# Explanations and Trustworthiness in Salary Analysis Classification: A case study

Valentina Ortega Pinto[1] and Martín Romero Romero[1]

University of Santiago de Compostela
{valentinaisabeldelamilagrosa.ortega,martin.romero.romero}@rai.usc.es

**Abstract.** This study analyzes the classification of salary analysis, highlighting the importance of explainable artificial intelligence models and reliable in considering sensitive attributes, such as gender, race or age, to guarantee equity and transparency in the structures organizational salaries. A variety of tools where employed, including InterpretML, SHAP values, and Explainable GAM, to gain insights into the influence of different attributes on our models' predictions.The results show that, although some models exhibit acceptable levels of fairness, challenges remain in mitigating bias, particularly in attributes such as educational level, pointing out the need for additional strategies to address these disparities.

**Keywords:** Explainable AI · Trustworthy AI · Fairness and Bias in AI · Sensitive Attributes · Salary Analysis Classification.

## 1 Introduction

In the world of workforce dynamics, the application of machine learning models to analyze and classify salaries has become an invaluable tool for organizations striving to ensure fairness and transparency in their compensation structures. However, as these models delve into sensitive attributes such as gender, race, or age, the need for ethical considerations, explainability, and trustworthiness becomes paramount.

This work delves into the intricate landscape of salary analysis classification, exploring the challenges posed by sensitive attributes and emphasizing the significance of explainable and trustworthy machine learning models. By navigating these considerations, organizations can not only comply with legal requirements but also proactively contribute to creating workplaces that champion diversity, equity, and inclusion.

## 2 Technical Issues

### 2.1 The Dataset

The Salary Prediction Classification dataset [1] , extracted by Barry Becker from the 1994 Census database, is designed for the task of predicting whether an individual earns over or under $50,000 annually. The dataset encompasses diverse socio-demographic and professional attributes, making it a comprehensive resource for exploring patterns in income distribution. Its features are:

- Age
- Workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- Fnlwgt
- Education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- Education-num: Continuous.
- Marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- Occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- Relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- Race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- Sex: Female, Male.
- Capital-gain: continuous.
- Capital-loss: continuous.
- Hours-per-week: continuous.
- Native-country: Includes several countries such as United-States, Cambodia, England, Puerto-Rico, Canada, ect.
- Salary: Higher or lower than 50k.

The dataset is a mixture of continuous and categorical variables, providing a rich source of information for building and evaluating classification models.

### 2.2 Dataset preparation and exploration

To lay the groundwork for our model development, our initial step involved a comprehensive analysis of the dataset. After a thorough evaluation, we made a strategic decision to omit certain columns from our analysis: capital-gain, capital-loss, education, and fnlwgt. Our decision to exclude these specific attributes was not made lightly. In particular, the choice to remove education; despite its apparent relevance; was based on a careful consideration of redundancy. We observed that the education-num column encapsulates the essence of the education attribute but in a numerical format that is more conducive to our models' processing needs.

The distribution of our target in the dataset presents a significant imbalance: 76% of the samples belong to individuals earning less than 50K dollars per year, while the remaining 24% are attributed to those earning more than 50K dollars. This skewed distribution poses a critical challenge in terms of model trustworthiness and systematic performance bias. This imbalance can lead to a model that is not only less accurate in identifying individuals with higher earnings but may also inadvertently reinforce systemic biases present in the data. Addressing this imbalance is crucial for developing a model that is both fair and reliable. Techniques such as resampling the dataset, are essential steps to mitigate these issues and enhance the trustworthiness of our predictive models.

