Forecasting Credit Spreads: A Machine Learning Approach

Rick (Yuankang) Xiong, Hui Cai, Israel Diego-Guerra, Yifei Lu, Xinye Xu, Yuan Yin

Abstract: In this paper, we developed an LSTM model with 88.9% out-of-sample accuracy that predicts the direction of next-day change in the credit spread. We constructed and compared the outputs of several different models with verified features provided by previous literature. The models include: Vasicek model, tree models such as random forest and BART (Bayesian Additive Regression Tree), and LSTM (Long Short-Term Memory) model. The model performance and predictive power were improved sequentially. BART model works well in predicting the magnitude of credit spread, but fails to predict the direction accurately. However, LSTM model yields the best prediction of next-day credit spread change with 3-day look-back period and feature variables. We captured two main characteristics of credit spread prediction. First, credit spreads are well explained by their own history. Second, the prediction model should be able to evolve and gradually forget older data, while also learning patterns from more recent history in order to capture regime changes. Specific machine learning details and economic interpretations of features are also discussed.

Keywords: credit spread, prediction, LSTM, BART, random forest

Contents

1. Introduction

1.1 Economic Significance

Credit spreads represent the difference in the yield of a risky security and a risk-free security of similar maturity. They act as good indicators for the credit rating of a firm and the condition of financial markets. Figure 1 demonstrates series for US Corporate AAA bond yield and 10-Year Treasury Constant Maturity Rate and Figure 2 shows the spread between them. In Figure 2, we see the credit spread decreased from 2001-2002, which implies that the market transitioned from recession to economic expansion. This transition is evidenced by the Dot-com bubble in 2000, where the U.S. government lowered the interest rate in order to stimulate the economy. The second grey area in Figure 2 corresponding to the 2008 great recession, shows that the credit spread spiked upward, indicating an economic downturn. Thus, predicting the direction and magnitude for credit spreads effectively plays an important role in financial markets.

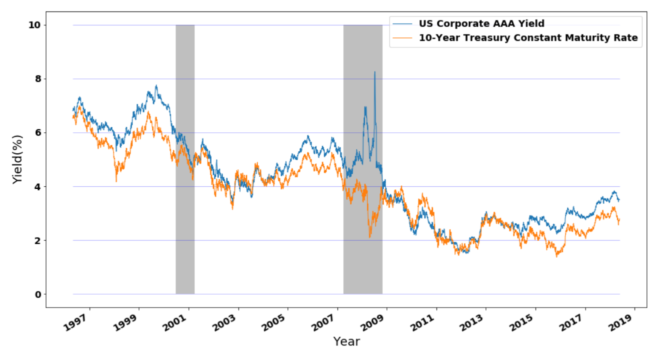

Fig. 1. Yield series of risky security(corporate AAA bond) versus that of a risk-free security(treasury bond) of similar maturity(10-year). Data source: St. Louis Fed data repository. Shaded area represents recession.

Fig. 2. Credit Spread. The difference between corporate AAA bond and treasury bond yield, derived directly from Figure 1.

Credit spreads can also serve as a useful tool for fixed income investors and corporate finance professionals. Fixed income investors can profit from abnormal trading by predicting credit spreads effectively. For example, if investors forecast that the credit spread will widen in the future, then they could sell the corporate bonds and buy government bonds.

Alternatively, if they forecast that the spread will tighten in the future, then they can sell government bonds and buy corporate bonds. In the end, investors will sell bonds with a higher price but buy with a lower price and the difference is the profit that investors earns.
Corporate finance professionals can reduce the cost of capital. For example, if a firm is planning to issue long-term corporate bonds and if the firm predicts that the credit spread will tighten in the future, to avoid the high cost of interest rate, the firm could fund with short-term debt and wait until credit spreads have actually narrowed.

1.2 Literature Review

Credit spreads are widely used as a symbol for systematic risk, liquidity, and conditions in financial economics. We find in previous papers that the direction-of-change and magnitude of the credit spread are explained well by bond characteristics, credit quality, and market conditions. The conclusions of the structural approach regarding the shape of the term structure of credit spreads are confirmed by the empirical work of Huang[1]. Option-implied volatility is also a measure for future cash flow uncertainty, Chiras and Manaster (1978)[2]. Pierre Collin-Dufresne, Robert S. Goldstein, and J. Spencer Martin (2001) summarized that changes in leverage has significant explanatory power for the movement of credit spreads[3]. The market jump risk factor constructed from high-frequency data also seems to capture low-frequency movements in credit spreads in terms of long-run trends, as shown by George Tauchena, Hao Zhou (2011)[4]. Tauchena and Zhou used quarterly S&P 500 index to capture changes in the overall economy.
Hai Lin, Sheen Liu, and Chunchi Wu developed a new methodology to better dissect the components of credit spreads and described their findings. They discover that 47% of yield spreads are explained by default risk while the remaining 53% is made up of nondefault risk[5].
Different patterns exist on different time-regimes. Over the short-run, credit spreads are negatively related to treasury rates. Initially, spreads narrow because a given rise in treasury rates produces a proportionately smaller rise in corporate rates. However, over the long-run, this relation is reversed. A rise in treasury rates eventually produces a proportionately larger rise in corporate rates. This widens the credit spread and induces a positive relation between spreads and treasury rates[6].

