



# **INFO-H420-MANAGEMENT OF DATA SCIENCE AND BUSINESS WORKFLOWS**

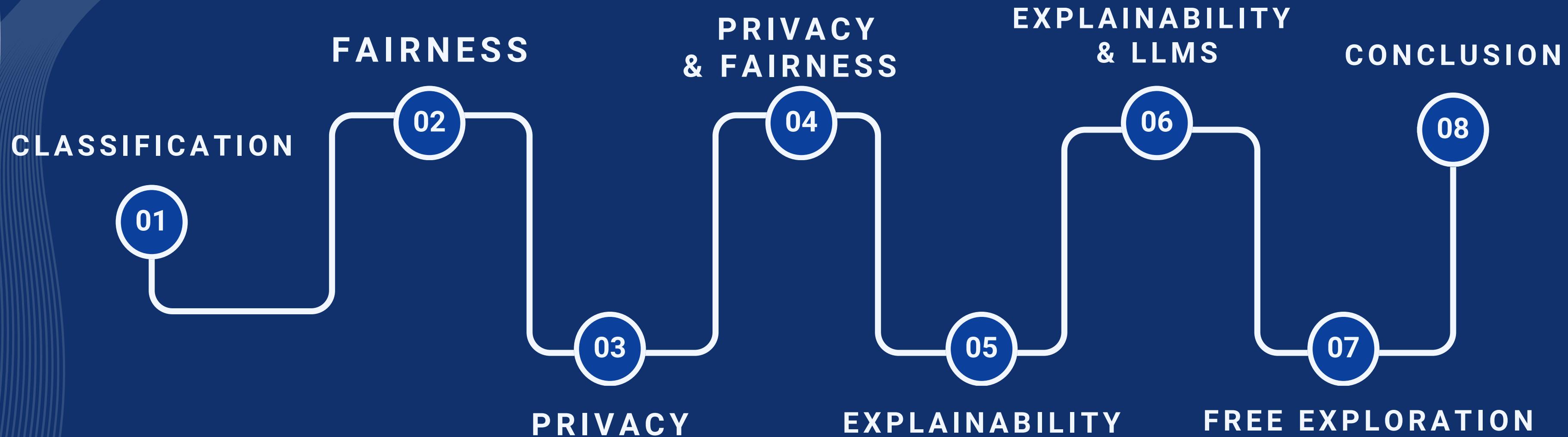
**AUTHORS :(GROUP P.13)**

**PHAM DANG PHI L. , ELEZI HERMA , BOUDJEMA MEHDI,  
COQUEREAU ARISTIDE, SY MOHAMED, DEMONCEAU QUENTIN**

**17 December, 2024**



# TABLE OF CONTENT



# CLASSIFICATION



## OBJECTIVE

To train the data and establish a baseline performance of the model using standard metrics.

## KEY TECHNIQUES

- **Data preprocessing:** cleaning and age binarization
- **Data splitting :** 70% Train, 15% Validation and 15% Test sets
- **Data scaling :** Standardization using StandardScaler
- **Logistic Regression:** Used as the classifier for the model training.
- **Performance metrics :** Accuracy - Precision - Recall - F1-score

# RESULTS

```
Performance Metrics: {'Accuracy': 0.8552712384851586, 'Precision': 0.7271750805585392, 'Recall': 0.5991150442477876, 'F1 Score': 0.6569626394953906}
```

## Classification Report

	precision	recall	f1-score	support
False	0.89	0.93	0.91	3755
True	0.73	0.60	0.66	1130
accuracy			0.86	4885
macro avg	0.81	0.77	0.78	4885
weighted avg	0.85	0.86	0.85	4885

## Confusion Matrix

```
[[3501 254]
 [ 453 677]]
```

## CLASSIFICATION REPORT

Breakdown of precision, recall, and F1-score for each class (False and True), along with the support (number of instances) for each class.

## CLASSIFICATION REPORT

shows the number of true positive, true negative, false positive, and false negative predictions.

76.63

# FAIRNESS



## Objective:

- Assess group fairness in a machine learning classifier predicting income ">50K" based on the **Adult dataset**.
  - Focus on ensuring equitable treatment for individuals across groups defined by protected attributes.
- 
- **Protected Attributes:**
    - Age: Split into binary categories (e.g., younger vs. older).
    - Sex: Binary attribute (Male/Female).
  - Evaluate how bias in the classifier affects different demographic groups.

## Fairness Metric:

- Demographic Parity Difference: Measures the disparity in the probability of positive outcomes across groups (e.g., younger vs. older or male vs. female).
  - Value close to 0 indicates fairness.
- **Demographic Parity Ratio:** Ratio of positive outcome probabilities across groups.
  - Value close to 1 indicates fairness.
- Why these metrics?
  - Provide insights into representation disparity.
  - Are simple and interpretable for auditing.

