Stevens Institute of Technology

Machine Learning - Final Project

Best ETF Performance

Arjun Koshal & Jeffrey Eng

December 5, 2022

1 Introduction

Stock performance evaluation has always been unique to each investor. Every person has different appetites for risk, plans for diversification, and investing strategies. The same can be said for evaluating stock performance. Every investor has different standards when it comes to evaluating stock performance, as one investor may expect an average annual return of 10%, whereas another investor may expect to choose stocks that are uncorrelated with each other. Overall, there are a few variables that we need to consider when evaluating the best stocks for our portfolio.

For the case of our study, we will be using exchange-traded funds (ETF), which are a type of pooled investment security that operates much like a mutual fund. A mutual fund is a financial vehicle that pools assets from shareholders to invest in securities such as stocks, bonds and other assets. Typically, ETFs will track a particular index, sector, commodity, or other assets. Unlike mutual funds, ETFs can be purchased or sold on a stock exchange similar to a regular stock, They can be structured to track anything from the price of an individual commodity to a large and diverse collection of securities. The first ETF was the SPY, which tracks the S&P 500 index and remains one of the most popular ETFs to this date.

The popularity of ETFs arise from the fact that ETFs are considered to be low-risk investments because they are low-cost and hold a basket of stocks or other securities, increasing diversification. When it comes to evaluating ETF performance, it requires more than simply looking at the change in price over time. We need to first calculate the daily log returns for 5 years of data and properly compare these returns to the SPY benchmark. Our goal is to accurately predict the 5 best performing ETFs in the next year from a list of 20 ETFs randomly selected. We want to compute the Annual Jensen's α, Market β, Treynor Ratio (TR), and Information Ratio (IR), and investigate the different types of machine learning methods to analyze the data. From the mean squared error and accuracy metrics, we aim to conclude which will be the top 5 performing ETFs in the upcoming year.

2 Data Collection and Discussion

Our first step was to obtain data for all 20 ETFs, which were chosen at random. Our data dated back to 5 years (2017) and the 20 ETFs were IVV, IJH, IJR, SUSA, VXUS, SCHD, VIG, DNL, PDBC, RYF, RYH, PBW, XLK, VGK, FBND, BSCQ, TOTL, ICSH, BKLN, VTEB. Our portfolio consisted only these 20 ETFs, so we proceeded to calculate the daily log returns for each ETF and then created a statistics summary for annualized mean returns, volatility, and sharpe ratio of our portfolio. Sharpe ratio compares the return of an investment with its risk.

	Mean Returns	Volatility	Sharpe Ratio
Q1	0.02560763	0.07531996	0.2689202
Median	0.07599233	0.19104728	0.4403905
Mean	0.07701182	0.17088563	0.4587464
Q3	0.11313356	0.20825237	0.5828274

Table 1: Statistics Summary for Annualized Mean Returns, Volatility, and Sharpe Ratio.

We then plotted the mean returns against volatility for each ETF to verify the risk-return tradeoff principle.

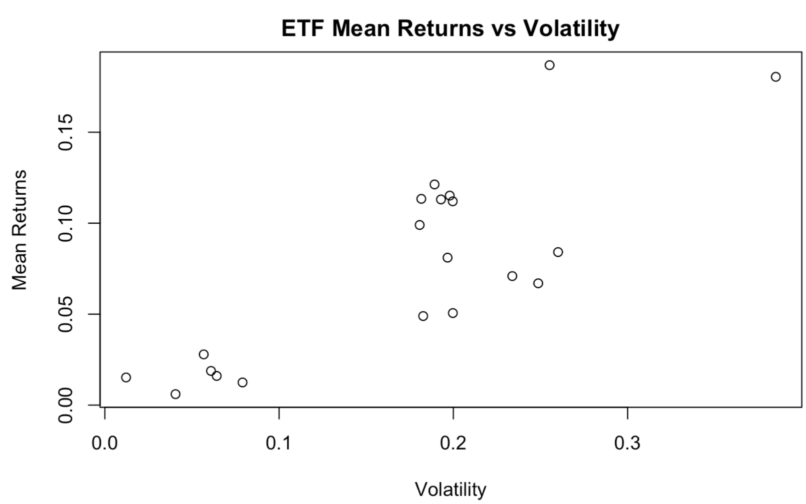

Figure 1: Plot of each individual ETF Mean Returns vs Volatility

We can see from the scatterplot above that there is a clear trend that as volatility increases, the mean returns increase as well. This makes sense as the bottom left implies that the lower the risk the investor is willing to take, then they can expect lower mean returns. The top right implies that the greatest the risk the investor is willing to take, then the higher the mean returns the investor can expect. Based on the risk-return tradeoff, the scatterplot above follows the principle.

