Teddy Mefford

Summary of Finance Research

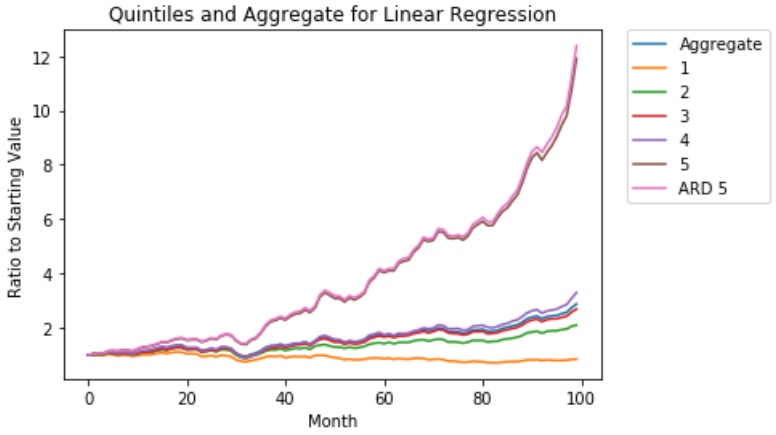
22nd November 2019

OVERVIEW

Our investment strategy identifies undervalued stocks based on a divergence of a company's peer-implied value estimate from its market value. We use machine learning models to predict the value of a certain stock based off of Compustat data to identify mispricing. We then track the returns for the most undervalued stocks for an investment length of one month. Further, we explore why certain models perform better than others. Previous work has only looked at the effectiveness of using a linear regression for price prediction.

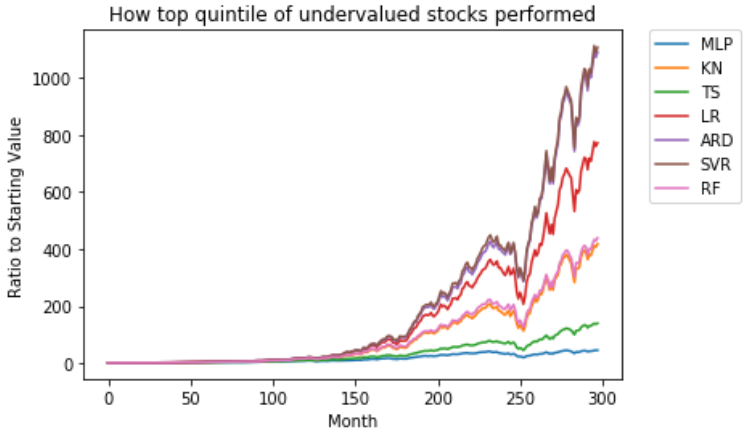
GOALS

1. Effectively identify undervalued stocks
2. Find the model that yields the greatest returns
3. Explore feature reduction
4. Explain the improved performance of models, such as SVR and ARD



SPECIFICATIONS

- Twenty three feature variables: Most important of which (determined by PCA and feature importance analysis) relate to company income. Data Size: 293945 rows × 54 columns
- Data from these features is used to develop a model to predict the price of a stock.
- Data is fed back into the model to obtain a "predicted value."
- Mispricing signal = (price predicted - actual price)/actual price
- Stocks sorted into quintiles and investment goes to the most undervalued quintile. Performance (net return on initial investment) is tracked from Nov 1987 to Dec 2012.



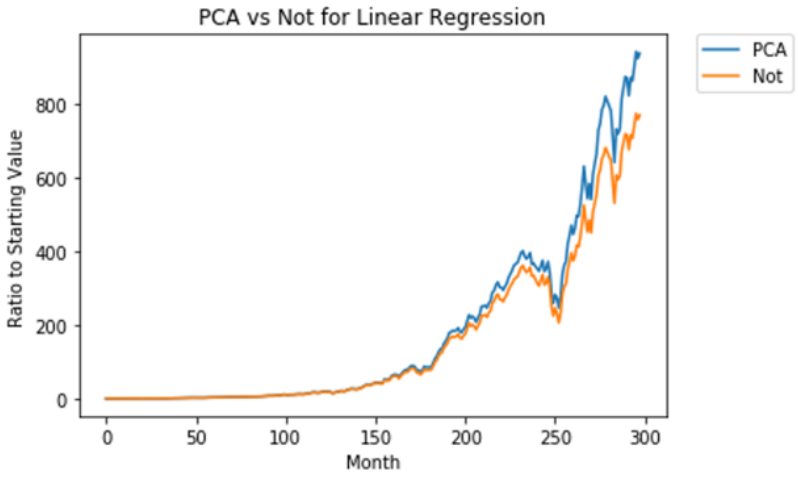
MILESTONES

Improved Model Selection

We have found that several machine learning algorithms, support vector regression (SVR) and automatic relevance determination (ARD), perform better than linear regression. The average monthly return on undervalued stocks for linear regression is 1.02414 whereas it is 1.02533 and 1.02525 for SVR and ARD respectively.

Feature Reduction

Using principal component analysis we reduce the number of features from 23 to 12 which results in a slight improvement in performance. The average monthly return for LR with PCA is 1.02475.



Model Selection Explanation

We are currently exploring the reason behind why specific models will perform better than others for this type of analysis. Based off of the mechanics for how these models work, it makes sense that we see the results that we do. SVR and ARD use Bayesian inference to obtain parsimonious solutions for regression. The key feature of these is that the target function attempts to minimize the number of errors made on the training set while simultaneously maximising the 'margin' between the two classes. This is an effective 'prior' for avoiding over-fitting, which leads to good generalization. To back this up analytically, we specifically identify the stocks that are different for LR and ARD in the undervalued quintile for each month. We can explore differences in the feature variables data to extract some understanding of why each model would choose that type of stock. This analysis is still underway.

This research was conducted under the guidance of Dr. Nicholas Tay, Ph.D, Professor of Finance, USF School of Management

Citation: Bartram, Sohnke and Grinblatt, Mark, "Agnostic Fundamental Analysis Works" Journal of Financial Economics, Volume 128, Issue 1, Pages 125-147, April 2018.