# RESULTS

Fairness Metrics Comparison (Original vs Reweighted Classifier):  
Original Classifier \

Metric	
Demographic Parity Difference (Age)	0.246116
Demographic Parity Ratio (Age)	0.087387
Demographic Parity Difference (Sex)	0.172528
Demographic Parity Ratio (Sex)	0.304549

Reweighted Classifier

Metric	
Demographic Parity Difference (Age)	0.100166
Demographic Parity Ratio (Age)	0.519463
Demographic Parity Difference (Sex)	0.140902
Demographic Parity Ratio (Sex)	0.368751

Performance Metrics Comparison (Original vs Reweighted Classifier):

Original Classifier    Reweighted Classifier

Metric	Original Classifier	Reweighted Classifier
Accuracy	0.855271	0.838076
Precision	0.727175	0.696864
Recall	0.599115	0.530973
F1 Score	0.656963	0.602712

## ORIGINAL CLASSIFIER

- Age: Difference = 0.246, Ratio = 0.087
- Sex: Difference = 0.173, Ratio = 0.305

## FAIR CLASSIFIER (REWEIGHTED)

- Age: Difference = 0.100, Ratio = 0.519
- Sex: Difference = 0.141, Ratio = 0.369

## IMPACT

- Applying fairness techniques significantly improved parity metrics.
- Minimal reduction in model performance:
  - Accuracy: 85.53% → 83.81%
  - F1 Score: 65.70% → 60.27%



# PRIVACY

Cross tabulation on  
the original dataset

sex	Female	Male
age		
17	186	209
18	268	282
19	356	356
20	363	390
21	329	391
...	...	...
85	1	2
86	1	0
87	0	1
88	1	2
90	14	29

Local  
differential  
privacy

Cross tabulation on  
the private dataset

sex	Female	Male
age		
16	8	3
17	184	203
18	265	275
19	349	371
20	368	380
...	...	...
87	0	1
88	0	3
89	0	1
90	14	27
91	1	0

Differences in results  
between classifiers on  
the original and new datasets

	original dataset	private dataset
Accuracy	0.8552712384851586	0.8528147389969294
Precision	0.7271750805585392	0.7236126224156693
Recall	0.5991150442477876	0.588495575221239
F1 score	0.6569626394953906	0.6490971205466081

# PRIVACY & FAIRNESS

Fairness Metrics Comparison (Original vs Reweighted vs Private+Fair Classifier):			
	Original Classifier	Reweighted Classifier	Private+Fair Classifier
Metric			
Demographic Parity Difference (Age)	0.246116	0.100166	0.106653
Demographic Parity Ratio (Age)	0.087387	0.519463	0.495613
Demographic Parity Difference (Sex)	0.172528	0.140902	0.079074
Demographic Parity Ratio (Sex)	0.304549	0.368751	0.620213
Performance Metrics Comparison (Original vs Reweighted vs Private+Fair Classifier):			
	Original Classifier	Reweighted Classifier	Private+Fair Classifier
Metric			
Accuracy	0.855271	0.838076	0.841146
Precision	0.727175	0.696864	0.704861
Recall	0.599115	0.530973	0.538938
F1 Score	0.656963	0.602712	0.610832

## BIAS REDUCTION

Improved fairness metrics for age and sex.

## IMPACT OF CONSTRAINTS

Privacy enforces better generalization.

## SUCCESSFUL BALANCE

Combines fairness with privacy preservation.

Privacy constraints in the Private+Fair Classifier do not significantly degrade performance.

# EXPLAINABILITY



Study the explainability of the private classifier:

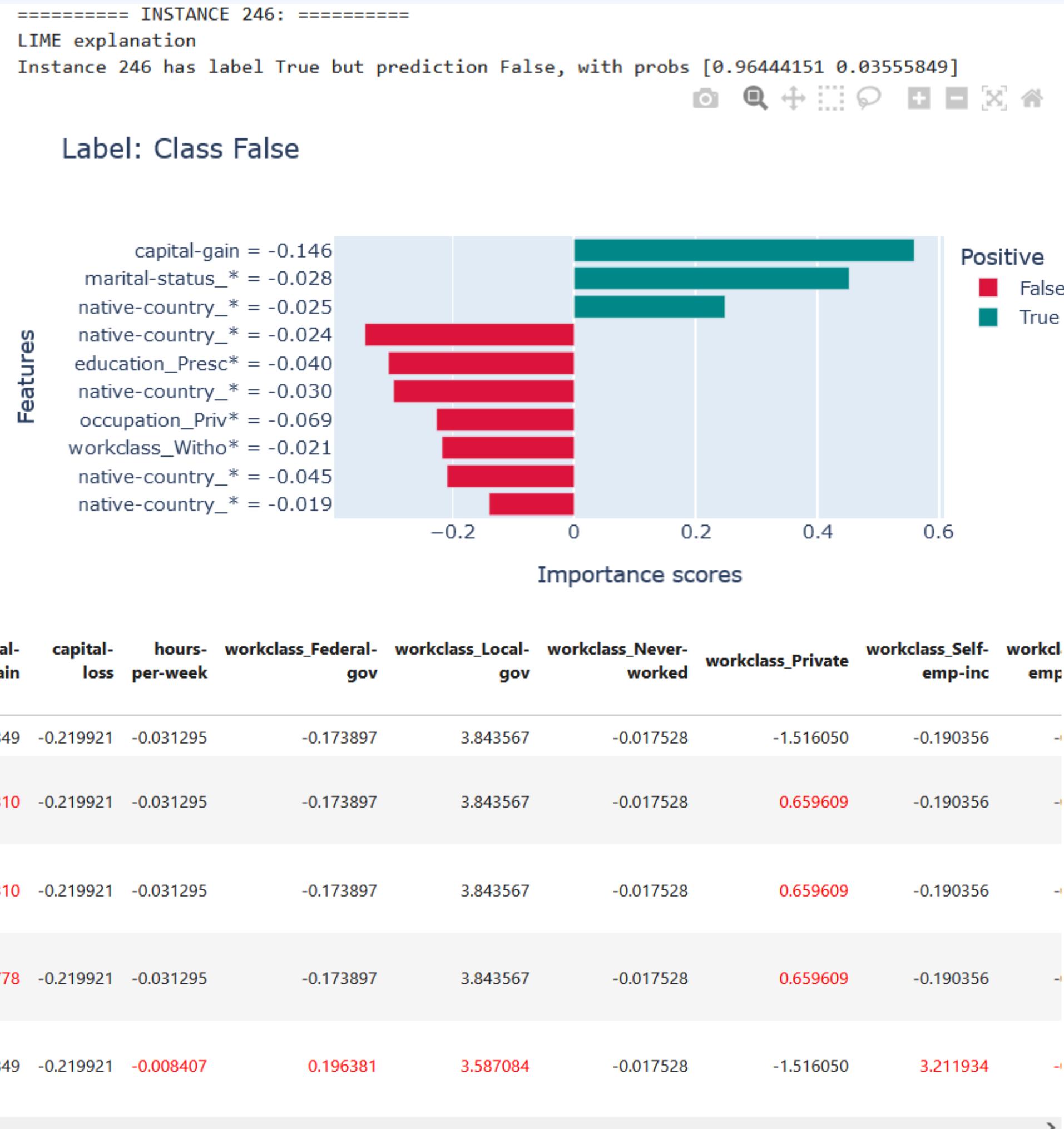
- Instances where the model is wrong but highly confident
- Explanation
  - LIME explainer
  - MACE explainer
- Are the noisy values for the sensitive values of Age and Sex attributes responsible for the model being confident and wrong?
  - Looking at all the test cases
  - Looking at only the cases where the model is wrong but confident

