Interpreting financial time series with SHAP values

Karim El mokhtari
Ayse Bener
elmkarim@ryerson.ca
Data Science Laboratory, Ryerson University
Toronto, Canada

ABSTRACT

Coming soon ...

KEYWORDS

SHAP values, times series, model interpretation

ACM Reference Format:
Karim El mokhtari and Ayse Bener. 2019. Interpreting financial time series with SHAP values. In *Proceedings of CASCON 2019.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Creating forecasts and comparing them to actual results is vital to every company in the market. Our industrial partner is a consumer goods company that sells hundreds of products to customers worldwide. These customers are primarily supermarket chains and retailers. To monitor the performance of every product, monthly reports are generated to show sales volumes by product and customer. These reports are then aggregated by brands of products, and total sales volumes are compared with the company forecast to show the discrepancy between actual and forecast. Experts analyze the discrepancies and write short commentaries to describe the reason of every alarming deviation and the main customers behind it. To validate their findings, they often turn to managers in various departments such as customer service or warehouse for inquiries on a specific product or brand.

This process relies heavily on the expert's knowledge of the business areas, but it depends on data living in the different silos of the company as well. That is why our partner's management expressed the need to create a model to learn from the existing commentaries and reveal any particular pattern triggering the generation of commentaries.

In our previous paper [1] our main focus was on learning and predicting commentaries using NLP tools and recurrent neural networks. However, due to the insufficiency of commentaries from different categories and the diversity in the writing style and semantics between experts, the model was only able to predict very basic commentaries.

CASCON 2019, November 4, 2019 - November 6, 2019, Toronto, ON, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

As many commentaries are related to sales performance, we discussed with our partner to evaluate the Point of Sales (POS) data as a candidate dataset in commentaries prediction. This dataset is extracted from the database used by customers to daily manage their retail stores. It contains the time and place of all retail transactions. Access to POS data requires a special agreement between our partner and the customer and poses multiple security concerns. That is why, in this work, we use POS data from only one important customer to assess its usefulness.

In addition, our partner is interested in interpreting how the different models used predict a commentary, and particular patterns in the monthly results trigger commentary generation.

To achieve both objectives, we have reformulated by categorizing commentaries into multiples classes. We evaluate multiple machine learning models in commentary prediction. Then, we use explanation models to understand how every model computes the output. There are many methods to interpret a model prediction, but we found that SHAP values proposed by Lundberg and Lee [7] are more consistent as we will explain in Section 2.

This paper is organized as follows. In the next section, we present briefly the methods used in explaining models and the motivation behind choosing SHAP values. In section 3, we explain the methodology followed in assessing the relevance of the POS dataset and in explaining how every model computes its output. In section 4, we describe the experimental part and show the results. The conclusion summarizes the findings and develops future works.

2 MODEL PREDICTIONS INTERPRETATION

It is easy to interpret and understand the prediction of linear models. However, for non-linear models, the model itself cannot be used in interpretation as the prediction process is often considered as a block box. In such cases, we define an explanation model as an interpretable approximation of the complex model as illustrated in Figure 1.

2.1 Additive feature attribution methods

As proposed in [2], local methods are designed to explain the prediction model f by the explanation model g. Model g uses simplified inputs x' mapped to the original inputs via a function $x = h_x(x')$ specific to every original input x. Local methods try to approximate $f(h_x(z'))$ by $g(z')$ when $z' \approx x'$. In additive feature attribution methods, the explanation model is a linear combination of M binary variables z' that represent a feature being observed when

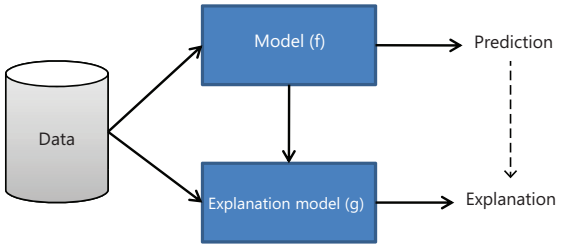

Figure 1: How an explanation model is used in predicting interpretation

$z' = 1$ or unknown otherwise:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{1}$$

M is the number of simplified input features. Coefficients ϕ_i are real numbers that show the effect of each feature. Summing up all the effects results in an approximation of the prediction $f(x)$ that we want to explain.

Several methods were proposed to match equation 1. The LIME method [2] finds ϕ_i by minimizing an objective function ξ described by equation 2 based on a squared loss L and a local kernel $\pi_{x'}$ to weight the simplified inputs. It uses a penalized linear regression in the minimization problem.

$$\xi = \underset{g \in \mathcal{G}}{\arg\min}\, L(f, g, \pi_{x'}) + \Omega(g) \tag{2}$$

DeepLIFT was proposed in [3, 4] as another additive feature attribution method that applies a recursive algorithm to explain deep neural networks prediction. It uses a "summation-to-delta" property to match equation 1.

