Strategic Predictions and Explanations By Machine Learning

The Prediction Model of Credit Default Swaps for the Telecommunication Service Sector

Caesar Wu,
SnT /Faculty of Science, Technology and Medicine (FSMT),
University of Luxembourg
Luxembourg
https://orcid.org/0000-0002-2792-6466

Jian Li, PhD,
Institute for Advance Economic Research, Dongbei University of Finance and Economics & University of Luxembourg
https://orcid.org/0000-0001-6865-0120

Jingjing Xu,
Faculty of Science, Technology and Medicine(FSTM),
University of Luxembourg
Luxembourg
https://orcid.org/0000-0002-0012-3911

Pascal Bouvry,
Faculty of Science, Technology and Medicine(FSTM),
University of Luxembourg
https://orcid.org/0000-0001-9338-2834

***Abstract*—Many machine learning (ML) models can make predictions regarding credit default swaps (CDS) for the telecommunication (telco) service sector. However, some ML algorithms can only offer a black-box model. It is crucial to explain the prediction result for strategic decisions. We compare various the state of arts, including deep learning (transformers), gradient boost machine (GBM), and Xgboost, plus different explainable tools: Variable Importance (VI) Partial Dependent Plots (PDP), Local Individual Conditional Expectation (LIME), Interpretable Model-agnostic Explanations (ICE), and Shapley values for the prediction model. Moreover, we also conducted a hyperparameter search of the prediction model by leveraging high-performance computing (HPC). Our experiment results show that the Xgboost provides the best solution with fewer constraints. We aim to find an optimal solution for strategic CDS investment decisions.**

***Keywords*—Machine Learning, Telecommunication Services, Credit Default Swap, Tree-Based Learning, Deep Learning, Transformer, Gradient Boost Machine, High-Performance Computing, Prediction**

I. Introduction

With growing digitization and increasing dataset sizes, various artificial intelligence (AI), machine learning (ML), and deep learning (DL) tools enable better strategic decision-making because ML tools can extract subtle and novel insights from data. However, a critical challenge facing many decision-makers [1] is dealing with ever-growing data size, selecting the right ML algorithm for the right dataset, and explaining results in the right context.

In addition, the more intricate question is how to explain the computational results. [2] Should we explain or interpret them? [3] Pasquale [4] suggested that we should explain the results from the perspectives of social, political, and economic implications because algorithms and automated decision-making processes now govern nearly all aspects of our lives and society. On the other hand, interpretations are justifiable and acceptable as long as they have reasons [5].

To a certain extent, the problem in explaining artificial intelligence (XAI) is the problem of causation. The Nobel laureate Guido Imbens [6] addressed the causation issue eloquently from an economic perspective. The Turing Award winner, Judea Pearl [8], explained the reasons why [7] from a computer sciences perspective. They have laid the foundations for XAI. In the recent literature review [9], Athey and Imbens highlighted some newly developed ML methods that can be applied to particular classes of econometric problems. Under the hood of supervised learning, they discussed four specific sets of ML methods: 1.) regularized linear regression: Least Absolute Shrinkage and Selection Operator (LASSO), Ridge, and Elastic Nets; 2.) regression tree and forest; 3.) deep learning and neural nets; 4.) boosting. Alberto et al. [10] proposed an empirical framework for value of Evidence-Based Decision-Making (EBDM) and the return on investment in statistical precision. They suggested adopting the current empirical tools to quantify the value of EBDM.

A. Motivations and Research Problem

We define our research question as: How to find an optimal ML solution that can make a better prediction of Credit Default Swaps (CDS) of the telecommunication (telco) service sectors for a given dataset (Fig. 1. A Scatter Plot of CDS for Telco Services Sector)? How can we explain the prediction results?

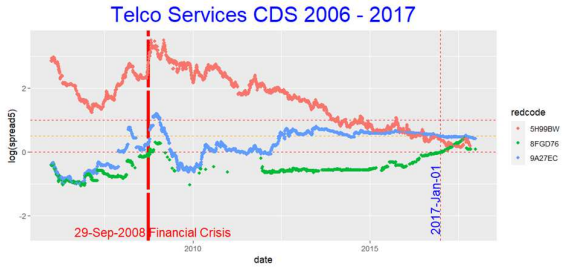

Fig. 1. : Scatter Plot of CDS for Telco Services Sector

The scatter plot represents a subset (7,842 observations) of a preprocessed large dataset (749,783 observations). It consists of 139 features, of which there are 126 trainable features. The y-axis exhibits a 5-year spread (or 5-year insurance contract rate in a 10-log scale), while the x-axis is a time domain from 3/Jan/2006 until 29/Dec/2017. The dataset only has three companies represented by "redcode". The initial plot illustrates that the 5-year spread rate trends between zero and one after the end of 2016. The reason for selecting a 5-year spread is that it is the most liquid spread in market trading. According to Kaggle's CDS dataset [11], there are ten spread values from 1 year to 10 years. These spreads are highly correlated; the closer the spreads, the higher their correlation.

