Importance of Big data in predicting stock returns and price

Gagan Arora
Indiana University
2709 E 10th St
Bloomington, Indiana 47401
gkarora@iu.edu

ABSTRACT

In this project, we will discuss the importance of big data in finance industry in predicting financial stock values. We will be using python libraries to fetch financial data from yahoo finance and will further predict the stock price returns of few selected technology companies such as Amazon, Yahoo depending on the historical data of x[16] years. Similarly, we will predict the returns based on y[10] years of data. The prediction will be based on SP 500 market return and market risk volatility. Here y is greater than x and then we will compare the predicted returns with the current returns. For the comparison we will be using the testing time frame as mentioned in the project later. This project will help us understand if more historic data helps in predicting the stock price returns or it adds noise. We will be using statistical approach and CAPM [capital asset pricing model] to predict stock price. Analysis will be done on the jupyter notebook

KEYWORDS

HID-301, Stock Price prediction, stock returns, SP500, risk free market, CAPM model, root mean square analysis, stock beta, Finance, Statistics,mean, variance, market premium, python,yahoo finance, i523

1 INTRODUCTION

By its nature of the business, the finance industry is always driven and dominated by data. The existence of Big data in the finance industry has exposed the big opportunity of growth and value extraction but at the same time imposed the various new challenges, which demand new skill set. [5] suggests that finance experts believe there is a huge potential in terms of value extraction from the financial big data. They also believe that finance industry can benefit more than any other industry. Historically, data was always there in some format either non-digital or digital. However, with digitalization, this data has fallen into the prevalence of high volume of information, which we call as Big Data. Dominant drivers for the actuality of big data in the finance industry are mainly customer call logs, social media, news feed, regulatory data etc. Call logs, news feed and etc. fall into the category of unstructured data which is identified as an area where we can extract vast amount of business value.

[4] talks about the three V of big data in finance industry: volume, velocity and variety. [6] clearly depicts the amount of financial data pouring in the daily basis. TechNavios forecast (Technavio 2016) predicts data will grow at a CAGR [compound annual growth rate] of 61 percent over the period of 2017-2021. According to the IDC financial insight 2016, every second there is around 10,000-payment card transaction and this number is expected to double by the end of this decade. The Capgemini/RBS Global payments study for 2012 suggests there was about 260 billion transactions in 2012 and is expected to grow between 15 and 22 percent for developing countries. Main drivers contributing to the big data in the finance industry are Data growth, increasing scrutiny from regulators, digitalization of financial products, changing the business model and increased customer insight platforms such as customer service. [4] shows 76 percent of banks say the business driver for embracing big data is to enhance customer engagement, retention, and loyalty and seventy one percent of banks say that to increase their revenue, they need to better understand customers and big data will help them to do so.

Thinking about the data strategy, the financial industry has taken the business-driven approach to a big data. According to the IBM report, all financial organizations are not keeping the same pace as peer industry is keeping. Today because of increased competition, customers always expect more personalized banking service and at the same time, there is increased regulatory surveillance which in result creates big pressure on finance industry to better utilize the value of Big data. To achieve better-personalized experience, many banks have started the initiative to utilize the information gained from the vast ocean of data to offer better-personalized products and gain competitive advantage. Despite the fact that financial industry is data-driven, there is a gap in the amount of initiative financial industry has taken to extract the value out of big financial data. Technavio 2016 report has shown only 26 percent of financial organizations has focused on understanding the principal notation of Big data and most of those 26 percent are still struggling to define the clear roadmap. This clearly concludes that finance industry lag behind their cross-industry peers in using more varied data types. A good example to support this fact is that there are very less research and domain knowledge in extracting value out of retail bank call logs.

Big data technologies not only help in extracting the effective business value but analysis of unstructured data in conjunction with a wide variety of data set also helps in extracting commercial value. Big data in finance industry does not necessarily decode to valuable or actionable information. The real benefit lies in developing the technologies, which

can be used to extract business and commercial value. [15] talks about what all advantage we can extract from the big data in the finance industry. Few examples are: Detection of false rumors that try to manipulate the finance market, Assessment of exposure to a reputational risk connected to consulting service offered by banks to their customer and Discover topic trends, detect events, or support the portfolio optimization or asset allocation. Big data based pattern recognition can also help in enhanced fraud detection systems and prevention capability systems. Other benefits of utilizing big data include building a machine learning based algorithm to achieve higher performance and accuracy in the trading algorithm and Enhanced market trading analysis. There has been proven research [12] which states more data increases accuracy and precision of simulations which is the backbone of financial modeling based analytics. This research [12] states modern modeling techniques are data hungry.In this project we will extract inference if more financial data can be used to have better prediction.

