Preliminary Study Report

YINGZHE LYU, Queen's University, Canada

ACM Reference Format:
Yingzhe Lyu. 2020. Preliminary Study Report. 1, 1 (November 2020), 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 EXPERIMENTS ON THE MODEL INTERPRETATION IN TIME PERIODS

1.1 Models used in prior works

In this preliminary study, we intend to test all the available models from prior works. In Table 1, we conclude the papers predict job failures on the Google cluster data and disk failures on the Backblaze data, the models they used, and the available model implementations in Python.

Table 1. ML models used in prior works.

Model	Type	Google Ref.	Backblaze Ref.	Implementation
LDA [1]	Whitebox	[5]		sklearn [5]
QDA	Whitebox	[5]		sklearn [5]
LR	Whitebox	[5]	[1, 3]	sklearn [5]
DT	Whitebox		[1, 3]	sklearn [5]
GBDT	Blackbox		[1]	XGBoost [6]
NN	Blackbox	[4] [3]	[3] [4]	sklearn [5]
RF	Blackbox	[2]	[1, 3]	sklearn [5]
SVM	Blackbox		[1, 3]	sklearn [5]
RGF	Blackbox		[1]	rgf-python [7]
FastTree [2]	Blackbox		[6]	microsoftml [8]

[1] Abbreviation for model names: LDA (Linear Discriminant Analysis), QDA (Quadratic Discriminant Analysis), LR (Logistic Regression), DT (Decision Tree), GBDT (Gradient Boosting Decision Trees), NN (multi-layer Neural Network), RF (Random Forest), SVM (Support Vector Machine), RGF (Regularized Greedy Forests).
[2] FastTree is a form of "Multiple Additive Regression Trees" (MART) gradient boosting algorithm.
[3] Rosa et al. [4] use a 2-level multilayer NN, which comprises two one-hidden-layer perceptrons.
[4] Mahdisoltani et al. [3] use a one-hidden-layer perceptron, with 100 nodes in the hidden layer.
[5] https://scikit-learn.org/stable
[6] https://xgboost.readthedocs.io/en/latest/
[7] https://github.com/RGF-team/rgf/tree/master/python-package
[8] https://docs.microsoft.com/en-us/machine-learning-server/python-reference/microsoftml/rx-fast-trees

Author's address: Yingzhe Lyu, ylyu@cs.queensu.ca, Queen's University, Software Analysis and Intelligence Lab (SAIL), Kingston, ON, Canada.

2020. XXXX-XXXX/2020/11-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1.2 Experimental design for obtaining feature importance

In this preliminary study, we mainly reuse the experimental design from our earlier paper draft *When and How Should We Maintain AIOps Models Under Data Evolution? An Exploratory Study*, Section 4. In brief, we carry out our experiments on model performance and feature importance as follows.

(1) We first partition the data into multiple time periods according to their timestamps. For the Google cluster trace data, we partition the entire 28-day trace data into 28 one-day time periods; for the Backblaze disk stats data, we partition the entire 3-year monitoring data into 36 one-month time periods.
(2) After separating the datasets into time periods, we build a model using samples from each period and calculate its feature importance.
(3) To obtain the performance of each individual model, we also test each model on the next time period to evaluate their performance, except for the model built with data from the last period as there are no future testing data available. Therefore, we have performance evaluated for the first 35 out of 36 periods on the Backblaze data and 27 out of 28 periods on the Google data.

Besides, we mainly have the following differences from the earlier paper.

(1) We now have a different set of models (i.e., LDA, QDA, LR, CART, GBDT, NN, RF, and RGF), which are all the available models in prior works. However, we have excluded two models: FastTree and SVM. In the paper that uses FastTree [6], they define the problem as a learning-to-rank regression problem and treat the top-ranked samples as failures. Their approach is different from other papers, which define the problem as a classification problem and use classifier models to predict failures. Therefore, we do not use the FastTree algorithm. Also, we choose not use the SVM model here, which could take quite a long time to execute (e.g., a few weeks for the Backblaze data).
(2) We use permutation feature importance than the `feature_importance_` variable natively available in RF and CART model for generalizability.
(3) Every time we built a model, we now use `RandomSearchCV` to tune the model hyperparameters on the training set first.
(4) We now make the models persistent on disk.

1.3 Performance Comparison

Figure 1 compares the AUC performance of different models on each dataset on each time period. We also have the separate AUC performance plot for each model available, but considering the volume (i.e., a total of 16 plots for 8 models on 2 datasets), we put it in the Appendix (Figure 6).

On the Google dataset, we observe that the model performance can be distinctively categorized into different performance, while on the Backblaze data, the similarity is not very prominent. We carry out a statistic test to group the models by their performance in the following sections.

In addition, we find that the LDA, QDA, and LR models have limited predictive ability, especially on specific periods (e.g., period 9 and period 19). The poor performance on the linear models could be that the feature space of the Google data cannot be linearly separated, thus requiring more complex models.

1.4 Grouping the models by their AUC performance

In order to investigate how consistent the interpretation is on the similarity and differently performed models, we first group the performance into different groups using a double Scott-Knott test. We carry out the experiment as follows.

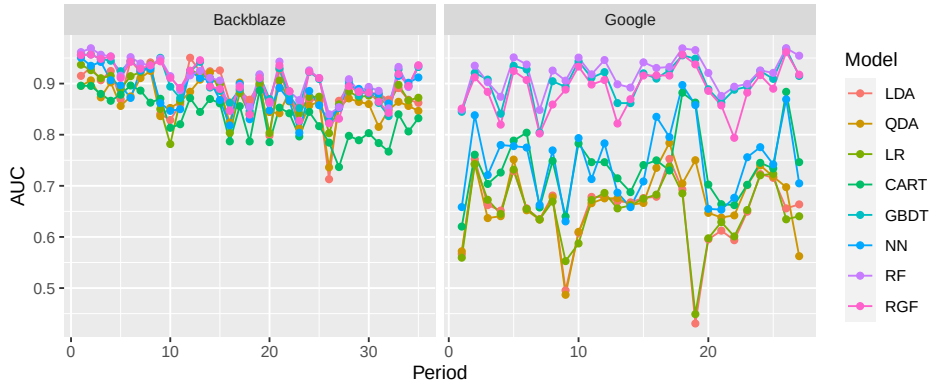
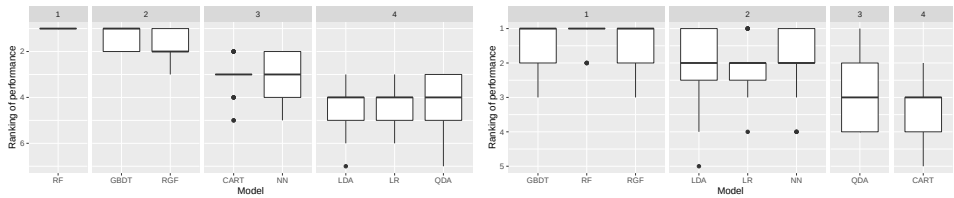