The motivation of this work is to define a prediction model for the telco CDS so that a strategic investor can make a wiser decision in the CDS derivative market. Moreover, we want to highlight some critical issues or meta-reasons in the telco industry.

B. Research Method

With the given and pre-processed dataset, we adopt a quantitative research method to describe, explore, understand, predict, explain, and evaluate the telco CDS data in a derivatives market. The method consists of five essential steps: 1.) Take a bird's eye view of the CDS data (See Fig. 1). 2.) Form a hypothesis. 3.) Test the hypothesis. 4.) Analyze the experiment results. 5.) Conclude and point out the future research direction.

C. Main Contributions

This study has yielded three primary contributions, as outlined through the analysis of experimental results:

- We find the optimal prediction model for the Telco CDS spread 5. To the best of our knowledge, this is the first time using various ML methods to select the best prediction model for the telco CDS 5-year spread. This finding implies that the prediction model can help many strategic decision-makers to invest wisely.
- In contrast to many traditional ML models that simply offer a black box model, this study offers five different ways of explanations for the results. We intend to build some causation inference power between a prediction result and its feature variables.
- In order to overcome the limited computational power for a hyperparameter search, we train our ML models on a large High-Performance Computing (HPC) platform, which lays the foundation for future research with a few hundred features and a large scale of longitudinal datasets under a cloud environment. The study provides a scalable approach for many intricate and strategic prediction problems.

D. Scope of this Research

The rest of the paper is organized into five parts: Section II is a literature review of three aspects: 1.) tree-based learning; 2.) Deep Learning/transformer; 3.) CDS derivatives. Section III provides experiment assumptions, configurations and results, including tree-based prediction and deep neural network models. In addition, we offer explanatory results for tree-based prediction models. Section IV is a detailed analysis with explanations from different perspectives. Section V presents conclusions and highlights the future works.

II. Literature Review

A. Tree-Based Machine Learning Models

The origin of the tree-based ML models can be traced back to a Decision Tree (DT) or Classification and Regression Trees (CART) [12]. The basic idea of the decision tree is to make inquiries intelligently based on available data and expect the result to be an accurate prediction. Compared with other nonparametric algorithms, the primary benefit of DT is that it offers a certain degree of transparency and explanation for the prediction model [13].

Historically, tree-based methods have evolved from the CART in the 1980s to bagging, random forest, and boosting iterations [14][15]. We can roughly divide this evolution into four phases: 1.) CART. 2.) Bagging bootstrap aggregation. 3.) Random Forests. 4.) Boosting iterations (See Fig. 2). The last three periods can be summarized as ensemble learning. The essence of ensemble learning is the "wisdom of crowds" [16]. Researchers have developed about ten major boosting strategies under the hood of boosting iterations. We can classify boosting models into three classes: 1.) Adaptive boosting, which is the earliest algorithm and its extension for different tasks. It is very slow in comparison to the next generation of models. 2.) Gradient Boosting Machine (GBM), which is based on Frieman's idea of greedy function approximation [17] and its extension. The Xgboost is one of GBM's extensions. The Xgboost method becomes popular because we can run the algorithm in parallel on the HPC platform. 3.) Boosting models for particular types of datasets.

