Project Proposal

Feng Wan & YiNing Wang

2024-03-04

Dow Jones Industrial Average Stock Market Price & Top 25 Daily News Analysis

We were inspired by Pragadeesh Suresh Babu with his work on predicting stock rise with news (Can be found at https://github.com/pragadeeshsureshbabu/Predicting-stock-rise-with-news). We want to try using different machine learning algorithms to also investigate the relationship between news and stock market change.

Dataset

DJIA Dataset

The DJIA dataset comes from the daily stock market price change of Yahoo Finance. The data was retrieved using the Yahoo Finance API. The data was collected by summing up the market value of the Dow Jones Industrial Market for each day. The date range of the dataset was from 2016/01/04 to 2024/12/30. There are six variables in the dataset, which are **Date**, **Volume** (The volume of the market), **Open** (The market value when the market opens), **Close** (The market value when the market close), **High** (The highest market value of the day), and **Low** (The lowest market value of the day).

```
# A tibble: 5 x 6
  Price      Close            High             Low              Open        Volume
  <chr>      <chr>            <chr>            <chr>            <chr>       <chr>
1 2016-01-04 17148.939453125  17405.48046875   16957.630859375 17405.4804~ 14806~
2 2016-01-05 17158.66015625   17195.83984375   17038.609375     17147.5     10575~
3 2016-01-06 16906.509765625  17154.830078125  16817.619140625 17154.8300~ 12025~
4 2016-01-07 16514.099609375  16888.359375     16463.630859375 16888.3593~ 17624~
5 2016-01-08 16346.4501953125 16651.890625     16314.5703125   16519.1699~ 14185~
```

News Dataset

The news dataset is retrieved from the worldnews subreddit (https://www.reddit.com/r/worldnews/about/) on reddit using the Reddit API. The top 25 most commented news are being downloaded for each date from 2016/01/01 to 2024/1/30. The dataset contains variables of **date**, **title** (news content), **score** (heat score on Reddit), **num_comments** (number of comments), **url** (the link to the discussion).

```
# A tibble: 5 x 5
  date       title                                          score num_comments url
  <date>     <chr>                                          <dbl>        <dbl> <chr>
1 2016-01-01 EU to help Ukraine replace Musk's Starlink 44338         1224 http~
2 2016-01-01 US Senator Mike Lee, Elon Musk calls for ~ 28655         3595 http~
3 2016-01-01 Europe to form 'coalition of the willing'~ 27960         1175 http~
4 2016-01-01 Norway rethinks €1.7 trillion sovereign f~ 26724         1557 http~
5 2016-01-01 Canada PM Trudeau begins talks with King ~ 20917         1845 http~
```

Combined Dataset

We then combine the two datasets together by the dates that present in both datasets. Each date would be represented by each row. The variables in the dataset are **date**, **volume** (The volume of the market), **open** (The market value when the market opens), **close** (The market value when the market close), **high** (The highest market value of the day), and **low** (The lowest market value of the day) and **title 1-25** (representing the top 25 news on a given date, ranks matter). The date range from 2016/01/04 to .2024/12/30.

```
# A tibble: 5 x 31
  date        close   high    low   open    volume `title 1` `title 2` `title 3`
  <date>      <dbl>  <dbl>  <dbl>  <dbl>     <dbl> <chr>     <chr>     <chr>
1 2016-01-04 17149. 17405. 16958. 17405. 148060000 EU to he~ US Senat~ Europe t~
2 2016-01-05 17159. 17196. 17039. 17148. 105750000 EU to he~ US Senat~ Europe t~
3 2016-01-06 16907. 17155. 16818. 17155. 120250000 EU to he~ US Senat~ Europe t~
4 2016-01-07 16514. 16888. 16464. 16888. 176240000 EU to he~ US Senat~ Europe t~
5 2016-01-08 16346. 16652. 16315. 16519. 141850000 EU to he~ US Senat~ Europe t~
# i 22 more variables: `title 4` <chr>, `title 5` <chr>, `title 6` <chr>,
#   `title 7` <chr>, `title 8` <chr>, `title 9` <chr>, `title 10` <chr>,
#   `title 11` <chr>, `title 12` <chr>, `title 13` <chr>, `title 14` <chr>,
#   `title 15` <chr>, `title 16` <chr>, `title 17` <chr>, `title 18` <chr>,
#   `title 19` <chr>, `title 20` <chr>, `title 21` <chr>, `title 22` <chr>,
#   `title 23` <chr>, `title 24` <chr>, `title 25` <chr>
```

Research Question

Is the sentiment of the News influencing the Dow Jones Industrial Average stock market price for a given day?

With strong interest in stock markets, we are interested whether the public hype, represented by the news, would influence the stock market. Earning in stock is all about predicting others' mindsets and it seems intuitive that people and their stock market decisions are influenced by the News. With this in mind, we want to investigate whether there is actually a relationship between news and stock market.

We put our focus on one of the biggest stock markets and the sentiments of the top 25 news on one of the largest discussion forums - Reddit, and analyze whether the sentiment of the news is affecting the average stock market price of the Dow Jones Market. We expect the relationship to be there based on our existing economic knowledge.

German Credit And Risk Analysis

Dataset

The dataset classifies a set of German people described by a set of attributes as good or bad credit risks. The dataset was found on the UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data) and was imported in python and saved locally. Each row represent a person's **attributes** and the labeled **credit classification** (binary variable with good credit score being 1 and bad credit score being 2). There are 20 attributes serving as independent variables for the model to use to predict the targets. There is a detailed code book in our github repository. The dataset was collected by Professor Hans Hofmann.

```
# A tibble: 5 x 21
  Attribute1 Attribute2 Attribute3 Attribute4 Attribute5 Attribute6 Attribute7
  <chr>           <dbl> <chr>      <chr>           <dbl> <chr>      <chr>
1 A11                 6 A34        A43              1169 A65        A75
2 A12                48 A32        A43              5951 A61        A73
3 A14                12 A34        A46              2096 A61        A74
4 A11                42 A32        A42              7882 A61        A74
5 A11                24 A33        A40              4870 A61        A73
# i 14 more variables: Attribute8 <dbl>, Attribute9 <chr>, Attribute10 <chr>,
#   Attribute11 <dbl>, Attribute12 <chr>, Attribute13 <dbl>, Attribute14 <chr>,
#   Attribute15 <chr>, Attribute16 <dbl>, Attribute17 <chr>, Attribute18 <dbl>,
#   Attribute19 <chr>, Attribute20 <chr>, target <dbl>
```

Research Question

How to determine good or bad credit classification based on different attributes?

The research question comes from our interest in credit and risk assessment. Nowadays, risk assessments has became extremely important for banks and financial institutes in deciding whether to loan debts to an individual or a commercial using mathematical models. Many of the models are deeply founded in data science and machine learning algorithms. With the motive to try building our own risk assessment model, we would wanted to use different machine learning algorithms to try to predict the credit label of an individual with different associated attributes.

Therefore, we found the German Credit Dataset, and looking forward to implement binary classification algorithms like XGBoost, Random Forest, or K-Neighbors and hopefully reach an acceptable performance.

Citations:

Hofmann, H. (1994). Statlog (German Credit Data) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5NC77.