# Using Machine Learning To Predict Stock Performance

Minchan Shi

12/6/2020

# What I Will Be Talking About

Introduction

Data Description and Diagnostics

Data Analysis and Main Results

Conclusion and Future Research

# Problem Statement and Motivation

**Investing involves risk, including possible loss of principal. An investor should consider investment objectives, risks, charges and expenses of the investment company carefully before investing.**

- Predicting stock price trends
- Popular methods-machine learning techniques
- Using supervised and unsupervised learning method forecasting stock price(GDX)

GDX®

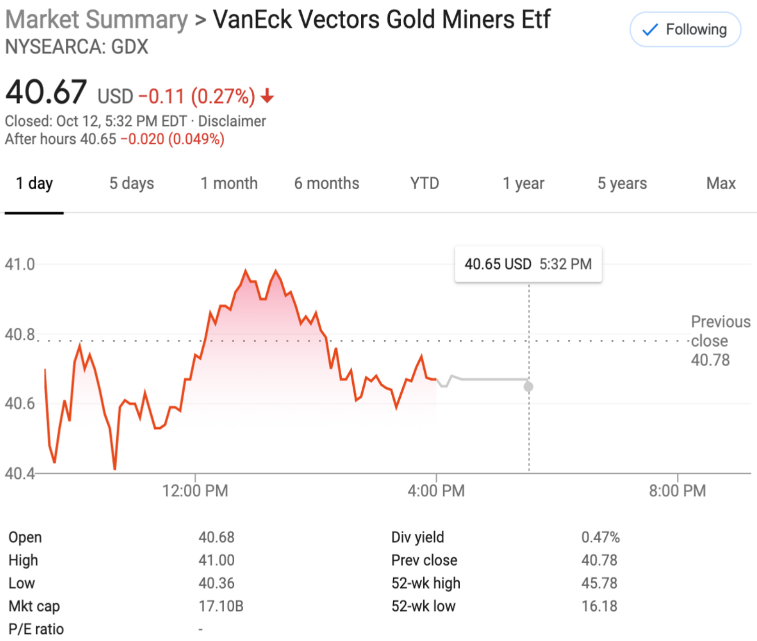
VanEck Vectors® Gold Miners ETF

# Data Description

The dataset is from yahoo finance historical database.<"https://finance.yahoo.com/quote/GDX/performance?p=GDX

VanEck Vectors Gold Miners ETF (GDX)

- ❖ 1259 sample size
- ❖ "Date" , "Open" , "High" , "Low" , "Close" , "Adj.Close", "Volume" . "Nextday", "Profit_In_Percentage", "Month" ,"Year" .



Market Summary > VanEck Vectors Gold Miners Etf
NYSEARCA: GDX

40.67 USD -0.11 (0.27%)
Closed: Oct 12, 5:32 PM EDT · Disclaimer
After hours 40.65 -0.020 (0.049%)

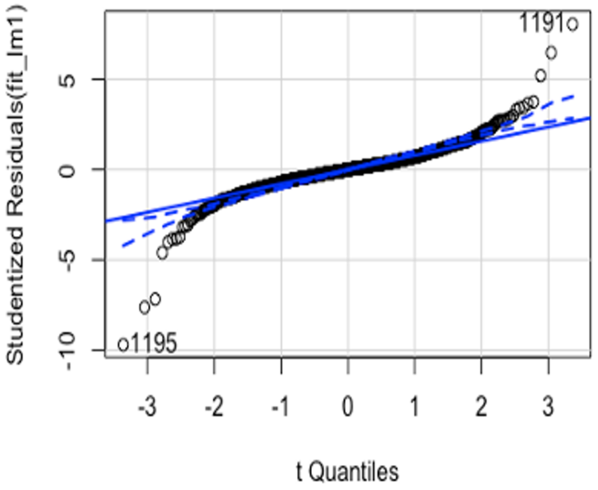
| | | | |
|---|---|---|---|
| Open | 40.68 | Div yield | 0.47% |
| High | 41.00 | Prev close | 40.78 |
| Low | 40.36 | 52-wk high | 45.78 |
| Mkt cap | 17.10B | 52-wk low | 16.18 |
| P/E ratio | - | | |

