Exploring Credit Spread Opportunities Using ETF Portfolios, A Machine Learning Approach

Abstract— **We utilize machine learning approach to predict BofAML US Corporate Investment-Grade Option-Adjusted Spread (OAS) Index and subsequently apply our findings to profitable systematic trading strategies.**

Economically, a large number of potential explanatory variables from different aspects have been examined, and many of which have been included in our analysis. Technically, a variety of effective machine learning models, including elastic net, general addictive model, random forest and gradient boosting, are implemented. By investigating the partial dependency, we attempt to resolve the non-transparency of these models and unfold the explanatory power of selected variables.

Based on predictions from our machine learning models, we construct tradable ETF portfolios to replicate credit spread, and develop profitable strategies. One of the strategies yields 40.96% annualized return with 1.4 Sharpe Ratio using prediction of AAA rating, which demonstrates the effectiveness, accuracy and practicability of our credit spread prediction.

I. Introduction

Corporate bond credit spread provides additional yield over treasuries to compensate investors for a variety of risks: 1) Default Risk - the risk that a debtor may be unable to make a payment or fulfill contractual obligations. 2) Market Risk - the exposure of bond price with respect to market variables, such as stock prices, interest rates, commodity prices. 3) Liquidity Risk - the risk that a credit bond transaction cannot be executed at market price. We believe these risks, which eventually drive credit spread, can be explained by various factors. Thus, sets of explanatory variables have been selected and tested in order to characterize above mentioned risks impacting credit spread.

II. Data

We collect economic and financial data from Jan 1997 to Feb 2019 to predict movements in credit spreads. The data and explanatory variables, which are divided into several categories based on their nature and frequency, have been obtained from different sources. They are summarized in Table III in appendix.

A. Credit Spread Data (Response)

BofAML US Corporate Investment-Grade Option-Adjusted Spread (OAS) Index tracks performance of USD denominated investment grade rated corporate bonds for all maturities. OASs are the calculated spreads between a computed index of all bonds in a given rating category and spot Treasury curve, net of embedded options. The sample includes 4 series for bonds rated AAA, AA, A and BBB. It's supposed to better represent the systematic risk component of credit spread than effective yield minus Treasury rates, because embedded options are mostly idiosyncratic [1].

We focus on predicting weekly spread changes, striking a balance between noise reduction and data availability since many explanatory variables, primarily the ones indicating macroeconomic trends and business cycles, are published monthly. Concretely, the models use explanatory variables to predict the weekly changes of spread after their announcements. These variables are divided into several categories such that their feature importance can be better interpreted.

B. Explanatory Credit Spread Data

Investment-grade Credit Spread Slope (Spread Slope) To use information contained in credit spread term-structure, we calculate the slope of credit spread as difference between BofAML US Corporate 10-15 Yr and 1-3 Yr OAS Index.

Speculative-grade Option Adjusted Spread (HYM) BofAML US High Yield Master II OAS is an equivalent index tracking credit spreads of speculative rated bonds (all ratings below BBB). Presumably speculative bonds are more sensitive to factors overshadowing corporate bond market, the rich information within their credit spread series can thus have desired explanatory power. The

difference between Investment-grade yields and Speculative yields are also calculated **(IG_HYM or AAA_HYM, AA_HYM, A_HYM, BBB_HTM)** and fed to models.

C. Technical Analysis Data

To make the most out of information contained in credit spread (OAS) series themselves, we further calculate the following technical indicators from their daily series.

Trailing 5-week Log Change (ret5) Rolling 5-week log change, a momentum-like indicator.

Trailing 5-week Realized Volatility (RV) Rolling realized volatility of OAS daily log changes.

Trailing 5-week Skewness of Spread Change (Skew) Skewness of the distribution of OAS daily log changes within 25-day window before a given date.

D. Macroeconomic Indicators

Non-farm Payroll (NFP) the number of jobs added or lost in the economy over the last month, excluding farming industry.

Consumer Price Index (CPI) a measure of the average monthly change in the price for goods and services paid by urban consumers.

Conference Board Composite Index of Coincident Indicators (COI) summary statistics for the U.S. economy, constructed by averaging 4 individual components: Employment, Personal income, Industrial production, Manufacturing and Trade Sales. Historically, this index have paralleled aggregate economic activity.

E. Financial Situation Indicators

Chicago Fed National Financial Conditions Index (NFCI) a measure of U.S. financial conditions in money markets, debt and equity markets and the traditional and shadow banking systems. Positive values indicates tighter-than-average financial conditions. The index can be broken into sub-components reflecting the themes of risk, credit, and leverage.

St. Louis Fed Financial Stress Index (STLFSI) a measure of financial stress in the markets and is constructed from 18 weekly data series covering interest rates, yield spreads and other financial indicators. Values below zero suggest below-average financial market stress.