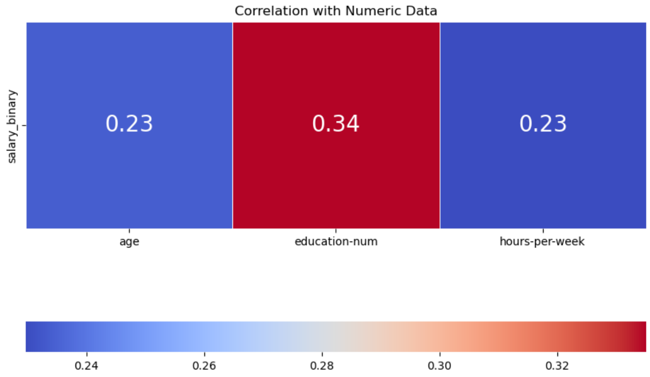

**Fig. 1.** Correlation of Salary with Numeric Data

To gain preliminary insights into which attributes might exert a more substantial influence on our model's predictions, we conducted a thorough analysis by generating a correlation matrix for the numerical variables and computing Cramer's V coefficient for the categorical variables. This analytical approach allowed us to quantitatively assess the relationships between each attribute and the model's output. As illustrated in Figure 1 and Figure 2, it became evident that 'education', 'relationship', and 'marital-status' attributes notably impact the model's predictions more than others. This observation not only guides us in identifying factors with potentially significant predictive power but also alerts us to areas where potential biases and discrimination may arise. By spotlighting these attributes, we underscore the necessity of a cautious approach in model development, aiming to mitigate bias and ensure a fair and trustworthy predictive framework.

In the process of preparing our dataset for modeling, particularly for algorithms that require strictly numerical input, we encountered the challenge of handling categorical attributes. Many of these attributes featured a broad array of categories without any inherent ordinal relationship. To address this, we opted for target encoding as our strategy for converting categorical data into a numerical format. This method was chosen because it not only efficiently condensed the vast categorical information into a numerical form that models can process

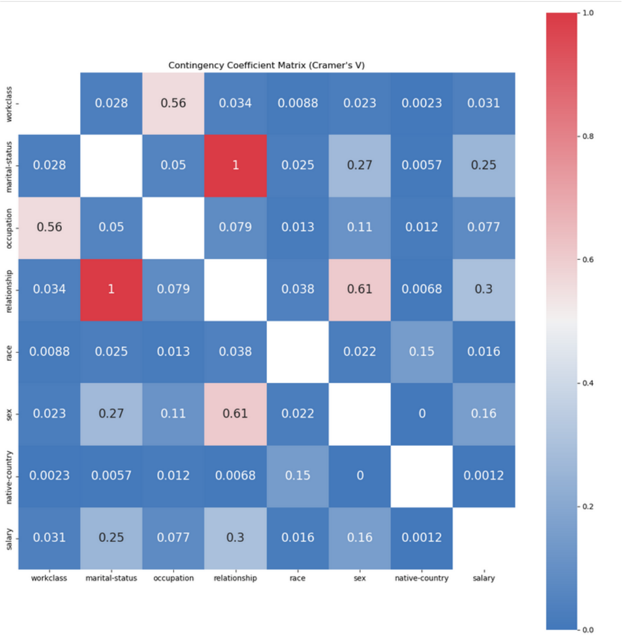

**Fig. 2.** Correlation of Salary with Categorical Data

but also did so in a way that preserved the underlying relationship between each category and the target outcome.

## 2.3 Models and Tools

### Explainability

Our initial approach focused on selecting a variety of models amenable to analysis via explainability tools [4]. These included the Explainable Boosting Machine (EBM), Gradient Activation Maps (GAM) Classifier, various Tree Classifiers, and a Random Forest Classifier. As detailed in Table 1, we meticulously recorded both training and test accuracy for these models to assess their performance rigorously. Among the models evaluated, the Tree Classifier with a predetermined depth of 5 emerged as the standout performer, boasting an impressive test accuracy of 83%. In contrast, the Full Tree Classifier, which does not restrict tree depth, lagged behind with the lowest test accuracy observed at 78%.

To describe the workings of the Explainable Boosting Machine (EBM), we used InterpretML, a comprehensive open-source Python package specifically designed to enhance the interpretability and comprehensibility of machine learning models. Our objective was to provide a global explanation focusing on the significance of various features in the model's predictions. By utilizing InterpretML,

**Table 1.** Training and Test Accuracy across Models.

| Model | Training Accuracy | Test Accuracy |
|---|---|---|
| EBM | 0.841 | 0.847 |
| GAM | 0.840 | 0.845 |
| TREE | 0.782 | 0.781 |
| TREE5 | 0.830 | 0.833 |
| RF | 0.823 | 0.826 |

we were able to systematically identify and illustrate the pivotal roles that features age, marital status, and education level played in influencing the model's decision-making process. As depicted in Figure 3, the analysis clearly highlights the disproportionate impact of these features, underscoring their critical importance in the predictive accuracy of the EBM. On a reassuring note, this initial review suggests that this model does not overly rely on race or gender to influence its predictions. An undue emphasis on these attributes could introduce potential biases and discrimination, compromising the model's fairness and ethical standards.