2 USE OF STRUCTURED FINANCIAL DATA

This reflects the data which has a higher degree of an organization such as a relational database where information/data is easily searchable and we can easily apply standard algorithm to extract patterns out of it. In this project we will be using Yahoo finance structured data. Examples of such data set include yahoo financial data, trading applications, enterprise finance resource planner, Retail banking systems, Credit history database systems and other financial applications that use legacy application systems. Structured data always has a big advantage of being easily entered, stored, queried and analyzed. Most of the personal banking financial statements are stored in a structured way. Structured dataset combined with the distributed systems can be leveraged to achieve structured big data set on which we can run optimized SQL queries to retrieve patterns. [9] discusses various SQL based ways to specify information quality in data which can be used to filter out the noise. In this project we will be using structured data.

3 VARIOUS CHALLENGES UTILIZING BIG DATA VALUE IN FINANCE INDUSTRY

There are multiple challenges and constraints in extracting value out of big financial data. The biggest challenge is old IT culture and infrastructure. The much financial organization still uses old IT infrastructure which is not compatible with the big data application thus fail to take advantage of big data. Other challenges include lack of skill set and data privacy and security. With the emergence of digitalization, customer data is saved persistently because of which there has been continued concern regarding the customer privacy. Regulatory bodies guidelines on customer data are always ill-defined because of which is there is always a concern regarding the use of customer data. In this project we will use standard python libraries to fetch financial data from yahoo finance. Analysis will be done on the jupyter notebook.

4 STOCK RETURNS PREDICTION - LITERATURE REVIEW

Authors of [1] discusse the importance of stock price and returns prediction based on the data extraction of historic data. This reserach [1] also shows historic financial data has definitive predtictive relationship to the future value of stocks. Stock prediction always help investors to decide perfect timing of buying or selling stocks. There are various data mining, artificial neural networks and machine learning techniques available for the stock price prediction based on the value extraction from the historic financial data. Based on the complexity of stock price matrix, pricing mechanism is essentially a non linear complex system. Authors of [14] and [13] state many predictictive algorithm is based on the fundamental analysis of macroeconomics and company fundamentals. [11] states problem with the fundamental analysis is that it is too much focused on the intrinsic and lacks the quantitative aspect of the historic financial data. On the broad category we can define stock prediction analysis is based on two types of analysis: qualitative and quantitative. Choice of analysis is mainly based on the fact if we want to have short term analysis or long term analysis. In this project we have have ten and sixteen years of training data and used close to one year of testing data. Since our analysis is based on the historic data we have choosen to do quantitative analysis. Quantitative anlysis is based on the pattern extraction, fact that history repeats and future financial drivers can be extracted based on the historic data. Advantage of using quantitative analysis is that we can use statistical confidence interval to validate the analysis.

There is a huge benefit of using machine learning algorithms in predicting stock prices. These algorithms made easy to cope up with the various financial events such as mergers acquisitions, bankruptcy, fraud, political changes, market crashes, housing bubble, dot net bubble and etc. In this project we have used hybrid approach of combined CAPM [Capital asset pricing] model and machine learning algorithm to mine data of sixteen and ten years of data and used close to 1 year of testing data. These machine learning algorithms can further be used to predict various financial events. Other approaches suchs as neural networks algorithms, SVM, logistic regression and multiple descriminant analysis can also be used to predict financial events. Example, [2] in their research they proved neural networks algorithms performs better in predicting financial events as compared to multiple descriminant analysis. There are other applications which use these algorithm to find predicted credit rating of a company. Credit rating plays a very important role doing qualitative analysis of the financial health of a company. On the other hand, accuracy of these algorithm is a big challenge because of the amount of huge data which it uses as input. Typically, accuracy of these algorithms is validated based on square root method.

In this project we have used several years of data for analysis which involves more than hundred thousands of rows with multiple columns. Then this data is analyzed two dimensionally with the same set of market return rows. Since this analysis is calculation intense, In the end we also have performed root mean square analysis.

Over the past few years there has been drastic changes in the way stock market operates. With the emergence of advance web services, there has been powerful enhancement in the data communication between various financial application. Because of which there is ocean of real time data is available, thus machine learning algorithm, neural networks algorithms, SVM, logistic regression and multiple descriminant analysis needs to be smarter. Forecasting stocks and financial parameter is of great interest to the investors. As discussed earlier these algorithms needs to modified depending on the fact if we want to have short term profit or long term profit.

[8] has shown the very interesting analysis of comparing the prediction of stock market with the random walk hypothesis. Author of [8] ran an experiment in which he tossed a coin and recordes the results and mapped head with the company profit and tail with the company loss. Then result of this experiment was shown to the investors pretending these are the actual market profit and loss. Looking at the result graphs, investors beleived it as a actual prediction. This research has shown the altogether different outlook which states stock price prediction and forecast can be fooled and stock prices are perfectly random in nature. On this theory many researchers have classified profit based on three hypothesis:

- Weak form Efficient Market Hypothesis: The weak form of the hypothesis states one can not generate profit by just looking at patterns and trends of stock market.
- Semi Strong Efficient Market Hypothesis: The semi strong form of the hypothesis states only possible way of generating profit is via inside trading.
- Strong form Efficient Market Hypothesis: The strong strong form of the hypothesis states its not possible to generate profit since stock market behaves in perfect random way.

