WP/19/292

IMF Working Paper

Completing the Market: Generating Shadow CDS Spreads by Machine Learning

by Nan Hu, Jian Li, Alexis Meyer-Cirkel

INTERNATIONAL MONETARY FUND

 WP/19/292

IMF Working Paper

Innovation Lab Unit

Completing the Market: Generating Shadow CDS Spreads by Machine Learning

Prepared by Nan Hu, Jian Li, Alexis Meyer-Cirkel

Authorized for distribution by Tristan Walker

December 2019

Abstract

We compared the predictive performance of a series of machine learning and traditional methods for monthly CDS spreads, using firms' accounting-based, market-based and macroeconomics variables for a time period of 2006 to 2016. We find that ensemble machine learning methods (Bagging, Gradient Boosting and Random Forest) strongly outperform other estimators, and Bagging particularly stands out in terms of accuracy. Traditional credit risk models using OLS techniques have the lowest out-of-sample prediction accuracy. The results suggest that the non-linear machine learning methods, especially the ensemble methods, add considerable value to existent credit risk prediction accuracy and enable CDS shadow pricing for companies missing those securities.

JEL Classification Numbers: C1, G1, G12, C52, C53

Keywords: Credit default swaps; Prediction; Machine Learning methods

Author's E-Mail Address: nan.hu@stud.uni-frankfurt.de; jianli.research@outlook.com; ameyercirkel@imf.org

I. INTRODUCTION

The Credit Default Swap (CDS) market has attracted considerable attention since its inception in the early 1990s. It has undergone a period of rapid growth and usage in the run-up to the 2008 Global Financial Crisis (GFC). Since then, the CDS market has experienced a cooling period as well as structural changes, but it still represents the third largest over-the-counter (OTC) derivatives market, with a gross market value of about $8 trillion US dollars (BIS, 2019[1]).[2]

By providing insurance against default, CDS enables loan lenders to hedge the default risk of borrowers, where CDS spread is dependent on the direct information about the creditworthiness of the entity named on the derivative security. After the 2008 financial crisis, CDS spreads have become the most closely monitored early warning signals for credit risk changes. The risk-neutral implied default probability estimated from CDS spreads are used to price credit securities, assess credit quality by rating firms, monitor systemic risk, and stress test financial systems by regulators (Chan-Lau 2006; Huang et al. 2009).

Compared with other credit risk measures such as bankruptcies, rating and bond yields, or general risk measures as stock volatility, CDS spreads have several advantages. First, CDS spreads are a continuous alternative to discrete credit assessments of rating agencies, which also incorporates market perceptions of default risk (Das et al., 2009). Unlike the rare credit events, the CDS market offers timely cross-sectional and time-series credit information, gauged by the market instead of a credit rating agency. Second, CDS spreads outperform ratings in capturing firm-specific default probability and also contains information on systematic risk. (Hilscher and Wilson, 2017). Third, CDS spreads contain credit information not included in stock prices or bond yields when important credit events occur, leading the price discovery on stock and bond market (Lee et al, 2019). Finally, CDS spreads are less affected by liquidity and tax effects compared to bond spreads (Elton et al., 2001), and are less sensitive to momentum than stock prices.

However, not all firms issue CDSs. Generating “shadow” CDS spreads for the firms without CDS can thus provide a useful credit risk measure, adding valuable insights for market participants. Das, Hanouna and Sarin (2009) find that both accounting-based and market-based information have explanatory power on CDS spreads. If the underlying structure between (economic/firm) fundamentals and CDS spreads is homogeneous across similar firms, one can artificially recover such a structure to the firms without CDS and generate "shadow" spreads. In this paper, we use the fundamentals to cross-sectionally nowcasting CDS spreads, test the validity, and generate the “shadow” spreads.

There has also been little research on forecasting CDS spreads to date. Two exceptions are

[1] https://www.bis.org/statistics/derstats.htm

[2] The regulations enforced through the Dodd-Frank Act, a financial reform legislation passed in response to the GFC, included registration requirements for market participants to trading, central clearing, and reporting of OTC derivative positions. The changes in the regulatory environment have led to quick reactions in the outstanding positions. While there is no mandatory central clearing regulation for the US single-name CDS contracts, the market activity is clearly transiting to clearing-eligible products, with an overall decrease in gross notional of single-name contracts outstanding (Boyarchenko, Costello, Shachar, 2019).

Guenduez and Uhrig-Homburg (2011), and Son et al (2016), who both predict firms' CDS spreads using historical spreads. No study has used economic and firm fundamentals to forecast CDS, while many researchers have used them to predict other credit risk measures, such as bankruptcies (e.g. Altman,1968; Ohlson, 1980; Altman, 2000; Hillegeist et al., 2004; Duffie et al., 2005; Agarwal and Taffler, 2008, and Duan et al., 2012), rating changes (Nickell et al., 2000; Duffie and Singleton, 2003; Jorion et al., 2009; Jones et al., 2015), bond yields (Huang et al., 2005; Collin-Dufresne et al.,2001; Longstaff and Rajan, 2006) and stock volatility (Christiansen et al., 2012; Mittnik et al., 2015). As discussed above, CDS spreads have multiple advantages over other risk measures. Hence, in this paper, we forecast future CDS spreads longitudinally using economic and firm fundamentals.

Our crosssectional nowcast and longitudinal forecast also incorporate the recent developments and applications of data-driven machine learning methods (MLs). In terms of credit risk, most studies using machine learning methods focus on bankruptcy and credit rating. Empirical evidence from these discrete measures suggests that recent classifiers such as gradient boost and random forest clearly excel compared to traditional LDA or probit/logit (Jones et al., 2015, Flavio et al., 2017). But there has not been equal scrutiny on the continuous measure of CDS spreads. What enables machine learning methods to outperform traditional approaches have not been investigated sufficiently. In this study, we "horserace" the predictive performances of traditional methods and a series of recent ML techniques in regards to their nowcasting and forecasting capabilities, and investigate the source of performance differences.

This paper aims at answering three specific questions:

> (1) Can we generalize the relationship between the fundamentals and CDS spreads cross-sectionally to other companies to construct "shadow" CDS spreads for those without actual CDS?
>
> (2) Can we generalize the relationship over time to forecast CDS spreads in the future?
>
> (3) What is the relative explanatory power of fundamental variables in predicting CDS, under traditional and Machine Learning approaches?

To answer these questions, we conduct nowcasting cross-sectionally, and the one-month ahead longitudinally forecast to predict CDS spreads. Our sample comprises monthly CDS spread data of 69 firms, with accounting-based, market-based, and macroeconomics series as input variables. We test a wide range of machine learning estimation techniques and use traditional credit risk model regressions as benchmark tools.

Our results indicate that machine learning methods can considerably enhance the prediction accuracy of CDS spreads both cross-sectionally and overtime when compared to traditional econometric models quantifying credit risk relationships. Ensemble methods including Bagging, Random Forest, and Gradient Boosting consistently outperform basic interpretable methods, such as Ridge, LASSO, and linear regression, in prediction accuracy and stability. The precision of linear regression fluctuates widely across randomly chosen estimation and test sets and leads to the weakest average out-of-sample prediction power.

We further assess the importance of regressors by using the LIME (Local Interpretable Model-Agnostic Explanations) method, to provide more thorough insights into the underlying reasoning for why ensemble MLs are more accurate in predicting CDS spreads, from the view of input variables. We find that linear regressions assign exceptionally high weights to interest rates and spreads, including treasury yields, term spreads, and long term bond yields. In contrast, ensemble ML methods rely mostly on the firm and economic fundamentals. The results pinpoint the most critical variables that predict CDS spreads and suggest that ensemble ML methods can identify authentic credit information for predicting CDS spreads.

The high cross-sectional and longitudinal precision of ensemble ML techniques suggests that the nonlinear relationship between the firm and economic variables and CDS spreads can be applied to other firms and also to the future. The corresponding generalizable relationship allows us to construct valid "shadow" CDS spreads for those companies without actual CDS, but with the firm and economic variables. We show that the constructed "shadow" CDS spreads can capture the main changing direction of the spreads, but are much less volatile. We are also able to predict future CDS spreads for those firms with CDS.

The remaining sections of the paper are organized as follows: Section 2 discusses the relevant literature, section 3 introduces the sample and imput variables, section 4 provides the discussion on methodology and empirical contexts, section 5 presents the results and provides the "shadow" CDS spreads that we have constructed for those firms who do not have "real" CDS, and section 6 provides LIME analysis on understanding why the nonlinear ensemble methods outperform the linear benchmark. In section 7, we design a specific case study using non-crisis periods as training sets and crisis periods as test sets, then we conclude.

II. LITERATURE REVIEW

The importance of using both accounting and market based variables in the modeling of credit risk has been intensively discussed in the credit risk literature. The pioneering works of Altman (1968) and Ohlson (1980) have used firm-specific financial ratios and other accounting variables to develop scores for predicting firm's default probability (Altman’s Z-score and Ohlson’s O-score).

The most widely recognized credit risk models in the field are based on market-based variables. Specifically, Merton (1974) has developed a distance to default (DTD) measure based on market information, assuming that the fundamental value of a firm follow a certain stochastic process and computes the default probability from the level and volatility of asset's market value.

As shown in Jarrow & Turnbull (1995) and Duffie & Singleton (1999), reduced-form models or intensity-based models assume that the default follows a process with stochastic intensity, and one can extract the default intensity from market securities. In such models, the conditional probability of failure of a firm depends purely on the distance to default, a variable calculated by market equity data and accounting data for liabilities. Empirically, the performance of these models are regarded to be superior to Altman’s Z-score and Ohlson’s O-score (Hillegeist et al., 2004).

Although structural models and reduced-form models have received great recognition both in the industry and academia - for example, structural models have been adopted by firms such as

Moody's KMV and CreditMetrics - the overemphasis of these models on distance to default raises concern. Duffie and Lando (2001) show that if markets are not fully efficient, DTD might cause filtering problems and other variables could provide additional information.

Hillegeist et al. (2004) find that DTD outperforms accounting information in predicting default. Hillegeist et al. (2004) and Duffie et al. (2007) conclude that accounting-based and macroeconomic variables are relevant as well in predicting corporate failure. Specifically, Das et al. (2009) find that models using accounting-based data and models using market-based information have performed similarly well in explaining CDS spreads. Bai and Wu (2016) combine DTD with multiple firm fundamentals and find that the fundamentals explain CDS spreads by an average 77% of R-square.

In the literature of corporate default prediction, the firm's failure intensity depends on the covariates, including firm-specific financial variables and macroeconomic variables. The prediction of forward intensity next period is conditional on the covariates observed on the present period. Duffie & Wang (2004) and Duan et al. (2012) incorporate all the accounting-based, market-based and macroeconomic variables to predict corporate default. This paper is in accordence with Duffie & Wang (2004) and Duan et al. (2012).

The application of machine learning methods in credit risk analysis and financial time series prediction has been pursued as separate strands of research in prior studies. For the credit market, most of the relevant work focuses on credit rating analysis. Huang et al. (2004) suggests that the rating analysis using artificial intelligence techniques choose input variables following the conclusion of traditional credit risk analysis. Jones et al. (2015) compares a range of classifiers from traditional techniques to fully non-linear classifiers including neural networks, support vector machines and more recent statistical learning techniques such as generalized boosting, Adaboost and random forest, to predict rating changes, using financial, market and macroeconomic variables as inputs. They find that new classifiers perform better than all other classifiers on both cross-sectional and longitudinal test samples.

