



Machine Learning Interpretability Techniques in Credit Risk Modeling

Craig Peters Sr Dir-Research Model Validation







Craig Peters

Senior Director - Research Model Validation Moody's Analytics

Research Motivation





ML algorithms improve prediction accuracy over traditional statistical models







Research Motivation





ML algorithms are often criticized as black-box models



MOODY'S ANALYTICS

Machine Learning Interpretability Techniques in Credit Risk Modeling 4

Agenda



- Dataset
- Generalized Additive Model (GAM) vs XGBoost (XGB)

2. Global Interpretability

- Feature Importance
 Feature Effect
 Feature Interaction
 Alternate GAM Model

3. Local Interpretability

- Local Interpretable Model-agnostic Explanation (LIME) Shapley value

4. Take-aways & Questions





Problem Setting

A Probability of Default Model

Problem Setting



Dataset

Category	Ratio Name	Ratio Description
Activity	A03	Inventories to Sales
	A08	Current Liabilities to Sales
	A18	Change in Working Capital over Sales
Debt Coverage	DC01**	EBITDA over Interest Expense
Growth	GROW01**	Sales Growth: Sales(t)/Sales(t-1) – 1
	GROW04	Change in ROA
Leverage	LEV12**	Retained Earnings to Current Liabilities
	LEV13**	LT Debt to (LT Debt plus Net worth)
Liquidity	LIQ05**	Cash and Marketable Securities to Total Assets
Profitability	PFT01**	ROA/ Net Income to Total Assets
Size	SIZE01**	Total Assets
Sector	SECTOR	14 Sectors
DUMDEF	PD	Default flag (1=default)

** feature of interest (to be covered later)

** important features (to be covered later)





Problem Setting

Methodology and Results











Global Interpretability SUMMIT 2019 **Feature Importance Most important Permutation Test** features produce the largest difference in AR Change in AR **Permute feature(s) Rank features** feature (j) permuted feat (j) RATIO_CASH_ASSETS RATIO_CHANGEINROA RATIO_SIZE RATIO_SIZE SECTOR 0.9044544 1.0090195 0.8865931 1.1517280 1 Utilities 2 1.1119519 0.8492434 1.0451281 Trade 1.3041865 3 0.5771833 1.0139390 0.9044544 Business_Products 0.9203573 4 1.2912324 0.9081427 0.9086526 0.9086526 Utilities 5 1.3745478 0.9419868 1.0192845 1.3041865 Consumer_Products 6 1.0090195 1.2018472 0.9044544 0.9203573 Services 7 1.0329275 0.8881080 Business_Products 0.9044544 0.9044544 8 0.6441740 1.0731902 0.9805385 0.9044544 Business Products 9 1.2912324 0.8708396 1.0192845 Utilities 1.0451281 10 0.5177905 Business Services 1.0731902 1.1517280 0.9805385

Feature Importance: Permutation Test





- The top 5 important features (LIQ05, DC01, GROW01, ...) are the same
- SIZE01 becomes more important in XGB vs GAM
- More area is covered by bar chart in XGB vs GAM

Feature Effects: Partial Dependence Plot (PDP)



PDP shows the marginal/partial effect of feature(s) on the predicted outcome.



MOODY'S ANALYTICS

Feature Effects: Partial Dependence Plot (PDP)



Analogous behavior (GAM and XGB)



LIQ05, LEV13 were among the top important (also common) features for both GAM and XGB

MOODY'S ANALYTICS

Machine Learning Interpretability Techniques in Credit Risk Modeling 15

Feature Effects: Partial Dependence Plot (PDP)





Non-analogous behavior (GAM vs XGB)

Size01 becomes more important in XGB, and **A03** has higher AR drop from permutation test in XGB

Machine Learning Interpretability Techniques in Credit Risk Modeling 16

Feature Effects: Accumulated Local Effect (ALE)







. . .

MOODY'S ANALYTICS

Machine Learning Interpretability Techniques in Credit Risk Modeling 17

Scatter Plot of DC01 and PFT01, correlation=0.68

0 00 00



Feature Effects: PDP vs. ALE--XGB



PDP and ALE show different effects of average PD changes in response to changes in **PFT01**

Feature Interaction



Friedman's H-statistic



Feature Interaction: Friedman's H-statistic





- All Way: Strong interaction of PFT01, DC01, LIQ05, GROW01 with rest of variables
- Two way: Pairwise SIZE01:PFT01, PFT01:DC01, LIQ05:GROW01... strong interaction observed

