HarvardX Professional Certificate in Data Science - Prediction of Amazon Stock

Ferdinand Pieterse

May 24, 2019

Contents

1 Introduction

This project is produced for the capstone for HarvardX Professional Certificate in Data Science. It a methodology to predict whether a particular stock will increase or decrease the following trading day.

In the ten years since the Global Financial Crisis in 2009 the United States stock market has recovered well. This recovery has, and continues to present traders and investors with opportunities to earn attractive returns if they are able to purchase stocks that are likely to increase.

One company that has frequently been in the headlines over the last few years is Amazon. In this paper an attempt is made to predict the stock of this company.

We will use the k Nearest Neighbours (kNN) machine learning algorithm for our prediction. The prediction system developed here makes the assumption that a stock's movement is related to other similar stocks. Amazon is in the Technology sector and we will use a number of well-known stocks in the Technology industry as features to try and predict whether the stock of Amazon will increase the following day.

1.1 Data Set

The historical price, dividend, and split data for most stocks can be obtained from Yahoo Finance (https://finance.yahoo.com/).

We will download historical data for our target stock Amazon, for the Dow Jones Industrial Average as well as for 10 other well known stocks in the Technology sector.

The following stocks are downloaded as individual data sets and then combined into data frames that can be used in our analysis.

Symbol	Name
DJI	Dow Jones
AMZN	Amazon
AAPL	Apple
CSCO	Cisco
GOOG	Google
HPQ	Hewlett Packard
IBM	IBM
INTC	Intel
MSFT	Microsoft
ORCL	Oracle
QCOM	Qualcom
TXN	Texas Intruments

1.2 Goal

The goal of this project is to create a stock forecasting system by developing a machine learning algorithm that predicts whether the price of a stock will increase or decrease the following day.

The algorithm used is k Nearest Neighbours or kNN. The success of the algorithm will be evaluated by calculating the accuracy of the prediction.

A prediction accuracy of 50% would be the same as randomly guessing whether the stock will increase. It is often extremely hard to predict the price of stocks. If we could make an improvement of 5% over random guessing, it will be deemed a successful model as this margin can make a difference given the amount of money at stake.

Therefore we will strive for a prediction accuracy in our validation (Testing) data set of at least 55%. The higher the prediction accuracy, the better the model.

1.3 Key Steps

- Download the individual stock datasets from Yahoo Finance
- Ensure all downloaded datasets has the same number of observations and no missing values
- Create a dataset for the Performance of the US Stock Market as a whole and for Amazon Stock
- Create a combined Dataset containing all 11 downloaded stocks (including Amazon Stock)
- Trim down the dataset to contain the past 10 years of data of Amazon and the 5 stocks most highly correlated to Amazon
- Split the dataset into:
  - Training - Nine years of data starting on 1 May 2009
  - Testing - One year of data starting on 1 May 2018
- Train the kNN model
- Determine the accuracy of the model

2 Analysis

2.1 Data Validation and Cleaning

Verify completeness of the data by making sure all downloaded data sets contain the same number of observations. Ensure none of the data sets contain missing values.