However, if we are running root mean square analysis we can surely compare the accuracy of various algorithm and arrive at conclusion which algorithm is viable for prediction.

5 FINANCIAL DATA EXTRACTION

In this section, we will discuss various technical requirements needed to achieve value extraction from the big data in the finance industry. There are various technical requirements such as data Acquisition, data quality, data extraction, data integration, decision support. In order to fulfill requirements, a hybrid approach combining computer science, algorithms, statistics, data mining, machine learning and pattern recognition study needs to be adopted. To explore the advantage of big data there have been initiatives like data virtualization, multi-document summarization, pattern recognition from LOGS and many start-ups have been emerged. All big companies such as Microsoft, Google, IBM and Amazon are investing heavily in this field to leverage business and commercial value out of it. There has been changed in the industry pattern where financial industry is resorting big data to strategize their business. According to [6] with a very rapid pace, the financial industry is utilizing big data advantage in investment analysis, econometrics, risk assessment, fraud detection, trading, customer interaction analysis and behavior modeling. If we look at the Big promise the Big data holds in the finance industry, progress in this field is still in nascent stage and we expect more growth in upcoming years. In this project we will discuss jupyter notebook based solution for Data extraction.

In this project we have used jupyter notebook and rich python libraries to fetch financial stock data. Later in this paper we will discuss the stock data extration in detail. Later we will also discuss what are different ways to fetch stock data and will discuss few important functions which python libraries

6 FETCHING FINANCIAL STOCK DATA

Fetching structured precise data is always a challenge. There are different ways to fetch the stock market data. In this project we will be fetching data from yahoo finance via python libraries which internally makes remote web service call to the yahoo webserver. There are also other ways to fetch data such as:

- Direct download of csv files from yahoo finance or google websites.
- Make web api call to download the data in the json/XML format
- Use python libraries to download data, which internally makes remote web service call to the yahoo webserver. This is preferred way of doing since it allows you to save data to system variables directly.
- Call yahoo or finance web service from the application.
- Calling VBA function in excel to fetch yahoo stock data
- Quandl best for using core financial data and this website also includes access to rich python libraries.
- Google sheet has feature to fetch real time stock prices
- Install stocks macros in excel

In this project we have exhaustively used python for data manipulation. Reasons for using python are:

- Sytax is super easy which comes with very level of readability as compared to other programming languages.
- It is free and supports cross platform as python code can be called from any version of machine.

- Python has strong community support so if any problem is encountered, support is available online.
- Python has powerful tools available such as statsmodels, matplotlib, Pandas,Numpy and SciPyfor calculation intense projects

Since we have exhaustively used the `get_data_yahoo` function from the `pandas_datareader` python library we will briefly discuss the parameters it takes. Please note we utilized only those arguments which are relevant to the project requirements. From [10] parameter list as listed below:

- symbols : string, array-like object (list, tuple,Series), or DataFrame Single stock symbol (ticker), array-like object of symbols or DataFrame with index containing stock symbols.
- start : string, (defaults to '1/1/2010' Starting date, timestamp. Parses many different kind of date representations (e.g., 'JAN-01-2010', '1/1/10', 'Jan, 1, 1980')
- end : string, (defaults to today) Ending date, timestamp. Same format as starting date.
- `retry_count` : int, default 3 Number of times to retry query request.
- pause : int, default 0 Time, in seconds, to pause between consecutive queries of chunks. If single value given for symbol, represents the pause between retries.
- session : Session, default None requests.sessions.Session instance to be used
- `adjust_price`: bool, default False If True, adjusts all prices in hist data ('Open','High', 'Low','Close') based on 'Adj Close' price. Adds 'Adj Ratio' column and drops 'Adj Close'.
- `ret_index` : bool, default False If True, includes a simple return index 'Ret Index' in hist data.
- chunksize : int, default 25 Number of symbols to download consecutively before intiating pause.
- interval : string, default 'd' Time interval code, valid values are 'd' for daily, 'w' for weekly, 'm' for monthly and 'v' for dividend.

