AMATH 521 Final Project

Topic: Sector Outperformance Analysis

Contributors:

1.) Luke Lee
2.) Wipada Wannasiwaporn

Summary:

The goal of this project was to build a model that predicts the sectors that tend to outperform the market return using logistic regression with different penalties and try out different models for comparison. In order to confine our scope of the project, we decided to focus on the two largest sector in S&P500, Technology and Financial.

Data:

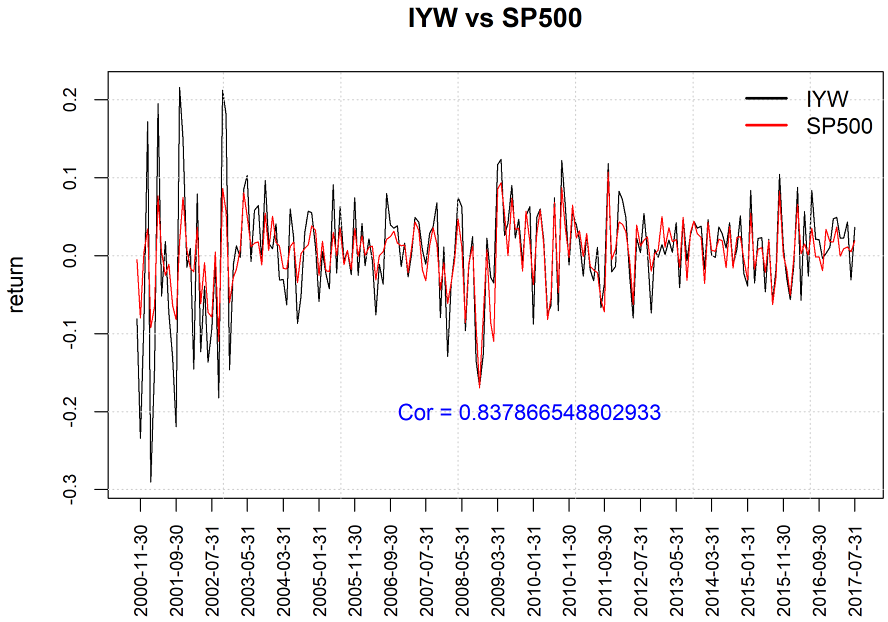
Our independent variables are basically the monthly macro economics data retrieved from FRED(https://fred.stlouisfed.org/). Some examples of these data are unemployment rate, FED funds rate, CPI, SP500 monthly return etc. We also added some sector related data such as Telecommunication export, (TODO: Financial Related)... to possibly add prediction power for a specific sector as well. We convert these raw data into a return space using monthly simple return for an Index-like data and use percentage conversion for a probability data and rate data. Here are some samples of our data.

Table 1. Data Example

Date	IYW return	GDP	CSUSHPINSA	DGS10	TEDRATE	FEDFUNDS
10/31/2000	-0.080775444	0.011088	0.005507	0.0577	0.0057	0.0651
11/30/2000	-0.234329233	0.011088	0.005198	0.0548	0.0068	0.0651
12/31/2000	-0.087222647	0.011088	0.004617	0.0512	0.0067	0.064
01/31/2001	0.172170997	0.003422	0.003953	0.0519	0.0056	0.0598

For the dependent variable, we use a separate model for each sector.

i) For Technology Sector, we use BlackRock's ETF(IYW) that tracks the Technology sector performance using Dow Jones as a benchmark. Overall, the returns of Technology sector moves strongly correlated with the return of S&P500 as seen in the high correlation value.



ii) For Financial Sector,

Methods & Algorithms:

