**A Machine Learning Approach to Pricing High Grade Corporate Bonds**

Alexander Y. Sedgwick

**Abstract:**

As opportunities to generate a return in financial markets have decreased many institutional investors such as pension funds and mutual funds are now turning to managing their costs as a way to boost performance. Equity investors have many tools available for measuring trading costs however fixed income traders lack such robust tools to managing their execution costs. The primary reason for this is the structural differences between the equity and bond markets and the simple fact that bonds don’t trade every day like stocks. This creates many problems when trying to establish benchmarks to measure execution against. By analyzing over half a million trades from 2013 we will show how a basic machine learning approach can significantly improve on existing vendor-based bond pricing and provide the foundation for a more comprehensive analysis of fixed income trading costs.

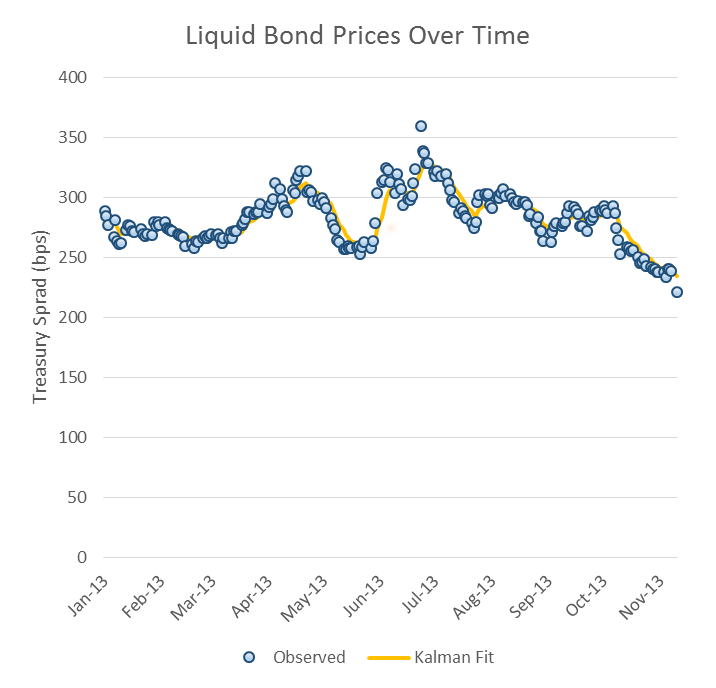
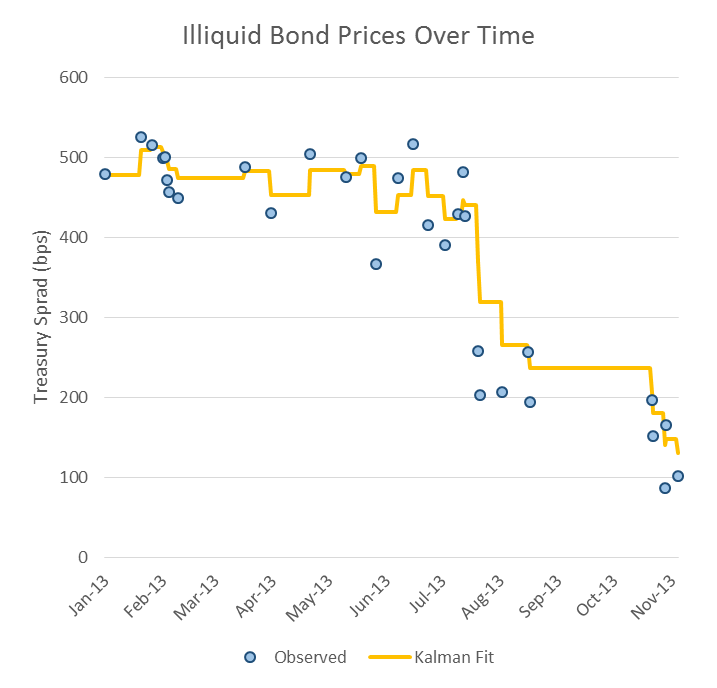
**Introduction:**

2013 has been a difficult investment environment for fixed income. The continued strength of the stock market, along with the threat of rising interest rates (which adversely affect bond prices) culminated in a large sell-off in May and June that pushed returns negative for the year. With the Federal Reserve pushing the possibility of ending their quantitative easing program to early 2014, the road ahead looks more ominous. These threats should be taken very seriously considering the rapid approach of retirement for many in the baby boomer generation – a demographic group who will increasingly rely on fixed-income investments to fund their retirement.

