The University of Texas at Austin
Texas McCombs
MS Business Analytics
McCombs School of Business

Predicting Volatility

Bull Stearns
Adam Hoard, Ari Chowdhury, Brandt Green, Ian Arzt

Presentation Link: Final Project: Predicting Volatility.pptx
Link To Code: Notebook link GitHub

Executive Summary

As Warren Buffett stated, “Be fearful when others are greedy and be greedy when others are fearful.” This quote is in-part the inspiration for this project. We wished to come up with a systematic methodology for trading on this idea. To do so, we needed to have some measure of fear and our metric of choice was volatility. Thus, the primary prediction task we focused on was accurately forecasting future volatility. With a forecast of volatility in hand, we implemented two trading strategies. Strategy 1 involved building sorted portfolios based on our predicted volatility, and strategy 2 entailed building sorted portfolios based on the difference between current volatility and future volatility.

We used a total of 13 different prediction models to forecast future volatility, and we believe our contribution is in-part, an illumination on the out of sample success of various, commonly used default estimations of stock volatility.

Our two primary hypotheses going in were that models that use only time series components will outperform models that use a cross-sectional approach and that portfolios sorted on the volatility metrics discussed above would result in positive, statistically significant alpha.

Result Summary:
We found that the best performing model was an OLS model that uses a cross-sectional approach to parameter estimation and incorporates all of the available variables. This model achieved an R^2 of .35 in explaining the out-of-sample variance in volatility of companies. This model’s predictions were used for the portfolio construction. Portfolio strategy 1 achieved a statistically insignificant CAPM alpha of -.005 as well as a CAPM beta of .92. Strategy 2 did result in a statistically significant alpha of -.006 with a beta of .27.

Table of Contents

Objective

The task we set out to achieve is simple in concept: predict the daily volatility of returns for a stock in the next month. For each stock, this is calculated as:

$$MonthlyVol_T = \sqrt{\sum_{t=1}^{n} \frac{(DailyReturn_t - MeanReturn_T)^2}{n}}$$

Equation 1: Monthly Volatility Equation

Where *n* is the number of days of returns in the month and *MeanReturn* is the mean return of the stock in that month.

Variable Selection

Throughout this course, our team has attempted to delve deep into selecting variables for our models. For this project, we have several different variable types including fundamental / company specific, unstructured data, and macroeconomic features.

Variables & Reasoning

- **Debt/Equity Ratio**: Company's long-term debt to book value of equity ratio. We expect that companies with more debt, should have higher volatility, due to the leverage effect.
- **Size (Market Cap):** We expect that larger companies should be less volatile. We expect this because larger companies should have a more diversified revenue stream, and therefore, their earnings should exhibit lower volatility. Additionally, it is simply common empirical knowledge that smaller companies are more volatile.
- **Lag Volatilities:** This is our most important predictor. We use a variety of different measures of historical volatilities, depending on which model we are using.
    - **Lag Volatility**: Lagged version of the dependent variable 'Monthly Realized Volatility'
    - **Lag Volatility 2-12:** These 11 variables represent the lagged volatilities over the previous 11 months before Lag Vol.

    - **Historical Volatility**: we use 4 different variables here, each with longer lookback periods. You get the benefit of more data, but you also lose the ability to capture nuances or changing volatility distributions. Our four lookback periods were:
        - **Historical Volatility 1 Year**: Daily volatility over the past year.
        - **Historical Volatility 2 Year:** Daily volatility over the past 2 years.
        - **Historical Volatility 5 Year:** Daily volatility over the past 5 years.
        - **Historical Volatility All:** Daily volatility over the entire stock's history at a given time.
- **Lag Return:** The stock's return in the previous month. Much research has shown that returns and volatility are negatively correlated. Some researchers again cite the leverage effect as the cause for this: negative returns cause the debt/equity ratio to increase on a market value basis.
- **Litigious Word Count:** The number of times a company uses a litigious word in its most recent filing. Uncertainty about the results of legal preceding can heighten volatility, we expect.
- **International Word Count:** The number of times a company uses an international word in its most recent filing. Like the size effect, we expect companies with operations in many regions to reap diversification benefits by selling to multiple markets.
- **VIX:** The 30-day forward looking volatility of the market derived from S&P 500 options. We expect this should be highly predictive, especially for stocks with high betas.
- **WTI Lagged Volatility:** One of the main benchmarks used by Oil Markets. We expect positive correlation, between oil volatility and stock volatility especially in certain industries, as highly volatile points in oil prices occur during times of uncertain economic phenomena.
- **3-Month Treasury Rate:** We expect a negative correlation between this variable and volatility because of the flight to safety during turbulent times. When people are highly risk averse, they buy treasuries, which pushes this yield down.

Data Overview

Sources

- WRDS
    - CompStat Fundamentals: company financial variables and pricing data.
- EDGAR
    - UT MSBA Google Drive: Edgar text filings of 10-Ks & 10-Qs.
- University of Notre Dame: Litigious word count for reports.

- Homemade Word Dictionary: the word dictionary we created to identify certain words in company filings.
- FRED:
    - VIX: CBOE Volatility Index
    - WTI: Crude Oil Prices: West Texas Intermediate
    - 3-Month: 3-Month-Treasury Rate
- Ken French Website: Factor returns data for portfolio evaluation

Data Wrangling

There were many steps to the data aggregation process, which we will discuss, but below you can see what we consider our "fundamental dataframe" meaning that this is the core data set that is used to build our other dataframes, depending on the model in question. This dataframe has a row for every day for every company that we have complete data for all our variables. We call this dataframe "df_daily" and it contains 492 unique companies from the period 2000-2020. There are 2,157,363 rows. Only one of our models (GARCH) uses daily data; the others use monthly statistics that are aggregated from this table.

Below, Table 1 shows a snippet of our data frame that is being used in our models.

tic	year	month	day	ret	vix	wti	three_month_rate	wti_ret	debt_to_equity	international_count	lm_litigious_count	mkt_cap
AAPL	2003	1	2	0.032272	25.39	31.97	1.20	0.024059	0.349794	227.0	509.0	5312.59320
AAPL	2003	1	3	0.006734	24.68	33.26	1.20	0.039558	0.349794	227.0	509.0	5348.48910
AAPL	2003	1	6	0.000000	24.91	32.29	1.19	-0.029598	0.349794	227.0	509.0	5348.48910
AAPL	2003	1	7	-0.003361	25.13	31.20	1.17	-0.034339	0.349794	227.0	509.0	5330.54115
AAPL	2003	1	8	-0.020409	25.53	30.66	1.17	-0.017459	0.349794	227.0	509.0	5222.85345

Table 1: DataFrame

Data Notes

S&P 500 Companies only: This choice was made to keep the data size and computational complexity at a manageable level and to reduce the problems of data issues found in smaller, less known companies.