Alternate GAM Model





Alternate GAM Model: Non-linearities





MOODY'S ANALYTICS

Alternate GAM Model: Interactions





MOODY'S ANALYTICS

Machine Learning Interpretability Techniques in Credit Risk Modeling 23



Local Interpretability

Local Interpretability

Local Interpretable Model-agnostic Explanations (LIME)

- Simulate points near specific observation
- Generate model predictions at these points
- **Use** model predictions as Y variable
- Weight new observations by proximity
- **Build** weighted linear regression (or other interpretable model)
- Interpret the local surrogate model

Advantage:



• Easy to interpret

Disadvantage:

- Simulating "good" nearby points
- Unstable results





Local Interpretability LIME Example



Firm Profile

Ratio	Value
A03	0.85
A08	1.04
A18	1.29
DC01	2.69
GROW01	0.95
GROW04	1.12
LEV12	1.86
LEV13	0.63
LIQ05	1.77
PFT01	2.40
SIZE01	1.05
SECTOR	Business
	Services



Local Interpretability

Shapley Value





Originally from game theory to attribute **the value of a team** effort to individual members



Unlike LIME, uses the **same original model** in a local space.



Explains:

- Individual vs. Average PD
- Feature contribution towards the difference

Local Interpretability

Shapley Value Example



Ratio	Value
A03	0.85
A08	1.04
A18	1.29
DC01	2.69
GROW01	0.95
GROW04	1.12
LEV12	1.86
LEV13	0.63
LIQ05	1.77
PFT01	2.40
SIZE01	1.05
SECTOR	Business
	Services







Take-aways

Take-aways





Interpretability techniques can help explain and predict black box model output



Model-Agnostic methods can be applied to any model enabling a broader range of methodologies



Interpretability techniques can help make today's black boxes tomorrow's interpretable models



The tradeoff between interpretability and accuracy is real and can only be mitigated

References



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

Partial Dependency Plot (PDP)

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Accumulated Local Effect (ALE)

Apley, D. W. (2016). Visualizing the effects of predictor variables in black box supervised learning models. *arXiv preprint arXiv:1612.08468*.

Friedman's H-statistics

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, *2*(3), 916-954.

Local interpretable model-agnostic explanations (LIME)

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). 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* (pp. 1135-1144). ACM.

Machine Learning Interpretability Techniques in Credit Risk Modeling 32



Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, *41*(3), 647-665.

Lundberg, S., & Lee, S. I. (2016). An unexpected unity among methods for interpreting model predictions. *arXiv preprint arXiv:1611.07478*.

Machine Learning Interpretability with H2O Driverless AI (K-LIME)

Patrick Hall, Navdeep Gill, Megan Kurka, & Wen Phan, Edited by: Angela Bartz

Discussions of Machine Learning Interpretation framework and taxonomy

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv* preprint arXiv:1606.03490.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 80-89). IEEE.

Molnar, C. (2018). Interpretable machine learning: A guide for making black box models explainable. *E-book at< https://christophm. github. io/interpretable-ml-book/>, version dated, 10.*



Appendix

Appendix

Partial Dependence Plots

• The partial dependence function is defined as:



 $\hat{f}_{x_S}(x_S) = E_{x_C} \left[\hat{f}(x_S, x_C) \right] = \int \hat{f}(x_S, x_C) d\mathbb{P}(x_C)$ integrate over all x_C x_C other features used in the model feature(s) of interest

Accumulated Local Effects

$$\hat{f}_{x_{S},ALE}(x_{S}) = \int_{z_{0,1}}^{x_{S}} E_{X_{C}|X_{S}} \left[\hat{f}^{S}(X_{s}, X_{c}) | X_{S} = z_{S} \right] dz_{S} = \int_{z_{0,1}}^{x_{S}} \int_{x_{C}} \hat{f}^{S}(z_{s}, x_{c}) \mathbb{P}(x_{C}|z_{S}) dx_{C} dz_{S}$$
where: $\hat{f}^{S}(x_{s}, x_{c}) = \frac{\delta \hat{f}(x_{S}, x_{C})}{\delta x_{S}}$
Differential/change in PD

MOODY'S ANALYTICS

Machine Learning Interpretability Techniques in Credit Risk Modeling 34

Appendix H-statistic



• Two Way:



• All Way:

$$H_{j}^{2} = \sum_{i=1}^{n} \left[\hat{f}(x^{(i)}) - PD_{j}(x_{j}^{(i)}) - PD_{-j}(x_{-j}^{(i)}) \right]^{2} / \sum_{i=1}^{n} \hat{f}^{2}(x^{(i)})$$
1-dim PDPs
(n-1) dim PDP
prediction

MOODY'S ANALYTICS

Machine Learning Interpretability Techniques in Credit Risk Modeling 35

MOODY'S

Craig Peters Sr Dir-Resrch Model Validation Craig.Peters@moodys.com © 2019 Moody's Corporation, Moody's Investors Service, Inc., Moody's Analytics, Inc. and/or their licensors and affiliates (collectively, "MOODY'S"). All rights reserved.