Commercial and Industrial Loans (TOTLAON) commercial and industrial Loans of all commercial banks.

F. Financial Markets Variables

US Treasury Yield (Treasury) US Treasury yields at 3M, 1Y, 2Y, 3Y, 5Y and 10Y maturities.

TED Spread (TED) Treasury-Eurodollar rate represents the difference between 3-month Treasury bill and 3-month USD LIBOR.

Merrill Lynch Treasury 3-month Option Volatility Estimate Index (MOVE3M) a measure tracking implied volatility on 3-month Treasury options, commonly referred to as 'MOVE'.

RUSSELL 2000 Index (RUSSELL) daily closing price series for RUSSELL 2000 Index; it's preferred over SP500 index because it covers a far wider range of cooperates, thus has more meaningful implication for lower rated bonds.

CBOE VIX Index (VIX) indicator of 30-day expected volatility of the U.S. stock market, derived from of SP500 option prices.

Crude Oil Price Benchmark (WTI) West Texas Intermediate price settlement point on NYSE, a grade of crude oil used as a benchmark in oil pricing. Its involvement intends capture energy prices' influence on credit spread as a cost to economic activities.

G. Alternative Data / Market Sentiment

The AAII Investor Sentiment Survey (Bullish, Neutral, Bearish) a polling result, giving percentage of individual investors who are bullish, bearish, and neutral on the stock market for the next six months, provided by *Quandl Inc.* The sum of these three indicators is one. To avoid perfect multi-collinearity, we only incorporate Bullish and Bearish in the models.

H. Fama-French Factor Returns

As an analogy to stock risk premium, credit spread indicates the extra compensation investors collect by holding risky assets. Consequently, Fama-French factor returns are incorporated in our models, in anticipation that the coincidence of financial markets would transfer their explanatory power for stock returns to credit bond market.

Small-minus-Big (SMB) average return on small stock portfolios minus the average return on big stock portfolios.

high-minus-Low (HML) average return on value portfolios minus the average return on growth portfolios.

Robust-minus-Weak (RMW) average return on robust operating profitability portfolios minus the average return on weak operating profitability portfolios.

Conservative-minus-Aggressive (CMA) average return on conservative investment portfolios minus the average return on aggressive investment portfolios.

To accomplish our prediction task, the aforementioned data sets are adjusted based on their release date or consequent revisions after initial announcement, preserving what-data-was-known-then for our manipulation.

III. Data Preprocessing

The sampling contains 1,109 weekly observations from Jan 1997 to Feb 2019. A 70%:30% training-test split was made, at July-20 2012.

A. Seasonality Test

We first attempt to detect possible seasonality pattern in the credit spread time series to decompose the possible source of predictive power. A regression model using dummy variables is constructed representing each months as predictors, to predict the series of interest.

TABLE I

Test of Seasonality (Rating A Credit Spread)

Month	t-value	p-value
January	0.446	65.6%
February	0.265	79.1%
March	0.274	78.4%
May	-0.149	88.2%
June	-0.289	77.3%
July	-0.236	81.4%
August	-0.061	95.2%
September	0.452	65.1%
October	1.303	19.3%
November	1.133	25.8%
December	1.024	30.6%

As seen from Table I (April was removed when performing the test), when examining credit spread of credit rating A, we see none of the dummy variables are proven to have predictive power under this measure.

B. Time Series Stationarization

Stationarity of time series data is to be guaranteed before feeding the predictive models. First, variables indicating or resembling financial securities prices (including OASs) are taken logarithm. Then using the standard approach, we assume the non-stationarity comes from the time series being autoregressive of some order p. Then using the augmented Dickey-Fuller test as described below, we proceed to test the whether the chosen time series data are stationary (existence of unit root).

$$\begin{aligned}\Delta y_t =& \alpha + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \dots \\ &+ \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t,\end{aligned} \tag{1}$$

where $\gamma = 0$ is the Null hypothesis and $\gamma < 0$ is the alternative hypothesis. If the null hypothesis was rejected (p-value less than α), the time series does not have a unit root and thus it is stationary. Unstationary series are differenced until they exhibits stationarity at 5% significance level.

C. Feature Correlation Heat Map

After aforementioned manipulation, training set exhibits the correlation relationships in Firgure 1.

