The Job Change of Data Scientists: a case study with imbalanced data

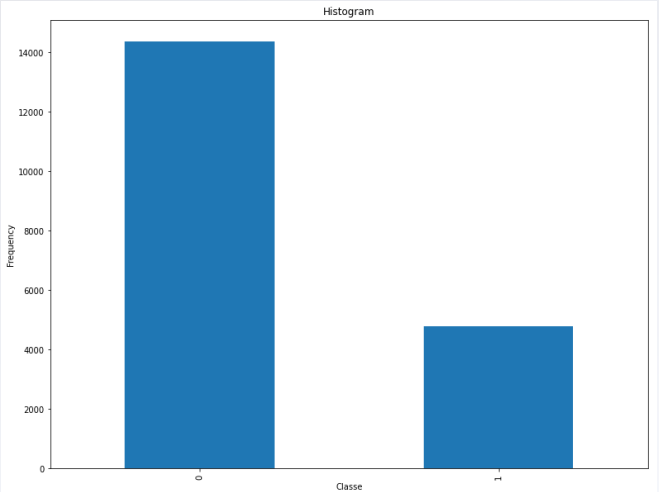
Lucas de A. G. Paixão[1], Arlindo Fraga Neto[2]
[1] Universidade Federal da Bahia
[2] Universidade Federal da Bahia

Abbreviated abstract: With the diffusion of data science across the industry, more and more companies are attempting to scoop worthy professionals. This present work attempts to predict a person's willingness to change career to that of a data scientist based on his/her features. Three models, logistic regression, KNN and random forests were tested on an imbalanced dataset and evaluated on accuracy, F1-score and AUC metrics, before and after rebalancing using SMOTE technique variants. The results showed that the rebalancing a dataset can significantly improve a model's performance.

3rd Conference on
Statistics and
Data Science
Salvador, Brazil (online)
October 28-30, 2021

Problem, Data, Previous Works

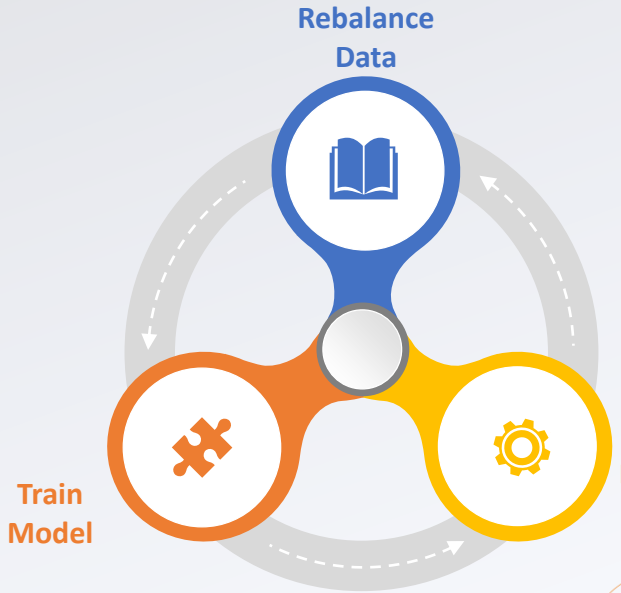
- Predict willingness to change job to that of a data scientist.
- Binary classification problem:
  - **Observations**: individuals interviewed (n = 19,158);
  - **Predictors**: features from those individuals (gender, experience, education level etc.) (p = 12);
  - **Response**: 'willing to change' ('one') and 'not willing to change ('zero');
- Dataset is imbalanced towards the negative class;
- Most of the predictors are categorical and imbalanced towards one category;
- Results evaluated before and after rebalancing the response class.



3rd Conference on
Statistics and
Data Science
Salvador, Brazil (online)
October 28-30, 2021

Methods

- Categorical predictors were encoded using OneHotEncoding technique;
- Numerical predictors were standardized using z-score;
- Missing values from predictors were replaced with the mode from the variable;
- Models tested with 5-fold Cross Validation;
- 3 different classifiers tested: Logistic regression, K-nearest Neighbors (KNN) and Random forests;
- 3 different synthetic rebalancing techniques: SMOTE, SMOTE + Tomek Links, ADASYN;



3rd Conference on Statistics and Data Science
Salvador, Brazil (online)
October 28-30, 2021

Results and Conclusions

	Average results from 5-fold cross-validation											
	Original Dataset			SMOTE			SMOTE+Tomek Links			ADASYN		
	Accuracy	F1-Score	AUC	Accuracy	F1-Score	AUC	Accuracy	F1-Score	AUC	Accuracy	F1-Score	AUC
Logistic Regression	0.7669	0.3673	0.6014	0.7606	0.4830	0.6564	0.7608	0.4889	0.6599	0.7605	0.6446	0.4626
K-nearest Neighbors (k=13)	0.7631	0.4317	0.6288	0.6046	0.4980	0.6653	0.5976	0.4912	0.6581	0.5565	0.6411	0.4766
Random Forests	0.7307	0.5221	0.6843	0.7285	0.5540	0.7109	0.7273	0.5460	0.7038	0.6927	0.7108	0.5487

- Accuracy is higher on the original dataset, however it isn't a good metric for imbalanced data;
- Performance is relatively similar for the different models, however Random Forests stay slightly ahead (ignoring confidence intervals);
- Smote provided the best AUC results, but ADASYN gave a more balanced precision and recall mean through the F1-score;
- Conclusively, the adoption of rebalancing techniques can help improve the results on imbalanced datasets.

3rd Conference on
Statistics and
Data Science
Salvador, Brazil (online)
October 28-30, 2021