Merging Daily and Quarterly Data: Most of our data was available on a quarterly basis, but some, such as debt/equity ratio are available only on a quarterly time frame. We combine these data sets by merging the fundamental data set onto the daily data set based on the release date of the fundamental data, which was extracted from EDGAR. This means that for every day in the df_daily data set, the quarterly variables will be equal to whatever the most recently released report contained, as of that day.

Input Data Matrix

The above df_daily is not in the right structure that is necessary for us to feed into our models. We envision models similar to the following form:

$$MonthlyVol_T = \beta_0 + \beta_1 MonthlyVol_{T-1} ... + \beta_{12} MonthlyVol_{T-12} + \beta_{13} Return_{T-1} + \beta_{14} Vix_{T-1} + \beta_{15} 3Month_{T-1} + \beta_{15} WtiVol_{T-1}$$

Equation 2: Monthly Volatility Regression Equation

Because of this, we need to take the daily data and convert it to monthly, aggregated values. This aggregation process involves calculating the monthly volatilities for all stock, and the end-of-month values for all relevant macro data. We label this new data frame 'DF_X'. See Appendix Figure 1 for a snippet of this dataframe. The data set contains all the variables discussed in the **Variables** section above.

This data set contains one row for every month where we have complete data for the S&P 500 companies. This data set contains 467 unique companies and is made up of 72,547 rows.

Terminology & Procedural Clarification

The way this data frame is set up, allows easy flexibility for us to shift back and forth between the two modeling paradigms: time series & cross-sectional. When we use the term "time series model", we are referring to a model that uses only a single company's rows to estimate the parameters of the model. When we use the term "cross sectional model", we are referring to a model that uses data from all companies to estimate the model parameters. These terms may not be correct in the technical sense, but it is how we have best been able to distinguish and communicate our results.

For example, if we are sitting at July 31st, 2015, and want to predict the volatility for the month of August 2015 for Apple, Inc. we could use either type of model. If we use the time series model, we will filter DF_X so that it only contains data on Apple through July 31st and train our model with just this data. Alternatively, if we went with the cross-sectional approach, we would keep all data for all companies available on July 31st and feed this into our model. With the time series approach, every single company will have a different model on every single prediction date each month, with different parameter estimates for each variable. With the cross-sectional approach, we create one model on every prediction date, and this same model is used to predict for all companies.

There are various pros and cons to each approach. The cross-sectional approach's main benefit lies in the training data set size. Because you lump all companies together you have much more data

with which to estimate the model parameters, but the downside is that you likely miss estimate the parameters for each individual company because the estimates are diluted by the aggregation of the data together. This is similar to the familiar bias-variance machine learning trade off. By using more companies, we increase the bias of our estimates for the model parameters for each individual company, but we reduce the variance and sensitivity to noise.

Exploratory Data Analysis

The volatility structure remains similar across most companies. In Figure 1 below, we plot the average daily volatility per month for Apple, along several of the lagged volatility variables.

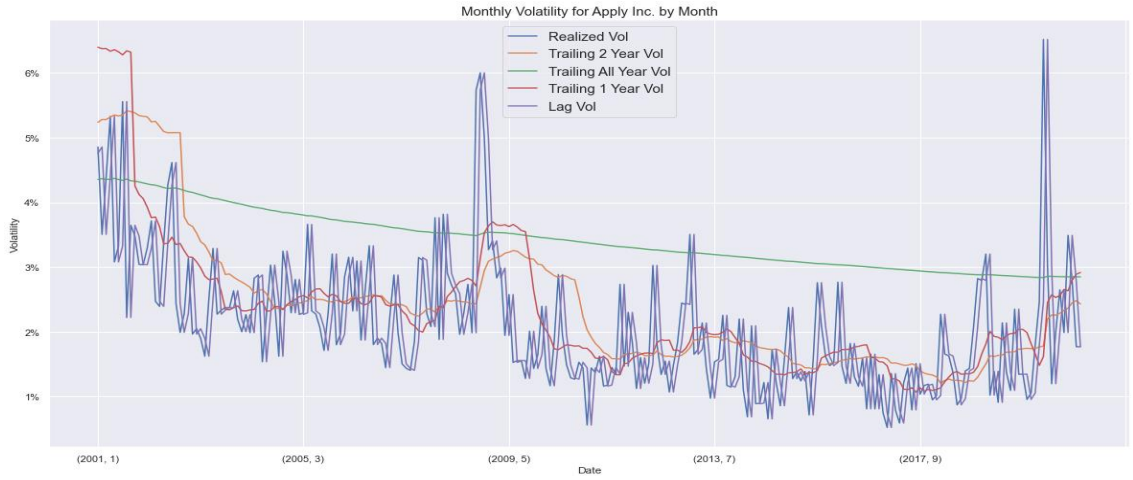

Figure 1: Realized Vol vs Lagged Variables for Apple

1-month lagged volatility is too close and noisy as a predictor for average daily volatility per month. Similarly, historical volatility (1-year and beyond) is too smooth and far to be significant predictors of volatility. This can be seen further from the correlation plot (in Figure 2 below) where realized volatility is highly correlated with immediate lagged variables but not as much with more historical variables. We thus need to find the sweet spot through a combination of multiple lagged volatility variables and financial and unstructured variables.