i) Linear Regression

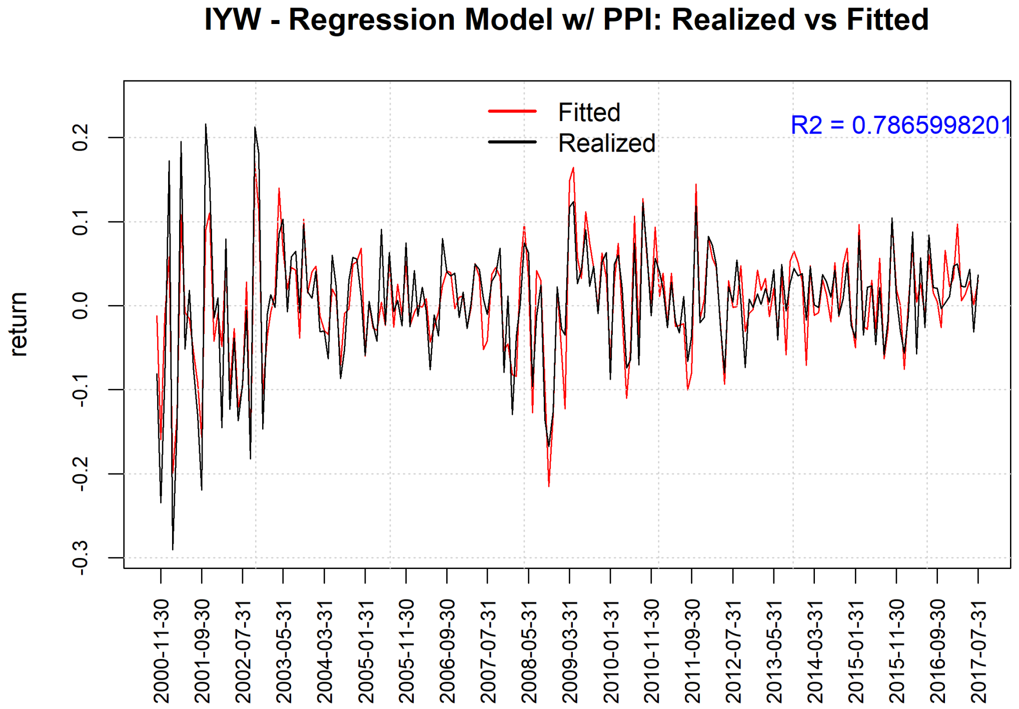
We first want to make sure that our data can explain the data well, so doing linear regression is one way of doing sanity check on our data. By using our whole data set, it turns out that the explained variance(R^2) is quite high for this data set, so we know our independent variables can somehow explain our dependent variable.

$$\mathbf{r} \in \mathbb{R}^T, \mathbf{F} \in \mathbb{R}^{T\times n}, \mathbf{x} \in \mathbb{R}^n$$

$$\mathbf{r} \sim F_1x_1 + F_2x_2 + ... + F_nx_n$$

where $\mathbf{r}$ represents our sector return(IYW - dependent variable) and $\mathbf{x}$ represents our economics data(independent variables).

The results of this linear regression shown a high coefficient of determination R^2.



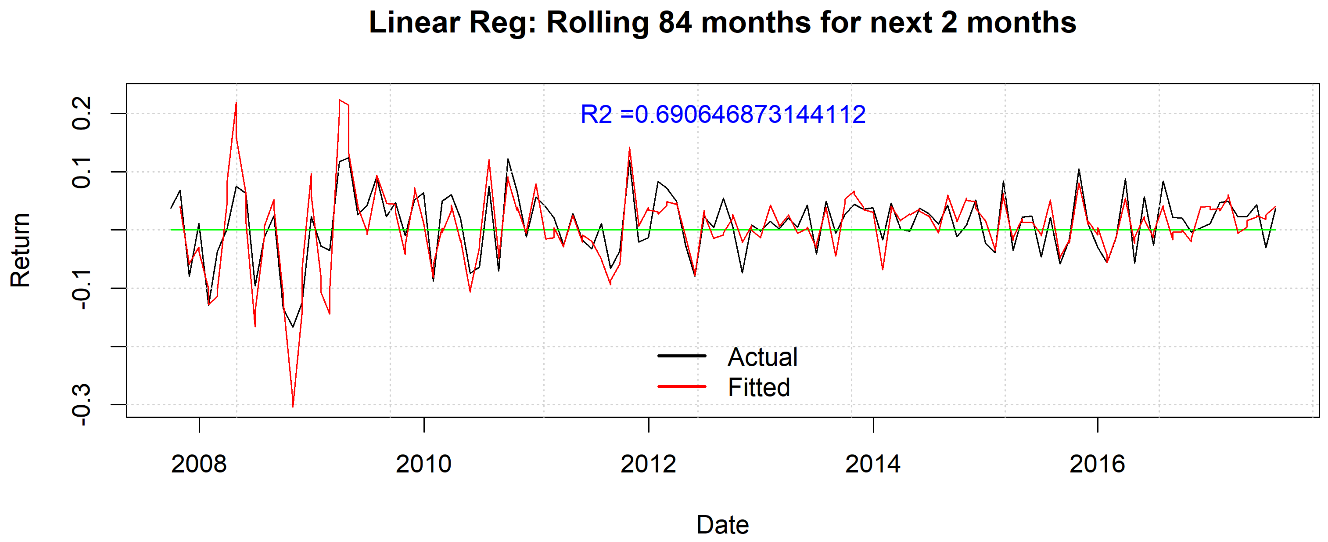
After doing some sanity check, we try to predict whether IYW will beat the market in the next period by assuming we know the exact values of our independent variables. We use the combination of rolling windows of 5, 6, 7, ..., 10 years of training samples to predict the next 1, 2, 3, ..., 6 months return.