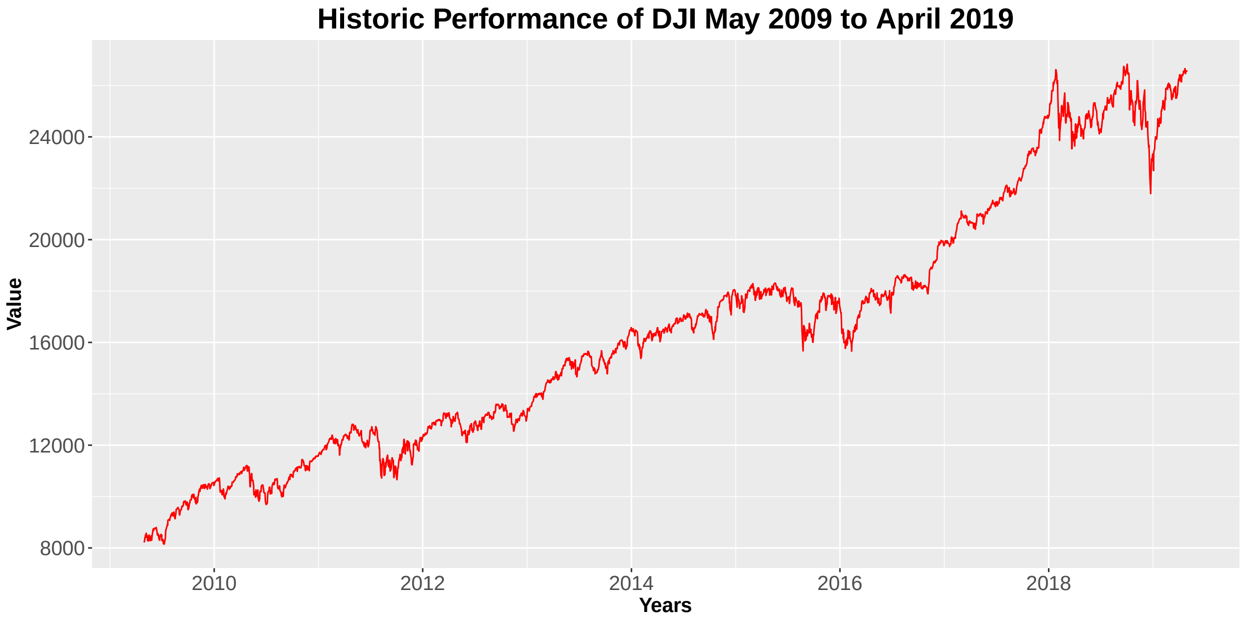
Name	Observations	Missing
Dow Jones	3119	0
Amazon	3119	0
Apple	3119	0
Cisco	3119	0
Google	3119	0
Hewlett Packard	3119	0
IBM	3119	0
Intel	3119	0
Microsoft	3119	0
Oracle	3119	0
Qualcom	3119	0
Texas Intruments	3119	0

2.2 Exploratory Data Analysis

Our analysis date range starts on 1 May 2009 and ends exactly 10 years later on 30 April 2019. All values are the "adjusted price" which means the closing daily stock price adjusted for dividends and stock splits.

The Dow Jones Industrial Average (DJI) is a widely accepted proxy for the performance of the entire United States stock market. During the time frame of our analysis the value of the DJI increased from `8212.41` on 1 May 2009 to `26592.91` on 30 April 2019, which represents a 10 year return of `224`% or `22.4`% per year.

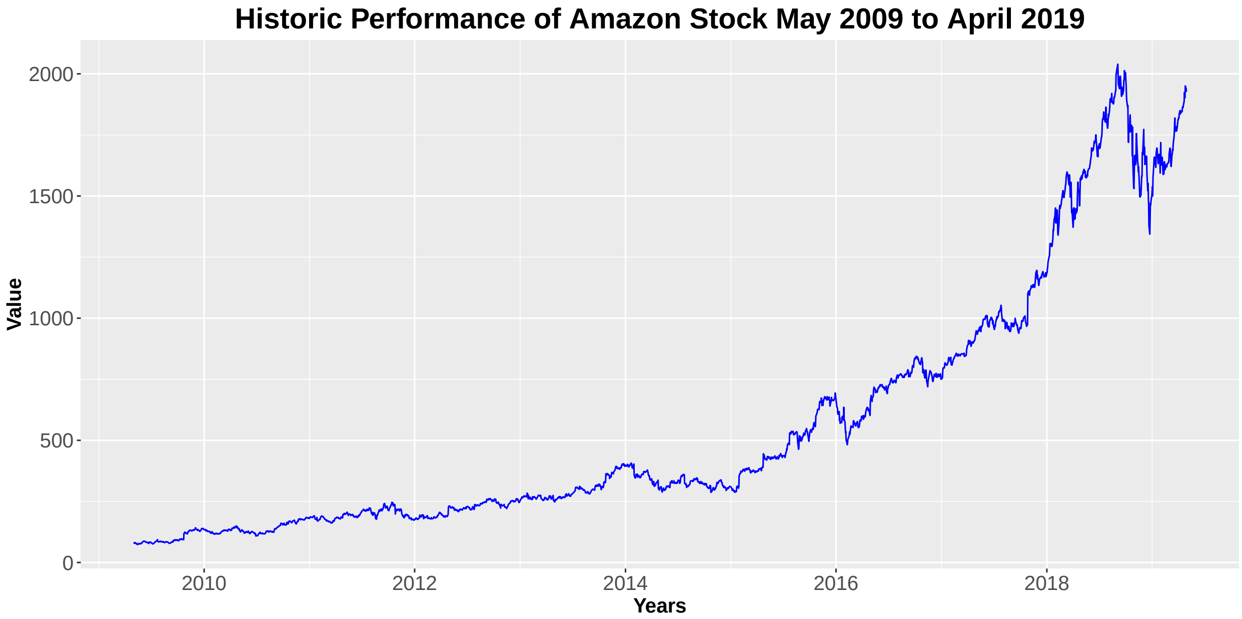
The value of the DJI over ten years were as follows:



Similarly the Stock of Amazon showed even more impressive returns: On 1 May 2009 the adjusted value of this stock was `$ 78.96` and it had moved to an adjusted value of `$ 1926.52` on 30 April 2019.