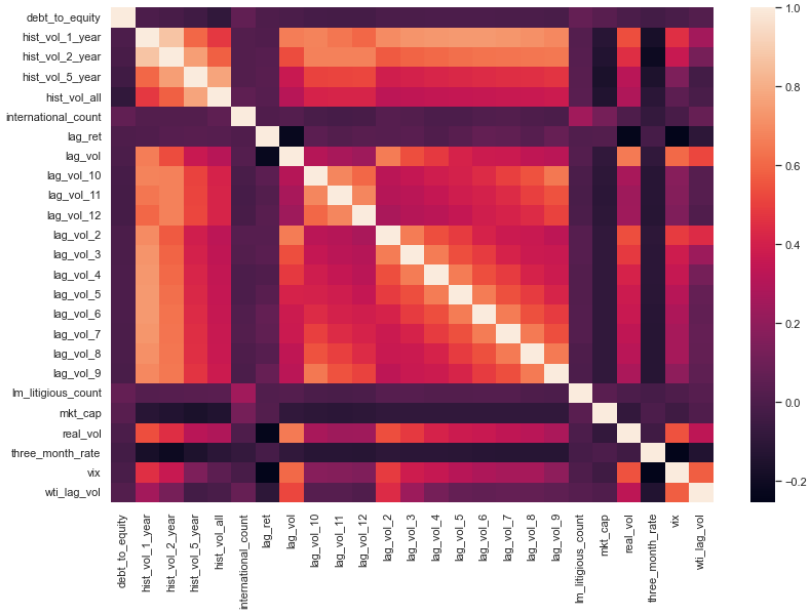

Figure 2: Correlation Matrix

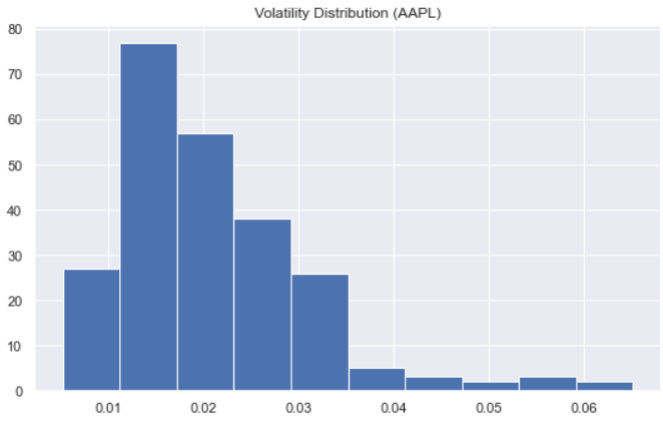

Figure 3: Volatility Distribution for Apple

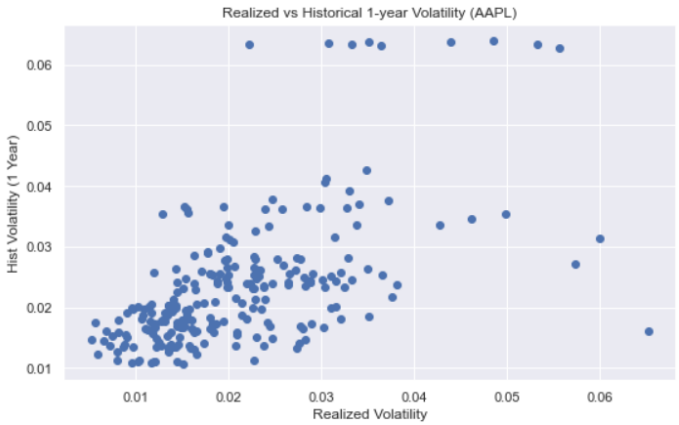

Figure 4: Realized vs Historical 1-year Volatility for Apple

From the distribution in Fig 4 above, average daily volatility is mostly around 1-3%. From Figure 4, there also seems to be a somewhat linear relationship between realized and historical 1-year average daily volatility per month. Thus, considering that volatility is mean-reverting and skewed, we can apply an autoregressive conditionally heteroskedastic time-series model along with the usual cross-sectional regression and ensemble of decision tree models.

In-Sample Data Analysis

Before diving into the out-of-sample modeling process where the primary goal is predictive accuracy, we would like to show the results of several in-sample OLS models, as this allows us to get a strong understanding for what features have historically best explained volatility over our time frame where we have complete data (2005-2020).

We first examine two versions of the in-sample OLS regression with a cross-sectional framework.

1. $Monthly\ Vol_t = \beta_0 + \beta_1 MonthlyVol_{t-1}$
    a. This model has an R^2 of .423, which does a surprisingly good job at explaining the data. See appendix Figure 2.

2. Model 2: Complete Model - This model uses all the variables listed in our variables section above.
    a. See appendix Figure 3. The model achieves an R^2 of .518
    b. In this model, we can observe the estimated coefficients of each of our variables to see if they align with our original expectations. All the variables in this model, are significant, likely due to the data set size, but most of the macro and fundamental variables do not appear to have any economic significance with only miniscule coefficients. Of note is that all variables, except for the volatility of WTI have the coefficient signs of what we expected.

We then felt it informative to build a separate model, using all predictor variables as in Model 2 above, for each company in our data set and then aggregate the individual results and present the summary statistics. The results are shown in appendix Figure 4. Worth pointing out, is that the average R^2 of the models is higher at .5846, than the R^2 of the total, cross-sectional model.

Modeling Process & Results: Out of Sample

For the out of sample prediction results, we used a total of 13 different models to generate volatility predictions. Note, some of the "models" such as historical two-year volatility, are not actually formal models, but they do represent what we think are reasonable approaches that one could take to estimating future volatility and thus, were included.

Model Overview

Note: The mode names presented below were chosen to match the names of the models in the code. Though these names are not the most descriptive, we think the consistency here makes the presentation of results more understandable when looking at the code output.

Time Series Models:

- **Hist_vol_1_year:** uses volatility of daily returns over the past year as an estimate for future volatility.
- **Hist_vol_2_year:** uses volatility of daily returns over the past 2 years as an estimate for future volatility.
- **Hist_vol_5_year:** uses volatility of daily returns over the past 5 years as an estimate for future volatility.
- **Hist_vol_all:** uses volatility of daily returns over the entire sample period of previous data.
- **Lag_vol:** predicts that next month's volatility will be the same as the previous month.
- **Ols_lag_vol_ts:** builds an AR(1) model where the only independent variable is the lag volatility.
- **Ols_lag_vol_12_ts:** builds an AR(12) model which uses the monthly lag volatilities for months 1-12.
- **Ols_max_ts:** OLS model is built using all of the lagged volatility terms 1-12, as well as all of the macro and fundamental variables.
- **Garch:** a garch(p=1, q=1) model is estimated using the daily return data available for the stock as of the prediction date.

Cross Sectional Models:

- **Ols_lag_vol_all:** this model only uses the lagged volatility as an independent variable in an ols estimation, but it uses all company data.
- **Ols_lag_vol_12_all:** this model uses the lagged volatilities for the past 12 months as independent variables in an ols estimation, but it uses all company data.
- **Ols_max_all:** OLS model, using all available data for all companies as of prediction date.
- **Rf:** random forest model, using all available data for all companies as of prediction date.

Results

Similar to the approach taken to evaluate the in sample results, we examine the results when lumping all of the predictions for all companies together and also, we explore the separated results on a company specific basis.