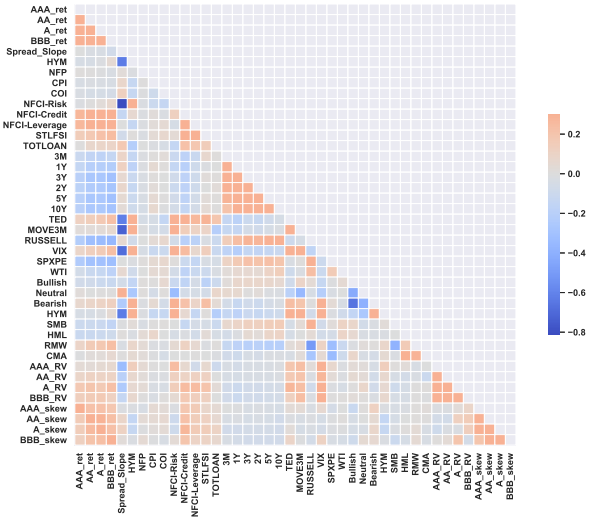

Fig. 1. Feature correlation heat map, training set

D. Dimension Reduction

High correlation within 3 subgroups of data series would bring collinearity and high feature dimensions to the models and destabilize subsequent predictions: Spread log-change (RET), Technical Analysis Data (RV and Skew) and Treasury Yield Data (3M-10Y). Standard trick of Principle Components Analysis for these subgroups can address such problems, and orthogonalize relevant features in the meantime.

We extracted first 3 principle components from Skewness and Realized Volatility data respectively, labeled PCA1, PCA2, PCA3; and first 3 principle components from Treasury Yield data, labeled Treasury-level, Treasury-slope and Treasury-cvx.

IV. Models

This section gives give a brief description of types of models we used, including linear model (Elastic Net Regularization), non-linear additive model (General Addictive Model) and Tree-Based Model (Random Forest, Gradient Boosting). They are preferred over more flexible models, because the number of observations are limited, and presumably signal-to-noise ratio within the sample is low.

In addition, to accommodate time-series data, nested cross validation is applied to tune hyperparameters of our models; different models are combined through averaging to produce an aggregated prediction.

A. Elastic Net Regularization

The first type is linear model. With proper regularization method, linear model can perform variable selections by itself. In this case, Elastic Net Regularization is used, which can be described by the following equation,

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(||y-\beta\mathbf{X}||^2+\lambda_1||\beta||_1+\lambda_2||\beta||_2^2). \quad (2)$$

where β is the linear model parameters.

Even though this method is more computationally expensive, compared to Lasso (L_1) and Ridge (L_2) regularization, the relatively small sample size of the data suggests that we can still explore the entire space of regularization parameters with reasonable computational cost. However, given probable non-linear nature of feature interactions and predictive power, following non-linear models are also examined.

B. General Addictive Model

Generalized Additive Model (GAM) extends multivariate linear regression by allowing non-linear relationships between the predictors and the response variable.

$$y = \beta_0 + \sum_{i=1}^{p} f_i(\mathbf{X}_i) + \varepsilon, \quad (3)$$

We imposed cubic smoothing splines ($f_i(x) = c_{0i} + c_{1i}x + c_{2i}x^2 + c_{3i}x^3$) on each feature to capture such non-linearity.

One major limitation of GAM comes from its additive assumption, which implies that each predictor's contribution should be added upon each other. More complicated interactions between features are left out. To account for these missing interactions, we also consider tree-based model.

C. Random Forest

The essence of tree-based models for regression is to partition the predictor space into different regions and fit simple model within each one. Specifically, we model spread with a constant c_m in each partition:

$$f(x) = \sum_{m=1}^{M} c_m \mathbb{I}(x \in R_m), \quad (4)$$

where $\{R_m\}_{1\le m\le M}$ is the set of partitions and c_m is the arithmetic average of the response within the partition R_m.

Unlike the linear model and additive model, tree-based models pick up interactions between predictors without specifying them. However one major problem with trees is their instability or high variance, due to hierarchical nature of the fitting process. To alleviate this, we apply Random Forest, which provides improvements over simple tree-based models via bagging and random selection of the predictors. By taking bootstrap samples to construct multiple different trees and using only a small subset ($\sqrt{p}$) of the predictors at each split, random forest model creates a massive collection of trees. Then by exploiting the low correlation between trees, random forest algorithm generates a massive reduction in variance and avoids preference for some variables.

D. Gradient Boosting

We also implement Gradient Boosting Model. Instead of averaging trees to reduce variance, Boosting Tree Model improves simple tree model by minimizing bias (the expected loss over some loss function $\mathscr{L}$). GBM performs prediction by building an ensemble containing a series of weak learners, each of which is built toward reducing the total loss by fitting to the gradient of the previous ensemble. Each iteration can be described by the following equation,

$$f_m(\mathbf{X}) = F_{m-1}(\mathbf{X}) - \gamma_m \sum_{i=1}^{n} \nabla_{f_{m-1}} L(y_i, F_{m-1}(x_i)). \tag{5}$$

By using an ensemble of high bias and low variance learners, GBM guarantees low variance by its construct and further reduce the bias at each iteration. However GBM has its own downfall. Unlike random forest which converges to a stable result (as it is averaging the prediction of strong learners), GBM overfits easily given its nature to optimized toward the minimum training loss and its high dimensional parameter space. However, if the model is tuned properly, GBM can generally outperform random forest.

