Master Thesis Information Sciences

Understanding and correcting bias in neural networks:
a credit default prediction case study

Piet Snel 2658151 | Supervisor: Sieuwert van Otterloo

Problem Definition and Motivation

Adoption of AI

Machine learning algorithms help us determine the route to drive to work, predict the weather and efficiently charge our phones.Recently, significant progress has been made in the performance of machine learning models and applications, driven by the development of new learning algorithms and theory and the enormous growth in the availability of online data and low-cost computing.

Complexity

Vision is what success looks like for your company. It is what your company aspires to be in the future. It is how the world will look like once you've accomplished your mission.

Impact

These are the guiding principles that will influence your actions to fulfill your company's mission and vision.

Main Research Question

How can we explain Neural Networks' outcomes in a financial decision making task and address bias in these networks?

- Neural network explainability
- Credit default prediction
- Practical advice
- Example notebook

Sub-Research Question 1

How can we correct bias in a neural network without increasing complexity?

- Bias correction methods
- Explainability
- Adressing black-boxes

Sub-Research Question 2

Is it desirable to counter any bias occurring in a neural network?

- Ethics
- Bias sources
- Understanding bias

Cases of bias in AI

- **Hiring**
  - Amazon
  - British Medical School
  - Modern Hire
- **Police**
  - Facial Recognition Software
  - Repetition Risk
  - Crime Prediction
- **Finance**
  - Default Prediction
  - Credit Scoring
- **Fraud**
  - Fraud Detection
  - Dutch "Toeslagenaffaire"

Related Work

Bias in Black-Box models

Bias in AI is *"...computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals ..."* . Often relatated to sensitive attributes such as Age, Gender, Nationality or Ethnicity. Most prominent cause is biased training data.

Bias Detection

Fragmented and complex method. Explainability methods such as LIME and SHAP. Current bias detection literature is complex and highly theoretical. Determining bias more a question of what you determine to consider bias to be. Four-fifths rule

Bias Mitigation

Pre-processing, post-processing techniqius or imposing fairness contstraints.

Research Strategy

Combining action research with a case-study

1. Diagnosis

Research both literature and practical sources on the impact and occurrence of bias and how this currently may be dealt with.

2. Planning

Evaluate if the credit default prediction model contains any bias and define measure to address this.

3. Intervention

Implement measures to mitigate bias in the model

4. Evaluation

Evaluate if the measures have indeed mitigated any bias in the model

5. Reflection

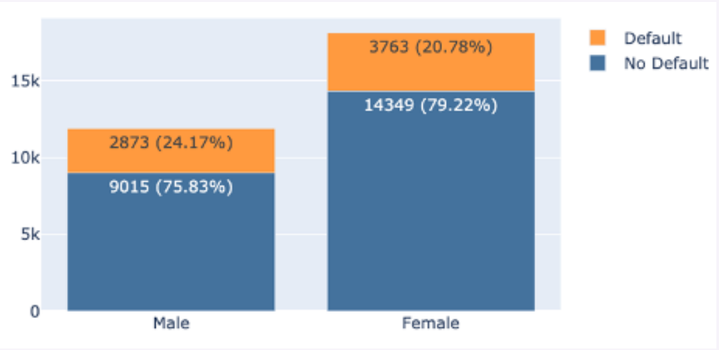
Reflect on the methods and provide a practical guide on how to address bias.

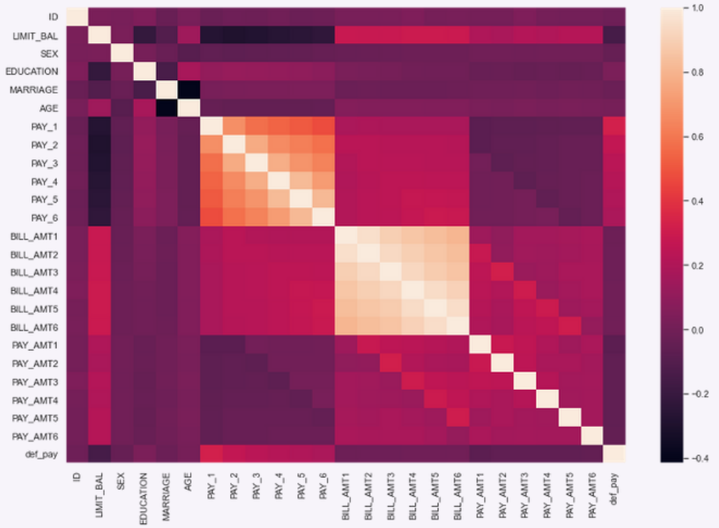
Taiwanese Credit Default Dataset

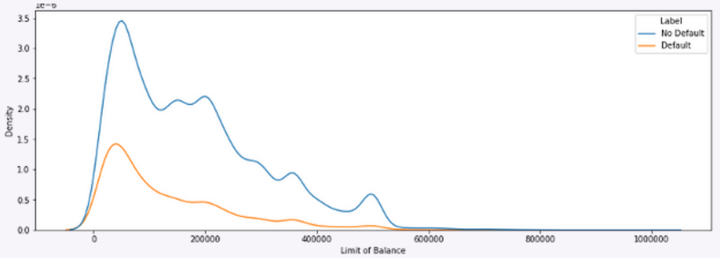
I. C. Yeh and C. H. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients.," Expert Systems with Applications, pp. 2473-2480, 2009.

Characteristics

- Generated by a Taiwanese credit institution in 2006
- **30.000 instances**
- 24 attributes such as Age, Gender, Limit of Balance and Payment status
- **11.888 males**, default rate: 24.17%
- **18.112 females**, default rate: 20.78%
- **High Limit of Balance** has a lower risk of default.



