# Finding S&P 500 Predictors from Dow Jones Transportation Components using Lasso Feature Selection

**Nicholas L. Trieu***
Department of Computer Science
Harvey Mudd College
Claremont, CA 91711
ntrieu@g.hmc.edu

## Abstract

One detriment of investing in diversified financial assets such as mutual funds is that they may rack up unnecessary transaction costs from churning through many different stocks. One way that these diversified funds may mitigate this issue is by investing only in stocks that matter most towards their objective. We apply Lasso Feature Selection to figure out what transportation sector stocks matter most in predicting the S&P 500. Our results indicate that the ground and delivery transportation sectors better predict the S&P 500 than air transportation.

## 1 Introduction

### 1.1 Churning

We begin by describing mutual funds and the churning problem. Mutual funds are investment pools that collect money from many investors in order to invest in a diversified portfolio. These funds are managed by specialized companies, such as the *Vanguard*, which manages more than $3 trillion in assets for more than 20 million investors (as of August 31, 2016). The benefit of investing with mutual funds is that pooling your money with many other investors allows you to diversify your portfolio, maintaining the expected returns while reducing risk.

However, these funds have "hidden" or implicit fees that arise from managing the fund. Among these fees are the transaction costs for buying and selling securities, which result in a reduction in the fund's rate of return. If the mutual fund has a high *churn rate* of buying and selling securities, these transactions may reduce the fund's yield not just from transaction costs, but also from the market impact of the fund's large transactions.

One way of mitigating transaction consequences is to simply reduce the churn rate. But this may not be feasible for funds that must rebalance long-term investment portfolios, such as retirement funds. When done right, even frequent rebalances may not necessarily be bad; for example, the Ab Nicholas investment firm has had 25% annual turnover rates while still beating the S&P 500 from 2008 to 2015. The key is that we must not only reduce the churn rate, but also maintain the benefits of diversification.

An alternative solution is to minimize the number of securities being transacted while maximizing the variance provided by these securities. We thus apply Lasso Feature Selection to determine which stocks contribute most to the objective.

*Project code hosted on Github at: https://github.com/NLTrieu/Math-189r-Big-Data

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Date | Time | Open | High | Low | Close | Volume | OC_Mag | HL_Mag | OC_Dir | OC_Chain | Per_Chang | Per_Chain |
| 2 | 8/19/2004 | 1155 | 50.06 | 50.31 | 49 | 49.05 | 2563134 | 1.01 | 1.31 | -1 | -1 | 2.017579 | -2.01758 |
| 3 | 8/19/2004 | 1156 | 49.05 | 49.42 | 48.4 | 48.43 | 1819175 | 0.62 | 1.02 | -1 | -2 | 1.264016 | -3.2816 |
| 4 | 8/19/2004 | 1157 | 48.43 | 48.48 | 48.03 | 48.46 | 831342 | 0.03 | 0.45 | 1 | 1 | 0.061945 | 0.061945 |
| 5 | 8/19/2004 | 1158 | 48.4 | 49.18 | 48.17 | 49.16 | 764878 | 0.76 | 1.01 | 1 | 2 | 1.570248 | 1.632193 |

Figure 1: Example data from GOOGL stock. Contrary to the screen shot, the data were loaded the data directly from text files because they were too big to fit in Excel.

### 1.2 Transportation and Dow Theory

One basic tenet of Dow Theory is that movements in the transportation sector correlate with future movements of the general market (Brown and Kumar [1998]). The idea is that as companies grow bigger or do more business, they need more transportation to transport goods. Thus, the general market (i.e. the S&P 500) should already adjust to transportation sector movements by the efficient-market hypothesis (Malkiel [2003]).

Suppose a mutual fund wants to invest in the Dow Jones Transportation Average in order to move with the general market. This leads us to the questions: Which transportation stocks matter most? What aspects of those stocks have the biggest influence?

In this project, we take 19 stocks comprising the Dow Jones Transportation Components from 2002 to 2016 and determine which stocks matter most in correlating with the S&P 500. We then analyze how the different indicators of a single stock's data correlate with each other.

## 2 Data processing

The given stock and ETF data had the following minute-by-minute features:

- $Date$: The year, month, and day, given as $Month/Day/Year$.
- $Time$: The time of day, given as $HourMinutes$ concatenated in a single string. Trading hours are from 9:30 AM to 3:59 PM.
- $Open$: The opening price at the start of the minute.
- $High$: The highest transaction price during the minute.
- $Low$: The lowest transaction price during the minute.
- $Close$: The closing price at the end of the minute.
- $Volume$: The trading volume, i.e. the number of shares transacted during that minute.