In responses to these trends many money managers have turned their attention to managing their costs and evaluating the quality of their trades.

For equity traders there are number of vendors who over comprehensive analytics products that measure transaction costs and execution quality, the most well known being ITG. For a stock trade this analysis is pretty straight forward, each trade has a commission as well as measures for delay cost and market impact (the change in price between when the strategy was implemented as well as a change in the stock’s price as a result of the order). Additionally, the high frequency with which stocks trade provides a large amount of observable market data to compare trades to.

This contrasts sharply with the bond market. Unlike the stock market, which has about 7,000 listed stocks, there are approximately 34,000 tradable bonds. In addition, of those 34,000 securities only 5,000 or so trade on any given day. This creates a significant problem for investors who want to measure their execution costs as they frequently can’t find an accurate benchmark. To illustrate this issue Figure 1 compares observable prices for a bond that traded almost every day (Liquid) in 2013 with one that has only traded sporadically (Illiquid). I’ve used R to generate a fit for both sets of bond prices over time using a Kalman filter - however even in the case of the illiquid bond the implied prices are choppy.



To combat the problem a number of companies have created “evaluated” pricing services that price every tradable bond on a daily basis. Companies like Interactive Data originally catered to mutual funds who had to price their portfolios daily. Now IDC as well as Elkins McSherry, Global Trading Analytics and MarkIt have begun applying evaluated pricing to trade evaluation. Unfortunately, many of these companies refuse to disclose their pricing methodologies and traders frequently point out that a large number of the prices are off-market.

Most recently 2011 saw the launch of Benchmark Solutions, a Warburg backed start up that priced several thousand bonds in 10 second increments throughout the day. The firm even launched a Kaggle competition to predict the next price that a corporate bond may trade at. Unfortunately operating costs consumed the firm and it shutdown in March 2013.

**Methodology:**

I began with a very modest goal, to predict the average treasury spread for a bond on a given day and to see whether I could improve upon the prices that a leading evaluated pricing service provided for a specific day.

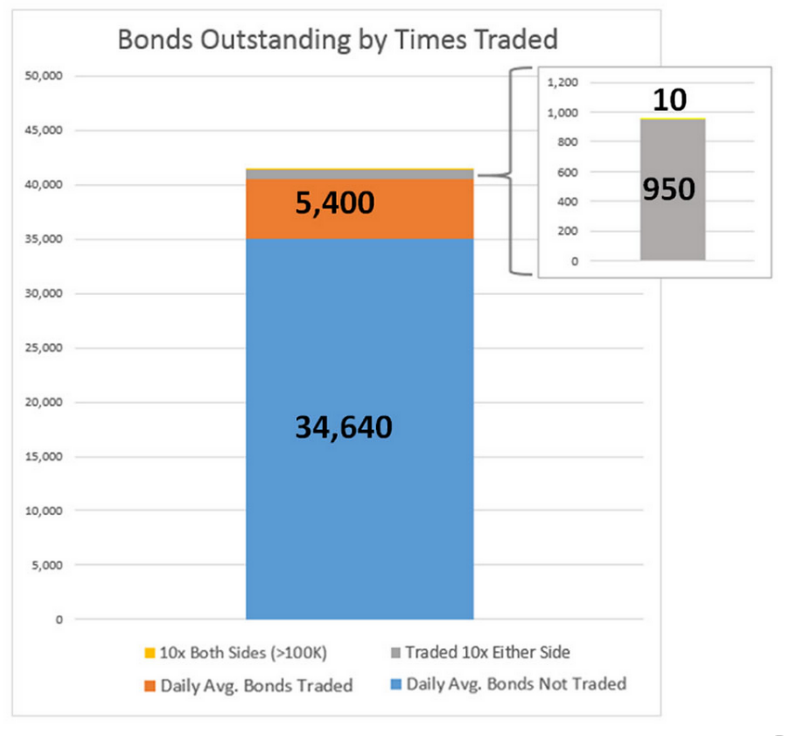
\**A quick note on how bonds are traded/priced. High grade bonds are traded on the basis of “spread”. Simply this is a measure of the additional yield the issuer must pay over the risk free rate to compensate an investor for the default risk of the company.*

*Consider this, if the US Government is currently paying interest of 2.78% on a 10-year bond and there is theoretically no chance the government would default (an assumption that has only very recently become contentious) then a company like Apple would have to pay more than 2.78% since they could potentially default. In reality they pay about .91% more than the 10 year US Treasury note. So…when we talk about Apple bonds we say they trade at 91 basis points (.91%=91 bps). Throughout this paper I will use price and spread interchangeably when referring to this convention.*

**Dataset description:**

The first step was to lay out the various datasets that I was going to use for inputs. The most important dataset to have is FINRA’s TRACE dataset which includes trade details for all dollar denominated bond trades done in the US – it effectively covers 99% of the high grade corporate bond market.