Relatedly, relevant research predicting financial time series have concentrated in stock market and achieved relatively accurate prediction results. Relevant studies have used technical input variables and fundamental variables to predict stock return (Chan te al.,1993, Cavalcante, 2016) or volatility (Charlotte et al., 2012, Mittnik et al.,2015). Specifically, studies predicting CDS spreads have only used historical spreads. Gündüz and Uhrig-Homburg (2011) analyze the ability of CDS spreads in predicting future CDS spreads using both traditional credit risk models and support vector machine regression. Son et al. (2016) expand Gündüz and Uhrig-Homburg's work by introducing more modeling methods with additional maturities.

In this paper, we conduct the prediction of CDS spreads using fundamental variables and compare the results with traditional benchmark models, and we fill the gap between two strands of the literature, the credit risk literature and the machine learning literature.

III. PRICING CDS SPREADS

In this section, we motivate our nowcast and forecast with a forward default intensity model. We model the pricing of the CDS spreads following Das et al. (2009) and Duan et al. (2012). We

model the default of a firm as an intensity process, λ_t, thus the probability to survive from staring time t=0 to default time t=τ is $s_\tau = \exp\left(-\int_0^\tau \lambda_t dt\right)$. In the model, the forward intensity λ_t depends on the firm and economic variables observed at time t (X_t) or beforehand ($X_{t-i}, i > 0$), and is of exponential affine form,

$$\lambda_t = \exp[B'_{t-i} X_{t-i}], i \geq 0 \text{ ,}$$

where $B_{t-i} = \left[\beta_{0(t-i)}, \ldots, \beta_{k(t-i)}\right]'$ is a vector of coefficients, and $X_{t-i} = \left[1, X_{1(t-i)}, \ldots, X_{k(t-i)}\right]$ is a vector of economic variables including both accounting-based, market-based firm-level, and macroeconomic variables. Assuming that conditional on the given economic variables vector X_{t-i}, the forward default intensity is a constant, as $E(\lambda_t | X_{t-i}) = \lambda$.

CDS enables market participants to shift the default risk on the firm from an insurance buyer to an insurance seller. The buyer pays a premium to guarantee future potential protection. Hence the premium and the protection legs both determine CDS spread together. The premium leg represents the expected present value of premium payment from the insurance buyer to the seller, while the protection leg indicates the expected present value of the default loss payment from the seller to the buyer. Fairly priced CDS equals the premium leg and the protection leg.

The premium leg is,

$$E\left[\int_0^T D_t s_t CS dt\right] \quad (1). \text{ ,}$$

and **the protection leg** is,

$$E\left[\int_0^T D_t s_t \lambda_t (1-\phi) dt\right] \quad (2).$$

where T is the maturity of CDS and CS is the CDS spread. $D_t = \exp\left(-\int_0^t r_s ds\right)$ is the discount rate at default time t, where r_t is the interest rate at time t. $s_t = \exp\left(-\int_0^t \lambda_s ds\right)$ is the probability that firm survive until default time t. λ_t is the default intensity that the firm default at t, and ϕ is the constant recovery rate.

Assume that the maturity T can be equally divided into n intervals, where Δt is the time interval between time t and t-1. The intervals are denoted by j = 1, 2, …, n. Note that conditional on the given economic variables vector X_{t-i}, the forward default intensity is a constant. Hence

$$\lambda = \lambda_j = \exp[B'_{t-i} X_{t-i}], i \geq 0, \ j = 1, 2, \ldots, n, ,$$

Equating the premium leg and protection leg under conditional constant intensity leads to fairly priced CDS spreads[3]:

[3] See Das, Hanouna and Sarin (2009) for detailed discussion.

$$CS = \frac{(1-\phi)(1-e^{-\lambda\Delta t})}{\Delta t} \quad (3).$$

.

Taking logarithm and employing the fact that $\lambda = \exp(B'_{t-i}X_{t-i})$ leads to a linear relationship between $logCS$ and firm and economic variables,

$$\begin{aligned} logCS &= \log\left(\frac{1-\phi}{\Delta t}\right) + \log\left(1 - e^{-\lambda\Delta t}\right) \\ &\approx \log\left(\frac{1-\phi}{\Delta t}\right) + \log(\lambda\Delta t) \\ &= \log\left(\frac{1-\phi}{\Delta t}\right) + B'_{t-i}X_{t-i}\Delta t \end{aligned} \quad .$$

Namely,

$$logCS \approx constant + B'X_{t-\mathrm{i}},\ i \geq 0.$$

In comparison, for Machine Learning methods, we assume a flexible relationship between economic variables and $logCS_t$,

$$logCS \approx f(X_{t-\mathrm{i}}),\ i \geq 0 \quad (4).$$

,

where the function form is determined by Machine Learning methods. We conduct nowcasting ($i = 0$) when generating shadow CDS spreads cross-secitonally and forecasting ($i > 0$) in longitudinal analysis.

IV. DATA, EMPIRICAL CONTEXT, AND METHODS

4.1 Sampling

We utilize the CDS contracts data obtained from MARKIT. Our sample is based on the CDS constituents in the CDX North American Investment Grade Index, which includes the most liquid 125 North American entities' CDSs with investment-grade credit ratings. The reason to focus on most liquid CDSs is that they have the most fairly-priced and informative spreads in the North American CDS market. We collect the 5-year CDS spreads of the constituents at the end of each month over the period 2006 to 2016 [4]. After merging the sample with the WRDS Monthly Finance Ratio database, Compustat, CRSP daily stock file database, and IBES analyst database, 69 entities remain in our sample with 6811 corresponding monthly CDS spreads.

4.2 Input variables

We collect firm-level accounting-based and market-based variables, analyst forecasts, financial markets, and macro-economic variables, details of which are presented in Table 1.

Accounting-based variables

[4] 5-year maturity CDS is the most liquid among all maturities (Gündüz & Uhrig-Homburg, 2011).

We use the monthly financial indicators from the WRDS Industry Financial Ratio database (WIFR). WIFR is developed by WRDS based on the Compustat, CRSP, and IBES databases, covering a wide range of most commonly used financial ratios. The ratios measure various aspects of firms' fundamental performance, including capitalization, efficiency, financial solvency, liquidity, profitability, and valuation. The WIFR carries forward the most recent quarterly or annual data and lags all variables by two months to guarantee that the data is available at the specific month. After removing variables with more than 10% empty values, 57 financial ratios remain in our sample and are described in Table 1. Following the previous credit risk literature (Hensher et al., 2007; Jorion et al., 2009; Ashbaugh et al., 2006; Jones, 2015), we expect that these variables measure the overall performance of a firm and have predictive power over CDS spreads.

Market-based variables

A. Equity market variables

We include several equity market variables, including the stock return, realized volatility, the change of realized volatility, as well as the trade volume, to measure a firm's performance on the stock market. We also include the variables to measure the general stock market performance, including S&P 500 return, VIX (CBOE Volatility Index), Fama-French four factors, and Pastor-Stambaugh liquidity factors. The equity market reflects the market perception of general firm performance besides credit risk. Griffen and Lemmon (2002) find that firms' credit risk is cross-sectionally priced on the stock market. Tang and Yan (2010) and Lee et al. (2019) find evidence that the change of stock return and volatility is correlated with CDS spreads. Consistent with this literature, we expect that the equity market variables have some predictive power over CDS spreads.

B. Analysts' recommendations and estimates

We also follow Jones et al. (2015) to include equity analysts' recommendations and estimates as input variables. The recommendation and estimates are based on analysts' thorough investigation on a firm, hence should have covered firms' financial performance and credit quality should be covered.

C. Interest rates, spreads and risk factors

This category captures the interest rate dimension following Welch and Goyal (2008), namely, the T-Bill rate, relative T-Bill rate, long term bond return, term spread, and default spread. The TED spread that measures the illiquidity of the bond market is also included. Duffee (1998), Collin-Dufresne et al. (2001) and Bharath and Shumway (2008) find that changes in interest rates negatively affect the changes in default risk. Moreover, since the underlying reference entities and obligations of CDSs are senior unsecured bonds issued by corporate, the spreads on the bond market could influence the pricing of CDS spreads. Hence, we expect the interest rate and spreads to have predictive power over CDS spreads.

D. Distance to default (DTD)

We further use the market-based credit measure distance to default (DTD) to measure the

probability of default[5]. DTD is the most frequently used market-based credit risk measure developed by Merton (1974). Bharath and Shumway (2008) find that DTD has predictive power on financial distress and default. Das et al. (2009) find evidence that DTD and financial ratios perform comparably in explaining CDS spreads. Thus, we also expect DTD to have predictive power over CDS spreads.

Macroeconomic variables

We use a range of monthly updated macroeconomic indicators, including the inflation rate, industrial production, housing starts, M1 growth, orders, return CRB spot, consumer confidence, and others to measure the overall economic condition. The macroeconomic variables are commonly used in default and rating change prediction (Dun et al. 2012; Jones et al., 2015). Bonfim (2009) find evidence that macroeconomic variables explain default probabilities. Thus, we expect macroeconomic variables to play a role in CDS spread forecasting.

Other variables

We further include the firm size proxy, industry dummies, credit rating, and CDS recovery as input variables. The firm size and industry dummies are commonly used as controls in credit risk research (Moody's, 2004; Bonfim, 2009). The rating is the long-term credit rating assigned to the entity by S&P, Moody's, or Fitch. Recovery rates are pre-populated based on the recovery rate set. We use the credit rating and CDS recovery rates reported by MARKIT.

4.3 Machine Learning Methods

In addition to the widely used linear regression methods, there are a series of parametric and nonparametric machine learning approaches, which are well established in the literature. In this paper, we compare the theory motivated linear regression with two parametric machine learning methods (Ridge and LASSO) and six nonparametric learning methods (Support Vector Regression, Neural Network, Regression Tree, Bagging, Random Forest and Gradient Boosting). In the nonparametric learning methods, Support Vector Regression, Neural Network, and Regression Tree are single methods, while Gradient Boosting, Bagging, and Random forest are ensemble methods. We briefly introduce the methods in **Appendix A**.

---

[5] Following Bharath and Shumway (2008), we calculate the DTD measure as,

$$DTD = \frac{\log\left[\frac{E+F}{F}\right] + \left(r - \frac{\sigma_v^2}{2}\right)T}{\sigma_v T^{\frac{1}{2}}}$$

a. E is the market value of equity calculated as "the number of shares outstanding" times "the end of the monthly closing stock price."
b. F is the face value of the firm's debt and is calculated as the debt level in current liabilities plus one-half of the long-term debt level reported in Compustat.
c. r is the expected return on firm assets (we have set r equal to the risk-free rate).
d. σ_v is the volatility of the firm's market value (we have calculated the volatility of the firm's equity value in the past 180 days).

4.4 Empirical Context

We focus on the out-of-sample predictive power of the accounting-based and market-based variables on CDS spreads using linear regression and machine learning methods, motivated by reduced-form forward intensity model. To fairly compare these methods, all of the models are estimated using the same set of input variables within the same dataset. To test the out-of-sample predictive performance, we divide the original dataset into an in-sample training set and out-of-sample test set. The methods are estimated on in-sample set to determine respective parameters and evaluated in the out-of-sample set. We follow Espinoza et al. (2012) to evaluate the predictive performance using root-mean-square error (RMSE), a frequently used measure that captures the difference between the predicted and observed values. Smaller RMSE indicates better predictive performance. We have split our CDS sample both *cross-sectionally* and *longitudinally*.