CREDIT RATINGS ISSUED BY MOODY'S INVESTORS SERVICE, INC. AND ITS RATINGS AFFILIATES ("MIS") ARE MOODY'S CURRENT OPINIONS OF THE RELATIVE FUTURE CREDIT RISK OF ENTITIES, CREDIT COMMITMENTS, OR DEBT OR DEBT-LIKE SECURITIES, AND MOODY'S PUBLICATIONS MAY INCLUDE MOODY'S CURRENT OPINIONS OF THE RELATIVE FUTURE CREDIT RISK OF ENTITIES, CREDIT COMMITMENTS, OR DEBT OR DEBT-LIKE SECURITIES. MOODY'S DEFINES CREDIT RISK AS THE RISK THAT AN ENTITY MAY NOT MEET ITS CONTRACTUAL, FINANCIAL OBLIGATIONS AS THEY COME DUE AND ANY ESTIMATED FINANCIAL LOSS IN THE EVENT OF DEFAULT, CREDIT RATINGS DO NOT ADDRESS ANY OTHER RISK, INCLUDING BUT NOT LIMITED TO: LIQUIDITY RISK, MARKET VALUE RISK, OR PRICE VOLATILITY. CREDIT RATINGS AND MOODY'S OPINIONS INCLUDED IN MOODY'S PUBLICATIONS ARE NOT STATEMENTS OF CURRENT OR HISTORICAL FACT. MOODY'S PUBLICATIONS MAY ALSO INCLUDE QUANTITATIVE MODEL-BASED ESTIMATES OF CREDIT RISK AND RELATED OPINIONS OR COMMENTARY PUBLISHED BY MOODY'S ANALYTICS, INC. CREDIT RATINGS AND MOODY'S PUBLICATIONS DO NOT CONSTITUTE OR PROVIDE INVESTMENT OR FINANCIAL ADVICE, AND CREDIT RATINGS AND MOODY'S PUBLICATIONS ARE NOT AND DO NOT PROVIDE RECOMMENDATIONS TO PURCHASE, SELL, OR HOLD PARTICULAR SECURITIES, NEITHER CREDIT RATINGS NOR MOODY'S PUBLICATIONS COMMENT ON THE SUITABILITY OF AN INVESTMENT FOR ANY PARTICULAR INVESTOR. MOODY'S ISSUES ITS CREDIT RATINGS AND PUBLISHES MOODY'S PUBLICATIONS WITH THE EXPECTATION AND UNDERSTANDING THAT EACH INVESTOR WILL, WITH DUE CARE, MAKE ITS OWN STUDY AND EVALUATION OF EACH SECURITY THAT IS UNDER CONSIDERATION FOR PURCHASE, HOLDING, OR SALE

MOODY'S CREDIT RATINGS AND MOODY'S PUBLICATIONS ARE NOT INTENDED FOR USE BY RETAIL INVESTORS AND IT WOULD BE RECKLESS AND INAPPROPRIATE FOR RETAIL INVESTORS TO USE MOODY'S CREDIT RATINGS OR MOODY'S PUBLICATIONS WHEN MAKING AN INVESTMENT DECISION. IF IN DOUBT YOU SHOULD CONTACT YOUR FINANCIAL OR OTHER PROFESSIONAL ADVISER.

ALL INFORMATION CONTAINED HEREIN IS PROTECTED BY LAW, INCLUDING BUT NOT LIMITED TO, COPYRIGHT LAW, AND NONE OF SUCH INFORMATION MAY BE COPIED OR OTHERWISE REPRODUCED, REPACKAGED, FURTHER TRANSMITTED, TRANSFERRED, DISSEMINATED, REDISTRIBUTED OR RESOLD, OR STORED FOR SUBSEQUENT USE FOR ANY SUCH PURPOSE, IN WHOLE OR IN PART, IN ANY FORM OR MANNER OR BY ANY MEANS WHATSOEVER, BY ANY PERSON WITHOUT MOODY'S PRIOR WRITTEN CONSENT.

All information contained herein is obtained by MOODY'S from sources believed by it to be accurate and reliable. Because of the possibility of human or mechanical error as well as other factors, however, all information contained herein is provided "AS IS" without warranty of any kind. MOODY'S adopts all necessary measures so that the information it uses in assigning a credit rating is of sufficient quality and from sources MOODY'S considers to be reliable including, when appropriate, independent third-party sources. However, MOODY'S is not an auditor and cannot in every instance independently verify or validate information received in the rating process or in preparing the Moody's publications.