Fig. 1. Comparison of the AUC performance before and after the redundancy and correlation test.

(1) We first group the AUC performance of each model into different rankings using the Scott-Knott test on each time period with observations from 10 runs.
(2) We then group the performance rankings of each model using the Scott-Knott test again, with the observations on each time period.
(3) Finally, the models in the same group of the second Scott-Knott test would have similar performance, while the models in different groups would have different performance.

Figure 2 shows the results of our double Scott-Knott test. We also analyze on how many periods the models have AUC performance in the same group, as shown in Figure 3. The results fit our double Scott-Knott grouping results (Figure 2).



(a) Google (b) Backblaze

Fig. 2. Double Scott-Knott of grouping the models by their ranking of AUC performance in time periods.

1.5 Consistency of the feature importance from different models

This section investigates how consistent the feature importance (i.e., model interpretability) is in different groups. Specifically, we calculate the consistency by measuring the proportion that two models have the same features in their top 1, top 3, and top 5 ranked features. We name such value as **overlapping value** and have experimented as follow.

(1) On each time period, we first rank the feature importance separately with a Scott-Knott test on the two models we choose.
(2) For the top N features (N=1, 3, or 5) on the two models, we calculate the **overlapping value** as the number of their intersection divided by their union. For example, if the top 1 ranked features on the first models are $F1, F2$ while the top 1 ranked features on the other model

(a) Google (b) Backblaze

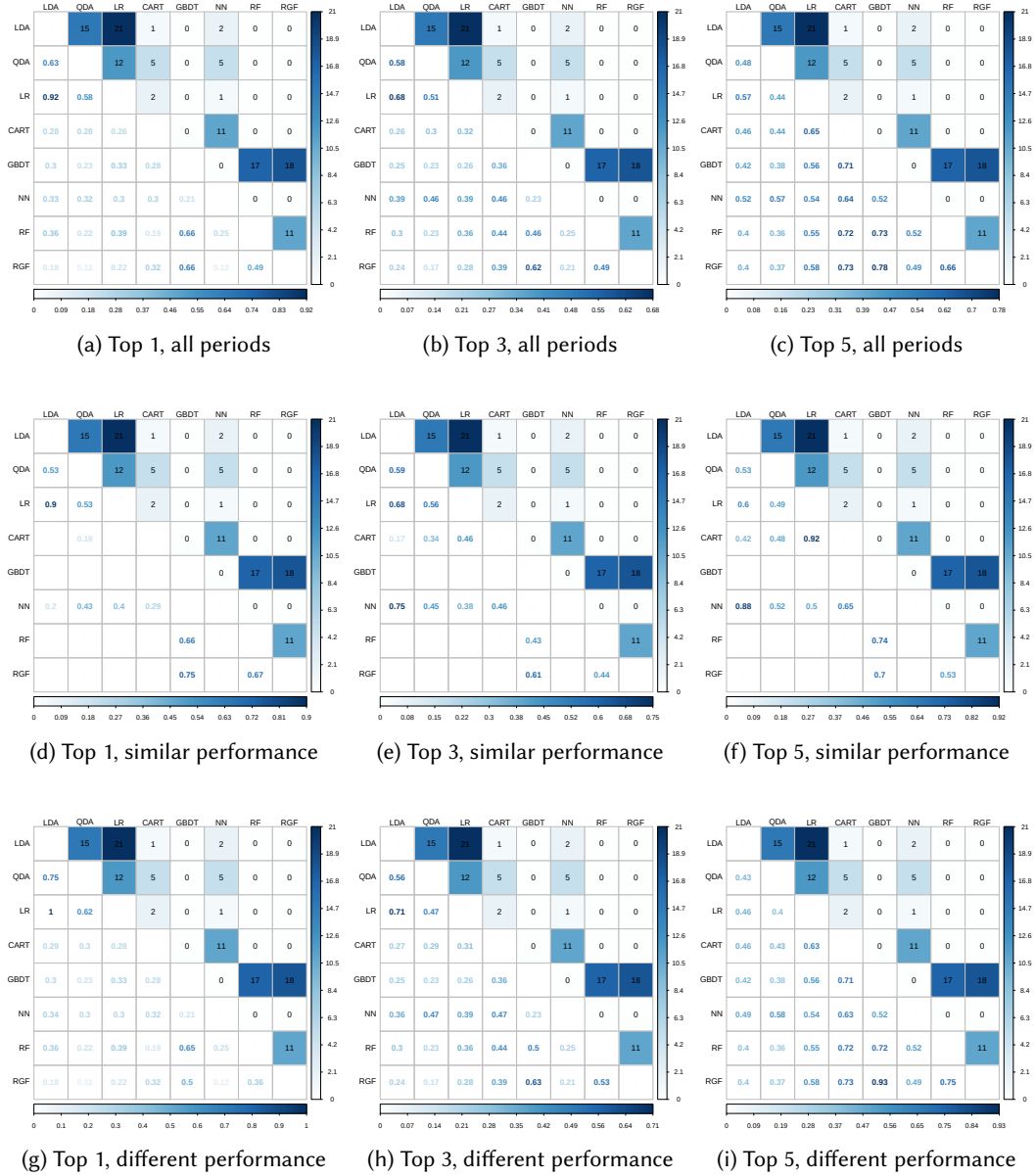
Fig. 3. Number of periods that two models have a same ranking of the AUC performance.

area $F1, F4, F6$, then the top 1 overlap value would be $1/4 = 0.25$. The overlapping value range is in $[0, 1]$, and the bigger the number is, the higher the overlap is.