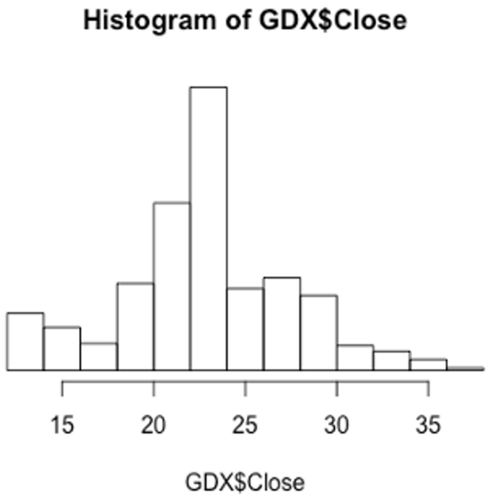
# Data Diagnostics

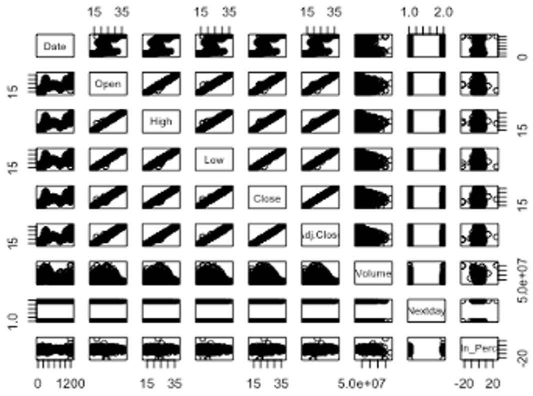
- *No empty values*
- *Outlier appears*

- *Close to normal distribution*

- *High correlation*







# Data Analysis and Main Results

The question I want to analyze is: what is the closing price of the stock the next day? Short-term predictions of whether it will go “up” or “down” tomorrow.

Long-term forecasting to understand what time should buy and sell.

Using data visualization to check and understand data central trends and outliers.

PCA,LDA,QDA was performed to understand that account for most of the variation in the variables and to predict whether the stock would go ‘up’ or ‘down’ for the next day, as well as to predict what the stock will performance in the future for long-term.

