A Bayesian Approach to Predicting Stock Performance Using Analyst Reviews

Elan Miller, Jon Gross, Louis Schlessinger

1 Introduction

Many of the trillions of dollars invested annually in the stock market are invested under the guidance or advisement of some financial professional. However there is much debate on whether these professionals have any better insight into financial markets than average citizens. Stock analysts, who assign ratings on a stock typically ranging from a 1 (strong sell) to a 5 (strong buy), are one notable group, whose recommendations are often considered by both self-directed and institutional investors. Barber aptly summarizes the controversy regarding the effectiveness of stock analyst recommendations in his paper on the matter:

> *"Academic theory and Wall Street practice are clearly at odds regarding this issue. On the one hand, the semi-strong form of market efficiency posits that investors should not be able to trade profitably on the basis of publicly available information, such as analyst recommendations. On the other hand, research departments of brokerage houses spend large sums of money on security analysis, presumably because these firms and their clients believe its use can generate superior returns"* [1]

Likewise empirical evidence on the matter has been mixed as well, with a study conducted by NerdWallet finding that in 2012, 49% of analyst ratings on large companies were incorrect, essentially meaning the analysts were on par with a coin flip. [2]

In this study, we investigate the effectiveness of stock analyst ratings though a Bayesian perspective. Bayesian methods are widely used in order to bring in *a priori* knowledge, such as the stock analyst recommendations, into empirical statistical models. In our case, we will be using the numerical stock analyst ratings to construct prior distributions that will be implemented in a variety of machine learning and statistical models. Models will be benchmarked against frequentist methods that do and do not consider the analyst ratings.

In this analysis, we consider the Dow 30 companies, which are the 30 mega-cap stocks which compose the Dow Jones Industrial Average (DJIA) index. Stock analyst ratings are obtained from Morningstar, a reputable investment research firm known for analysis of stocks and securities. Typically only larger companies are analyzed on a regular basis, hence our choice of the Dow 30 companies. [3] Furthermore, historical ratings on stocks are limited to 3 months, which was January 1 - March 31 2018 at time of data collection. As a result, we obtained daily stock price data for this same time period from Yahoo! Finance for the Dow 30 companies to serve as model evidence in our statistical models. A daily return value for each stock can be trivially calculated as the percent change between the adjusted closing prices of a stock on sequential days. We generally only models that use a time series of past price/return data as predictors. There is justification for this in the stock picking world, and is commonly referred to as technical analysis, which is in turn dependent on an efficient market hypothesis - that a price of a stock reflects all available information about the stock. An overview of the stock return data is shown in Fig. 1, while an example of stock analyst ratings can be seen in Fig. 2.

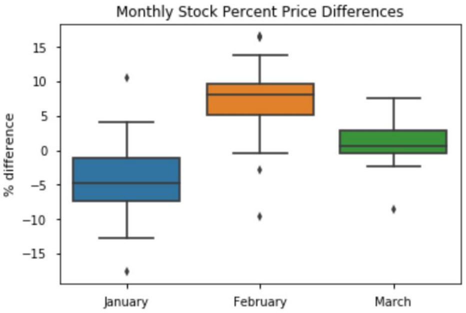
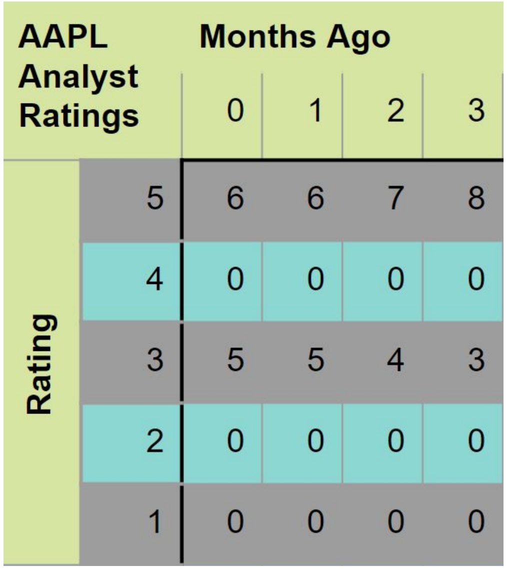


Fig. 1. Overview of monthly stock returns.