(3) Finally, we calculate the mean overlapping value on all periods for each model as their final overlapping value. We calculate the overlapping in three scenarios: 1) on all time periods; 2) on periods only when two models have similar performance (i.e., in the same SK group); 3) and on periods only when two models have different performance (i.e., in the different SK groups).

The overlapping of feature importance is positively correlated with the similarity of model performance. Figure 4 and Figure 5 shows our results on the overlapping values. We observe that the overlapping values are positively correlated with the similarity of performance (i.e., the number of periods two models have performance in the same rank and the grouping of model performance). For example, on the Google dataset, the RGF and GBDT models are in the same performance group on 17 out of 27 periods, and their overlapping values are 0.66, 0.62, and 0.78 for the top 1, top 3, and top 5 ranked features, way above the average overlapping value of 0.35, 0.36, and 0.55. In contrast, the RGF and LDA models share no same grouping across all time periods, and their overlapping values are only 0.18, 0.24, and 0.4 for the top 1, top 3, and top 5 ranked features, respectively.

The models tend to have bigger overlapping when they have similar performance. We also observe that the overlapping values are bigger when on time periods where they have a similar performance. On the Google dataset, the mean top 1, top 3, and top 5 overlapping values across all the time periods are 0.35, 0.36, and 0.55, while the same values on periods that two models have the performance in the same group are 0.43, 0.49, and 0.61, and the same values on periods that two models have different performance groupings are 0.35, 0.36, and 0.54, respectively. Similarly, on the Backblaze dataset, the mean top 1, top 3, and top 5 overlapping values on all the time periods are 0.82, 0.47, and 0.49, while the same values on periods that two models have the performance in the same group are 0.90, 0.45, and 0.49, and the same values on periods that two models have different performance groupings are 0.79, 0.47, and 0.50, respectively.

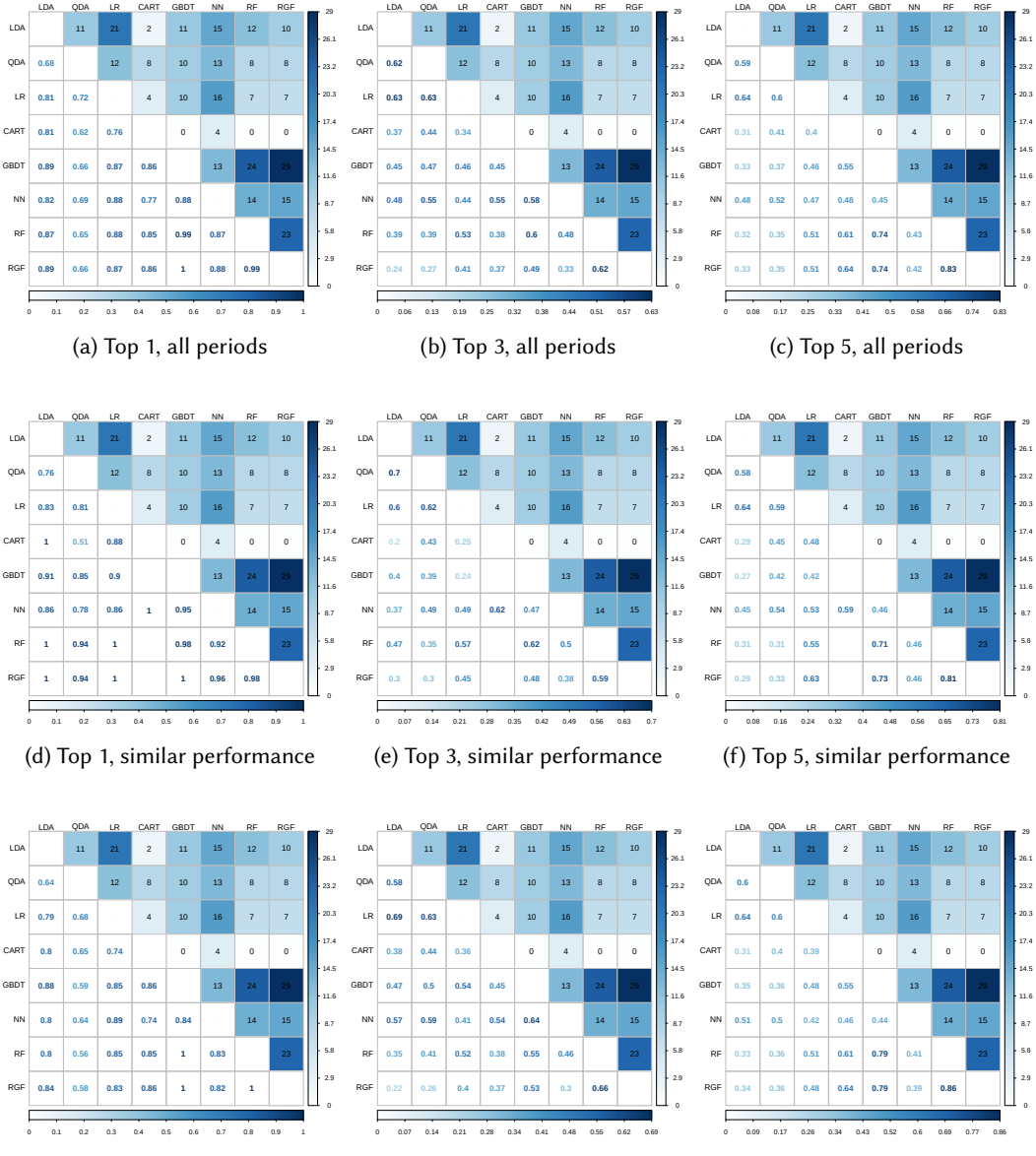


(a) Top 1, all periods (b) Top 3, all periods (c) Top 5, all periods

(d) Top 1, similar performance (e) Top 3, similar performance (f) Top 5, similar performance

(g) Top 1, different performance (h) Top 3, different performance (i) Top 5, different performance

Fig. 4. The overlapping values between two models on all periods, periods they have the same ranking, and periods they have different rankings on the Google dataset. The number in each cell in the upper triangle indicates how many periods two models have AUC performance in the same group. The number in each cell in the lower triangle indicates the mean value of the two models' overlapping values on time periods described in the subfigure caption.



