Identifying early help referrals for local authorities with machine learning and bias analysis*

(*) available at Journal of Computational Social Science (open access).

Dr Eufrásio Lima Neto(1), Prof Georgina Cosma(2) and Dr Axel Finke(2), Jonathan Bailiss(2), Jo Miller(3)
De Montfort University(1) Loughborough University(2) Leicestershire County Council(3)

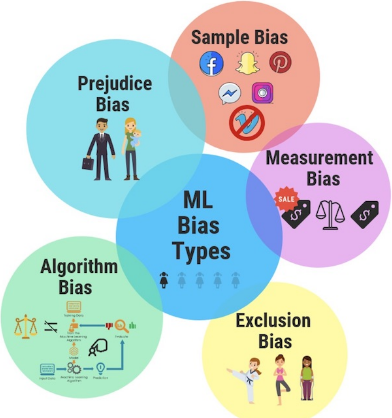
Paris, 24th May 2024

Outline

1. Motivation
2. Introduction
3. Dataset and Preprocessing
4. Machine learning models and bias analysis
5. Findings
6. Conclusions
7. Going Forward

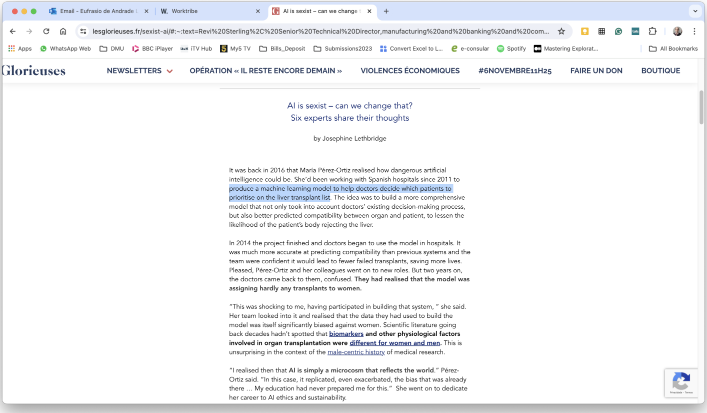
Motivation

Algorithm bias: the lack of fairness in the output generated by an algorithm. These biases may include age discrimination, gender bias, and racial bias, among others (Equality Act 2010).



Motivation: algorithm bias

- Article: AI is sexist – can we change that? (https://lesglorieuses.fr/sexist-ai/)



AI is sexist – can we change that?
Six experts share their thoughts

by Josephine Lethbridge

It was back in 2016 that María Pérez-Ortiz realised how dangerous artificial intelligence could be. She'd been working with Spanish hospitals since 2011 to produce a machine learning model to help doctors decide which patients to prioritise on the liver transplant list. The idea was to build a more comprehensive model that not only took into account doctors' existing decision-making process, but also better predicted compatibility between organ and patient, to lessen the likelihood of the patient's body rejecting the liver.

In 2014 the project finished and doctors began to use the model in hospitals. It was much more accurate at predicting compatibility than previous systems and the team were confident it would lead to fewer failed transplants, saving more lives. Pleased, Pérez-Ortiz and her colleagues went on to new roles. But two years on, the doctors came back to them, confused. **They had realised that the model was assigning hardly any transplants to women.**

"This was shocking to me, having participated in building that system, " she said. Her team looked into it and realised that the data they had used to build the model was itself significantly biased against women. Scientific literature going back decades hadn't spotted that **biomarkers and other physiological factors involved in organ transplantation were different for women and men.** This is unsurprising in the context of the male-centric history of medical research.

"I realised then that **AI is simply a microcosm that reflects the world.**" Pérez-Ortiz said. "In this case, it replicated, even exacerbated, the bias that was already there … My education had never prepared me for this." She went on to dedicate her career to AI ethics and sustainability.

Introducing the problem

Problem: Local authorities in England, such as the Leicestershire County Council (LCC) provide **Early Help services** that may be offered at any point in a young person's life, when they are experiencing some difficulties that universal services, such as schools, alone cannot support.

Introducing the problem

Problem: Local authorities in England, such as the Leicestershire County Council (LCC) provide **Early Help services** that may be offered at any point in a young person's life, when they are experiencing some difficulties that universal services, such as schools, alone cannot support.