1.3 Model Evolution

Although financial markets are generally unpredictable, there are still many successful methods proven to capture dynamics of financial markets. Mathematicians use stochastic models to explain the dynamics of financial markets. Geometric Brownian Motion (GBM) is one of the earliest and simplest models used to describe stock prices and is still used today by financial firms. CIR and Vasicek models describe dynamics of interest rates and the Heston model describes volatility of an underlying asset. Since credit spreads are based on the interest rates of government and corporate bonds, it is natural to use mathematical equations like Vasicek model to fit the dynamics of credit spread.
A well-defined stochastic differential equation is a good starting point, but it typically cannot capture the whole dynamics. One reason is the limitation of parameters. We turn to machine learning, another popular technique used in financial markets. Financial markets are some of the earliest adopters of machine learning methods, which are still one of the mainstream tools in market analysis. Given the success of machine learning methods on many market related processes, we seek suitable machine learning methods for predicting credit spreads. We will focus on some models that are proven to perform well in financial markets.
Decision tree is one of the popular machine learning methods in analyzing financial markets. A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. We will test our predictions first with the random forest model in order to get an easy interpretable result. In order to improve the tree model performance, we also introduce Bayesian additive regression tree (BART) model. Supporting evidence from Hugh A. Chipman, Edward I. George and Robert E. McCulloch (2008)[7] shows that BART is robust to small changes in the prior distribution and the choice for number of trees.
Finally, Long Short-Term Memory (LSTM) is also used as another candidate model for credit spread prediction. Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal (2015) show that by modeling the normal behavior of a time-series via stacked LSTM networks, we can accurately detect deviations from normal, also suggesting that LSTM-based prediction models may be more robust compared to RNN-based models[8].
The above methodologies lay the foundation for our analysis on credit spread prediction. In addition, some potential shortcomings of each approach are also covered.

2. Data

In order to have a good understanding of historical movements of credit spread, the first task lies in determining which historical period we should choose. To capture previous economic expansions as well as recessions, we set our time horizon from December 1996 to January 2019.

We use the following variables to conduct our empirical experiment. Most of the variables have a major capability to predict the movement of the credit spread based on previous literature.

- **CS:** Credit Spread of US investment grade corporate bonds (e.g. BofAML US Corporate AAA) over US Treasuries (e.g. 10-Year Treasury Constant Maturity Rate). Data is obtained from the St. Louis Fed data repository and daily observations are used for the period December 1996 through January 2019.

- **SLOPE:** Slope of terms, which is the yield on 10-year Treasury bonds minus yield on 2-year Treasury bonds, which is calculated by taking the difference between 10-Year Treasury Constant Maturity Rate and 2-Year Treasury Constant Maturity Rate. Daily observations are obtained from the St. Louis Fed data repository.
- **VIX:** Cboe volatility index on the S&P 500 index. Daily data is retrieved from Bloomberg. VIX is a barometer of equity market volatility. It is based on real-time prices of options on the SP 500 Index (SPX) and is designed to reflect investors' consensus view of future (30-day) expected stock market volatility. The VIX Index is often referred to as the market's "fear gauge". VIX can be correlated to the credit spread, which basically captures the future probability of default as a common forward-looking risk metric.
- **SP500_R:** Return on the S&P 500 stock index. S&P500 return is calculated with data from Yahoo finance by dividing the difference of current day's S&P 500 stock index and previous day's S&P 500 index by previous day's S&P 500 index. The equity market could be relevant to corporate bond credit spreads in several ways: as an alternate investment, as a measure of capital markets investment levels, or as a yield-producing security. The equity market as an alternative to fixed-income investment is the most plausible choice in today's environment.
- **SWAP_SPREAD:** A swap spread is the difference between the fixed-rate component of a given swap and the yield on a treasury item or other fixed-income investment with a similar maturity. Here a corresponding 10-year maturity is chosen and daily data is retrieved from Bloomberg. Swap spread is highly correlated with the credit spread because the swap rate is a proxy for credit rate. As the swap market is significantly deeper and more liquid than that for corporate bonds, swap rates may provide a forward indication of the direction of credit spreads. As swap spreads increase, so should credit spreads[9].
- **SKEW:** The Cboe SKEW Index ("SKEW") is an index derived from the price of SP 500 tail risk. The price of SP 500 tail risk is calculated from the prices of SP 500 out-of-the-money options. The SKEW Index can reflect the tail risk in the equity market.

