Preliminary Study Report regarding Jack's Comments.

YINGZHE LYU, Queen's University, Canada

ACM Reference Format:
Yingzhe Lyu. 2020. Preliminary Study Report regarding Jack's Comments.. 1, 1 (December 2020), 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 GENERAL EXPERIMENTAL DESIGN

In this preliminary study, we mainly use similar approaches as in the prior preliminary study. We list the main points of our experimental design in the following part of this section.

1.1 Time period and feature importance extraction

Our study uses the natural time intervals (i.e., one-day periods for the Google dataset and one-month periods for the Backblaze dataset) to split the data into different time periods. We choose such a time window size as prior works have applied similar update strategies. We further have the following experiment design to extract the model interpretation in each period by using 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 various models (i.e., LDA, QDA, LR, CART, GBDT, NN, RF, and RGF) using samples from each period and calculate its feature importance with permutation feature importance.
(3) To know how well each individual model performs, 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.

1.2 Overlapping of feature ranking

We mainly measure the similarity between two model interpretations with the overlapping value, as in our previous report. The overlapping values represent the consistency between two model interpretations by measuring the proportion that two models have the same features in their top 1, top 3, and top 5 ranked features and is calculated as follows.

(1) On each time period, we first rank the feature importance separately with a Scott-Knott test on the two groups of feature importance value observations we choose.
(2) For the top N features (N=1, 3, or 5) on the two observations, 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 observations are $F1, F2$ while the top 1 ranked features on the other model 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.

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

2020. XXXX-XXXX/2020/12-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(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 observations have similar performance (i.e., in the same SK group); 3) and on periods only when two observations have different performance (i.e., in the different SK groups).

2 COMPARING THE PERFORMANCE OF TUNED VS. UNTUNED MODELS

2.1 Approach

In this section, we study the relationship between model performance and its interpretation to check whether the model quality matters on the explainability results. Specifically, we compare the overlapping of feature ranking from a model tuned with random search (i.e., the tuned model) and a model using only the default configurations (i.e., the untuned model). We carry out our analysis by calculating the overlapping values between the same model before and after hyperparameter tunning on each time period, then using the mean value among all the time period as the final overlapping value.

2.2 Results

Table 1 shows the top 1, top 3, and top 5 overlapping values of feature ranking between tuned and untuned models on the eight models we used on both the Google and Backblaze datasets. We also analyze the AUC performance by calculating the mean AUC difference in each period (the "Diff" column) and the mean AUC performance of untuned and tuned models in each period.

Table 1. The top 1, top 3, and top 5 overlapping between untuned and tuned models.

Dataset	Model	Overlapping			AUC		
		Top 1	Top 3	Top 5	Diff	Untuned	Tuned
Google	LDA	0.90	0.78	0.92	0.01	0.61	0.63
Google	QDA	0.63	0.65	0.79	0.05	0.72	0.75
Google	LR	0.79	0.62	0.89	0.02	0.69	0.71
Google	CART	0.23	0.47	0.85	0.02	0.63	0.63
Google	GBDT	0.77	0.73	0.97	0.01	0.61	0.65
Google	NN	0.58	0.60	0.68	0.03	0.86	0.87
Google	RF	0.87	0.81	1.00	0.00	0.89	0.89
Google	RGF	0.76	0.53	0.72	0.01	0.86	0.86
Backblaze	LDA	0.59	0.53	0.65	0.05	0.85	0.85
Backblaze	QDA	0.35	0.48	0.69	0.12	0.85	0.86
Backblaze	LR	0.81	0.69	0.88	0.01	0.80	0.81
Backblaze	CART	0.85	0.48	0.64	0.02	0.81	0.86
Backblaze	GBDT	1.00	0.81	0.94	0.00	0.72	0.84
Backblaze	NN	0.81	0.61	0.62	0.01	0.88	0.88
Backblaze	RF	0.99	0.86	0.99	0.00	0.88	0.88
Backblaze	RGF	1.00	0.52	0.81	0.01	0.88	0.88

We find that overlapping (i.e., the extent of consistency between the two interpretations) are negatively related to quantity of their performance difference. For example, on the Backblaze data, the mean AUC difference between the untuned and tuned model in each period on the GBDT model is only 0.00 (less than 0.005), and the corresponding top 1, top 3, and top 5 overlapping values are 1.00, 0.81, 0.94, which is relatively high. While the mean AUC difference on the QDA model is 0.12, and its overlapping value is only 0.35, 0.48, 0.69. Our results suggest that the model quality matters when practitioners want to utilize model interpretation, as even on the same model, differently performed models would have a big discrepancy in their model interpretations.