Goal: Provide a data-driven solution to classify young people about Early Help services based on social-demographic features and/or educational indicators.

Introducing the problem

Problem: Local authorities in England, such as the Leicestershire County Council (LCC) provide **Early Help services** that may be offered at any point in a young person's life, when they are experiencing some difficulties that universal services, such as schools, alone cannot support.

Goal: Provide a data-driven solution to classify young people about Early Help services based on social-demographic features and/or educational indicators.

LCC's triage is categorised as:

- EH SUPPORT: Early Help support – the most intense type of intervention; ($\approx$ 33%)
- SOME ACTION: referral to less intensive services such as group activities or schemes that run during the school holidays, or to external services; ($\approx$ 57%)
- NO ACTION: additional support is not currently required. ($\approx$ 10%)

Introducing the problem

The main aspects addressed in this work were:

- ML models were implemented and their performance was evaluated across different training and validation sets;
- A fairness analysis was conducted and mitigation algorithms were applied to reduce bias in the ML models;
- The models identified 83 % and 85 % of all young people that need EH SUPPORT and SOME ACTION, respectively, and presented a low false negative rate (FNR) (15 % and 17 % respectively) on the test set.
- The study revealed that certain educational indicators such as fixed-term exclusion and free school meals are important to predict the need for Early Help services (EHS).

Dataset and Preprocessing

- Data of **young people under 18 years old** and were assessed for Early Help support between **April 2019 and August 2022**.
- LOCALITY DECISION (EH SUPPORT, SOME ACTION and NO ACTION) holds Early Help outcomes.
- Input features:
  - Social-demographic: AGE AT LOCALITY DECISION, GENDER, IDACI.
    - *IDACI: Income Deprivation Affecting Children Index measures the proportion of all children living in income-deprived families.*
  - Educational indicators grouped into topics such as Absence, Exclusion, School Transfer, Free School Meal (FSM), Special Educational Needs and Disabilities (SEND), Pupil Referral Unit (PRU), Home Education, Missing, Not in Education, Employment or Training (NEET), Early Years Funding (EYF).

Dataset and Preprocessing

- The initial dataset:
  - 15 976 records and 149 features
  - Sparse and noisy ⇒ missing values and 'not applicable' (NA) cells.
  - Example: Not in Education, Employment or Training (NEET) is not applicable to those under 16 years.
- After pre-processing:
  - 14 360 rows ⇒ records with more than 30 % of missing values were removed (10 % of the data).
    - The percentage of missing values (5%).
    - The percentage of NA cells (20%).
  - For all features with NA values ⇒ one-hot encoding method was used ⇒ each feature was paired with another feature called FEATURE NAME NA that received the value of 1 if the original feature was not relevant to the record ⇒ number of features increased to 363.

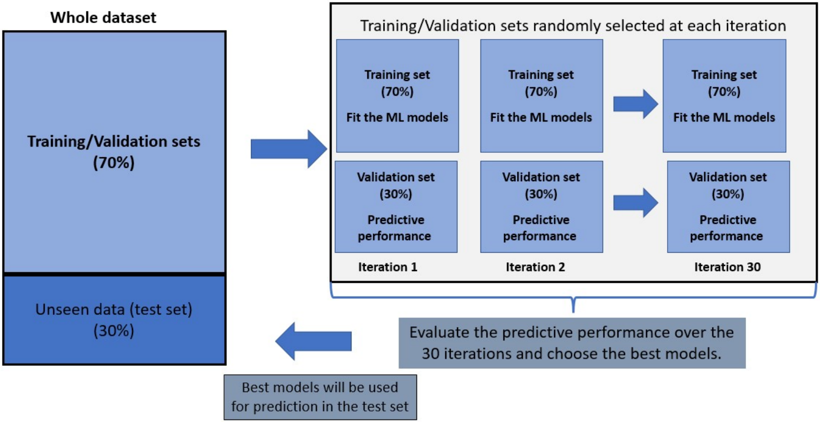
Machine learning models and bias analysis

1. A ML model was created for each LOCALITY DECISION outcome.
   - Model 1 predicts whether a young person needs EH SUPPORT
   - Model 2 predicts whether a young person needs SOME ACTION
   - Model 3 predicts whether a young person needs NO ACTION

2. A single model (**multiclass model**) to predict the three LOCALITY DECISION outcomes.
   - This model did not give acceptable results.
3. Fairness assessment and bias mitigation strategies
   - Fairness assessment in social-demographic features.
   - Bias mitigation algorithms.