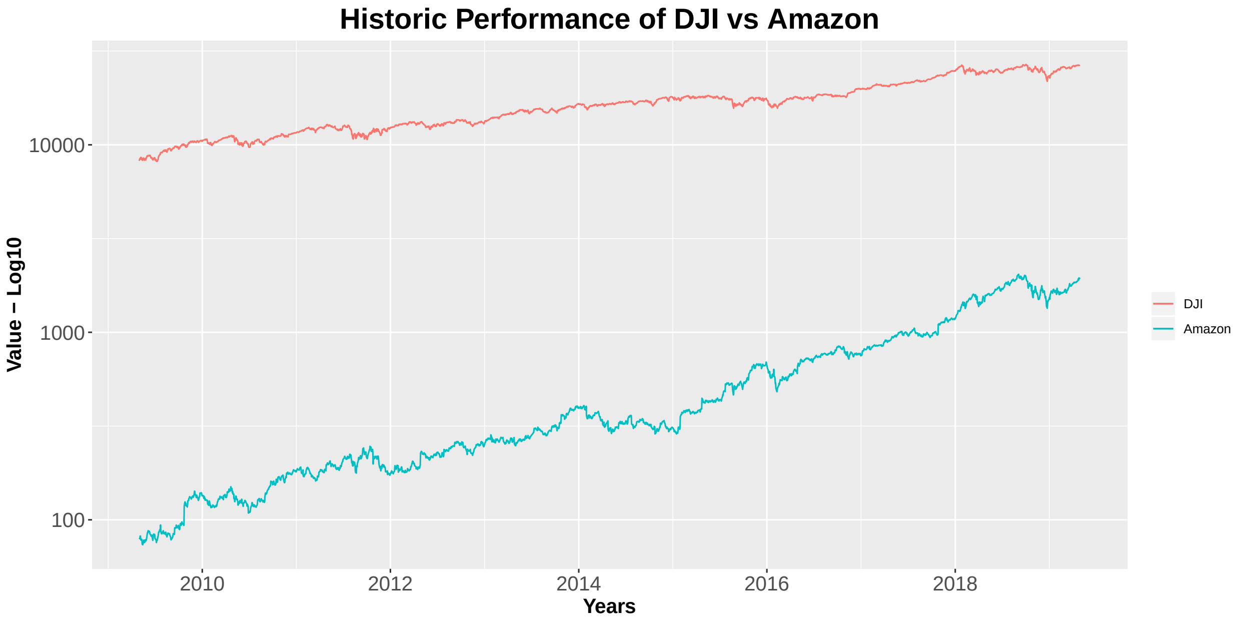
This equates to a 10 year return of `2340`% or `234`% per year

The performance chart of Amazon over ten years were as follows:



When the performance charts of the Dow Jones are combined with Amazon it is clear how impressive the performance of Amazon was.

Due to the fact that the value of the Dow Jones is much higher than that of Amazon we perform a log transformation to enable a more meaningful comparison.



2.3 Modelling Approach

2.3.1 Explanation of the k Nearest Neighbours (kNN) algorithm

The kNN algorithm is a non-parametric algorithm that can be used for either classification or regression. Non-parametric means that it makes no assumptions about the underlying data distribution.

For each data point, the algorithm finds the k closest observations (neighbours), and then classifies the data point to the majority. Usually, the k closest observations are defined as the ones with the smallest Euclidean distance to the data point under consideration.