The following data are generated in order to achieve two purposes. First, as shown in Figure 3, the collinearity between the variables above will result in poor performance of parameter estimation, for example, the correlation between VIX and spread is more than 0.5; therefore, we take first-differences of the original data to make the time-series more stationary and less collinear. Second, historical credit spread may act as a good feature in prediction. Here, we provide the partial autocorrelation function (PACF) plot. PACF gives the partial correlation of a time series with its own lagged values, controlling for the values of the time-series at all shorter lag orders. As shown in Figure 4, the partial autocorrelation of lag-1, lag-2 and lag-3 are significant. Hence lag-i values of spread are also listed.

- **DGS10_DIFF1:** The difference between current day's and yesterday's 10-Year Treasury Constant Maturity Rate.
- **VIX_DIFF1:** The difference between current day's and yesterday's Cboe VIX Index.
- **SPREAD_lag_i:** Credit spread of previous days. (SPREAD_lag_1 denotes yesterday's spread)

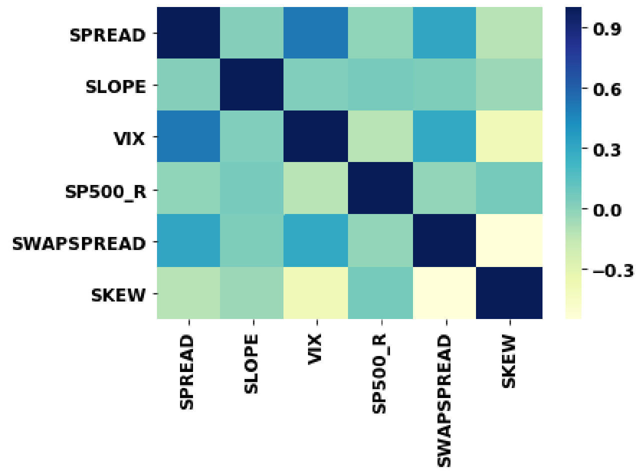

Fig. 3. Variable Correlation

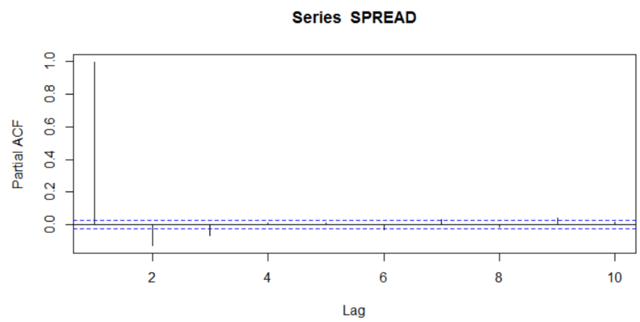

Fig. 4. PACF of credit spread, the solid vertical line shows the significance of partial autocorrelation at certain lag

3. Methodology and Model Evaluation

3.1 Vasicek Model

The Vasicek model is a mathematical model describing the evolution of interest rates. It is a type of one-factor short-rate model that describes interest rate movements driven by only one source of market risk. The model can be used in the valuation of interest rate derivatives and has also been adapted for credit markets. It was introduced in 1977 by Oldřich Vašíček and can be also seen as a stochastic investment model. The model specifies that the instantaneous interest rate follows the stochastic differential equation:

$$dr_t = a(b - r_t)dt + \sigma dW_t \quad (1)$$

- b: "long term mean level". All future trajectories of r will evolve around a mean level b in the long run.
- a: "speed of reversion", which characterizes the velocity at which such trajectories will regroup around b in time.

- σ: "instantaneous volatility", which measures instant by instant the amplitude of randomness entering the system. Higher σ implies more randomness

The model calibration is conducted as follows:
First, the continuous SDE is discretized as follows:

$$CS_{t+\Delta t} - CS_t = a(b - CS_t)dt + \sigma\sqrt{\Delta t}\epsilon_t \qquad \textbf{(2)}$$

provided that ϵ_t is $i.i.d.\ N(0,1)$. Then a simple linear regression is used to yield an estimation of the parameters above. To be more specific, the model is calibrated every 100 days through a rolling window; in this way, the parameters are allowed to change over time. Figure 5 shows the movement of σ and b.

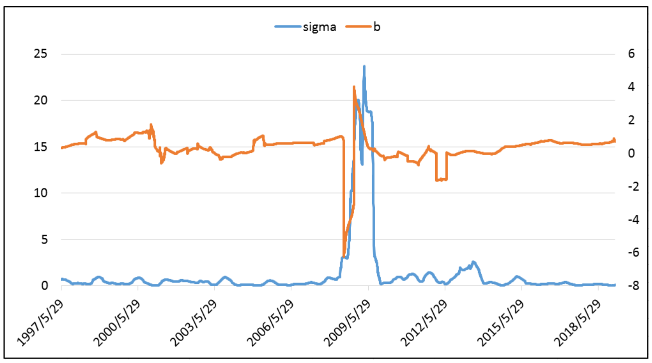

Fig. 5. Vasicek Model Parameter Estimation, the dynamics of $sigma$ and b with the changing of time

Based on Figure 5, we can see that parameter values change suddenly at certain time periods, especially around year 2008, and then go back to normal quickly. This indicates that the Vasicek model is not time-homogeneous for the credit spread, in other words, one single set of parameters does not hold for all time regimes. Using only one set of parameters could be detrimental to our predictions, because the estimated parameters would not adapt to regime changes in future times. Therefore, in order to capture this property, we need more complex models.

