LOCATION INFORMATION AND RACIAL BIASES IN CREDIT SCORING

Anonymous authors
Paper under double-blind review

ABSTRACT

We design a series of experiments on credit scoring and employ SHAP values to demonstrate that the use of location information may introduce racial biases. The analysis relies on race statistics collected from Brazilian Institute of Geography and Statistics (IBGE) and on fully anonymized credit information. We argue against using location information for credit scoring and discuss how to track racial biases when protected attributes are not available.

1 CONTEXT

Even before the current wave of interest on fairness (see Gajane & Pechenizkiy (2017 - https://arxiv.org/abs/1710.03184)), the machine learning community has been focusing on the risk of social discrimination in automated credit scoring systems of Governors of the Federal Reserve System (2007). The interest is justified by the obvious social implications involved in credit worthiness assessment, with impacts ranging from access to consumption and investment opportunities to home ownership, education and health. It has now become clear that the pattern recognition capabilities made possible by the large scale use of machine learning technologies can easily become an amplifier of socioeconomic inequalities, with credit systems being at the core of this process.

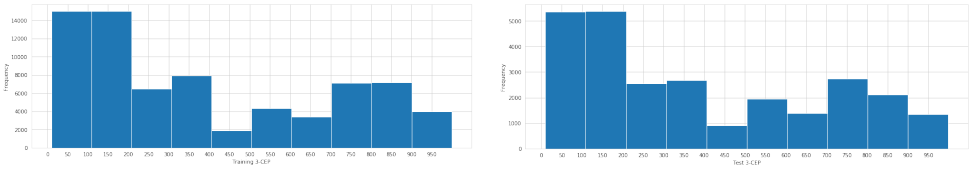
A credit assessment system relies on designing scores that can effectively rank counterparts according to their probability of default. Modern machine learning techniques allow training complex and powerful classifiers Chen & Guestrin (2016) at the cost of less explainable outputs. The standard score development consists of a manual feature engineering stage followed by automatized feature selection, that may be conducted by ranking features, almost exclusively, according to their importance for the performance of the classifier.

Here we exemplify the dangers of the common practice by building an experimental credit system based on gradient boosting decision trees, explained by SHAP values Lundberg et al. (2018 - https://ui.adsabs.harvard.edu/abs/2018arXiv180203888L). In particular we, discuss the use of location information, represented in our experiment by the first 3 digits (out of 8 digits) of the Brazilian postal code (CEP-3). Those 3 digits specify geographical regions larger than individual neighbourhoods and smaller than whole states, depending on the postal granularity of each region. When used for credit scoring, CEP-3 ends up among the most important features performance-wise. However, we show that this feature implicitly introduces severe racial biases into the model.

2 THE DATA

The data used to build the model consisted of a sample of size 98 698. Each entry was individually identified by an anonymized ID and a data collection date. Collection dates ranged from April 1st, 2017 to March 31st, 2018. The training dataset was composed by 72 271 samples collected before January 1st, 2018, the remaining 26 427 samples were used as the test dataset. Payment delays longer than 90 days within a one year period after the reference date were annotated as a credit event and used as targets.

Each sample consisted of ten features: age, financial behavior indicators and CEP-3. Figure 1 shows the CEP-3 distribution both in training and test datasets. Figure 2 shows Brazilian regions represented by the first digit of the CEP-3. Regions starting with 0 and 1 correspond to of the state of São Paulo (SP in the map), the most populated region and also the most represented in the dataset.



(a) Training data (b) Test data

Figure 1: 3-CEP distribution by first digit

Figure 2: Regions corresponding to first digit of CEP-3. The South region has 21.5% of its population composed by self-declared not-whites, Southeast has 43.3%, Northeast, 71.2% and North has 76.4%.

In our study we also used racial data from the Brazilian Institute of Geography and Statistics (IBGE) and a proprietary dataset with the spatial distribution of default rates based on 10 million samples.