In our analysis, for the symbol parameter we are passing ticker symbol one at a time. Though, we have an option to pass multiple tickers as an array argument. We are using `get_data_yahoo` function and utilizing only first three paramters: symbols, start and end. This function returns YahooDailyReader object which can further be manipulated to get Open, High, Low, Close, Adj Close and Volume stock values. Since default number of retry count is three we will be using this default value. Default value of pause which is zero is also good with respect to our requirment so we will not pass this as argument. Session argument should be used when we are handling multiple request in parallel in the code since our project we just need one session so we will not use this argument. `adjust_price` is not required in our analysis since we are interested only in returns which can be fetched using `pct_change()` function. Since return index is of no use in calculating the returns, we will not use this argument. Argument chunksize is used to modify number of consecutive downloads of stocks since we are just using single ticker so this argument is of no use. This function uses interval also as a parameter since we are only intereseted in daily values and daily value is the default interval so we didn't pass this argument in the function call.We could also use the contemporary google function which is `get_data_google`. Arguments which goes to the `get_data_google` are symbols, start, end, `retry_count`, pause, chunksize and session since we are not using `get_data_google` function in our project we will not discuss these in detail.

7 INTRODUCTION TO CAPM MODEL

CAPM [Capital asset pricing model] model was developed by William Sharpe and John Lintner in 1964. This model is considered so powerful that it is being used in current prediction models. There are few advantages of using CAPM model as compared to other pricing models:

- This model is a single dimensional model and easy to use, still powerful to model capital asset pricing.
- Since this model is based on the market portfolio and risk free rate, this model removes unsystematic risk.
- We can run root mean square algorithm to validate the algorithm.
- This model provides a flexibility to utilize various risk free rates and run model for various time range.
- This model can be applied to various financial objects such as stocks, put option, call option, bonds, and etc

This model can be used to evaluate the theoretical expected return on a security, security can be any financial object such as stocks, put option, call option, bonds, and etc. In CAPM model we evaluate how much financial object is sensitive to the market using statistical analysis. Then this sensitivity which is also known as beta is used to find the expected return on security. This expected return can be on daily basis, weekly basis, monthly basis or yearly. Here is the formula to evaluate expected return:

$$E(R_i) = r_f + \beta_i(E(r_m) - r_f)$$

Where

- $E(R_i)$ is expected return
- r_f is risk free interest rate example: Government bond
- $E(r_m)$ is return on market example SP 500
- β_i is sensitivity of stock with respect to market

β_i can further be defined how much stock is sensitive to the stock market. Example if β_i for a particular stock is two it means if market goes up by five percent then stock will go up by ten percent and if market goes down by two percent then stock will go down by ten percent. In terms of statistics β_i is defined as:

$$\beta_i = \frac{Cov(R_i, r_m)}{Var(r_m)}$$

Where covariance and variance are defined as

- $cov_{x,y} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{N-1}$
- $var^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$

β_i matrix can be used to illustrate β_i in a following way:

$$\begin{array}{c} \\ Row1 \\ Row2 \\ Row3 \\ Row4 \end{array} \begin{array}{c} \begin{array}{ccc} \beta_i & MarketReturn & ExpectetReturn \end{array} \\ \begin{pmatrix} +2 & +5\% & +10\% \\ -2 & +5\% & -10\% \\ +0.5 & +4\% & +2\% \\ +0.5 & -4\% & -2\% \end{pmatrix} \end{array}$$

Above matrix suggests how expected returns can be correlated with the the β_i. Example if for certain company has β_i of +2 and market returns is +5 % then company's expected returns can be predicted as +10 %. Please note β_i can be positive as well as negative.

8 PROPOSED ANALYSIS

In this project we will utilize structured data and use CAPM [Capital asset pricing model] to statistically find the expected daily return of selected technological stocks: Amazon and Yahoo. This daily expected return can be used to predict next day stock value given the condition we have current stock price. Following formula can be used to predict next day stock price:

Next Day stock price =: Today stock price * (1+Daily expected return)

Daily expected return will be calculated using CAPM model. Daily expected return sensitivity in CAPM terminology is also known as beta. In this project beta will be calculated based on two time frames:

Time frame 1: [01/01/2000 to 12/31/2016] 16 years of data

Time frame 2: [01/01/2006 to 12/31/2016] 10 years of data

Thus we we will have 2 β_i:

$$\beta_1 = \frac{Cov(R_1, r_{m1})}{Var(r_{m1})}$$

$$\beta_2 = \frac{Cov(R_2, r_{m2})}{Var(r_{m2})}$$

Where

- β_1 is β based on time frame 1
- β_2 is β based on time frame 2
- R_1 is actual return based on time frame 1
- r_{m1} is a mean market return based on time frame 1
- R_2 is actual return based on time frame 2
- r_{m2} is a mean market return based on time frame 2

Above two time frames will be our training data set.We will run two analysis: one on training time frame 1 and other on training time frame 2 to arrive at predicted CAPM variables. Then we will use this training data set to predict stock returns for test data set which will comprise of time frame:

Test data time frame: 01/01/2017 to 11/16/2017

Then we will run the statistically analysis on the test data to evaluate if 16 years of training data produced more accurate result or else it added noise compared to 10 years of training data. Please note this is purely a quantitative analysis not qualitative. Actual returns can also be impacted by a qualitative factors such as mergers acquisitions, bankruptcy, fraud, political changes, market crashes, housing bubble, dot net bubble and etc.