Overall Results

	R^2	RMSE	MAE	Max_Resid
hist_vol_1_year	0.165832	0.010278	0.005964	0.113221
hist_vol_2_year	0.18211	0.010178	0.00603	0.076095
hist_vol_5_year	-0.122957	0.011926	0.007638	0.060283
hist_vol_all	-0.385679	0.013248	0.010076	0.040459
lag_vol	0.047514	0.010983	0.006168	0.313851
ols_lag_vol_12_all	0.286933	0.009503	0.005309	0.169476
ols_lag_vol_12_ts	-0.104928	0.01183	0.005815	1.086638
ols_lag_vol_all	0.235211	0.009842	0.005681	0.23133
ols_lag_vol_ts	0.266157	0.009641	0.005539	0.185784
ols_max_all	0.348274	0.009085	0.005087	0.157701
ols_max_ts	-0.215182	0.012406	0.006332	1.116691
rf	0.008384	0.011207	0.005508	0.169449
garch	-91623.487304	3.406502	0.032798	590.834293

Table 2: Summarized Results for All Models

The models incorporating data from all companies (cross-sectional) perform better out-of-sample compared to individual model-based models. A simple OLS run on all data points (ols_max_all) tends to perform the best when predicting volatility with an R^2 of 34.8%. A regression with the last-12-month lagged volatilities gives us an R^2 of 28.7%, indicating that most of the predicted volatility comes from lagged variables itself. The standard GARCH model, surprisingly, provides absurd results with negative R^2 values.

Individual Results

To further evaluate our predictions, we calculate the R^2 values for each company and evaluate the summarized results for the same. For our predictions, we only include companies for which the number of data points is greater than 30. See below for summary statistics:

metric	MAE	Max_Resid	RMSE	R^2
garch	0.0324	2.6151	0.2572	-230634.2793
hist_vol_1_year	0.0061	0.0209	0.0099	-0.0262
hist_vol_2_year	0.0061	0.0156	0.0097	0.0064
hist_vol_5_year	0.0076	0.0178	0.0113	-0.3917
hist_vol_all	0.0102	0.0175	0.0126	-0.8477
lag_vol	0.0063	0.0425	0.0105	-0.1812
n	109.0973	109.0973	109.0973	109.0973
ols_lag_vol_12_all	0.0054	0.0205	0.0091	0.1057
ols_lag_vol_12_ts	0.0060	0.0285	0.0100	-0.1198
ols_lag_vol_all	0.0058	0.0263	0.0094	0.0421
ols_lag_vol_ts	0.0056	0.0239	0.0093	0.0863
ols_max_all	0.0052	0.0181	0.0087	0.1856
ols_max_ts	0.0065	0.0348	0.0104	-0.2542
rf	0.0056	0.0495	0.0107	-0.4761

Table 3: Summarized Results for Individual Models

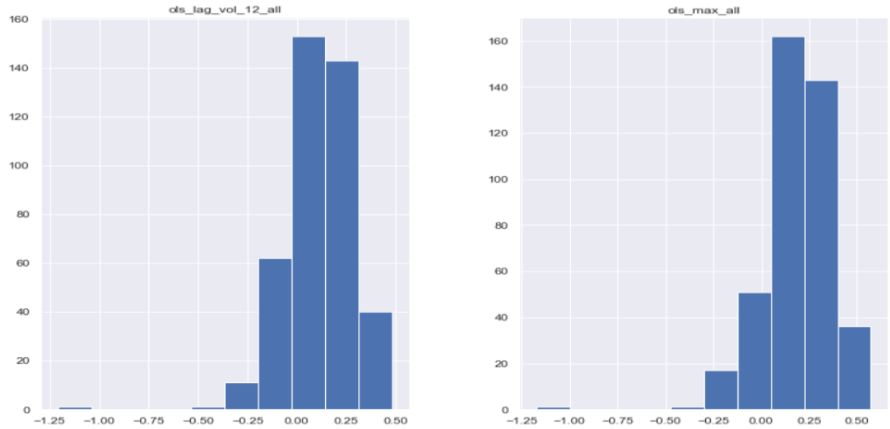

Figure 6: R^2 Distribution for OLS Models

The same two OLS models incorporating all variables perform better than all others. Figure 6 above shows the distribution of R^2. The values are above 30% for around a quarter of the companies. See Figure 7 below for an industry distribution comparison between companies we predict well (R^2 greater than or equal to 30%) and the companies we don't (R^2 less than 30%).

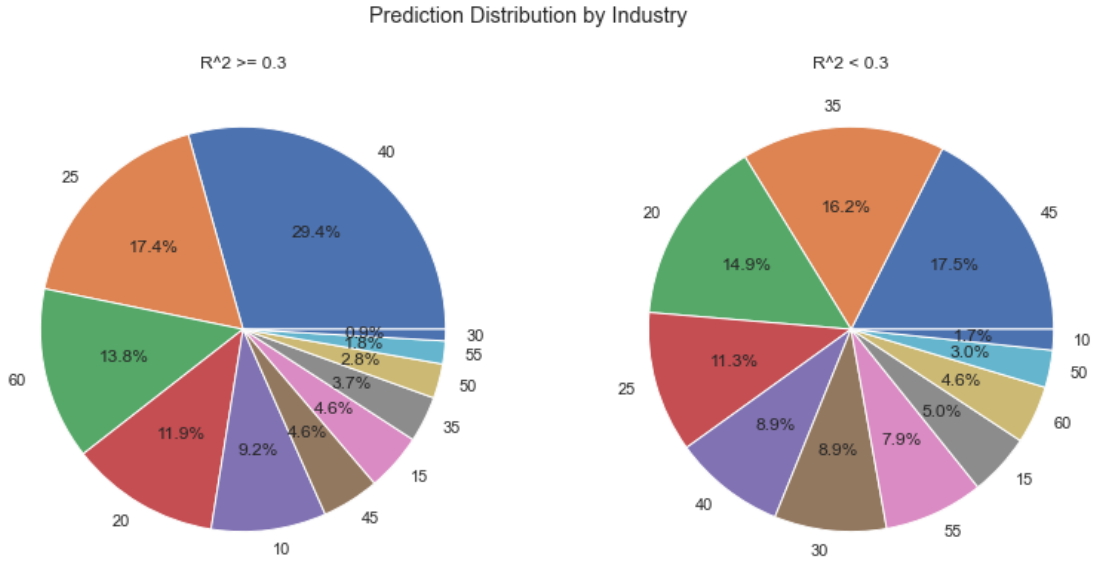

Figure 7: Prediction Distribution by Industry (gsector)

The companies our models predict well for are mainly distributed across Financials, Consumer Discretionary, and Real Estate. Similarly, our models do not do justice to Industrials, Healthcare, and Financial Technology. This brings up the interesting notion that perhaps, we should incorporate these statistics into our trading strategy by only trading in the sectors where our prediction accuracy is highest.