E. Nested Cross Validation

To prevent overfitting, a proper validation technique is necessary to tune hyper-parameters and measuring model performance within training set. Generally, for most machine learning problems, cross validation is preferred over setting-aside validation set since it retains more data for training and testing. However, traditional cross-validation (k-fold and leave-one-out) methods do not apply to time series prediction problem. As chronological nature of data dictates that training - CV split cannot be completely random nor arbitrary without causing data leakage.

Therefore we apply a Nested Cross Validation (NCV) technique, e.g Figure 2, which was proposed in [2]. Unlike traditional cross validation methods which perform train-test split randomly, NCV performs a series of partitions, which split the data into successive folds according to their chronological order. For example, in our model, we perform a five-fold NCV on the data, we start off by splitting the data into five equally weighted

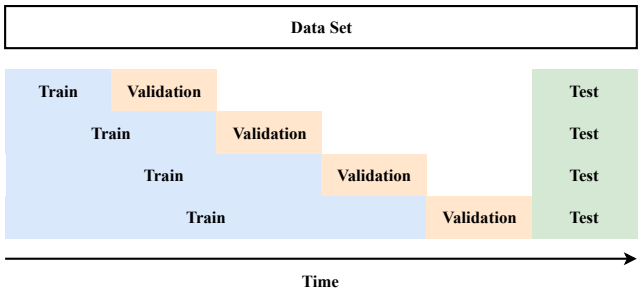

Fig. 2. Nested Cross Validation Procedure

folds according to their order. For the first iteration, we use the first fold as train data and the following fold as the validation data. For each subsequent iteration, we continue to expand the train data by adding in the previous training data and use the subsequent fold as the validation data, until all data has been used, and the error at each splits are then averaged to be the final validation error.

F. Ensemble Averaging

Rather than choosing a single model to perform prediction, we want to combine the regularization and parameter selection of the elastic net, the non-linearity of the GAM, the interaction structure of the random forest and the sequential bias reduction feature of gradient boosting. Ensemble averaging then becomes our best choice to best preserve the initial model structure and the robustness of the method.

Suppose we have $\hat{g}_m, m = 1, \ldots, M$ be a discrete set of estimators from each of the M models. Let $\{w_m\}$ be a set of weights for the mth estimator. Let $\mathbf{w} = (w_1, \ldots, w_M)$ be the vector of weights. Typically we require $0 \leq w_m \leq 1$ and $\sum_{m=1}^{M} w_m = 1$. Our final estimator is given by:

$$\hat{g}_{\mathbf{w}}(\mathbf{X}) = \sum_{m=1}^{M} w_m \hat{g}_m(\mathbf{X}). \tag{6}$$

where g_m is the mth model and w_m is its corresponding weight.

A naive method to combine the model is to take the arithmetic average. However, we want to put more weight on models with lower MSE(mean squared error) on the validation set, $w_m \propto \frac{1}{\text{MSE}_{m,\text{validation}}}$.

V. Model Results

This section examines the performance of each models and summarizes some of the common characteristics of predictors with the strongest predictive power. Also we discuss the potential problems of machine learning techniques and their possible solutions.

A. Model Performance

According to aforementioned train-test split, we start off by looking at the credit spread with class A rating. The performance of all four models and their ensemble averaging model are measured within the test data . Using MSE as a measure, GBM exhibit the strongest performance among all four individual models from Figure 3. The ensemble averaging further reduces the MSE by 3%.

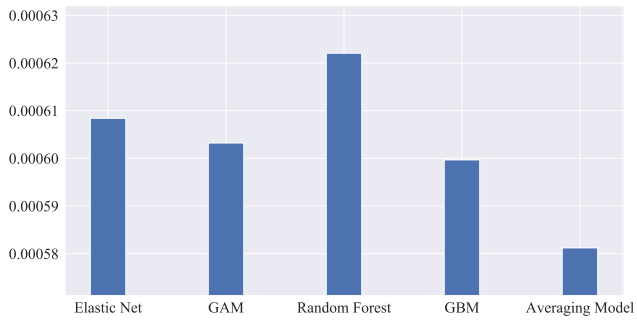

Fig. 3. Mean Square Error of All Models, Prediction of Rating A

Apart from MSE, as our main interest is to predict direction of movement in credit spread and possible application of this predictive model using the sign of its prediction as a trading signal, we define our own utility function,

$$U(\Delta y_t, \Delta \hat{y}_t) = |\Delta y_t| * \text{sign}(\Delta y_t \Delta \hat{y}_t), \qquad (7)$$

where Δy_t is realized change in credit spread and $\Delta \hat{y}_t$ is predicted change in credit spread. As defined in (7), if realized direction is as predicted, the utility is positive (negative if the prediction is off) and its magnitude is higher if the magnitude of change is high. It follows that the cumulative utility over a certain time period become a measure of how the model's directional prediction has performed over that time period.