The above methods rely on equations from cooperative game theory to compute prediction explanations using Shapley regression values [6], Shapley sampling values [4] and Quantitative input influence [5]. Shapley regression values consists in retraining the model f on every feature subset S from $F \setminus \{i\}$, the set of all features F excluding i. To calculate the effect of feature i, two models are trained, one including the feature $f_{S\cup\{i\}}$ and the second f_S withholding it. The difference between the output of both models $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$ computed for a specific input x_S shows the effect of feature i on the prediction. As the order of withholding feature i can affect the prediction in case feature i depends on other features in subset S, the model is retrained for all possible subsets $S \subseteq F \setminus \{i\}$. Equation 3 calculates Shapley values as a weighted average of all possible differences.

$$\phi_i = \sum_{S \subseteq F\setminus\{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \tag{3}$$

Shapley sampling values [4] reduce the high computational load of retraining the model of all combinations of S in equation 3. The method approximates the effect of excluding a feature by integrating over samples from the training set.

Lundberg and Lee [7] numbered three properties in the class of additive feature attribution methods: local accuracy, missingness and consistency. They demonstrated that methods not based on Shapley values violate at least one of these properties. They proposed a unified measure of feature importance called SHAP (SHapley Additive exPlanation) values that we explain in the next section.

2.2 SHAP values

It is demonstrated in [7] that equation 4 corresponds to the only explanation model that satisfies local accuracy, missingness and consistency.

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)] \tag{4}$$

$|z'|$ is the number of non-zero entries in z', and $z' \subseteq x'$ stands for all z' vectors where the non-zero entries are a subset of the non-zero entries in x'.

They defined SHAP values as a solution to this equation where f_x is a conditional expectation function of the original model: $f_x(z') = f(h_x(z')) = E[f(z)|z_S]$ with S the set of non-zero indexes in z'. However, exact computation of SHAP values is challenging. That's why they proposed several approximations in [7] assuming model linearity and feature independence. A version of SHAP values adapted to Tree ensembles was proposed recently in [8].

In this paper, we use KernelSHAP that is based one LIME [2]. The choice of the objective function parameters in equation 2 is not made heuristically, which violates the consistency property, instead it uses the following parameters proposed in [7] to recover the Shapley values:

$$\begin{aligned} \Omega(g) &= 0 \\ \pi_{x'} &= \frac{(M-1)}{(M \text{ choose } |z'|)|z'|(M - |z'|)} \\ L(f, g, \pi_{x'}) &= \sum_{z' \in Z} [f(h_x^{-1}(z')) - g(z')]^2 \pi_{x'}(z') \end{aligned}$$

—- Figure of vectors for SHAP (SHapley Additive exPlanation) values (1) ——

3 METHODOLOGY

In this study, we are interested in predicting commentary generation on financial data using binary classifiers. We use SHAP values theory to interpret the output of the prediction model. Additionally, we want to show the impact of the new POS dataset on the prediction accuracy. As mentioned above, this dataset requires access to the customer sales database and requires special agreement and security considerations between our partner and his customers. SHAP values are applied to explain the output computation for

Table 1: The VAR dataset describes discrepancies observed by brand for a chosen customer from January 2016 to present in Millions of CAD

CUSTOMER 1

Brand	Jan. 2016	Feb. 2016	...	June 2019
Brand 1	+0.10	-0.01	...	+0.05
Brand 2	-0.08	-0.05	...	+0.12
...	...	...	...	...

CUSTOMER 2

Brand	Jan. 2016	Feb. 2016	...	June 2019
Brand 1	-0.03	+0.11	...	-0.03
Brand 2	+0.14	-0.08	...	+0.06
...	...	...	...	...

Table 2: The COM dataset is an aggregation of VAR by brand and month, it shows discrepancies in Millions of CAD for each brand along with the financial expert commentaries

January 2016

Brand	Variance	Commentary
Brand 1	+0.50	Variance driven by CUSTOMER 10 and CUSTOMER 25, caused by over delivery
Brand 2	-0.71	CUSTOMER 15: Promotion in brand did less than the expected result
...	...	...

June 2019

Brand	Variance	Commentary
Brand 1	-0.39	CUSTOMER 12: High inventory
Brand 2	+0.63	CUSTOMER 10 and CUSTOMER 25: declining faster than seen in the market
...	...	...

every model.

We use the discrepancy dataset denoted VAR extracted from the company's ERP (Enterprise Resource Planning). It shows the monthly variance from January 2016 to present by brand and customer (see Table 1). The experts commentaries are provided monthly in the COM dataset shown on Table 2. The commentaries explain the performance of one or many customers, with a brief description of the root cause of the discrepancy observed for each one of them. Our goal is to create a model that learns to generate a commentary from the VAR dataset.