In the cross-sectional case, we conduct nowcast and evaluate whether our approach can provide precise CDS spreads prediction cross-sectionally and hence potentially generate effective shadow CDS spreads for the firms without CDS. We follow the 80/20 sample division arrangement to randomly allocate 80% of the firms into the in-sample set and the remaining 20% into the out-of-sample set. The division is replicated ten times to avoid biased allocation. For each replicate, we generate 1000 bootstrapped RMSEs and then calculate the average out-of-sample RMSEs across the ten replicates to measure the predictive power of the methods.

In the longitudinal case, we generate one-month forward forecasting and test the intertemporal predictive ability of model-motivated linear regression, and machine learning methods on CDS spreads[6]. We separate the in-sample training set from the out-of-sample test set with boundary year rolling from 2011 to 2016. To mimic the actual data available at the end of each boundary year, we include the observations before the year in the training set, within the year in the test set, and abandon the rest observations. Such a longitudinal arrangement can provide the intertemporal validation that is missing in the cross-sectional case, in which the test set is drawn from the same sample period of the training set (see Jones and Hensher, 2004). The rolling windows generated by rolling boundaries can also avoid biased allocation and hence provide an adequate test of a model's intertemporal predictive ability. Similar to the cross-sectional case, we produce 1000 bootstrapped RMSEs for each window and calculate the average out-of-sample RMSEs across all rolling windows to measure the methods' predictive performance.

Finally, the hyperparameters of a model might strongly influence the performance of prediction outcomes, as well as the degree to which the model overfits the data. Overfitting indicates that the model fits well in the in-sample training set but performs poorly in the out-of-sample test set. We use a standard 10-fold cross-validation method with the loss function RMSE to adjust the hyperparameters of models and intend to avoid the overfitting problem.[7]

---

[6] In this paper, we focus specifically on filling a gap for those firms that do not have a CDS market and on the near-term precision of CDS forecasting. However, we have still tested the sensitivity of the findings for additional forecasting windows, as 3, 6 and 12 months ahead, results are available upon request.

[7] Our 10-fold cross-validation method select hyperparameters by dividing the original training set equally into 10 subsets, estimating a model with certain hyperparameter based on 9 subsets and score the model's performance using

(continued…)

V. RESULTS

This section describes the empirical performance of model-motivated linear regressions and alternative machine learning methods. We first provide the descriptive details of the sample. Among all the 6811 log CDS spreads in our sample, the average log spread is -5.026, namely 65 bps in original spreads. Figure 1 demonstrates that our sample has covered a wide range of spreads. The log spreads range from -7.67 to -2.43, which is 4.6 bps to 880.3 bps for spreads. Our sample is representative since 94.8% of the spreads for all CDX indices constituents fall into our sample range, including 99.7% of investment-grade CDX.NA.IG constituent, 93.9% of the CDX.NA.XO[8] constituents, and 87.9 % of the high yield CDX.NA.HY constituents.

Figure 1: The Distribution of The log of Five year CDS spreads

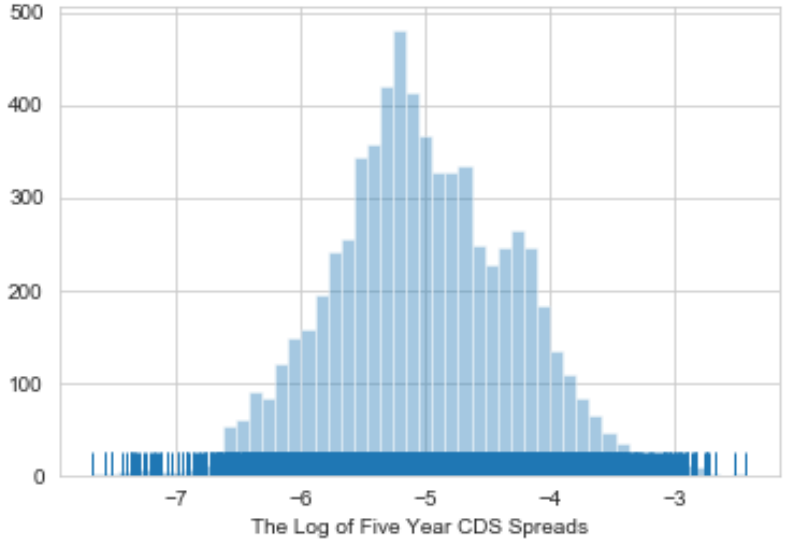


Figure 2 displays the box plots for the bootstrapped RMSE of each model across cross-sectional and longitudinal test samples. The box plots provide insights on the predictive stability of each method over different data subsamples. For the cross-sectional case, the box plots show bootstrapped RMSEs calculated from the ten randomly selected test samples; for longitudinal, the bootstrapped RMSEs on all rolling boundary years are displayed. The extreme RMSEs are showed as outliers in box plots. We consider the methods with more outlier RMSEs and more substantial variance as less stable.

---

the out-of-sample RMSE on the rest subset. The procedure is conducted recursively until all subsets have been used to score. The chosen hyperparameters are determined based on all the scores. The hyperparameters for each method are different. Specifically, for LASSO and Ridge regression, the hyperparameter is the regularization penalty degree; for SVR, there are the penalty parameter of the error term, kernel coefficient and epsilon in the epsilon-SVR model; for neural network, it's the regularization penalty degree and size of hidden layer. The hyperparameters for regression tree, gradient boosting and random forest are the maximum depth of a tree, the maximum number of leaves in a tree, the minimum number of samples required to split a tree, and the minimum samples at each leaf node; for bagging, there are the number of base estimators in bagging and the maximum input variables used in each base estimator.

[8] CDX.NA.XO index contains CDSs that are at the crossover point between investment and junk grade.

Table 2 summarizes 1) the average overall RMSE and ranking of each method across all test samples as well as the average RMSE and ranking of each model over cross-sectional and longitudinal test samples. Table 3 alternatively demonstrates the variances and ranking of RMSE across all test samples and separately for cross-sectional and longitudinal test samples.

The overall results displayed in Table 2 indicate that ensemble machine learning models, including Random Forest, Bagging, and Gradient Boosting, have outperformed all other methods, both in cross-sectional and longitudinal samples. Gradient Boosting and Bagging have overall average RMSE at 0.397 and 0.413, with the former performs slightly better than the latter. The overall RMSE of Random Forest is around 0.454, and the Regression tree follows by 0.554. The accuracy of support vector regression and Lasso regression decrease relatively large and have RMSE above 0.7. Ridge regression has further worse accuracy with an RMSE 1.818. Among all the methods, theory-motivated linear regression is the weakest over the whole subsample, with a very large RMSE of 3.433. The neural network is slightly better with RMSE of 3.167.

An interesting result presented in Figure 2 is that inflexible methods including Linear and Ridge regression can forecast comparably well along the lines of ensemble machine learning methods in some cases, but can also perform quite poorly in other cases. The three methods have more outlier RMSEs and a wider range of RMSEs.

Table 3 confirms that Linear regression is the most unstable method with an overall variance of 22.51, followed by Ridge regression (16.787) and neural network (3.450). In comparison, ensemble methods, including Gradient Boosting, Bagging, and Random Forest, have provided forecasts with very low variance (0.006, 0.007 and 0.010), indicating that their predictive performances are remarkably stable across different subsamples. Though support vector regression does not provide very accurate prediction (RMSE=0.701, rank 5), it has the lowest RMSE variance among all the methods. [9]

[9] Jones et al. (2015) find that inflexible methods such as LDA, probit and logit can predict discrete rating changes respectively good compared to more flexible methods. Our results partly confirm their findings on continuous CDS spreads, however, we also find that model performance is not robust over all multiple training and test sets.

(continued…)

Figure 2: The Bootstrapped RMSEs in Cross-sectional and Longitudinal Test Samples[10]

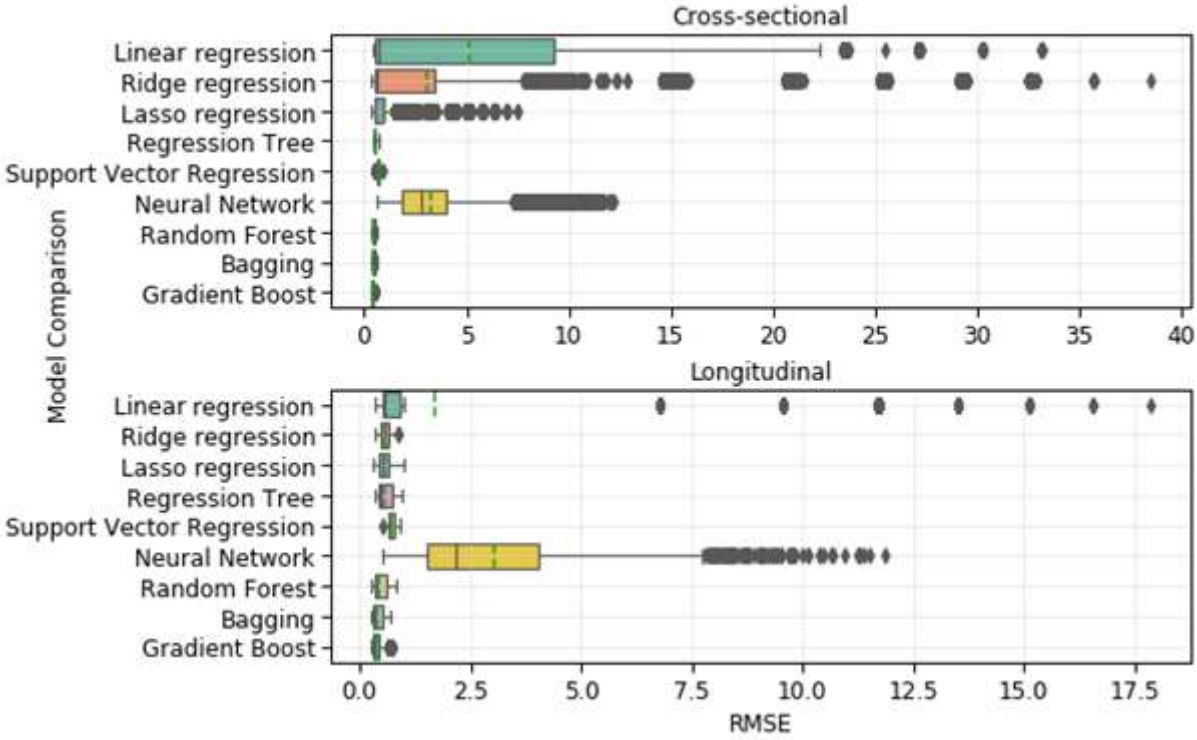


Table 2: The Average RMSEs in Cross-sectional and Longitudinal Test Samples

Overall Performance			Overall Performance (Cross-sectional)		Overall Performance (Longitudinal)	
Methods	Average RMSE	Rank	Average RMSE	Rank	Average RMSE	Rank
Linear Regression	3.433	9	5.177	9	1.688	8
Ridge Regression	1.818	7	3.066	7	0.570	5
Lasso Regression	0.775	6	0.982	6	0.567	4
Support Vector Regression	0.701	5	0.689	5	0.713	7
Neural Network	3.167	8	3.284	8	3.05	9
Regression Tree	0.554	4	0.537	4	0.571	6
Random Forest	0.454	3	0.458	3	0.450	3
Bagging	0.413	2	0.434	2	0.391	1
Gradient Boosting	0.397	1	0.401	1	0.393	2

[10] The box plots regard the RMSEs larger or smaller than the average RMSE ± 1.5 standard deviation as outliers. The solid black line within the box indicates the average RMSE without outliers, while the green dash line suggests the average RMSE with outliers.