9 PROPOSED ALGORITHM

Code is written purely in python language and used the powerful rich python libraries such as statsmodels, matplotlib, Pandas,Numpy and SciPyfor. We have used jupyter notebook as interpreter tool to python. Code is started by importing above mentioned rich python libraries. Since we are interested only in technological stocks: Amazon and yahoo we need to initialize they stock ticker with the python variable. In CAPM model we need to know the market return in order to know the stock sensitivity we will also initialize market ticker with SP 500 index. As discussed above we will be using `get_data_yahoo` function from the `pandas_datareader` and in this project we will be only utilizing only first three parameters which is stock ticker, start date and end date. For first iteration we will be using `get_data_yahoo` to fetch stocks and market returns for time frame 1. For having better understanding of how the data looks when fetched using `get_data_yahoo` function, we will have amazon financial data matrix calculated like:

$amazonData = dr.$`get_data_yahoo`$(amazon,$ `start_date`$,$ `end_date`$)$

Where

- dr is `pandas_datareader`.data class
- amazon is stock ticker for amazon which is 'AMZN '
- `start_date` is start date of time frame 1: 01/01/2000
- `end_date` is end date of time frame 1: 12/31/2016

and synopsis of above amazon data looks like:

$$\begin{array}{c} \\ 23-Dec-16 \\ 27-Dec-16 \\ 28-Dec-16 \\ 29-Dec-16 \\ 30-Dec-16 \end{array} \begin{array}{c} \begin{array}{cccccc} Open & High & Low & Close & Adj & Volume \end{array} \\ \begin{pmatrix} 764.54 & 766.50 & 757.98 & 760.59 & 760.59 & 1976900 \\ 763.40 & 774.65 & 761.20 & 771.40 & 771.40 & 2638700 \\ 776.25 & 780.00 & 770.50 & 772.13 & 772.13 & 3301000 \\ 772.40 & 773.40 & 760.84 & 765.15 & 765.15 & 3153500 \\ 766.46 & 767.40 & 748.28 & 749.86 & 749.86 & 4139400 \end{pmatrix} \end{array}$$

Similarly using `get_data_yahoo` we will fetch Yahoo and market returns. Since we are interested in daily return, we fetched the daily data from yahoo finance which is evident from the above result data set. Now lets find the percentage change on the daily Close value to get the percentage change array which in finance terminology will be daily return on stock. For finding the percentage change we are using `pct_change()` function on the close column of result set. This function can be elaborated as follows:

$$\texttt{return\_amazon} = amazonData.Close.pct_change()[1:]$$

$$\texttt{return\_yahoo} = yahooData.Close.pct_change()[1:]$$

$$\texttt{return\_market} = marketData.Close.pct_change()[1:]$$

`return_amazon`, `return_yahoo` and `return_market` are two dimensional arrays and we need to convert them to single dimensional array in order to run statistical analysis. We can use dot values method to extract single dimensional array out of 2 dimensional array. This operation can be elaborated as follows:

$$\texttt{X\_amazon\_actualReturns} = \texttt{return\_amazon\_testing}.values$$

$$\texttt{X2\_yahoo\_actualReturns} = \texttt{return\_yahoo\_testing}.values$$

$$\texttt{Y\_market\_actualReturns} = \texttt{return\_market\_testing}.values$$

Please note these are actual returns - fetched from yahoo finance. Now in order to evaluate expected return for the testing period based on the calculated beta we need to calculate the risk free rate r_f as mentioned above in the CAPM formula. Please note `get_data_yahoo` formula will fetch the annualized rate but here we are dealing with the daily returns so this needs to be normalized to daily rate. Here we are using Treas Yld Index-10 Yr Nts bond. Ticker symbol for Treas Yld Index-10 Yr Nts bond is `^TNX`. Please note `get_data_yahoo` will return columns: Open, High, Low, Close, Adj Close and Volume. Dot values will convert to 2 dimensional array and then used index [0][4] to fetch annual rate. Detailed code with comments is mentioned on jupyter notebook.