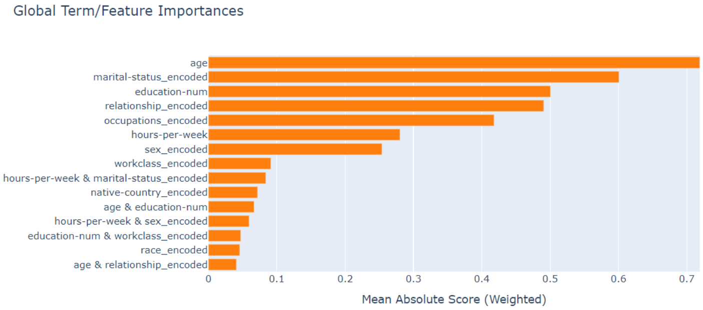

**Fig. 3.** Feature Importance from InterpretML for EBM

To delve deeper into the intricacies of the XGB Classifier, we employed SHAP (SHapley Additive exPlanations) values, a powerful tool for understanding the contribution of each feature to the model's predictions. Our analysis, visually encapsulated in Figure 4, reveals a consistent pattern across various models, with age and education emerging as significantly influential factors in shaping the model's outcomes. Notably, for the XGB Classifier, the relationship attribute also stood out as a key determinant, exerting a considerable impact on the predictions.

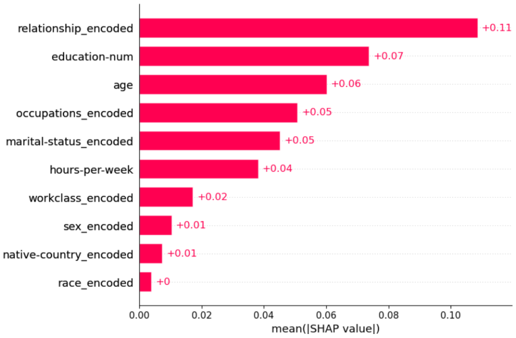

**Fig. 4.** SHAP Values for XGB Classifier

For the Gradient Activation Maps (GAM) Classifier, our analysis extended to investigating counterfactual explanations, providing insights into how slight changes in input features could affect predictions. The outcomes of this exploration are detailed in Figure 5, where we observe a notable trend: the model's predictions are less influenced by sensitive features such as race and gender, compared to more significant determinants like education and age. This pattern reaffirms the model's focus on attributes that are more directly relevant to the predictive task, reducing concerns over bias linked to sensitive demographic characteristics.

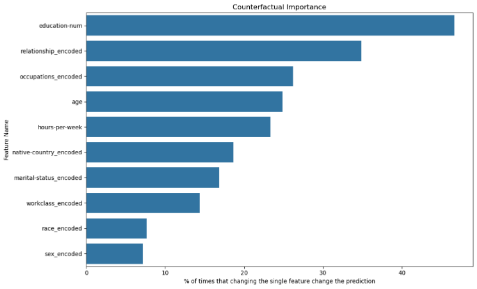

**Fig. 5.** Counterfactual Importance for GAM

## Trusworthiness

In this section, our objective is to meticulously assess the dataset for any potential biases across its various characteristics. Should we identify any such biases, our goal will be to actively minimize their impact within the models we develop..

To detect if there is bias, we have used the ”fatf” library and we have applied the sampling bias detector and systematic performance bias to our dataset and to a model trained with this dataset, paying special attention to the features of the dataset that we believe may contain more bias. Specifically, we focused on the following characteristics: gender, race, education and age.

- *Sex*
In the characteristic ”sex” sampling bias has been found, indicating that the data sample does not represent men and women equally. This could be due to different reasons, such as imbalances in the original population, biases in data collection or in the way individuals were selected to form the sample. As for systematic performance bias, no bias has been detected.