Table 3: The RMSE Variances in Cross-sectional and Longitudinal Test Samples

Overall Performance			Overall Performance (Cross-sectional)		Overall Performance (Longitudinal)	
Methods	RMSE Variance	Rank	RMSE Variance	Rank	RSME Variance	Rank
Linear Regression	22.551	9	28.622	9	6.998	9
Ridge Regression	16.787	8	25.377	8	0.014	2
Lasso Regression	0.624	6	0.949	6	0.035	6
Support Vector Regression	0.005	1	0.004	5	0.007	1
Neural Network	3.450	7	4.571	7	2.082	8
Regression Tree	0.014	5	0.002	1	0.037	7
Random Forest	0.010	4	0.003	3	0.025	5
Bagging	0.007	3	0.004	4	0.014	4
Gradient Boosting	0.006	2	0.002	2	0.014	3

5.1 Cross-sectional sample

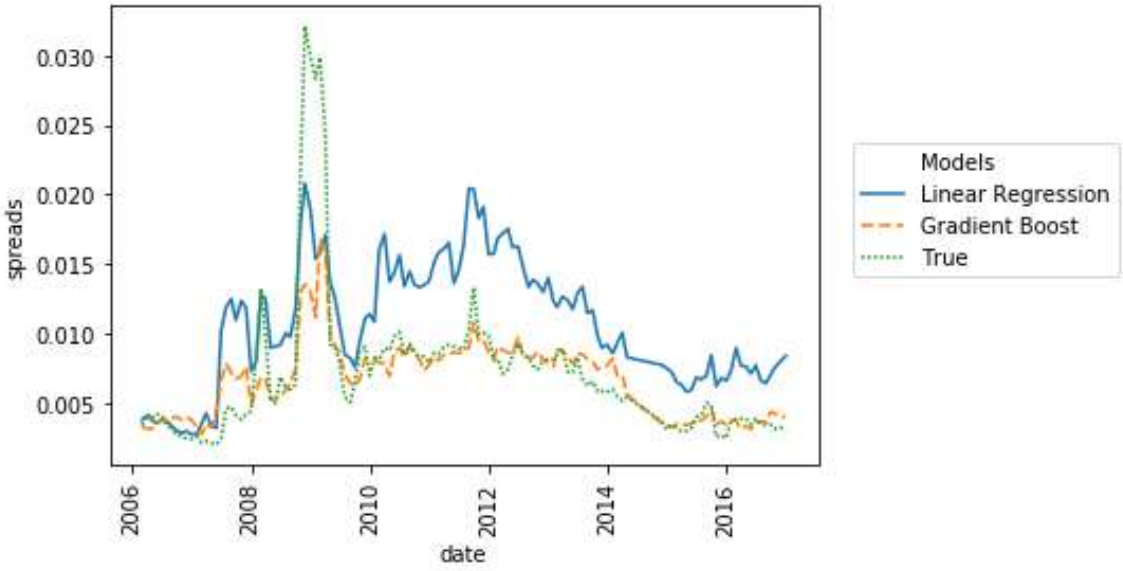
In the cross-sectional sample, among the 69 firms, we randomly select 56 firms as part of the training set and 13 firms as the test set with ten replications and calculate the average RMSEs. Table 2 summarizes the overall average performance of used methods across cross-sectional test samples, and the performance ranking is consistent with the ranking for average RMSE across both cross-sectional and longitudinal samples. Results from Table 2 indicate that the ensemble machine learning methods, including Random Forest, Bagging, and Gradient Boosting, provide the most accurate nowcasting predictions, with Gradient Boosting outperforming all other methods with the average RMSE at 0.401.

Linear methods, namely OLS, Ridge, and LASSO, perform relatively worse compared to ensemble machine learning methods. Both Ridge and LASSO outperform OLS. The predictive accuracy of OLS is substantially worse than all other of the methods used.

Ensemble methods combine a range of weak estimators to produce a strong one. The weak estimators are assessed on multiple subsamples extracted from the original dataset, and the final prediction is a weighted average of the predictions generated by all the weak estimators (see the differences of Random Forest, Bagging, and Gradient Boosting in Appendix A). Hence ensemble methods are much more stable for various training and test set pairs.

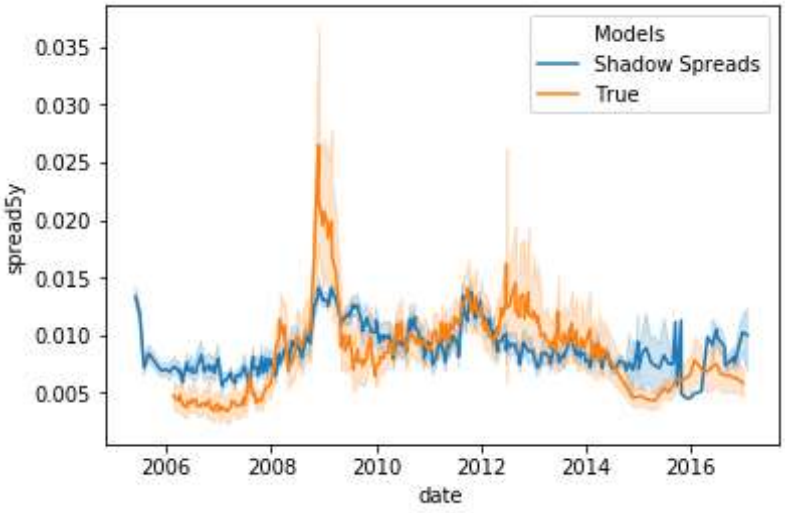
Figure 3 demonstrates the shadow spreads generated by out-of-sample nowcasting using the Omnicom Group as an illustrating example. We have compared the nowcasting result with the original data series (green ellipsis line): the Gradient Boosting method (orange dash line) has generated the best accuracy in terms of the lowest RMSE (0.401) and the Linear Regression Model is our comparison benchmark. We can see that the Gradient Boosting method is much better than the benchmark model in terms of fitting the actual data points.

Figure 3: Nowcasting Shadow Spreads , an Illustrating Example on Omnicom Group



While our gradient boost "shadow" CDS spreads can achieve relatively high accuracy, it cannot fully describe the movement of actual spreads. Tang and Yan (2017) find that besides the fundamental factors, the supply-demand imbalance and liquidity in the CDS market also moves CDS spreads. Since our "shadow" CDS spread is generated based on a wide range of fundamentals and factors on equity and bond markets, the spreads only capture the fundamental and "external" part of actual CDS spreads.

Figure 4: Aggregate Shadow Spreads for firms



The solid lines indicates the average CDS spreads of all firms, and shadow area around the solid line describes the interval of individual spreads.

Our shadow spreads generated by Gradient Boosting can also play an essential role in the existing CDS on missing times. As a derivative, CDS do not necessarily have continuous spreads

for every month. Figure 4 summarizes the actual spreads on existing months for all firms, and the corresponding shadow spreads on the missing months using the actual spreads and input variables as the training set. Our shadow CDS spreads generated by Gradient Boosting managed to capture the main moving direction of actual spreads while being less volatile, consistent with its fundamental property.

5.2 Longitudinal sample

Joy and Tollefson (1975) notes that the test set created from the same period as the training/estimation set will not provide intertemporal validation, and thus cannot provide an adequate test on a model's predictive ability (see also Jones and Hensher, 2004).

To ease the above concern, in this subsection we use longitudinal samples on rolling windows, which are rolling from 2011 to 2016. While the cross-sectional sample separates different firms into training and test sets, the longitudinal sample have all firms in both training and test set, and the division is on the data time. Taking the year 2011 as an example, we allocate the observations for all firms before 2011 into the training set, observations on 2011 into the test set, and drop the data points after 2011. Such a procedure is applied in the year 2011 until the year 2016.

Table 4: RMSEs in Longitudinal Test Samples

RSME	OLS	Ridge	Lasso	Tree	RF	Bagging	Gradient Boosting	SVR	Neural Network
2011	0.916	0.449	0.434	0.464	0.353	0.340	0.341	0.689	1.820
2012	7.078	0.469	0.436	0.453	0.356	0.349	0.341	0.692	4.446
2013	0.523	0.564	0.420	0.378	0.311	0.286	0.280	0.588	4.909
2014	0.496	0.605	0.659	0.734	0.671	0.555	0.607	0.702	3.340
2015	0.586	0.776	0.900	0.873	0.629	0.528	0.451	0.842	2.497
2016	0.525	0.558	0.553	0.525	0.379	0.288	0.339	0.767	1.289

Table 4 demonstrates the performance score of methods used in the longitudinal test sample. Compared with Machine Learning models, especially Random Forest, Bagging, and Gradient Boosting, linear models still perform relatively worse but deliver much better results than the cross-sectional samples. On average, the Bagging method and the Gradient Boosting method have generated the best accuracy in terms of the lowest RMSE (0.391 for Bagging, and 0.393 for Gradient Boosting on average, see Table 2).

To summarize, Table 4 shows that unlike the cross-sectional case, Ridge and Lasso regression perform comparably well to other Machine Learning methods on average for longitudinal samples. However, linear regression still suffers from extreme RMSE outliers.

VI. CHANNELS

After testing the cross-sectional and longitudinal samples, we apply the LIME module to different

methods. LIME is short for "Local Interpretable Model-Agnostic Explanations," the LIME module has two major advantages: (1) it is able to detect and improve untrustworthy models; and (2) it allows insights into different models. In this section, we use LIME to provide a locally faithful explanation for linear and non-linear algorithms..[11]

We first select the top 50 most important observations. To obtain a representative explanation of the overall dataset, we select 50 "true" observations using the *submodular* pick method (Ribeiro, Singh, Guestrin, 2016). (Ribeiro, Singh, Guestrin, 2016).[12] The advantage of the submodular pick is in explaining the model globally by combining local explanations, namely to select observations that give the most different input variable importance to capture the heterogeneity from the raw data set.

Among these 50 most important observations, for each selected observation y_i, we first generate perturbed input variables values $X\_i\_perturbed_{(n*k)}$ around y_i. For variables with numerical values, we perturb them by sampling from a standard normal distribution and implementing the inverse operation of mean-centering and scaling, according to the means and standard deviations in the training data. For variables with categorical values, we perturb them by sampling according to the training distribution and construct a binary variable that is 1 when the value is the same as the instance being explained. Then we calculate the prediction of the trained algorithm using the perturbed variables values, $y\_i\_predict_{(n*1)}$.

With the newly created local dataset ($X\_i\_perturbed_{(n*k)}$, $y\_i\_predict_{(n*1)}$), we use weighted Ridge regression to find the top 10 most important variables. The weight for each perturbed observation is the kernel distance of the observation to the true observation around which we build the local dataset. The top 10 variables are selected using the highest weights, namely, selecting the top 10 input variables that have the highest product of absolute coefficient and the variable value of the original data point.