# **Instances where the model is wrong but highly confident:**

- Find instances where the predicted label is different from the actual label (→ missed instance)
- Create the prediction probabilities
- Find all missed instances that have a maximum probability above a set "high\_confidence\_threshold" (by default at 95%)

# Explanation:

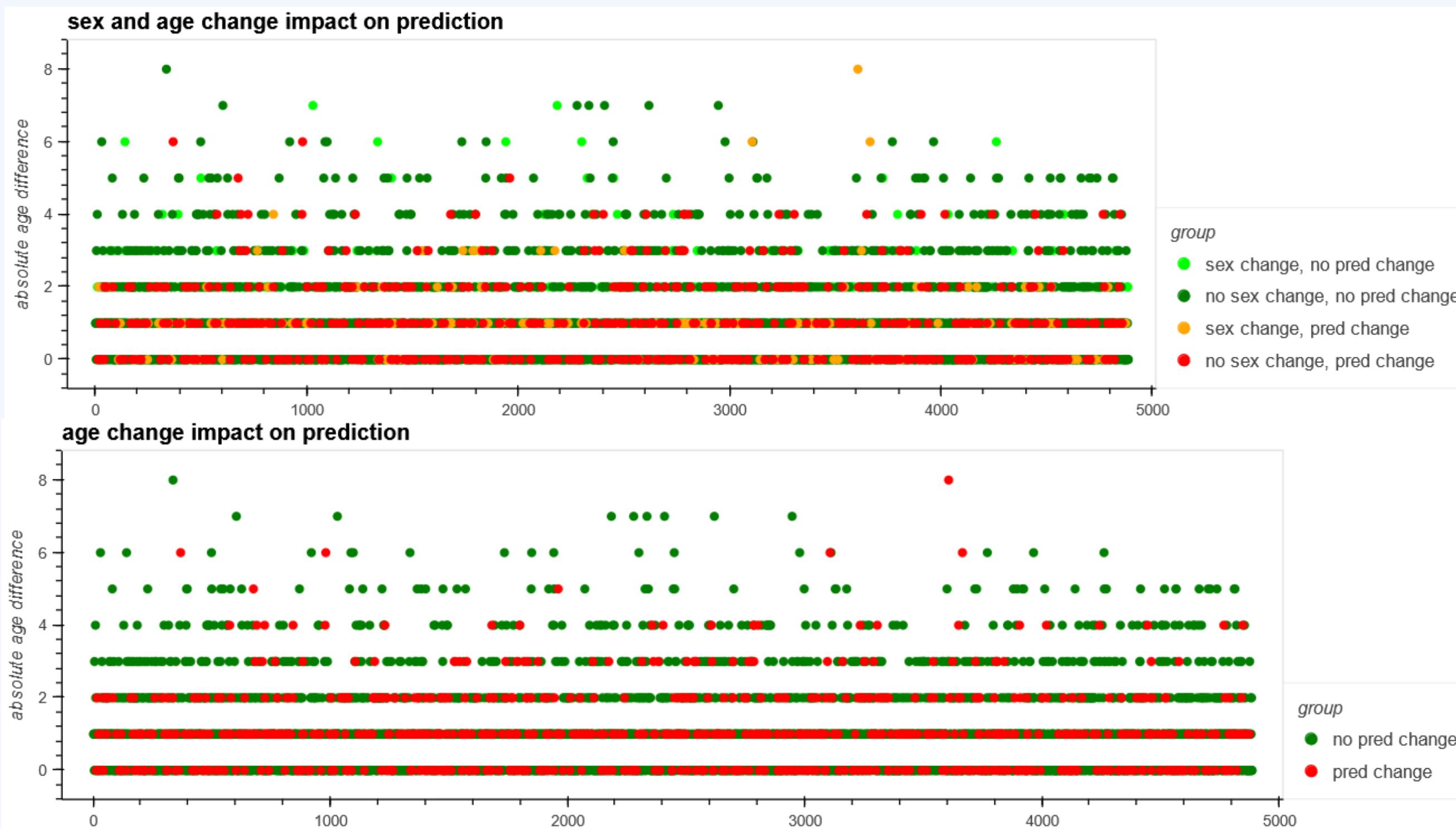
## LIME Explainer:



