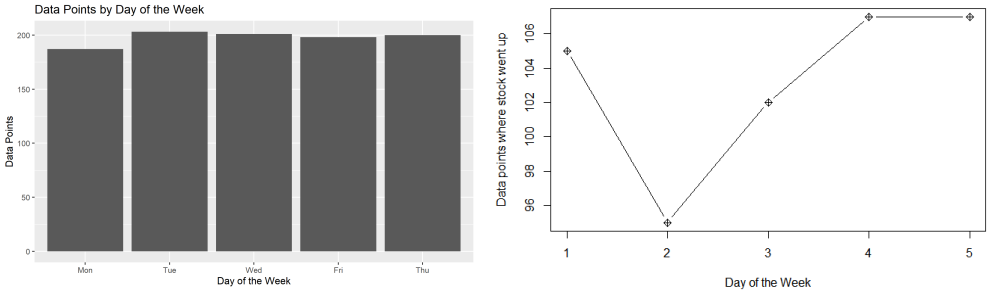

Figure 2: Bar plot of days of the week present in stock data and plot for days of the week and number of times the stock went up on that day.

A candlestick chart was also created for the Spotify stock data and can be seen in Figure 3. This type of financial chart is used by traders to determine possible price movement based on past patterns and incorporates the four price points included in our dataset: open, close, high, and low.

Figure 3: Candlestick chart representing open, high, low, and close values for each day.

Analysis

Data Partitioning

We initially split our dataset so that 90% of the data points could be used as a training set for our models and the remaining 10% could be used as a validation set to see if the models are performing well. After running our initial models and looking at the preliminary results, however, it was discovered that in order to forecast more accurately, the rolling window method should be implemented. The older data was actually having a negative impact on the accuracy of our models due to outdated characteristics and factors that were not applicable in the present time or useful in forecasting future opening stock prices. To implement the rolling window method, a nested for loop was used to iterate through the data and identify the window size in which the model best fit the test set. The window with the lowest mean absolute percentage error was selected to build the models. In using the rolling window approach, we were able to take recent lag values into account when forecasting for the open price of the next day. This method significantly improved the accuracy and reliability of each model in which it was utilized.

Time Series Decomposition

To better understand the data, the team began by decomposing the time series. As can be seen in Figure 4, there was an observable trend in the data, along with what appeared to be a very rapidly changing seasonality. This seasonality, however, was contrary to the seasonality we had seen before. While we generally see monthly or annual seasonality, we noticed that this stock had rapid, daily seasonality which was indicative of weekday trends. Keeping the decomposition of the time series in mind, we then went on to check the accuracy of different models that used time series.

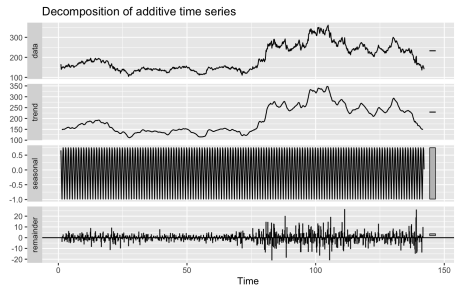

Figure 4: Decomposition of additive time series

Exponential Smoothing Model

We first trained an exponential smoothing model for the data and checked its accuracy. This was a very primitive model, and consequently the results did not surprise us. As can be seen in Figure 5, this was not at all an accurate model or one which we would want to take inspiration from. To improve the accuracy, we thought we would try to implement the rolling forward method in this model as well. However, we saw that since the ets function only accepts univariate time series, we were not able to add the lag components of the data to create the rolling window. Accordingly, this did not provide satisfactory results.

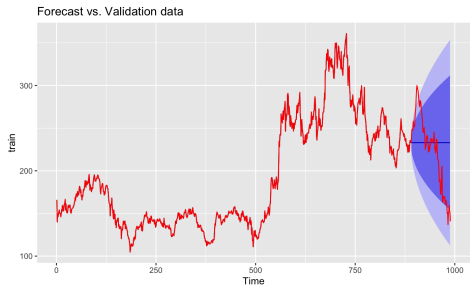

Figure 5: Exponential smoothing forecast vs. validation data

Linear Regression Model

In linear regression, one models the relationship between the target variable and the explanatory variable(s) by using a linear approach. When we used the linear regression model on our data, we received much more accurate results than with the previous smoothing model. Unlike with a smoothing model, we could incorporate the rolling forward method here as the tslm function accepts multivariate time series to train the model. In our tslm model, we used the lag variables and the volume of the stock as predictors to forecast our target variable.