We conduct the above local interpretations for each observation and aggregate all the interpretations. Each estimated local weighted Ridge provides the top 10 most important input variables and their coefficients. After aggregating our explanations, we build up two measures:

[11] LIME is a local method because it is based on specific observations. Using the estimated non-linear algorithm (e.g. a trained Random Forest) as the data-generating process, LIME conducts sampling in the neighborhood of one specific observation and generates virtual observations. The virtual observations provide a local dataset around the true observation, and allow us to estimate an interpretable model, such as linear regression. We use the default weighted ridge regression as the interpretable model. Such estimated interpretable model provides corresponding input variables' coefficients, which can be regarded as the local explanation of the non-linear algorithm(s). Such explanation has local fidelity around the specific observation. To generate a global explanation, one needs to select a set of representative observations, construct an explanation matrix and combine their local explanations.

[12] Following Ribeiro, Singh, Guestrin (2016), we use submodular pick to avoid selecting observations with similar local explanations. Such a pick will maximize a weighted coverage function to keep picking the observation with the largest marginal coverage gain.

1. **Importance probability**: it measures the frequency of a variable to appear in the top 10 most important variables among the selected 50 observations; e.g., for linear regression, "Long-term Debt Spreads" is among the top 10 most important variables for all the 50 observations, thus the probability is 50/50 = 100%. "Distance to default" is not so important in the context of linear regression, since among the 50 observations, only 5 observations pick the variable as the important top 10, hence the importance probability is 5/50 = 10%.

2. **Coefficient**: The coefficient for a variable is the average coefficient of the variable in the local weighted Ridge of 50 observations.

In the following, we mainly discuss the LIME results of the Linear Regression model (our benchmark model) and the Gradient Boosting model (our best-performing model) in the context of cross-sectional nowcasting.

For the linear regression model, we find that three variables have the most significant impacts on prediction results[13]: (1) the three-month T-Bill rate; (2) the long term bond return; (3) the term spreads. All the above three variables belong to the financial market information subset, which is in accordance with the theoretical model that we have used in section 3. The variable importance probability matrix is reported in Table 5. We calculate the average importance probability across 10 randomly selected training/testing set of firms in our cross-sectional sample.

In Table 6 we report the variables importance probability matrix of the Gradient Boosting Model. Under the non-linear model of Gradient Boosting, we find that the most important variables are the macro economic variables and firm specific balance sheet variables, as (1) Unemployment Rate, (2) Credit Rating Category, (3) Size proxy, (4) Distance to Default, and (5) Inflation Rate. The significantly different pattern compared with the LIME results from the linear model, is that the balance sheet information (firm dependent information) becomes more important in the non-linear model, which is believed to be the key driving factor why the non-linear model's prediction power could outperform the linear models in our setup.

It is interesting to see that the "Unemployment Rate" and "Inflation Rate" appear in the top 5 ranking variables, indicating the important role of the current state of the business cycle or, more broadly, the macroeconomic environment. The "Credit Rating Category" and "Distance to Default" are direct measures for evaluating the health of a specific firm, and it is not surprising that these two variables play important role in our nowcasting & forecasting exercises.

[13] This is consistent through cross-sectional and longitudinal samples.

Table 5: Variables Importance Probability – Linear Regression Model
(Cross-sectional Nowcasting)

Variables Label	0	1	2	3	4	5	6	7	8	9	Average Importance Prob.
3 Month Treasury Bill Rate	1	1	1	1	1	1	1	1	1	1	100.0%
Long Term Bond Return	1	1	1	1	1	1	1	1	1	1	100.0%
Term Spreads	1	1	1	1	1	1	1	1	1	1	100.0%
Capitalization Ratio	0.9	0.92	0.86	0.92	0.94	0.98	1	0.94	0.96	0.94	93.6%
Long-term Debt/Invested Capital	0.84	0.84	0.76	0.88	0.92	0.88	1	0.76	0.96	0.88	87.2%
Common Equity/Invested Capital	0.82	0.78	0.72	0.76	0.94	0.84	0.92	0.64	0.9	0.9	82.2%
Total Debt/Capital	0.76	0.32	0.76	0.18	0.54	0.28	0.42	0.36	0.82	0.56	50.0%
Total Debt/Total Assets	0.84	0.36	0.04	0.26	0.52	0.48	0.34	0.08	0.12	0.56	36.0%
Unemployment Rate	0.14	0.24	0.28	0.1	0.16	0.06	0.14	0.36	0.32	0.2	20.0%
Long-term bond yield minus its 12 month moving average	0.12	0.24	0.2	0.18	0.1	0.22	0.1	0.16	0.24	0.28	18.4%
Distance to Default	0.16	0.36	0.26	0.12	0.16	0.14	0.1	0.24	0.18	0.08	18.0%
T-Bill rate minus its 12 month moving average	0.14	0.18	0.14	0.2	0.08	0.16	0.16	0.16	0.18	0.08	14.8%
Industrial Production Growth, YoY	0.08	0.22	0.16	0.1	0.06	0.1	0.08	0.12	0.12	0.06	11.0%
Pre-tax Profit Margin	0.3	0.7	0.34	0	0.52	0.32	0.1	0.02	0.48	0.16	29.4%
Price/Operating Earnings (Basic, Excl. EI)	0.08	0.12	0.26	0.76	0.18	0	0.62	0.6	0.14	0.06	28.2%
Price/Operating Earnings (Diluted, Excl. EI)	0.06	0.08	0.22	0.76	0.12	0	0.64	0.6	0.08	0.06	26.2%
Gross Profit Margin	0	0.1	0.4	0.26	0.42	0.22	0.26	0.08	0.22	0.62	25.8%
M1 Growth, YoY	0.02	0.02	0.1	0.02	0.02	0.02	0	0.12	0.14	0.12	5.8%
Default spreads	0.04	0.06	0.1	0.04	0	0.02	0.02	0.06	0.06	0.04	4.4%
Risk-Free Return Rate (One Month Treasury Bill Rate)	0.02	0.08	0	0.06	0.04	0.06	0.02	0.04	0.02	0.06	4.0%

Table 6: Variables Importance Probability – Gradient Boosting Model
(Cross-sectional Nowcasting)

Variables Label	0	1	2	3	4	5	6	7	8	9	Average Importance Prob.
Unemployment rate	1	0.96	0.98	0.94	0.96	0.9	0.96	0.92	0.9	0.9	94.2%
Credit Rating	0.66	0.96	0.9	0.94	1	1	0.96	0.96	0.82	1	92.0%
Size proxy: total asset/ average total asset of sample time	0.92	0.78	0.94	0.94	0.8	0.64	0.82	0.6	0.82	0.78	80.4%
Distance to Default	0.8	0.84	0.74	0.72	0.78	0.66	0.78	0.8	0.8	0.7	76.2%
Inflation Rate, YoY	0.72	0.82	0.56	0.56	0.6	0.58	0.38	0.44	0.64	0.42	57.2%
Dividend Yield	0.58	0.3	0.84	0.48	0.38	0.64	0.48	0.58	0.48	0.34	51.0%
After-tax Interest Coverage	0.3	0.56	0.44	0.46	0.24	0.7	0.76	0.68	0.28	0.54	49.6%
Enterprise Value Multiple	0.36	0.28	0.44	0.66	0.26	0.36	0.68	0.48	0.34	0.28	41.4%
Interest Coverage Ratio	0.4	0.62	0.16	0.48	0.14	0.44	0.38	0.12	0.36	0.58	36.8%
Price/Sales	0.12	0.42	0.16	0.18	0.6	0.12	0.18	0.62	0.5	0.22	31.2%
RelT-Bill Rate	0.52	0.28	0.3	0.3	0.32	0.3	0.12	0.38	0.32	0.22	30.6%
Monthly log returns of the S&P 500	0.28	0.36	0.38	0.12	0.02	0.26	0.08	0.38	0.26	0.06	22.0%
Interest/Average Total Debt	0.12	0.04	0.06	0.34	0.2	0.04	0.06	0.12	0.12	0.12	12.2%
Median Estimate	0.24	0.22	0.22	0.36	0.5	0	0.58	0.06	0.3	0.26	27.4%
Log realized variance	0.26	0.14	0.14	0	0.46	0.36	0.1	0.14	0.2	0.38	21.8%
Research and Development/Sales	0.44	0.3	0.14	0.12	0.42	0.02	0.24	0.02	0	0.12	18.2%
Industrial Production Growth, YoY	0.14	0.14	0.32	0.12	0.12	0.14	0.2	0.04	0	0.12	13.4%
Price/Cash flow	0.02	0.08	0.02	0	0.1	0.02	0.02	0.18	0.44	0.12	10.0%
Forward P/E to Long-term Growth (PEG) ratio	0	0.08	0.1	0.14	0.08	0.12	0.2	0.06	0.08	0.02	8.8%
M1 Growth, YoY	0.1	0	0.14	0.1	0.08	0.14	0.14	0.14	0.14	0	9.8%
Shillers Cyclically Adjusted P/E Ratio	0	0.12	0.08	0.14	0.14	0	0.02	0.02	0.02	0.1	6.4%
Long Term Bond Return	0.14	0.56	0.38	0.1	0	0.04	0	0.58	0	0.02	18.2%
Common Equity/Invested Capital	0.16	0	0.04	0	0.3	0.04	0.12	0	0.1	0.62	13.8%
Price/Book	0.06	0.08	0.02	0	0	0.04	0.14	0	0.28	0.3	9.2%
Labor Expenses/Sales	0.04	0.02	0.02	0.06	0.06	0	0.12	0	0	0.14	4.6%
Pre-tax Profit Margin	0	0.16	0.02	0	0.04	0	0.02	0.02	0.16	0.02	4.4%
Long-term Debt/Total Liabilities	0.04	0.04	0.14	0.02	0	0	0.04	0.12	0	0.02	4.2%

It is equally surprising that the Linear Regression Model do not seem to properly capture the important explanatory power of "Credit Rating Category" and "Distance to Default" separately or simultaneously; they have not even entered the top 10 ranking list. These results seem to suggest that the nonlinear relationship between "Credit Rating Category" & "Distance to Default" and the CDS spreads could be the major reason why non-linear models (Gradient Boosting or Bagging) can generate high prediction accuracy compared to the linear benchmark (including Ridge and

Lasso model).[14]

VII. CRISIS VS. NON-CRISIS PERIOD

Scrutinizing Table 4, we see that if one uses the most recent data as test set (2015 or 2016) one will not see much of a difference between OLS and ML algorithms, the relatively large RMSE difference appears with using year 2011 or 2012 as the rolling window boundaries. Since particularly 2011 is still very close to the global financial crisis period and a time of considerable adjustments of both economic processes and financial systems, it seems useful to dig deeper and really assess whether there is a considerable difference in the forecasting/nowcasting abilities of traditional linear models versus Machine Learning models when using crisis versus non-crisis times as training and test sets separately.

Fine tuning based on our previous longitudinal sample with rolling windows, we further conduct a specific case study by separating the observations of all firms into two groups: crisis and non-crisis. We include the observations within non-crisis periods as training sets (2006, 2013, 2014, 2015, 2016) and crisis periods (2008, 2009, 2010, 2011) as test sets. Since 2007 and 2012 are the transitory years between crisis and non-crisis periods, to fully separate the two periods, we don't include the two years in training or test sets. Although we have perturbed the time periods due to the division of non-crisis vs crisis periods, our main goal of designing such a test is to evaluate the linear vs ML algorithms when a structural break is present. Given the very nature of the non-linearity brought by crisis as a structural break, we expect that the ML algorithms especially the ensemble methods are able to behave relatively well in terms of forecasting performance.