3 COMPARING THE OVERLAPPING OF MODEL INTERPRETATION IN DIFFERENT TIME PERIODS.

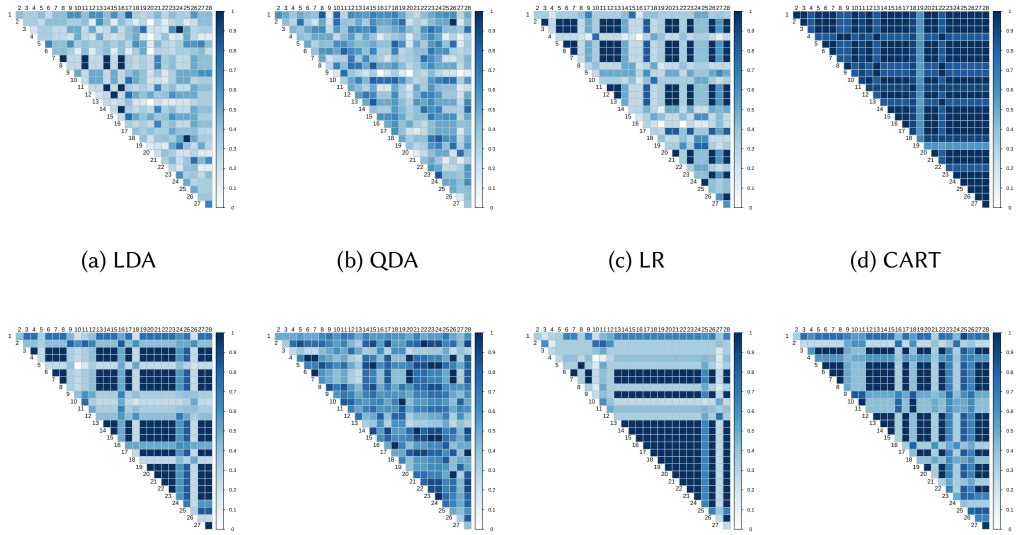
3.1 Approach

In this section, we study to what extent the model interpretation changes in different time periods. This experiment aims to analyze whether the model interpretation changes much in different time periods. If the model interpretation does not change much from time to time (i.e., interpretation stability), then practitioners could update the knowledge from model interpretation less often.

We conduct the experiment as follows. For each feature importance ranking extracted from the model in each time period, we calculate its overlapping values with feature importance rankings extracted from models built on the following time periods.

3.2 Results

Figure 1 and Figure 2 shows the heatmap of top 3 overlapping values on the Google and Backblaze data, respectively. We assume that the top 3 would be the most interesting target as it contains a moderate amount of features, while the top 1 usually only has 1 or 2 features, and the top 5 could contain all the features. We put the figures for the top 1 and top 5 overlapping in the Appendix.