The dates for a single stock or ETF usually ranged between 2002 and 2016, but some newer stocks and ETFs only had data over 5-7 years. Notably, the data did not include the best bid or ask for the stocks and ETFs; not knowing these potentials limited the analysis we conducted.

Several features were derived from these stats (see Figure 1):

- $OC\_Mag$: The Open-Close difference magnitude, actually calculated across minutes as $Close - PreviousClose$ due to data inconsistency issues.
- $OC\_Dir$: The Open-Close direction (the sign of $Closed - Open$, including 0).
- $Per\_Change$: The percentage change magnitude (calculated as ${}^{OC\_Mag}/_{Open} * 100$).

There were several issues with the data that needed to be addressed before performing PCA. We summarize these issues and our resolution methods as follows:

- Time skips: The stock and ETF data have missing minutes here and there. These time skips throw off comparisons between different securities. The last known feature values were reused to fill in these gaps.

- Differing time ranges: Not all stocks had dates ranging from 2002 to 2016. Some started in 2006, 2007, or even as late as 2012. Obviously the same gap-filling trick does not apply here, so we set the latest start date to 2007 and threw away any stocks that started later in order to keep above one million data points.

- Open-Close, High-Low constants: Some stocks have the same opening, closing, high, and low prices within a minute for a consecutive sequence of minutes (see Figure ). Confusingly, the prices do change over the sequence of minutes (i.e. $Open! = PreviousClose$) and the trading volume is nonzero, so shares are being traded at different prices but the $Open$, $Close$, $High$, and $Low$ are simply being treated as one value.

  To mitigate this issue, we calculate the change in prices across minutes instead of intra-minute. This means that $OC\_Mag = Close - PreviousClose$ (notably not $OC\_Mag = Open - PreviousClose$ because the majority of data points had $Open \neq PreviousClose$). We effectively ignore the $high$ and $low$ features because too many data points simply set them equal.

- Stock splits and multiples: When a stock's share price becomes too high for traders to trade, the stock will split its shares (i.e. generating two shares for every old share) and halve its price. Similarly, a stock may multiply its price at dangerously low levels (to avoid volatility). Clearly, these splits and multiples will throw off any price difference calculations across minutes.

  To resolve this issue, we compute a hidden multiplier if the stock's price drastically changes (i.e. doubling or halving) and base our features off of that multiplier.

The original data files for each stock and index were first filtered by time (conforming them to a common time range and filling time gaps), then had the derived features added. For example, the $ALK.txt$ file for Alaska Air Group was read, filtered into $ALK\_filled$, then feature-modified into $ALK\_filled\_modded$.

## 3 Lasso Implementation

### 3.1 Rationale

We use Lasso L1 regularization to encourage sparse solutions, knocking out weak predictors and selecting a subset of the transportation stocks. By looking at the regularization path, we can compare the optimal weights for each stock. Given the diversity of the industries represented by the Dow Jones transportation components (see Table 1), we hypothesize that no one stock should dominate. However, we can try interpreting the Lasso optimized weights and see if any patterns arise.

### 3.2 Algorithm

Our algorithm performs Lasso feature selection. We consider the regularized empirical risk minimization problem

$$\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1$$

whose gradient is given by

$$\nabla_{\mathbf{w}} \left[\|X\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1\right] = 2X^\top X\mathbf{w} - 2\mathbf{y}^\top X + \lambda \text{sign}(\mathbf{w})$$

and solve using proximal gradient descent. Our proximal wrapper for the gradient descent step is the threshold function

$$\text{prox}_r(\mathbf{x}_i) = \begin{cases} \mathbf{x}_i - r, & \mathbf{x}_i > r \\ 0, & |\mathbf{x}_i| \leq r \\ \mathbf{x}_i + r, & \mathbf{x}_i < -r \end{cases}$$

where $r$ is the learning rate.

Table 1: Dow Jones Transports Components (Dec 5, 2016). Excluded: MATX (too few data points).

| Corporation | Ticker | Industry |
|---|---|---|
| Alaska Air Group, Inc. | ALK | Airlines |
| American Airlines Group Inc. | AAL | Airlines |
| Avis Budget Group, Inc. | CAR | Rental and Leasing Services |
| C.H. Robinson Worldwide, Inc. | CHRW | Trucking |
| CSX Corp. | CSX | Railroads |
| Delta Air Lines | DAL | Airlines |
| Expeditors International | EXPD | Delivery Services |
| FedEx Corporation | FDX | Delivery Services |
| JB Hunt Transport Services, Inc. | JBHT | Trucking |
| JetBlue Airways Corp. | JBLU | Airlines |
| Kansas City | KSU | Railroads |
| Kirby Corp. | KEX | Marine Transportation |
| Landstar System, Inc. | LSTR | Trucking |
| Norfolk Southern Corp. | NSC | Railroads |
| Ryder System, Inc. | R | Transportation Services |
| Southwest Airlines, Inc. | LUV | Airlines |
| Union Pacific Corp. | UNP | Railroads |
| United Continental Holdings | UAL | Airlines |
| United Parcel Service, Inc. | UPS | Delivery Services |

### 3.3 Implementation

Our data set $X$ consists of the price percentage changes ($Per\_Change$) for the transport stocks in Table 1.