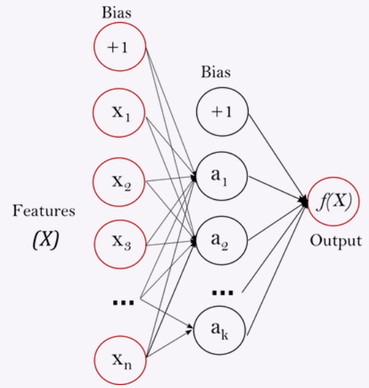




Model Preparation

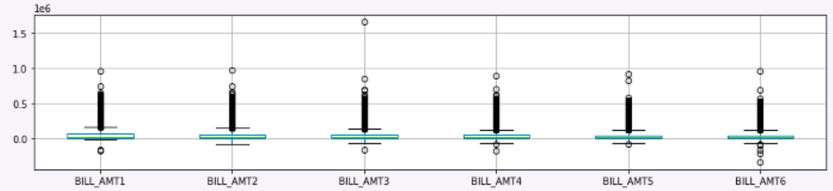
Model Selection

Artificial neural networks are the only algorithms able to accuratetely predict this datasets probability of default (Yeh and Lien, 2009). Multi-Layer Perceptron in Sci-Kit learn using the MLPClassifier function



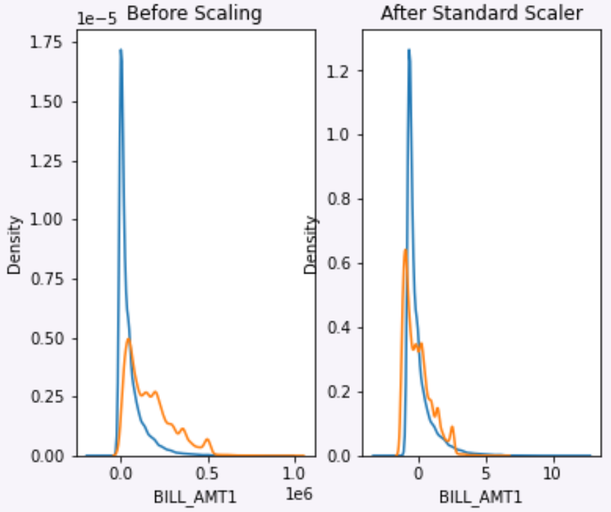
Feature Selection

- PAY_1
- LIMIT_BAL
- PAY_AMT2
- PAY_2
- PAY_4
- BILL_AMT2
- PAY_AMT5
- PAY_3

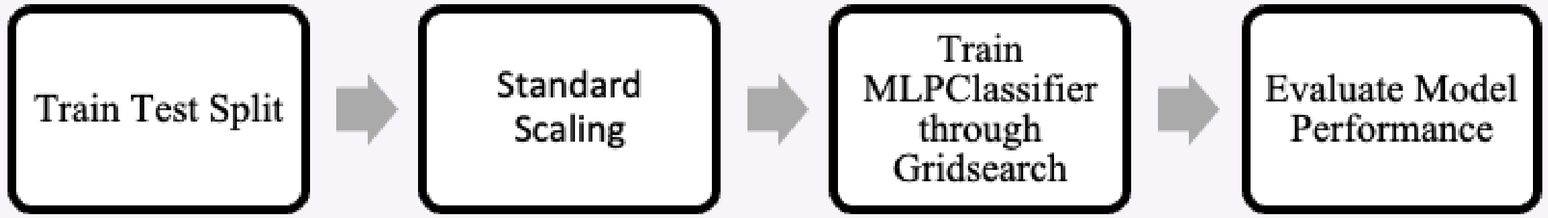


Feature Scaling

Execution is how you will employ the tactics you've chosen. It includes measurable outcomes, such as timelines and deliverables.