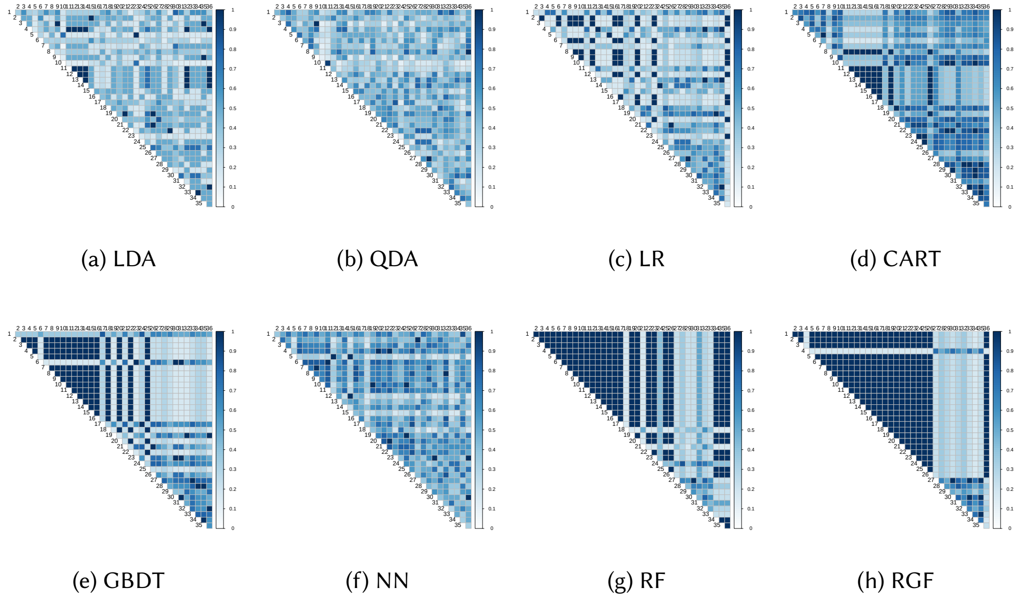


(a) LDA (b) QDA (c) LR (d) CART

(e) GBDT (f) NN (g) RF (h) RGF

Fig. 1. The top 3 overlapping values between different periods on the same model on the Google dataset.

We find that the stability of model interpretation depends highly on the models. We notice that specific types of models have relatively stable interpretation compared with others. For



(a) LDA

(b) QDA

(c) LR

(d) CART

(e) GBDT

(f) NN

(g) RF

(h) RGF

Fig. 2. The top 3 overlapping values between different periods on the same model on the Backblaze dataset.

example, the RF, GBDT, RGF, and CART models have stable interpretation, showing a great number of periods with high overlapping values (i.e., the cells in dark blue color). The reason behind this could be the model performance or model design.

We also find that the change of model interpretation related to time periods. We can clearly notice that the heatmap shows multiple stripes, which means models built on specific time periods have relatively different interpretations compared with other periods. The reason behind this could be the concept drift, which means the distribution of the data and the relationship between the variables evolve over time.

4 COMPARING THE PERFORMANCE OF MODEL TRAINED WITH ONLY TOP RANKED FEATURES.

4.1 Approach

In this section, we conduct experiment to investigate the quality of model interpretation and how could we Based on the feature importance rankings in each period, we now build models using only the top 3 ranked groups of features to evaluate its performance compared with the full models. Prior works have applied similar approach in investigating model interpretability [1]. We conduct experiments as follows.

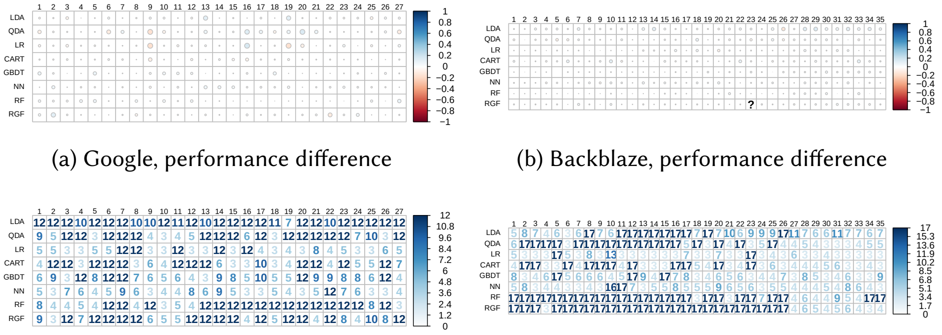
(1) We first compare the performance before and after reducing the number of features of the same model on the immediate successor period, which would be the most concerned scenario in production environment. For each model on each period i, we calculate the relative difference of performance between the models using all features and using only top 3 ranked features that trained on period i and tested on period $i + 1$.
(2) We also compare the performance of model using reduced number of features on all subsequent periods with models using all features training on the immediate predecessor period of

the testing period to analyze how stable the model interpretation could be. For each model, on each period i and each of its subsequent period $j(j > i)$, we calculate the relative difference of performance between the model using only top 3 ranked features that trained on period i and tested on period j with the model using all features and trained on period $j - 1$ and tested on period j.