Converstion of annualized return to daily return can be done using following formula:

$$riskFreeDailyRate = (1+riskFreeAnnualRate)^{(1/365)}-1$$

Now we need to copy the content of `X_amazon_actualReturns` to new array `X_amazon_predictedReturns` and initialized each element in `X_amazon_predictedReturns` using CAPM model as discussed above in Introduction:

$$\texttt{X\_amazon\_predictedReturns} = list(\texttt{X\_amazon\_actualReturns})$$

We will do the same for Yahoo stocks:

$$\texttt{X2\_yahoo\_preddictedReturns} = list(\texttt{X2\_yahoo\_actualReturns})$$

In the code we have run the while loop and each element of `X_amazon_predictedReturns` and `X2_yahoo_preddictedReturns` is assigned the value based on CAPM model. Now we have two returns arrays for amazon stocks based on sixteen years of data:

- `X_amazon_actualReturns` are the actual returns
- `X_amazon_predictedReturns` are returns based on the CAPM model.

Similarly we have two returns arrays for yahoo stocks based on sixteen years of data:

- `X2_yahoo_actualReturns` are the actual returns
- `X2_yahoo_preddictedReturns` are the returns based on the CAPM model.

Now we can utilize `mean_squared_error` function from the sklearn.metrics python library to find how predicted returns are deviated from the actual returns. We will run this function on both stocks, amazon and yahoo:

$$a1 = \texttt{Y\_market\_actualReturns}$$

$$a2 = \texttt{X\_amazon\_predictedReturns}$$

$$y1 = \texttt{Y\_market\_actualReturns}$$

$$y2 = \texttt{X2\_yahoo\_preddictedReturns}$$

$$\texttt{rms\_amazon} = sqrt(\texttt{mean\_squared\_error}(a1, a2))$$

$$\texttt{rms\_yahoo} = sqrt(\texttt{mean\_squared\_error}(y1, y2))$$

Here is the root mean square values for both the stocks under sixteen years of data case:

- Root mean square error for Amazon stocks analysis based on 16 years of data 0.0013770 or 0.137 percent
- Root mean square error for Yahoo stocks analysis based on 16 years of data 0.0014313 or 0.143 percent

Now we run the same analysis as discussed above for the ten years of data and will validate how much predicted stocks returns based on the ten years of data are deviated from the actual returns using root mean square method. Please note testing data set remains the same we are just using different training data set. This will let us compare if sixteen years of data is of more worth in predicting stock returns or it added noise to the analysis:

- Root mean square error for Amazon stocks analysis based on 10 years of data 0.0005310 or 0.053 percent
- Root mean square error for Yahoo stocks analysis based on 10 years of data 0.0014910 or 0.149 percent

Above analysis is purely quantitative and does not include any elements of qualitative analysis. It shows predicting yahoo stock price or its returns based on the sixteen years of data or ten years of data - both resulted in almost same results. However things are totally different for the amazon stocks, recent ten years of amazon stocks data produced more accurate results as compared to using recent sixteen

years of data. Author of [7] agrees with the fact that most recent financial data are the better predictors of the future price returns. Though, in the [7] author has used the neural networks and support vector machine for prediction. Author also stressed that neural networks algorithm produced better accuracy than other machine learning algorithms.

10 THREE PARADIGMS OF PREDICTION

Data prediction and analysis done in this project is purely quantitative. However there are other paradigms of predictions also which we will discuss here. Example in above analysis we totally missed the qualitative aspect of the data. This is why it explains recent data on amazon stocks produced better results. Here are the other prediction paradigms explained by the author of [3] :

- Quantitative research based prediction. This is a method where we utilize statistical tools to arrive at predictictive value based on data
- Quantitative research based prediction. This is a method where we utilize conceptual knowledge to arrive at predicted value. Example of such study would be prediction of stocks based on the events such as mergers acquisitions, bankruptcy, Fraud, political changes, market crashes, housing bubble, dot net bubble and etc.
- Mixed research based prediction. This is a hybrid method where we utilize both qualitative and quantitative results to predict result.

In this project we used quantitative based approach to validate the fact if more data is good for prediction or it adds noise. Result of this project also showed there is a importance of recent data in predicting results. This is also validated by the research done under [7]

11 FUTURE WORK

In this project, analysis is based on two technological stocks: yahoo and amazon. We can extend our study to more diverse portfolio by including more stocks from various industries. Technological stocks tend to be more volatile than other stocks. Since this project is purely quantitative based prediction we deliberately choosen the technological stocks to leverage their volatility. More accurate prediction could also be made by encapsulating qualitative based prediction in the analysis which is more like a hybrid approach. Such hybrid approaches includes assigning weight to each predictions and taking cumulative result. As the part of future work we can also compare results across industry and arrive at conclusion which industry is more stable in prediction. Comparison can based on the root mean square analysis which is discussed in this project report.