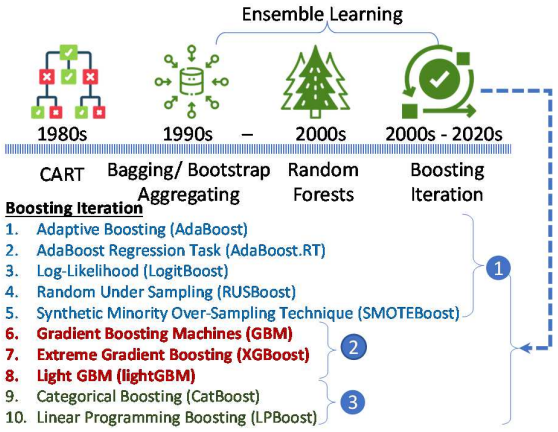

Fig. 2. : Evolution of Tree-Based Learning Methods

We mainly focus on GBM and Xgboost in this paper because GBMs have proven to be one of the most successful ML algorithms in many competitions[11]. The essence of GBM is an iterative learning process. The "G" or gradient, represents a steep descent. The "B" or boosting means boosting from weak models. The concept of GBM can be mathematically represented [32] as follows:

$$f^* = \underset{f}{\operatorname{argmin}}\, L(f)\,;\; L(f) = \sum_{i=1}^{N} L\,|y_i, f(x_i)| \tag{1}$$

$$f_B = \sum_{b=0}^{B} f_b,\; f_b \in \mathbb{R}^N;\; g_b = \left\{ \left[\frac{\partial L(f)}{\partial f} \right]_{f=f_{b-1}(x_i)} \right\}_{i=1}^{N} \tag{2}$$

Where f^* is an optimal prediction function based on a genetic function f. $L(f)$ is a loss function. x_i $(i = 1, 2, \ldots N)$ means "i" observation and $y_i$is a predicted result. f_B means the sum of "B" or the overall boosting functions based on N-features and f_b represents a weak learner of boosting. g_b is the steepest descent.

According to [12], a good modelling tool for data mining should at least: 1.) identify which features are more important than others; 2.) disclose the relationship between independent (predictors) and dependent (features) variables; 3.) be scalable or handle large datasets; 4.) can tolerate missing values well; 5.) show how the independent variables interact; 6.) display a big picture of the dataset; 7.) reveal any novelty and outlier cases. Tree-based learning has all these advantages. However, it comes with costs, especially the GBM algorithm, because it is: 1.) computationally expensive; 2.) potentially for

overfitting; 3.) hyperparameter tuning is very challenging; 4.) sensitivity to noise and outliers; 5.) biased towards strong learners; 6.) requiring one-hot encoding if it contains categorical data; 7.) struggling to handle imbalanced datasets; 8.) less effective to work with high-dimensional data; 9.) lacks transparency.

To overcome transparency issues, a Partial Dependence Plot (PDP) [17] is a visualizing tool that can help interpret the relationship between the prediction result and a subset of features. However, the PDP could be misleading if influenced features result in a highly intertwined prediction. To solve this issue, Goldstein et al. [18] proposed Individual Conditional Expectation (ICE) plots. ICE plots show the estimated relation between the predicted result and each observation. Both PDP and ICE tools provide a global interpretation approximately. To interpret the prediction result locally, Ribeiro et al. [19] introduced Local Interpretable Model-agnostic Explanations (LIME), which is a visualization tool that helps explain individual predictions. Similarly, Shapley [20] used game theory to interpret individual predictions, which assumes each feature is independent.

B. Deep Learning and Transformer Models

Deep learning and transformer architecture [22] have generated attention due to their remarkable performance in many applications, especially in time series. The transformer of time series can be classified into three categories: forecasting, anomaly detection, and classification [23]. Researchers have developed many models, such as baseline or vanilla transformers [22], patch time-series transformers (PatchTST) [24], and time-series neural networks (TimesNet) [25]. However, most of these models primarily focus on developing novel techniques for better performance. In contrast, we aim to apply these models to the real-world dataset, namely the telco CDS, for investment decisions.

The advantages of Transformer models are parallelization, long-range dependencies (capturing long-range dependencies in input sequences), scalability, flexible input length, transfer learning, and multi-modal applications. However, it also comes with costs: 1.) The architecture is more complex than the traditional models like Recurrent Neural Network (RNN). 2.) It requires a large dataset and more computational resources, such as GPUs and TPUs. 3.) While self-attention is powerful, it introduces quadratic complexity regarding input sequence length. 4.) It does not inherently preserve the sequential order of tokens in the same way as RNNs. 5.) Limited local context. 6.) Hyperparameter searching can sometimes be challenging because the model might be hard to interpret or control, especially for credit risk applications.