(a) Top 1, all periods (b) Top 3, all periods (c) Top 5, all periods

(d) Top 1, similar performance (e) Top 3, similar performance (f) Top 5, similar performance

(g) Top 1, different performance (h) Top 3, different performance (i) Top 5, different performance

Fig. 5. The overlapping values between two models on all periods, periods they have the same ranking, and periods they have different rankings on the Backblaze dataset. The number in each cell in the upper triangle indicates how many periods two models have AUC performance in the same group. The number in each cell in the lower triangle indicates the mean value of the two models' overlapping values on time periods described in the subfigure caption.

1.6 Consistency of the feature importance by performance groupings.

We also compare how consistent the interpretation in the same performance group and in different performance groups. We calculate the statistics (i.e., mean, median, variance, and standard

deviation) of overlapping values from model pairs from the same performance group and from different performance groups as follows.

(1) **Models in the same performance group**. We calculate the statistics in all combinations of pairs in the same groups. For example, the LDA, QDA, and LR models are in the same performance groups on the Google dataset, so we calculate the mean, median, variance, and standard deviation of the overlapping values from all possible pairs (i.e., LDA~QDA, LDA~LR, and QDA~LR). If there is only one model in one performance group, we put a "-" on all statistics since no available model pair exists. If there are two models in one performance group, then we put a "-" on the variance and standard deviation as there is only one pair of models.
(2) **Models in different performance groups**. We also calculate the statics in all combinations of pairs in differently performed groups. For example, on the Google data set, the GBDT and RGF are in one performance group while CART and NN are in another group, so we calculate the statistics of overlapping values from all combinations between the two groups (i,e, GBDT~CART, GBDT~NN, RGF~CART, and RGF~NN).

Models in the same performance grouping tend to have higher overlapping of feature importance ranking than models in different performance groups. Table 2 shows the overlapping values in the same performance group, and Table 3 shows the results in different groups. We observe that models tend to have a higher overlapping value when they are from the same performance group. For example, on the Google dataset, the average overlapping value for the group LDA, QDA, LR (i.e., the mean value of overlapping values from pairs LDA~QDA, LDA~LR, and QDA~LR) on the top 1, top 3, and top 5 ranked features are 0.70, 0.58, and 0.50, as shown in Table 2. However, the mean overlapping value between the same group LDA, QDA, and LR and other groups have smaller overlapping values (e.g., between the group LDA, QDA, and LR and group CART, NN the overlapping values for top 1, top 3, and top 5 ranked features are only 0.30, 0.36, and 0.53), as shown in Table 3.

Moreover, the overlapping values from the same performance group are all above average on the Backblaze dataset and only 2 overlapping values from the same performance group on the Google dataset are lower than the average (overlapping value for the top 1 ranked features for the CART, NN group, and the top 5 ranked features for the LDA, QDA, LR group), as shown in Table 2. In contrast, the overlapping values from different model groups on the Backblaze data shows a relatively lower value, with only 1, 1, and 3 overlapping values out of the 6 group pairs are above average on the mean values of top 1, top 3, and top 5 ranked features are on the Google dataset, and only 2, 1, and 2 overlapping values out of 6 are above average on the Backblaze dataset, as shown in Table 3.

Models with better performance tend to have more consistent interpretation than models have worse performance. We also observe that the better-performed models tend to produce more consistent feature importance rankings. For example, although RF is in a different performance group than the GBDT and RGF models on the Google dataset, they still share an average overlapping value on top 1, top 3, and top 5 ranked features of 0.59, 0.47, and 0.71, all above the average value on the Google dataset. However, the RF model have share only 0.21, 0.34, and 0.63 overlapping values with the CART, NN group; 0.32, 0.30, and 0.45 overlapping values with LDA, QDA, LR group.

Table 2. Interpretation consistency in the same performance groups. A bold text in the mean values indicates an above average overlapping value.

Dataset	Group	Mean			Median			Var			Sd		
		Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
Google	RF	-	-	-	-	-	-	-	-	-	-	-	-
Google	GBDT, RGF	**0.68**	**0.62**	**0.78**	0.68	0.62	0.78	-	-	-	-	-	-
Google	CART, NN	0.31	**0.47**	**0.64**	0.31	0.47	0.64	-	-	-	-	-	-
Google	LDA, QDA, LR	**0.70**	**0.58**	0.50	0.61	0.57	0.48	0.022	0.005	0.002	0.15	0.07	0.05
Backblaze	GBDT, RF, RGF	**0.99**	**0.57**	**0.76**	0.99	0.59	0.73	0.000	0.003	0.003	0.01	0.05	0.05
Backblaze	LDA, LR, NN	**0.84**	**0.53**	**0.52**	0.83	0.50	0.49	0.001	0.006	0.005	0.03	0.08	0.07
Backblaze	QDA	-	-	-	-	-	-	-	-	-	-	-	-
Backblaze	CART	-	-	-	-	-	-	-	-	-	-	-	-

Table 3. Interpretation consistency in different performance groups. A bold text in the mean values indicates an above average overlapping value.