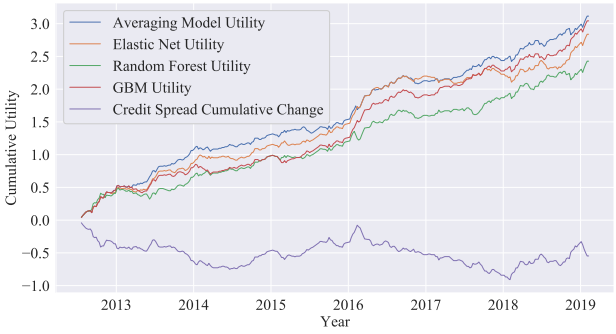

Fig. 4. Cumulative Utility of All Models Over Time, Prediction of Rating A.

In Figure 4, Cumulative log change of credit spread shown here is equivalent to utility of a naive model predicting $\delta y_t > 0$ all the time. One can note that the cumulative utilities have consistently increased across all predictive models. the ensemble averaging model has also performed more consistently, especially during the period 2013 - 2014, when most of the other models suffer from a sudden down fall.

B. Cracking the Black-box

Machine learning models suffer from non-transparency, which means it is hard to identify the relation between the predicted variable and the predictors. We mitigate this problem by using feature importance heatmap and partial dependence plot.

To make comparisons between models, we use MSE reduction ranking in the following heatmap.

Darker color implied a higher importance, or specifically higher MSE reduction ranking within the model. We find several features with high importance across almost all models. These features include RUSSELL, NFCI-Credit, NFCI-Leverage, IG_HYM, RET_PCA_1.

High importance of these predictors is expected. RUSSELL is the log weekly return of RUSSELL 2000, which represents the equities market condition. When the equities market is bearish, it is more likely for bonds to default. The opposite is true when the equities market is bullish. As a result, we interpret this predictor as the reflection of default risk as well as market risk incorporated in the credit spread.

Unlike RUSSELL 2000 return describes the relation between credit spread and equities market,

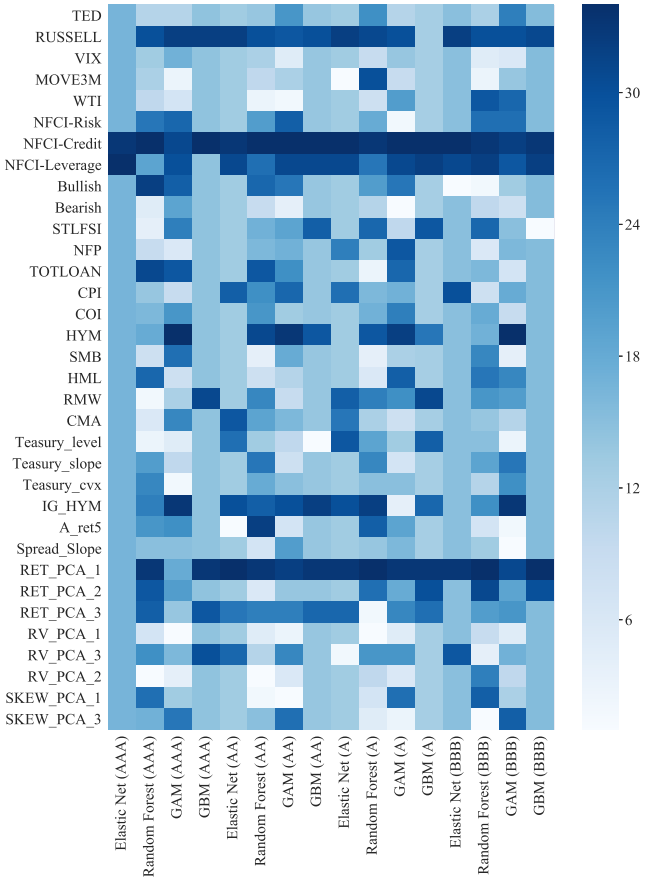

Fig. 5. Features Importance Heatmap Across Models and Ratings. Features are ranked by MSE reduction they contributed to each model. Importance of 34 means the variable provide most MSE reduction, importance of 1 means it's the least important.