Credit Defaults Swap for Telco Service Sector

The topic of how to model credit risk has been intensively discussed in the credit risk literature since the 1960s and the 1980s. Early accounting-based models, such as Z-score [26] and O-score [27], have been widely adopted to predict a firm's default (failure to make payment) risks. The next generation of credit risk models is known as market-based variables or distance to default (DTD), developed by Merton [28]. It calculates the default distance, a conditional probability variable between the market equity and accounting data for the firm's liabilities. Although the model has been widely recognized in academia and industry, it overemphasizes the distance to default. Duffie and Lando [29] argued that if the markets are not fully transparent, DTD could filter out some critical information. Bai and Wu [30] found that if we can combine DTD with firms' fundamentals, the combination model can predict and explain 77% of the CDS spreads.

The result leads to data-driven prediction and analysis. Guenduez and Uhrig-Homburg [31] did research on CDS spreads. The study intends to extend the previous research by leveraging ML models for the CDS spread. However, a gap still remains in finding an optimal solution by leveraging a hyperparameter search. It leads to our proposed solution.

EXPERIMENTAL DESIGN AND RESULTS

We run two tree-based learning (GBM and Xgboost) experiments with the same data. We first split the telco CDS dataset into a 70:30 ratio, which is 70% for training and 30% for testing with 5-fold cross-validation. And then, we use PDP, ICE, LIME, and Shapley values to explain the prediction results.

C. Experiment with Assumptions and Set up

A medium HPC cluster is set up with 128 cores and 256 GB of memory. While Amazon Web Services (AWS) cloud resources are an option, the HPC infrastructure should possess sufficient capacity to handle the relatively small size of the telco CDS dataset. Before starting the experiments, we 1.) removed all duplication and missing values plus zero and no variance features. 2.) increased the CDS spread value by 100.

D. Tree-Based Learning Experiments

a) GBM experiment

The parameters of the first GBM experiment are based on an educated guess. We set four parameters to train the GBM model: 1.) The total number of trees to fit equals 10,000. 2.) The maximum depth of each tree is one. 3.) The learning rate is 0.001. 4.) The cross-validation is five-fold. The experiment only takes about two minutes on a laptop and about one second on the HPC cluster. It is tolerable to wait two minutes. The model prediction error is about 1.183 in root mean square error (RMSE). However, the initial parameters are not optimal. To find an optimal solution for the prediction model, we have to do a hyperparameter search. HPC demonstrates its computational power for the search task. Table I shows that HPC is about two times faster than a single machine in the GBM model for 81 grid points of hyperparameter search regarding learning rates, tree deep, end nodes, and bagging fraction. Each parameter has three different values.

The final and optimal GBM prediction model with the training dataset offers a 0.317 RMSE of prediction result. It is about 3.7 times better than educated guess parameters.

a) Xgboost experiments

Extreme Gradient Boosting (Xgboost) usually is about 8-10 times faster than GBM. The main advantage of Xgboost is the ability to parallel computation even on a single machine. As evident from Table I, the computational (user) time of HPC is significantly larger than the elapsed time compared to a single machine. Table 1 shows that HPC is about eight times faster than a single machine for 243 grid points hyperparameter searches. The final optimal Xgboost model produces 0.238 RMSE in training, about 24.92% better than the GBM. The Mean Absolute Errors (MAE) and Mean Squared Error (MSE) are also shown in Table I. One of the remarkable features of Xgboost is an early stopping mechanism if the cross-validated error does not improve for "k" continuous trees.

TABLE I. TREE-BASED GBM EXPERIMENTS COMPARISON

GBM	GBM Hyperparameter (81 grid points)			Entire Dataset (final)	
Time	User	System	**Elapsed**	MAE	MSE
Laptop (sec.)	1,146.810	3.970	**2,175.600**	0.154	0.147
HPC (sec.)	1,218.238	0.047	**1,220.565**	0.148	0.085
Xgboost	Xgboost Hyperparameter (243 grid points)			MAE	MSE
Laptop (sec.)	9,516.195	19.432	**13,275.214**	0.099	0.061
HPC (sec.)	207,043.845	30.898	**1,641.976**	**0.121**	**0.057**

E. Transformer Models Experiments

We can also use DL to train a prediction model. To fit the transformer architecture along the time domain, we separate the entire dataset into three subsets by a company's redcode. One company has a higher data missing rate regarding the New York Stock Exchange (NYSE) Calendar. Table II summarizes the details of each subset.