Summary Statistics:

* 28,369 trades per day on average
* $12.9 billion traded per day on average
* An average of 4,650 high grade bonds traded per day in 2013 down from 2012 (below)

The daily trading activity provides us with a useful dataset for a supervised learning project. Ultimately it provides us with useful information like the trade sizes, buys vs. sells and descriptive information on the bond for our inputs. It also provides us with executed prices which we can use to evaluate the accuracy of our pricing vs. the market price and the IDC provided evaluated price.

In addition to the trade prices MarketAxess also receives dealer pricing runs. This is where large banks provide pricing schedules that show where they will buy or sell a given bond. However these levels are indicative and as such frequently when one calls on what appears to be an attractive price a dealer will change the level offered for execution (almost always to a less attractive price). As such, these prices, while they provide a useful pricing guide are rarely reliable on their own.

**Bond Pricing Determinants:**

Another distinguishing feature of bonds, relative to equities, is that each bond represents a “loan” made by the company with a unique set of features such as the interest rate paid and the time the company will take to repay the loan. For example, bonds for the same company will likely pay more interest if they are maturing in 30 years vs. 2 years because the future solvency of the company is less clear further out into the future – hence there is a higher risk the company will default on their bonds. For example:

Age: Older bonds trade less frequently, they are more likely to have been bought by an insurance company or investor that will hold them to maturity

Maturity: As mentioned, long-maturity bonds will trade at a higher yield/lower price than short term bonds

Trade Size: Though not explicitly used in the model, if a trade is over $750,000 in size it gets better pricing because it is easily traded in the institutional market. However, if a trade is too large 50MM or more it may see a less attractive price as banks will have trouble trading it all in one large block.

Rating: Ratings serve as a rough guide to the default risk of various companies. The higher the risk the higher the required yield on the bond needed to attract investors.

Issue Size: Investors are very sensitive about liquidity – they prefer to hold more liquid bonds so that if they need to trade them quickly there is an active market for them. As such they will prefer to hold bonds that were part of a very large bond deal (issuance of a lot of bonds) since it is conducive to active trading. If an issuer only issued a small number of bonds with a particular set of characteristics, it is highly likely there isn’t an active market for them and they will be harder to trade.

**Economic Factors:**

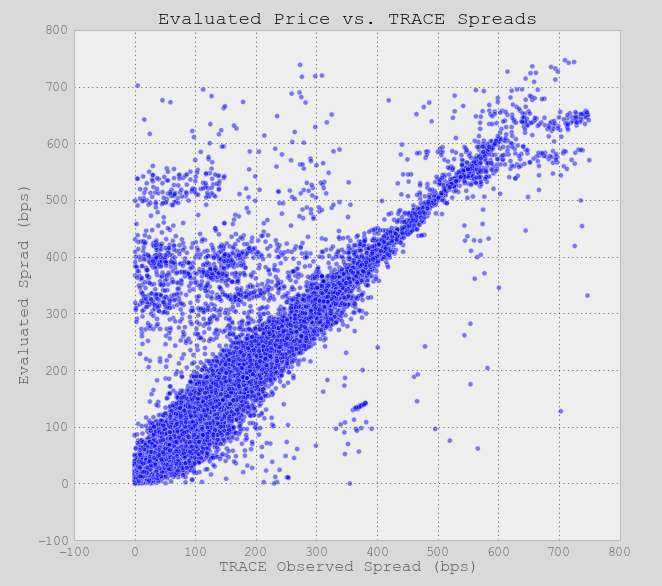
There are also a number of economic factors that affect bond prices. Most importantly is the perception of market risk (i.e. are companies likely to default in the current economic environment) and the general direction of bond prices. In order to incorporate these factors into our model I pulled historical pricing for the VIX or the CBOE Volatility Index (sometimes referred to as a “fear gauge”) which we can use as a proxy for risk perception among market participants. Bond index data is expensive to get, however number of providers have built Exchange Traded Fund products that aim to replicate index performance. Since they are exchange traded securities I pulled historical prices from Yahoo Finance for LQD (a liquid bond ETF) using the pandas datareader to scrape the website.

