

An Empirical Comparison of Model-Agnostic Techniques for Defect Prediction Models

Gichan Lee and Scott Uk-Jin Lee
*Department of Coumputer Science and
Engineering
Hanyang University
Ansan, Republic of Korea
{fantasyopy, scottlelee}@hanyang.ac.kr*

発表者：坂本

欠陥予測

| 欠陥予測

- ソフトウェア品質を高めるために混入し得る欠陥を予測
- 実務ではモデルの出力が説明可能であることが重要^[1,2]
 - 欠陥予測モデルがブラックボックス化に対処したい

| モデル診断技術

- LIME^[3], SHAP^[4], BreakDown^[5] など
 - XAI（説明可能なAI）領域から派生したモデル診断技術
- 予測に寄与した特徴量を含む，局所的説明を提供可能

[1] H. K. Dam, T. Tran, A. Ghose, “Explainable software analytics,” in Proceedings of the 40th international conference on software engineering: New ideas and Emerging results, 2018, pp. 53–56.

[2] C. Tantithamthavorn, J. Jiarapakdee, and J. Grundy, “Explainable AI for Software Engineering,” arXiv preprint arXiv:2012.01614, 2020.

[3] M.T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?: Explaining the Predictions of Any Classifier,” in Proceedings of the International Conference on Knowledge Discovery & Data Mining (KDD), 2016, pp. 1135–1144.

[4] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in Advances in Neural Information Processing Systems (NIPS), 2017, pp. 4765–4774.

[5] M.StaniakandP.Biecek, “Explanationsofmodelpredictionswithlive and breakDown packages,” arXiv preprint arXiv:1804.01955, 2018.

局所的説明

| 特定の入力に対するモデルの予測の根拠を提示することでモデルを説明する方法

- 各特徴量がどのように予測に寄与しているか
- 大局的説明：複雑なモデルを可読性の高いモデルで表現することで説明する方法

| 局所的説明の必要性

- 実務では説明可能性が重要^[1,2]
- 欠陥予測に対して実務者が取るべき行動のヒントが得られる

モデル診断手法

| モデル診断手法

- LIME, SHAP, BreakDownなどが存在
- 局所的説明の特徴を分析

| 異なるモデル診断手法の違いは？

- 局所的説明に違いがあれば実務者は混乱
- それぞれの手法の信頼性にも関わる

目的

それぞれのモデル診断手法の局所的説明の不一致を特定する

リサーチクエスチョン

| RQ1：局所的説明の違いによって上位にランクされる特徴量の寄与度はどの程度違うか

- モデル診断手法によって生成された局所的説明の情報の違いを分析
- 既存研究では全ての特徴量を考慮して各手法を比較

| RQ2：局所的説明の違いによる特徴の寄与度のランキングや貢献度の方向性はどの程度違うか

- 上位特徴量の寄与度を局所的説明の類似度測定の際に考慮
- 局所的説明の中の上位特徴量が寄与する方向

分析方法

RQ1

| 寄与度

- 実務者が特に注目する上位特徴量の平均寄与度を算出

RQ2

| 寄与度ランキングの比較 $W_{RBO}(1:d) = 1 - p^{d-1} + \frac{1-p}{p} \cdot d \cdot \left(\ln \frac{1}{1-p} - \sum_{i=1}^{d-1} \frac{p^i}{i} \right) (1)$