## MACE Explainer:

# Is the Age and Sex noise responsible for the model being confident and wrong?

Looking at all the test cases

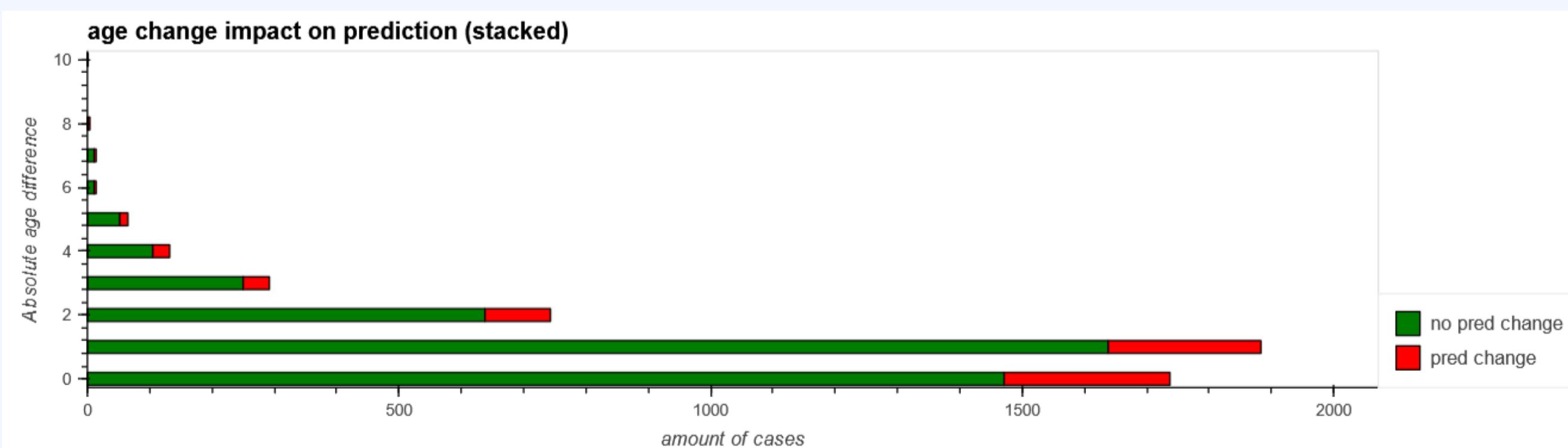
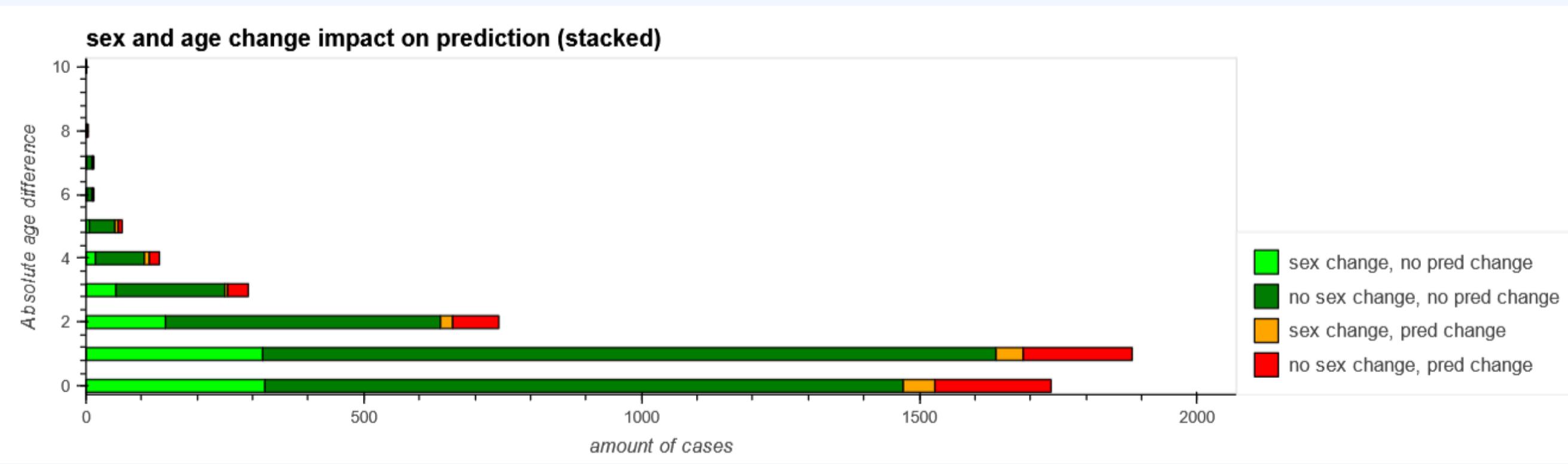


pred changes	no pred changes	age
285	1600	1
241	1516	0
38	258	3
22	110	4
121	608	2
4	17	6
2	53	5
1	1	8
0	8	7