Euclidean distance is the shortest distance between two points in a plane. By using this formula, you can calculate distance between two points no matter how many features you are given. The formula to calculate it is:

$$\sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

Where n is the number of features

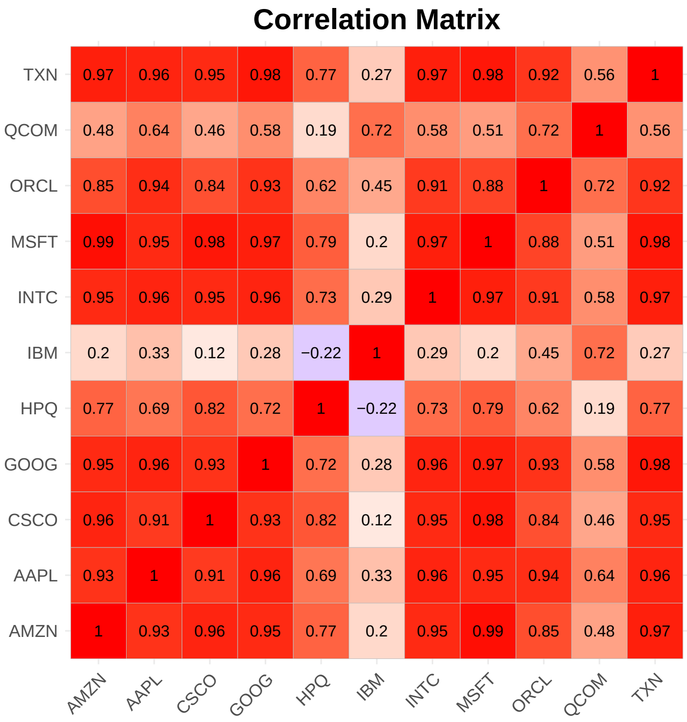
2.3.2 Create outcome field

An additional field is created for Amazon, indicating “TRUE” if the stock rose from the prior day to the current day and indicating “FALSE” if not. This column is the outcome which we wish to predict.

2.3.3 Determine which features to use

k-Nearest Neighbors is computationally heavy, particularly in high dimensionality, so to save processing resources and create a more robust model, only the 5 downloaded stocks that are the most highly correlated to Amazon are identified and used with the Amazon stock price as features to predict the outcome.

Correlation Matrix



	AMZN	AAPL	CSCO	GOOG	HPQ	IBM	INTC	MSFT	ORCL	QCOM	TXN
TXN	0.97	0.96	0.95	0.98	0.77	0.27	0.97	0.98	0.92	0.56	1
QCOM	0.48	0.64	0.46	0.58	0.19	0.72	0.58	0.51	0.72	1	0.56
ORCL	0.85	0.94	0.84	0.93	0.62	0.45	0.91	0.88	1	0.72	0.92
MSFT	0.99	0.95	0.98	0.97	0.79	0.2	0.97	1	0.88	0.51	0.98
INTC	0.95	0.96	0.95	0.96	0.73	0.29	1	0.97	0.91	0.58	0.97
IBM	0.2	0.33	0.12	0.28	−0.22	1	0.29	0.2	0.45	0.72	0.27
HPQ	0.77	0.69	0.82	0.72	1	−0.22	0.73	0.79	0.62	0.19	0.77
GOOG	0.95	0.96	0.93	1	0.72	0.28	0.96	0.97	0.93	0.58	0.98
CSCO	0.96	0.91	1	0.93	0.82	0.12	0.95	0.98	0.84	0.46	0.95
AAPL	0.93	1	0.91	0.96	0.69	0.33	0.96	0.95	0.94	0.64	0.96
AMZN	1	0.93	0.96	0.95	0.77	0.2	0.95	0.99	0.85	0.48	0.97

As can be seen from the correlation matrix above, the 5 stocks that has the highest correlation and that will therefore be used with the Amazon stock price to predict the outcome is:

Symbol	Name
CSCO	Cisco
GOOG	Google
INTC	Intel
MSFT	Microsoft
TXN	Texas Intruments