4.2 Results

The interpretation of models are predictive. Figure 3a and Figure 3b show the relative performance difference between the models trained with all the features and the same model trained with only the top 3 ranked groups of features. We notice that the relative performance difference is at most 16.2% on the Google dataset and 15.5% on the Backblaze dataset The average relative difference of performance among all models is only 0.6% on the Google dataset and 1.3% on the Backblaze dataset.

As shown in Figure 3c and Figure 3d, on difference periods and models, the amount of features in the top 3 ranked groups change significantly, and some of them include all the features, so we do not compare the performance difference among different models as it would not be a fair comparison.



(a) Google, performance difference

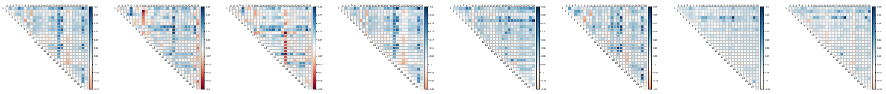
(b) Backblaze, performance difference

(c) Google, features used

(d) Backblaze, features used

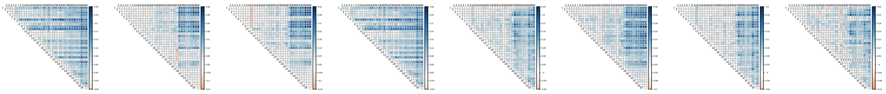
Fig. 3. The performance difference between the models using all features and only top 3 ranked groups of features tested on the immediate successor period and the number of feature in the top 3 ranked groups. **Note:** the experiment for RGF and GBDT models are not finished yet, we are using the results from first few iterations so on some periods the results could be irregular.

The performance of interpretation could change drastically in the future data while the better performed models could have more generalizable interpretation. Figure 4 and Figure 5 shows the relative performance difference of models using only top 3 ranked groups of features in all subsequent time periods compared with models using all features and trained on the immediate predecessor period before the testing period on the Google and Backblaze data, respectively. We notice that the performance of models using only top 3 ranked features could drop drastically on the future time periods. We also notice that the better performed models could have better performance using only top 3 ranked features in the future time periods. For example, the RF and RGF model on the Google dataset (Figure 4g and Figure 4h), which are the top-performed models, have better performance on the subsequent periods than other models.



(a) LDA (b) QDA (c) LR (d) CART (e) GBDT (f) NN (g) RF (h) RGF

Fig. 4. The relative performance difference of models using only top 3 ranked groups of features in all subsequent time periods compared with models using all features and trained on the immediate predecessor period before the testing period on the Google dataset.



(a) LDA (b) QDA (c) LR (d) CART (e) GBDT (f) NN (g) RF (h) RGF

Fig. 5. The relative performance difference of models using only top 3 ranked groups of features in all subsequent time periods compared with models using all features and trained on the immediate predecessor period before the testing period on the Backblaze dataset. **Note:** the experiment for RGF and GBDT models are not finished yet, we are using the results from first few iterations so on some periods the results could be irregular.

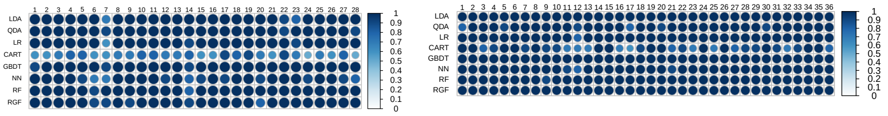
5 COMPARING THE OVERLAPPING OF THE SAME MODEL IN THE SAME TIME PERIOD BY THE PERFORMANCE

5.1 Approach

In this section, we compare the model interpretations among the multiple iterations of experiments. We conduct the experiment as follows. For each model and each period, we sort the feature importance values in each iteration separately and measure the consistency of the top 1 features by calculating how much of them are the same (e.g., if 7 of the 10 iteration have F1 as the most appeared top 1 feature, then the value would be 0.7). We do not conduct experiment on top 3 and top 5 features as there is no proper approach to measure the overlapping in 10 iterations.