For the performance metrics, we used Annual Jensen's α, Market β, Treynor Ratio (TR), and Information Ratio (IR), where the benchmark we used was the SPY. The Annual Jensen's α is the difference in how much a person returns vs. the overall market. Market β is a measure of the volatility, or systematic risk, of a security or portfolio compared to the market as a whole. In this case, the market was the SPY ETF. A β greater than 1.0 can be interpreted as more volatile than the benchmark, and less than 1.0 can be interpreted as less volatile. When β is greater than 1.0, the ETF will move with more momentum than the SPY, and the opposite can be said if β is less than 1.0. The Treynor Ratio, also known as reward-to-volatility ratio, is a performance metric for determining how much excess return was generated for each unit of risk taken on by a portfolio. Excess return in this sense refers to the return earned above the return that could have been earned in a risk-free investment. The Information Ratio is measurement of portfolio returns beyond the returns of a benchmark compared to the volatility of those returns. The table below shows the performance metrics for our portfolio.

	Annual Jensen Alpha	Market Beta	TR	IR
Q1	-0.021575	0.20660	0.044450	-0.456975
Median	0.006800	0.87265	0.106900	-0.340900
Mean	0.000100	0.69415	0.245990	-0.225160
Q3	0.017350	1.03180	0.201075	0.033275

Table 2: Statistics Summary for Performance Metrics against SPY Benchmark.

After analyzing each ETF individually and comparing to overall portfolio performance metrics, it appears that the best performing ETF from our selected 20 ETFs appears to be XLK. XLK is the technology select sector SPDR Fund. XLK had an annualized Jensen alpha of 0.0002, a Treynor Ratio of 0.1364, and an Information Ratio of 0.7147. Comparing these performance metrics to the summary table that was produced, we can state that XLK was in the top 25 percentile for almost all three of these metrics. The Jensen alpha shows that XLK risk-adjusted returns are independent of the benchmark S&P 500 and current outperforming the market. The Treynor Ratio shows that XLK is able to generate more excess returns over the benchmark for each unit of risk taken on by a portfolio. The Information Ratio shows that XLK's returns were beyond the returns of the S&P 500 compared to the volatility of those returns. Seeing XLK's Treynor Ratio and Information Ratio were extremely high relative to the other ETFs, we can state with confidence that XLK is the best performer.

3 Machine Learning Methods

The ability to successfully and accurately forecast the trend in stock market price movements is crucial for investors and traders. Machine Learning algorithms have been implemented to accomplish this task due to their high prediction accuracy of stock prices and stock price direction. Our goal is to utilize 4 different machine learning methods to understand both individual stock movements and see the consequences of grouping certain stocks together. The machine learning methods are linear regression, logistic regression, decision trees, and random forest. We evaluate model performance using the following criteria:

1. Root Mean Squared Error: Since the models are estimated at the stock level, it is natural to compare the average magnitude of prediction errors for individual stock returns. To facilitate interpretation, we use the root mean squared error (RMSE), which has the same unit as returns. RMSE is defined as:

$$RMSE = \frac{1}{N}\sum_{i=1}^{N}(r_i - \hat{r_i})^2$$

   in which r_i is the actual return for stock i and $\hat{r_i}$ is the forecasted return.

2. Confusion Matrix: We use confusion matrices for classification problems and compare the actual target values with those predicted by the machine learning model. For logistic regression, we will use the stock price directions and calculate the number of true positives, true negatives, false positives, and false negatives, to see how accurate our model is.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

3.1 Linear Regression

Linear regression denotes a linear relationship between the dependent variable and the independent variable or variables. The dependent variable is the target variable in the model. The target variable in our study would be the future returns of the ETFs. We use a ”simple linear regression” to understand the individual ETF movements. Our historical data was utilizing each ETF’s daily log returns. We use this data to create a model that predicts the future returns and compare the training mean squared error and testing mean squared error.

3.2 Logistic Regression

Logistic regression is also used to estimate the relationship between a dependent variable and one or more independent variables, but it is used to make a prediction about a categorical variable versus a continuous one. A categorical variable can be true or false, yes or no, 1 or 0, etc. In our study, our goal was to use the ETF directions and run a logistic regression on these directions. We will evaluate the performance of this model with the test accuracy and confusion matrix.