3 EXPERIMENTS

The model we implemented was gradient boosting trees trained with XGBoost Chen & Guestrin (2016). The model employed a regularized binary logistic objective function. The final model had 250 trees of maximal depth of 3. The model attained a performance of 0.76 and 0.74 of ROC-AUC in train and test, respectively. We then used SHAP-values for analyzing the behavior of the model.

SHAP-values allows one to assess how the trained model uses the features it is fed with, ascribing a number to the intensity of the impact of each feature in each individual prediction while also indicating if the feature drove the prediction up or down from the average model output. It does so by taking a weighted account of model output using all possible combinations of presence and absence of features. For an input vector $\mathbf{x}$ and a model f, the impact ϕ_i of the i-th feature of $\mathbf{x}$ on the model's output $f(\mathbf{x})$ is given by

$$\phi_i(f, \mathbf{x}) = \frac{1}{|N|} \sum_{S \subseteq N \setminus \{i\}} \binom{|N|-1}{|S|}^{-1} \left[f(\mathbf{x}_{S \cup \{i\}}) - f(\mathbf{x}_S) \right], \tag{1}$$

where N is the number of features, S represents subsets of features that do not include the i feature and $\mathbf{x}_S$ represents a vector containing only the input features in the set S. As a proxy to the model's output when it should have access only to a subset of the features, the technique proposes that $f(\mathbf{x}_S)$ should be evaluated as $E[f(x)|x_S]$ (see Lundberg & Lee (2017) for more on this). We can see the result on figure 3, where each individual sample is represented by one point in each line. The place of the point on the horizontal axis tells the impact of the feature on the model's output: positive SHAP values indicates the feature value drove the model's default probability assessment up and negative SHAP values means it drove the default probability down. The color that each scatter point receives refers to the feature value inside the sample used to build the plot - relatively high values are red

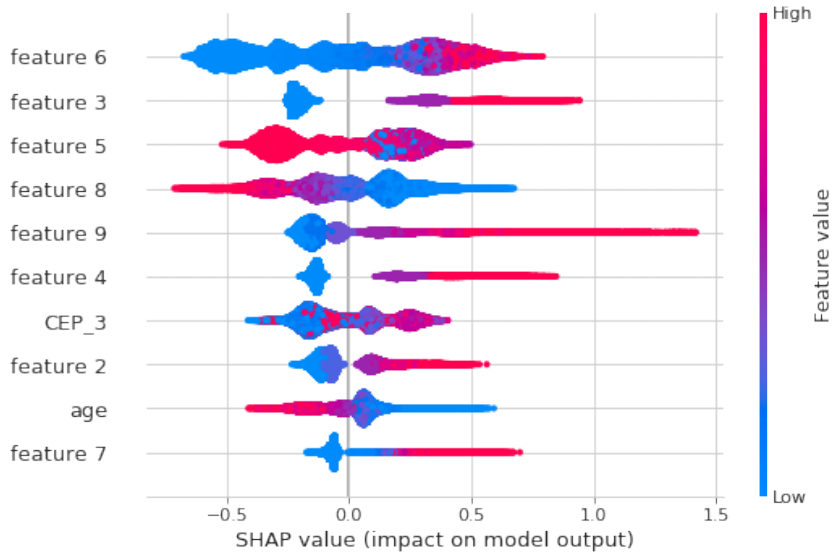
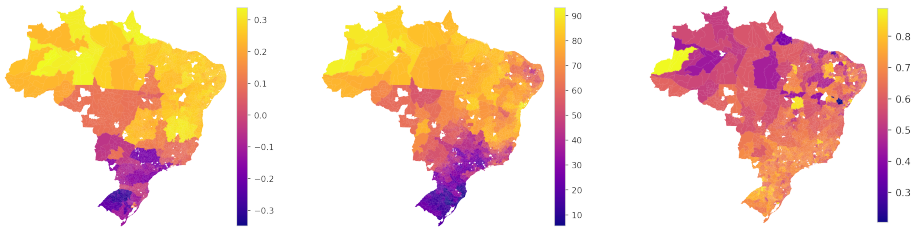


Figure 3: Summary feature impacts over the test dataset. For example, typically, high values of "age" drive the probability of default down (and the credit score up), while low values of "age" do the opposite.

while relatively low values are blue. It is interesting to note that almost all features in our credit model behave monotonically, namely, either high (low) values of the given feature drive the score up (down) or the exact opposite. In our scenario, this monotonic behavior on SHAP values is due to the use of preprocessing techniques devised to be optimal for linear models. The location variable, CEP-3, emerges among the most important features.