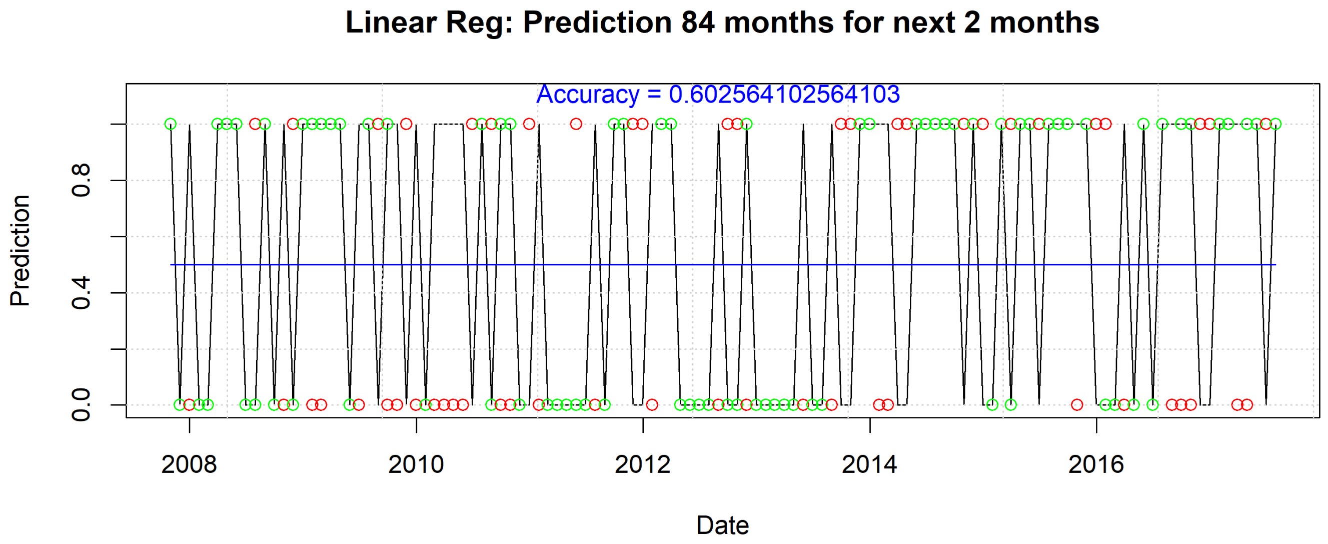
After we get our predicted return, we check whether this return is greater than the market and use this as a boolean result of our linear regression. Using this boolean output, we can assess our outperformance accuracy by using

$$Accuracy = \frac{Numbers\ of\ correct\ predictions}{Numbers\ of\ total\ predictions}.$$