3.2 Random Forest

Vasicek model only takes the historical data of credit spread into account. In order to predict the future direction of credit spread, we use a classification method and add more variables mentioned in our literature review such as S&P 500 return, VIX, and so on. The decision tree is fundamental to our random forest.

Classification and regression trees or CART models, also called decision trees, are defined by recursively partitioning the input space, and defining a local model in each resulting region of input space. This can be represented by a tree, with one leaf per region.

The model can be written in the following form:

$$f(\boldsymbol{x}) = \mathbb{E}[y|\boldsymbol{x}] = \sum_{m=1}^{M} w_m \mathbb{I}(\boldsymbol{x} \in R_m) = \sum_{m=1}^{M} w_m \phi(\boldsymbol{x}; \boldsymbol{v}_m) \qquad \textbf{(3)}$$

Here, vector $\boldsymbol{x}$ represents our features, the label as y, R_m is the m-th region, w_m is the mean response in this region, and v_m encodes the choice of variables to split on.

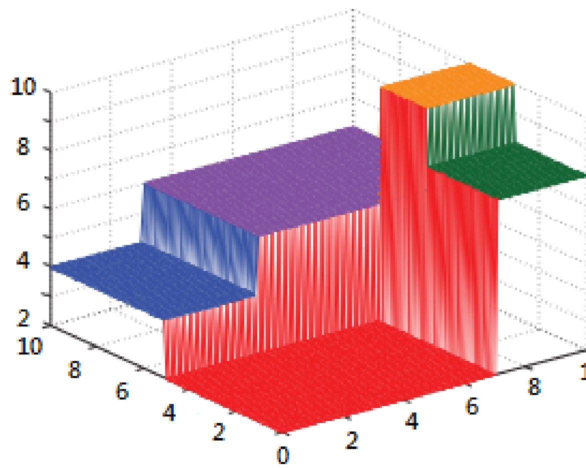

Fig. 6. Decision Tree Model Visualization

We can associate a mean response with each of these regions, resulting in the piecewise constant surface [10] shown in Figure 6. The technique known as random forests (Breiman 2001a) tries to decorrelate the weak learners by learning trees based on a randomly chosen subset of input variables, as well as a randomly chosen subset of data cases.

Random forest is a bagging method. Bagging (stands for Bootstrap Aggregating) is a way to decrease the variance of prediction by generating additional data for training from the original data set using combinations with repetitions to produce multi-sets of the same cardinality size as the original data. We cannot improve the model's predictive accuracy by increasing the size of the training set, but only decrease the variance, narrowly tuning the prediction to the expected outcome.

Bagging is a frequentist concept, but it is possible to adopt a Bayesian approach to learning trees. In particular, (Chipman et al. 1998; Denison et al. 1998; Wu et al. 2007) perform approximate inference over the space of trees (structure and parameters) using MCMC. This reduces the variance of the predictions.

In order to predict the next-day direction-of-change for the random forest model, we use the following 9 variables: SLOPE, VIX_DIFF1, SP500_R, SWAPSPREAD, SKWE, DGS10_DIFF1, SPREAD, SPREAD_lag1 and SPREAD_lag2.

Proportion of Predicted		TRUE VALUE	
		Positive	Negative
PREDICTED VALUE	Positive	56.9%	43.1%
	Negative	55.9%	44.1%

Proportion of Actual		TRUE VALUE	
		Positive	Negative
PREDICTED VALUE	Positive	58.9%	57.9%
	Negative	41.1%	42.1%

Fig. 7. Confusion Matrix for random forest. For each grid, the row label indicates the true label; column label indicates the predicted label and the numbers inside represent the accuracy. For example number 56.9% in first row first column of first table indicates that 56.9% of sample with positive label (credit spread change greater than or equal to 0) is classified as positive by the model. The second table is similar to the first, but instead of each row, now each column has a summation of 1.

We find that random forest does not predict the direction well. The confusion matrix tells us that the random forest has almost no effect on predicting the direction. The result is close to a random guess. The accuracy (total true positive divided by total predicted positive) is 56.9% and the negative predictive value (total true negative divided by total predicted negative) is only 44.1%.