3.3 Decision Tree

Decision Tree is used to deal mostly with classification problems. We start with a root node where our data is situated at. The data then is split into two or more mutually exclusive child nodes depending on different classes, each child node is in turn split into grandchild nodes and so on. The trees that are descended from the root node called sub-trees. Each node in the decision tree acts as a test case for some attribute, and each sub-nodes descending from the node corresponds to the possible responses to the test case. We will build a decision tree based on the training data and then evaluate the performance by comparing the training mean squared error and testing mean squared error.

3.4 Random Forest

Random Forest builds multiple decision trees, often denoted as forests, to builds a large collection of de-correlated trees by randomly selecting a subset of predictors for splitting at each branch. We created a random forest with 250 trees and 2 predictors for each ETF. The final outcome in the random forest method is reached by simply combining the outcomes of the multiple decision trees that are created randomly and then taking the average of all the outcomes obtained from all these decision trees in the forests based on the respective parameters that have been used in each tree. We will build a random forest based on the training data and then compare the training mean squared error and testing mean squared error.

4 Results

For Linear Regression, Decision Tree, and Random Forest, we looked to compare the training and testing MSE to check if overfitting occurred. Overfitting occurs when the training and testing MSE significantly differ and the MSE is not close to zero. For Logistic Regression, we calculated the accuracy from the confusion matrices constructed that were focused on the direction of the stock price.

4.1 Linear Regression

Using our 20 ETFs, we computed the 2-day lags for each ETF and computed the log returns for the lags. We then used a 80-20 train test split on our data. After creating a data frame consisting the log returns along with the 2 day lags, we begun building a linear regression model for each ETF. Our goal was to find the MSE for our model, and the train/test MSE to see if the model was not over fitting or over fitting. If our MSE was very small or close to 0, then we can state that our model is a good fit. Our results from the MSE alone can not tell us whether our model is good. We needed to check if our model was over fitting through comparing the train and test MSE. If the train MSE is greater than the test MSE, then we can conclude our model is not over fitting, which is good. However, if the train MSE is less than the test MSE, then our model is over fitting, which is not good. Below are the results:

Ticker	Training MSE	Testing MSE	Ticker	Training MSE	Testing MSE
IVV	1.449718e-06	1.946325e-06	RYH	2.477085e-05	2.703632e-05
IJH	5.674613e-06	6.572765e-06	PBW	1.799857e-04	1.927175e-04
IJR	1.260333e-05	1.576633e-05	XLK	1.102945e-05	1.504484e-05
SUSA	2.695511e-06	2.755554e-06	VGK	9.924502e-06	1.074057e-05
VXUS	4.750525e-06	4.601869e-06	FBND	2.529168e-06	1.742265e-05
SCHD	8.372009e-06	1.133287e-05	BSCQ	3.702204e-06	1.204186e-05
VIG	3.615916e-06	5.438634e-06	TOTL	1.932715e-06	4.123096e-06
DNL	9.992703e-06	1.328232e-05	ICSH	1.846753e-07	1.298915e-06
PDBC	9.617753e-05	1.053121e-04	BKLN	7.305342e-06	8.838036e-06
RYF	2.608987e-05	2.891369e-05	VTEB	2.882545e-06	1.367884e-05

Table 3: Linear Regression: Training and Testing Mean Squared Error

Based on the MSE for all of the ETFs, they are all close to zero and very small, so it is a good fit. However, using the Linear Regression model, the 5 best ETFs that are IJR, VXUS, SUSA, RYF, and PBW because they are the ETFs where the training and testing MSE do not significantly differ compared to the other ETFs.

4.2 Logistic Regression

For our logistic regression, we computed the directions of the ETFs individually and implemented a logistic regression for the direction of the ETF as a function of the 1 and 2 lagged returns. We constructed 20 confusion matrices below to display the accuracy of our algorithm.