Finally, bond prices are sensitive to interest rates, as rates go up bond prices fall. In order to factor that into my model I pulled interest rates for the 2 year, 10 year and 30 year treasury bonds from the Federal Reserve (FRED) database – also using datareader.

**Benchmark:**

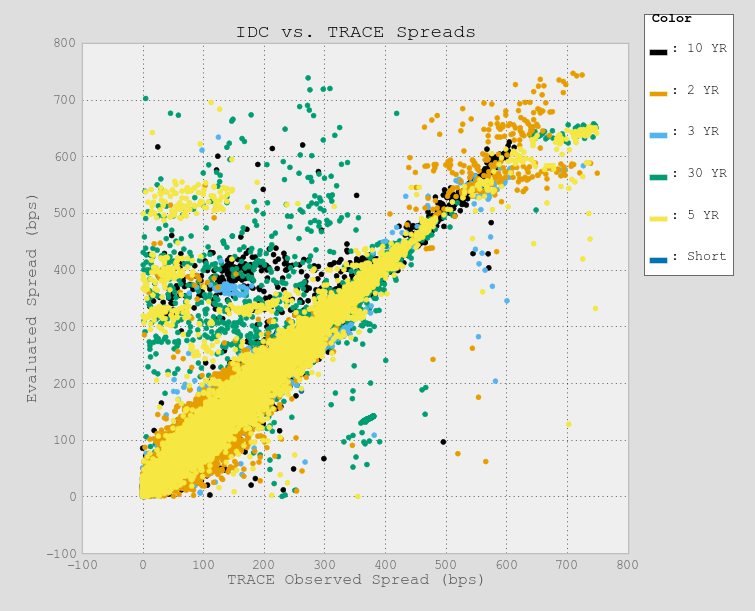
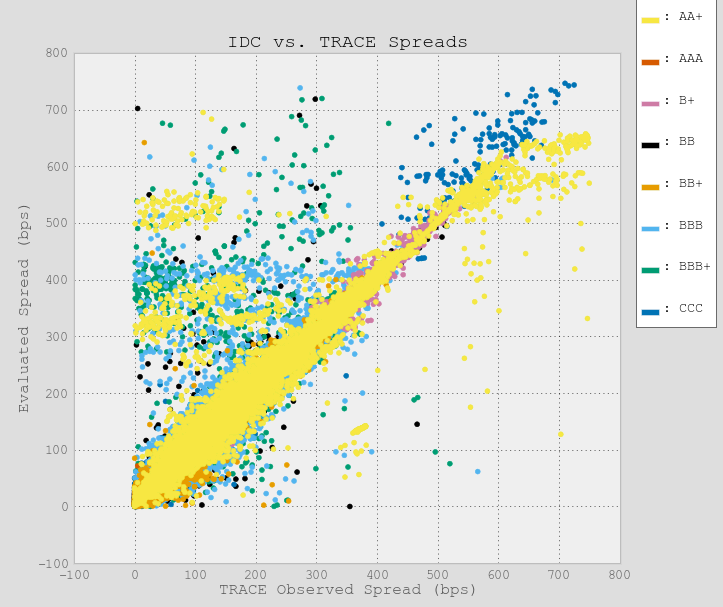
The benchmark in this case is bond price/yield data from the largest provider of evaluated pricing, Interactive Data Corp. IDC provides a bid-side quote for each bond at the end of the day that includes price and yield. Since I am focusing on treasury spreads I converted the IDC yield to an implied spread. This was accomplished by matching the IDC quoted yield to a benchmark yield and finding the difference (as noted earlier, the spread is simply the additional yield required over the risk free rate or treasury yield to attract investors).

