Stocks' Pricing

Fadhel ALeid

Abstract

The aim of this project is to use prediction modeling to figure out the best variables impacting the stocks' prices for the companies listed in Saudi Trade market (Tadawul). The data has been collected from Yahoo finance using yfinance library and stored in Google drive as listed below:

1- Four years trade data

2- Balance sheet trade data

3- Cashflow trade data

In this project the data has been merged and cleaned in order to be utilized in the prediction models using python. The output of this project has been communicated via a powerpoint presentation and python script loaded to github.

Design

This project has been done to demonstrate the skills gained in Data Science Bootcamps by SADIA Academy. By using the collected data from yahoo finance for the listed companies in Tadawul, it is aimed to find what are the top variables that could be used to predict the closing price at the end of each day. The data has been merged into one data frame, cleaned and standardized then utilized in 3 different prediction models.

Data

The collected data covers 3 areas (day trades, balance sheet and cash flow statements) for 4 years for 200 companies. This includes around 50 variables and around 200,000 rows. As some of the companies are not four years old, their data were excluded from the data used in this project. Using the day trade data, new features were introduced to smooth velocity of the model predictions.

Algorithms

1. Colab has been used to write the code to implement this project to utilize ready to use environments as well as its accessibility from different channels.
2. The following libraries have been imported in order to use them in the project
    - **pandas, numpy, seaborn, matplotlib, sklearn**
3. The project is mounted to google drive in order to read the data from google drive
4. The data has been read using panada's read_csv function
5. The data has been explored to check the distribution among the companies' data
6. Companies data that does not met the minimum requirements has been excluded (4 years data for the 3 areas)
7. Columns names have been lowered and data types have been converted to the proper format as part of the cleaning phase
8. Additionally columns with no data or rare updated ones have been removed as well in the cleaning phase.
9. New features such as (5 days moving average prices) have been calculated using the existing one to smooth the velocity of the model.
10. Prior merging the data, a forign key has been created for each row in each data frame using the following logic:
    - **Day trade row forign key equals the concatenation of the ticker of the company , hyphen and the year of trade day**
    - **Balance sheet and cash flow forign key equals the concatenation of the ticker of the company , hyphen and the year of the financial data +1**
11. The data has been merged using the defined forign keys above using inner join to make sure the merged data have the required from all the areas.
12. Further exploration has been done to see the correlation between the features and the closing price and some graph has been built to visualize that.

Modeling:

1. 3 feature selection groups have been defined to be the input for the model as following:
    - **All features, All high correlated features, All high correlated financial features**
2. The data has been splitted to different data sets for training and testing (30% for testing)
3. The new sets have been utilized to validate the linear regression model
4. The data has been standardized using Standard scaler before it got used in the following models:
    - **Linear Regression , Lasso and Ridge**
5. The models have been ran over the 3 defined groups for the features and showed that Linear Regression and Ridge are much better than Lasso as shown below

Features	All		
Model Name	LinearRegression	Lasso	Ridge
Absolute percentage error	0.96%	3.71%	0.96%

Features	High Correlated (All)		
Model Name	LinearRegression	Lasso	Ridge
Absolute percentage error	0.95%	3.71%	0.95%

Features	High Correlated (Financial)		
Model Name	LinearRegression	Lasso	Ridge
Absolute percentage error	72.09%	73.31%	72.09%

Tools:

- Colab, pandas, numpy, seaborn, matplotlib, sklearn

Communication:

Google slides have been used to communicate the findings and results generated by colab and python libraries such as seaborn.

Findings:

Close price correlation varies across the different companies, specifically for the financial data