Portfolio Strategy

Once we have made our predictions for next month's volatility, we can implement multiple trading strategies. Our group focused on two strategies. The first strategy operates under the assumption that investors are disproportionately afraid of downside risk. Therefore, securities that are too volatile will be underpriced. We can achieve alpha by buying high predicted volatility stocks and shorting lower volatility stocks. At the end of each month, we predict next month's volatility. On that day we will sort and form our long-short portfolio, holding this portfolio for a month.

Below is a summary of our results:

Strategy 1	**CAPM**	**Fama - French**
Beta	0.92	0.73
Alpha	-0.005	-0.002
T-Stat	-1.5	-0.52
Sharpe Ratio	0.36	

Table 4: Summarized Portfolio Results

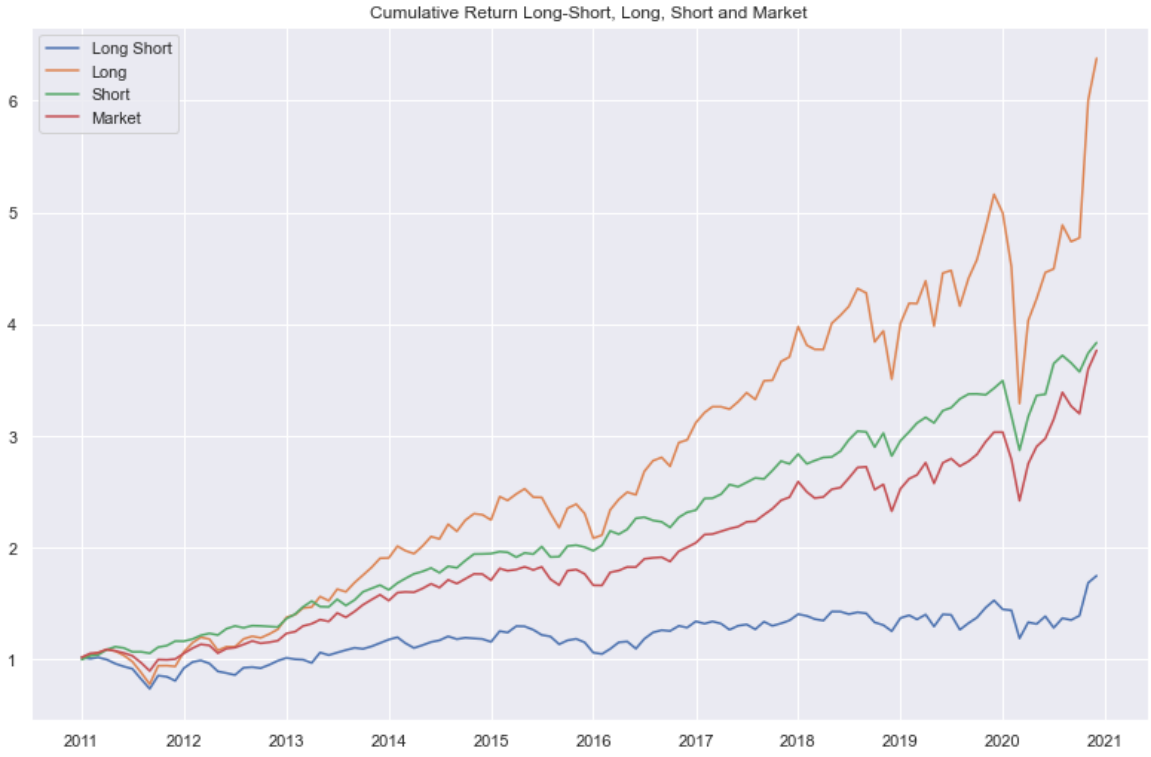

Figure 8: Cumulative Portfolio Returns

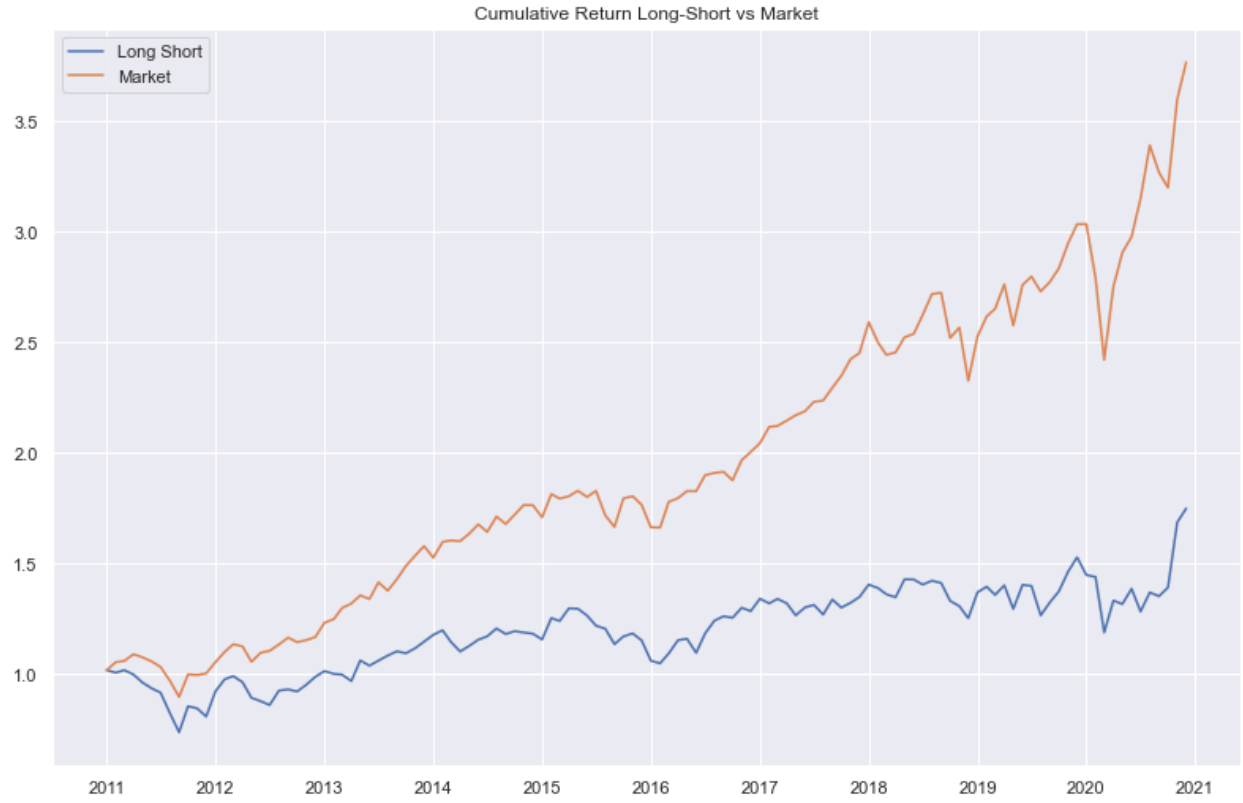

Figure 9: Long-Short Portfolio vs Market

Strategy 1 did not produce significant alpha either positively or negatively. However, the Beta was very high for a long-short portfolio. After investigating this, we discovered that our long portfolio has a beta of around 1.4 and our short portfolio has a beta around 0.5. This explains the 0.9 beta for the long-short portfolio. Strategy 1 longs stocks that have a high predicted