TABLE II. DATA SIZE AND DATA MISSING RATE

Redcode	**5H. (5H99BW)**	**8F. (8FGD76)**	**9A. (9A27EC)**
Data points	2,938	1,929	2,975
NYSE Trading Days	2,978	**3,019**	3,019
Missing Data Rate	1.34%	**36.10%**	1.46%

First, we split each subset into a 70:10:20 ratio for training, validation, and testing. Second, we employ three transformer models – baseline, TimesNet, and PatchTST- for experiments on a standalone server with one GPU and two cores. Table III shows that TimesNet generally outperforms other models because it can deal with the significant missing data. In comparison, other models struggle with this higher missing data rate. Due to the GPU resource constraints, we did not implement a hyperparameter search in this experiment.

TABLE III. TRANSFORMER-BASED EXPERIMENTS COMPARISON

Models/ Results	5H.		8F.		9A.		Average of all three		Total Run Time (sec.)
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	
PatchTST	0.002	0.035	1.722	1.140	0.006	0.401	0.577	0.412	63.628
TimesNet	0.006	0.053	1.517	1.034	0.012	0.357	0.512	0.392	183.932
baseline	0.633	0.664	16.033	3.708	0.468	4.123	5.711	1.652	109.132

F. Features Explanation & Variable Contributions

After finding the optimal solution, the question is how to explain the result produced by the Xgboost model. We use five different tools to explain the results: variable importance (VI), partial dependence plots (PDP), individual conditional expectation (ICE), local interpretable model-agnostic explanations (LIME), and Shapley values.

The VI is the relative influence plot regarding each feature (See Fig. 3). It shows which feature contributes to the prediction model. The highest contribution feature is the ratio of debt and the earnings before interest taxes depreciation amortization (EBITDA). The second is a ratio of the total debt to assets (debt_at). The third one is slightly subtle because if we include all categorical variables, it is price book (ptb) (See Fig. 3). However, if we exclude all categorical variables (i.e. redcode), cash_debt (cash flow versus debt) will become the third on the list. It might suggest that the cash_debt feature is highly correlated with other features. We can use Accumulated Local Effect (ALE) to examine this correlation. However, due to space constraints, we leave this issue to future research.

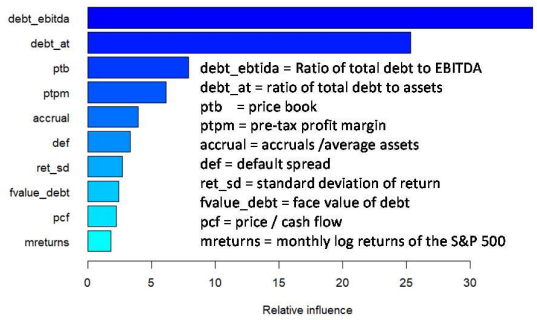

Fig. 3. : Top Ten Variables Importance Plots

The second explanation tool is PDP. It illustrates the detailed variation between the CDS prediction and a particular feature. For example, if the ratio of debt_ebidta is around 5, the CDS risk increases significantly (Refer to Fig. 4). However, PDP assumes that features are not correlated.

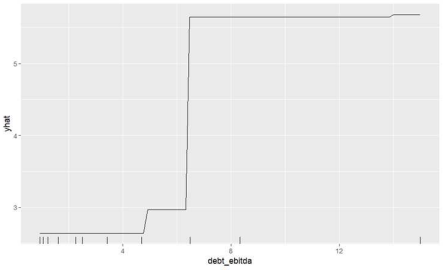

Fig. 4. : Partial Dependence Plots (PDP) for debt_ebtida

We can also use the ICE curve plot to see how each observation contributes to the overall prediction model shown in Fig. 5. Notice that we have two ICE curve plots. One is a stack (left), and the other is a centre (right). The central ICE plot can highlight heterogeneity in our prediction model.

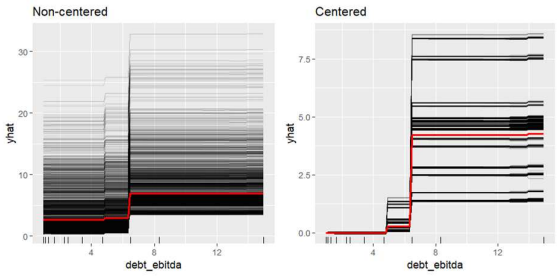

Fig. 5. : ICE Plot explanation for debt ebitda

In other words, the PDP (Fig. 4) is the average plot of several clusters of instances regarding the CDS risk (< 2.5, 5, and > 7.5). However, the cluster of 5.5 dominates the overall prediction result.