- *Race*
For the characteristic ”race” a sampling bias was found, indicating that the distribution of the different racial categories in the sample does not accurately reflect the racial composition of the general population. This bias can arise due to a variety of factors, such as lack of adequate representation of certain racial groups in the sample selection process, discrimination in data collection, or under representation of certain communities. It is crucial to address this bias to ensure that the model does not perpetuate inequities or discrimination based on race and that its predictions are equitable and accurate for all racial communities represented in the population of interest. As for systematic performance bias, no bias has been detected.

- *Education*
In the characteristic ”education” a sampling bias has been detected, suggesting that the sample does not adequately reflect the distribution of educational levels in the general population. This could be problematic, since people’s characteristics and behavior may vary significantly according to their educational level. A bias in this variable could lead to erroneous conclusions or unreliable models. In addition, ”systematic performance bias” has been found, suggesting that the model may have difficulty generalizing effectively across all educational levels, which could be due to the lack of equal representation of each educational level in the sample or significant differences in behavior between these groups.

- *Age*
For the characteristic ”age” we found ”sampling bias” indicating that the distribution of ages in the sample does not match the distribution of ages in the underlying population. This could be due to non-random selection of individuals for the sample or to specific age subgroups that are over- or under-represented in the sample. Systematic performance bias has also been found, suggesting that model performance varies with the age of individuals in the data set. This could indicate that the model may not be equally effective in predicting outcomes for all ages, which could be due to differences in behavior or relevant characteristics

between different age groups.

After analyzing the presence of bias, we have calculated the percentage of the population earning more than 50K divided by the above sensitive characteristics, and we see that there is quite a difference depending on the group to which one belongs, as the presence of ”sampling bias” indicated. For example, in Figure 6, where we can see the fraction of people earning more than 50K as a function of educational level.

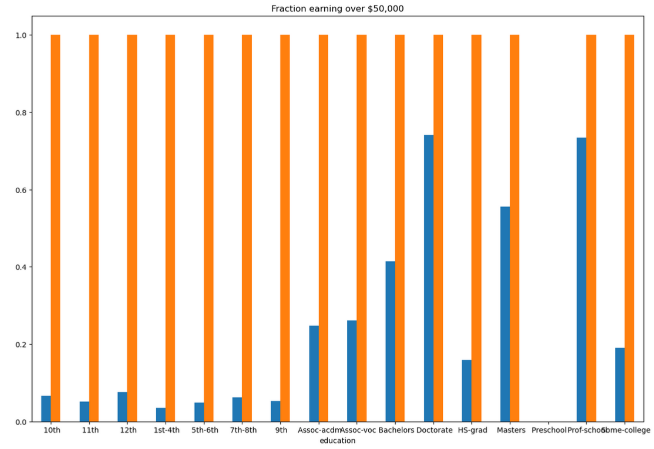

**Fig. 6.** Fraction over 50K distributed by education

Once the data have been analyzed, and the presence of bias has been observed, it is inevitable that the models trained with these data will also contain the biases of the dataset. Therefore, based on the sensitive characteristics, we have tried to minimize the presence of these biases by training models using the equity criterion of ”Demographic Parity” as a restriction. The models trained with this criterion, have obtained a lower amount of bias than the models that have not been trained without this criterion. Moreover, there was hardly any loss of accuracy in the models obtained, as can be seen in Figure 7.

The only exception has been in the model that has tried to minimize bias by using education as a sensitive characteristic. Specifically, the supposedly ”less biased” model shows a much larger bias than the original model and suspiciously with ”more bias”. We cannot be totally sure of the reasons, but we believe that the original model relies quite heavily on the information provided by the educational level of the person, and by minimizing the bias with this characteristic, it relies more on characteristics such as sex or race, being a model with more bias.