volatility, which is highly correlated with high past volatility. Strategy 1 is in effect longing high beta S&P 500 stocks and shorting low beta S&P 500 stocks. Unsurprisingly this averages out to close to 1.

Our second theory is that investors are disproportionately afraid of previously volatile stocks, but the fear will correct more quickly. Therefore, as a stock becomes less volatile, investors become less 'afraid' of it. Our target variable to sort on, is previous periods volatility minus predicted volatility. To demonstrate this visually, we have included a hypothetical example of a stock's daily returns below:

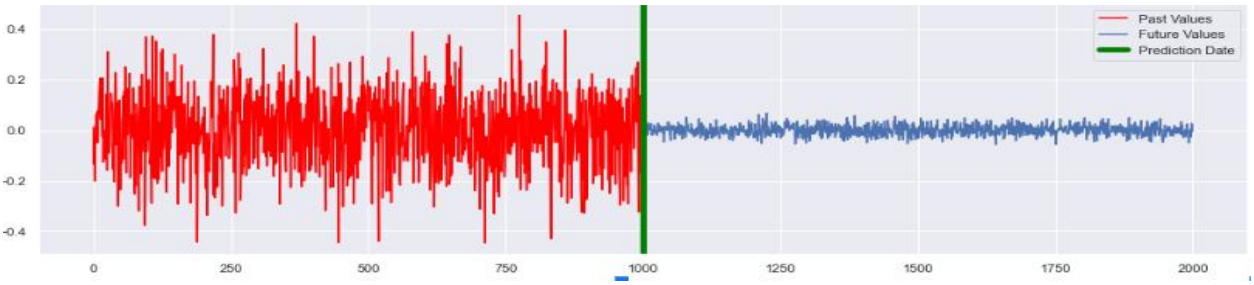

Figure 10: Hypothetical Example of Stock's Daily Returns

The previous month is extremely volatile, which we theorize would scare investors away. The green line represents our prediction date, and the blue line represents our prediction. The securities that have the biggest drop from previous volatility to predicted volatility, would be the securities that we buy. Similar to the last strategy, we will make a prediction for the next month at the end of each month and hold the long-short portfolio for one month.

Below is a summary of the results of this strategy:

Strategy 2	**CAPM**	**Fama - French**
Beta	0.27	0.22
Alpha	-0.006	-0.005
T-Stat	-2.9	-2.6
Sharpe Ratio	-0.36	