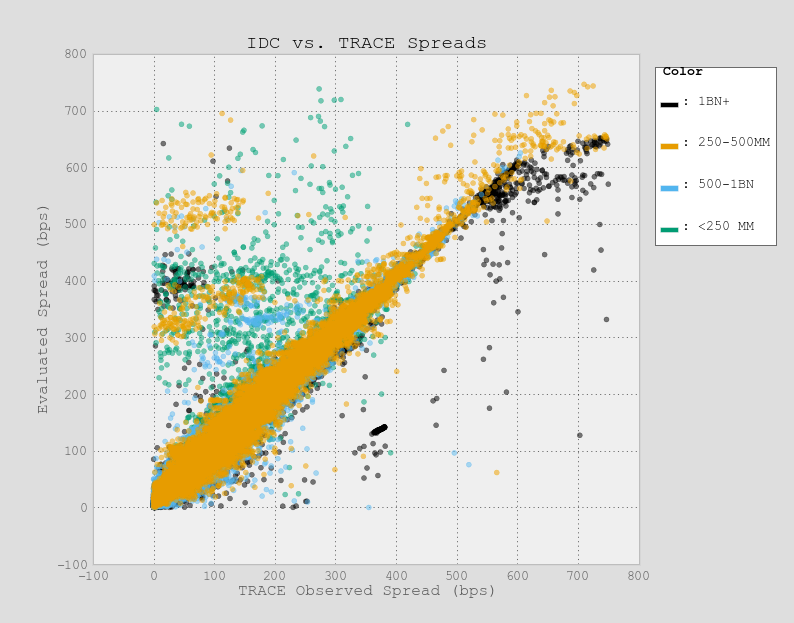
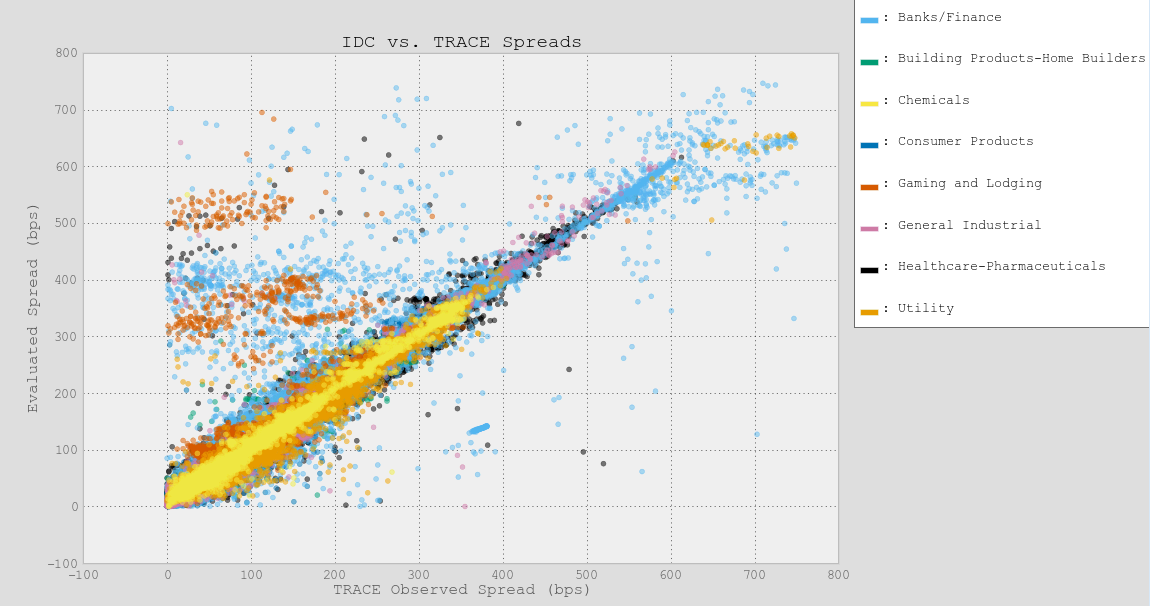
The first thing I did was try to get a sense of my benchmark, in other words, what was the relationship between the IDC pricing and corresponding TRACE price. So I did a simple scatter plot:



Probably the most notable item that stands out is that there are a large number of cases where the evaluated price is significantly higher than the market observed spread. Note the large “cloud” of observations where the evaluated spread is between 300-550 basis points and the bond is trading at less than 250. Calculating the MSE for their estimates I get 466.4 which approximates to an average error of 21 basis points. To put this in context the average dollar value of a basis point is $800 per $1 million traded so that would amount to a mispricing of $16,800 per 1 MM traded. Given the average trade size is about $500,000 for the bond market, this is a material number.

The next step in my analysis was to consider what were some of the variables that may be leading to the mispricings in the benchmark set of data. I ran several scatterplots grouping points by different criteria:

Several things stuck out in the cluster of mispriced data. First, there were a large number of 30 year bonds and bonds issued under 250MM in size. Additionally, the mis-priced bonds looked to be lower rated. Sector seemed to be less important, likely because Bank/Finance bonds make up a significant portion of the market. Based on my initial understand of factors that affect bond prices, as well as the output of these scatter plots I put together an initial dataset.