AAPL Analyst Ratings		Months Ago			
		0	1	2	3
Rating	5	6	6	7	8
	4	0	0	0	0
	3	5	5	4	3
	2	0	0	0	0
	1	0	0	0	0

Fig. 2. Apple stock analyst ratings.

2 Methods

Since stock analyst recommendations are made in relative to the rest of the market, and since experts commonly agree that it is near impossible to predict exact stock prices with a meaningful level of accuracy, we will not be trying to build models that predict the exact future price or return of stocks. [4] Rather, we will consider the more general classification problem of whether we expect a stock to outperform or underperform the market in terms of return. The market in this context is measured in terms of the NASDAQ index.

In order to solve this classification problem, a total of 13 models are constructed. Some will be discussed in mathematical detail below. Other models, especially those constructed using pre-existing functionality from Python and R packages will only be discussed at a higher level.

To measure accuracy, we divide our data into training and testing data. For most models, January and February were used a training data, and March was used for testing data. We then generate a prediction for whether a stock will beat the market in March, and likewise calculate an accuracy rating based on how many predictions were correct.

2.1 **Model 1**: Bernoulli/Beta Predictive Distribution

We first consider the Posterior Predictive Distribution generated by a Bernoulli likelihood function with a Beta prior distribution on the probability parameter, θ. A beta distribution is typically parameterized by two hyperparameters, α and β, which can used to calculate the mean and variance of a beta distribution as shown by Equations 1 and 2.

$$E[\theta] = \frac{\alpha}{\alpha+\beta} \quad [1]$$

$$var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \quad [2]$$

We are trying to build our prior distribution to reflect analyst ratings. In order to effectively do this, we first scale our analyst ratings to a 0-1 scale, meaning an analyst rating of a 1 (strong sell) corresponds to a scaled rating of 0, an analyst rating of a 5 corresponds to a scaled analyst rating of a 1, with other ratings proportionately scale as well. We then calculate the mean and variance of the scaled analyst ratings. Now, since we want our prior distribution to have the same mean and variance as our scaled analyst ratings, we can apply the method moments, to algebraically solve equations 1 and 2 for α and β, as shown in Equations 3 and 4.

$$\alpha = \frac{-E[\theta](E[\theta]^2 - E[\theta] + var(\theta))}{var(\theta)} \quad [3]$$

$$\beta = \frac{(E[\theta]-1)(E[\theta]^2 - E[\theta] + var(\theta))}{var(\theta)} \quad [4]$$

Where $E[\theta]$ and $var(\theta)$ are equivalent to the mean and variance of the scaled analyst ratings respectively.

In this model, we assume our likelihood function follows a Bernoulli distribution with probability parameter θ which describes the probability of a successful trial. We consider each day as one Bernoulli trial, with a success, $x_i = 1$, being characterized by the stock return beating the market return on that day, or equivalently the difference of the stock return and the market return being greater than 0. Using this conjugate prior model, we know the posterior distribution follows a beta distribution with updated parameters shown by equations 5 and 6 below:

$$\alpha' = \alpha + \sum x_i \qquad [5]$$

$$\beta' = \beta + n - \sum x_i \qquad [6]$$

Since our end goal is create a classification prediction on whether the stock will outperform the market, we next find the probability we observe a success in the future, $p(\tilde{x} = 1|x, \alpha', \beta')$, or more generally, since we assume θ parameter will remain the same for the immediate future, $p(\theta > 0.5 \mid x)$; $p(\theta > 0.5 \mid x) > 0.5$, then we have a belief that the stock will outperform the market in the future, and would thus predict the stock to outperform the market.