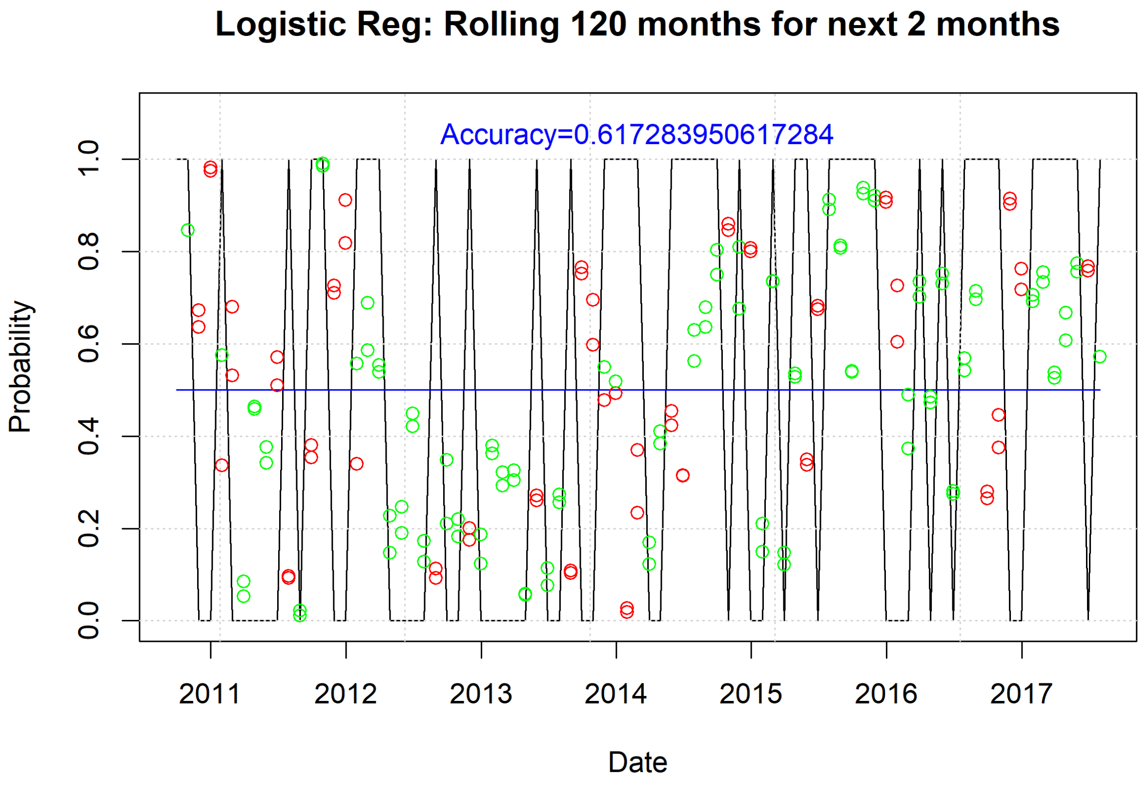
Using accuracy as our selected criterion, we ran different combination of windows size and as a result, using 7 years of training data and 2 months of prediction period will give us the highest accuracy for linear regression method.

The plots below show the result of prediction using 7 years training data to predict the next 2 months value.

