A Novel Methodology in Credit Spread Prediction Based on Ensemble Learning and Feature Selection

Abstract

The credit spread is one of the most critical indicators in bond investment, the forecast of which can provide significant help for fix income investors to develop trading strategies. We propose a novel credit spread forecasting model based on the ensemble learning method. The feature selection method with mutual information is employed to enhance the prediction. Empirical results show that our proposed methodology can provide a more accurate prediction for the credit spread. Furthermore, we provide the forecast trend for future credit spread with current data.

Keywords: Credit Spread Forecast, Ensemble Learning, Feature Selection, Mutual Information

1 Introduction

Credit spread has always been an important concern for investors, and investment grade corporate bonds have received even more focus. Therefore, prediction on the change direction and magnitude of credit spread has been a research topic studied for years. Credit spread represents the difference in yield between a risky security and a risk-free security (usually U.S. treasury bond). A higher credit spread often indicates a lower quality bond, which has more chance of the issuer defaulting. As a result, it varies from one security to another based on the credit ratings. Moreover, there are many other factors related to the change of credit spread. It would be practical to establish a model predicting the change of credit spread in terms of investment grade corporate bonds.

There has already been abundant studies on macro-factors being the determinants of credit spread change. These studies are divided into two directions, either identifying a positive relation or a negative relation. Krueger and Kenneth (2003) applied a model of rational Bayesian and regression analysis to study the positive relation between unexpected rise in employment and benchmark treasury rate [8]. Wu and Zhang (2008) propose an internally consistent approach to quantifying the linkages between market prices of systematic macroeconomic risks and the term structure of credit spread [11]. The hypothesis test of Davies (2008) provided evidence on the effects of high inflation on the poor performance of corporate bonds [5]. For negative relations, Gertler (1991) concluded the connection between credit spread and GNP growth with a simple reduced-form test [6]. Similarly, Tang and Yan (2010) found that credit spread narrows with an increase in GDP growth rate and Consumer Confidence Index (CCI) through their emprical analysis [9]. A Study by Collin (2001) included a bench of macro-level factors in the regression model to analyze this issue, in terms of both directions [3].

Despite many studies, there still exist plenty doubts about the determinate in credit spread. The most obvious one would be current factors and regression models can only explain a quarter of the change [3]. Another deficiency is the influence of these determinants being mostly restricted to qualitative analysis. To overcome these deficiencies, the ensemble learning method is introduced. Since recently, more ensemble learning methods are employed for financial forecast and received

better performance than traditional machine learning algorithms [1]. It would be helpful to apply the ensemble learning method to forecast credit spread change, which has not been studied so far. Therefore, we propose a methodology of predicting the change in credit spread combining the ensemble learning methods with feature selection. Empirical results show that our methodology can achieve a high accuracy in credit spread forecast. The rest of the paper is organized as followed. In Section II, we discuss the methodology, including feature description (describing the potential factors affecting the change in credit spread), feature selection (getting rid of features with low significance), and forecast model (explaining how the machine learning algorithm is combined). In Section III, we conduct an empirical study and analyze the results to provide evidence for robustness of our model. In Section IV, we make our conclusion.

2 Methodology

We establish a credit spread forecast model based on the ensemble learning method. Firstly, we get features from six aspects closely related to the credit spread, which construct our raw feature set. Considering the raw feature set contains useless information which may involve unnecessary noises into our prediction, we introduce a feature selection filter based on mutual information. Mutual information also implies the contribution of the features to the credit spread prediction. The filtered features set and the historical behavior of the credit spread will be the input for our prediction model. Next, we employ several machine learning tools and ensemble learning methods for the forecasting of future credit spread. Figure 1 presents an overview of our methodology framework.