- RBO (Rank Biased Overlap) を使用
 - 不特定多数のランク付されたリスト間の類似性測定

| 上位特徴量の寄与方向

- ex) LIMEとSHAPの上位特徴量の違いを比較, 不一致を特定

データセット

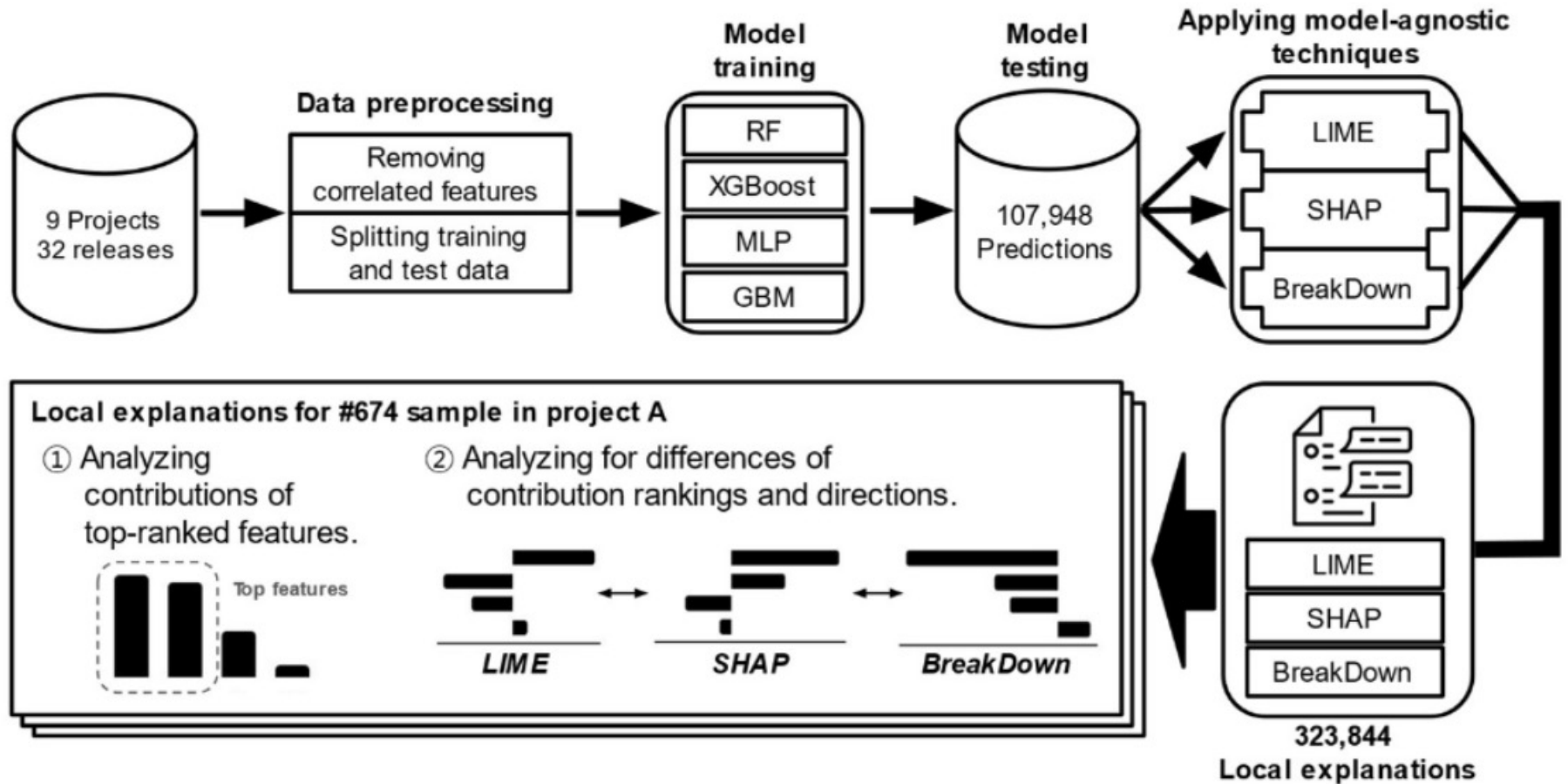
| Yatishら^[6]のデータセット

- 9つのApache OSSシステムの32リリースから構成
- AutoSpearman関数を適用し、相関のある特徴量を除去
- 標本外ブートストラップサンプリングで学習、テストに分割