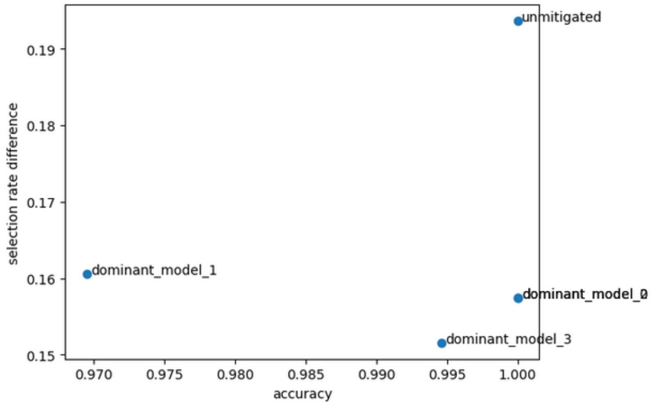

**Fig. 7.** Relation between accuracy and selection rate difference with race

## 3 Ethical, Legal and Socio-economic Issues

For our analysis, we meticulously evaluate our models using the seven critical aspects outlined in the Assessment List for Trustworthy Artificial Intelligence (ALTAI):

### 3.1 Ethical Considerations

Human Agency and Oversight (1): Our approach ensures support for human autonomy by allowing human oversight in the model's development and deployment phases. Our decision in using interpretable and explainable models and tools; suchs as InterpretML, SHAP Values and counterfactual explanations; upholds the principle of respect for human autonomy, enabling stakeholders to understand and, if necessary, contest or make informed decisions based on the model's predictions.

Diversity, Non-discrimination, and Fairness (5): Upon thorough investigation, it became evident that our models exhibit a commendable level of fairness regarding demographic attributes such as native-country and race, and similarly uphold impartiality concerning gender. Nonetheless, we identified a discernible bias pertaining to education level, indicating a degree of discrimination that warrants attention. In response to this finding, we undertook measures to address and mitigate this bias, specifically through the strategies of relabeling and reweighting [2]. Despite these efforts, the outcomes fell short of our expectations, revealing the complexity and stubborn nature of the bias embedded within the education-related attributes.

### 3.2 Legal and Socio-Economic Considerations

Privacy and Data Governance (3): This dataset, has taken stringent measures to safeguard the privacy of individuals by not disclosing any personally identifiable information. This ensures that the identities of the participants remain confidential, reflecting the commitment to upholding privacy.

Accountability (7): Our project embeds accountability mechanisms by transparently documenting the model's development process, including choices made to avoid or mitigate bias. This transparency ensures that the project can be audited and understood by third parties, aligning with legal requirements for accountability in AI systems.

### 3.3 Socio-Economic and Cultural Issues

Technical Robustness and Safety (2): The project relies on model accuracy in order to achieve technical robustness and safety. While we acknowledge that the test accuracy achieved by our models leaves room for improvement, it's important to note that enhancing this metric falls beyond the current project's scope. Nonetheless, pursuing higher accuracy levels presents a valuable avenue for future work, offering the potential to significantly bolster the models' reliability and safety.

Transparency (4): Our project's focus on explainability and transparency, particularly through the use of tools like SHAP values and InterpretML, ensuring traceability and open communication about the model's limitations [3].

## 4 Conclusions

The utilization of explainability tools, such as InterpretML, SHAP, and Explainable GAM, shed light on the significant influence of attributes like education and age on model predictions. This insight enabled us to delve deeper into the model's trustworthiness, verifying our initial suspicions. Upon examining for sampling bias, it was found that the early versions of our models exhibited this type of bias in terms of sex, race, education, and age. However, systematic performance bias was predominantly observed in education and age. In efforts to address these biases, our mitigation strategies successfully reduced bias concerning age, race, and gender. Despite these successes, we encountered challenges in effectively neutralizing bias within the education attribute, probably due to the task's nature.

## References

1. Salary Prediction Classification. Available at https://www.kaggle.com/datasets/ayessa/salary-prediction-classification
2. Kamiran, F., Calders, T. Data preprocessing techniques for classification without discrimination. Knowl Inf Syst 33, 1–33 (2012). https://doi.org/10.1007/s10115-011-0463-8
3. Ferrara E. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. Sci. 2024; 6(1):3. https://doi.org/10.3390/sci6010003
4. Teso S, Alkan Ö, Stammer W, Daly E. Leveraging explanations in interactive machine learning: An overview. Front Artif Intell. 2023 Feb 23;6:1066049. https://doi.org/10.3389/frai.2023.1066049. PMID: 36909207; PMCID: PMC9995896.