Table 5: Summarized Strategy Results

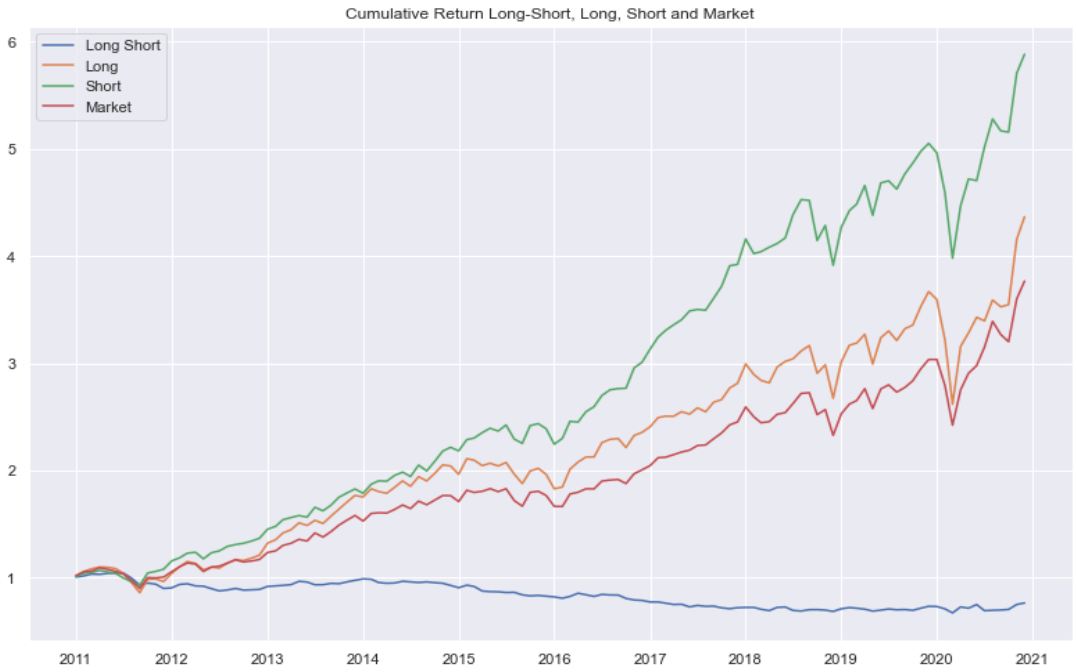

Figure 11: Cumulative Returns

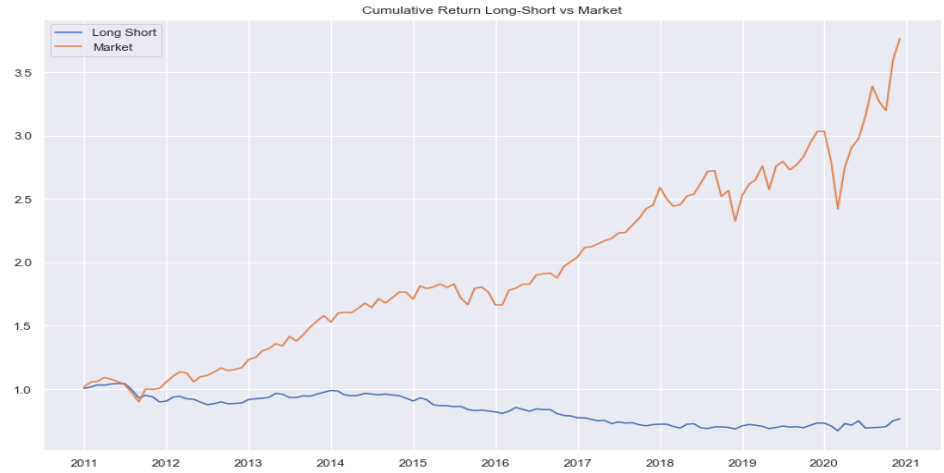

Figure 12: Cumulative Long-Short Returns

While the results for both portfolio theories were opposite our hypothesis, we have produced significant alpha for strategy 2, even though it is negative. To achieve a positive alpha, we would simply flip the long and short positions.

We thought about why our theory for strategy 2 was backwards. We thought that highly volatile stocks would scare investors away, and we could benefit from this by predicting securities' 'volatility drop'. In actuality, the securities that had a volatility increase saw the best returns. We determined that this might be a result of the investors who are attracted to the 'get rich quick stocks'. A company who has seen a drastic increase in volatility, must have had something occur, good or bad. Since these companies are seeing rapid fluctuations in price, this might attract more investors' eyes and increase returns for that specific period.

Our second thought was that our initial theory that investors are too scared of downside risk is backwards. Investors might be too interested in upside volatility. That is why strategy 2's long position which favors volatility decreases, performed worse than its short counterpart. The market is more interested in the upside of volatility, so trading on a volatility increase might lead to better returns.

Lastly, it is important to remember our sample of stocks. We are looking at S&P 500 companies for this project. These companies are highly successful and for the most part would be considered 'Safe' or 'Winners'. Therefore, investors might be looking for volatility increases for these companies since it is more likely than not, the volatility will work in their favor. It would be interesting to conduct a similar study on smaller companies, where risk of bankruptcy is more realistic.

Conclusion

Accurately capturing the volatility of volatility is no simple task. There are many sensible forecasting choices and modeling types to choose when trying to forecast. We found, after much experimentation, that the best overall results are achieved by using a cross-sectional model that utilizes all available data at any point in time. The primary reason for this model's success is because the data set size engenders a robustness to the model parameters. On any given company, there is likely an "optimal" model choice that is not the cross-sectional model, but it is difficult to choose that model in a systematic way, without foresight.

The portfolios we built were unsuccessful in generating statistically significant alpha as we expected, but a future opportunity for research may be to only include companies in our investable universe when prediction confidence is above a certain threshold or for certain industries. Additionally, our initial goal was to create an estimate of when people are fearful and greedy. Perhaps volatility is too blunt an estimate of market sentiment and we can get market sentiment directly in other ways.

Appendix:

Figure 1: DF_X Sample:

This data frame is transposed, so that it will fit on the page. Each column is a data point, and the rows represent the variables.

	0	1	2	3	4
year	2005.000	2005.000	2005.000	2005.000	2005.000
month	8.000	8.000	8.000	8.000	8.000
gvkey	2184.000	4058.000	5047.000	5125.000	7146.000
debt_to_equity	0.568	0.477	0.825	0.421	0.618
hist_vol_1_year	0.018	0.013	0.009	0.011	0.012
hist_vol_2_year	0.020	0.014	0.010	0.010	0.010
hist_vol_5_year	0.038	0.019	0.020	0.013	0.014
hist_vol_all	0.038	0.019	0.020	0.013	0.014
international_count	186.000	23.000	6.000	9.000	57.000
lag_ret	0.111	0.126	-0.004	0.108	0.062
lag_vol	0.017	0.013	0.008	0.010	0.010
lag_vol_10	0.011	0.015	0.009	0.022	0.009
lag_vol_11	0.017	0.011	0.009	0.007	0.014
lag_vol_12	0.017	0.011	0.011	0.010	0.011
lag_vol_2	0.031	0.009	0.009	0.008	0.010
lag_vol_3	0.013	0.010	0.008	0.007	0.010
lag_vol_4	0.019	0.018	0.010	0.010	0.007
lag_vol_5	0.013	0.011	0.010	0.009	0.017
lag_vol_6	0.018	0.014	0.006	0.015	0.009
lag_vol_7	0.017	0.017	0.009	0.008	0.018
lag_vol_8	0.016	0.010	0.009	0.009	0.009
lag_vol_9	0.017	0.009	0.009	0.010	0.011
lm_litigious_count	260.000	42.000	32.000	25.000	76.000
mkt_cap	25150.997	8405.817	365904.240	7988.066	4656.659
real_vol	0.017	0.011	0.008	0.010	0.011
three_month_rate	3.340	3.340	3.340	3.340	3.340
vix	11.570	11.570	11.570	11.570	11.570
wti_lag_vol	0.018	0.018	0.018	0.018	0.018

Figure 2: In-sample Lagged Variable Model Results

OLS Regression Results

Dep. Variable:	real_vol	R-squared:	0.423
Model:	OLS	Adj. R-squared:	0.423
Method:	Least Squares	F-statistic:	5.320e+04
Date:	Sun, 15 May 2022	Prob (F-statistic):	0.00
Time:	14:27:27	Log-Likelihood:	2.2943e+05
No. Observations:	72547	AIC:	-4.589e+05
Df Residuals:	72545	BIC:	-4.588e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0064	6.38e-05	99.640	0.000	0.006	0.006
lag_vol	0.6500	0.003	230.656	0.000	0.644	0.656

Omnibus:	75730.630	Durbin-Watson:	1.174
Prob(Omnibus):	0.000	Jarque-Bera (JB):	22724775.483
Skew:	4.688	Prob(JB):	0.00
Kurtosis:	89.197	Cond. No.	74.1