Model evaluation

Monte Carlo Experiments



Whole dataset

Training/Validation sets (70%)

Unseen data (test set) (30%)

Training/Validation sets randomly selected at each iteration

Training set (70%)
Fit the ML models

Validation set (30%)
Predictive performance

Iteration 1

Training set (70%)
Fit the ML models

Validation set (30%)
Predictive performance

Iteration 2

Training set (70%)
Fit the ML models

Validation set (30%)
Predictive performance

Iteration 30

Evaluate the predictive performance over the 30 iterations and choose the best models.

Best models will be used for prediction in the test set

Model evaluation

- 14 different ML models were evaluated.
- The best models were chosen based on their performance in identifying the outcomes (metrics: AUC, Recall and Precision).
- Training set (70 % ≈ 10,052 individuals) and Test (unseen data) set (30 % ≈ 4,308 individuals).
- 10-fold stratified cross-validation used to validate models.
- Models were run 30 times, each time a different subset of the dataset was used for training and validation. To evaluate the performance of the models across different sub-populations.
- Python (`pycaret` library)

Model evaluation

The following machine-learning techniques were considered:

1. Ridge Classification
2. Logistic Regression
3. SVC (Support Vector Classification: Linear and Kernel)
4. KNN Classifier
5. Gaussian Naive Bayes
6. Decision Tree
7. Random Forest Classifier
8. Gradient Boosting Classifier
9. Extreme Gradient Boosting
10. Ensemble Methods (Ada boost, Catboost)
11. Discriminant Analysis (Linear and Quadratic)

Model evaluation

- Model fairness and bias mitigation algorithms were carried out and techniques were applied to correct any detected bias.
    - Socio-Demographic features: GENDER, CLASS AGE, ATTENDANCE, and IDACI.
    - The FNR was used as a metric to mitigate bias since it represents those who would benefit from the EH SUPPORT (or SOME ACTION or NO ACTION) but were not predicted as such.
    - The two-sample Z-test for proportions statistical test was applied to evaluate whether there is a significant difference between the FNR of the categories for a given feature.

Machine learning models: results

Best models (predictive performance in the test set):

- The EH SUPPORT model correctly identifies 83 % of young people that received EHS.
- The SOME ACTION model correctly identifies 85 % of young people that received SOME ACTION.
- The NO ACTION model correctly identifies 61 % of young people that received NO ACTION.
- The **multiclass model** did not perform well, correctly identifying only 46 % of young people who received EHS*.

Table: Predictive performance in the test set

Model	ML	AUC	Precision	Recall
EH SUPPORT	LR	0.63	0.38	**0.83**
SOME ACTION	GBC	0.60	0.59	**0.85**
NO ACTION	LR	0.56	0.12	0.61
Multiclass	LR	0.59	0.42*	0.46*

Fairness assessment and bias mitigation in ML models

Warning:

Young people that would benefit from EH SUPPORT or SOME ACTION but are not recommended for it may experience allocation **harm**. In the context of the classification scenario, these are called **false negatives.**

Goals:

- To identify groups that may be disproportionately negatively impacted by the ML model.
- To identify if the social-demographic variables are **sensitive features** and **mitigate the bias** reducing the false negative rate (FNR).
  - Two bias mitigation algorithms: `ThresholdOptimizer` (Postprocessing) and `ExponentiatedGradient` (Reductions).

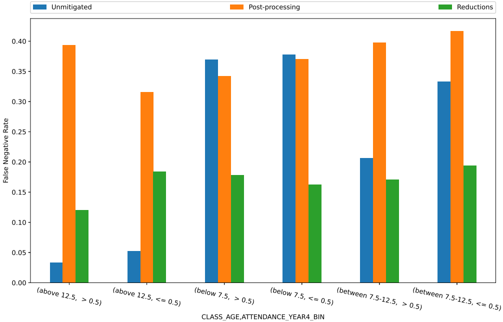
Bias mitigation strategies

- ThresholdOptimizer algorithm
  - Proposed by Hardt, Price, and Srebro (2016).
  - Postprocessing strategy: applied after training the model.
  - Mitigate bias in the predictions.
  - Based on the false negative rate parity which requires that all the groups have similar false negative rates.