5.2 Results

Figure 6 shows the top 1 values on each time period, model, and dataset among multiple iterations.



(a) Google (b) Backblaze

Fig. 6. Consistency of the top 1 features among multiple iterations.

5.3 Concerns about this section

I personally have some concerns regarding the approach used in this section.

(1) The approach used in this section is completely different from other sections in this report and in the earlier report. It could add confusion for the audience.
(2) In this section, we are only using the top N features according to the feature importance values, not the top N ranked groups of features using Scott-Knott test. Without statistical test, it would be hard to persuade audience and add uncertainty to the conclusion.
(3) There is no good approach for measuring 10 groups of observations, and our overlapping value is only for two groups of observations.
(4) It is not reasonable enough to separate the 10 iterations. The feature importance rankings are only meaningful when calculating on multiple iterations, and the top N feature in one iteration could be easily disturbed by random factors (e.g., randomness in the model architecture, downsampling) and not trustworthy enough.

6 FURTHER SUGGESTIONS

(1) We use the term explainability and interpretability interchangeably in this and our previous preliminary study report. However, I notice that the two words have very subtle differences, as the interpretability is the extent to which **a cause and effect** can be observed within a system while the explainability is the extent to which **the internal mechanics** of a machine or deep learning system can be explained in human terms [1]. We may need to make adjustments to the words we use.
(2) During the experiments in Section 3 and Section 4, I find that the grouping of feature importance ranking generates less than five groups of observation. I was wondering whether it would be a problem when we are calculating the top 5 overlapping values (e.g., the importance from the 12 features on the Google dataset in some occasion could only be categorized into 3 groups, then we only have top 3 rankings instead of top 5).
(3) When comparing the interpretation between good and bad iterations, I find it hard to separate the 10 iterations into two groups. For now, I choose the top 5 performed as the first group and the other 5 as the second group by the AUC performance. If we group the 10 iterations with Scott-Knott clustering, we cannot guarantee it could generate exactly two groups, and there could be only one or a few iterations in one group, which is not ideal for ranking feature importance.
(4) Noticed in Section 4, the top 3 ranked features could cover all features in some occasion, so maybe we could try Scott-Knott ESD to increase the groups of ranks. However, one concern could be inconsistency with the model groupings used in earlier preliminary study, where we managed to group the 8 models into 3 differently performed groups while the Scott-Knott ESD would generate much more groups, rendering the analysis impossible.

7 CONCLUSION

Based on our experiment results, we could answer the comments provided by Jack.

(1) **Comment 1: The impact of model quality.** We find that the model quality does affect the model interpretation, and the better-performed models (i.e., tuned models) can provide better interpretations, which suggest practitioners still need better-performed models when extracting model interpretation.
(2) **Comment 2: Whether interpretation obtained from old data suits the new data.** We notice that specific models like CART and RF have more stable interpretation over time. However, we also find evidence suggesting that the concept drift issue would affect the

[1] https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html

model stability over time. We may incorporate concept drift detection or other techniques to mitigate the concept drift problem.