In our previous paper [1], the focus was on generating plain English commentaries with RNN structures. In this paper, we categorize commentaries into different classes then we experiment

Table 3: Binary classification performance with VAR dataset

	kNN	RF	SVM	XGB	LSTM
F1-score	0.72	0.69	0.71	0.67	0.66

several models: kNN, XGBoost, SVM, Random Forest and LSTM. Each one uses a different approach in computing the prediction.

We assume that a commentary depends mainly on the variance recorded in the current month along with data coming from the twelve months preceding its generation. Consequently, we process the VAR dataset as follows. For each brand b, and for each month m, every commentary $C_m^{(j,b)}$ generated for customer j, we compute an input vector $V_m^{(j,b)}$ of 13 items described as follows:

$$V_m^{(j,b)} = \left[v_{m-12}^{(j,b)}, v_{m-11}^{(j,b)}, \ldots, v_{m-1}^{(j,b)}, v_m^{(j,b)}\right] \quad (5)$$

where $v_{m-k}^{(j,b)}$ is the value of the discrepancy recorded for brand b and customer j during month $m-k$.

The POS dataset is also aggregated by month and brand and for each commentary $C_m^{(j,b)}$ generated for customer j, we construct a vector $P_m^{(j,b)}$ as follows:

$$P_m^{(j,b)} = \left[p_{m-12}^{(j,b)}, p_{m-11}^{(j,b)}, \ldots, p_{m-1}^{(j,b)}, p_m^{(j,b)}\right] \quad (6)$$

where $p_{m-k}^{(j,b)}$ is the total value of sales recorded for all products under brand b and sold by customer j during month $m-k$.

We train the models first with the VAR dataset alone, then with both VAR and POS datasets. Each model performs a binary classification where 1 indicates that a commentary needs to be generated. We measure the performance of every model using F1-score defined as the harmonic mean of precision and recall.

We explain every model using SHAP values
First transformation:
Brand, forecast, actual, variance, commentary to Brand, forecast, actual, variance, customer, class, commentary
Commentary generation is done on the ba
(1) A unified apporaoch
(2) Consistent ... Tree ensembles

4 EXPERIMENT

- Binary classification with kNN, SVM, SGB, RF, Single LSTM, bi-LSTM with
  - Variance data alone
  - Variance and POS data
- Balance the dataset - Comparison of F1-score for all classifiers
- Calculation of SHAP values
- Figure of SHAP values for the best classifier
- Comparison with feature importance provided by sklearn (?)
- Permutation importance

-Experiment with multiclass prediction using VAR+POS and VAR (if possible)

Table 4: Binary classification performance with VAR and POS datasets

	kNN	RF	SVM	XGB	1-LSTM	2-LSTM
F1-score	0.53	0.65	0.61	0.64	0.51	0.52

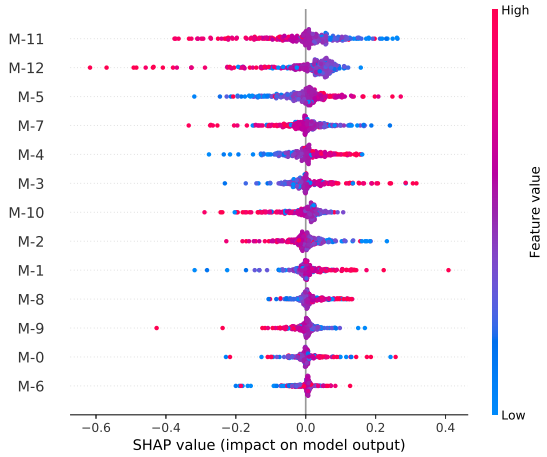

Figure 2: SHAP values plot explaing the SVM classifier prediction

ACKNOWLEDGMENTS

Funding SOSCIP here.

REFERENCES

[1] El Mokhtari K., Maidens J., Bener A. (2019) Predicting Commentaries on a Financial Report with Recurrent Neural Networks. In: Meurs MJ., Rudzicz F. (eds) Advances in Artificial Intelligence. Canadian AI 2019. Lecture Notes in Computer Science, vol 11489.

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "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. ACM. 2016, pp. 1135-1144.

[3] Avanti Shrikumar et al. "Not Just a Black Box: Learning Important Features Through Propagating Activation Differences". In: arXiv preprint arXiv:1605.01713 (2016).

[4] Erik Strumbelj and Igor Kononenko. "Explaining prediction models and individual predictions with feature contributions". In: Knowledge and information systems 41.3 (2014), pp. 647-665.

[5] Anupam Datta, Shayak Sen, and Yair Zick. "'Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems"'. In: Security and Privacy (SP), 2016 IEEE Symposium on. IEEE. 2016, pp. 598-617.

[6] Stan Lipovetsky and Michael Conklin. "Analysis of regression in game theory approach". In: Applied Stochastic Models in Business and Industry 17.4 (2001), pp. 319-330.

[7] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems 30. Curran Associates, Inc., 4768-4777. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[8] Lundberg, S.M., Erion, G.G., Lee, S.I.: Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888 (2018)

Example of appendix