Dataset	1st Group	2nd Group	Mean			Median			Var			Sd		
			Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
Google	RF	GBDT, RGF	**0.59**	**0.47**	**0.71**	0.59	0.47	0.71	0.007	0.000	0.001	0.08	0.02	0.03
Google	RF	CART, NN	0.21	0.34	**0.63**	0.21	0.34	0.63	0.001	0.009	0.010	0.03	0.09	0.10
Google	RF	LDA, QDA, LR	0.32	0.30	0.45	0.35	0.31	0.41	0.006	0.003	0.007	0.08	0.05	0.08
Google	GBDT, RGF	CART, NN	0.23	0.30	**0.62**	0.24	0.30	0.62	0.005	0.006	0.012	0.07	0.08	0.11
Google	GBDT, RGF	LDA, QDA, LR	0.23	0.25	0.46	0.22	0.25	0.41	0.005	0.001	0.008	0.07	0.04	0.09
Google	CART, NN	LDA, QDA, LR	0.30	0.36	0.53	0.29	0.36	0.53	0.001	0.005	0.005	0.03	0.07	0.07
Backblaze	GBDT, RF, RGF	LDA, LR, NN	**0.88**	0.43	0.42	0.88	0.46	0.43	0.000	0.009	0.006	0.01	0.09	0.08
Backblaze	GBDT, RF, RGF	QDA	0.64	0.38	0.36	0.64	0.39	0.35	0.000	0.007	0.000	0.01	0.08	0.02
Backblaze	GBDT, RF, RGF	CART	**0.85**	0.40	**0.59**	0.85	0.39	0.60	0.000	0.001	0.001	0.01	0.03	0.03
Backblaze	LDA, LR, NN	QDA	0.68	**0.61**	**0.57**	0.67	0.63	0.59	0.000	0.001	0.001	0.02	0.03	0.03
Backblaze	LDA, LR, NN	CART	0.77	0.42	0.40	0.77	0.37	0.40	0.000	0.008	0.005	0.02	0.09	0.07
Backblaze	QDA	CART	0.62	0.44	0.41	0.62	0.44	0.41	-	-	-	-	-	-

1.7 Consistency of the feature importance by model types and performance groupings.

In Table 2, we noticed that even the models are from the same performance grouping, their mean overlapping value could still below the average (e.g., top 5 ranked features for the LDA, QDA, LR group on the Google dataset). We would wonder whether their model type (i.e., whitebox or blackbox model) would contribute to such a difference. Therefore, we further include the model type as another dimension and further investigate its impact on interpretation consistency. We first summarize the groupings into Table 4 regarding both their performance groupings and model types (i.e., whitebox or blackbox model).

Table 4. Models categorized by the model type and performance grouping.

Dataset	Model type	Performance groups			
		Group 1	Group 2	Group 3	Group 4
Google	Whitebox			CART	LDA, QDA, LR
	Blackbox	RF	GBDT, RGF	NN	
Backblaze	Whitebox		LDA, LR	QDA	CART
	Blackbox	GBDT, RF, RGF	NN		

The model type could not be a significant factor for the consistency of model interpretation. Table 5 and Table 6 shows the statistics of overlapping values from the same and different performance grouping considering the model type. The intuition is that the models in the same model type would generate a more consistent interpretation. However, out of our expectation, we do not observe higher overlapping values in the same performance group or lower overlapping values in different performance groups when considering the model types. For example, we separate the performance group LDA, LR, NN into two groups (NN, and LDA, LR) considering the model types, but we observe the overlapping value in group LDA, LR is actually lower than including NN (the mean top 1 overlapping values drops from 0.84 on group LDA, LR, NN to 0.81 on group LDA, LR).

We also notice that two performance groups comprise models from both whitebox and blackbox types: the CART, NN group on the Google dataset and the LDA, LR, NN group on the Backblaze dataset. As shown in Table 6, when we compare the member models considering their model type, we do not find a significant difference with other groups that are both whitebox models or blackbox models. For example, the top 1 and top 3 overlapping values between the CART and NN model are even higher than between the CART and GBDT, RGF model. Our results show that the type of models may not significantly affect the interpretation.

Table 5. Interpretation consistency in the same performance group considering the model type. A bold text in the mean values indicates an above average overlapping value.

Dataset	Model Type	Group	Mean			Median			Var			Sd		
			Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
Google	Whitebox	CART	-	-	-	-	-	-	-	-	-	-	-	-
Google	Whitebox	LDA, QDA, LR	**0.70**	**0.58**	0.50	0.61	0.57	0.48	0.022	0.005	0.002	0.15	0.07	0.05
Google	Blackbox	RF	-	-	-	-	-	-	-	-	-	-	-	-
Google	Blackbox	GBDT, RGF	**0.68**	**0.62**	**0.78**	0.68	0.62	0.78	-	-	-	-	-	-
Google	Blackbox	NN	-	-	-	-	-	-	-	-	-	-	-	-
Backblaze	Whitebox	LDA, LR	0.81	**0.64**	**0.63**	0.81	0.64	0.63	-	-	-	-	-	-
Backblaze	Whitebox	QDA	-	-	-	-	-	-	-	-	-	-	-	-
Backblaze	Whitebox	CART	-	-	-	-	-	-	-	-	-	-	-	-
Backblaze	Blackbox	GBDT, RF, RGF	**0.99**	**0.57**	**0.76**	0.99	0.59	0.73	0.000	0.003	0.003	0.01	0.05	0.05
Backblaze	Blackbox	NN	-	-	-	-	-	-	-	-	-	-	-	-

1.8 Evolution of feature importance

To investigate the changes of model interpretation across different time periods, we also analyze the change of the rankings on the most influential features. Figure 7 shows the evolution of ranking on the top-ranked features. We use the Scott-Knott test to rank the feature importance in each period and choose the top-five highest-ranked features averaged on all the periods. Since the figure takes much space, we place it in the Appendix section.

Most of the features have their ranking changes from one time period to another. As shown in Figure 7, the ranking of most features are not stable across time periods, and the feature importance could at the most change from rank 1 in one period to rank 9 in another period.

Models with better performance tend to have more stable feature rankings. We notice that the better-performed models (e.g., RF, RGF, and GBDT) tend to produce more stable interpretation. For example, on the Backblaze dataset, the top 5 ranked features on the Google dataset only fluctuate between rank 1 and rank 5, while other worse performed models have a more drastic change in the feature ranking in different time periods (e.g., the LDA model have feature importance ranking fluctuate between rank 1 and rank 9). In addition, we observe that the better-performed

Table 6. Interpretation consistency in different performance groups considering the model type. A bold text in the mean values indicates an above average overlapping value.