In addition to the ICE curve plot, LIME also helps us see inside the black box of the prediction model for some individual cases. We select two sets of cases. One set is before

the 2008 financial crisis, and the other is near 2017. Each case set consists of four cases (See Fig. 6).

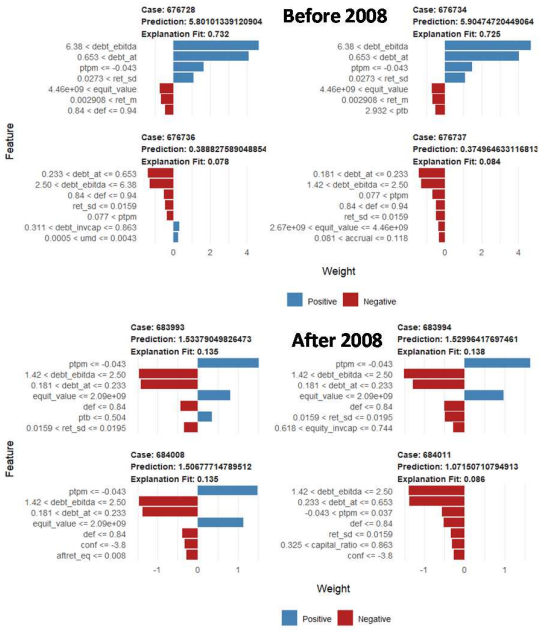

Fig. 6. : LIME Plot

Compared with Fig. 1, there are two types of telco companies. One is a high-risk CDS shown on the top row. The other is a lower-risk type, shown in the second row. Nevertheless, all telco companies converged after 2017, as shown at the bottom of the two rows. Similarly, we can also use Shapley values to explain the prediction result. The explanation tool answers the prediction results with why and by how much (See Fig. 7).

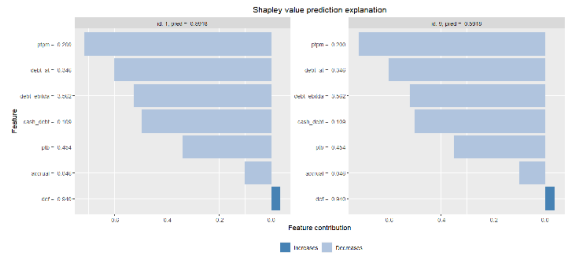

Fig. 7. : Shapley Value Plot

Based on the VI (Refer to Fig. 3), we select the top six and the ratio of "cash to debt" features for Shapley value analysis. However, different distribution approaches could lead to different analysis results. Therefore, we use a combination of empirical, Gaussian, coupla, and conditional trees (ctree) distribution for the analysis shown in Fig. 7.

III. Result Analysis

Drawing from the preceding experiments, key concerns in modelling telco CDS spreads come to light: 1.) How do we understand the data at first glance? 2.) How do we pre-process the dataset? 3.) How do we select the right ML algorithm for training, validation, and testing? 4.) How do we implement hyperparameter search? 5.) How do we leverage HPC or cloud resources? 6.) How do we use different ML tools to explain and interpret the prediction results?

A. Understanding Dataset and The Goal of Analysis

Before delving into ML, it is critical to gain a clear understanding of the dataset and the goal. This research shows the overall plot of the 5-year spread (see Fig. 1). In comparison with the technology (37,526 observations) and utility (65,765 observations) sectors (see Fig. 8), the CDS of the telco sector varies between -1 and +3 (in a log scale) in 2008 and stabilize the variation just between 0 and +0.5 after the end of 2016 while utilities and technology's CDS continues to fluctuate between -2 and +2. Utilities are between -2 and 1, but the overall trend is decreasing.

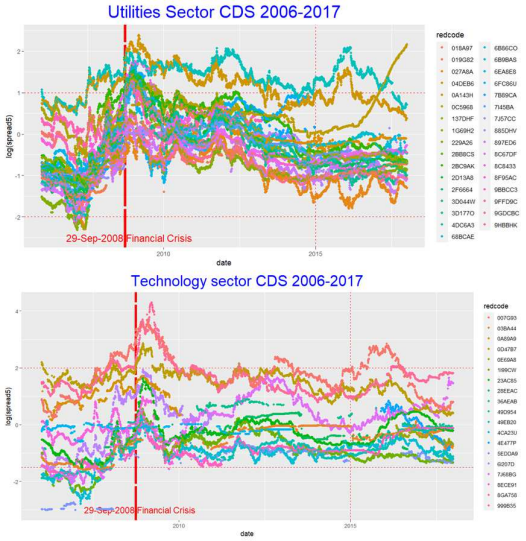

Fig. 8: Technology and Utility Sectors of CDS

The reason for comparing the technology and utilities sectors is that the Telco company can be considered both a utility and a technology firm. The main goal of this study is to understand the meta-reasons and underlying factors driving this trend. An interpretative meta-reason or speculative factor could be attributed to lower levels of innovation and competition within the telco sector.