- ExponentiatedGradient algorithm
  - Proposed by Agarwal et. al (2018).
  - Reduction strategy: applied during model fitting.
  - Rather than requiring that false negative rates be equal, it is possible to specify the maximum allowed difference or ratio between the largest and the smallest value.

Bias mitigation: SOME ACTION model

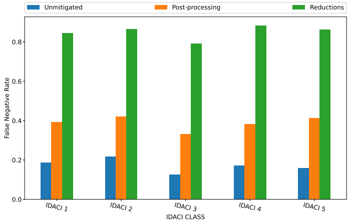
- Groups of young people **were disproportionately negatively impacted** (in terms of FNR) in the features CLASS AGE and ATTENDANCE.



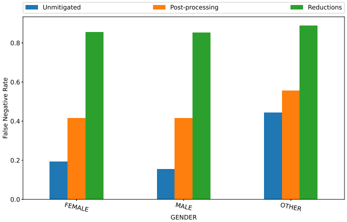
- The model with `reductions algorithm` correctly identified 85 % of young people who need SOME ACTION

Bias mitigation: EH SUPPORT model

- Young people **were not negatively impacted** by the ML model according to IDACI and GENDER.



IDACI: FNR between the unmitigated model and the mitigation strategies.

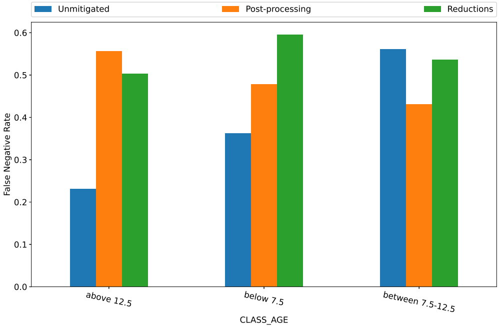


GENDER: FNR between the unmitigated model and the mitigation strategies.

- The bias mitigation algorithms did not reduce the FNR for all groups.

Bias mitigation: NO ACTION model

The use of the bias mitigation algorithms did not reduce the FNR for all groups.



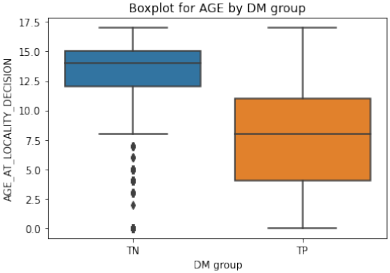
Machine learning models: 2-step procedure

- The binary models for EH SUPPORT and SOME ACTION presented a good predictive performance in validation and test sets.

- An **2-step procedure** is proposed to classify young people as EH SUPPORT, SOME ACTION or NO ACTION by the LCC:
    - **Step 1:** apply the EH SUPPORT model and IF the model classifies the individual as label 1, proceed with EHS. Otherwise, go to Step 2.
    - **Step 2:** apply the SOME ACTION model and IF the model classifies the individual as label 1, proceed with Some Action Services. Otherwise, NO ACTION is required.

Findings: Young people who required EH SUPPORT

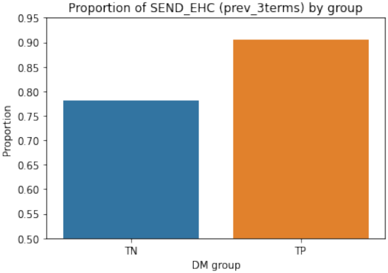
- The **median age** of those who did not require EH SUPPORT was 14 years, compared to a median age of 8 years for those who did require EH SUPPORT.

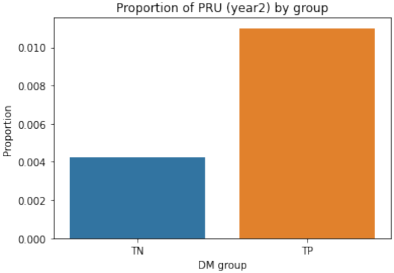


TN (true negative) refers to a young person who did not require EH SUPPORT and was predicted as such. TP (true positive) refers to a young person who required EH SUPPORT and was predicted by the model as such.

Findings: Young people who required EH SUPPORT

- A **higher proportion** of young people who required EH SUPPORT attended a PRU or received SEND EHC support compared to those who did not require EH SUPPORT.