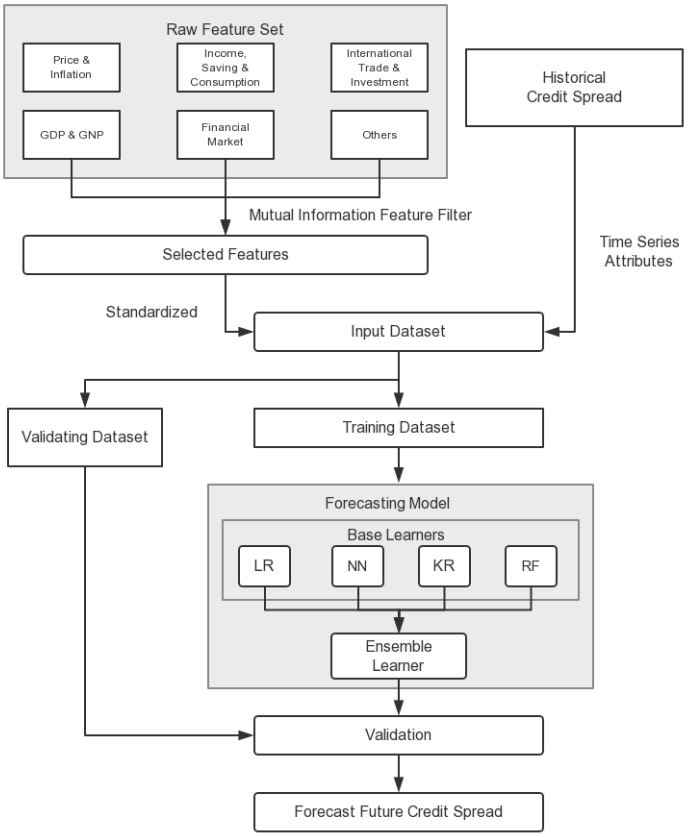


Figure 1: The Framework of Credit Spread Forecasting Model

2.1 Feature Description

We found 34 features having effects on credit spread of corporate bond. For the convenience of analysis, we divide these features into 6 categories with explanation.

First, let us do a quick review of all the features we used to train our model. Gertler (1991) used GNP as a factor to changes of credit spreads [6]. A study by Collin (2001) summarized in the literature review that a combination of several financial market data decide the change of credit spread, which are change in yield on 10-year Treasury, change in 10-year minus 2-year Treasury yields, change in implied volatility of S&P 500 and return on S&P 500 [3]. Christiansen (2002) analyzed the relationship between macroeconomics index such as Producer Price Index (PPI) and changes of credit spreads [2]. Krueger and Kenneth (2003) illustrated the relationship between benchmark treasury interest rate which is a component of credit spreads and unemployment rate [8]. Wu and Zhang (2008) showed the relationship between GDP growth rate and changes of credit changes [11]. Davies (2008) analyzed the relationship between inflation risk presented by CPI index and credit spread [5]. Cúrdia (2010) studied the relationship between credit spreads and monetary policy and used index like M1 and M2 money stock, government purchase and government revenue [4]. The findings of Gilchrist (2010) verified the relationship between some significant fluctuation of macroeconomy presented by industrial production, personal disposable expenditure, personal income and credit spread [7]. Tang and Yan (2010) found the relationship between GDP growth rate, Consumer Confidence Index (CCI) and credit spread [9]. Tsai (2010) used some international trade index to analyze the tendency of credit spread [10]. Table 1 lists financial and economic index in some of the aforementioned studies.

Author (Year)	Features
Gertler (1991)	GNP
Collin-Dufresne et al. (2001)	Change in yield on 10-year Treasury; Change in 10-year minus 2-year Treasury yields; Change in implied volatility of S&P 500; Return on S&P 500
Christiansen (2002)	Produce Price Index (PPI)
Krueger & Kenneth (2003)	Unemployment rate
Davies (2008)	Inflation risk (CPI)
Wu & Zhang (2008)	GDP; Real GDP
Cúrdia (2009)	M1 and M2 money stock; Government purchase; Government revenue
Gilchrist (2010)	Industrial Production; Personal disposable expenditure; Personal income
Tang & Yan (2010)	Consumer Confidence Index (CCI)
Tsai (2010)	Net export volume; Export price index; Import price index; Total trading volume

Table 1: Features Associated with Credit Spread

We have collected data for 34 features. Most data of the 34 features are obtained from the St. Louis Fed data repository, and the rest are from the IMF Data and some other databases.

Second, based on the classification provided by U.S. Bureau of Economic Analysis, we group the 34 features into 6 categories, which are price and inflation, income, saving and consumption, international trade and investment, GDP and GNP, financial market, and other factors. Details are shown in Table 2. We also take the difference of each feature as a separate variable in the

selectino process.