IVV (Accuracy = 60.29%)

Predicted \ Actual	0	1
0	7	1
1	109	160

IJH (Accuracy = 55.23%)

Predicted \ Actual	0	1
0	12	7
1	117	141

IJR (Accuracy = 54.15%)

Predicted \ Actual	0	1
0	15	11
1	116	135

SUSA (Accuracy = 53.99%)

Predicted \ Actual	0	1
0	7	2
1	125	142

VXUS (Accuracy = 54.15%)

Predicted \ Actual	0	1
0	13	13
1	114	137

SCHD (Accuracy = 52.71%)

Predicted \ Actual	0	1
0	3	1
1	130	143

VIG (Accuracy = 61.01%)

Predicted \ Actual	0	1
0	2	0
1	108	167

DNL (Accuracy = 56.68%)

Predicted \ Actual	0	1
0	1	0
1	120	156

PDBC (Accuracy = 54.15%)

Predicted \ Actual	0	1
0	0	0
1	127	150

RYF (Accuracy = 55.96%)

Predicted \ Actual	0	1
0	21	18
1	104	134

RYH (Accuracy = 50.54%)

Predicted \ Actual	0	1
0	0	3
1	134	140

PBW (Accuracy = 49.10%)

Predicted \ Actual	0	1
0	0	0
1	141	136

XLK (Accuracy = 57.76%)

Predicted \ Actual	0	1
0	16	22
1	95	144

VGK (Accuracy = 58.12%)

Predicted \ Actual	0	1
0	5	3
1	113	156

FBND (Accuracy = 52.71%)

Predicted \ Actual	0	1
0	1	7
1	124	145

BSCQ (Accuracy = 50.54%)

Predicted \ Actual	0	1
0	4	1
1	136	136

TOTL (Accuracy = 52.71%)

Predicted \ Actual	0	1
0	63	46
1	85	83

ICSH (Accuracy = 54.35%)

Predicted \ Actual	0	1
0	138	121
1	5	12

BKLN (Accuracy = 54.71%)

Predicted \ Actual	0	1
0	151	125
1	0	0

VTEB (Accuracy = 54.15%)

Predicted \ Actual	0	1
0	42	43
1	84	108

Table 4: Logistic Regression: Confusion Matrices

From our results, we notice that the algorithm was predicting primarily an increasing for the ETF, when in reality the ETF moved either up or down. The top 5 ETFs from logistic regression were VIG, IVV, VGK, XLK, and DNL. The accuracy of these stocks were all above 55%.

4.3 Decision Tree

For the decision tree, we calculated the training and testing MSE and compared the values. Below are the mean squared errors.

Ticker	Training MSE	Testing MSE	Ticker	Training MSE	Testing MSE
IVV	1.568076e-04	1.902792e-04	RYH	1.469118e-04	1.810647e-04
IJH	2.163091e-04	2.268949e-04	PBW	5.454987e-04	7.927279e-04
IJR	2.623293e-04	2.336575e-04	XLK	2.520830e-04	3.165514e-04
SUSA	1.559078e-04	2.258836e-04	VGK	1.585150e-04	1.783620e-04
VXUS	1.432308e-04	1.369483e-04	FBND	2.527512e-05	3.178746e-05
SCHD	1.537137e-04	1.416429e-04	BSCQ	1.676635e-05	1.019022e-05
VIG	1.362591e-04	1.621486e-04	TOTL	7.728490e-06	1.158101e-05
DNL	1.590870e-04	1.752902e-04	ICSH	5.750837e-07	9.647586e-07
PDBC	1.108707e-04	1.750218e-04	BKLN	3.402670e-05	4.051174e-05
RYF	3.265265e-04	3.813630e-04	VTEB	2.198505e-05	1.338796e-05

Table 5: Decision Tree: Training and Testing Mean Squared Error

We can see that for each ETF, the MSE are all close to zero and very small. Using the Decision Tree model, the 5 best ETFs are DNL, IJH, VXUS, SCHD, and VGK. The training and test MSE do not significantly differ compared to the other ETFs.

4.4 Random Forest

For the random forest, we also calculated the training and testing MSE and compared the values. Below are the mean squared errors.