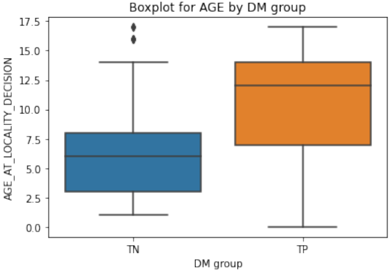




TN (true negative) refers to a young person who did not require EH SUPPORT and was predicted as such. TP (true positive) refers to a young person who required EH SUPPORT and was predicted by the model as such.

Findings: Young people who required SOME ACTION

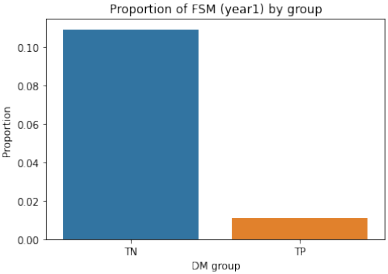
- The **median age** of those who did not receive SOME ACTION was 6 years compared to 12 years for those who did receive it.

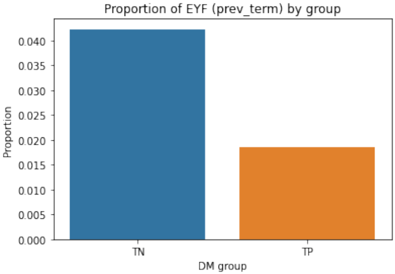


TN (true negative) refers to a young person who did not require SOME ACTION and was predicted as such. TP (true positive) refers to a young person who required SOME ACTION and was predicted by the model as such.

Findings: Young people who required SOME ACTION

- A **higher proportion** of young people who received EYF or were eligible for FSM did not receive SOME ACTION.

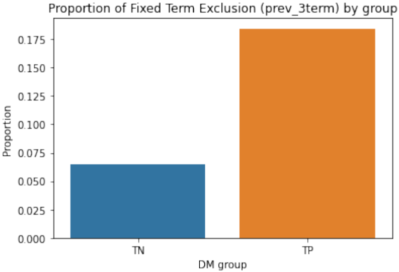




TN (true negative) refers to a young person who did not require SOME ACTION and was predicted as such. TP (true positive) refers to a young person who required SOME ACTION and was predicted by the model as such.

Findings: Young people who required SOME ACTION

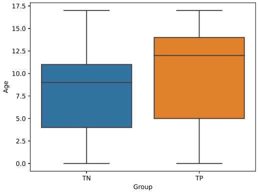
- The **average number** of FIXED-TERM EXCLUSION sessions **was higher** among those who received SOME ACTION (0.184 average) compared to those who did not (0.065 average).

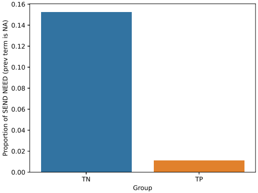


TN (true negative) refers to a young person who did not require SOME ACTION and was predicted as such. TP (true positive) refers to a young person who required SOME ACTION and was predicted by the model as such.

Findings: Young people who required NO ACTION

- A **high proportion** of young people who had NA values in the SEND NEED feature did not belong to the NO ACTION category (hence they had received EH SUPPORT or SOME ACTION).
- No difference in the distribution of the AGE AT LOCALITY DECISION for those belonging to the TN group (median age of 9 years) and TP group (median of 12 years).





TN (true negative) refers to a young person who did not require NO ACTION and was predicted as such. TP (true positive) refers to a young person who required NO ACTION and was predicted by the model as such.

Conclusions

- The EH SUPPORT model correctly identified 83 % of all young people that need EH SUPPORT.
- The SOME ACTION model correctly identified 85 % of all those that need SOME ACTION.

Conclusions

- The EH SUPPORT model correctly identified 83 % of all young people that need EH SUPPORT.
- The SOME ACTION model correctly identified 85 % of all those that need SOME ACTION.
- The fairness assessment (bias check) identified groups of young people negatively impacted in the features CLASS AGE and ATTENDANCE.
    - The use of bias mitigation algorithms reduced the bias in all the groups and improved the model's predictive performance.