(a) Average SHAP Values by CEP-3.

(b) Self-declared non-whites by CEP-3.

(c) Spatial distribution of mean scores.

Figure 4: Spatial distributions of SHAP values, racial composition and mean scores.

Figures 4a and 4b show a strong correlation between scores and racial composition: regions with a majority of self-declared whites of Geography & IBGE (2010 - http://dados.gov.br/dataset/cgeo_vw_per_pessoas_brancas) tend to have higher scores. In fact, the SHAP-values of the CEP-3 region and the percentage of self-declared non-whites[1] population exhibits a Pearson's correlation of 82% and arithmetically normalized mutual information of **99%**. This means that CEP-3 is a very good proxy of racial composition.

A look at how the score distribution would be seen on the map points to how important SHAP-values is for identifying racial bias in this experimental model. While the pattern of feature impact makes

[1]Which includes people of African descent, Indigenous people and people of Asian-descent. The African-descent Brazilian population corresponds to 50.94% of the total population

the racially harmful effects self-evident, a simple look at the mean score distribution would not be sufficient to raise concerns about the model as figure 4c reveals.

Two further experiments make the case stronger. First, we built counterfactual examples by simulating what would happen to the score of people living in the countrie's Southeast region if they moved to the Northeast region, age and financial behavior kept the same. Figure 5 we show that only in **0.15%** the credit scores would not decrease.

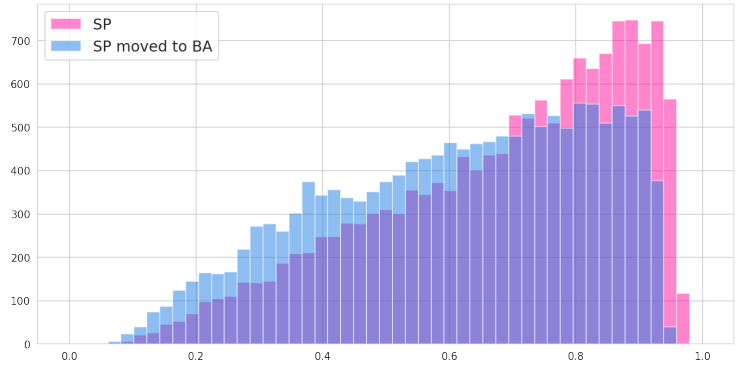


Figure 5: Counterfactual simulation: moving from São Paulo (SP, Southeast) to Bahia or Sergipe (BA, Northeast).

Second, the very high normalized mutual information between the impact of CEP-3 on model decisions and the racial distribution indicates that a one can build a model as good by replacing CEP-3 with the percentage of non-whites living in the associated region. The result is a model with the same performance (train ROC-AUC: 0.76, test ROC-AUC: 0.74). The explanation outputs of the other features did not change and the new racial variable takes exactly the place of CEP-3, with the same importance, as can be seen in figure 6. This relates intuitively to the fact that the feature's model impact did not show much interaction with any of the other features' model impact. The difference this time was the monotonic behavior of the impact of the new feature: the greater the percentage of white people living around, the better the score.