2.2 **Model 2**: Bayesian Ridge Regression and ARD Regression

We next consider Bayesian ridge regression and ARD (automatic relevance determination) regression. Classical ridge regression combines ordinary least squares regression with an L2-regularizer. It solves the following problem:

$$\widehat{w} = argmin_w \|Xw - y\|^2 \; + \lambda \|w\|^2 \qquad [7]$$

This weight estimate may also be solved exactly:

$$\widehat{w} = \left(X^T X + \lambda I\right)^{-1} X^T y \qquad [8]$$

Our goal in this model is to map analyst ratings to stock percent price differences for a given month. To learn this function, we create a different model for each stock, using mean analyst ratings as features. After learning this mapping, we then can determine if the prediction was correct if the sign of actual stock percent price difference matches the predicted sign of the stock percent price difference. This allows us to formulate an accuracy metric from the real-valued prediction.

We then include Bayesian ridge regression into our pipeline. This model solves the following optimization problem:

$$\widehat{w} = argmin_w \alpha||Xw - y||^2 + \lambda||w||^2 \qquad [9]$$

We place a Gamma prior on α and λ, which is a conjugate prior. These hyperparameters chosen from a uniform prior. We also place a (spherical) Gaussian prior on the weights w with mean zero and an isotropic diagonal covariance $\lambda^{-1}I$. This is given by:

$$p(w|\lambda) = N(w; 0, \lambda^{-1}I) \qquad [10]$$

We estimate the model parameters using evidence maximization.

Lastly, we also look at ARD regression. It is similar to Bayesian ridge regression, but the regularization term is changed. The penalty term changes from $\lambda||w||^2$ to $w^T\Lambda w$, which gives each weight its own precision λ_i. Consequently, the prior over weights is no longer a spherical Gaussian and is instead is an axis-parallel, elliptical normal distribution, given by:

$$p(w|\Lambda) = N(w; 0, \Lambda^{-1}) \qquad [11]$$

ARD regression solves the following problem:

$$\widehat{w} = argmin_w \alpha||Xw - y||^2 + w^T\Lambda w \qquad [12]$$

where $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_p)$.

The estimation of parameters (α and λ) is also done using evidence maximization.

2.3 **Bayesian Logistic Regression:**

In our implementation of a Bayesian Logistic Regression we use the a binary (-1,1) series of past data as predictors, with a one indicating a stock beat the market on a given day, and a negative one indicating a market outperform the stock on that day. The weights placed on each day follow a gaussian distribution with a mean and variance matching that of the scaled analyst ratings. To implement our bayesian logistic regression, we used a linear formulation with a logit link function to derive a Bayesian GLM, than put the Bayesian GLM through the logit function. This was accomplished using the arm package in R.

2.4 **Naive Benchmark:**

In order to benchmark our statistical and machine learning models, we also create two very simple frequentist benchmarks. In one benchmark model, we track the fraction of days where a stock beat the market; if this fraction is greater than .5, we predict a stock will outperform the market. In another benchmark model, we simply identify whether the total return of a stock across the entire training period was greater than the total market return across the same period; if the stocks total return outperformed the total market, return, we predict a stock will outperform the market. Note that neither of these models is Bayesian nor reflects any stock analyst ratings.

2.5 **Gaussian Processes:**

Using an RBF and Matern kernel, we were able to achieve a 50.3% and 52.6% accuracy respectively. This was using 10-fold cross validation and using all of our data. To try and develop a more accurate model, we added, as additional features on which to build our models, the stock's growth from the data before as well as the market's growth on the previous day. Choosing a random stock, we trained a gaussian process with an RBF kernel on data from January, February and the first half of march. Then, we iterated over the next half of March, making predictions for the day and updating our model with the actual results. This yielded an even lower test accuracy than earlier: 45%.
Thus, we did not pursue this strategy.