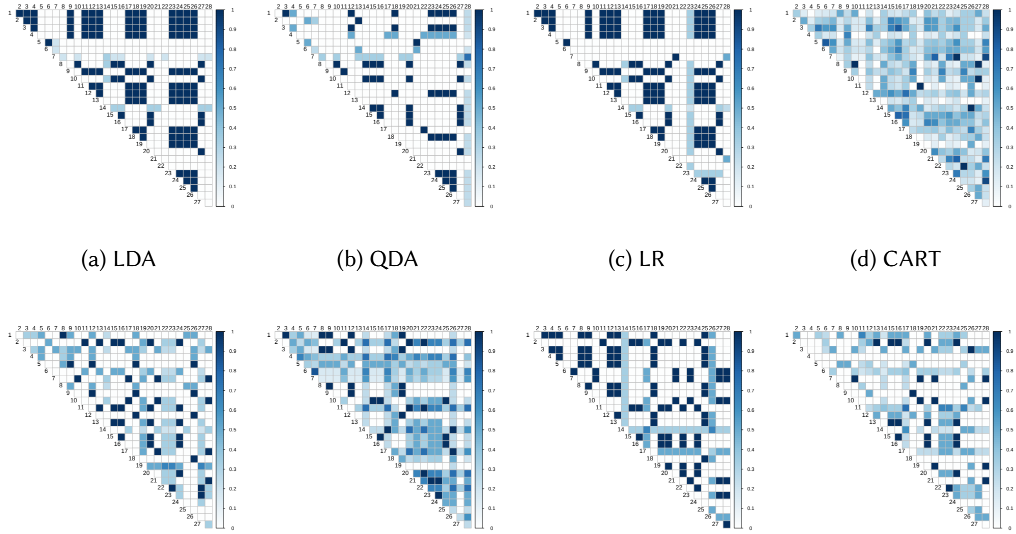
(3) **Comment 3: How trustworthy is the model interpretation.** We find that when only using the top 3 ranked groups of features instead of all the features, the performance only drops a bit, which proves that the model interpretation on the immediate successor period is trustworthy enough. However, we notice that the performance on the subsequent (i.e., more faraway future time periods) drops drastically, indicating the interpretation may not fit the future samples. Again, we should take concept drift into consideration.

(4) **Comment 4: The stability of model interpretation in different iterations.** The performance difference in each iteration does not have a strong level of connection to the discrepancy of model interpretations.

REFERENCES

[1] Yangguang Li, Zhen Ming Jiang, Heng Li, Ahmed E. Hassan, Cheng He, Ruirui Huang, Zhengda Zeng, Mian Wang, and Pinan Chen. 2020. Predicting Node Failures in an Ultra-large-scale Cloud Computing Platform: an AIOps Solution. *ACM Transactions on Software Engineering and Methodology* (2020).

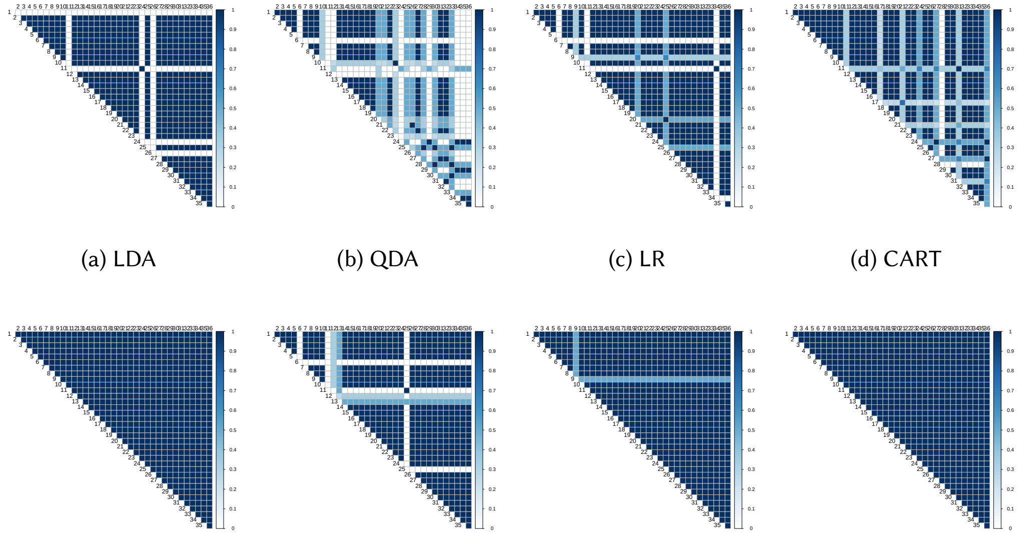
A FIGURE FOR THE TOP 1 AND TOP 5 OVERLAPPING ON THE SAME MODEL IN DIFFERENT TIME PERIODS.