B. GBM, Xgboost and Hyperparameters Analysis

The GBM method is considered ensemble learning (See Fig. 2), which is essentially a meta-learning algorithm that learns how to combine the final prediction from various ensemble members. However, the main disadvantage of GBM is that it is computationally too expensive. To overcome this issue, we adopt the Xgboost, which makes it possible to run a parallel computation, especially for hyperparameter searching.

Our experiment illustrates that we should take a small hyperparameter search with a few parameters on a single machine and then implement a full-scale parameters search on the HPC, which could result in an optimal solution [21]. However, most grid points could add little value to the final prediction model. It is not worth wasting much computational power for minor improvement during training.

C. Predicting Results and Interpret Tools

For training, we can achieve the best RMSE of 0.238. However, the final prediction model in a testing dataset is 0.303. If we want an overall picture to explain the prediction result, VI, PDP, and ICE are good explanation tools. Nevertheless, PDP is the average of all observations,

assuming independent features. ICE provides the details of each observation's contribution to the prediction model.

LIME is a good tool for local interpretability but is too sensitive to sampling. Its explanation result can vary with even very small input changes. Shapley value analysis can reflect a fair allocation of contribution for each feature. It is relatively constant, but it is computational complexity. It does not capture nonlinear interactions due to its additive nature.

IV. Conclusions and Future Direction

A. Conclusions

The experimental results demonstrate that Xgboost is a better model for the telco CDS dataset than other ML models. If we just focus on a transformer, TimeNet is the better model without hyperparameter tuning, but PatchTST performs better if the dataset has a few missing data. Although every investor understands that the ratio of debt and EBITDA is an important metric for the CDS, VI illustrates that it is the most crucial metric. PDP tells us that the risk is significantly high if this ratio exceeds five. The result offers insight for some strategic investors in a derivative market. However, the caveat is that this conclusion is based on the limited data size (only three telco companies). More research needs to be done.

B. Future Research Direction

In future, we will focus on a generalized prediction model for different applications, especially to find a better solution with hyperparameter search for transformer models by better utilizing HPC resources. Furthermore, we intend to devise a universal approach for explaining and understanding the prediction outcomes concerning strategic decisions, such as the 10-year CDS spread investment or other strategic options.

Acknowledgement

This research was funded by the Luxembourg National Research Fund (FNR), grant ID C21/IS/16221483/CBD and grant ID 15748747. For open access, the author has applied a Creative Commons Attribution 4.0 International (CC BY 4.0) license to any Author Accepted Manuscript version arising from this submission.

References

[1] Papadakis V and Barwise P, editors. "Strategic decisions", Springer Science & Business Media; 2012 Dec 6. pp. 35-50 https://doi.org/10.1007/978-1-4615-6195-8

[2] Marcin Milkowski, " Explaining the Computational Mind", The MIT Press, April 2013. pp. 97–149, *https://doi.org/10.7551/mitpress/9339.003.0001,*

[3] Rudin C. "Stop explaining black-box machine learning models for high-stakes decisions and use interpretable models instead. Nature machine intelligence." 2019 May; 1(5) : 206-15, https://doi.org/10.1038/s42256-019-0048-x

[4] Pasquale F, "The black box society: The secret algorithms that control money and information." Harvard University Press; 2015 Dec 31. https://doi.org/10.4159/harvard.9780674736061

[5] C Wu et al. "Survey of Trustworthy AI: A Meta Decision of AI", arXiv:2306.00380, 2023, https://doi.org/10.48550/arXiv.2306.00380

[6] Imbens GW, Rubin DB. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press; 2015 Apr 6. https://doi.org/10.1017/CBO9781139025751

[7] Pearl J, Mackenzie D. The book of why: the new science of Cause and Effect. Basic books; 2018 May 15. https://doi.org/10.1080/14697688.2019.1655928