Our $\mathbf{y}$ to predict consists of the price percentage changes for the SPY index, which has traditionally tracked the S&P 500.

We consider the Dow Jones transportation components from their latest start date of "05/03/2007" to the earliest end date "10/07/2016" excluding MATX, which starts in 2012. This gives us around one million minute-by-minute data points to work with.

## 4 Experiments

It is important that we use normalized price movements (i.e. the percentage differences) and not the unnormalized prices to give each stock weight the same meaning. Figure 2 shows what happens when we try learning based on the opening prices: one stock gets an absurdly large weight while the others go to 0. Due to inflation and economic growth, the S&P 500 index has had a general upward trend from 2007 to 2016. The same positive trend applies to at least one of the stocks in the Dow Jones transportation components. As a consequence, the Lasso model over-aggressively zeros the weights of all but one stock, scales the growth of that stock to match the S&P 500, and then adds a constant to match the means. This interpretation with the extreme weights would tell us that one stock represents the transportation sector, which is absurd.

In contrast, the regularization path plot in Figure 3 shows the distribution of weights for each stock after learning based on the percentage changes. The weights in this plot seem much more reasonable than those in Figure 2; some stocks are near-zero, having been eliminated as weak predictors, while the others are more evenly distributed across the spectrum with no severe outliers (also see Table 2, which orders the stocks from biggest weight to smallest, where the weights come from optimizing the Lasso objective $\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1$). The one negative weight in the regularization path plot corresponds to the constant bias term appended to $X$; all of the stock weights are positive.

The regularization path plot in Figure 4 shows the distribution of weights for the stocks and the Dow Jones Transportation Average index IYT after learning based on the percentage changes. Including the index results in a similar plot to the plot without the index; one distinction, however, is that the LSTR stock's weight is now negative (see Table 3).