# Data Visualization

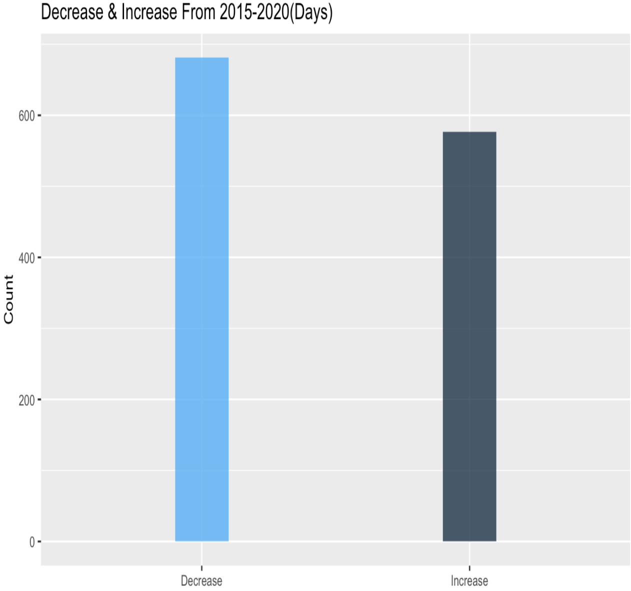


Figure 1

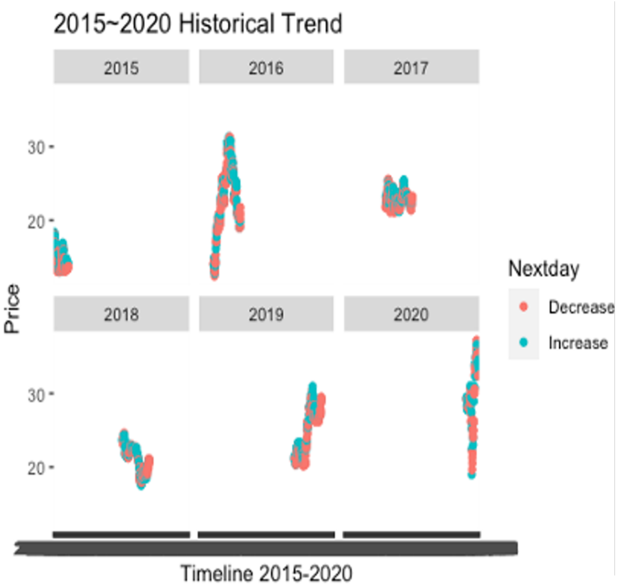


Figure 2

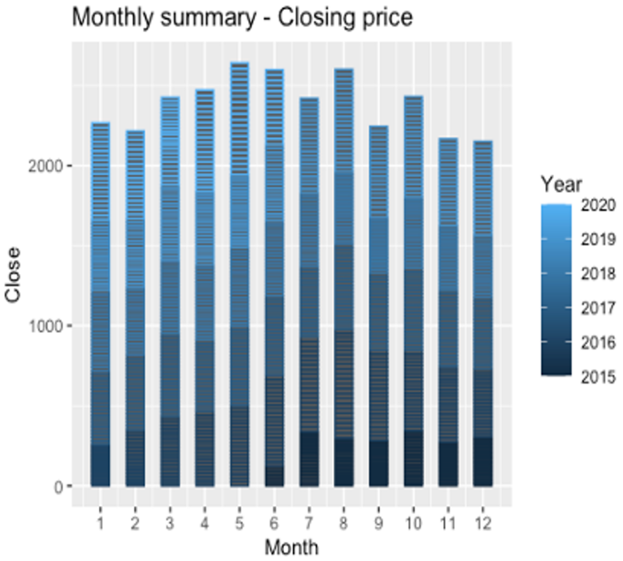


Figure 3

# Generalized Linear Model

- From Generalized Linear Model we can see that the closing price of a stock can be predicted by establishing a regression model, however, the real value is far away from the model result. Around 25% error rate in the regression model compare with the real closing price. I used the model generated on the training set to predict outcomes for our validation set and calculated the test error rate, or the percentage of time the model misclassified an outcome as compared to the observed results. In general, linear regression analysis does not seem to be suitable for predicting stock prices.
- I use logistic regression to estimate that the model correctly predicted that the market opening price order to go up on 13 days and that it order to go down on 669 days, for a total of 669 + 13 = 682 correct predictions. Among them, the error rate reaches 1-0.5421304=0.4578696, which means the error rate of logistic regression is about 46%. According to the observation, we can know that the Generalized Linear Model is not very accurate in predicting stocks.

# Linear Discriminant Analysis

- Now we will perform LDA and QDA for classification analysis. Here again I used split dataset into training and validation sets and fit LDA and QDA model to training set. Linear Discriminant Analysis corrected predict 666 days market went done and 11 days market go up. The LDA output indicates that ˆπ1 = 0.5488 and ˆπ2 = 0.4512; In other words, almost 55% of the training observations correspond to days during which the market went down, nearly 45% error rate in the model,the prediction accuracy is 55% .

# Quadratic Discriminant Analysis

- Quadratic Discriminant Analysis corrected predict 548 days market went done and 127 days market go up. The QDA output indicates that ˆπ1 = 0.5488 and ˆπ2 = 0.4512; ˆπ1 and ˆπ2 outPut are same as LDA. Also nearly 54% of the training observations correspond to days during which the market went down, nearly 46% error rate in the model,the prediction accuracy is 54% .

## After compare three model, We can see the LDA,QDA and logistic regression predictions are almost identical.

- LDA - Prediction Accuracy is 55% .
- QDA - Prediction Accuracy is 54% .
- Logistic Regression - Prediction Accuracy is 54% .

# Principal Component Analysis

- The rotation matrix provides the principal component loadings; each column of pr.out$rotation contains the corresponding principal component loading vector, we see that there are six distinct principal components. The standard deviations for six distinct principal components are : 2.25, 0.96, 0.08, 0.04, 0.04, 0.02. After I compute the proportion of variance explained by each principal component, it indicates that the first principal component explains 85% of the variable in the data, and the next principal component explains 15% of the variance, and the rest of other principal component explains almost 0% of the variable. Meanwhile, we know that first two component are contains most of the variable in the data.

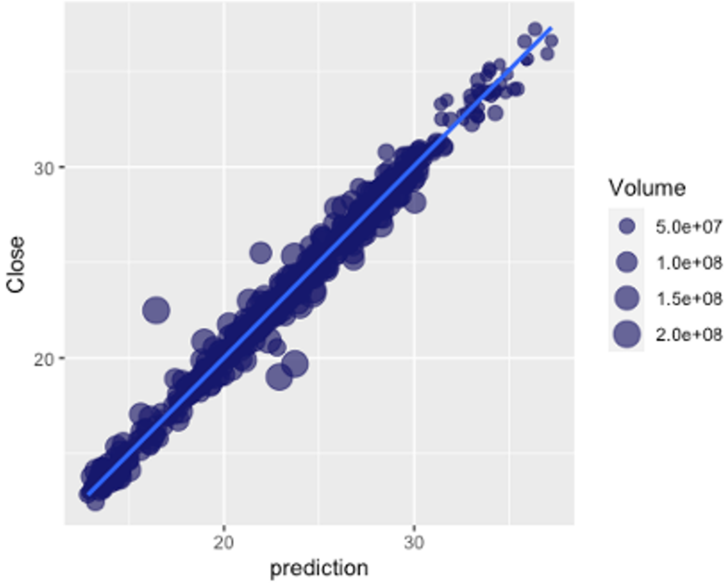
# Conclusion and Future Research

## Insight gain

- 2015 - 2020 shows an average daily rise and fall of -0.9% to 0.97%
- Short-term investment in this stock is not an ideal strategy
- November and December are low buying times

## Future Research

- Other creative and effective methods
- Model can be modified
- Study the short selling mechanism



# Video Link for the presentation

https://uri.techsmithrelay.com/YTFc

# Thank you!