A brief introduction of the six features categories is given as follow:

1. The category with price and inflation is used to quantify the economy's general price level or a cost of living and to measure the rate of inflation in an economy.
2. The category with income, saving and consumption is the personal income and consumption to reveal the economic situation in the U.S..
3. The category with international trade and investment is to record changes in the price which firms and countries receive for products to reflect the economic situation in global.
4. The category with GDP and GNP is self-explanatory.
5. The category with financial market includes some indices treated as common symbols of financial market performance and activity.
6. The category with other factors includes some other features that we and some formers believe could have practical meaning and the capability to represent the economic trend in different facets, but not worthy having a separate category.

Category	Features Name
Price and Inflation	CPI, Producer Price Index (PPI), Total Wholesale Trade Industries (WPI) Consumer Confidence Index (CCI), Producer Price Index (PPI) GDP Deflator (GDP_D)
Income, Saving and Consumption	Personal Consumption Expenditures (PCE), Personal Saving Rate (PSR) Disposable Personal Income (DPI)
International Trade and Investment	Export Price Index (EPI), Export volume index (EVI), Export rate (ER) Import Price Index (IPI), Import volume index (IVI), Import rate (IR) Price Index of Machinery product export Quantity Index of Machinery product export Electric product export order, Electric machinery product export order Information and communication product export order Net Exports of Goods and Services (NEGS), Total trading volume Index (TTVIND)
GDP and GNP	GDP, Real GDP (R_GDP), GNP, Percentage change in GDP (ΔGDP) Percentage change in Real GDP (ΔRGDP), Percentage change in GNP (ΔGNP), Industrial production Index (INDPRO)
Financial Market	10-Year Treasury Constant Maturity Minus 2-Year Treasury Constant Maturity (T10Y2YM), 10-Year Treasury Constant Maturity Rate (GS10) VIX index (VIX), S&P 500 return (S&P500), Discount Rate (DR) M1, M2, USD Index (USDIND)
Other Factors	Leading Index (LIND), Lagging Index (LIND2), Unemployment rate (UR)

Table 2: Category of Features

2.2 Feature Selection

In our feature selection method, the criterion is the amount of information a feature can bring to the forecast system. The more information it can bring, the more critical the feature will be. For a given feature, due to whether or not the forecast system including the feature, the amount of information will change. The difference of information amount between the system with and without the feature is the information gain brought by the feature. This information gain can help us to make our forecasting decision. We use differential entropy to measure the amount of information for the forecasting problem.

Given a random variable X with probability density function $f(x)$, the differential entropy $h(X)$ is defined as

$$h(X) = -\int f(x) \log f(x) dx. \tag{1}$$

For two random variables X and Y, suppose they have a joint pdf $f(x, y)$, their conditional differential entropy $h(X|Y)$ is defined as

$$h(X|Y) = -\int f(x, y) \log f(x|y) dx dy \tag{2}$$

and the mutual information between X and Y is given as

$$I(X; Y) = h(X) - h(X|Y). \tag{3}$$

Notice that $I(X; Y) \geq 0$ and $h(X|Y) \leq h(X)$ where the equalities hold if and only if X and Y are independent. If we regard the future credit spread as the random variable X, then the entropy measures the randomness of X, the inclusion of any feature Y will reduce the randomness of X and consequently give us a better prediction of the credit spread. The higher the mutual information between X and Y, the more information to the prediction can be provided by the feature Y. Thus, we now want to calculate the mutual information between the features and the credit spread using the historical data, and select features with the highest mutual information value as the input for the prediction model.