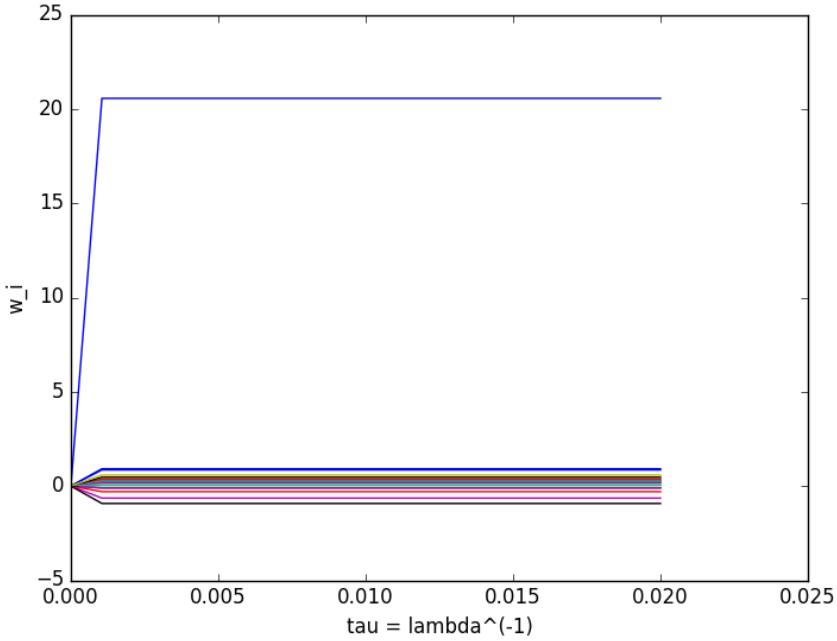


Figure 2: Lasso Regularization Path: Transport Stocks $Open$ Predicting SPY

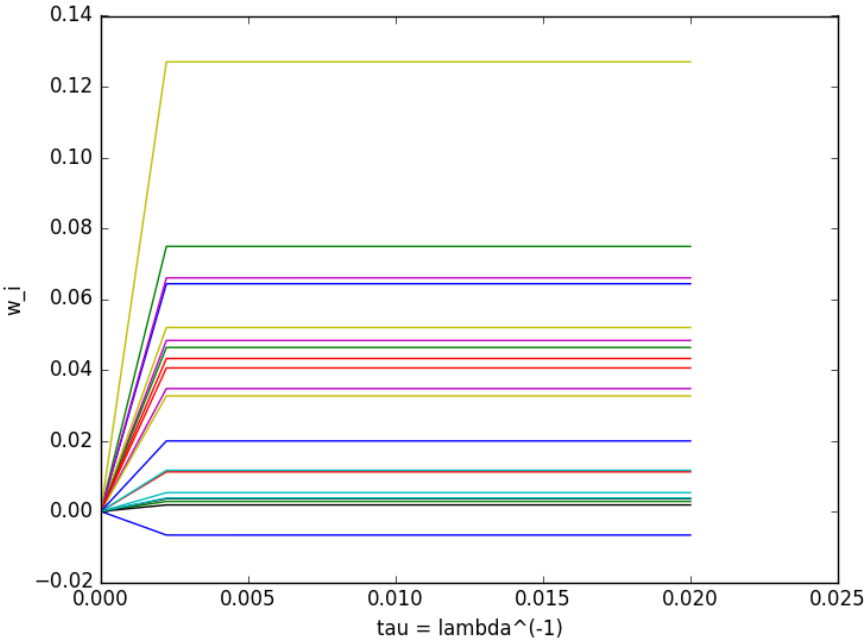


Figure 3: Lasso Regularization Path: Transport Stocks $Per\_Change$ Predicting SPY

Table 2: Dow Jones Transports Components$Per_Change$ Weights.

| Corporation | Ticker | Industry | Weight ($w_i$) |
|---|---|---|---|
| United Parcel Service, Inc. | UPS | Delivery Services | 0.126998024737 |
| FedEx Corporation | FDX | Delivery Services | 0.0748592406015 |
| Union Pacific Corp. | UNP | Railroads | 0.0659628826315 |
| Expeditors International | EXPD | Delivery Services | 0.06430810349 |
| CSX Corp. | CSX | Railroads | 0.0519410568611 |
| C.H. Robinson Worldwide, Inc. | CHRW | Trucking | 0.0482897806767 |
| Norfolk Southern Corp. | NSC | Railroads | 0.0463079329717 |
| Ryder System, Inc. | R | Transportation Services | 0.0431943007092 |
| JB Hunt Transport Services, Inc. | JBHT | Trucking | 0.0405205437987 |
| Kirby Corp. | KEX | Marine Transportation | 0.0346942025747 |
| Kansas City | KSU | Railroads | 0.0326286114762 |
| Southwest Airlines, Inc. | LUV | Airlines | 0.0198953182821 |
| JetBlue Airways Corp. | JBLU | Airlines | 0.0116079269819 |
| Alaska Air Group, Inc. | ALK | Airlines | 0.0111592409286 |
| Avis Budget Group, Inc. | CAR | Rental and Leasing Services | 0.0052952656847 |
| United Continental Holdings | UAL | Airlines | 0.00377363673377 |
| Landstar System, Inc. | LSTR | Trucking | 0.00359611247851 |
| American Airlines Group Inc. | AAL | Airlines | 0.0028201725868 |
| Delta Air Lines | DAL | Airlines | 0.00182567356804 |

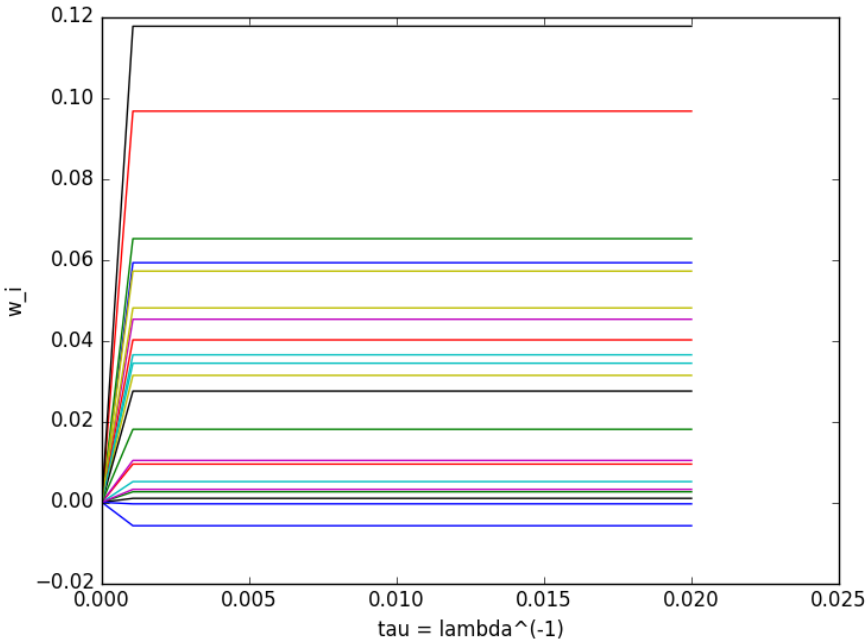


Figure 4: Lasso Regularization Path: Transport Stocks and IYT $Per\_Change$ Predicting SPY