In figure 6, one can see how the forecasted trend is very similar to the actual trend as well. The actual trend is represented by black and the forecasted trend is in red. In the graph on the left, one can see the actual trend during the training as well as validation period, along with the forecast for the validation period. The figure on the right is a maximized version of the trend particularly in the validation period (so that the trend is easier to see). We see that the linear regression model is able to predict the downward trend in the recent months which is encouraging to see.

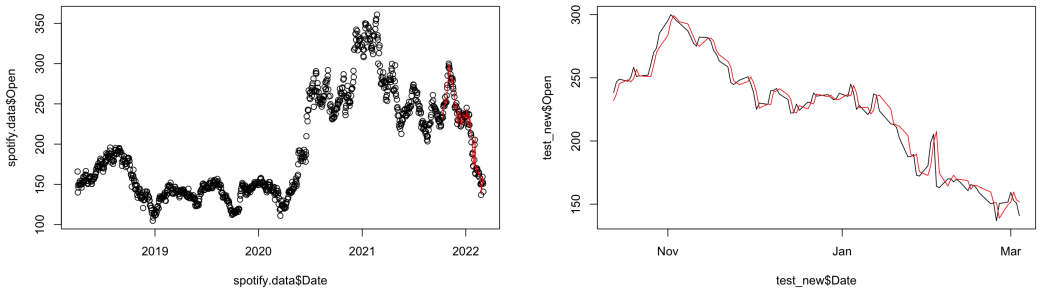

Figure 6: Linear regression model validation period

ARIMA Model

An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or predict future trends. A statistical model is autoregressive if it predicts future values based on past values. We chose to use an ARIMA model due to its power in forecasting for time series based on lag values. A rolling window approach was used to find the best ARIMA model and the corresponding values of p,d,and q that give the best accuracy for the test set. The ARIMA(1,1,0) model, seen in Figure 7, performed the best on the validation set and resulted in the lowest MAPE value. The autocorrelation function (ACF) plot of the residuals from the ARIMA model showed no significant autocorrelation, as seen in Figure 7, so we could use this model to predict future stock prices. One significant advantage of the ARIMA model is its ability to capture linear relationships which proved particularly useful in capturing the trend of the stock prices. The ARIMA model lacks flexibility in predicting sharp turning points, however, and hence the forecasts were not completely accurate when there was a sharp downward or upward trend.

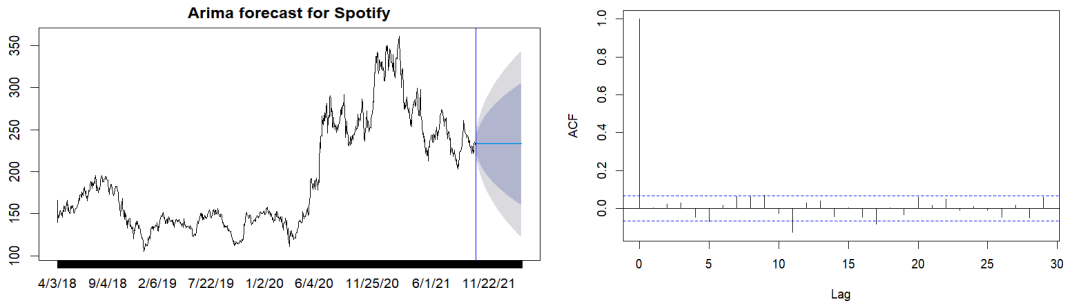

Figure 7: ARIMA (1,1,0) model forecast and corresponding ACF plot

Neural Network

To provide more flexibility to our model and overcome the shortcomings of the ARIMA model, a simple neural network with one hidden layer, seen in Figure 8, was developed. The “nnetar” function in the “forecast” package for R fits a neural network model to a time series with lagged values of the time series as inputs. It is a nonlinear autoregressive model and hence can forecast the time series accurately even for sharp turning points.

