Examining Statistical and Machine Learning Techniques to Predict the Stock Market

PURDUE UNIVERSITY | Krannert School of Management

Alex Roell, Aakash Shambwani, David Mennen, Yashmanth Chintamaneni, Yuvraj Malhotra, Matthew A. Lanham
Purdue University, Krannert School of Management
roella@purdue.edu; ashambwa@purdue.edu; dmennen@purdue.edu; ychintam@purdue.edu; malhotry@purdue.edu; du131@purdue.edu; lu796@purdue.edu; lanhamm@purdue.edu

ABSTRACT

In this study we examine stock market prediction using statistical and machine learning approaches. The motivation for this study is that understanding the stock market and being able to predict when stocks will rise and fall can be very rewarding if done even remotely accurately. We investigate using the h2o R library to shortlist 6 best models for the prediction.

INTRODUCTION

For an investor, predicting the stock market can be a serious challenge. Only 55% of adults invest in the stock market which could be because they are not confident investing in the stock market. Turns out these stocks rise, and an investor has forgone hundreds to thousands of dollars. Our study will help investors be more confident in investing in the stock market.

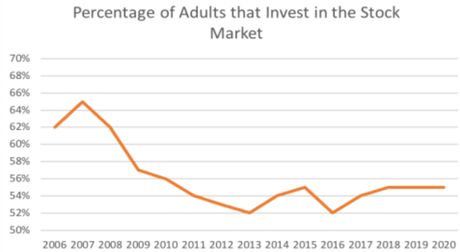

Fig 1. Percentage of Adults that Invest in the Stock Market

Source:https://www.statista.com/statistics/270034/percentage-of-us-adults-to-have-money-invested-in-the-stock-market/

Research Objective:

- **How can we maximize the returns on an investment portfolio?**

LITERATURE REVIEW

Many of the studies utilized an ensemble of different machine learning models and statistics. Unlike the rest of the studies, we utilized an XGBoost model, along with generalized linear models. Especially by using the XGBoost model, we believe our study goes above and beyond and adds predictive performance the others do not.

Study	ARIMA	XGBoost	RF	LM	Ensemble
(2006) Qian			x		x
(2014) Ariyo	x				
(2015) Rather				x	x
(2016) Khaidem			x		x
(2020) Our Study		x		x	x

METHODOLOGY

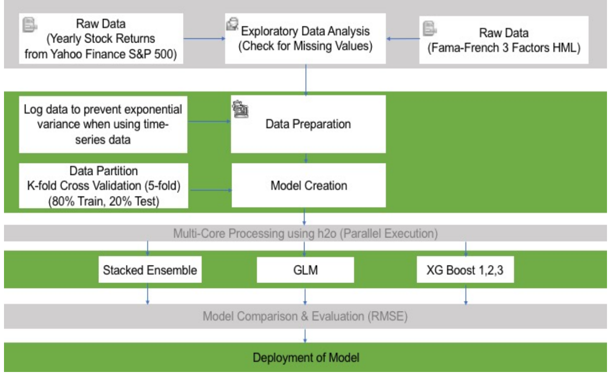

Fig 2. Methodology

- We used the h20 library to train and identify the best candidate model.
- H2O does the data pre-processing and normalization steps.

STATISTICAL RESULTS

The following graph shows the difference in the RMSE of the test and the training datasets. As shown in the graph the difference is bare minimum between them for GLM. Thus we can concur that the data is not overfitting. Out of the 6 models we had chosen using h2o's AutoML functionality which led to the GLM model being the best since it has the lowest test RMSE.

Model	Mean Residual Devaince	RMSE
GLM	6.84E-06	0.00262
XGBoost 1	2.37E-05	0.00487
XGBoost 2	3.21E-05	0.00567
XGBoost 3	1.53E-05	0.00391
StackedEnsemble 1	6.90E-06	0.00263
StackedEnsemble 2	6.96E-06	0.00264

Fig 3 Models and RMSE

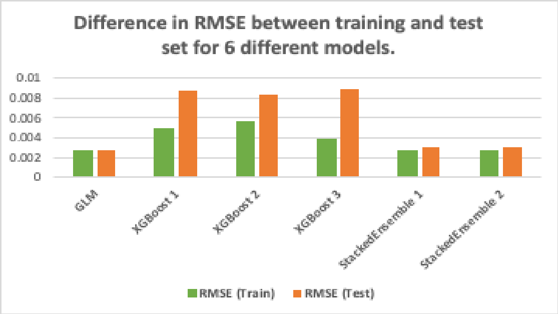

Fig 4. Difference in RMSE for Test and Train

EXPECTED BUSINESS IMPACT

The Dow Jones Industrial Average Index makes an average return of 5.42% per year. The average investor in the United States invests about 10%-15% of their annual salary.

Assuming an investor deposits $6,000 at the beginning of the first year. If their deposit is in for 10 years being compounded yearly, we would expect the investor to earn more using our best model with 64% accuracy versus a 60% accurate model by **$371.66** on average.

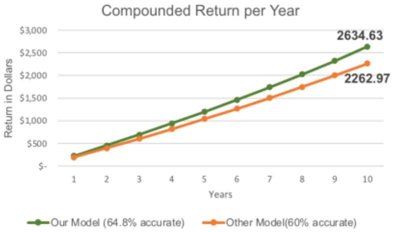

Fig 5. Impact on the Returns for an Investor

CONCLUSIONS

According to our model we predict an accuracy of **64.8%** using the Dow Jones Industrial Average Index. Through this model an investor can increase their returns significantly. For every **1%** increase in model accuracy it will raise the return on the investment for 1 year by **0.0542%.**

ACKNOWLEDGEMENTS

We would like to thank Professor Matthew Lanham, Xinyu Wang, Theo Ginting, and our graduate student mentors Zeyu Du and Zhaotian Lu for their guidance and support on this project.