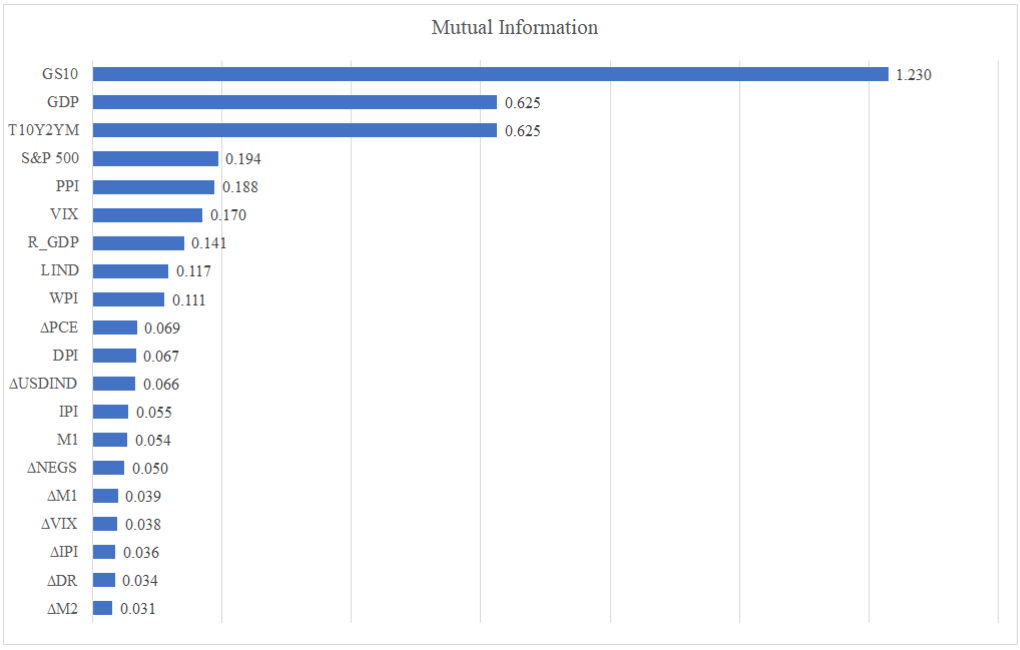


Figure 2: Mutual Information of Selected Features

Figure 2 shows features with the highest mutual information. We select 20 features from the raw feature set to construct our prediction model so that we can avoid the disturb from the features of insignificance. Since the mutual information reflects the significance of features to the credit spread, features with the higher mutual information are more helpful for us to determine the future spread. Therefore, from Figure 2, features with the highest influence to credit spread are the 10-Year Treasury Constant Maturity Rate, GDP, 10-Year Treasury Constant Maturity Minus 2-Year Treasury Constant Maturity, S&P 500 index, PPI and VIX index, which correspond to our analysis in the previous section.

2.3 Prediction Model

In this section, we will set up a two-layer prediction model. The forecasting model consists of several base learners including Multi-layer Perceptron regressor(MLP), random forest regressor, k-nearest neighbors regressor (K-NN) and an ensemble learner as the second layer.

In supervised learning algorithms of machine learning, our goal is to develop a stable model that performs well in all aspects, but the actual situation is often not so ideal, sometimes we can only get multiple weak supervised models, which perform better in certain aspects. Ensemble learning is to combine these weak models here in order to get a better and more comprehensive model. When spread price changes dramatically, ensemble learning prevents the original model from being affected by outlier values. There are several ensemble learning techniques available, and stacking will be used in our work. The algorithm in Table 3 summarizes stacking.

Algorithm: Stacking
Input: training data $D = \{x_i, y_i\}_{i=1}^{m}$
Ouput: ensemble classifier H
Step 1: learn base-level regressor
for $t = 1$ to T **do**
learn h_t based on D
end for
Step 2: construct new data set of predictions
for $t = 1$ to T **do**
$D_h = \{x_i^{'}, y_i\}$ where $x_i^{'} = \{h_1(x_i), ..., h_T(x_i)\}$
end for
Step 3: learn a meta-regressor
learn H based on D_h
return H

Table 3: Stacking Algorithm

In the first layer of our model, we should get predictions from three base learners. These predictions will be passed as features to the ensemble learner and will be trained in the second layer. Specifically, our predicting process is as follows:

The spread price itself reveals much information above the future movement of the spread price, in addition to those features we have mentioned in Section II, we include the average spread over the past few months. While this limits the length of time we can predict, it will correct the future predicted value by updating the spread price. The optimal duration for the average spread price is calculated by minimizing the mean square error. Next, we will conduct PCA whitening with the recent average spread price and the other financial indicators. The goal here is to reduce the correlation between all features, as financial data are usually highly correlated. Then, the

whitened data will be passed to Multi-layer Perceptron regressor (MLP), Random Forest regressor, K-NN regression respectively, and the predictions from these regressor will be used to train the Kernel Ridge regressor which is the second layer of our model.