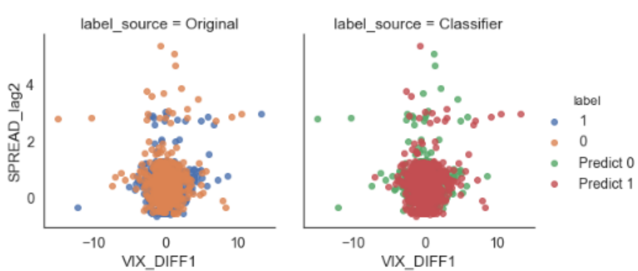

Fig. 8. Comparison Plot. The x-axis and y-axis are 2 dimensions of all features. In other words, this figure is a 2-dimension slice of the feature space. The left graph shows the original data points, which are colored into yellow and blue based on true value. The same data are colored to green and blue on the right graph based on predicted values.

As shown in Figure 8, we can see numerous points were predicted to be different from their true value. For example, the point at the bottom left corner of both graphs has a value of 1 but its prediction is 0. Therefore, random forest may not be a good model in predicting the next-day direction of credit spread.

When we label the original credit spread as 0-1 variable, we lose some important information of the continuous label. Later, we will use the boosting method instead of bagging and turn to the regression model rather than the classification model.

3.3 BART Model

Since the random forest model has poor performance on prediction, we want to improve our tree model with better ensemble methods. Thus, we introduce a boosting tree model called: Bayesian Additive Regression Trees (BART) Model. In contrast with the random forest model, BART uses a boosting method to train its model. BART is a stronger learner for each tree in the model. Additionally, boosting is a great method for dealing with underfitting.

In the BART model, each tree is constrained by a regularization of the prior to be a weak learner. Fitting and inference are accomplished via an iterative Bayesian back-fitting MCMC algorithm that generates samples from a posterior. In the following, we will briefly introduce the BART model.

First, we define inputs as $x = (x_1, \ldots, x_p)$ which is a p-dimensional vector. Then, an unknown function f used to predict output Y. We have:

$$Y = f(x) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2) \tag{4}$$

We consider f as a conditional expectation $f(x) = E(Y|x)$, and it is approximated by a sum of m regression trees $f(x) \approx h(x) = \sum_{j=1}^{m} g_j(x)$ where each g_i denotes a regression tree. Thus, we approximate the equation above by a sum-of-trees model:

$$Y = h(x) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2) \tag{5}$$

The essential idea for this model is to modify the sum-of-trees model by imposing a prior that regularizes the fit by keeping the individual tree effects small. Even though every tree explains a small and different part of f, the sum produces a strong explanation for the output.

We use T to denote a binary tree that contains a set of interior node decision rules and a set of terminal nodes. Also, $M = \{\mu_1, \mu_2, \ldots, \mu_b\}$ denotes a set of parameter values in each of the b terminal nodes of T. With the sequence of decision rules from top to bottom, each input x is assigned to the μ_i value associated with this terminal node. Thus,

$$Y = g(x; T, M) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2) \tag{6}$$

is a single tree model.
With the notation from above, the BART model can be expressed as

$$Y = \sum_{j=1}^{m} g(x; T_j, M_j) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2) \tag{7}$$

where m denotes the total number of trees.

As for specification of the parameters of the sum-of-trees model, namely $(T_1, M_1), \ldots, (T_m, M_m)$ and σ, the model imposes a prior over all the parameters. The model also has independence and symmetry properties. We omit the details for specification of the parameters to save space.

Now we use BART to make predictions on the credit spread. The input x has the 9 features we mentioned before. The credit spread data is divided into two parts: training set and test set, and the proportion is 7:3. One of BART's advantages is that it can be used for model-free variable selection. Thus, we do not need a validation set. We get predictions for both training set and test set. We compare predictions for each model in Figure 16.

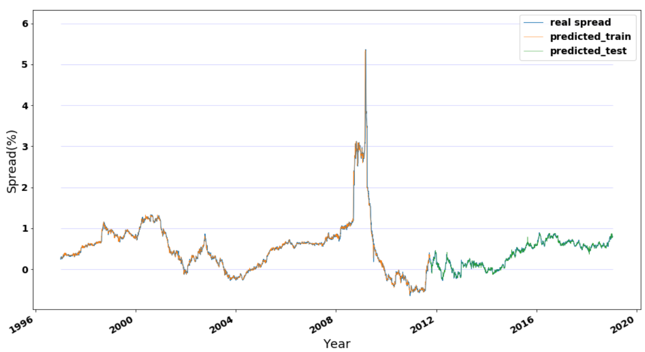

Fig. 9. BART Model Prediction: It shows all data between year 1996-2019, comparing real credit spread with fitted value by BART model

Both RMSEs for training data and test data are satisfactory, where the RMSE for training data is 0.026 and 0.029 for test data. BART proves to be effective at result predicting the magnitude of credit spread. However, if we focus on predictions of the test data, we can see that BART model will not predict well when there are frequent turning points. Instead, BART is flatter at these moments.