2.6 **Other (Frequentist) Models:**

Other frequentist-based linear models were implemented as well for comparison, which include linear regression, cross-validated Ridge Regression, Lasso regression, Huber regression. These were implemented using the scikit-learn library in Python. A frequentist logistic regression was implemented as well, again using the scikit-learn library. In all models, the stock returns time series was still used as predictors, but now we also consider stock analyst ratings as predictors, instead of using them to inform a prior.

3 Results

The overall accuracy of each model is presented in Fig. 3 below. Bayesian methods are shown in green, while frequentist methods are shown in blue. Generally, frequentist models and bayesian models performed at a similar level, with most models in both cases have a 53% accuracy (predicting 16/30 stocks correctly). The best models overall were the naive benchmark model comparing overall period returns and the bayesian logistic regression, both having an accuracy of 63% (19/30 correct). This indicates that the naive benchmark model outperformed all frequentist and most bayesian models. We can also note that the Bernoulli/Beta predictive distribution model performed the worst, with an accuracy of 43%.

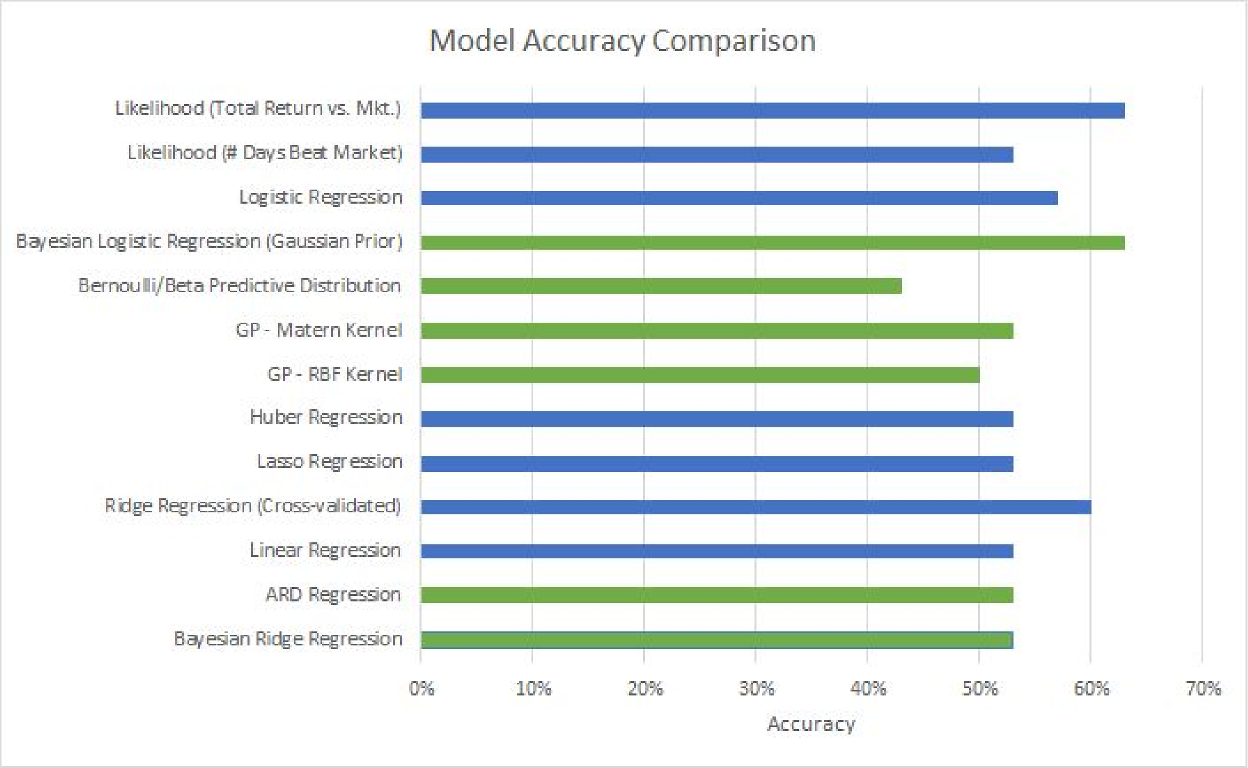


Fig. 3. A plot of final results. Green indicates that the model is Bayesian.