12 CONCLUSION

Main objective of doing this project is to know the importance of big data in predicting financial variables. Analysis of this project is based on two stocks: amazon and yahoo. We started this project report with the discussion of importance of big data in financial industry. In the introduction we discussed how various industries are investing in the Big Data to attain higher standards in terms of quality and customer satisfaction. Then we discussed what are the various types of data available: structured and unstructured. Since in this project we utilized only structured data so it was discussed deeply. This project report also touch base with various challenges financial industry takes in utilizing the value of big data. As the part of literature review we reviewed various researches done in the field of stock returns prediction. In the financial data extraction section we reviewed various technical requirements need for financial data extraction. As the part of data analysis for this project we discussed what are the various ways to fetch live stock data from yahoo or google server. In this project we used the rich financial python libraries for the analysis so we discussed them in details in this report. Financial model which we chose for the prediction is the CAPM model which is explained theoretically in this report. There are two different section where we discussed the proposed analysis and proposed algorithm. We finally concluded the report by discussing three paradigms of prediction. In the end we also mentioned what further can be done under future work section.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and TAs for their support and suggestions to write this paper. TAs and professor are very good in terms of providing valuable guidance and suggestion in a very prompt fashion.

REFERENCES

[1] Qasem Al-Radaideh, Adel Abu Assaf, and Eman Alnagi. 2013. Predicting Stock Prices Using Data Mining Techniques. (12 2013). https://www.researchgate.net/publication/281865047_Predicting_Stock_Prices_Using_Data_Mining_Techniques

[2] A. F. Atiya. 2001. Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks* 12, 4 (Jul 2001), 929–935. https://doi.org/10.1109/72.935101

[3] Adam Chu. 2017. Quantitative, Qualitative, and Mixed Research. (2017). https://www.bcps.org/offices/lis/researchcourse/images/lec2.pdf

[4] Daniel D. Gutierrez. 2014. *Big Data for Finance.* Technical Report. Dell & Intel. https://whitepapers.em360tech.com/wp-content/files_mf/1427803213insideBIGDATAGuidetoBigDataforFinance.pdf

[5] Kazim Hussain and Elsa Prieto. 2015. *Big Data in Finance.* chapman and chapman and hall/crc, https://www.cs.helsinki.fi/u/jilu/paper/bigdataapplication04.pdf, Chapter 17, 329–356.

[6] Kazim Hussain and Elsa Prieto. 2016. *Big Data in the Finance and Insurance Sectors.* Springer, Cham, "https://link.springer.com/content/pdf/10.1007/", Chapter 12, 209–223.

[7] Hui Lin. 2014. Stanford. (2014). https://pdfs.semanticscholar.org/56f0/59ea400f31b60bfde4d59aea71bd7b411553.pdf

[8] Burton G. Malkiel. 2015. *A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing.* Recorded Books on Brilliance Audio. https://www.amazon.com/Random-Walk-Down-Wall-Street/dp/1501260375?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=1501260375

[9] A. Parssian, W. Yeoh, and M. S. Ee. 2015. Quality-Based SQL: Specifying Information Quality in Relational Database Queries.

Computer 48, 9 (Sept 2015), 69–74. https://doi.org/10.1109/MC.2015.264

[10] Kevin Sheppard. 2017. daily.py daily.py. (2017). https://github.com/pydata/pandas-datareader/blob/master/pandas_datareader/yahoo/daily.py

[11] Philip M. Tsang, Paul Kwok, S.O. Choy, Reggie Kwan, S.C. Ng, Jacky Mak, Jonathan Tsang, Kai Koong, and Tak-Lam Wong. 2007. Design and implementation of NN5 for Hong Kong stock price forecasting. *Engineering Applications of Artificial Intelligence* 20, 4 (2007), 453 – 461. https://doi.org/10.1016/j.engappai.2006.10.002

[12] Tjeerd van der Ploeg, Peter C. Austin, and Ewout W. Steyerberg. 2014. Modern modelling techniques are data hungry a simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology* 14, 1 (22 Dec 2014), 137. https://doi.org/10.1186/1471-2288-14-137

[13] Martin Walker and Mamoun Al-Debi'e. 2000. Fundamental Information Analysis: An Extension and UK Evidence. *ACS Biomaterials Science & Engineering* 31 (02 2000).

[14] Muh-Cherng Wu, Sheng-Yu Lin, and Chia-Hsin Lin. 2006. An effective application of decision tree to stock trading. *Science Direct* 31 (08 2006), 270–274.