**To summarize we have the following inputs to our model:**

* Ticker (represents the issuer)
* Coupon
* Maturity
* Option Flag (does it have an embedded option)
* Issue Amount
* Duration (Maturity)
* Maturity Bucket
* Bond Age
* Trading Sector
* Market – Investment Grade, High Yield, Floating Rate
* Median Dealer Inventory Sell Spread
* Median Dealer Inventory Buy Spread
* LQD (Proxy for HG Liquid Bond Indices)
* VIX (Proxy for perception of Market Risk)
* 2, 10, 30 Year US Treasury Rates (Proxy for interest rates and interest rate views)

**Dependent Variable:** Volume(Trade Size) Weighted Average Spread for the bond on a given day

**Set-Up and Analysis:**

The most time consuming aspect of my project was cleaning and prepping the data. I began by pulling using pulling the data out of a SQL database and starting to clean it. There were a couple of items that I wanted to approach first. There were several variables that we needed to consolidate or adjust:

Trading Sector: I removed Unassigned bonds, Investment Grade Asian Bonds and Latin American bonds. Unassigned bonds don’t have a sector and are few and far between, likely a chance for some data clean up. Asian and LATAM bonds both trade on price along with other Emerging Market Bonds and would likely skew or analysis since we are looking at High Grade bonds that trade on Spread

Rating: S&P has Approximately 25 different rating levels, these could easily be cut in half by looking at broader groupings that eliminated the “+” or “-“ associated with each rating.

Issue Amount: The market is widely viewed in terms of bonds that are issued in sizes less than $250MM, 250MM-1BN and 1BN+. I went ahead and added two intermediate buckets for the sake of granularity (250-500MM and 500M-1BN).

Maturity: While I retrieved the information from the database it is worth noting that I broke bonds down into groups based on maturity. As an industry standard bonds are priced (that is the spread is calculated) to one of the most actively traded treasury bonds either the 2 YR, 3 YR, 5 YR, 10 or 30 YR. For the purposes of our analysis I grouped the bonds the same way.

I also grouped bonds by whether they were investment grade, high yield, floating rate notes (the interest rate changes) or bonds issued by Fannie Mae or Freddie Mac. For all of these fields I used the get\_dummies function to create the appropriate dummy variables.

Finally I trimmed the data set down even further. By and large bonds in the investment grade universe trade with a spread above 0 and below 750 basis points (where a basis point is 1/100th of percent). So I restricted the dataset to observations where both the observed spread and IDC spread was >0 and <=750 bps.

Once I had cleaned the data I ran several different models:

Benchmark: I ran the MSE for the evaluated prices vs. the observed prices in the market. The MSE was 466.4 which is equivalent to a mis-pricing of 21.6 bps. My goal is to do better than that.

Linear Regression:

I wanted to run a baseline model and just see how bad it could be. Using a Ridge Regression model I generated an MSE of 685.3 which translates to an error of 26.2 basis points.

Basic Decision Tree:

The next model that I selected was a basic decision tree model. Given the sophistication of the relationships between the data and the features I was not expecting a great result here, partly because, as has been noted in class – why would you use one tree if you can have a forest? Still, I think that one important aspect of the exercise is seeing how much I gain with different models. The MSE for this model was 244 or approximately 15.6 bps error.

Extra Tree Regressor:

The next model I moved to was an Extra Trees Regressor which yielded an MSE of 232.6 or 15.3 bps. This was a modest improvement over the basic decision tree, but I am looking for a more substantial drop.

Random Forest Regressor:

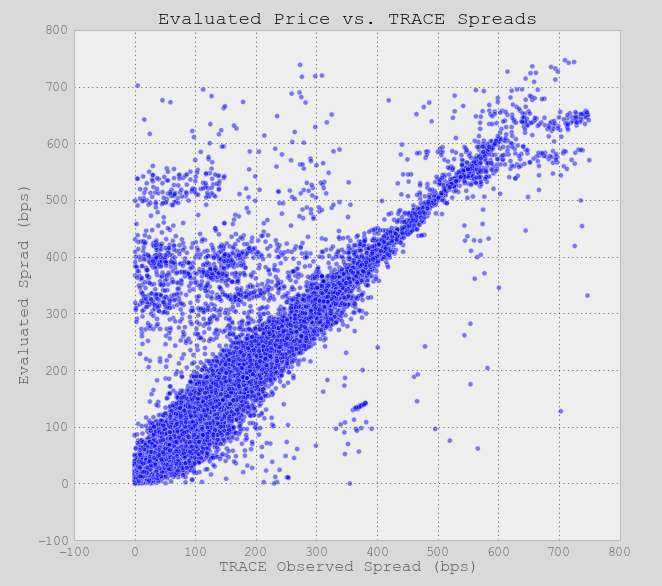
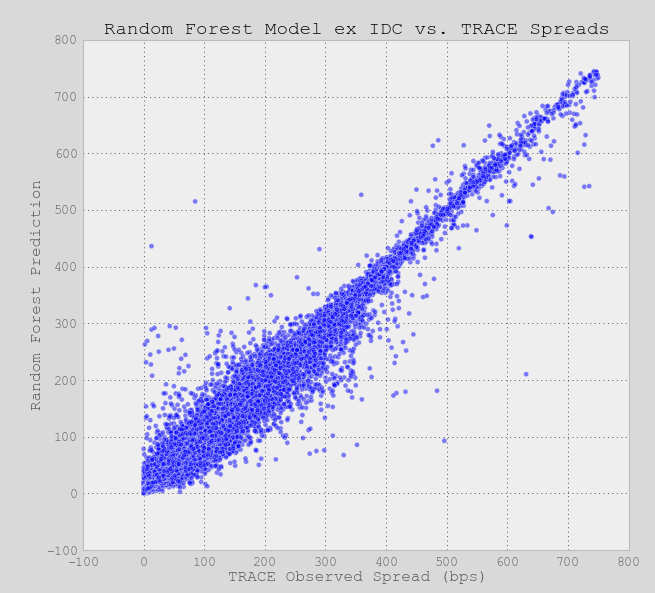
Finally, I ran two random forest models. Used all of my inputs except for the evaluated prices (the benchmark). This model resulted in an MSE of 160 or approximately 12.6 bps – almost half as much as my benchmark.

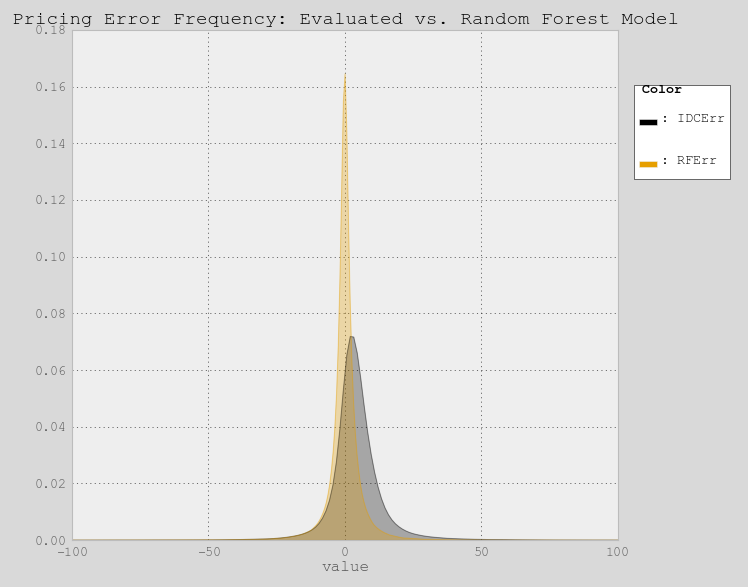
Gradient Boosting Regressor:

I was hoping that GBM would result in a better fit though it returned an MSE of 573 – significantly worse than all of the other models.

In the end I settled with a random forest model as my “winning predictor”

To investigate how my predicted levels compared with the benchmark I graphed a new scatterplot with my predicted levels side by side. We can see that there is no longer a cloud of mis-priced securities.

Next I also looked at the dispersion of errors to see whether my MSE was driven by a few large errors or a large number of small by consistent errors.

Based on the graph it appears as though the evaluated price errors show fat tails or tend to have large significant error rates. In contrast the Random Forest model has both a lower average error as well as a more consistent set of small errors. This indicates that it could be a reliable pricing model with little risk of severe mispricings.

**Conclusion and Future Research:**

I believe I have successfully shown how machine learning models can be used to improve existing estimates of bond prices. More importantly such an approach has a significant economic and social impact by allowing pension funds and mutual funds to have a reliable benchmark to compare their trades to. Over the long run, one would hope that this would drive down trading cost and reduce fees for retirement investment vehicles like mutual funds.

In terms of future research, there are a number of things that need further investigation. Most importantly I’ve completely ignored the prior trading activity of the bond. A good example of this approach is the Kaggle competition run by Benchmark Solutions.

The Kaggle competition used a dataset comprised of the a target price (observed) and a set of 10 prior trades along with the size and side of the trade. I actually intend on continuing my research in this area and so I began by pulling a sample of over 300,000 trades and running both a random forest and GBM model to see how that approach compared with my initial research.