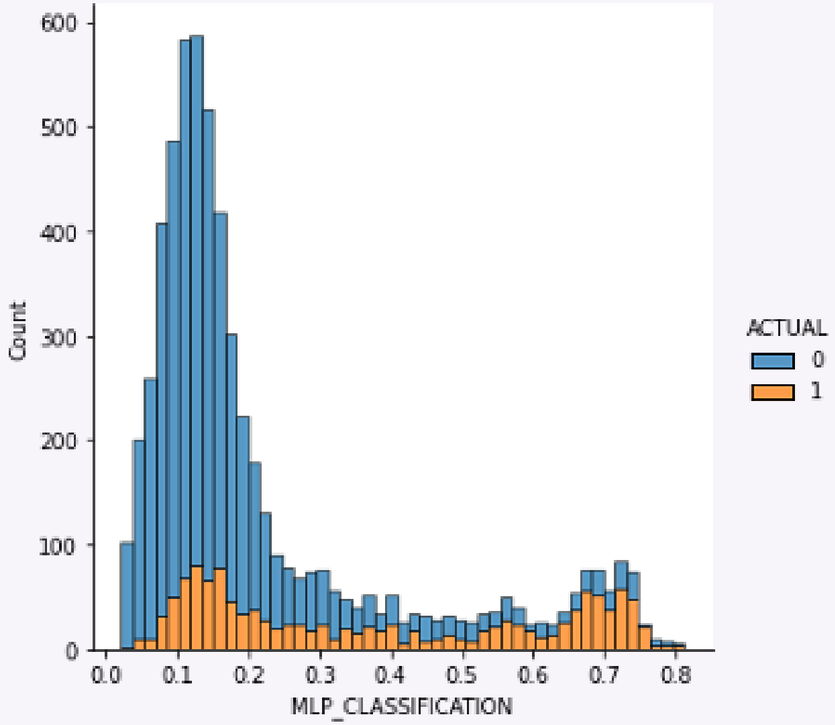


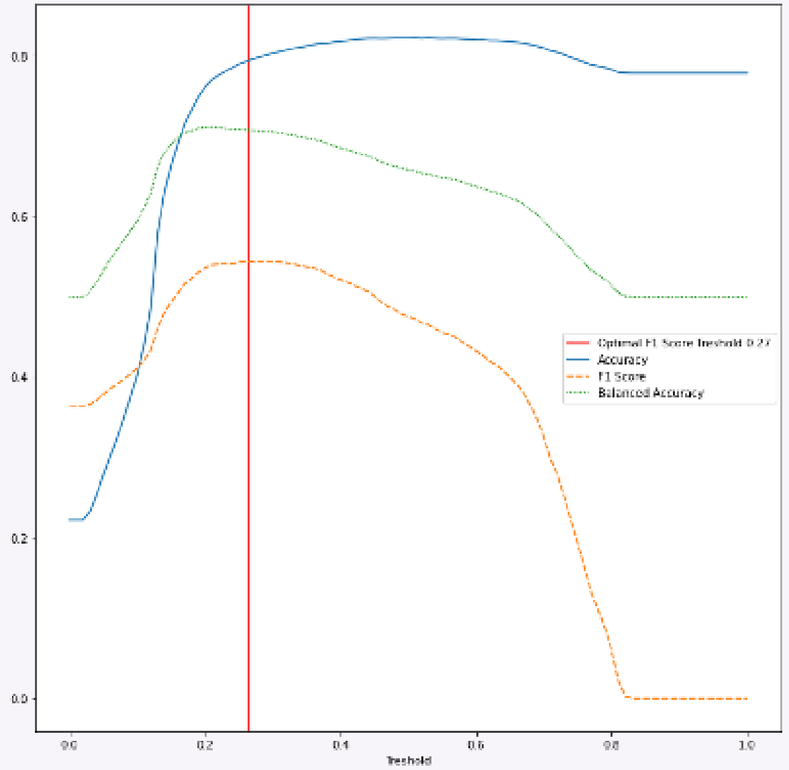
Model Training and Performance



Gridsearch

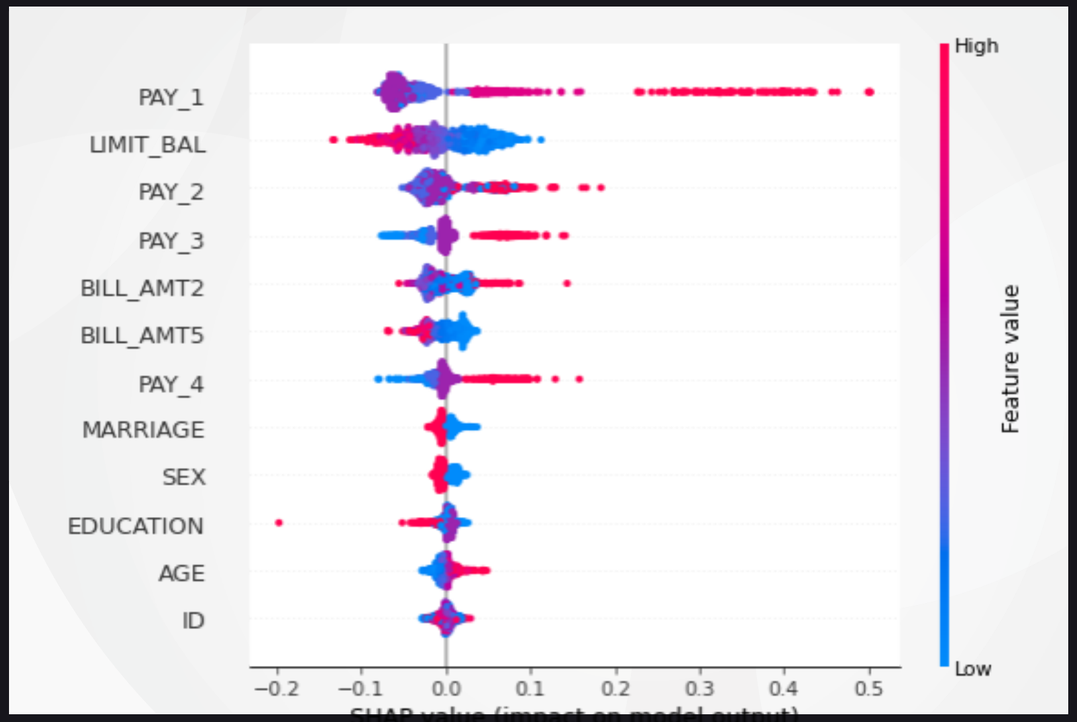
parameters = {'activation': 'logistic',
'alpha': 0.0001,
'hidden_layer_sizes': (12, 17, 12),
'learning_rate': 'constant',
'max_iter': 10000,
'solver': 'adam'}