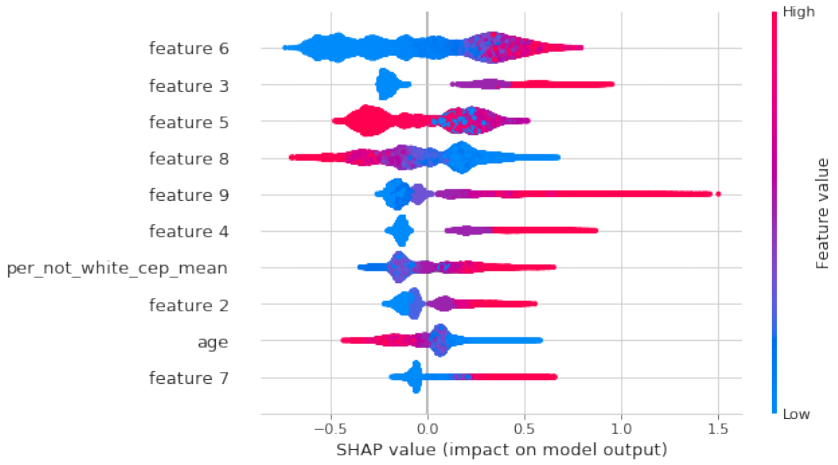


Figure 6: Feature impact summary for a model using regional racial composition as a feature.

We also calculated the geographical distribution of default rate using a large anonymized dataset (10 million samples) that was not used in the original experimental setup. The normalized mutual information between this default rates and the distribution of whites is again very high (**95%**) - and

Pearson's correlation coefficient of 60%. So it is very likely that any model built using location would exhibit the same sort of racial bias and consistently decrease scores of people living in predominantly not-white regions and would feed back historical disparities. Nevertheless, things get more interesting once we compare performances in two regions with distinct racial makeups.

We compared performances conditioned to location. The city of São Paulo (Southeast) has around 40% of non-whites, while the Northeast is 71% non-white. The model performances were 0.74 (train) and 0.72 (test) when conditioned on data from the Northeast and of 0.77 (train) and 0.75 (test) when evaluated on data from the Southeast. As figure 7 shows, the model shows a worse true positive rate in Northeast than in São Paulo for every approval threshold.

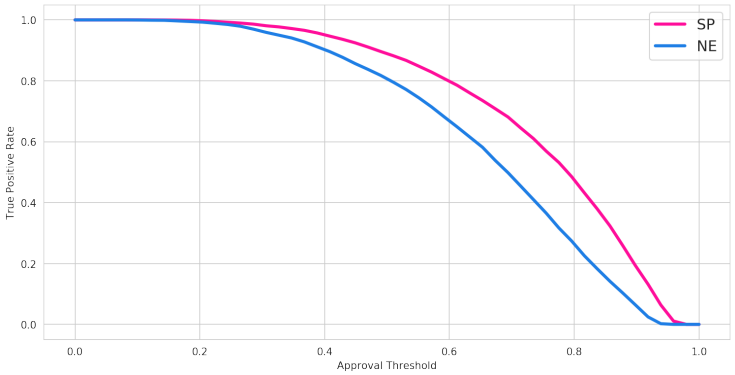


Figure 7: True positive rates for different thresholds for data from the Northeast and from São Paulo.

4 Conclusion

Often, when a machine learning model is found to be unfair, people rush to point fingers at the data used to build the model raising the question if better data collection would be the way to control negative implications. Here, we provided an example of how a naive use of location information can be harmful to fairness by reinforcing biases against historically unprivileged groups while underperforming in these populations.

In the modern world, credit is key for a satisfying life. We see as worthwhile goal to find ways to build models that are fair, performing equally in different populations, and are also positively discriminatory, allowing for the maximum credit offer with minimum risk.

REFERENCES

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning, 2017 - https://arxiv.org/abs/1710.03184.

S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv e-prints*, 2018 - https://ui.adsabs.harvard.edu/abs/2018arXiv180203888L.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Brazilian Institute of Geography and Statistics IBGE. *Percentual de pessoas residentes de cor ou raça branca*, 2010 - http://dados.gov.br/dataset/cgeo_vw_per_pessoas_brancas.

Board of Governors of the Federal Reserve System. *Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit*, 2007. URL https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/creditscore.pdf.