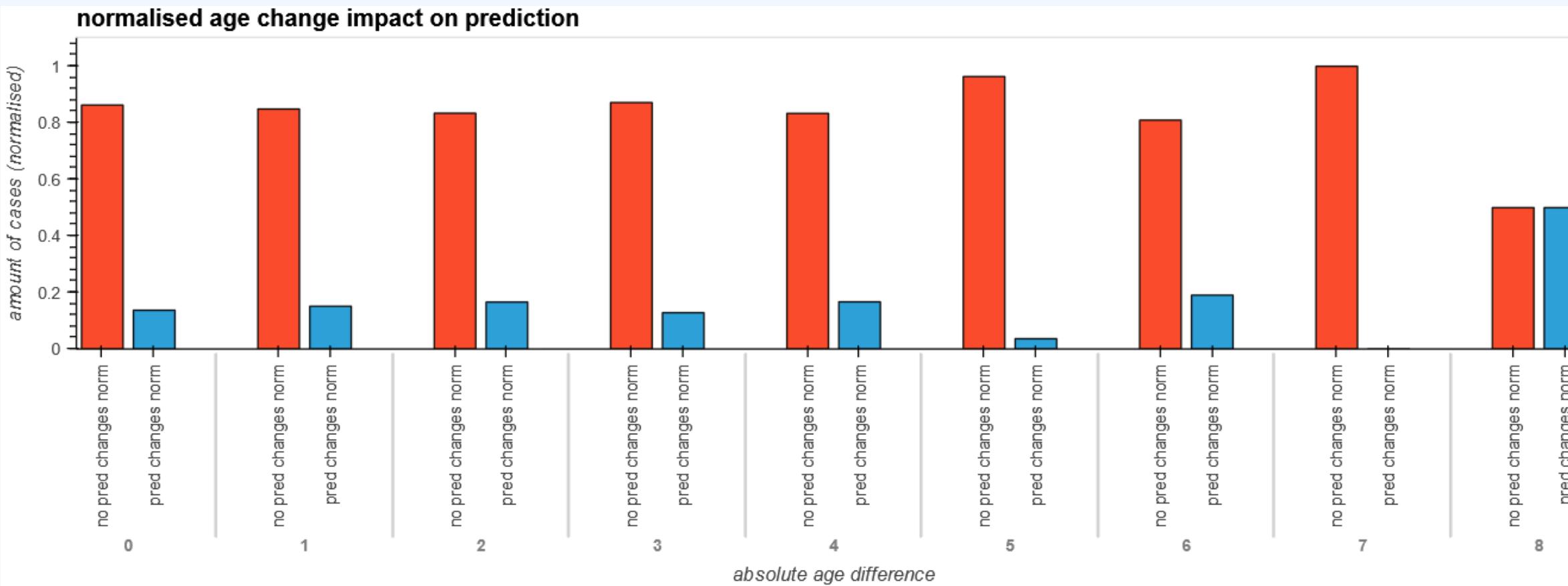
Proportions seem similar between age differences