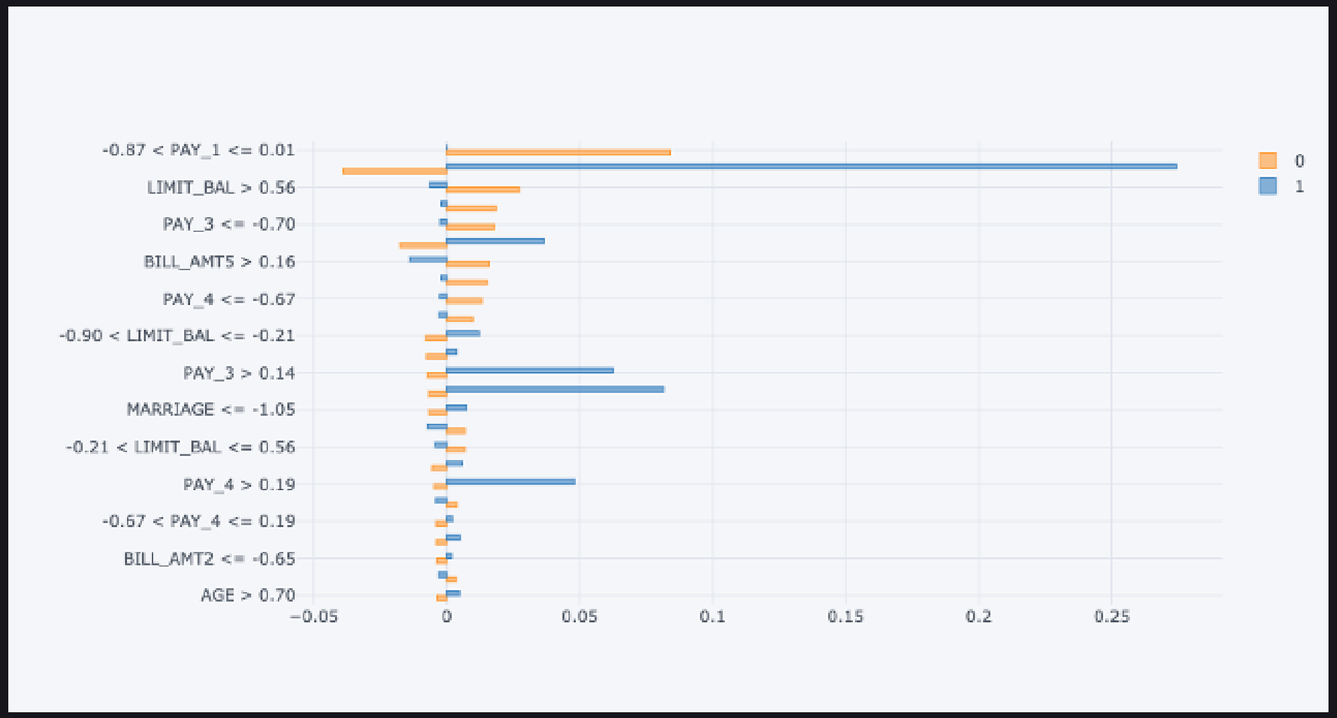


Understanding through LIME and SHAP

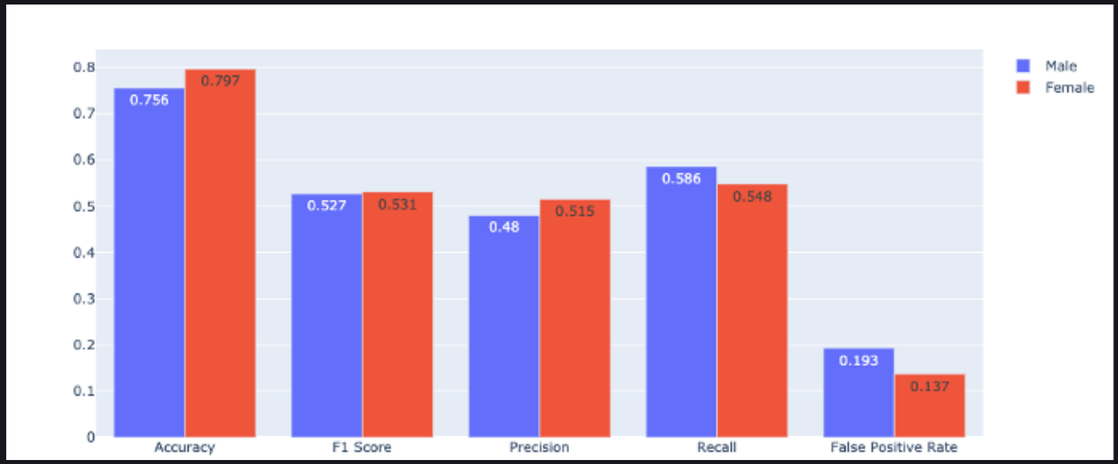
SHAP

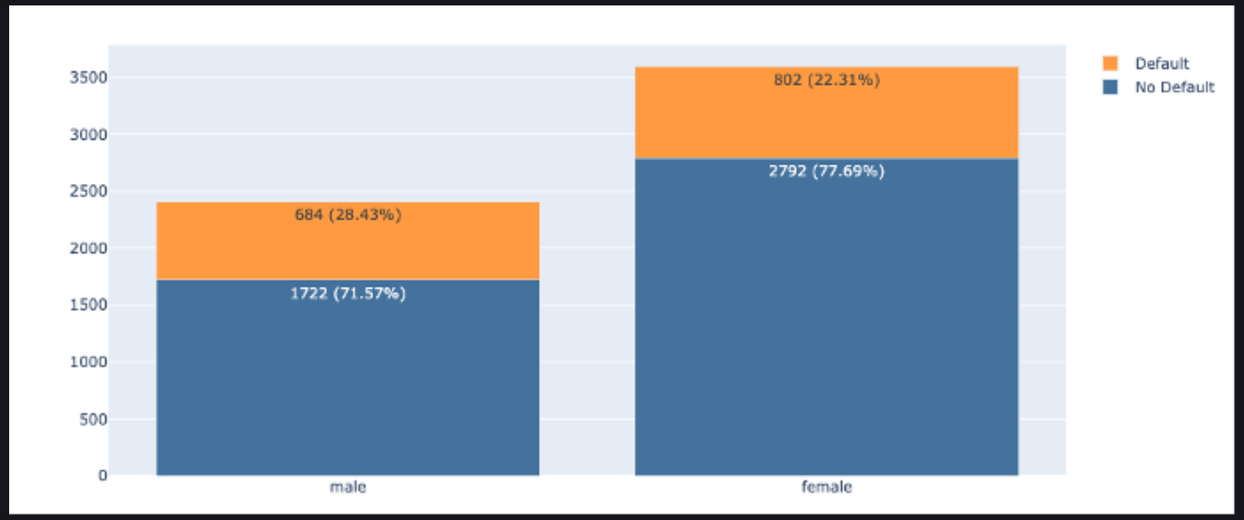


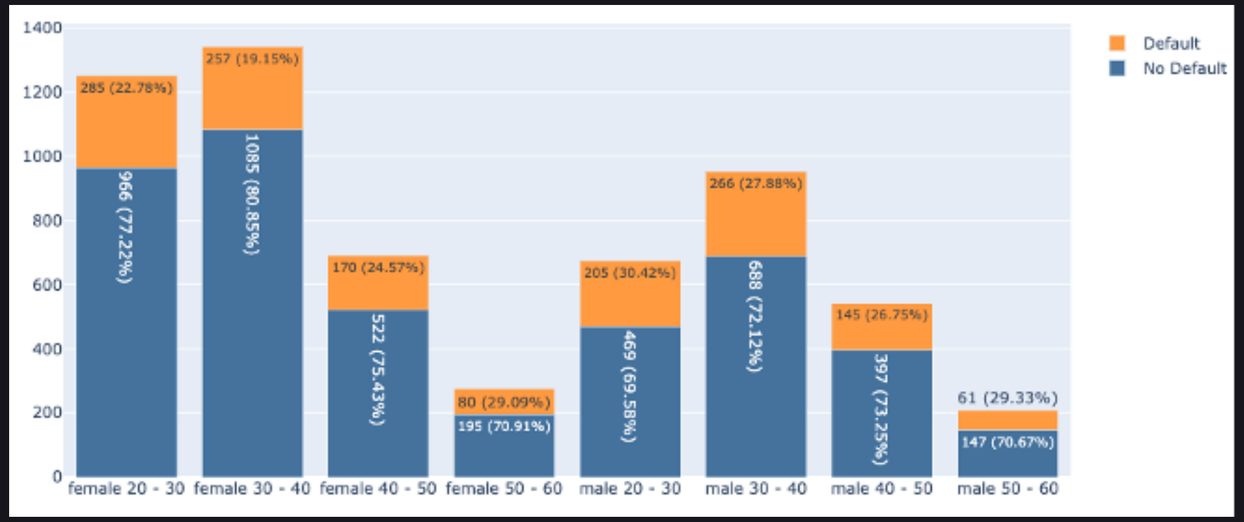
