Predicting Future Stock Returns Through Ratio Analysis and XG Boost Regression

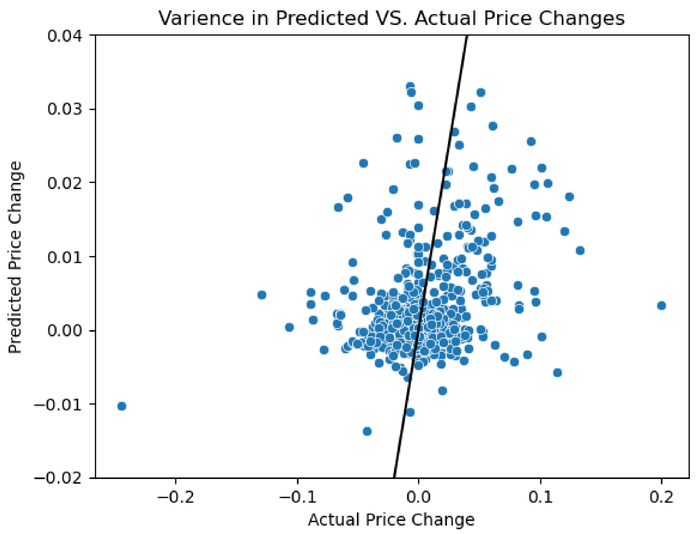
Eric Schneider

Abstract

This project utilizes XG Boosting Regression and ratio analysis to predict future stock returns based on current financial statement information. The model uses the YFinance, Pandas, SciKitLearn, and XGBoost python packages to create a model capable of multiple regression for non-normally distributed, non-correlated financial data.

Introduction

XGBoost, otherwise known as Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) open source machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems. XGBoost is an example of an ensemble learning algorithm, which combines multiple machine learning algorithms to obtain a better model. Because of this, the model is better optimized for performance and accuracy in multi variable based regression, such as the primary model in this project used to predict future stock returns.



Materials and Methods

- **Data Collection:** **YFinance and Pandas were the primary python packages which were used to build the overall dataset. YFinance was used to scrape ratio data from the financial statements published from publicly traded companies on the New York Stock Exchange. After sampling, this method yielded a dataset with 8113 rows.**

- **Data formatting:** **Pandas was then utilized to form the ratio data into a dataset which could be formatted and fitted into a model**

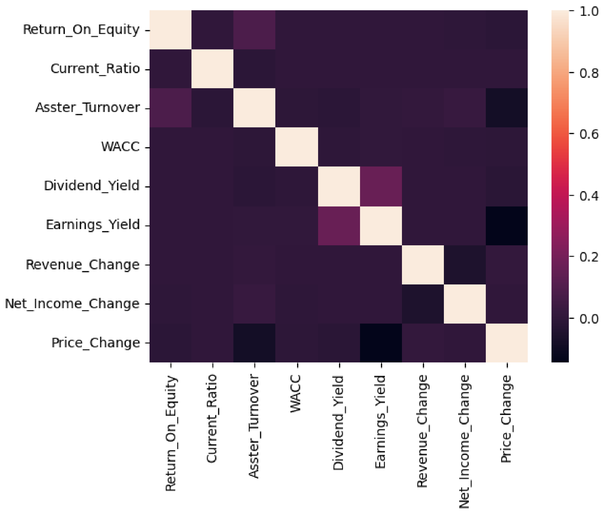
- **Train-test split:** **In order to provide sufficient data available for testing and evaluating the model, scikitlearn was utilized to split the data into the three neccesary train, test, and evaluation datasets utilized in the project. This model was trained with 77.5% of the dataset, 14% used for testing and optimization purposes, and the remaining 8.5% was used for the final evaluation of the model**

- **The XG Boost model:** **The model used a predictive method known as XG Boost Regression. This is a tree based regression model. The parameters of max depth and parallel tree count were optimized to reach the lowest possible standard error.**

- **Evaluation of Model:** **Standard error between predicted and actual price change of stock data was utilized to evaluate the model. This can be considered the expected possible deviation from the true returns on a stock when compared to the percentage returns predicted by the model.**

References

https://www.nvidia.com/en-us/glossary/xgboost/



Objectives

The main objective for this project is to create a model with the capacity to predict the returns for a publicly traded stock utilizing financial ratios collected from annual financial statements.

Results

After optemizating the model, the final standard error for the model was 3.084%.

Conclusion and Future works

While the final standard error for the model reached as low as 3%, a larger dataset utilizing a wider array of companies for a longer period of time could yield better results through the same method of regression.

Contact Info

eschneider@bellarmine.edu