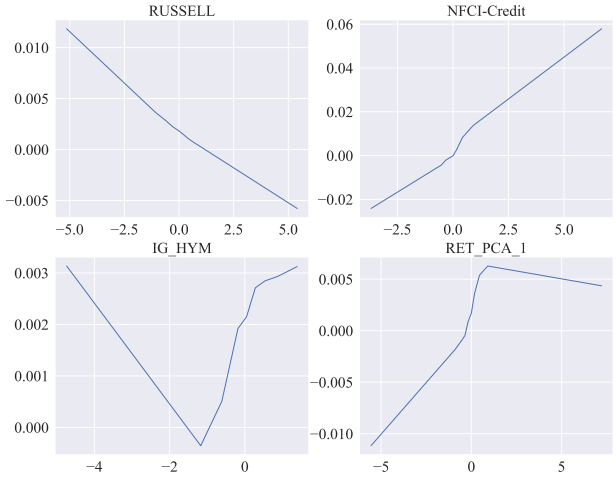

Fig. 6. Partial Dependence Plot, Prediction of Rating A. Y-axis, RUSSELL, IG_HYM and RET_PCA on log change scale. Partial dependence shows the marginal effect of one feature x_S on prediction of machine learning model $\hat{f}$, defined as $\hat{f}_{x_S}(x_S) = E_{x_S}(\hat{f}(x_S, x_C))$

NFCI-Credit and NFCI-Leverage describe the relation between credit spread and financial stress. NFCI-Credit is a leading index to measure the overall credit condition. The NFCI-Leverage measures the overall leverage of US economy. Both variables can be linked to the market risk embedded in credit spread.

IG_HYM is the difference between the investment grade credit spread and the high yield credit spread. This variable can explain the additional default risk of lower rated bond versus higher rated. RET_PCA_1 is the first principle component of log change of four IG spreads. The model concludes a quadratic relationship between these 2 measures and the response.

C. Confronting Regime Shifts Issues

Regime shifts is a way of describing structural changes in a data series. In our case, the underlying distribution that generates the credit spread data may vary over time.

There are many methods to detect regime shifts, such as hidden Markov model. However, the parameter estimation for such models require large amount of data and it is hard to specify the number of regimes because the choice should be based onp economic arguments instead of from data itself. From Figure 4, no regime shift in the test set is detected because the cumulative utility function is positive and increasing over time, which means the model's directional prediction performs well. As we obtain more data set in future, we can detect regime shift if there is a drastic change in the cumulative utility function, and should retrain our models accordingly.

VI. Trading Strategy

This section focuses on application of credit spread predictive models in trading strategy formulation in credit and treasury ETFs.

A. Replication of Credit Spread

Although credit spreads are not tradable on exchange, the change of the credit spreads can be tracked with some portfolio of bond ETFs and treasury ETFs, since these ETFs should contain price information of relevant corporate bonds and treasuries. In order to create such replicating

portfolio, we select ETFs based on following requirements: (1) sufficiently liquid (2) has abundant historical data (3) has easy market access. As a result, we pick 3 representative US Investment-grade Bond ETFs: QLTA (ISHARES TR/AAA A RATED CORP Bond), LQD (iShares IBoxx Invest Grade Corp Bond Fund) and VCLT (Vanguard Long-Term Corporate Bond ETF), 2 US Treasury ETFs: TLT (iShares Barclays 20+ Yr Treasury Bond) and VGSH (Vanguard Short-Term Treasury ETF). Price series of these ETFs are obtained from 2012-02-26 to 2019-02-25. As for their weights, our method is to use a rolling-window linear regression with window size of 50 weeks. Rolling window regression captures the latest variability better than simple OLS:

$$\Delta_d = \sum_{i=1}^{p} \beta_{id} \mathbf{e}_{id} + \varepsilon_{id}, \quad d = \{1, \ldots, D\} \tag{8}$$

Δ_d vector is log change of credit spread in the d'th window and $\mathbf{e}_{id}$ is the return of the i'th bond ETFs in the d'th window. And β_{id} is the weights of ETFs in the portfolio in the d'th window. Suppose the d'th window covers week $t-50$ to week t, then the $t+1$'th week predicted credit spread log change is calculated by:

$$\hat{\Delta}_{t+1} = \sum_{i=1}^{p} \beta_{id} e_{i,t+1} \tag{9}$$

By trading this replicating portfolio, we can trade the credit spread indirectly.

First, we generate the weekly trading signal by taking sign of ensemble averaging prediction:

$$s_{t+1} = \begin{cases} 1, & \text{if } \hat{g}_{t+1}(\mathbf{w}) > 0 \\ -1, & \text{if } \hat{g}_{t+1}(\mathbf{w}) \leq 0 \end{cases} \tag{10}$$

Then, our strategy commits all capital to long or short the constructed portfolio based on the signal, and holds the position for a week. The $t+1$'th week's return is given by:

$$r_{t+1} = \hat{\Delta}_{t+1} s_{t+1} \tag{11}$$

Given the weekly log return, the net simple return after subtracting transaction cost is give by:

$$R_{t+1} = e^{r_{t+1}} - 1 - tc \sum_{i=1}^{p} \left| \beta_{i,t+1} - \beta_{i,t} \right| \tag{12}$$

where tc denotes the transaction cost.