LIME



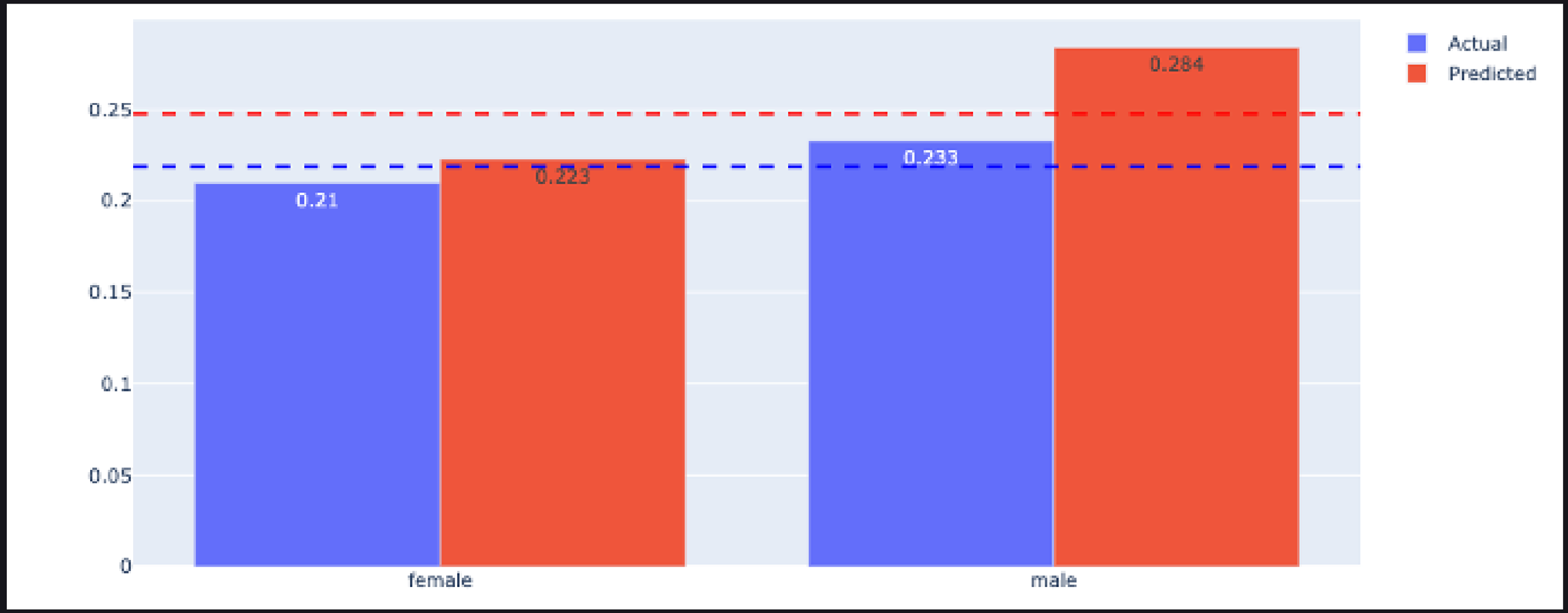
Review of *initial model*



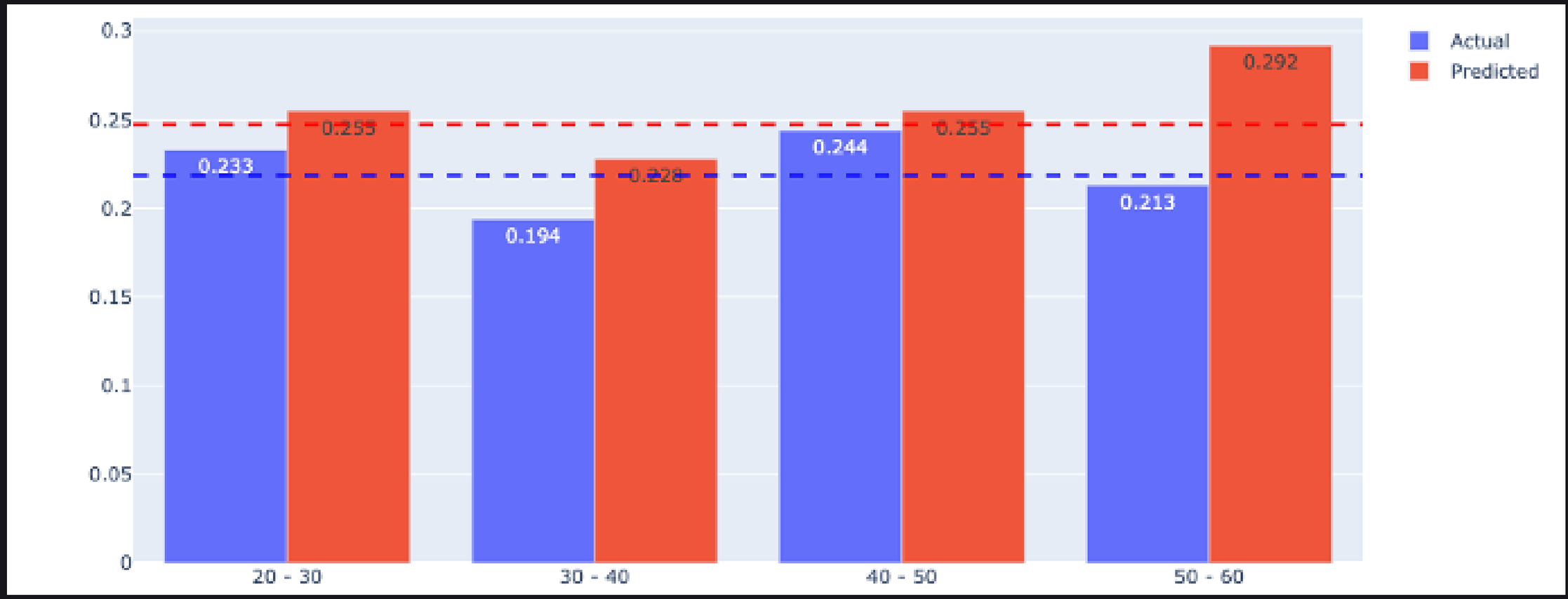




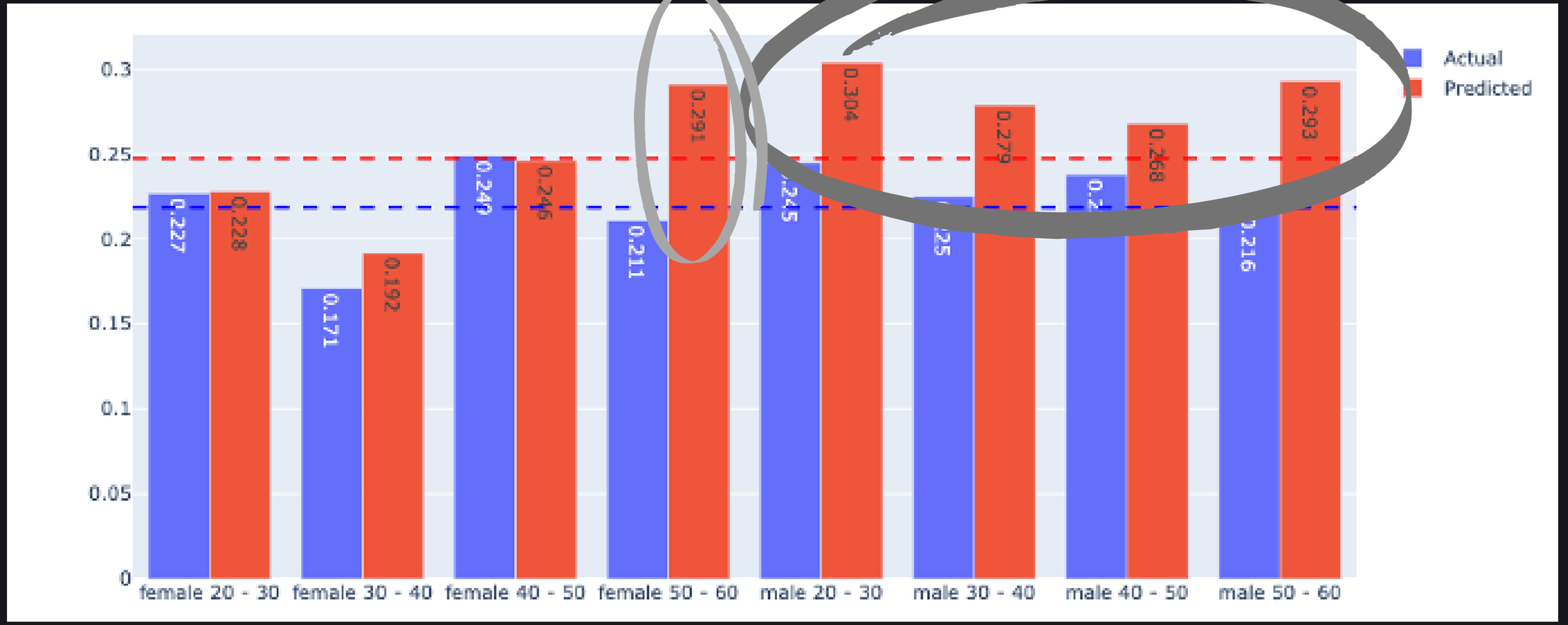