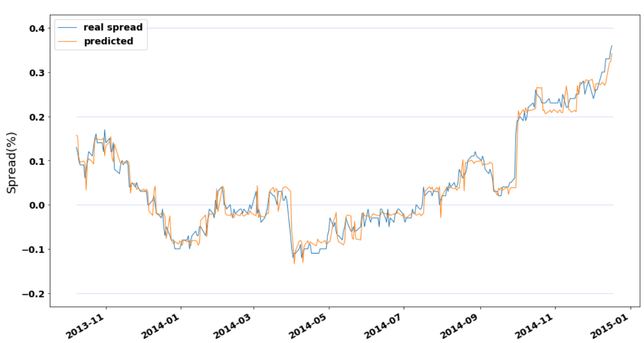

Fig. 10. BART Model Prediction (Zoom In): The plot above shows details for test data fitted value by BART model. We can see that with a closer look at predicted output, there are still some underfitting periods

To explain this characteristic intuitively, we need to check the predicted direction-of-change for credit spread with real data. We build a confusion matrix to illustrate the result clearly.

Proportion of Predicted		TRUE VALUE: Positive	TRUE VALUE: Negative
PREDICTED VALUE	Positive	41.1%	58.9%
	Negative	41.3%	58.7%

Proportion of Actual		TRUE VALUE: Positive	TRUE VALUE: Negative
PREDICTED VALUE	Positive	49.9%	50.0%
	Negative	50.1%	50.0%

Fig. 11. Confusion Matrix for BART Model. Similar to **Figure 7**, it implies the accuracy for BART model to some extent

From the confusion matrix in Figure 11, we see that BART predicts true negatives better than true positives. However, both predictions are around 50% correction. Again this prediction accuracy is akin to a coin toss. Thus, we need a better model that can predict both magnitude and direction-of-change of the credit spread.

3.4 LSTM Model

In order to capture the direction-of-change for the credit spread, we continue to test other models. In this case, Neural Network is a good choice. In the conventional feed-forward neural networks, all test cases are considered to be independent. That is, when fitting the model for a particular day, there is no consideration for the data on previous days. In order to capture the time-dependency, we can use Recurrent Neural Networks (RNN). However, RNN only remembers for small durations of time. Therefore, we can use Long short-term memory (LSTM), which proves to be an effective solution to this shortcoming.

A recurrent neural network is simply a type of densely connect neural network as shown in Figure 12.

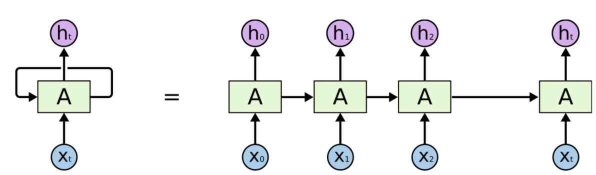

Fig. 12. . Structure of a Basic Recurrent Neural Network(RNN)

However, simply using one activation function in each layer can cause problems, for example, vanishing or exploding gradient which makes prediction hard for time series data. Therefore, we introduce LSTM to overcome this shortcoming; the key difference to normal feed forward networks is the introduction of time - in particular, the output of the hidden layer in a recurrent neural network is fed back into itself.

The structure of a typical LSTM cell is shown in Figure 13. Below we illustrate this structure:

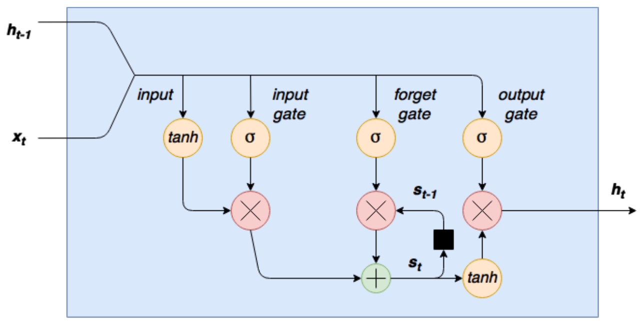

Fig. 13. . Structure of a Typical LSTM Cell

- "tanh" denotes the hyperbolic tangent function:

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad \textbf{(8)}$$

  which is used to normalize the input x to a value range from -1 to 1.
- σ denotes the sigmoid function:

$$\sigma(x) = \frac{e^x}{1 + e^x} \quad \textbf{(9)}$$

- x_t and h_t denote the input and output at time t respectively.
- $\times$ and $+$ denotes the element-wise multiply and add respectively.

The basic concept is as follows: By concatenating x_t and h_{t-1}, passing through input, forget and output gates respectively, the time-dependence and effects of previous inputs are well controlled (remembered or forgotten) and this leads to h_t, which is the output at time t. Below we describes the data flow within the LSTM cell:

- Input
  At first, the input is normalized between -1 and 1 using $tanh$ activation function as follows

$$g = tanh(b_g + x_t U_g + h_{t-1} V_g) \quad \textbf{(10)}$$

  where U_g and V_g are the weights for the input and previous cell output respectively, and b_g is the input bias.
- Input Gate
  The input gate is a hidden layer of sigmoid-activated nodes, which is denoted as a σ in Figure 13. Mathematically the expression can be written as

$$i = \sigma(b_{input} + x_t U_{input} + h_{t-1} V_{input}) \quad \textbf{(11)}$$