[15] Sonja Zillner, Tilman Becker, and Munn. 2016. *Big Data-Driven Innovation in Industrial Sectors*. Springer International Publishing, Cham, Chapter 4, 169–178. https://doi.org/10.1007/978-3-319-21569-3_9

A HID 301:GAGAN ARORA

- Identified Project topic.
- Collected the python financial libraries.
- fetched data from yahoo finance
- Studied, designed and reviewed CAPM model
- Implemented CAPM model using python libraries
- Created project report

B CODE REFERENCE

All code, notebooks and files for this project can be found in the github repository: https://github.com/bigdata-i523/hid301/blob/master/project/finalProject.ipynb

C BIBTEX ISSUES

Warning–entry type for "Ref7" isn't style-file defined

–line 169 of file report.bib

Warning–entry type for "Ref8" isn't style-file defined

–line 187 of file report.bib

Warning–entry type for "Ref14" isn't style-file defined

–line 341 of file report.bib

Warning–entry type for "Ref15" isn't style-file defined

–line 359 of file report.bib

Warning–empty address in Ref13

Warning–page numbers missing in both pages and numpages fields in Ref10

(There were 6 warnings)

D ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

D.1 Assignment Submission Issues

DONE:

Do not make changes to your paper during grading, when your repository should be frozen.

D.2 Uncaught Bibliography Errors

DONE:

Missing bibliography file generated by JabRef

DONE:

Bibtex labels cannot have any spaces, _ or & in it

DONE:

Citations in text showing as [?]: this means either your report.bib is not up-to-date or there is a spelling error in the label of the item you want to cite, either in report.bib or in report.tex

D.3 Formatting

DONE:

Incorrect number of keywords or HID and i523 not included in the keywords

DONE:

Other formatting issues

D.4 Writing Errors

DONE:

Errors in title, e.g. capitalization

DONE:

Spelling errors

DONE:

Are you using *a* and *the* properly?

DONE:

Short form of verbs is for spoken language. Do not use them in scientific writing. example: can't is incorrect, use cannot

DONE:

Do not use phrases such as *shown in the Figure below.* Instead, use *as shown in Figure 3*, when referring to the 3rd figure

DONE:

Do not use the word *I* instead use *we* even if you are the sole author

DONE:

Do not use the phrase *In this paper/report we show* instead use *We show.* It is not important if this is a paper or a report and does not need to be mentioned

DONE:

If you want to say *and* do not use *&* but use the word *and*

DONE:

Use a space after . , :

DONE:

When using a section command, the section title is not written in all-caps as format does this for you

`\section{Introduction}` and NOT `\section{INTRODUCTION}`

D.5 Citation Issues and Plagiarism

DONE:

It is your responsibility to make sure no plagiarism occurs. The instructions and resources were given in the class

DONE:

Claims made without citations provided

DONE:

Need to paraphrase long quotations (whole sentences or longer)

DONE:

Need to quote directly cited material. Are you sure you have quoted all of them?

DONE:

The citation mark should not be in the beginning of the sentence or paragraph, but in the end, before the period mark. example: ... a library called Message Passing Interface(MPI) [7].

DONE:

Put a space between the citation mark and the previous word

D.6 Character Errors

DONE:

Erroneous use of quotation marks, i.e. use "quotes" , instead of " "

DONE:

To emphasize a word, use *emphasize* and not "quote"

DONE:

When using the characters & # % _ put a backslash before them so that they show up correctly

DONE:

Pasting and copying from the Web often results in non-ASCII characters to be used in your text, please remove them and replace accordingly. This is the case for quotes, dashes and all the other special characters.

DONE:

If you see a gure and not a figure in text you copied from a text that has the fi combined as a single character

D.7 Structural Issues

DONE:

Acknowledgement section missing

DONE:

Incorrect README file

DONE:

In case of a class and if you do a multi-author paper, you need to add an appendix describing who did what in the paper

DONE:

The paper has less than 2 pages of text, i.e. excluding images, tables and figures

DONE:

The paper has more than 6 pages of text, i.e. excluding images, tables and figures

DONE:

Do not artificially inate your paper if you are below the page limit

D.8 Details about the Figures and Tables

DONE:

Capitalization errors in referring to captions, e.g. Figure 1, Table 2

DONE:

Do use *label* and *ref* to automatically create figure numbers

DONE:

Wrong placement of figure caption. They should be on the bottom of the figure

DONE:

Wrong placement of table caption. They should be on the top of the table

DONE:

Images submitted incorrectly. They should be in native format, e.g. .graffle, .pptx, .png, .jpg

DONE:

Do not submit eps images. Instead, convert them to PDF

DONE:

The image files must be in a single directory named "images"

DONE:

In case there is a powerpoint in the submission, the image must be exported as PDF

DONE:

Make the figures large enough so we can read the details. If needed make the figure over two columns

DONE:

Do not worry about the figure placement if they are at a different location than you think. Figures are allowed to oat. For this class, you should place all figures at the end of the report.

DONE:

In case you copied a figure from another paper you need to ask for copyright permission. In case of a class paper, you must include a reference to the original in the caption

DONE:

Remove any figure that is not referred to explicitly in the text (As shown in Figure ..)

DONE:

Do not use textwidth as a parameter for includegraphics

DONE:

Figures should be reasonably sized and often you just need to add columnwidth

e.g.

```
/includegraphics[width=\columnwidth]{images/myimage.pdf}
```