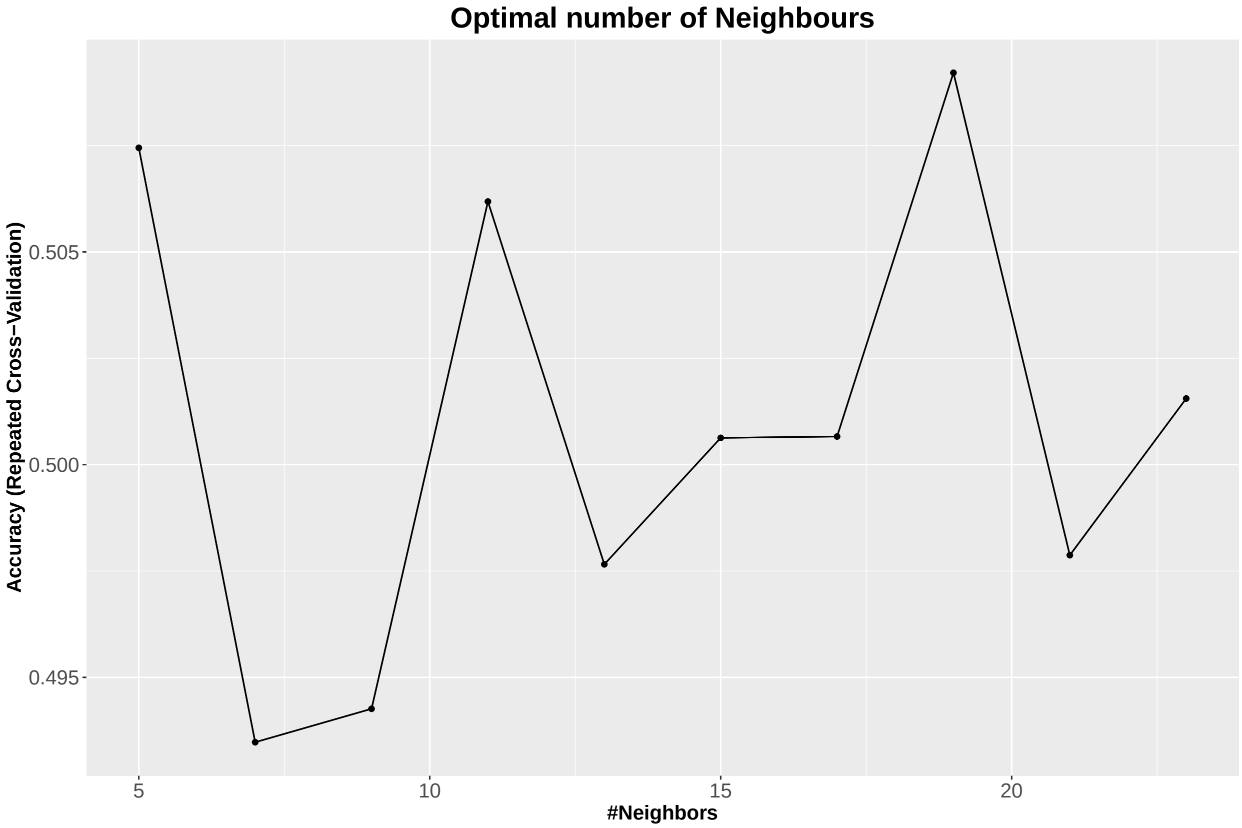
2.3.4 Split data into Training and Testing data sets

We will use the first nine years i.e. 1 May 2009 to 30 April 2018 for the Training data, and the last year i.e. 1 May 2018 to 30 April 2019 for Testing data.

2.4 Train the kNN model

The model is trained with the knn algorithm implementation of the caret package.

The caret package automatically selects the number of neighbours that delivers the highest accuracy of the training set.



From the chart above it is clear the optimal number of neighbours to use is `19`.

3 Findings / Results

The optimal neigbours equals an accuracy in the training set of `50.9`%, which does not seem to be very useful as it is just barely better than randomly guessing. However when we run the model against the validation set it produces the following confusion matrix and statistics:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE    13    6
##      TRUE     96  136
##
##                Accuracy : 0.5936
##                  95% CI : (0.5301, 0.655)
##     No Information Rate : 0.5657
##     P-Value [Acc > NIR] : 0.2042
##
##                   Kappa : 0.0852
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.11927
##             Specificity : 0.95775
##          Pos Pred Value : 0.68421
##          Neg Pred Value : 0.58621
##              Prevalence : 0.43426
##          Detection Rate : 0.05179
##    Detection Prevalence : 0.07570
##       Balanced Accuracy : 0.53851
##
##        'Positive' Class : FALSE
##
```

The high number of predicted increases (Prediction is TRUE) when the stock actually retreated (Reference is FALSE), can most likely be ascribed to the fact that we are in a bull (upward trending) market. However from the statistics above, it is clear the overall accuracy comes to `59.4`%.

4 Conclusion

Since we have achieved an accuracy of `59.4`% in our Testing set, which beats our stated goal of 55% by `4.4`%, the model is usable and should be able to provide an edge in trading the stock of Amazon.