Review of *initial model* : Gender



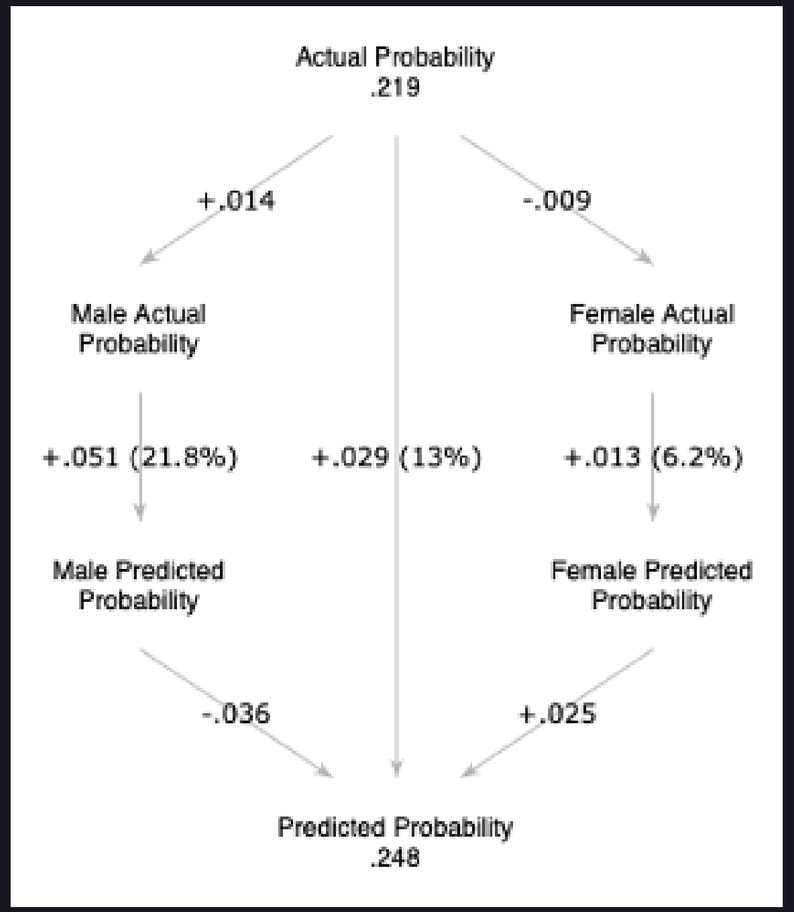
Review of *initial model* : Gender and Age



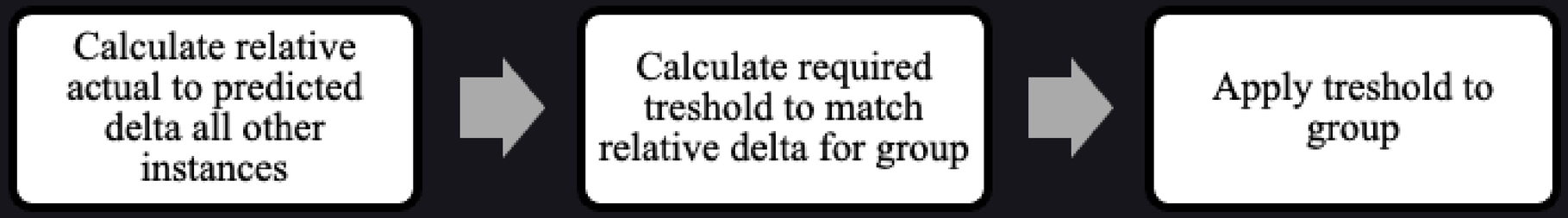
Review of *initial model* : Gender and Age