Figure 5 displays the box plots on the bootstrapped RMSEs of our test sample, which provides a clear presentation on the RMSEs across different methods. Table 7 also summarizes the overall RMSEs and the corresponding ranking. Not surprisingly, the ensemble machine learning models including Random Forest, Bagging and Gradient Boosting have outperformed all other methods.

[14] We find strong consistency for results using cross-sectional nowcasting, cross-sectional forecasting and longitudinal forecasting under the Gradient Boosting model, here we report the top 5 importance ranking variables across different samples.

	Gradient Boosting Model		
Importance Ranking	CS-Nowcasting	CS-Forecasting	Longitudinal
1	U.Rate	U.Rate	Credit Rating
2	Credit Rating	Credit Rating	Dis. To Default
3	Size	Size	R&D/Sales
4	Dis. To Default	Dis. To Default	U.Rate
5	inflation	inflation	Size

Figure 5: The RMSEs in Non-Crisis Training vs Crisis Test Sample

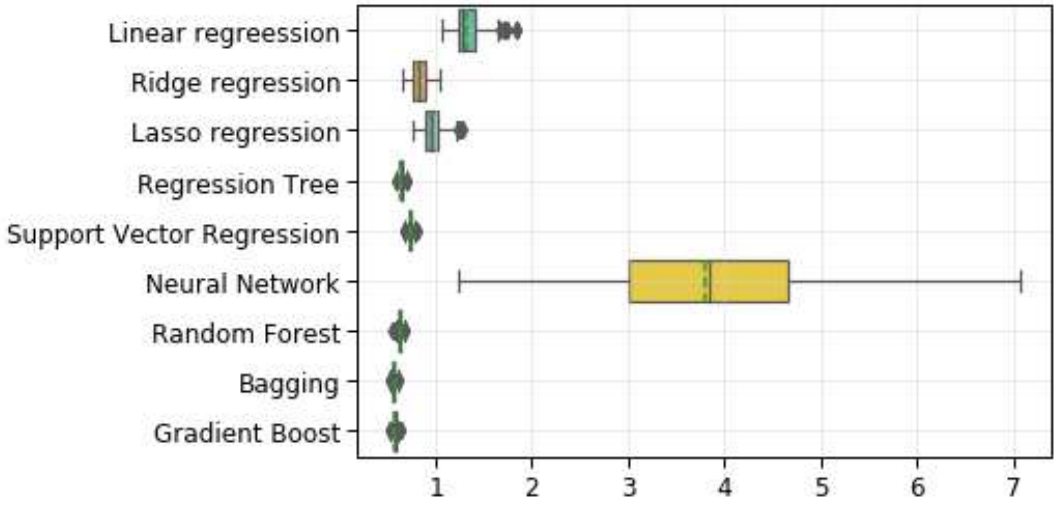


Table 7: The Average RMSEs in Non-Crisis Training vs Crisis Test Sample

Performance		
Methods	Average RMSE	Rank
Linear Regression	1.313	8
Ridge Regression	0.827	6
Lasso Regression	0.957	7
Support Vector Regression	0.738	5
Neural Network	3.790	9
Regression Tree	0.641	4
Random Forest	0.625	3
Bagging	0.563	1
Gradient Boosting	0.576	2

Table 8 summarizes the RMSEs of the top two ranked methods (Bagging and Gradient Boosting) across all the test samples. As what we have expected, the RMSE of the non-crisis training vs crisis test sample is much higher than the longitudinal test samples and the longitudinal/cross-sectional test samples all together. It is not surprising that the non-linearity brought by crisis as a structural break is the driver behind the spike in RMSEs.

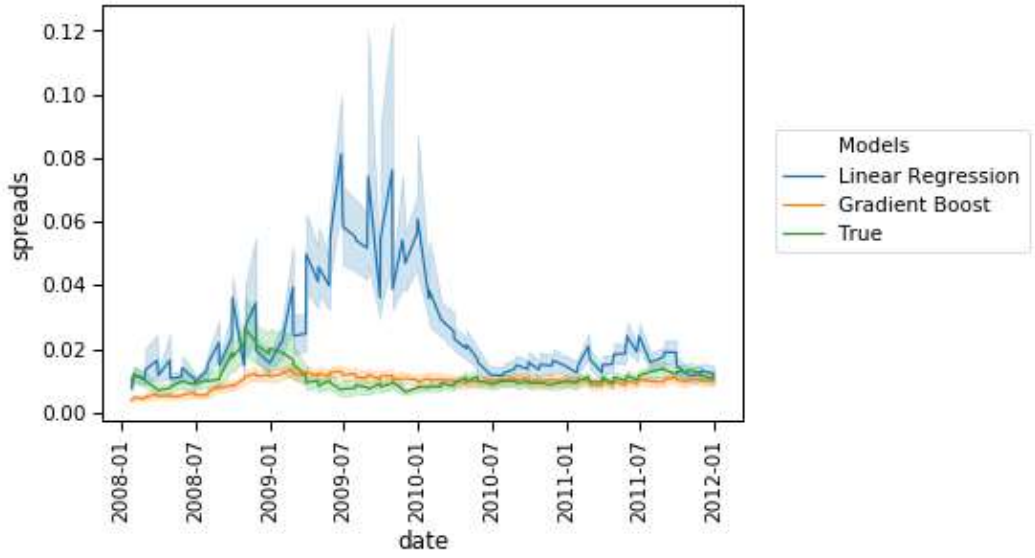
We also apply LIME module to the bagging method (first ranked method for the non-crisis vs crisis test sample). According to the variables importance probability generated by LIME, the top ten most important variables are: (1) unemployment rate, (2) credit rating, (3) M1 growth rate (YOY), (4) size proxy measured by total asset over average total asset of the sampling period, (5) after-tax interest coverage, (6) distance to default measure, (7) term spread, (8) multiple of enterprise value to EBITDA, (9) interest coverage ratio, and (10) dividend yield. The results of applying LIME module to the gradient boosting model are similar in terms of the selection of the top ten most important vaiables.

Table 8: The Average RMSEs across All Test Samples (Bagging/Gradient Boosting)

Longitudinal/Cross-sectional Test Samples			Non-crisis vs. Crisis Test Sample		Longitudinal Test Samples	
Methods	Average RMSE	Rank	Average RMSE	Rank	Average RSME	Rank
Bagging	0.413	2	0.563	1	0.391	1
Gradient Boosting	0.397	1	0.576	2	0.393	2

Interestingly, and consistent with our previous findings in section 5, the LIME results of linear model are quite different. Leaving the large RMSE and variance aside, the top ten important variables are: (1) capitalization ratio, (2) three-month T-Bill rate, (3) long term bond return, (4) term spread, (5) common equity/invested capital, (6) monthly capacity utilization, (7) long-term debt/invested capital, (8) monthly industrial production growth, (9) unemployment rate, and (10) total debt to capital ratio. The inability of the linear model to properly capture the importance of "Credit Rating Category" and "Distance to Default" is again evident in this specific case. Hence, it is evident that linear models are less reliable in properly capturing

Figure 6: CDS Spreads Forecasting during Crisis Period



In conclusion, Figure 6 describes the CDS Spreads forecasting during crisis period. The forecasting of Linear regression shows an extraordinary excess increase during crisis time compared with the true spreads. In comparison, Gradient Boost provides smooth prediction which is less volatile but captures the direction of true spreads.

VIII. CONCLUSIONS

In this paper, we analyze the predictability of CDS spreads cross-sectionally and longitudinally using accounting based, market based, and macroeconomics variables. We first compare the

nowcasting and one-step ahed predictive power of traditional credit risk model and various machine learning models, and find that machine learning models can strengthen the prediction accuracy of CDS spreads both cross-sectionally and over time horizons. Among all the machine learning models, ensemble methods including Bagging, Random Forest and Gradient Boosting consistently outperform other interpretable methods. The high cross-sectional and longitudinal precision of ensemble MLs suggests that the nonlinear relationship between economic variables and CDS spreads can be used for constructing "shadow" CDS spreads for those companies without actual CDS.

Using LIME, the "Local Interpretable Model-Agnostic Explanations", we calculate the importance of right hand side variables, which allows insights into the underlying reasoning for why ensemble methods are more accurate in predicting the variable of interest. The application of LIME is particularly interesting in order to shed potential light into the reasoning why non-linear Machine Learning techniques outperform traditional estimation procedures in nowcasting and forecasting CDS spreads during crisis periods. In times of higher volatility and potential structural breaks, prediction accuracy seems particularly driven by non-linear firm specific credit risk and broader economic conditions, which are not properly captured by traditional estimation procedures such as OLS.

To summarize, our results present three valuable contributions to the literature: (1) Machine learning techniques are able to add considerable value in the prediction of CDS spreads. (2) We are able to map the relationship between available market and firm-specific information and CDS spreads to other companies, thus constructing "shadow" CDS spreads for those companies without actual CDS. (3) By using LIME, we are able to unpack some of the "black box" around Machine Learning techniques, and obtain insights into the explanatory power of different variables in predicting the CDS spreads.

References

1. Aldasoro, Inaki, and Torsten Ehlers. “The credit default swap market: what a difference a decade makes.” BIS Quarterly Review, June (2018).

2. Altman, Edward I. "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy." The journal of finance 23.4 (1968): 589-609.

3. Ashbaugh-Skaife, Hollis, Daniel W. Collins, and Ryan LaFond. "The effects of corporate governance on firms credit ratings." Journal of accounting and economics 42.1-2 (2006): 203-243.

4. Bai, Jennie, and Liuren Wu. "Anchoring credit default swap spreads to firm fundamentals." Journal of Financial and Quantitative Analysis 51, no. 5 (2016): 1521-1543.

5. Barboza, Flavio, Herbert Kimura, and Edward Altman. "Machine learning models and bankruptcy prediction." Expert Systems with Applications 83 (2017): 405-417.

6. Bonfim, Diana. "Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics." Journal of Banking & Finance 33.2 (2009): 281-299.

7. Boyrchenko, Nina, A. Costello and O. Shachar. “The Long and Short of It: A Primer on Corporate CDS Positions Data.” Federal Reserve Bank of New York Staff Report No. 879 (2019)

8. Bharath, Sreedhar T., and Tyler Shumway. "Forecasting default with the Merton distance to default model." The Review of Financial Studies 21.3 (2008): 1339-1369.

9. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

10. Cavalcante, Rodolfo C., et al. "Computational intelligence and financial markets: A survey and future directions." Expert Systems with Applications 55 (2016): 194-211.

11. Collin-Dufresne, Pierre, and Bruno Solnik. "On the term structure of default premia in the swap and LIBOR markets." The Journal of Finance 56. 3 (2001): 1095-1115.

12. Chan, Louis KC, Yasushi Hamao, and Josef Lakonishok. "Can fundamentals predict Japanese stock returns?." Financial Analysts Journal 49, no. 4 (1993): 63-69.

13. Christiansen, Charlotte, Maik Schmeling, and Andreas Schrimpf. "A comprehensive look at financial volatility prediction by economic variables." Journal of Applied Econometrics 27, no. 6 (2012): 956-977.

14. Das, Sanjiv R., Paul Hanouna, and Atulya Sarin. "Accounting-based versus market-based cross-sectional models of CDS spreads." Journal of Banking & Finance 33.4 (2009): 719-730.