4 Discussion

Overall we find that no models were able to meaningfully outperform the market, and no model incorporating analyst ratings is able to outperform our naive benchmark. Based on this, we fail to conclude that analysts ratings provide any meaningful insight into investment decisions. Given, that Morningstar, the investment research firm that our ratings came from charges customers $200/year, we do not believe that this is a worthwhile service to purchase, and even with a large portfolio, do not believe this investment would be recouped.

It is worthwhile to note a few limitations in our project. First, a buy rating provided by an analyst may not necessarily indicate that the analyst believes the stock will outperform the market in the future; the rating is an indicator that at the time of evaluation, the analyst believed a stock to be under or overvalued. Using this rating in a predictive nature, as we do in this study, depends on the assumption that we have an efficient market that will try to move a stock's price to its true underlying value; for example if a we believe a stock is undervalued, this assumption would indicate that market forces would try to correct this undervaluation in a finite and relatively short amount of time, forcing the stock to grow at an above average rate to reach its proper valuation.

Without this assumption, but still assuming we are able to identify a true valuation, an analyst rating could not be logically in a predictive capacity.

One other major limitation was the limited scope of stocks (30 stocks) and dates (3 months) considered. This was implicitly restricted in our project due to availability of ratings. However, without testing across more companies and more times, it is ultimately very difficult to make any strong conclusion, as changing just one stock leads to large changes in results, meaning our results have a high variability. Furthermore, we only looked at mega-cap companies. However, mega-cap companies have empirically been proven to outperform small-cap companies in the long run, a factor commonly known as the small-cap premium. [5] So, by restricting ourselves to mega-cap companies, we may be implicitly biasing our models. By broadening our scope to look at ratings from other firms, and not just Morningstar, we would likely have enough information to look consider a wider range of companies, although this would require substantial funding to source all these recommendations.

Additionally, all our models utilized return data as predictors. While there was rationale behind this decision, it would be interesting to explore other models that consider other factors as well, such as volume, earnings, beta (volatility), etc.

Despite none of our models being overly successful, nor improving with the incorporation of analyst ratings, we feel that this study overall highlights the complexity of the stock market; even full time professional who have dedicated years toward the study of the stock market have been successful at making good predictions.

References

[1] Brad Barber, Reuven Lehavy, Maureen McNichols, and Brett Trueman. 2002. Can Investors Profit from the Prophets? Security Analyst Recommendations and Stock Returns. (December 2002). Retrieved May 1, 2018 from http://onlinelibrary.wiley.com/doi/10.1111/0022-1082.00336/abstract

[2] Jonathan Hwa. 2017. Study: 49% of Analyst Ratings on the Dow 30 Were Incorrect in 2012. (May 2017). Retrieved May 1, 2018 from https://www.nerdwallet.com/blog/investing/investing-data/investment-stock-analyst-ratings-stockpicking-research-wrong/

[3] Maureen McNichols and Patricia O'Brien. 1997. Self-Selection and Analyst Coverage. Journal of Accounting Research, 35 (1997), 167-199. DOI: 10.2307/2491460

[4] Mark Leung, Hazem Daouk and An-Sing Chen. 2000. Forecasting stock indices: a comparison and classification of level Estimation Models. International Journal of Forecasting, 2 (June 2000), 173-190. DOI: https://doi.org/10.1016/S0169-2070(99)00048-5

[5] Rajiv Kalra. 1999. An Analysis of Mutual Fund Design: The Case of Investing in Small-Cap Stocks. CFA Digest 29, 3 (1999), 36–38. DOI:http://dx.doi.org/10.2469/dig.v29.n3.516