These three base learners are chosen because of their excellent predictive effect on the historical spread price. MLP is a class of feedforward artificial neural network, and it allows us to solve problems stochastically, which makes it a good regression method for the spread price. Random Forest regressor is chosen for its popularity among research scientists and its high accuracy. K-NN works perfectly around local values. The limitation of each method drives us to set up a second layer (Kernel Ridge) to balance the predictions as well as reduce the noise.

3 Empirical Analysis

The empirical data in our work comes from websites of the Bureau of Economic Analysis and the Federal Reserve. We collect 120 pieces of monthly historical data lasting from Jan 2008 to Dec 2017 to construct our experimental data set. Each piece of data contains the features mentioned in Table 2 as the input parameters and the monthly credit spread as the target of the learning model. We divide our data set into two parts where the first 70 % data construct the training set for the modeling process and the latter 30 % construct the testing set for validation. For each prediction model, three indicators, the Mean Average Error (MAE), Mean Square Error (MSE) and R^2 score, are employed as the criteria for the performance evaluation of models. Given n pairs of actual value y_i and the predicted value $\hat{y}_i$, the indicator is described as the following:

$$\begin{aligned} MAE &= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \\ MSE &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ R^2 &= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \end{aligned}$$

3.1 Experimental Result

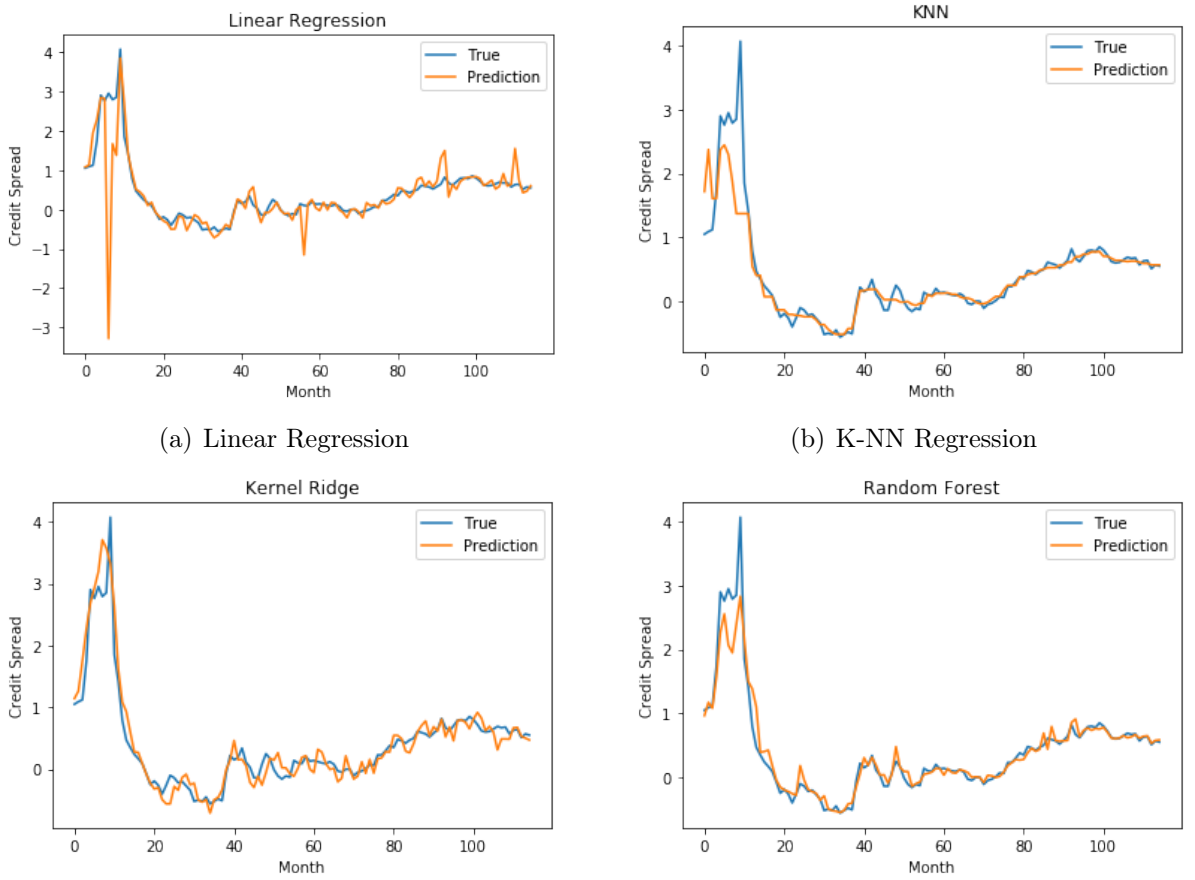
We compare the prediction results of four base learners and the ensemble learner. For each learner, we consider cases of using the raw features and using the selected feature set. The results of the training and testing performances are displayed in Table 4.

