"CS210-Data Science Project Phase#3"

2023-2024 Spring

Osman Şah Yılmaz

Introduction

This project explores the correlation between social media signals and stock market indicators, focusing specifically on the weight of tweets containing the word "GPT" and its relation to Nvidia's stock market performance. By analyzing daily data, this study aims to uncover the interplay between online sentiment and financial metrics, providing insights into how digital discourse can potentially influence stock volatility and trading patterns.

Datasets Employed

The analysis utilizes two main datasets sourced from Kaggle:

- US Stock Market Data (2019-2024): Contains daily trading information for Nvidia, including price and volume.
- Tweets Containing "GPT": Comprises a collection of tweets that include the word "GPT," used to gauge social media sentiment and its potential impact on the stock market.

Exploratory Data Analysis

The exploratory data analysis is set up with essential Python libraries for data manipulation and visualization (Pandas, Numpy, Seaborn, Matplotlib). The environment is prepared using Google Colab and Google Drive to access and store datasets efficiently.

Key Operations:

- Data Loading: The Nvidia stock dataset is loaded into a DataFrame to inspect the initial and final entries, providing a snapshot of the dataset's structure and data points.
- Initial Observations: Displays the head and tail of the dataset, which includes dates, stock prices, and trading volumes, showcasing the data's granularity and range.
- Variable Transformation: Since the followers weight of accounts mentioned “GPT” is important, followers of accounts grouped and made new variable. Also taken square root of some variables since it will be hard to visualize and understand data.

Statistical Analysis

Hypothesis Testing:

The hypothesis posits that an increase in the weight of tweets containing "GPT" correlates with a rise in Nvidia's stock volume.

Analysis Techniques:

Pearson Correlation Test: Conducted to measure the strength and direction of the relationship between tweet weight and stock volume. The results indicate a correlation coefficient of 0.41, suggesting a moderate relationship. The p-value (<0.05) implies statistical significance, allowing the rejection of the null hypothesis.

Linear Regression Model

A linear regression model is developed to quantitatively assess the relationship between the variables.

- Scatter Plot: Visualizes the relationship between the square root transformations of tweet weight and trading volume, highlighting data dispersion and trends.
- Model Fitting: The regression model is fitted with the transformed data, and results are summarized to describe the line of best fit.
- Regression Line Plot: The plot displays the regression line with the observed data points, illustrating the model's fit and predicting the response variable based on the predictor.

Model Evaluation:

- The model's coefficients (slope and intercept) are presented, along with the R^2 value, which quantifies the variance explained by the model, supporting the analysis with quantitative evidence of the relationship strength.

Machine Learning Models

k-Nearest Neighbors (kNN) Regressor:

Performance: kNN's performance can vary significantly with the choice of the number of neighbors (k). In this case, using 1 neighbor (k=1) essentially means the model is using the nearest single data point to make predictions, which can lead to overfitting, especially with noisy data.

Features: kNN tends to perform well when the feature space is well-distributed and not overly complex. With just one feature (SqrtTweetWeight), the model's simplicity is maximized, but it might struggle with more complex relationships in the data.

Decision Tree Regressor:

Performance: Decision Trees are capable of capturing non-linear relationships by splitting the data based on feature values. The identical MSE value indicates that, for this dataset, the complexity of the model isn't providing an advantage over the simple nearest neighbor approach of kNN.

Features: Decision Trees can handle both numerical and categorical data and perform feature selection implicitly. However, with only one feature, the model's capacity to split the data and create a meaningful structure is limited, leading to similar performance to the kNN with k=1.

Features and ML Techniques:

Feature Importance: In this case, SqrtTweetWeight is the only feature available. To enhance the performance of both models, additional features that could provide more information about SqrtTradeVol (such as other relevant market indicators or more derived features from tweet data) would likely improve model accuracy.

Results

kNN Mean Squared Error: 1,599,861.31

Decision Tree Mean Squared Error: 1,599,861.31

Both models have identical MSE values, indicating that they perform equally in terms of prediction accuracy for this specific dataset and feature set.