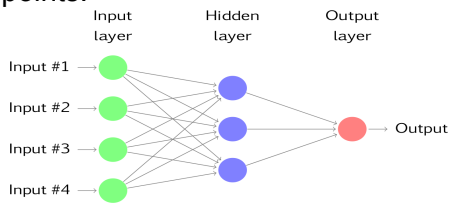

Figure 8: Simple neural network architecture

A rolling window approach was used to find the best lag input value. We used a for loop and iterated through various combinations until we found that an input size of 33 would be optimal. These 33 lag values of the time series were then used to forecast the next data point, and the results can be seen in Figure 9. One of the shortcomings of this model was that our validation set consisted of a series with a continuous downward linear trend.The model was not able to predict this accurately and hence the MAPE value was high on the validation set.

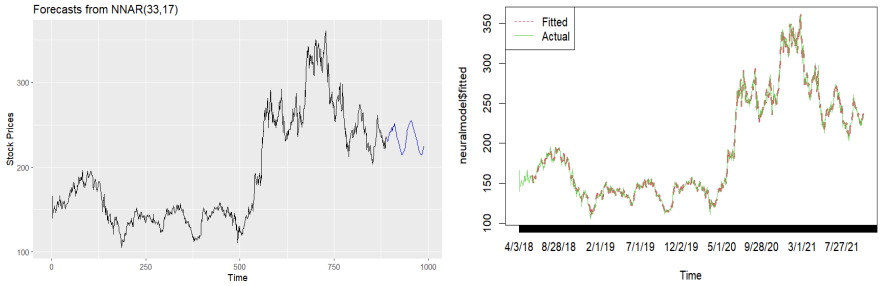

Figure 9: Forecasts of Neural network model

Deep Learning

After determining the inaccuracy of the ARIMA model, a deep learning model with three hidden layers consisting of 64 nodes and one output layer was developed. We used grid search hyperparameter tuning to find the best input size, batch size and activation function for the model. Over 35 values for the input size, seven values of batch size, and two activation functions (“Relu” and “Sigmoid”) were tested to find the optimal hyperparameters for our model. We evaluated the hyperparameters on the test set to evaluate their performance by using a for loop. It was determined that a batch size of 25, an input size of 13, and the activation function set as “Relu,” were the best hyperparameters and gave the lowest MAPE value on the test set. The default settings were maintained for the other hyperparameters and an Rmsprop optimizer was used in our model. This deep learning model had one of the best performances and was able to capture the inherent and volatile characteristics of the time series; it was able to capture the linear downtrend of our test set accurately and the results are pictured in Figure 10.

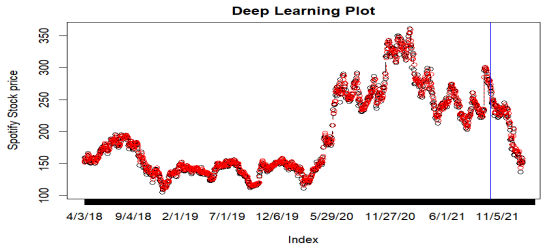

Figure 10: Deep learning model with rolling window forecast

Logistic Regression

To simplify our model and attempt to understand if any gain could be gleaned from the lag variables, we created a logistic regression model to see if we could predict stock direction. Using our added direction variable, we used the volume for the previous day and the open price for the five previous days to predict the direction of the next day. Looking at the resulting confusion matrix and ROC curve in Figure 11, an extremely high false positive rate was found in the validation set. The area under the curve was 0.562, which was just slightly better than random guessing.

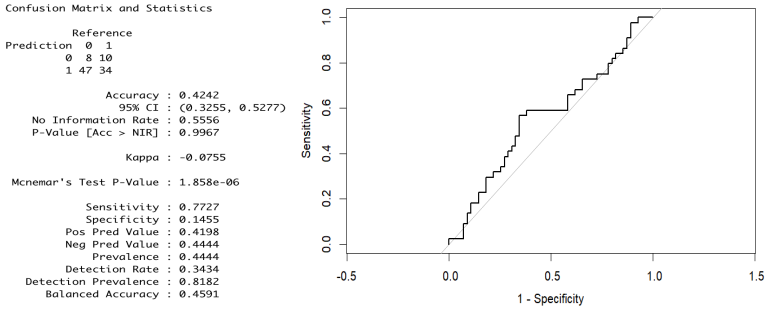

Figure 11: Confusion matrix and ROC curve for logistic regression model

Summary of Results

In comparing the RMSE and MAPE values of our various models, which are presented in Table 1, we were surprised to find that the linear regression model performed best on the validation set, with the deep learning model following close behind. These two models had the lowest error values. The success of the linear regression model can likely be attributed to the recent sharp downturn of stock, mimicking a linear trend. The deep learning model may be better suited to future predictions over a larger planning horizon, however the additional computational power required for the deep learning model presents a tradeoff between the two models. Not pictured in Table 1 are the results of the logistic regression model, which proved to be no better than random guessing, as it had an accuracy of 42.42%.