The result of that very basic initial model showed an MSE of 353.6 or about 18.7 basis points. Though it only included trade data and ignored many of the bond characteristics I factored into the main models I used. I should point out that it was difficult to generate the data because I needed to run the model as for each trade I needed to pull the prior 10 trades as well. In short, the fact I could do so much better with a dataset that was easier to generate was quite impressive.

**Bibliography:**

Asquith, Paul and Au, Andrea S. and Covert, Thomas R. and Pathak, Parag, The Market for Borrowing Corporate Bonds (October 20, 2011). Available at SSRN: http://ssrn.com/abstract=1631260 or <http://dx.doi.org/10.2139/ssrn.1631260>

Asquith, Paul and Covert, Thomas R. and Pathak, Parag, The Effects of Mandatory Transparency in Financial Market Design: Evidence from the Corporate Bond Market (September 4, 2013). Available at SSRN: http://ssrn.com/abstract=2320623 or <http://dx.doi.org/10.2139/ssrn.2320623>

Bao, Jack and Pan, Jun and Wang, Jiang, Liquidity of Corporate Bonds (July 9, 2008). Available at SSRN: http://ssrn.com/abstract=1106852 or <http://dx.doi.org/10.2139/ssrn.1106852>

Bessembinder, Hendrik and Kahle, Kathleen M. and Maxwell, William F. and Xu, Danielle, Measuring Abnormal Bond Performance. Review of Financial Studies, Forthcoming. Available at SSRN: http://ssrn.com/abstract=650883 or <http://dx.doi.org/10.2139/ssrn.650883>

Bessembinder, Hendrik and Maxwell, William F. and Venkataraman, Kumar, Market Transparency, Liquidity Externalities, and Institutional Trading Costs in Corporate Bonds. Journal of Financial Economics, Forthcoming. Available at SSRN: <http://ssrn.com/abstract=827984>

Biais, Bruno and Green, Richard C., The Microstructure of the Bond Market in the 20th Century (August 1, 2007). Available at <http://repository.cmu.edu/cgi/viewcontent.cgi?article=1133&context=tepper>

Chakravarty, Sugato and Sarkar, Asani, Liquidity in U.S. Fixed Income Markets: A Comparison of the Bid-Ask Spread in Corporate, Government and Municipal Bond Markets (March 1999). FRB of New York Staff Report No. 73. Available at SSRN: http://ssrn.com/abstract=163139 or <http://dx.doi.org/10.2139/ssrn.163139>

Cici, Gjergji and Gibson, Scott and Merrick, John J., Missing the Marks: Dispersion in Corporate Bond Valuations Across Mutual Funds (July 14, 2010). Journal of Financial Economics, Forthcoming. Available at SSRN: http://ssrn.com/abstract=1104508 or <http://dx.doi.org/10.2139/ssrn.1104508>

Dick-Nielsen, Jens, Peter Feldhütter, and David Lando, 2012, Corporate Bond Liquidity Before and Af-ter the Onset of the Subprime Crisis, Journal of Financial Economics 103, 471-492.

Edwards, Amy K. and Harris, Lawrence and Piwowar, Michael S., Corporate Bond Market Transparency and Transaction Costs (September 21, 2004). Fifteenth Annual Utah Winter Finance Conference. Available at SSRN: http://ssrn.com/abstract=593823 or <http://dx.doi.org/10.2139/ssrn.593823>

Goldstein, Michael A. and Hotchkiss, Edith S. and Sirri, Erik R., Transparency and Liquidity: A Controlled Experiment on Corporate Bonds. Review of Financial Studies, Vol. 20, No. 2, pp. 235-273, March 2007. Available at SSRN: <http://ssrn.com/abstract=979069>

Hendershott, Terrence and Madhavan, Ananth, Click or Call? Auction Versus Search in the Over-the-Counter Market (September 4, 2012). Journal of Finance, Forthcoming. Available at SSRN: http://ssrn.com/abstract=1961350 or <http://dx.doi.org/10.2139/ssrn.1961350>

Jankowitsch, Rainer and Nashikkar, Amrut J. and Subrahmanyam, Marti G., Price Dispersion in OTC Markets: A New Measure of Liquidity (April 2008). EFA 2008 Athens Meetings Paper. Available at SSRN: http://ssrn.com/abstract=1100704 or <http://dx.doi.org/10.2139/ssrn.1100704>

Maxwell, William F. and Bessembinder, Hendrik, Transparency and the Corporate Bond Market. Journal of Economic Perspectives, 2008. Available at SSRN: <http://ssrn.com/abstract=1082459>