Conclusions

- The EH SUPPORT model correctly identified 83 % of all young people that need EH SUPPORT.
- The SOME ACTION model correctly identified 85 % of all those that need SOME ACTION.
- The fairness assessment (bias check) identified groups of young people negatively impacted in the features CLASS AGE and ATTENDANCE.
  - The use of bias mitigation algorithms reduced the bias in all the groups and improved the model's predictive performance.
- The features GENDER and IDACI (Income Deprivation Affecting Children Index) did not bias the models with regards to predicting LOCALITY DECISION (i.e. EH SUPPORT, SOME ACTION, NO ACTION).

Conclusions

- The LIME analysis identified important features in the EH SUPPORT and SOME ACTION models.
  - Median age (AGE AT LOCALITY DECISION) of those who did not require EH SUPPORT was 14 years, compared to a median age of 8 years for those who did require.
  - Median age (AGE AT LOCALITY DECISION) for those who did not receive SOME ACTION was 6 years compared to 12 years for those who received SOME ACTION.
  - PRU services, SEND support, benefits such as EYF or FSM and FIXED-TERM EXCLUSION sessions represent important features to the need for EHS or contribute to the need for less intensive services (SOME ACTION).

Going Forward

Limitations:

- A decision was taken early on to restrict to young people aged 18 and under at the point of assessment.
- This does leave a blind spot in the study of young people aged 18 and under who were never referred for assessment (but possibly should have been).

Future works:

- Evaluate the factors relating to primary and secondary school young people who require EH SUPPORT.
- To incorporate new demographic features to uncover new insights and findings.
- With EHS being a whole family intervention service there is a future piece of work to understand those interrelationships and identify requirements at a family level.

References

- Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J. and Wallach, H. A reductions approach to fair classification. *In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research*, **80**, 60–69, 2018.
- Ali, M. *PyCaret: An open-source, low-code machine learning library in Python*. PyCaret version, 2020.
- Equality Act 2010. Available at: `https://www.gov.uk/guidance/equality-act-2010-guidance`.
- Hardt, M., Price, E. and Srebro, N. Equality of opportunity in supervised learning. *Advances in neural information processing systems* **29**, 2016.
- Lima Neto, E.A., Bailiss, J., Finke, A., Miller, J and Cosma, G. Identifying early help referrals for local authorities with machine learning and bias analysis. *Journal of Computational Social Science*, 1-19, 2024.

Mercy !!!

Additional information

The following metrics will be used to compare the predictive performance between the models:

- Accuracy: measures how many observations, both positive and negative, were correctly classified.
- Area under the curve (AUC): represents an aggregated metric that evaluates how well a classification model classifies positive and negative outcomes.
- Precision: represents the number of true positives divided by the number of true positives plus the number of false positives. False positives are cases the model incorrectly labels as positive that are negative. Precision expresses the proportion of the data points our model says existed in the relevant class that were indeed relevant.
- Recall: represents the number of true positives divided by the number of true positives plus the number of false negatives. False negatives are cases the model incorrectly labels as negatives that are positives. Recall expresses the ability to find all relevant individuals of a class in a data set.
- F1-score: represents the harmonic mean of precision and recall. It combines precision and recall into a single value. The higher the precision and recall, the higher the F1-score.

Additional information

Let us consider the true labels (individuals that did not receive EH SUPPORT = 0 and individuals that received EH SUPPORT = 1) and the respective predicted labels (0 and 1). The decision-making of an ML algorithm produces FOUR decisions:

- The true label is 0 and the model classifies the individual as 0 (**True negative**)
- The true label is 1 and the model classifies the individuals as 1 (**True positive**)
- The true label is 0 and the model classifies the individual as 1 (**False positive**)
- The true label is 1 and the model classifies the individual as 0 (**False negative**)

Additional information

In the LCC scenario, we could consider two metrics for quantifying harms/benefits:

- false negative rate: fraction of individuals that require EH SUPPORT but the models classified that they do not require; this quantifies harm
- selection rate: overall fraction of individuals that are recommended for EH SUPPORT; this quantifies benefit; here the assumption is that all patients benefit similarly from the EHS.

LIME (Local Interpretable Model-agnostic Explanations)

- Lime can explain any "black-box" classifier, with two or more classes. By interpreting the weights of each feature it is possible to identify the most/less relevant features for a given class/category and the impact of those features on the predicted probability.