Table 3: Dow Jones Transports Components and IYT $Per\_Change$ Weights.

| Corporation | Ticker | Industry | Weight ($w_i$) |
|---|---|---|---|
| United Parcel Service, Inc. | UPS | Delivery Services | 0.117660544381 |
| iShares Dow Jones Transport. Avg. (ETF) | IYT | Exchange Traded Product | 0.0966779643449 |
| FedEx Corporation | FDX | Delivery Services | 0.0651563154648 |
| Expeditors International | EXPD | Delivery Services | 0.0592153834167 |
| Union Pacific Corp. | UNP | Railroads | 0.0571421595973 |
| CSX Corp. | CSX | Railroads | 0.0480401072606 |
| C.H. Robinson Worldwide, Inc. | CHRW | Trucking | 0.0452393253124 |
| Norfolk Southern Corp. | NSC | Railroads | 0.040143297956 |
| Ryder System, Inc. | R | Transportation Services | 0.036453092742 |
| JB Hunt Transport Services, Inc. | JBHT | Trucking | 0.0343717495153 |
| Kirby Corp. | KEX | Marine Transportation | 0.0313717091875 |
| Kansas City | KSU | Railroads | 0.0274816703841 |
| Southwest Airlines, Inc. | LUV | Airlines | 0.0180378182726 |
| JetBlue Airways Corp. | JBLU | Airlines | 0.0103502320302 |
| Alaska Air Group, Inc. | ALK | Airlines | 0.00943822842662 |
| Avis Budget Group, Inc. | CAR | Rental and Leasing Services | 0.00510799412822 |
| United Continental Holdings | UAL | Airlines | 0.00319945764963 |
| American Airlines Group Inc. | AAL | Airlines | 0.0025910135552 |
| Delta Air Lines | DAL | Airlines | 0.000977802693483 |
| Landstar System, Inc. | LSTR | Trucking | -0.000391734850494 |

## 5 Discussion

The $Per\_Change$ weights in Table 2 include the sign of the weights, not just the magnitudes. Interestingly, we see that all of the stock weights are positive (balanced by the negative constant weight - see Figure 3). This implies that none of the transports components have gone against the general market for the past 10 years, perhaps due to inflation. In contrast, the $Open$ weights are evenly positive and negative, giving further proof that the two features do not yield the same results.

Table 3 of the $Per\_Change$ weights including the transportation index IYT is similar to Table 2. This is as expected, since the only difference is that the transportation index ETF has been included with the transport stocks. All the stocks except for LSTR have positive weights and the order of the corporations is mostly the same (FDX and EXPD switch, LSTR falls to the bottom). The index beats all of the stocks except for UPS, which emphasizes the necessity of diversity.

From Table 2, we notice a pattern between the transport industries and their respective $Per\_Change$ weights. Airlines occupy the bottom half with weights less than 0.02. Delivery services occupy three of the top four spots with weights greater than 0.06, about three times the airline weights. Land transportation (railroads, trucking) forms the middle, with the exception of the LSTR trucking company.

These results suggest that the market associates price movements in delivery and ground transportation sectors with the general market more so than price movements in the air transportation sector. The heavier weights for delivery services support the Dow theory model, which hypothesizes that as businesses grow, they need more transportation specifically to deliver their goods. The lighter weights for airlines imply that the majority of goods are transported by ground or sea as opposed to by air, which may be more expensive or unfeasible for large quantities.

So suppose a mutual fund wants to invest in the Dow Jones Transportation Average as a predictor of the general market. From this analysis, we conclude that the fund should focus on ground transportation and delivery services over airlines.

**Acknowledgments**

We thank Professor Gary Evans, Harvey Mudd College, for supplying the data used in this project.

## References

William N. Goetzmann Brown, Stephen J. and Alok Kumar. The dow theory: William peter hamilton's track record reconsidered. *The Journal of Finance*, 53.4:1311–1333, 1998.

Burton G. Malkiel. The efficient market hypothesis and its critics. *The Journal of Economic Perspectives*, 17.1:59–82, 2003.