Model Name	***RMSE***	***MAPE***
Exponential Smoothing	43.074	17.06
Linear Regression	**8.289**	**2.874**
ARIMA	43.064	17.05
Neural Network (1 Hidden Layer)	37.756	15.54
Deep Learning	**8.858**	**2.95**

Table 1: RMSE and MAPE for each model tested

Potential Improvements

As the inaccuracy of the logistic regression model was realized, our team discussed possible ways to improve our modeling with a binary predictor. It is possible that a deep learning model could be used to more accurately forecast with this binary factor of whether or not stock price will increase, so this would be a useful addition to our analysis. One of the other key takeaways from this project was that technical factors are not sufficient on their own to predict stock prices. The stock market is pretty volatile and technical factors do not provide enough information for accurate predictions. If this project

were to be continued, a potential improvement to this analysis would be to use a stock market index, like the S&P 500, as a feature to see if it has an effect on the prices of Spotify stock. We could also compare the stock prices of other music streaming platforms to see if they have a relationship with the stock prices of Spotify. Additionally, dummy variables could be created to account for outside factors, such as a pandemic, economic recession, festival, or holiday, to see if they have an effect on the stock prices. Another way that this analysis could be improved would be through the evaluation of the "day of the week" factor as a predictor to determine the probability that the stock price will increase or decrease on a specific day. Lastly, the most important thing that could be done to improve this project would be to include a sentimental analysis of social media, news articles, and other events in popular culture that could cause a huge oscillation in the stock market. We have recently seen the power of social media and how it is used to manipulate the stock prices of a certain company, so this analysis will prove very significant in forecasting spotify stock prices.

Lesson Learned

Through the completion of this project, our team learned much about the application of temporal analytics in the real world. First and foremost, we came to the conclusion that stock prices are extremely difficult to forecast. The stock market is extremely volatile, so it can be challenging to model - especially when events like the COVID-19 pandemic disrupt any previous trends. Additionally, if it was easy to model and forecast stock prices, everyone would be rich! When it comes to modeling the data, we recognize that, as George Box once said, "All models are wrong, but some are useful." The models that we have developed can be used to guide and support decision making with stocks, but should not be relied upon. Additionally, it was extremely useful to learn about how a large training set can sometimes be detrimental to a forecasting model, so a rolling window should be used to select the proper training and validation sets.

Conclusion

Returning to the original goals of this project, it was possible to analyze the current trends of Spotify stock prices and come to a conclusion about the value of this investment. As Spotify stock appears to be on the downturn, stock prices are fairly cheap, so it could be a good time to buy; however, an investor in this stock must be looking for long-term growth rather than short-term profitability, because the prices will likely continue to decrease or simply level out in the near future. As per the results of our modeling, the team determined that the linear regression model and the deep learning model are the best for forecasting Spotify stock prices. The linear regression model produced the lowest RMSE and MAPE values, which is likely due to the recent negative, linear trend that open prices have experienced. It is likely that this model would be best for predicting the open price for the next day. The deep learning model would be better suited for a larger planning horizon, but it requires significantly more computing power, so it could be more difficult to use. Due to this tradeoff, our team concluded that the simpler model could be sufficient, but it could also be useful to utilize both models in tandem. Unfortunately, it is not possible to accurately predict whether Spotify stock prices will increase or decrease on a daily basis, as seen in the logistic regression model - interested parties will have nearly the same luck guessing. Overall, this analysis was extremely useful, as it provided our team with the opportunity to investigate how temporal analytics can be applied to real-world data.

References

https://www.nasdaq.com/market-activity/stocks/spot/historical
https://newsroom.spotify.com/company-info/
https://www.midiaresearch.com/blog/music-subscriber-market-shares-q2-2021