<i>Project Name</i>	<i>Description</i>	<i>Studied Releases</i>	<i># of features</i>	<i># of training samples</i>	<i># of testing samples</i>
ActiveMQ	Messaging and Integration Patterns server	5.0.0, 5.1.0, 5.2.0, 5.3.0, 5.8.0	26-28	1884-3420	691-1258
Camel	Enterprise Integration Framework	1.4.0, 2.9.0, 2.10.0, 2.11.0	25-26	1515-8846	554-3293
Derby	Relational Database	10.2.1.6, 10.3.1.4, 10.5.1.1	28	1963-2705	723-995
Groovy	Java-syntax-compatible OOP for JAVA	1.5.7, 1.6.0.Beta_1, 1.6.0.Beta_2	26-28	757-884	274-319
HBase	Distributed Scalable Data Store	0.94.0, 0.95.0, 0.95.2	25-27	1059-1834	380-674
Hive	Data Warehouse System for Hadoop	0.9.0, 0.10.0, 0.12.0	24-27	1416-2662	514-981
JRuby	Ruby Programming Lang for JVM	1.1, 1.4, 1.5, 1.7	29-31	731-1614	263-595
Lucene	Text Search Engine Library	2.3.0, 2.9.0, 3.0.0, 3.1.0	27-31	805-2806	292-1036
Wicket	Web Application Framework	1.3.0.beta1, 1.3.0.beta2, 1.5.3	27	1672-2578	610-953

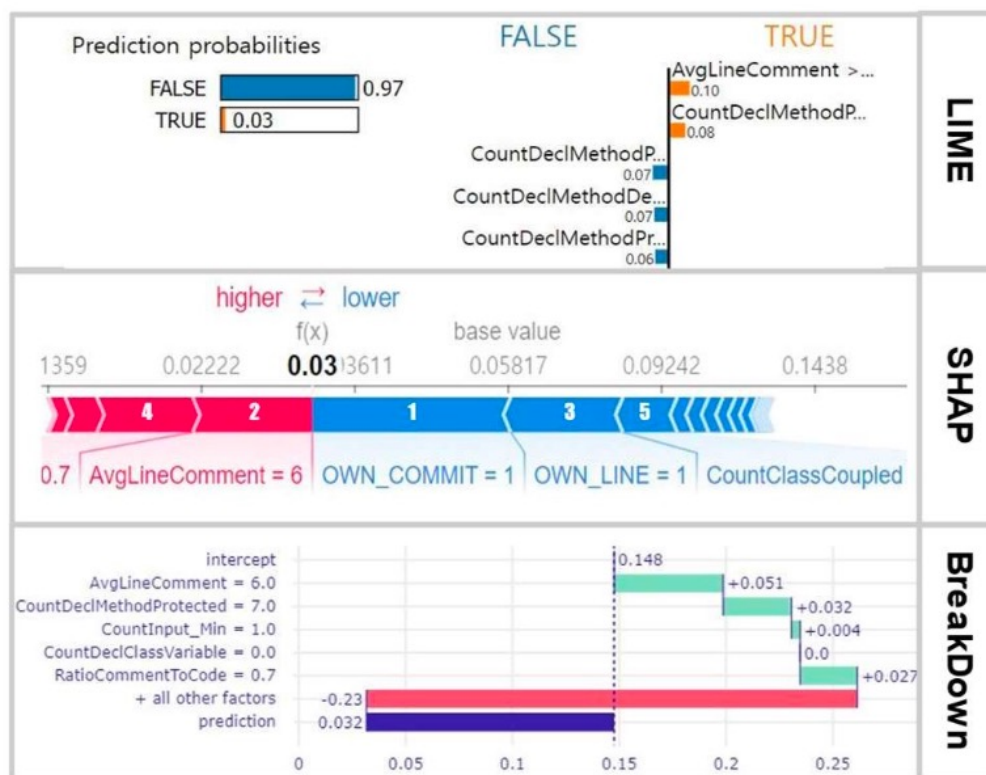
[6] S. Yatish, J. Jiarpakdee, P. Thongtanunam, and C. Tantithamthavorn, “Mining Software Defects: Should We Consider Affected Releases?” in Proceedings of the International Conference on Software Engineering (ICSE), 2019, pp. 654–665.