# Is the Age and Sex noise responsible for the model being confident and wrong?

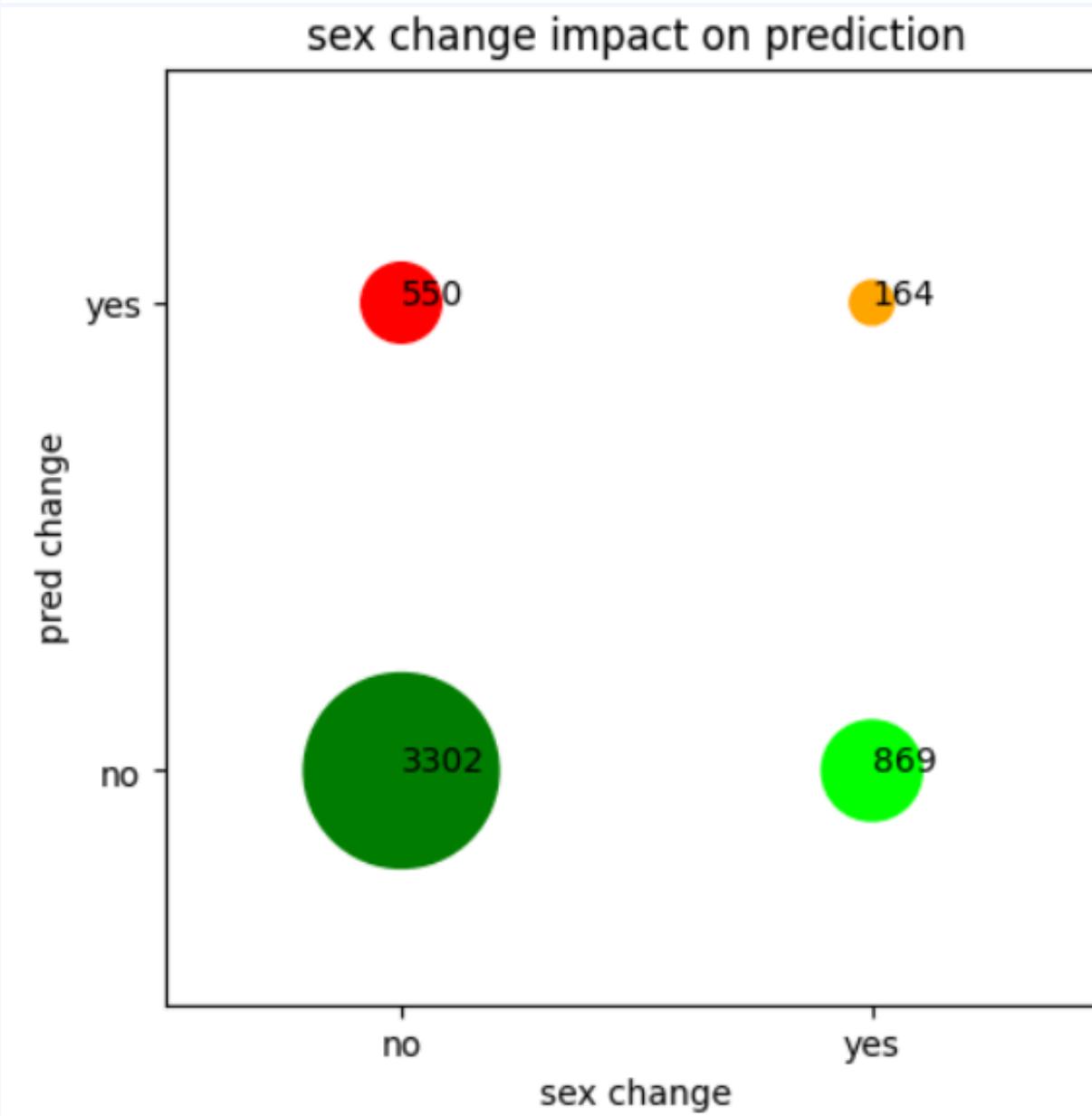
Looking at all the test cases



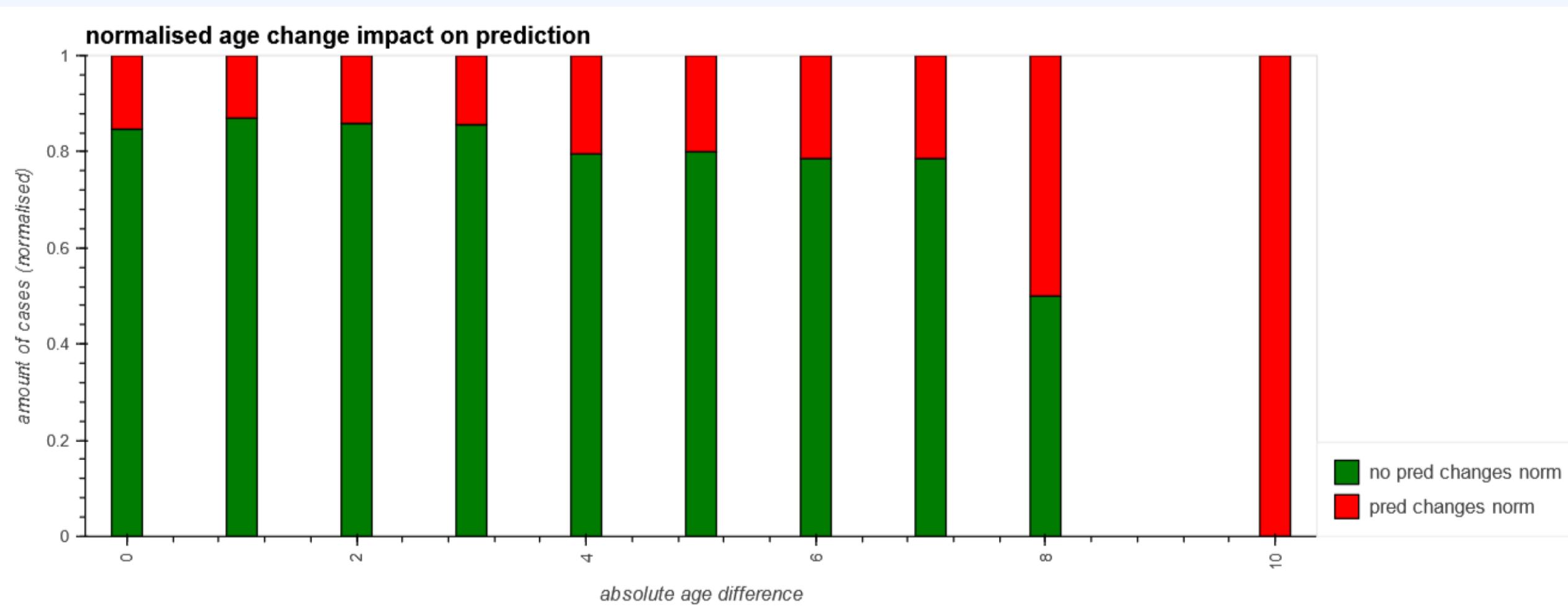
Proportions seem similar between age differences