Review of *initial model* : Probability increase



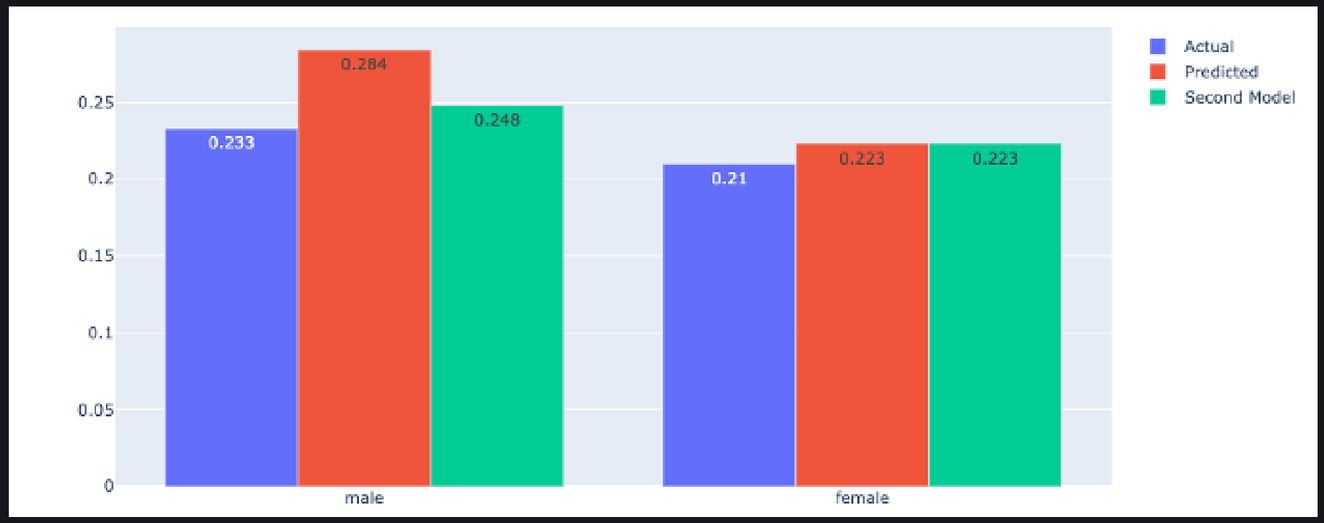
Bias Mitigation Technique

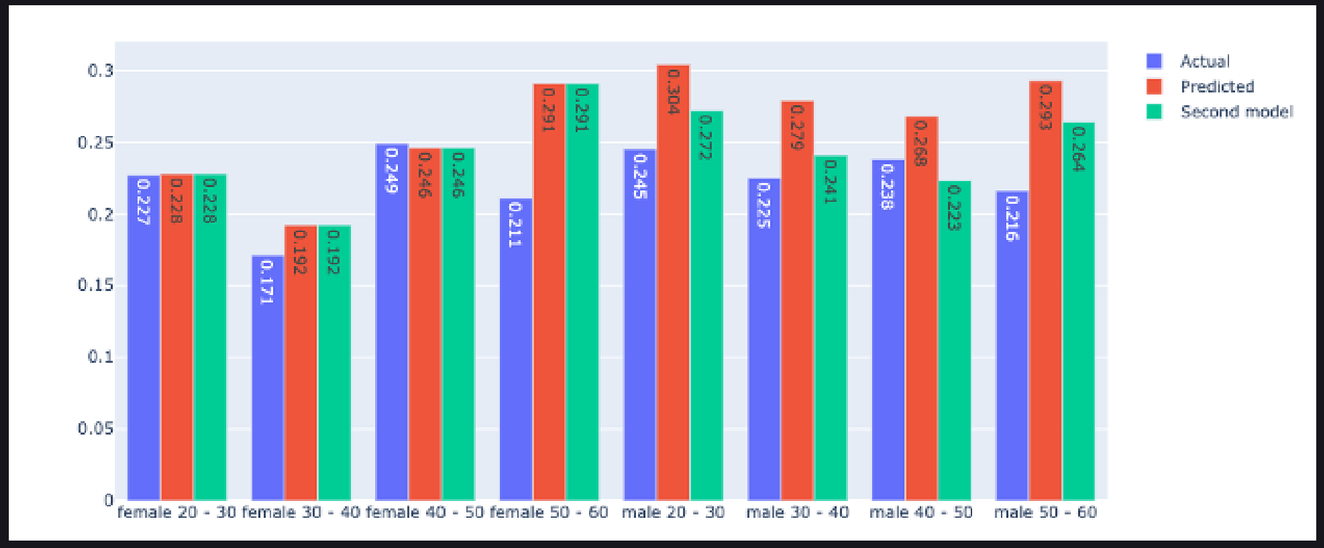


$$TargetProbability = \frac{MeanPredictedProbability \notin Group}{MeanActualProbability \notin Group} \times MeanActualProbability \in Group$$

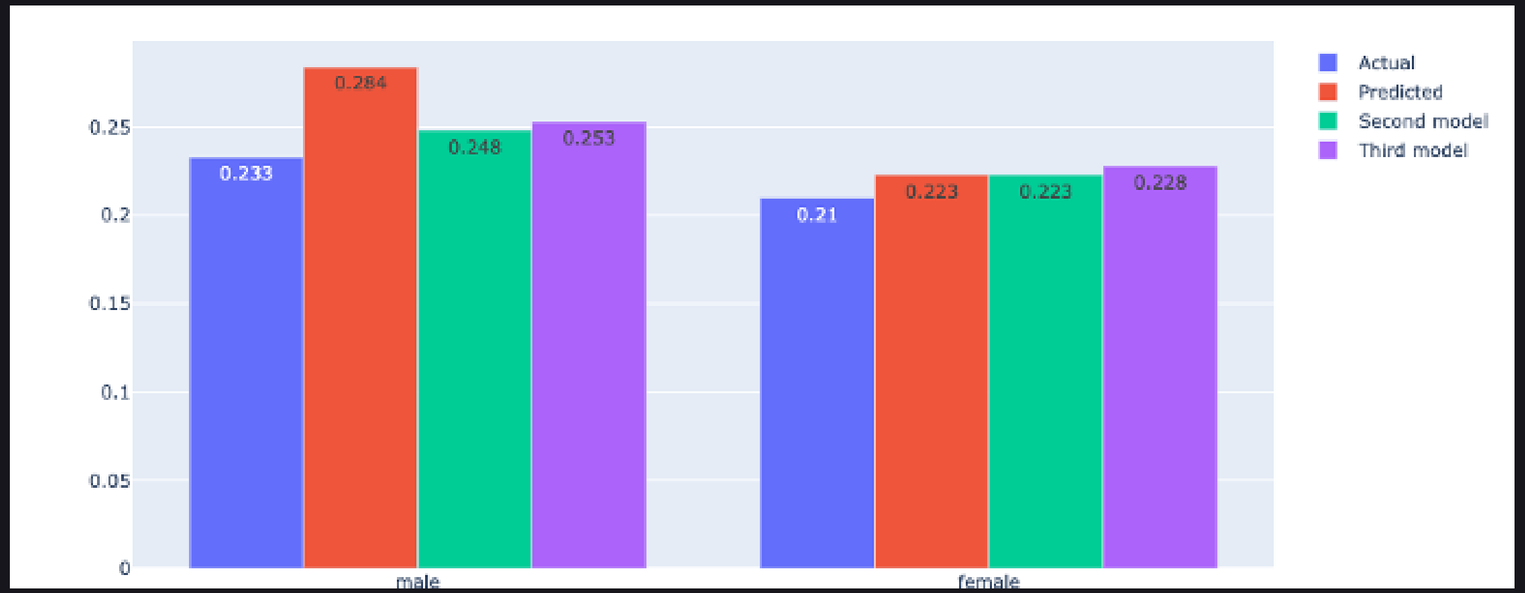
```
column = 'SEX'
group = 'male'

actual_probability = df[df[column] != group].ACTUAL.mean()
predicted_probability = df[df[column] != group].PREDICTED.mean()

relative_delta = predicted_probability / actual_probability
target_probability = df[df[column] == group].ACTUAL.mean() *
relative_delta
```