Figure 3: In-Sample All Variables Model Results

OLS Regression Results

Dep. Variable:	real_vol	R-squared:	0.518
Model:	OLS	Adj. R-squared:	0.518
Method:	Least Squares	F-statistic:	3897.
Date:	Sun, 15 May 2022	Prob (F-statistic):	0.00
Time:	14:28:16	Log-Likelihood:	2.3595e+05
No. Observations:	72547	AIC:	-4.719e+05
Df Residuals:	72526	BIC:	-4.717e+05
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.329e-05	0.000	-0.038	0.969	-0.001	0.001
lag_ret	-0.0181	0.000	-44.096	0.000	-0.019	-0.017
wti_lag_vol	-0.0566	0.003	-22.208	0.000	-0.062	-0.052
np.log(mkt_cap)	-0.0001	3.16e-05	-3.787	0.000	-0.000	-5.77e-05
vix	0.0004	5.84e-06	64.046	0.000	0.000	0.000
three_month_rate	0.0007	2.55e-05	28.235	0.000	0.001	0.001
debt_to_equity	0.0003	0.000	2.542	0.011	7.92e-05	0.001
international_count	8.957e-07	2.63e-07	3.409	0.001	3.81e-07	1.41e-06
lm_litigious_count	-1.562e-07	3.1e-08	-5.033	0.000	-2.17e-07	-9.53e-08
lag_vol	0.3412	0.004	81.608	0.000	0.333	0.349
lag_vol_2	0.1357	0.004	32.822	0.000	0.128	0.144
lag_vol_3	0.1182	0.004	29.818	0.000	0.110	0.126
lag_vol_4	-0.0337	0.004	-8.445	0.000	-0.042	-0.026
lag_vol_5	4.701e-05	0.004	0.012	0.991	-0.008	0.008
lag_vol_6	0.0648	0.004	16.308	0.000	0.057	0.073
lag_vol_7	0.0170	0.004	4.281	0.000	0.009	0.025
lag_vol_8	0.0166	0.004	4.187	0.000	0.009	0.024
lag_vol_9	0.0055	0.004	1.370	0.171	-0.002	0.013
lag_vol_10	0.0072	0.004	1.607	0.108	-0.002	0.016
lag_vol_11	-0.0425	0.004	-9.457	0.000	-0.051	-0.034
lag_vol_12	0.0617	0.004	15.056	0.000	0.054	0.070

Omnibus:	84833.092	Durbin-Watson:	1.204
Prob(Omnibus):	0.000	Jarque-Bera (JB):	37136285.294
Skew:	5.668	Prob(JB):	0.00
Kurtosis:	113.259	Cond. No.	2.14e+05

Figure 4: In Sample OLS

The results presented below are based on the summary statistics of all the models estimated separately. The variables have their coefficients aggregated, and the metrics of RMSE and R^2 are also averaged.

	mean	median	std
lag_ret	-0.0085	-0.0090	0.0264
wti_lag_vol	-0.0325	-0.0194	0.1233
np.log(mkt_cap)	-0.0001	0.0004	0.0177
vix	0.0009	0.0008	0.0008
three_month_rate	0.0005	0.0007	0.0078
debt_to_equity	0.0299	0.0099	0.1770
international_count	-0.0000	-0.0000	0.0002
lm_litigious_count	0.0000	-0.0000	0.0001
lag_vol	-0.0004	0.0057	0.3796
lag_vol_2	-0.0196	-0.0166	0.1767
lag_vol_3	0.0604	0.0665	0.1434
lag_vol_4	-0.1138	-0.1026	0.1504
lag_vol_5	-0.0439	-0.0248	0.1657
lag_vol_6	-0.0135	0.0013	0.1705
lag_vol_7	-0.0138	-0.0018	0.1307
lag_vol_8	-0.0779	-0.0459	0.2601
lag_vol_9	-0.0012	0.0027	0.1824
lag_vol_10	0.0047	0.0091	0.1482
lag_vol_11	-0.0929	-0.0627	0.3619
lag_vol_12	-0.0380	-0.0136	0.3383
R^2	0.5846	0.5679	0.1601
RMSE	0.0080	0.0074	0.0032
n	155.3469	175.0000	47.0598

Figure 5: Litigious Word Dictionary (Sample)

LITIGATIONS	DELEGABLE	THEREAFTER	SUBPOENAED	ESCHEATED	AMENDS
INDICTABLE	REBUTS	INDEMNITORS	DISTRIBUTEES	USURIOUS	WHENSOEVER
CROSSCLAIM	APPEALABLE	PLAINTIFFS	ARBITRATED	RECOUPMENT	SUBLEASEE
SUBDOCKET	CONVICTING	NONFORFEITABLE	CONTRACTUAL	ADJUDICATORS	CONVEYANCE
ENDORSEE	ABSOLVING	USURPATION	PREJUDICE	PECUNIARILY	INCHOATE
REGULATIONS	HEREBY	NOTARY	REGULATED	ARREARAGES	INDICTABLE
THENCEFORTH	OVERRULED	EVIDENTIAL	SETTLEMENT	CLAIMS	UNDISCHARGED
CONTRAVENING	ATTESTATIONS	CODIFICATIONS	APPURTENANCES	INSOFAR	ENFORCEABLE
NOTARIES	ENCUMBRANCE	OBLIGEE	CRIMINALLY	CONVICTION	INDEMNITOR
LITIGATING	CONSENTING	AGGRIEVED	INTERROGATES	JUDICIALLY	DISTRIBUTEES
USURY	CHOATE	NONFIDUCIARY	CONSENTED	LAWFULNESS	DOCKETS
CONSENTING	COURT	BREACHES	PLEADINGS	APPEAL	DECLARANT
ASCENDANT	PREJUDICIAL	UNDEFEASED	DEROGATING	VENDEE	UNAPPEALED
DULY	HEREON	COUNTERSUIT	PROSECUTOR	RECOURSES	RULINGS
PARI	INCAPACITY	POSTCLOSURE	ESTOPPEL	ENCUMBERED	INTERROGATES
INDEMNITOR	TORTIOUSLY	OBLIGEE	DOCKET	ATTESTED	EXECUTORS
APPOINTOR	INTERROGATORIES	ADJUDGED	ADJOURNS	CONTRAVENTIONS	DECREE
CODIFIES	IMPLEADED	SEVERANCES	DISPOSSESSION	NULLIFICATION	DELEGEES
REGULATE	THENCEFORTH	PRORATA	NOTARIAL	JURISDICTIONAL	PROSECUTE
MEDIATOR	FORWHICH	REQUESTOR	LICENSABLE	ATTEST	ASSIGNATION
SENTENCING	CONTRACTIBLE	AFOREDESCRIBED	THEREON	CESSION	PERPETRATING
INTERROGATOR	OBLIGORS	INTERROGATION	AFORESTATED	CONSTITUTION	PLEDGORS
LAWFULNESS	WARRANTOR	NONFORFEITABILITY	APPELLATE	JURISDICTION	ACQUITTING