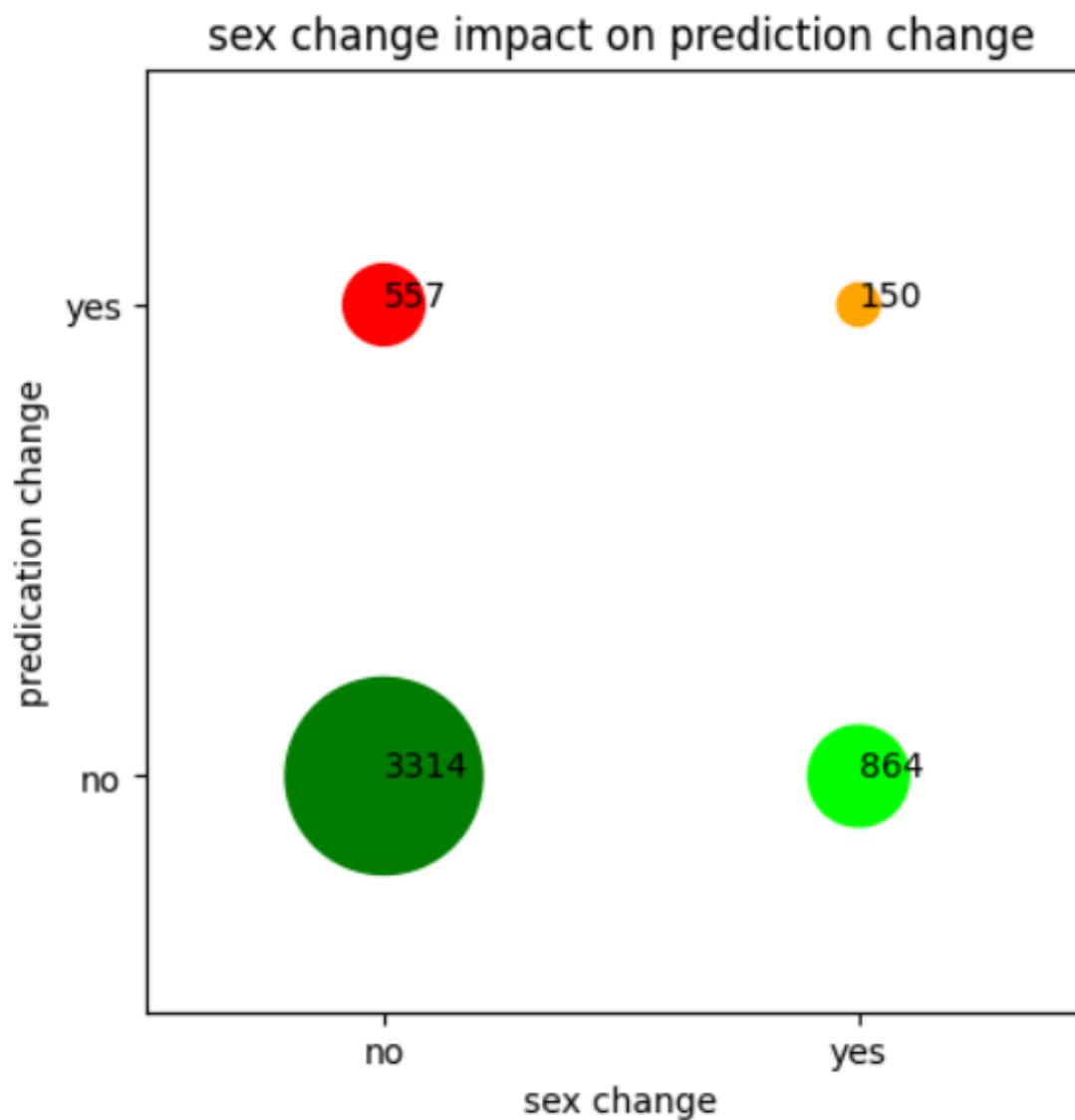
Normalisation confirms similarity



17% and 19% for  
prediction change/no prediction change



Normalisation confirms similarity



~17% for both cases of prediction change/no prediction change

## Looking at only the cases where the model is wrong but confident

```
Average age change while 'miss but confident': 1.0416666666666667  
While 'miss but confident' age was changed 12 times on 24  
While 'miss but confident' sex was changed 5 times on 24  
While 'miss but confident' both were changed 1 times on 24
```

No significant impact visible,  
but fewer cases -> less precise



# EXPLAINABILITY & LLMS

## SETUP ENVIRONMENT :

- **Llm3.2-3b-instruct**
- **LIME** method
- *User role:* provide features, importance scores, and labels.
- *System role:* specify how the natural language should answer, this means the context.

```
explain_method="LIME"
context_system= ("You're an interpreter of a explainability method for a classifier."
                 " The method used is " + explain_method + "."
                 " You need to explain in a simple way anyone can understand what the values of the")
```