  After input and input gate, the results are multiplied element-wise as

$$g \circ i \quad \textbf{(12)}$$

  to be simple, the output i from input gate acts as the weights for input g.
- Internal State S
  This is the essential part of LSTM model and is achieved by a gating mechanism called GRU, gated recurrent unit(2014)[11]. The internal state plays an important role in generating next-step prediction result. The state is delayed by 1 and is added into $g \circ i$ to provide an internal recurrence loop to learn the relationship between inputs separated by time. It connects the forget gate result and the final prediction and allows the model to decide "how long it should keep the memory".
- Forget Gate
  The forget gate is another layer of sigmoid-activated set of nodes and is multiplied element-wise by S_{t-1}. As the name suggests, an output close to 1 means to remember and an output close to 0 means to "forget". The result f can be expressed as:

$$f = \sigma(b_{forget} + x_t U_{forget} + h_{t-1} V_{forget}) \quad \textbf{(13)}$$

  above follows by

$$s_t = s_{t-1} \circ f + g \circ i \quad \textbf{(14)}$$

- Output Gate:
  The output gate is the final layer of sigmoid-activated set of nodes.

$$o = \sigma(b_{output} + x_t U_{output} + h_{t-1} V_{output}) \quad \textbf{(15)}$$

  So the final output can be expressed as:

$$h_t = tanh(s_t) \circ o \quad \textbf{(16)}$$

For our analysis, multiple LSTM models with different input variables are tested. For the first model, we use the same features as in BART model and also a three-day look-back period. In other words, the credit spread at time $t-1, t-2$, and $t-3$ are used to feed the model in order to get the prediction of CS_t. The new model makes a relatively good prediction shown in Figure 14, however it seems to be worse than the BART model.

In the test set, we define direction-of-change as the difference between the next-period credit spread and the current spread and take the sign of this value. The formula is shown below:

$$\text{predicted direction of CS} = \text{sign}(\hat{\text{CS}}_{t+1} - \text{CS}_t) \quad \textbf{(17)}$$

In the confusion matrix in Figure 15, we can see that in the test, LSTM performs extremely well. The accuracy (total true positive divided by total predicted positive) is 96.1%

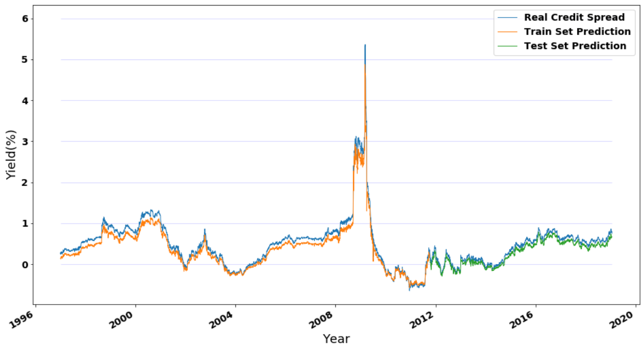

Fig. 14. LSTM Model Prediction on whole data between year 1996-2019

Proportion of Predicted		TRUE VALUE	
		Positive	Negative
PREDICTED VALUE	Positive	96.1%	3.9%
	Negative	18.4%	81.6%

Proportion of Actual		TRUE VALUE	
		Positive	Negative
PREDICTED VALUE	Positive	83.8%	4.7%
	Negative	16.2%	95.3%

Fig. 15. Confusion Matrix for LSTM Model

and the negative predictive value (total true negative divided by total predicted negative) is 81.6%.

Figure 16 compares the last two model results on the test set. Below are some observations:
1. Although BART captures the average credit spread well, it actually performs poorly in predicting direction-of-change. That is to say, BART model is shortsighted and puts too much weight on recent observations, which undermines its predictive power on credit spread movements.
2. LSTM with features performs better than the one without features and this supports our inclusion of feature variables.
3. Although LSTM can effectively predict the direction-of-change, LSTM tends to underestimate the magnitude of the credit spread.

The accuracy below is calculated by test set correctly predicted direction of credit spread number divided by the total data number in test set.
Usually, a machine learning model is evaluated by residuals between predicted outcomes and true outcomes. Here we use

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - f(x_i))^2} \qquad \textbf{(18)}$$

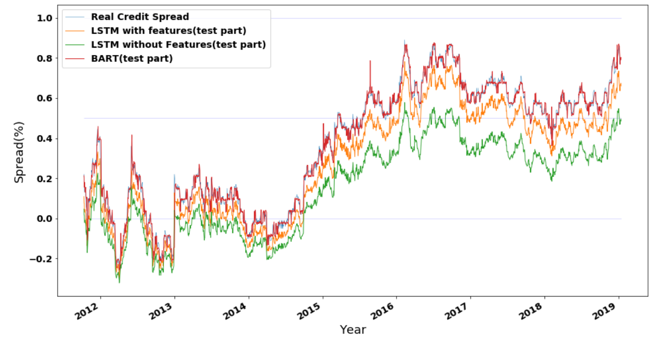

Fig. 16. Model Prediction Comparison

specifically.