As we can see from Table 4, the best prediction on training set and testing both appear in the stacking model. Notice that in most algorithms, although feature selection doesn't improve the results on the training set, the results on the testing set are better, this is because the over-fitting problem on the original data set is avoided by feature selection, and the anti-noise ability of the model has also been improved. Moreover, besides from Stacking algorithms, all methods have their own limitations. For linear regression, although it fits the training set well, the predictions are even worse than using than mean prices on the testing data set; for K-NN regression, the drawback lies in its sensitivity to the local structure of the dataset; for kernel ridge, the results are not bad, but it still get a room for improvement; The random forest also has decent predictions, but it subjects to randomness and is not robust on the data set. Also, we plot a simulated price prediction and provide a short analysis on each base learner.

The linear regression could be seen as a baseline in this work, but it does not produce a convincing result from the plot, the predictions fluctuate around the true prices. K-NN regression smoothes

Learning Method	Feature Selection	Training Set Result			Testing Set Result		
		MAE	MSE	R^2 Score	MAE	MSE	R^2 Score
Linear Regression	No	0.082	0.010	0.981	0.528	1.378	-0.787
	Yes	0.086	0.011	0.978	0.463	1.163	-0.508
K-NN regression	No	0.119	0.127	0.752	0.180	0.120	0.844
	Yes	0.119	0.127	0.752	0.180	0.120	0.844
Kernel Ridge	No	0.119	0.027	0.948	0.235	0.111	0.856
	Yes	0.121	0.028	0.945	0.231	0.107	0.861
Random Forest	No	0.104	0.064	0.874	0.238	0.165	0.786
	Yes	0.098	0.064	0.875	0.204	0.106	0.862
Stacking	No	0.003	1e-4	0.999	0.168	0.071	0.908
	Yes	0.004	1e-4	0.999	0.155	0.062	0.920

Table 4: The Performance of Credit Spread Prediction Models



(a) Linear Regression

(b) K-NN Regression

(c) Kernel Ridge Regression

(d) Random Forest Regression

Figure 3: Prediction of Base Learners

the historical spread and provides a better estimate in the period of relatively stable price. For the estimation of the spread price with a high volatility during the financial crisis period where, the estimation is not satisfying because of the local structure. Kernel ridge is an improved version of linear regression method here, and we want to add a penalty term to the features to avoid over-fitting problem of regression. Therefore, we have a better output than the linear regression. The random forest is actually an ensemble learning method, and it indeed provides a decent result, which gives us confidence to believe that stacking algorithm should work in this context.

The stacking method combines the output combines the output from other predictors, hence

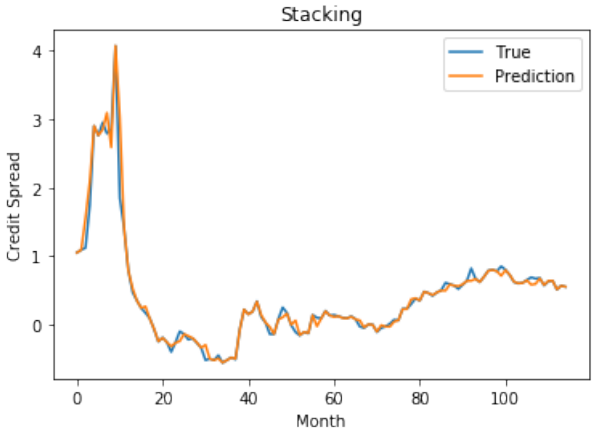


Figure 4: Stacking

it provides us with an algorithm robust enough to handle most situations, also it has the best outcome. As we can see from the plot, not only it does well on the training set, but it also works very well on the testing set.