15. Duan, Jin-Chuan, Jie Sun, and Tao Wang. "Multiperiod corporate default prediction: A

forward intensity approach." Journal of Econometrics 170.1 (2012): 191-209.

16. Duffie, Darrell, and Kenneth J. Singleton. "Modeling term structures of defaultable bonds." The review of financial studies 12.4 (1999): 687-720.

17. Duffie, Darrell, and David Lando. "Term structures of credit spreads with incomplete accounting information." Econometrica 69.3 (2001): 633-664.

18. Duffie, Darrell, Leandro Saita, and Ke Wang. "Multi-period corporate default prediction with stochastic covariates." Journal of Financial Economics 83.3 (2007): 635-665.

19. Emery, Kenneth, R. Cantor, S. Oh, R. Solomon, and P. Stumpp. "Credit Loss Rates on Similarly Rated Loans and Bonds." Moody's Investors Service, Global Credit Research, Special Comment (2004): 12-12.

20. Espinoza, Raphael, Fabio Fornari, and Marco J. Lombardi. "The role of financial variables in predicting economic activity." Journal of Forecasting 31, no. 1 (2012): 15-46.

21. Freund, Yoav, and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting." Journal of computer and system sciences 55.1 (1997): 119-139.

22. Griffin, John M., and Michael L. Lemmon. "Book-to-market equity, distress risk, and stock returns." The Journal of Finance 57.5 (2002): 2317-2336.

23. Guendz, Yalin, and Marliese Uhrig-Homburg. "Predicting credit default swap prices with financial and pure data-driven approaches." Quantitative Finance 11.12 (2011): 1709-1727.

24. Hensher, David A., Stewart Jones, and William H. Greene. "An error component logit analysis of corporate bankruptcy and insolvency risk in Australia." Economic Record 83.260 (2007): 86-103.

25. Huang, Xin, Hao Zhou, and Haibin Zhu. "A framework for assessing the systemic risk of major financial institutions." Journal of Banking & Finance 33.11 (2009): 2036-2049.

26. Huang, Zan, Hsinchun Chen, Chia-Jung Hsu, Wun-Hwa Chen, Soushan Wu. "Credit rating analysis with support vector machines and neural networks: a market comparative study." Decision support systems 37.4 (2004): 543-558.

27. Hillegeist, Stephen A., Elizabeth Keating, Donald Cram, Kyle Lundstedt. "Assessing the probability of bankruptcy." Review of accounting studies 9.1 (2004): 5-34.

28. Jarrow, Robert A., and Stuart M. Turnbull. "Pricing derivatives on financial securities subject to credit risk." The journal of finance 50.1 (1995): 53-85.

29. Jones, Stewart, and David A. Hensher. "Predicting firm financial distress: A mixed logit

model." The Accounting Review 79.4 (2004): 1011-1038.

30. Jones, Stewart, David Johnstone, and Roy Wilson. "An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes." Journal of Banking & Finance 56 (2015): 72-85.

31. Jorion, Philippe, and Gaiyan Zhang. "Credit contagion from counterparty risk." The Journal of Finance 64.5 (2009): 2053-2087.

32. Joy, M., Tollefson, J., 1975. On the financial applications of discriminant analysis. Journal of Financial and Quantitative Analysis, 723–739.

33. Jones, S., Hensher, D.A., 2008. Advances in Credit Risk Modelling and Corporate Bankruptcy Prediction. Cambridge University Press, Cambridge, UK; New York.

34. Lantz, B. (2013). Machine learning with R. Packt Publishing Ltd.

35. Lee, Jongsub, Andy Naranjo, and Guner Velioglu. "When do CDS spreads lead? Rating events, private entities, and firm-specific information flows." Journal of Financial Economics 130. 3 (2018): 556-578.

36. Merton, Robert C. "On the pricing of corporate debt: The risk structure of interest rates." The Journal of finance 29.2 (1974): 449-470.

37. Mittnik, Stefan, Nikolay Robinzonov, and Martin Spindler. "Stock market volatility: Identifying major drivers and the nature of their impact." Journal of Banking & Finance 58 (2015): 1-14.

38. Ohlson, James A. "Financial ratios and the probabilistic prediction of bankruptcy." Journal of accounting research (1980): 109-131.

39. Pal R. Predictive modeling of drug sensitivity. Academic Press; 2016 Nov 15.

40. Son, Youngdoo, Hyeongmin Byun, and Jaewook Lee. "Nonparametric machine learning models for predicting the credit default swaps: An empirical study." Expert Systems with Applications 58 (2016): 210-220.

41. Tang, Dragon Yongjun, and Hong Yan. "Market conditions, default risk and credit spreads." Journal of Banking & Finance 34.4 (2010): 743-753.

42. Welch, Ivo, and Amit Goyal. "A comprehensive look at the empirical performance of equity premium prediction." The Review of Financial Studies 21. 4 (2007): 1455-1508.

Appendix A - Machine Learning Models

Ridge Regression and LASSO Regression

In variable selection and regularization, Ridge and LASSO regressions are two common used methods. They are developed specifically to solve the problem of collinearity in datasets with many variables. They are based on standard linear regression plus a regular term to reduce the model variance. Both Ridge and LASSO regression use all of the variables in the dataset, and adjust the coefficient estimates of non-significant variables to "shrink" towards the zero. The main difference between the two methods is that Ridge keeps all variables, but LASSO allows the penalty to force some parameters to equal zero (thus, LASSO has variable selection features and produces a reduced model). The hyperparameter of the regularization penalty degree is λ.

Cost function of OLS: $f(\omega) = \sum_{i=1}^{m} (y_i - {x_i}^T \omega)^2$

Cost function of Ridge: $f(\omega) = \sum_{i=1}^{m} (y_i - {x_i}^T \omega)^2 + \lambda \sum_{i=1}^{n} {\omega_i}^2$

Cost function of LASSO: $f(\omega) = \sum_{i=1}^{m} (y_i - {x_i}^T \omega)^2 + \lambda \sum_{i=1}^{n} |\omega_i|$

Support Vector Regression (SVR)

Support Vector Regression (SVR) is the regression version of Support Vector Machines classifier (SVM). The original SVM algorithm was invented in 1963 by Vladimir Vopnick and Alexei Zefan Rangers. In SVM, a hyper-plane is used to divide p-dimensional feature space into two halves. A good separation is achieved when the hyper-plane has the largest distance to the nearest training data point of any class. In contrast, SVR is trying to find a hyperplane that minimizes the distance of all data to this hyperplane. The task of the SVR is to cover as many sample points as possible with a fixed-width stripe (the width is controlled by the parameter ϵ and is called margin) so that the total error is as small as possible. The data points in the margin are considered as no error; ξ_i and ξ_i^* capture the error of data points falling out of the stripe from above and below respectively.

The problem of SVR is to solve:

$$\begin{aligned} \min \quad & \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \\ s.t. \quad & y_i - \omega x_i - b <= \epsilon + \xi_i \\ & \omega x_i + b - y_i <= \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* >= 0 \end{aligned}$$

Where C indicates the tolerance level for misclassification degree, which are represented by the slack variables ξ_i and ξ_i^*. For multiple variables case, SVR use kernel functions to enlarge the feature space. In this study we adopt a widely used kernel function called radial kernel.

Neural Network

Neural network is an operation model consisting of a large number of nodes (or neurons) connected to each other. A classic neural network model has at least two layers: input layer and output layer,

and the intermediate hidden layers capture the complexity of the system. Nodes are located on layers, whereby each node represents a specific output function called an activation function. The connection between each two nodes represents a weighted value for the signal passing through the connection, called the weight, which is equivalent to the memory of the artificial neural network. Neural network has many hyper-parameters: the number of layers, number of nodes on each layer, drop rate of layers and so on. Just like other methods, we are using cross-validation method to tune and set the hyper-parameters to obtain a good enough predictive accuracy.

Regression Tree

Regression tree is the regression version of decision tree. The tree method seeks to split the data recursively into subsets so as to find linear solutions within the subset that can improve the overall fit. By dividing data into homogeneous subsets to minimize the overall standard deviation, this method uses a top-down approach to choose the best attribute to divide the space. The basic idea is to construct a tree with the fastest decline in entropy value based on information entropy, whereby the entropy value at the leaf node being zero.

Bagging

Bagging is an abbreviation of bootstrap aggregating. It is a method of sampling with replacement, possibly with duplicate samples. Bagging starts by extracting the training set from the original sample set. Each round draws n training observations from the original sample set using Bootstraping. A total of k rounds of extraction were performed, resulting in k independent training sets. Each time a training set is used to obtain a model, a total of k models are obtained for k training sets. For the regression problem, the mean value of the above model is calculated as the final result, with all models having the same importance.

Random Forest

Random forest was first proposed by Leo Breiman (2001) . It is a classifier/regression model containing multiple decision trees, and is built to deal with the overfitting problem of decision/regression trees. The tree method has good in-sample performance but relatively bad out-of-sample performance. Random forests assist to solve the problem by combining the concept of bagging with random feature selection (Pal, 2017). Random Forest further conducts random feature selection on the subsamples generated from original dataset, and estimate a regression tree on each subsamples.. When forecasting, each tree predicts a result, and all the results are weighted to avoid overfitting.

Gradient Boosting

The Boosting algorithm optimizes the regression results through a series of iterations. The idea behind boosting is to combine the outputs of many models to produce a powerful overall voting committee. AdaBoost is an abbreviation of "Adaptive Boosting", which was put forward by Yoav Freund and Robert Schapire in 1995. In the Adaboost process the first model is trained on the data where all observations receive equal weights. Those observations misclassified by the first weak model will receive a higher weight, while correct observations have a lower weight. The newly

added second model will thus focus more on the error of first model. Such iteration keeps adding weak models until the desired low error rate is achieved.

Gradient Boosting is the generalized version of AdaBoost. Gradient Boosting selects the direction of the gradient drop during iteration to ensure that the final result is best.The loss function is used to describe the degree of "flight" of the model. It is assumed that the model is not overfitted. The greater the loss function, the higher the error rate of the model. If our model can make the loss function continue to decline, then our model is constantly improving, and the best way is to let the loss function in the direction of its gradient.