Dataset	Model Type	1st Group	2nd Group	Mean			Median			Var			Sd		
				Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
Google	Both whitebox	CART	LDA, QDA, LR	0.27	0.30	0.52	0.28	0.30	0.45	0.000	0.001	0.010	0.01	0.03	0.10
Google	Both blackbox	RF	GBDT, RGF	**0.59**	**0.47**	**0.71**	0.59	0.47	0.71	0.007	0.000	0.001	0.08	0.02	0.03
Google	Both blackbox	RF	NN	0.24	0.25	0.53	0.24	0.25	0.53	-	-	-	-	-	-
Google	Both blackbox	GBDT, RGF	NN	0.16	0.23	0.51	0.16	0.23	0.51	0.002	0.000	0.000	0.04	0.01	0.02
Google	Whitebox vs. blackbox	CART	RF	0.19	**0.43**	**0.73**	0.19	0.43	0.73	-	-	-	-	-	-
Google	Whitebox vs. blackbox	CART	GBDT, RGF	0.29	0.38	**0.73**	0.29	0.38	0.73	0.000	0.000	0.000	0.02	0.02	0.01
Google	Whitebox vs. blackbox	CART	NN	0.31	**0.47**	**0.64**	0.31	0.47	0.64	-	-	-	-	-	-
Google	Whitebox vs. blackbox	LDA, QDA, LR	RF	0.32	0.30	0.45	0.35	0.31	0.41	0.006	0.003	0.007	0.08	0.05	0.08
Google	Whitebox vs. blackbox	LDA, QDA, LR	GBDT, RGF	0.23	0.25	0.46	0.22	0.25	0.41	0.005	0.001	0.008	0.07	0.04	0.09
Google	Whitebox vs. blackbox	LDA, QDA, LR	NN	0.32	**0.42**	0.55	0.33	0.40	0.55	0.000	0.001	0.000	0.01	0.04	0.02
Backblaze	Both whitebox	LDA, LR	QDA	0.68	**0.64**	**0.59**	0.68	0.64	0.59	0.001	0.000	0.000	0.02	0.01	0.01
Backblaze	Both whitebox	LDA, LR	CART	0.78	0.36	0.36	0.78	0.36	0.36	0.001	0.000	0.002	0.02	0.01	0.05
Backblaze	Both whitebox	QDA	CART	0.62	0.44	0.41	0.62	0.44	0.41	-	-	-	-	-	-
Backblaze	Both blackbox	GBDT, RF, RGF	NN	**0.88**	0.47	0.44	0.88	0.47	0.43	0.000	0.011	0.001	0.01	0.10	0.02
Backblaze	Whitebox vs. blackbox	LDA, LR	GBDT, RF, RGF	**0.88**	0.42	0.41	0.88	0.44	0.39	0.000	0.007	0.009	0.01	0.09	0.09
Backblaze	Whitebox vs. blackbox	LDA, LR	NN	**0.85**	**0.48**	0.47	0.85	0.48	0.47	0.001	0.000	0.000	0.03	0.02	0.01
Backblaze	Whitebox vs. blackbox	QDA	GBDT, RF, RGF	0.64	0.38	0.36	0.64	0.39	0.35	0.000	0.007	0.000	0.01	0.08	0.02
Backblaze	Whitebox vs. blackbox	QDA	NN	0.67	**0.56**	**0.52**	0.67	0.56	0.52	-	-	-	-	-	-
Backblaze	Whitebox vs. blackbox	CART	GBDT, RF, RGF	**0.85**	0.40	**0.59**	0.85	0.39	0.60	0.000	0.001	0.001	0.01	0.03	0.03
Backblaze	Whitebox vs. blackbox	CART	NN	0.77	**0.55**	0.48	0.77	0.55	0.48	-	-	-	-	-	-

models (i.e., RF, RGF, and GBDT) constantly have one feature (i.e., RSC, Reallocated Sectors Count) on the top 1 position, while other models have its rank varied.

2 CONCLUSION

Our experiments mainly yield the following findings.

(1) **The similarity of model performance positively affect the consistency of model interpretations.** Our results in Section 1.5 shows that when models are in the same performance group (i.e., have similar performance), they tend to have a higher consistency of interpretation (i.e., have more features in the same rank). Our results in Section 1.6 further prove that by comparing the statistics of overlapping on models divided into different performance groups.
(2) **The better-performed model tend to have similar model interpretation.** Our results in Section 1.6 further prove that, when models have better performance, the overlapping of their feature importance ranking is higher than the models have inferior performance.
(3) **The model type could not be a significant factor in the consistency of model interpretations.** In Section 1.7, we find that the model type (i.e., whitebox or blackbox model) will not affect the overlapping of feature importance ranking much, which indicates it could not be a very significant factor on model interpretation.
(4) **The feature importance ranking are constantly changing, and better performed models tend to have smaller range of variation.** In Section 1.8, we analyze the ranking of the most important features in each period.

We also list our findings to research questions that can be fully or partly answered by this preliminary study's results.

(1) **Q3: Does using a black-box vs. white-box model change the generated explanations?** As presented in our Section 1.7, the model type is not a significant factor to the model interpretation. Instead, we find that the model performance could be a more significant factor.
(2) **Q4: Do AIOps models in different performance ranges produce different explanations?** Yes, results in Section 1.8 shows that models that have different model performance

could have very different model interpretation. Our results of the overlapping feature importance ranks in Section 1.6 could also serve as further proof.

(3) **Q6: Do the computed feature importance ranks vary in a cyclical fashion? And if so, do they correlate with something specific to the data?** We find the feature importance ranking changing over time (Section 1.8). However, no recurrent trend has been found.

(4) **Q7: Can we generate globally relevant rules across time windows to simplify AIOps models?** We can possibly build a rule-based model derived from the best-performed models on the Backblaze data, as the feature importance are relatively stable. However, the feature ranking is frequently evolving on the Google dataset, adding challenges to the construction a stable global rule-based model (could because of a more severe level of concept drift in the data).