(a) LDA (b) QDA (c) LR (d) CART

(e) GBDT (f) NN (g) RF (h) RGF

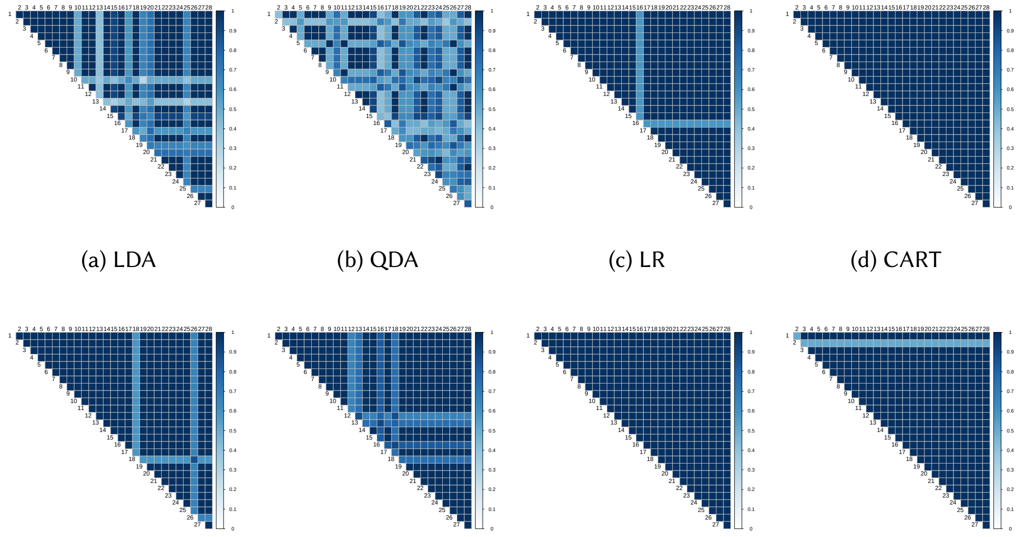
Fig. 7. The top 1 overlapping values between different periods on the same model on the Google dataset.



(a) LDA (b) QDA (c) LR (d) CART

(e) GBDT (f) NN (g) RF (h) RGF

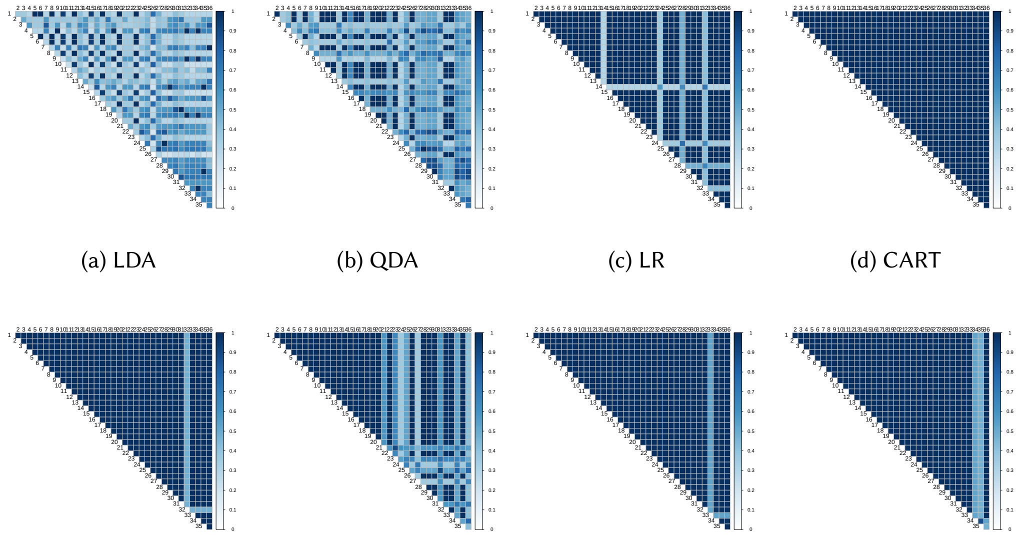
Fig. 8. The top 1 overlapping values between different periods on the same model on the Backblaze dataset.



(a) LDA (b) QDA (c) LR (d) CART

(e) GBDT (f) NN (g) RF (h) RGF

Fig. 9. The top 5 overlapping values between different periods on the same model on the Google dataset.



(a) LDA (b) QDA (c) LR (d) CART

(e) GBDT (f) NN (g) RF (h) RGF

Fig. 10. The top 5 overlapping values between different periods on the same model on the Backblaze dataset.