手法概要



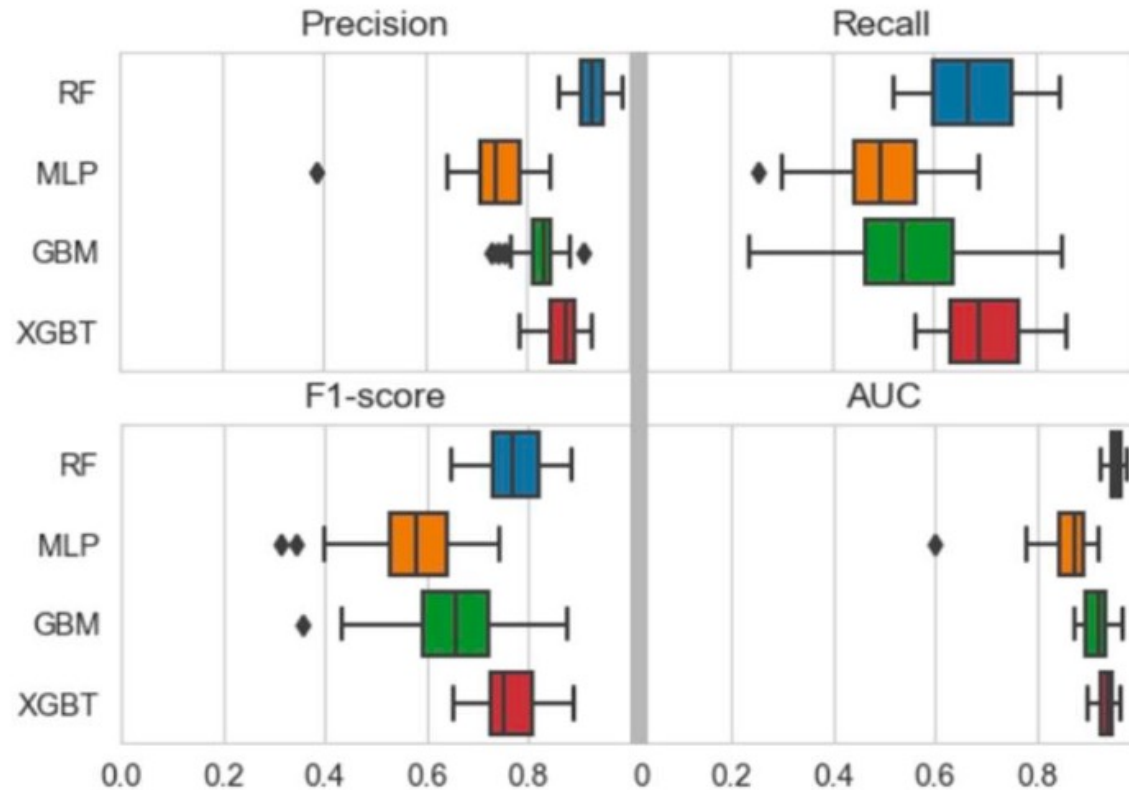
使用するモデル診断技術

| LIME, SHAP, BreakDownを使用



Model-agnostic	Top 5 features	Raw values (%)
LIME	AvgLineComment	0.097 (12%)
	CountDeclMethodProtected	0.077 (9%)
	CountDeclMethodPublic	-0.073 (-9%)
	CountDeclMethodDefault	-0.066 (-7%)
	CountDeclMethodPrivate	-0.062 (-7%)
SHAP	OWN_COMMIT	-0.647 (-27%)
	AvgLineComment	0.389 (16%)
	OWN_LINE	-0.365 (-15%)
	RatioCommentToCode	0.312 (13%)
	CountClassCoupled	-0.177 (-7%)
BreakDown	OWN_COMMIT	-0.120 (-34%)
	AvgLineComment	0.050 (14%)
	OWN_LINE	-0.035 (-10%)
	CountDeclMethodProtected	0.032 (9%)
	RatioCommentToCode	0.027 (7%)