3.2 Prediction

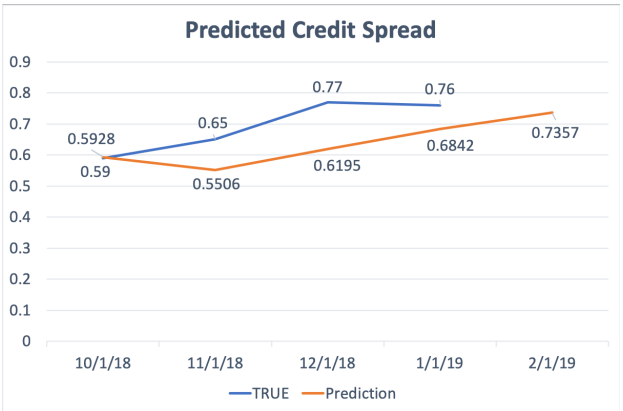


Figure 5: Prediction of Credit Spread

From the previous discussion, we know that the ensemble learning model has the best perform with feature selection techniques included. Therefore, given current data, we use the stacking method with selected feature set to predict the future behavior of the credit spread. With the data from 2018, we have our prediction for the credit spread in the first quarter of 2019 are given in Figure 5. The actual credit spreads for the previous four months validate our prediction with errors less than 15 bps. We also provide a forecast for the credit spread of February 2019 to be 73 bps. In general, the predicted value is less than the true value, this is because we have some null values in some features, which affects the accuracy. However, the predictions on the direction and magnitude of credit spread movement is quite accurate.

4 Conclusion

This paper illustrates a novel prediction method for the monthly credit spread based on the ensemble learning method. We innovatively include the feature selection method using mutual information into forecasting credit spread. The empirical result shows that the ensemble learning method performs better than the traditional machine learning method. We predict the credit

spread February 2019 to be 73 bps. Moreover, feature selection before prediction not only explains the rationality of features but also improves the accuracy and robustness of the prediction results.

However, our work has some limitations. First, although mutual information shows the significance of selected features, it is hard to find the cross relationship within different features, which requires further feature engineering method. Second, we need the cross-section data for the prediction, which means we have to get access to the most recent monthly data to forecast the credit spread of the next month. This fact of our prediction model limits the length of time we can predict in the future. Future works for our work can include more efficiently time series analysis techniques to our model to get better performance and more extended forecasting period for forecasting.

References

[1] Michel Ballings, Dirk Van den Poel, Nathalie Hespeels, and Ruben Gryp. Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications*, 42(20):7046–7056, 2015.

[2] Charlotte Christiansen. Credit spreads and the term structure of interest rates. *International Review of Financial Analysis*, 11(3):279–295, 2002.

[3] Pierre Collin-Dufresn, Robert S Goldstein, and J Spencer Martin. The determinants of credit spread changes. *The Journal of Finance*, 56(6):2177–2207, 2001.

[4] Vasco Cúrdia and Michael Woodford. Credit spreads and monetary policy. *Journal of Money, credit and Banking*, 42:3–35, 2010.

[5] Victor AB Davies. Postwar capital flight and inflation. *Journal of Peace Research*, 45(4):519–537, 2008.

[6] Mark Gertler, R Glenn Hubbard, and Anil Kashyap. Interest rate spreads, credit constraints, and investment fluctuations: an empirical investigation. Technical report, National Bureau of Economic Research, 1990.

[7] Simon Gilchrist and Egon Zakrajšek. Credit spreads and business cycle fluctuations. *American Economic Review*, 102(4):1692–1720, 2012.

[8] Alan B Krueger and Kenneth N Fortson. Do markets respond more to more reliable labor market data? a test of market rationality. *Journal of the European Economic Association*, 1(4):931–957, 2003.

[9] Dragon Yongjun Tang and Hong Yan. Market conditions, default risk and credit spreads. *Journal of Banking & Finance*, 34(4):743–753, 2010.

[10] Chih-Fong Tsai and Yu-Chieh Hsiao. Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems*, 50(1):258–269, 2010.

[11] Liuren Wu and Frank Xiaoling Zhang. A no-arbitrage analysis of macroeconomic determinants of the credit spread term structure. *Management Science*, 54(6):1160–1175, 2008.