Ticker	Training MSE	Testing MSE	Ticker	Training MSE	Testing MSE
IVV	4.376286e-05	2.244206e-04	RYH	4.613888e-05	2.219638e-04
IJH	6.954818e-05	3.169826e-04	PBW	1.708551e-04	9.840916e-04
IJR	7.690994e-05	2.920931e-04	XLK	6.427096e-05	3.878972e-04
SUSA	4.364739e-05	2.754985e-04	VGK	5.147345e-05	2.198474e-04
VXUS	4.298925e-05	1.543738e-04	FBND	6.479572e-06	2.742761e-05
SCHD	4.267506e-05	1.638741e-04	BSCQ	4.749250e-06	1.358457e-05
VIG	3.829530e-05	1.793762e-04	TOTL	2.199943e-06	1.868954e-05
DNL	4.502723e-05	2.287150e-04	ICSH	1.474787e-07	1.053154e-06
PDBC	4.087895e-05	2.153785e-04	BKLN	7.447293e-06	3.826441e-05
RYF	8.432627e-05	4.177540e-04	VTEB	5.923151e-06	1.871042e-05

Table 6: Random Forest: Training and Testing Mean Squared Error

We noticed that the training and testing MSE differed a little for each ETF. This can be attributed to the fact that random forest modeling is prone to overfitting, so the accuracy may be high for our training data, but on new data the accuracy would be lower. From the training and testing MSE, it is difficult to state which MSEs were differed the least, so we computed the percent errors between each MSE. The 5 ETFs were the lowest MSE were BSCQ, VTEB, VXUS, IJR, and SCHD. All of the percent errors for the MSE for each ETF lied between 0.65 and 0.88, proving that the random forest MSE for all ETFs performed well.

5 Conclusions and Further Discussion

From our 4 different types of machine learning methods, we can state the 5 top performing ETFs. Below are tables with the best ETFs from our linear regression, logistic regression, decision tree, and random forest models.

Ticker	Training MSE	Testing MSE
IJR	1.260333e-05	1.576633e-05
VXUS	4.750525e-06	4.601869e-06
SUSA	2.695511e-06	2.75555e-06
RYF	2.608987e-05	2.891369e-05
PBW	1.799857e-04	1.927175e-04

Linear Regression: Top 5 Performing ETFs

Ticker	Accuracy
VIG	61.01%
IVV	60.29%
VGK	58.12%
XLK	57.76%
DNL	56.68%

Logistic Regression: Top 5 Performing ETFs

Ticker	Training MSE	Testing MSE
DNL	1.590870e-04	1.752902e-04
IJH	2.163091e-04	2.268949e-04
VXUS	1.432308e-04	1.369483e-04
SCHD	1.537137e-04	1.416429e-04
VGK	1.585150e-04	1.783620e-04

Decision Tree: Top 5 Performing ETFs

Ticker	Training MSE	Testing MSE
BSCQ	4.749250e-06	1.358457e-05
VTEB	5.923151-06	1.871042e-05
VXUS	4.298925e-05	1.543738e-04
IJR	7.690994e-05	2.920931e-04
SCHD	4.267506e-05	1.638741e-04

Random Forest: Top 5 Performing ETFs

From the performance metrics calculated from the data collection, we know that XLK is the best performing ETF; however, we can select the other 4 best ETFs from our different types of models. We can see that VXUS appears 3 times, VGK appears 2 times, IJR appears 2 times, SCHD appears 2 times, and DNL appears 2 times. VXUS was a top performer for 3 of our methods, so we can include VXUS in our selection of the best 5 ETFs. Comparing VGK, IJR, SCHD, and DNL, both IJR and SCHD were the top performers in the random forest model, which seemed to have indications of overfitting. Therefore, we can include VGK and DNL in our portfolio for the top 5 performing ETFs. With only IJR and SCHD remaining, since IJR had a higher accuracy from the logistic regression, we can also include IJR in our portfolio. Thus, from our 20 ETFs, XLK, VXUS, VGK, DNL, and IJR were the top performers.

For the next steps, we could have potentially created portfolios and seen which portfolio outperforms the market. Since individual ETFs are volatile on their own, constructing multiple portfolios that contained different combinations of ETFs would have been beneficial for our project. We could have also utilized different types of algorithms such as Naive Bayes Classifier with the ETF price directions or a support vector machine using a radial basis kernel. Both of these methods could have been implemented to determine if there were other ETFs that could have been the top 5 performers.

Appendix

The work in this document was contributed by both Arjun Koshal and Jeffrey Eng. Jeffrey worked on the introduction and selected 20 ETFs at random. The data analysis, financial ratios, and performance metrics were computed and described by Arjun. Jeffrey worked on the write up for Annual Jensen Alpha and Market Beta and Arjun worked on the write up for Treynor Ratio and Information Ratio. We then worked together on implementing the machine learning methods, where Jeffrey focused on linear regression and logistic regression, and Arjun focused on decision tree and random forest. Lastly, the conclusion and next steps were completed together by both Arjun and Jeffrey.