REFERENCES

[1] Mirela Madalina Botezatu, Ioana Giurgiu, Jasmina Bogojeska, and Dorothea Wiesmann. 2016. Predicting Disk Replacement Towards Reliable Data Centers. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. 39–48.

[2] Nosayba El-Sayed, Hongyu Zhu, and Bianca Schroeder. 2017. Learning from Failure Across Multiple Clusters: A Trace-Driven Approach to Understanding, Predicting, and Mitigating Job Terminations. In *37th IEEE International Conference on Distributed Computing Systems (ICDCS' 17)*. 1333–1344.

[3] Farzaneh Mahdisoltani, Ioan Stefanovici, and Bianca Schroeder. 2017. Proactive error prediction to improve storage system reliability. In *2017 USENIX Annual Technical Conference, USENIX ATC 2017*. 391–402.

[4] Andrea Rosà, Lydia Y. Chen, and Walter Binder. 2015. Catching failures of failures at big-data clusters: A two-level neural network approach. In *2015 IEEE 23rd International Symposium on Quality of Service (IWQoS' 15)*. 231–236.

[5] Andrea Rosà, Lydia Y. Chen, and Walter Binder. 2015. Predicting and Mitigating Jobs Failures in Big Data Clusters. *Proceedings - 2015 IEEE/ACM 15th International Symposium on Cluster, Cloud, and Grid Computing, CCGrid 2015*, 221–230.

[6] Yong Xu, Kaixin Sui, Randolph Yao, Hongyu Zhang, Qingwei Lin, Yingnong Dang, Peng Li, Keceng Jiang, Wenchi Zhang, Jian-Guang Lou, Murali Chintalapati, and Dongmei Zhang. 2018. Improving Service Availability of Cloud Systems by Predicting Disk Error. In *2018 USENIX Annual Technical Conference (USENIX ATC 18)*. USENIX Association, Boston, MA, 481–494.

A THE SEPARATE AUC PERFORMANCE FIGURE

Figure 6 shows the AUC performance on each model on each time period in separate plots.

B FIGURE FOR THE EVOLUTION OF FEATURE IMPORTANCE

Figure 7 shows the evolution of the top-ranked feature importance. We use the Scott-Knott test to rank the feature importance in each period and choose the top-five highest-ranked features averaged on all the periods.

C THE COMPLETE TABLE OF INTERPRETATION CONSISTENCY

Table 7 shows the complete table of all pairs of the overlapping values on the two datasets.

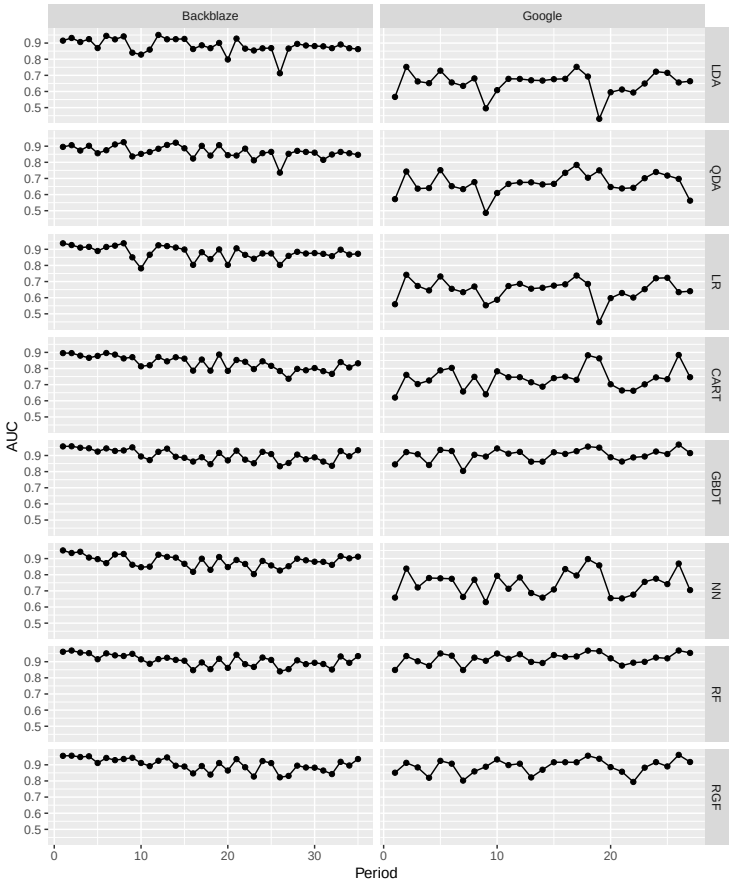


Fig. 6. Comparison of the AUC performance before and after the redundancy and correlation test.