Finally, the periodic simple return from time 0 to time T is:

$$R_0^T = \prod_{t=0}^{T} (1 + R_t) - 1 \tag{13}$$

B. Performance of Replicating Portfolios

Below Figure 7 displays log change of replicating portfolio and actual log change f spread on test data set. The model replicates the actual change quite well with $R^2 = 62.18\%$.

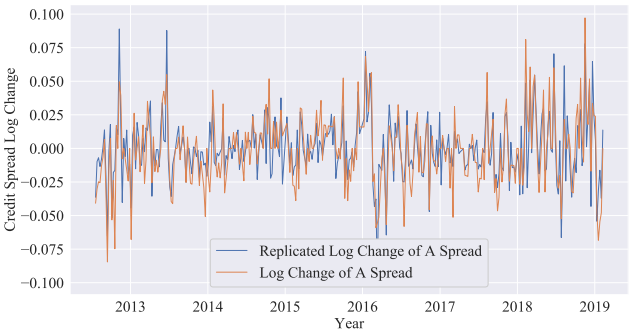

Fig. 7. Replicated Log Change versus. Actual Log Change

The same replicating procedure is repeated for AA, AAA and BBB spread and we obtain R^2 of 52.85%, 47.45% and 61.78% respectively. The method of replicating portfolio construction is robust across different ratings.

C. Performance of Trading Strategies

The strategies' annualized absolute returns, Sharpe ratio and performance for different transaction cost (tc) assumptions are shown in Figure 8, 9, 10, 11. The performance is consistently strong for different ratings, and particularly resilient in AAA class.

TABLE II

STRATEGY PERFORMANCE (TRANSACTION COST 0.1%)

	Sharpe Ratio	Annualized Return
AAA spread	1.40	40.96%
AA spread	1.17	27.87%
A spread	1.49	32.28%
BBB spread	1.42	28.68%

