# Replication project

Source code properties of defective infrastructure as code scripts

Nicolas Legros

Montréal, Canada
nicoals.legros@polytml.ca

Thomas Trépanier

Montréal, Canada
thomas.trépanier@polytml.ca

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—component, formatting, style, styling, insert**

## I. Problem statement

1) The original study's problem statement.

## II. Research questions

### *A. RQ1: What source code properties characterize defective infrastructure as code scripts?*

### *B. RQ3: How can we construct defect prediction models for infrastructure as code scripts using the identified source code properties?*

## III. Approach

3) Methodologies for answering each RQ (How to mine the repositories, the tools employed, and ML models they've used).

### *A. Repository mining*

### *B. RQ1: What source code properties characterize defective infrastructure as code scripts?*

For this research question we used the reported data from the paper[1]. We used the *Mann-Whitney U* test with the Scikit Learn package to evaluate which properties had the biggest influence on defective files. The null hypothesis is that the property is not different between defective and neutral files, and the alternative hypothesis is that the property is larger for defective than neutral files. As in the paper, we consider a significance level of 95% which means we reject the null hypothesis when $p - value < 0.05$.

We also used *Cliff's Delta*[2] to measure how large the difference between the distribution of each characteristics for defective and neutral files is.

[1]https://figshare.com/s/ad26e370c833e8aa9712
[2]https://github.com/neilernst/cliffsDelta

### *C. RQ3: How can we construct defect prediction models for infrastructure as code scripts using the identified source code properties?*

Before using statistical learners, we completed a PCA analysis to determine what properties should be used. We only used the principal component that accounted for at least 95% of the total variance as the input for the statistical learners. We can see in Table I that only one or two principle components account for 95% of the total variance depending on the dataset.

With the component created, we than used it as the input for the different statistical learners. Like the paper, we used Scikit Learn packages to construct the models. The learners that were used are Classification Tree (CART), K Nearest Neighbor (KNN), Logistic Regression (LR), Naive Bayer (NB) and Random Forest (RF).

To evaluate the performance of the different classification models, we used the same metrics as the paper (i.e. precision, recall, AUC, F-measure).

## IV. Results and discussion

### *A. RQ1: What source code properties characterize defective infrastructure as code scripts?*

### *B. RQ3: How can we construct defect prediction models for infrastructure as code scripts using the identified source code properties?*

TABLE I
Number of principle components for the models

| Dataset | Number of components |
|---|---|
| Mirantis | 1 |
| Mozilla | 1 |
| Openstack | 2 |
| Wikimedia | 2 |

## V. Conclusion

5) conclusion

TABLE II
CROSS-VALIDATION RESULTS FOR MIRANTIS

| | RF | NB | LR | KNN | CART |
|---|---|---|---|---|---|
| AUC | 0.701661 | 0.714252 | 0.750981 | 0.693334 | 0.659597 |
| Recall | 0.707425 | 0.407360 | 0.649691 | 0.672546 | 0.707425 |
| Precision | 0.701199 | 0.846909 | 0.798322 | 0.667389 | 0.701199 |
| F1-measure | 0.695896 | 0.541781 | 0.708236 | 0.663964 | 0.698448 |

TABLE III
CROSS-VALIDATION RESULTS FOR MOZILLA

| | RF | NB | LR | KNN | CART |
|---|---|---|---|---|---|
| AUC | 0.731664 | 0.699599 | 0.756323 | 0.713161 | 0.691230 |
| Recall | 0.649017 | 0.392519 | 0.565923 | 0.619417 | 0.627106 |
| Precision | 0.642651 | 0.831862 | 0.706600 | 0.604170 | 0.642550 |
| F1-measure | 0.645764 | 0.532261 | 0.626990 | 0.608864 | 0.633393 |

## REFERENCES


[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.




TABLE IV
CROSS-VALIDATION RESULTS FOR OPENSTACK

| | RF | NB | LR | KNN | CART |
|---|---|---|---|---|---|
| AUC | 0.647741 | 0.694343 | 0.659972 | 0.659195 | 0.574832 |
| Recall | 0.667616 | 0.368902 | 0.731321 | 0.687022 | 0.660176 |
| Precision | 0.653112 | 0.847009 | 0.643218 | 0.661449 | 0.655685 |
| F1-measure | 0.660440 | 0.512676 | 0.682243 | 0.673287 | 0.657360 |

TABLE V
CROSS-VALIDATION RESULTS FOR WIKIMEDIA

| | RF | NB | LR | KNN | CART |
|---|---|---|---|---|---|
| AUC | 0.664721 | 0.709438 | 0.736270 | 0.699140 | 0.583164 |
| Recall | 0.591171 | 0.366128 | 0.586200 | 0.627349 | 0.587493 |
| Precision | 0.664007 | 0.885945 | 0.774041 | 0.732415 | 0.666620 |
| F1-measure | 0.628276 | 0.515651 | 0.663515 | 0.673132 | 0.623400 |