[8] Pearl J. Causality. Cambridge university press; 2009 Sep 14

[9] Athey S and Imbens GW. Machine learning methods that economists should know about, Annual Review of Economics. 2019 Aug 2;11:685-725. https://doi.org/10.1146/annurev-economics-080217-053433

[10] Abadie A, et al., Estimating the Value of Evidence-Based Decision Making. arXiv preprint arXiv:2306.13681. 2023 Jun 21. https://doi.org/10.48550/arXiv.2306.13681

[11] https://www.kaggle.com/datasets/debashish311601/credit-default-swap-cds-prices

[12] Breiman L. Classification and regression trees. Routledge; 2017 Oct 19. https://doi.org/10.1201/9781315139470

[13] Lundberg SM, et al. From local explanations to global understanding with explainable AI for trees. Nature machine intelligence. 2020 Jan;2(1):56-67. https://doi.org/10.1038/s42256-019-0138-9

[14] Mayr A, et al., The evolution of boosting algorithms. Methods of information in medicine. 2014;53(06):419-27 https://doi.org/10.48550/arXiv.1403.1452

[15] He Z, et al., Gradient boosting machine: a survey. arXiv preprint arXiv:1908.06951. 2019, https://doi.org/10.48550/arXiv.1908.06951

[16] Surowiecki J. The wisdom of crowds. Anchor; 2005 Aug 16.

[17] Friedman JH. Greedy function approximation: a Gradient boosting machine. Annals of statistics. 2001 Oct 1:1189-232. https://doi.org/10.1016/S0167-9473(01)00065-2

[18] A Goldstein et al., .Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation, Journal of Computational and Graphical Statistics, 24(1):44–65,2015, https://doi.org/10.1080/10618600.2014.907095.

[19] Ribeiro MT, et al. "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778

[20] Lloyd S. Shapley, A Value for n-Person Games, pages 307–318. Princeton Uni. Press,2016 https://doi.org/10.1515/9781400881970018

[21] Smith LN and Topin N. Super-convergence: Very fast training of neural networks using large learning rates. In Artificial intelligence and machine learning for multi-domain operations applications 2019 May 10 (Vol. 11006, pp. 369-386). SPIE. https://doi.org/10.1117/12.2520589

[22] A.Vaswani, et al. "Attention is all you need," Advances in neural information processing systems, 2017, https://doi.org/10.48550/arXiv.1706.03762.

[23] Q.Wen, et al., Transformers in time series: A survey," arXiv preprint arXiv:2202.07125, 2022, https://doi.org/10.48550/arXiv.2202.07125

[24] H. Wu, et al., "Timesnet: Temporal 2d-variation modelling for general time series analysis," arXiv preprint arXiv:2210.02186, 2022. https://doi.org/10.48550/arXiv.2210.02186

[25] Y. Nie et al., "A time series is worth 64 words: Long-term forecasting with transformers," arXiv preprint arXiv:2211.14730, 2022. https://doi.org/10.48550/arXiv.2211.14730

[26] Altman, Edward I. "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy." The Journal of Finance 23.4 (1968): 589-609 https://doi.org/10.2307/2978933

[27] Ohlson, James A. "Financial ratios and the probabilistic prediction of bankruptcy." Journal of Accounting Research (1980): 109-131 https://doi.org/10.2307/2490395

[28] Merton, Robert C. "On the pricing of corporate debt: The risk structure of interest rates." The Journal of Finance 29.2 (1974): 449-470, https://doi.org/10.2307/2978814

[29] Duffie, Darrell, and David Lando. "Term structures of credit spreads with incomplete accounting information." Econometrica 69.3 (2001): 633-664, https://doi.org/10.1111/1468-0262.00208

[30] Bai, Jennie, and Liuren Wu. "Anchoring credit default swap spreads to firm fundamentals." Journal of Financial and Quantitative Analysis 51, no. 5 (2016): 1521-1543 https://www.jstor.org/stable/44157830

[31] Guenduez, Yalin, and Marliese Uhrig-Homburg. "Predicting credit default swap prices with financial and pure data-driven approaches." Quantitative Finance 11.12 (2011): 1709-1727, https://doi.org/10.1080/14697688.2010.531041

[32] James G, Witten D, Hastie T, Tibshirani R. "An introduction to statistical learning", New York: Springer; 2013 Jun 24 https://doi.org/10.1007/978-1-4614-7138-7