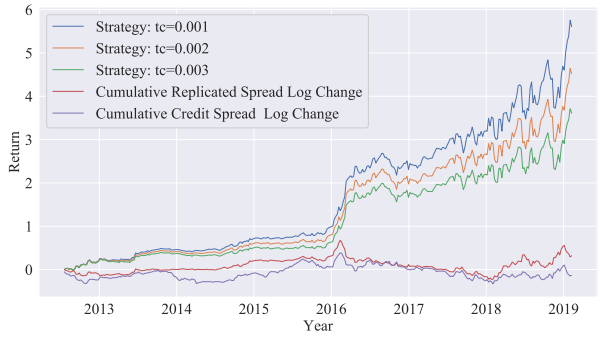

Fig. 8. Strategy Performance for AAA Spread

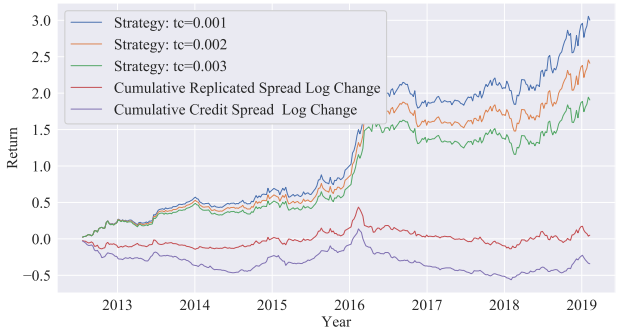

Fig. 11. Strategy Performance for BBB Spread

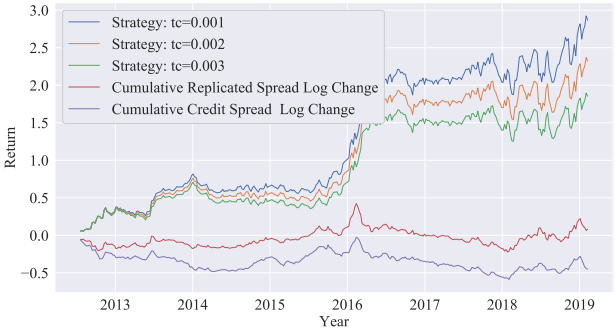

Fig. 9. Strategy Performance for AA Spread

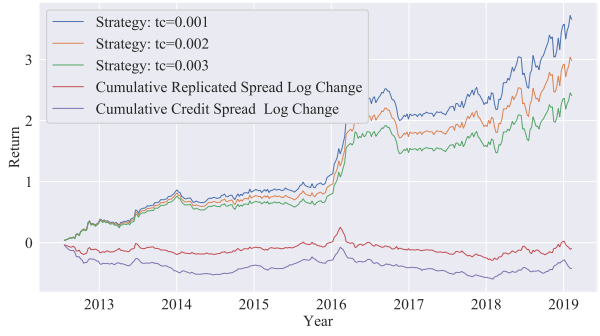

Fig. 10. Strategy Performance for A Spread

VII. CONCLUSIONS

This paper applies several machine learning models to predict weekly change of investment grade credit spread. We develop a subsequent trading strategy by replicating non-tradable credit spread with ETFs. To increase the prediction accuracy, averaging ensemble is applied to combine the unique strength of each model. To prevent overfitting, we apply Nested Cross Validation to tackle challenges of time series ad, and add regularization term in all models. To unpack the black-boxes, we demonstrate the relationship between predictors and predicted variable by MSE-reduction-rank heat map and partial dependence plot. Based on the statistical models, there are 5 explanatory variables are exceptionally effective: RUSSELL 2000 return, NFCI-Credit, NFCI-Leverage, IG-HYM (yield difference between the investment and specualtive grade) and finally RET_PCA_1 (the first component of our PCA analysis).

All machine learning based strategies are subject to regime shifts risk. We thus design a utility function to detect such shifts, and from the monotonic cumulative utility plot one can confirm the consistently strong performance of our model within available sample. We thus believe no significant regime shifts exist in our test data. Cautions are to be taken, should the model be applied to out-sample tasks.

We construct portfolios of bond ETFs, with weekly weights calculated with rolling regressions to replicate credit spread changes. Such portfolios are used as a proxy for credit spread to make use of the signal generated by machine learning models. This strategy proves to be effective as the back-testing annualized portfolio returns ranging from 40.96% (AAA spread) to 28.68% (BBB spread) with Sharpe Ratio from 1.49 (A spread) to 1.17 (AA-spread), indicating effectiveness and practicability of our machine learning prediction results.

APPENDIX

TABLE III
DATA DESCRIPTION

Category	Variable Name	Label	Source	Frequency
Credit Spread Data	BofAML US Corporate Investment Grade Option-Adjusted Spread Index, log change	RET	FRED	Weekly
	Credit Spread Slope (BofAML IG OAS 10-15 Yr - 1-3Yr)	Spread_Slope	FRED	Weekly
	Speculative-grade Option Adjusted Spread (BofAML High Yield OAS)	HYM	FRED	Weekly
	Spread between Investment-grade and Speculative	IG_HYM	FRED	Weekly
Macroeconomic Indicators	Non-farm Payroll	NFP	Bureau of Labor Statistics	Monthly
	Consumer Price Index	CPI	FRED	Monthly
	Conference Board Composite Index of Coincident Indicators	COI	Conference Board	Monthly
Financial Situation Indicators	Chicago Fed National Financial Conditions Index	NFCI-Credit NFCI-Liability NFCI-Risk	Fed Chicago	Weekly
	St. Louis Fed Financial Stress Index	STLFSI	Fed St.Louis	Weekly
	Commercial and Industrial Loans	TOTLAON	FRED	Weekly
Financial Markets Variables	US Treasury Yield	Treasury	FRED	Daily
	TED Spread	TED	FRED	Daily
	Merrill Lynch Treasury Option 3-Month Volatility Estimate index	MOVE3M	Bloomberg	Daily
	RUSSELL 2000 Index	RUSSELL	Bloomberg	Daily
	CBOE VIX Index	VIX	CBOE	Daily
	SP500 Price to Earnings Ratio	SPXPE	Bloomberg	Dailt
	WTI Crude Oil	WTI	Bloomberg	Daily
Market Sentiment	AAII Investor Sentiment Survey	Bullish, Neutral, Bearish	Quandl	Weekly
Fama-French Factor Returns	Small Minus Big	SMB	French Data Library	Daily
	High Minus Low	HML		
	Robust Minus Big	RMB		
	Conservative Minus Aggressive	CMA		
Technical Analysis of Credit Spread Data	Trailing 5-week Log Change	Ret5	Daily	
	Trailing 5-week Realized Volatility	RV	FRED	Daily
	Trailing 5-week Skewness of Spread Change	Skew	FRED	Daily

REFERENCES

[1] Huang, Jing-Zhi Jay and Kong, Weipeng, March 2003, Explaining Credit Spread Changes: Some New Evidence from Option-Adjusted Spreads of Bond Indices. Stern School of Business Working Paper No. FIN-03-013.

[2] Bergmeir, C. and Bentez, J.M., 2012. On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, pp.192-213.

[3] Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2), pp.301-320.

[4] Hastie, T. and Tibshirani, R., 1990. Generalized Additive Models. Chapman and Hall, London.

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

[6] Friedman, J.H., 2002. Stochastic gradient boosting. Computational statistics & data analysis, 38(4), pp.367-378.