Here are the features and their score with LIME method:

```
Name of feature: capital-gain, Importance score: 0.5226965078276797
Name of feature: education_Preschool, Importance score: -0.36581589319378627
Name of feature: occupation_Priv-house-serv, Importance score: -0.3354421907663196
Name of feature: workclass_Without-pay, Importance score: -0.2924377615117225
Name of feature: occupation_Armed-Forces, Importance score: -0.2740265598030081
Name of feature: marital-status_Married-AF-spouse, Importance score: 0.25956166029686994
Name of feature: native-country_Hong, Importance score: 0.22241719269040844
Name of feature: native-country_Cambodia, Importance score: 0.21128383610602153
Name of feature: native-country_Columbia, Importance score: -0.20444894299948574
Name of feature: native-country_Honduras, Importance score: -0.06931146789191218
```

For the example 246, true label was True and prediction label was False. Can you explain to me

## RESULTS :

- Top 5 features most influential
- Provide a potential real explanation of each feature.
- Remind to be mindful and what LIME focus on

Let's analyze the LIME results for example 246.

Since the true label was "True" but the predicted label was "False", we're looking for features that likely contributed to the inco

```
1. **education_Preschool**: -0.36581589319378627
   * This feature represents education level as preschool. Individuals with lower education levels might be less likely to be married.
2. **occupation_Priv-house-serv**: -0.3354421907663196
   * This feature categorizes individuals into private household servants, which is an occupation often associated with limited social mobility.
3. **workclass_Without-pay**: -0.2924377615117225
   * This feature represents individuals without a regular pay, indicating they might be part of the informal economy or have limited access to stable employment.
4. **occupation_Armed-Forces**: -0.2740265598030081
   * Similar to the previous point, this feature indicates military personnel, who may have limited access to stable employment and family life.
5. **native-country_Columbia**: -0.20444894299948574
   * This feature represents individuals from Colombia, which might be associated with cultural or socio-economic factors that influence marriage rates.
```

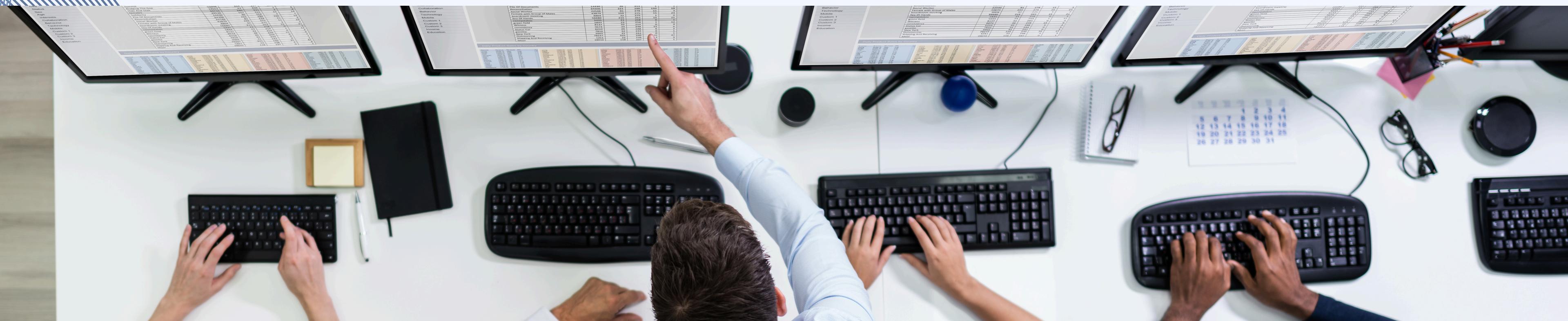
These features collectively suggest that example 246 was mispredicted because:

```
* They have a relatively low education level (preschool), which could indicate limited social capital and stability.
* They work as private household servants, which is an occupation often associated with instability and limited access to resources.
* They don't have regular pay, indicating they might be part of the informal economy or have limited financial security.
* They are military personnel, who may have limited access to stable employment and family life.
* They are from Colombia, a country with potentially different cultural or socio-economic factors that influence marriage rates.
```

The model may have misinterpreted these features, leading to a prediction error. However, it's essential to note that LIME explanations are just one possible interpretation, and there could be other factors at play. The results should be considered in context.

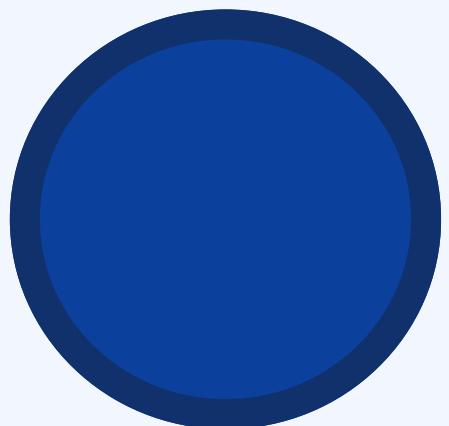
Keep in mind that this is just one possible interpretation, and there could be other factors at play. The results should be considered in context.

# FREE EXPLORATION



In the small time we have given to this task, we couldn't manage to find anything relevant to mention.

# PLEASE ASK QUESTIONS





**THANK  
YOU  
FOR YOUR ATTENTION**