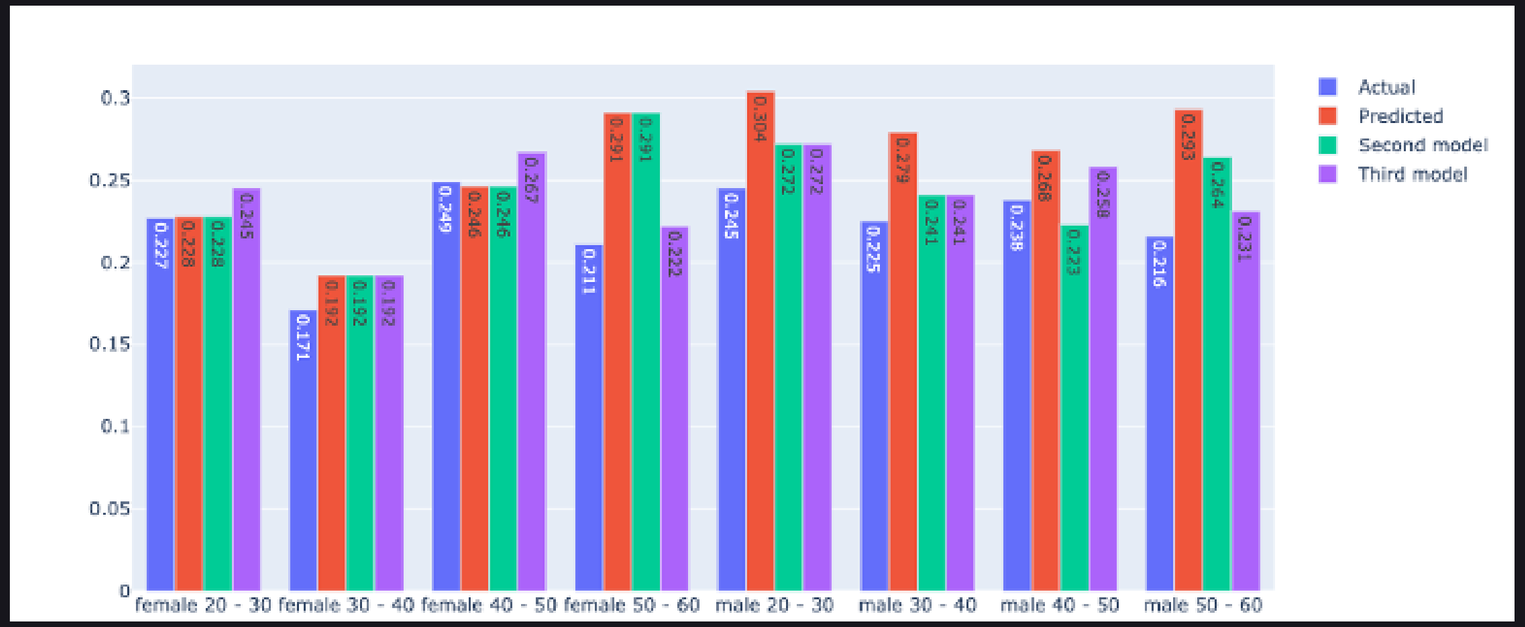
Review of *second model*





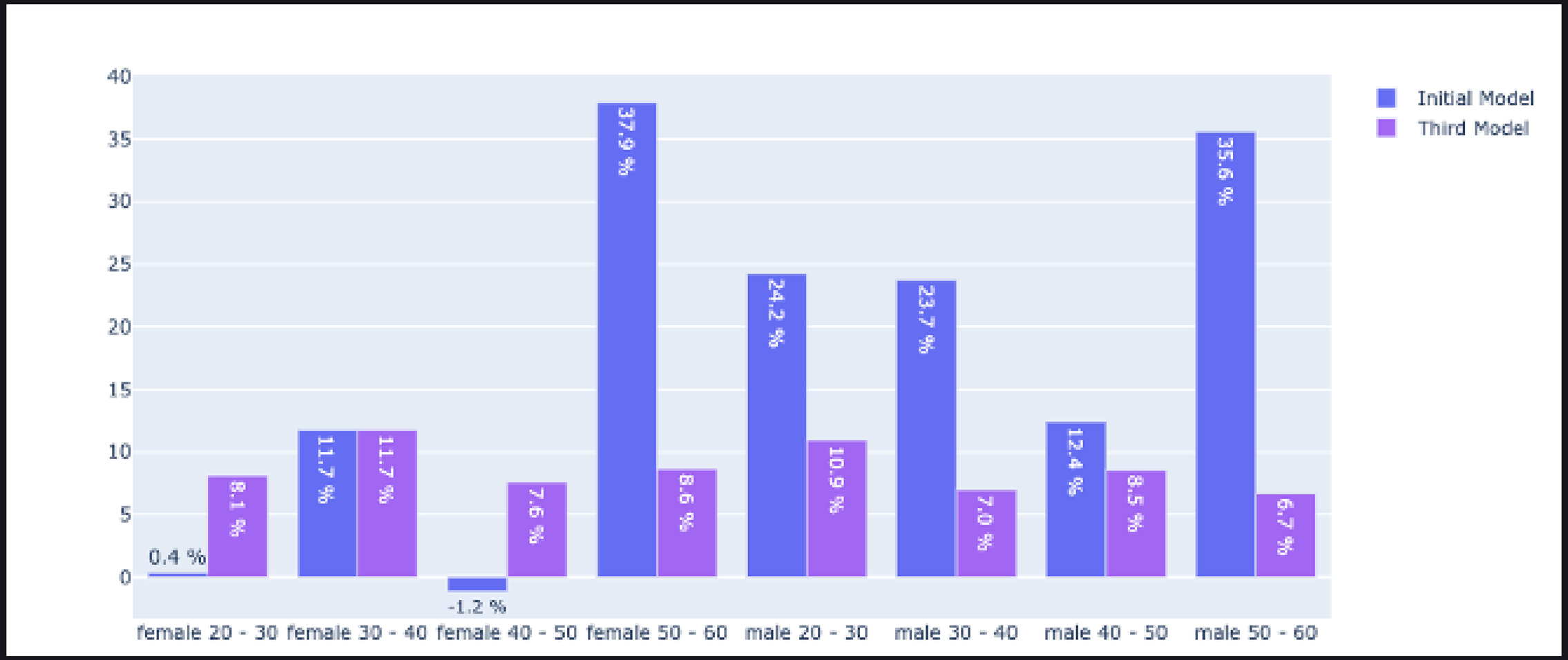
Review of *third model*





Review of *performance and impact*

Third Model overall Result



Conclusion and Discussion

WIP, dependent on feedback of report