欠陥予測モデルの性能



中央値

Precision : 0.74 - 0.93

F1-score : 0.58 - 0.77

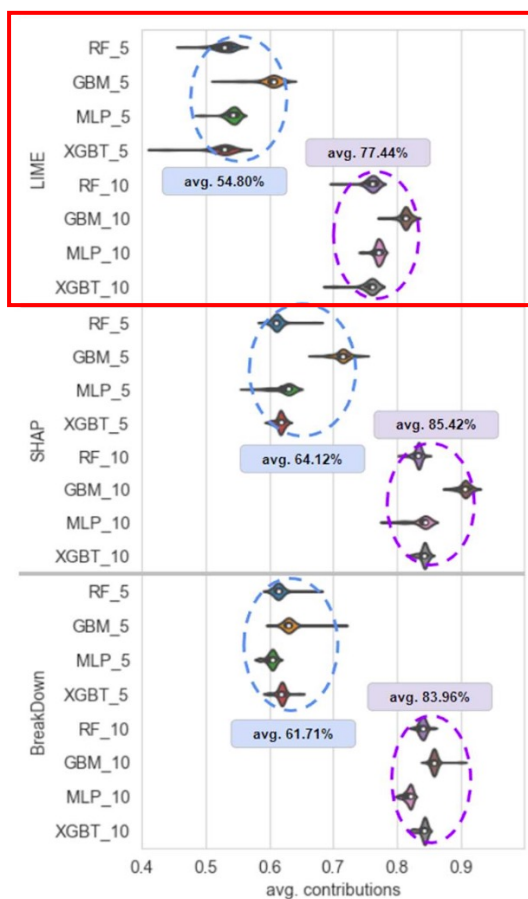
Recall : 0.50 - 0.69

AUC : 0.87 - 0.96

RQ1

特徴量ランキングの上位5位と上位10位を考慮

– リリースごとに平均寄与度を算出



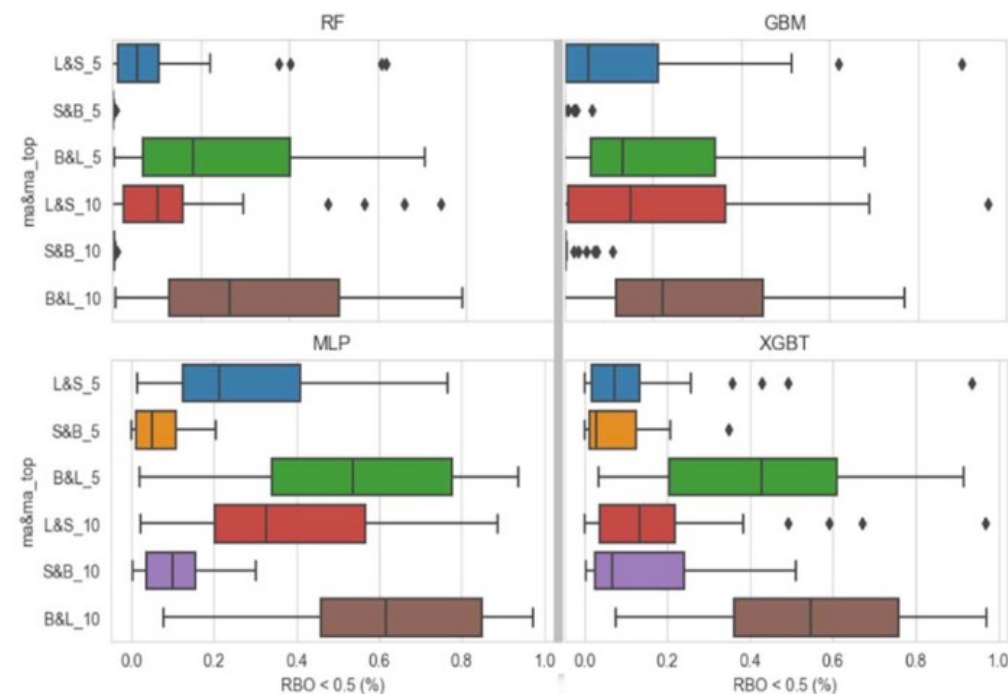
寄与度

– LIMEは上位特徴量の平均寄与度が低い

RQ2

類似度

- 3つのモデル診断手法の全てのペアの組み合わせの局所的説明の寄与度ランキングのうち、 $RBO < 0.5$ の割合で測定
 - 1ならランキングが全く違う



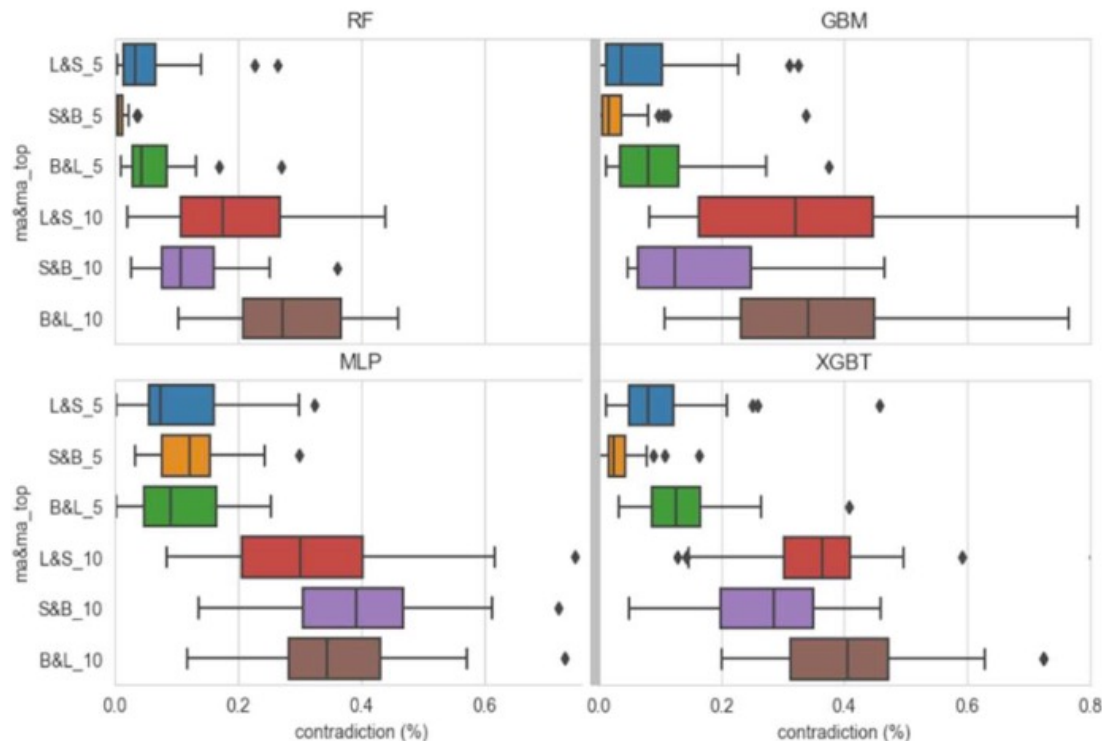
RFのLIME&BreakDownでは、約18%で説明が異なる

手法によって説明が違うことは多め
→特にLIME

RQ2

| 寄与の方向性

- 異なるモデル診断手法の貢献度ランキングに含まれる特徴量を比較後、矛盾が生じる比率
 - どれだけ特徴量が異なるか



LIMEは他の手法と矛盾が存在

まとめ

| LIMEの上位特徴量の平均寄与度が低い

- 上位以外の特徴量も考慮する必要がある可能性

| 特徴量の寄与度ランキングが異なる局所的説明が生成

- LIMEとその他との組み合わせがで $RBO < 0.5$ が多い
- 実務者は欠陥予測に向けて異なる行動をとる可能性がある

| 寄与度の矛盾

- 各手法が寄与方向が異なる局所的説明を生成
- 選択する手法によって欠陥予測に対して様々な行動をとる

複数のモデル診断手法を用いて保守活動を行うことが大事