	TRAIN RMSE	TEST RMSE	ACCURACY
lookback=0; all features	0.7	0.31	58.5%
lookback=1; all features	0.28	0.12	86.7%
lookback=2; all features	0.27	0.16	83.0%
lookback=3; all features	0.15	0.1	88.9%
lookback=3; no features	0.27	0.22	85.6%
lookback=4; all features	0.17	0.07	88.4%
lookback=5; all features	0.19	0.11	86.3%

Fig. 17. Model Comparison. We tuned the parameters and ran different models to get the best parameters and features. A smaller train and test RMSE and higher accuracy indicate a better model. Test RMSEs are smaller because the test set does not contain the credit spread spike during the recession.

As Figure 17 shows, we tune the number of features and look-back period parameters in order to find the model with highest accuracy and lowest RMSE. We find that the best model has all features and look-back period of 3. The model with look-back period of 3 and no features also performs well, but has lower accuracy and higher RMSE than the model with all features included. Models with a higher look-back period than 3, begin to lose accuracy and RMSE increases. We believe that a look-back period of 3 or 4 and containing all features can provide decent predictions.

4. Conclusion

In this paper, we used different mathematical and machine learning methods to predict the direction and magnitude of credit spread movement, and managed to get a final accuracy of almost 90%. In general we followed a trial and error process, trying to fix the model drawback by introducing new techniques. Below we provide our result as well as some potential improvements.

4.1 Results

At first, we find that the credit spread follows different dynamics in different regimes by fitting the data to a Vasicek model in moving time windows. We could see that the parameters are not time-homogeneous.
Then random forest model is introduced. However it cannot predict well for the direction-of-change. This could be due to the labeling process, because we might lose some information of the original data.

To solve the missing value problem as well as improving the tree model performance, Bayesian Additive Regression Tree (BART) model is used and provides a good fit and prediction for next day's credit spread. However, the direction of credit spread cannot be predicted well based on the result of our confusion matrix.
To tackle the new problem, Long Short-Term Memory (LSTM) model does a great job in predicting the direction of credit spread. Part of the reason is that the credit spread follows different patterns among different regimes. LSTM can gradually "forget" the inputs from older data in order to put more weight on current features.
The features mentioned in the literature review play an important role in predicting the credit spread movement and can improve model performance significantly.
Overall, the LSTM model with look-back period equal to 3 or 4 is the best model with both relatively low RMSE and high accuracy for out-of-sample predictions of the direction-of-change of credit spreads. Incorporating additional lag terms does not help much in predicting the out-of-sample direction.

4.2 Potential Improvement

1. In the future, we can include some latent variables to partition different time regimes and detect regime shifts.
2. We may combine the direction prediction from LSTM and magnitude prediction from BART into one model to improve the accuracy of prediction.

Acknowledgment

We would like to especially thank Professor Romesh Saigal, Professor Yang Chen and Rishi Sonthalia for their valuable suggestions.

Reference

[1]Bedendo, M., Cathcart, L., & El-Jahel, L. (2007). The slope of the term structure of credit spreads: An empirical investigation. Journal of Financial Research, 30(2), 237-257.

[2]Chiras, D. P., & Manaster, S. (1978). The information content of option prices and a test of market efficiency. Journal of Financial Economics, 6(2-3), 213-234.

[3]Collin-Dufresn, P., Goldstein, R. S., & Martin, J. S. (2001). The determinants of credit spread changes. The Journal of Finance, 56(6), 2177-2207.

[4]Tauchen, G., & Zhou, H. (2011). Realized jumps on financial markets and predicting credit spreads. Journal of Econometrics, 160(1), 102-118.

[5]Lin, H., Liu, S., & Wu, C. (2011). Dissecting corporate bond and CDS spreads. The Journal of Fixed Income, 20(3), 7.

[6]Morris, C., Neal, R., & Rolph, D. (1998). Credit spreads and interest rates: A cointegration approach (No. 98-08). Federal Reserve Bank of Kansas City.

[7]Chipman, H. A., George, E. I., & McCulloch, R. E. (2008). June, 2008. arXiv preprint arXiv:0806.3286.

[8]Malhotra, P., Vig, L., Shroff, G., & Agarwal, P. (2015, April). Long short term memory networks for anomaly detection in time series. In Proceedings (p. 89). Presses universitaires de Louvain.

[9]Duffie, D., & Huang, M. (1996). Swap rates and credit quality. The Journal of Finance, 51(3), 921-949.

[10]Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

[11]Cho, Kyunghyun; van Merrienboer, Bart; Gulcehre, Caglar; Bahdanau, Dzmitry; Bougares, Fethi; Schwenk, Holger; Bengio, Yoshua (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv:1406.1078