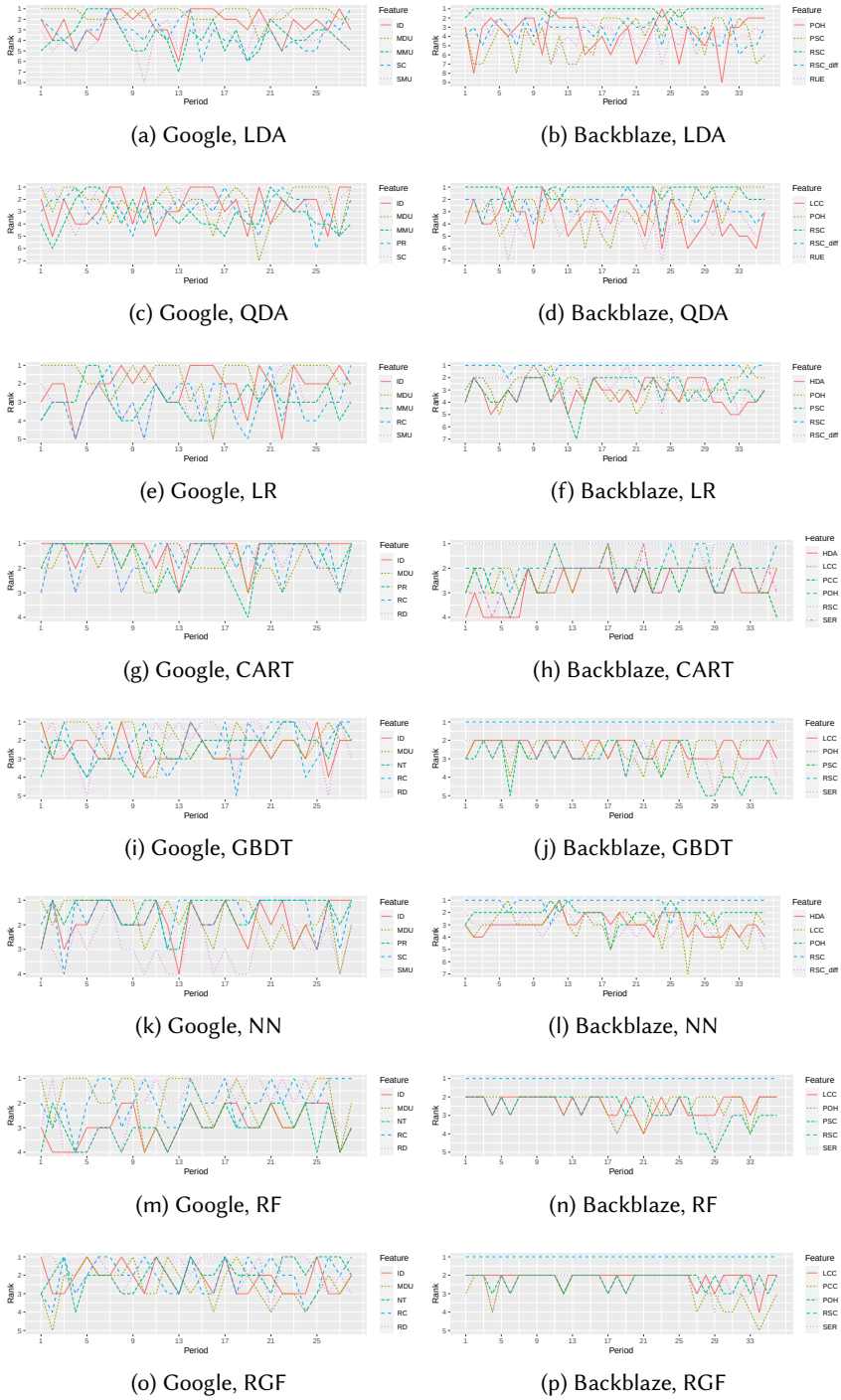


Fig. 7. Evolution of the feature importance ranking in different time periods on the most important features.

Table 7. Interpretation consistency between each pair of models

(a) Google

Dataset	Model 1	Model 2	Top 1	Top 3	Top 5
Google	LDA	CART	0.28	0.26	0.45
Google	LDA	GBDT	0.29	0.26	0.42
Google	LDA	LR	0.90	0.68	0.56
Google	LDA	NN	0.33	0.39	0.52
Google	LDA	QDA	0.61	0.57	0.48
Google	LDA	RF	0.35	0.31	0.41
Google	LDA	RGF	0.17	0.24	0.40
Google	QDA	CART	0.28	0.30	0.44
Google	QDA	GBDT	0.22	0.23	0.38
Google	QDA	LR	0.57	0.50	0.45
Google	QDA	NN	0.33	0.47	0.57
Google	QDA	RF	0.21	0.22	0.37
Google	QDA	RGF	0.12	0.18	0.38
Google	LR	CART	0.26	0.33	0.66
Google	LR	GBDT	0.34	0.28	0.58
Google	LR	NN	0.30	0.40	0.55
Google	LR	RF	0.39	0.36	0.56
Google	LR	RGF	0.23	0.29	0.59
Google	CART	GBDT	0.27	0.36	0.72
Google	CART	NN	0.31	0.47	0.64
Google	CART	RF	0.19	0.43	0.73
Google	CART	RGF	0.31	0.39	0.74
Google	GBDT	NN	0.20	0.24	0.53
Google	GBDT	RF	0.67	0.46	0.74
Google	GBDT	RGF	0.68	0.62	0.78
Google	NN	RF	0.24	0.25	0.53
Google	NN	RGF	0.12	0.22	0.50
Google	RF	RGF	0.51	0.49	0.67

(b) Backblaze

Dataset	Model 1	Model 2	Top 1	Top 3	Top 5
Backblaze	LDA	CART	0.80	0.37	0.31
Backblaze	LDA	GBDT	0.89	0.46	0.33
Backblaze	LDA	LR	0.81	0.64	0.63
Backblaze	LDA	NN	0.83	0.50	0.49
Backblaze	LDA	QDA	0.66	0.63	0.60
Backblaze	LDA	RF	0.88	0.39	0.32
Backblaze	LDA	RGF	0.89	0.25	0.32
Backblaze	QDA	CART	0.62	0.44	0.41
Backblaze	QDA	GBDT	0.64	0.48	0.39
Backblaze	QDA	LR	0.70	0.64	0.59
Backblaze	QDA	NN	0.67	0.56	0.52
Backblaze	QDA	RF	0.63	0.39	0.34
Backblaze	QDA	RGF	0.64	0.28	0.35
Backblaze	LR	CART	0.75	0.35	0.40
Backblaze	LR	GBDT	0.87	0.47	0.45
Backblaze	LR	NN	0.88	0.46	0.46
Backblaze	LR	RF	0.88	0.52	0.53
Backblaze	LR	RGF	0.87	0.41	0.53
Backblaze	CART	GBDT	0.85	0.45	0.55
Backblaze	CART	NN	0.77	0.55	0.48
Backblaze	CART	RF	0.84	0.39	0.60
Backblaze	CART	RGF	0.85	0.37	0.63
Backblaze	GBDT	NN	0.88	0.59	0.47
Backblaze	GBDT	RF	0.99	0.59	0.72
Backblaze	GBDT	RGF	1.00	0.49	0.73
Backblaze	NN	RF	0.87	0.47	0.43
Backblaze	NN	RGF	0.88	0.33	0.41
Backblaze	RF	RGF	0.99	0.62	0.83