**Linear Regression Project Write-up**

**Predicting the stock price using financial statistics**

**Abstract**

* Goal of the project: Can stock price be explained/forecasted by financial statistics? Or how much can it be explained?
* Famous factor that influence stock price: news, overall market performance, risk-free rate, correlation between target stock return and market return, company news, momentum of the stock price…
* Observation: these factors are more or less relative to company financial statistics/ stock price is largely impacted whenever there is a financial report release

**Design**

This project originates from the idea that “Can stock price be explained/forecasted by financial statistics? Or how much can it be explained?”. The data is provided by Yahoo finance and presents key financial statistics of Nasdaq listing companies. Regress the last quarter stock price against key financial features of the company and tune the model by feature engineering and cross validation and lasso-regression.

**Data**

The dataset contains 4759 waterpoints with 25 features, which are all numerical. A few feature highlights include market capital, total debt, earnings per share.

Web-scraping data from yahoo finance, tab ‘Statistics’ (29 features)

Data cleaning: Delete 2 features due to large portion of NAs/ Delete rows that contain NA entries. Final list has 1579 data point

**Algorithms**

*Feature Engineering*

*Log features*

*Polynomial features*

*Create ‘Cash/debt’ feature by dividing total cash by total debt*

*Models*

OLS Linear regression model

Cross Validation

Lasso regression

*Model Evaluation and Selection*

The entire training dataset of 1579 records was split into 80/20 train vs. holdout, and all scores reported below were calculated with 5-fold cross validation on the training portion only. Predictions on the 20% holdout were limited to the very end, so this split was only used and scores seen just once.

* **Final linear regression 5-fold CV scores:** 3 features Market Cap, Enterprise\_Value and og(Total\_Debt)
* **Key performance:**

**R-square:0.30 in 5-fold CV and 0.28 in test data**

**P-value of variables: all under 0.1**

**Tools**

* Numpy and Pandas for data manipulation
* Scikit-learn for modeling
* Matplotlib and Seaborn for plotting

**Communication**

slides and visuals are presented through a zoom meeting.