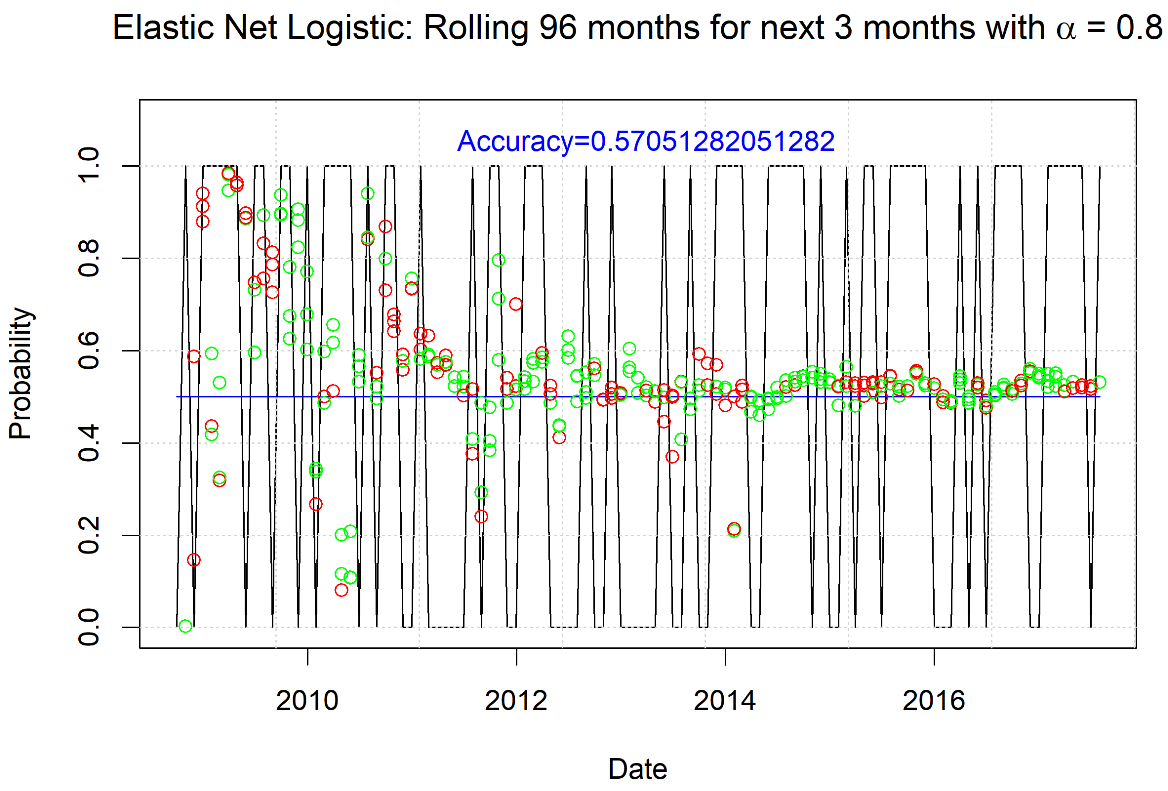




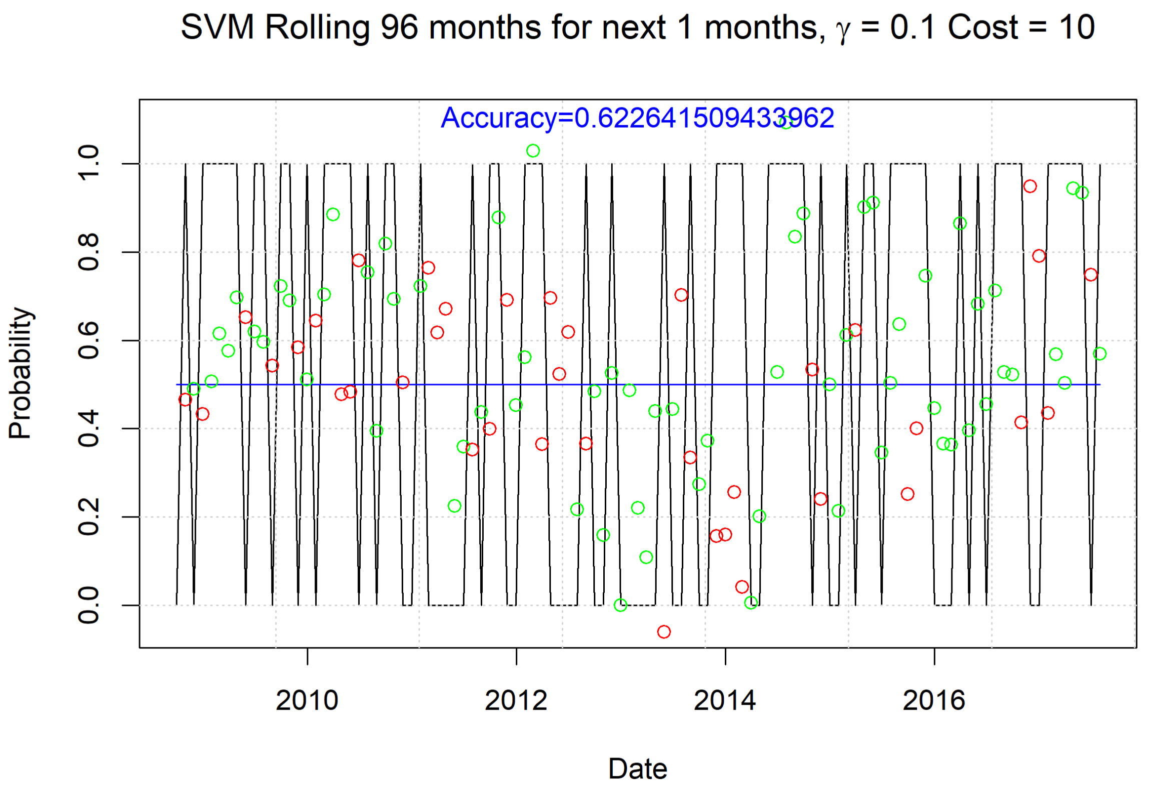
ii) Logistic Regression



iii) Logistic Regression with Elastic Net



iv) Support Vector Machine


SVM Rolling 96 months for next 1 months, γ = 0.1 Cost = 10
Accuracy=0.622641509433962
Probability
Date