To the extent permitted by law, MOODY'S and its directors, officers, employees, agents, representatives, licensors and suppliers disclaim liability to any person or entity for any indirect, special, consequential, or incidental losses or damages whatsoever arising from or in connection with the information contained herein or the use of or inability to use any such information, even if MOODY'S or any of its directors, officers, employees, agents, representatives, licensors or suppliers is advised in advance of the possibility of such losses or damages, including but not limited to: (a) any loss of present or prospective profits or (b) any loss or damage arising where the relevant financial instrument is not the subject of a particular credit rating assigned by MOODY'S.

To the extent permitted by law, MOODY'S and its directors, officers, employees, agents, representatives, licensors and suppliers disclaim liability for any direct or compensatory losses or damages caused to any person or entity, including but not limited to by any negligence (but excluding fraud, willful misconduct or any other type of liability that, for the avoidance of doubt, by law cannot be excluded) on the part of, or any contingency within or beyond the control of, MOODY'S or any of its directors, officers, employees, agents, representatives, licensors or suppliers, arising from or in connection with the information contained herein or the use of or inability to use any such information.



Moody's Investors Service, Inc., a wholly-owned credit rating agency subsidiary of Moody's Corporation ("MCO"), hereby discloses that most issuers of debt securities (including corporate and municipal bonds, debentures, notes and commercial paper) and preferred stock rated by Moody's Investors Service, Inc. have, prior to assignment of any rating, agreed to pay to Moody's Investors Service, Inc. for appraisal and rating services rendered by it fees ranging from \$1,500 to approximately \$2,500,000. MCO and MIS also maintain policies and procedures to address the independence of MIS's ratings and rating processes. Information regarding certain affiliations that may exist between directors of MCO and rated entities, and between entities who hold ratings from MIS and have also publicly reported to the SEC an ownership interest in MCO of more than 5%, is posted annually at www.moodys.com under the heading "Investor Relations — Corporate Governance — Director and Shareholder Affiliation Policy."

Additional terms for Australia only: Any publication into Australia of this document is pursuant to the Australian Financial Services License of MOODY'S affiliate, Moody's Investors Service Pty Limited ABN 61 003 399 657AFSL 336969 and/or Moody's Analytics Australia Pty Ltd ABN 94 105 136 972 AFSL 383569 (as applicable). This document is intended to be provided only to "wholesale clients" within the meaning of section 761G of the Corporations Act 2001. By continuing to access this document from within Australia, you represent to MOODY'S that you are, or are accessing the document as a representative of, a "wholesale client" and that neither you nor the entity you represent will directly or indirectly disseminate this document or its contents to "retail clients" within the meaning of section 761G of the Corporations Act 2001. MOODY'S credit rating is an opinion as to the creditworthiness of a debt obligation of the issuer, not on the equity securities of the issuer or any form of security that is available to retail investors. It would be reckless and inappropriate for retail investors to use MOODY'S credit ratings or publications when making an investment decision. If in doubt you should contact your financial or other professional adviser.

Additional terms for Japan only: Moody's Japan K.K. ("MJKK") is a wholly-owned credit rating agency subsidiary of Moody's Group Japan G.K., which is wholly-owned by Moody's Overseas Holdings Inc., a wholly-owned subsidiary of MCO. Moody's SF Japan K.K. ("MSFJ") is a wholly-owned credit rating agency subsidiary of MJKK. MSFJ is not a Nationally Recognized Statistical Rating Organization ("NRSRO"). Therefore, credit ratings assigned by MSFJ are Non-NRSRO Credit Ratings. Non-NRSRO Credit Ratings are assigned by an entity that is not a NRSRO and, consequently, the rated obligation will not qualify for certain types of treatment under U.S. laws. MJKK and MSFJ are credit rating agencies registered with the

Japan Financial Services Agency and their registration numbers are FSA Commissioner (Ratings) No. 2 and 3 respectively.

MJKK or MSFJ (as applicable) hereby disclose that most issuers of debt securities (including corporate and municipal bonds, debentures, notes and commercial paper) and preferred stock rated by MJKK or MSFJ (as applicable) have, prior to assignment of any rating, agreed to pay to MJKK or MSFJ (as applicable) for appraisal and rating services rendered by it fees ranging from JPY200,000 to approximately JPY350,000,000.

MJKK and MSFJ also maintain policies and procedures to address Japanese regulatory requirements.



MOODY'S ANALYTICS