Appendix B - Variables Description

Variable	Abbrev.	Description	Source
I. Accounting-based variables			
A. Capitalization: measures the debt component of a firm's total capital structure.			
Capitalization Ratio	capital_ratio	Total Long-term Debt as a fraction of the sum of Total Long-term Debt, Common/Ordinary Equity and Preferred Stock	WRDS
Long-term Debt/Invested Capital	debt_invcap	Long-term Debt as a fraction of Invested Capital	WRDS
Common Equity/Invested Capital	equity_invcap	Common Equity as a fraction of Invested Capital	WRDS
Total Debt/Invested Capital	totdebt_invcap	Total Debt (Long-term and Current) as a fraction of Invested Capital	WRDS
B. Efficiency: captures the effectiveness of firm's usage of assets and liability			
Asset Turnover	at_turn	Sales as a fraction of the average Total Assets based on the most recent two periods	WRDS
Payables Turnover	pay_turn	COGS and change in Inventories as a fraction of the average of Accounts Payable based on the most recent two periods	WRDS
Receivables Turnover	rect_turn	Sales as a fraction of the average of Accounts Receivables based on the most recent two periods	WRDS
Sales/Stockholders Equity	sale_equity	Sales per dollar of total Stockholders' Equity	WRDS
Sales/Invested Capital	sale_invcap	Sales per dollar of Invested Capital	WRDS
C. Financial Soundness & Solvency: captures the firm's ability to meet long-term obligations			
Cash Flow/Total Debt	cash_debt	Operating Cash Flow as a fraction of Total Debt	WRDS
Cash Balance/Total Liabilities	cash_lt	Cash Balance as a fraction of Total Liabilities	WRDS
Cash Flow Margin	cfm	Income before Extraordinary Items and Depreciation as a fraction of Sales	WRDS
Total Debt/EBITDA	debt_ebitda	Gross Debt as a fraction of EBITDA	WRDS
Long-term Debt/Book Equity	dltt_be	Long-term Debt to Book Equity	WRDS
Free Cash Flow/Operating Cash Flow	fcf_ocf	Free Cash Flow as a fraction of Operating Cash Flow, where Free Cash Flow is defined as the difference between Operating Cash Flow and Capital Expenditures	WRDS
Interest/Average Long-term Debt	int_debt	Interest as a fraction of average Long-term debt based on most recent two periods	WRDS

Interest/Average Total Debt	int_totdebt	Interest as a fraction of average Total Debt based on most recent two periods	WRDS
Long-term Debt/Total Liabilities	lt_debt	Long-term Debt as a fraction of Total Liabilities	WRDS
Total Liabilities/Total Tangible Assets	lt_ppent	Total Liabilities to Total Tangible Assets	WRDS
Short-Term Debt/Total Debt	short_debt	Short-term Debt as a fraction of Total Debt	WRDS
Total Debt/Equity	de_ratio	Total Liabilities to Shareholders' Equity (common and preferred)	WRDS
Total Debt/Total Assets	debt_assets	Total Debt as a fraction of Total Assets	WRDS
Total Debt/Capital	debt_capital	Total Debt as a fraction of Total Capital	WRDS
After-tax Interest Coverage	intcov	Multiple of After-tax Income to Interest and Related Expenses	WRDS
Interest Coverage Ratio	intcov_ratio	Multiple of Earnings Before Interest and Taxes to Interest and Related Expenses	WRDS
D. Profitability: measures the ability of a firm to generate profit			
After-tax Return on Average Common Equity	aftret_eq	Net Income as a fraction of average of Common Equity based on most recent two periods	WRDS
After-tax Return on Total Stockholders Equity	aftret_equity	Net Income as a fraction of average of Total Shareholders' Equity based on most recent two periods	WRDS
After-tax Return on Invested Capital	aftret_invcapx	Net Income plus Interest Expenses as a fraction of Invested Capital	WRDS
Effective Tax Rate	efftax	Income Tax as a fraction of Pretax Income	WRDS
Gross Profit Margin	gpm	Gross Profit as a fraction of Sales	WRDS
Net Profit Margin	npm	Net Income as a fraction of Sales	WRDS
Operating Profit Margin After Depreciation	opmad	Operating Income After Depreciation as a fraction of Sales	WRDS
Operating Profit Margin Before Depreciation	opmbd	Operating Income Before Depreciation as a fraction of Sales	WRDS
Pre-tax Profit Margin	ptpm	Pretax Income as a fraction of Sales	WRDS
Return on Assets	roa	Operating Income Before Depreciation as a fraction of average Total Assets based on most recent two periods	WRDS
Return on Capital Employed	roce	Earnings Before Interest and Taxes as a fraction of average Capital Employed based on most recent two periods, where Capital Employed is the sum of Debt in Long-term and Current Liabilities and Common/Ordinary Equity	WRDS

Return on Equity	roe	Net Income as a fraction of average Book Equity based on most recent two periods, where Book Equity is defined as the sum of Total Parent Stockholders' Equity and Deferred Taxes and Investment Tax Credit	WRDS
Gross Profit/Total Assets	gprof	Gross Profitability as a fraction of Total Assets	WRDS
E. Valuation: estimates the attractiveness of a firm's stock (overpriced or underpriced)			
Book/Market	bm	Book Value of Equity as a fraction of Market Value of Equity	WRDS
Shillers Cyclically Adjusted P/E Ratio	capei	Multiple of Market Value of Equity to 5-year moving average of Net Income	WRDS
Dividend Yield	divyield	Indicated Dividend Rate as a fraction of Price	WRDS
Dividend Payout Ratio	dpr	Dividends as a fraction of Income Before Extra. Items	WRDS
Enterprise Value Multiple	evm	Multiple of Enterprise Value to EBITDA	WRDS
Price/Cash flow	pcf	Multiple of Market Value of Equity to Net Cash Flow from Operating Activities	WRDS
P/E (Diluted, Excl. EI)	pe_exi	Price-to-Earnings, excl. Extraordinary Items (diluted)	WRDS
P/E (Diluted, Incl. EI)	pe_inc	Price-to-Earnings, incl. Extraordinary Items (diluted)	WRDS
Price/Operating Earnings (Basic, Excl. EI)	pe_op_basic	Price to Operating EPS, excl. Extraordinary Items (Basic)	WRDS
Price/Operating Earnings (Diluted, Excl. EI)	pe_op_dil	Price to Operating EPS, excl. Extraordinary Items (Diluted)	WRDS
Price/Sales	ps	Multiple of Market Value of Equity to Sales	WRDS
Price/Book	ptb	Multiple of Market Value of Equity to Book Value of Equity	WRDS
Forward P/E to 1-year Growth (PEG) ratio	peg_1yrforward	Price-to-Earnings, excl. Extraordinary Items (diluted) to 1-Year EPS Growth rate	WRDS
Forward P/E to Long-term Growth (PEG) ratio	peg_ltgforward	Price-to-Earnings, excl. Extraordinary Items (diluted) to Long-term EPS Growth rate	WRDS
F. Others:			WRDS
Avertising Expenses/Sales	adv_sale	Advertising Expenses as a fraction of Sales	WRDS
Labor Expenses/Sales	staff_sale	Labor Expenses as a fraction of Sales	WRDS
Accruals/Average Assets	accrual	Accruals as a fraction of average Total Assets based on most recent two periods	WRDS
Research and Development/Sales	rd_sale	R&D expenses as a fraction of Sales	WRDS
G. Controls:			

Firm size	avsize	Proxied by the total asset as a fraction of average total asset	COMPUSTAT
Industry dummy		A series of 7 dummy variables coded for each major industry defined by SIC (Insurance, Mining, Manufacturing, Retail Trade, Wholesale Trade, Services, Transportation & communication).	CRSP
Firm rating	rating	Long-term credit rating assigned to the entity by S&P, Moody's or Fitch: including A, AA, AAA, B, BB, BBB	MARKIT
CDS Recovery rate	recovery	Pre-populated based on the recovery rate set for the Ticker + Tier combination	MARKIT
II. Market-based variables			
A. Equity Market Variables			
Stock return	ret	Monthly Log stock return	CRSP
Stock realized variance	lrvar	Monthly Log realized variance	CRSP
Change of stock realized variance	lrvar_c	Change of the log realized variance	CRSP
Trading volumn	avtrd	Monthly trading volumn of firm's stock as a fraction of average trading volumn of sample time	CRSP
S&P 500 return	mreturns	Monthly log returns of the S&P 500	CRSP
CBOE Market Volatility Index	mvix	Monthly log returns of the implied volatility of S&P 500 index options	CBOE
Excess Return on the Market	mktrf	Fama–French's market factor: U.S. stock market return minus one-month T-Bill rate	Fama French
Small-Minus-Big Return	smb	Fama–French's SMB factor: Return on small stocks minus return on big stocks	Fama French
High-Minus-Low Return	hml	Fama–French's HML factor: Return on value stocks minus return on growth stock	Fama French
Momentum	umd	Fama–French's momentum factor: Average return on the two high prior return portfolios minus the average return on the two low prior return portfolios	Fama French
Levels of aggregate liquidity	ps_innov	Pastor-Stambaugh's level Liquidity measure: cross-sectional average of individual-stock liquidity measures	Pastor Stambaugh

Innovations in aggregate liquidity	ps_level	Pastor-Stambaugh's innovation Liquidity measure: the residual of a second-order autoregression of level liquidity measure	Pastor Stambaugh
Traded liquidity factor	ps_vwf	Pastor-Stambaugh's Liquidity Factors: Return on stocks with low liquidity minus return on high liquidity stocks	Pastor Stambaugh
Distance to default	did	The market-based risk measure developed by Merton(1975) and simplified by Bharath & Shumway.	Bharath & Shumway
B. Analyst Forecasting variables			
Median Recommendation	medrec	The median of analysts' Recommendation scale. The scale is: 1. Strong Buy 2. Buy 3. Hold 4. Underperform 5. Sell	IBES
Number of Recommendations	numrec	Number of Analysts Recommendations	IBES
One year forward Median Estimate EPS	eps_medest	The median estimate of one year forward EPS	IBES
C. Interest Rates and Spreads			
One-month T-Bill rate	rf	One-month T-Bill rate	Goyal Welch
Three-month T-Bill rate	t_b	Three-month T-Bill rate	Goyal Welch
Rel.T-Bill Rate	rtb	T-Bill rate minus its 12 month moving average	Goyal Welch
Long Term Bond Return	ltr	Rate of return on 10 year government bonds	Goyal Welch
RelBond Rate	rbr	Long-term bond yield minus its 12 month moving average	Goyal Welch
Term Spread	t_s	Difference of long-term bond yield and three-month T-Bill rate	Goyal Welch
Default spread	def	Measure of default risk: BAA minus AAA corporate bond yields	Goyal Welch
TED spread	ted	Measure of illiquidity: LIBOR minus T-Bill rate	Datastream
III. Macroeconomic Variables			
Unemployment rate	unrate	Country's official annual unemployment rate.	Datastream
Industrial Production Growth, YoY	ipga	Year-over year (log) growth rate of U.S. industrial production	Datastream
Industrial Production Growth, Monthly	ipm	Monthly (log) growth rate of U.S. industrial production	Datastream
Inflation Rate, YoY	infa	Year-over year (log) growth rate of the U.S. consumer price index	Datastream
Inflation Rate, Monthly	infm	Monthly (log) growth rate of the U.S. consumer price index	Datastream
Return CRB Spot	crb	Commodity price spot index; annual log difference	Datastream

Federal budget	avdgt	Federal budget as a franction of average federal budget of sample time	Datastream
Capacity Utilization, Monthly	cap	Level to which the productive capacity is used	Datastream
Diffusion Index	diff	Philadelphia Fed Business Outlook Survey Diffusion Index	Datastream
Housing Starts	h_s	Monthly change in housing started	Datastream
M1 Growth, YoY	m1a	Year-over-year (log) growth rate of U.S. M1	Datastream
M1 Growth, Monthly	m1m	Monthly (log) growth rate of U.S. M1	Datastream
Orders, YoY	orda	New orders, consumer goods and materials; year-to-year growth rate	Datastream
Orders, Monthly	ordm	New orders, consumer goods and materials; monthly growth rate	Datastream
Chicago PM Business Barometer	pmbb	Leading indicator of economic health; survey of purchasing managers	Datastream
ISM PMI	pmi	Monthly change in purchasing manager index	Datastream
Consumer Confidence	conf	Monthly change in consumer confidence index	Datastream
Consumer Sentiment	sent	Monthly change in University of Michigan consumer sentiment	Datastream