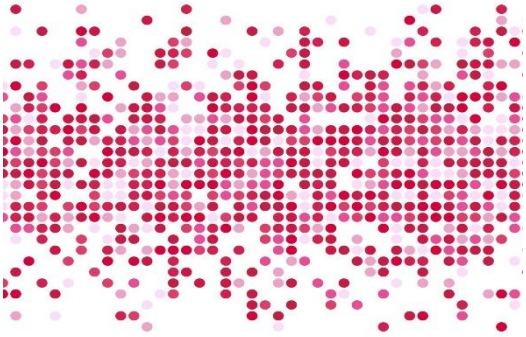
Investigating Bias in Machine Learning Models

"A Health Insurance Cross-Sell Case Study"

Artificial Intelligence Ethics and Applications

(CIS4057)



by

Krishna Gopal Sharma
S3454618

Word Count: 2098

Abstract

This report investigates gender bias in a machine learning model predicting health insurance cross-sell acceptance. Using fairness metrics—accuracy, demographic parity, and equal opportunity—the study found performance discrepancies between genders, favoring female accuracy and male positive prediction rates. LIME was used to interpret model decisions, revealing feature influences. The findings emphasize the need for fairness evaluations and transparency in AI decision-making.

1. Introduction

Artificial Intelligence (AI) and Machine Learning (ML) have transformed various industries by automating decision-making processes and improving efficiency. However, one critical issue in AI is the presence of bias, which can lead to unfair and discriminatory outcomes. In sectors like healthcare, finance, and insurance, biased AI models can perpetuate existing inequalities, leading to adverse consequences for marginalized groups.

In the context of health insurance, biased models can reinforce disparities in access to services or unfairly target specific demographic groups. For example, if a model trained on biased data and favours one gender or age group, then it offers better opportunities to that group while overlooking others. This can result in unfair treatment and exacerbate existing social and economic disparities.

AI bias often arises from several factors, including biased training data, flawed algorithms, and the failure to account for protected characteristics such as gender, race, or age. In this report, we explore the presence of such biases in a machine learning model used for predicting the likelihood of customers accepting a health insurance cross-sell offer. The fairness of this model is evaluated using three key fairness metrics:

1. **Accuracy.**
2. **Demographic parity**
3. **Equal opportunity.**

With ‘Gender’ as the protected characteristic. By evaluating these metrics, my aim is to highlight any potential biases present in the model and discuss their implications for fairness in AI-driven decision-making. AI decisions are not random, there are logic behind each decision, and to understand these hidden logics, this study used LIME (Local Interpretable Model-agnostic Explanations), technique which used to explain the predictions of machine learning models. It creates simple, interpretable models locally around a specific prediction to help understand why the model made a particular decision, regardless of the complexity of the underlying model.

2. Model Development and Application of Fairness Criteria

2.1 Data Exploration and Preprocessing

The dataset used in this study, "Health Insurance Cross-Sell Prediction", contains customer demographic and behavioural data. Key features include age, gender, vehicle age, vehicle damage, annual premium, and a binary target variable ‘Response’ (whether the customer accepted the health insurance offer).

Data preprocessing included handling categorical variables (like gender, vehicle age, and vehicle damage) using ‘One-Hot-Encoding’, transforming them into numerical features for model input. For instance, gender was encoded as 0 for male and 1 for female. Vehicle age categories were mapped to numerical values, and vehicle damage was transformed into binary values (0 for "No" and 1 for "Yes").

Numerical features such as age, annual premium, and vintage were scaled using ‘StandardScaler’ to ensure the model's performance was not biased due to differing feature scales. The dataset was split into a training set (80%) and a testing set (20%) to evaluate the model's performance.

2.2 Model Development

This study is based on the 'Support Vector Machine (SVM)' machine learning model as the classification algorithm for predicting whether a customer would accept the insurance offer or not. SVM is a popular and effective method for binary classification tasks, especially when dealing with high-dimensional data. The model was trained on the scaled training set using the SVC (Support Vector Classifier) with the probability=True parameter to allow probability estimation, which helps in evaluating the dataset over by predicting the value and then compare then with the test results create a confusion matrix.

The model's performance was evaluated on the test set using multiple metrics, including accuracy, precision and recall. The confusion matrix was also computed to gain deeper insights into how the model performed in terms of true positives, false positives, true negatives, and false negatives

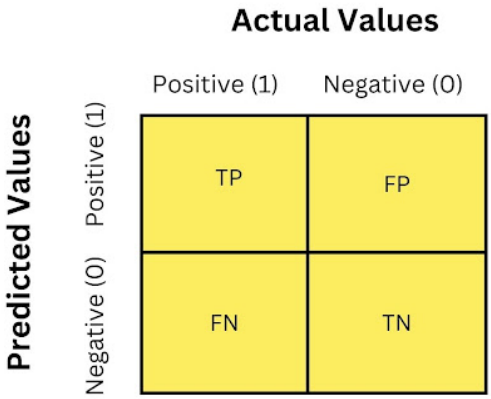


Fig 1: The Confusion Matrix

2.3 Performance Evaluation

The evaluation metrics for the SVM model are as follows:

- Accuracy: 0.8125
- AUC: 0.8307
- Classification Report: Precision, recall, and F1-score were computed for both the positive and negative classes.

These metrics provide a detailed picture of the model's performance. The confusion matrix for the overall model shows how the model is predicting both classes (0 = No, 1 = Yes) across all instances in the test set.

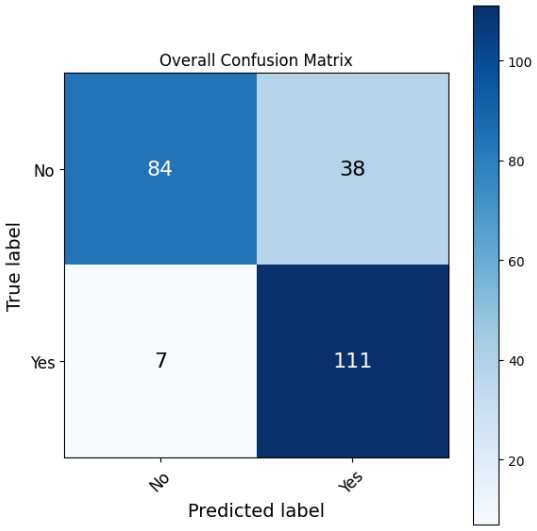


Fig 2: Confusion Matrix for Overall Model

- **Confusion Matrix (Overall)**:
  - True Positive (TP): 111
  - False Positive (FP): 38
  - False Negative (FN): 7
  - True Negative (TN): 84

2.4 Fairness Evaluation Based on Gender

To investigate any potential bias, we performed fairness evaluations based on the **gender** attribute, which was treated as the protected characteristic. The following fairness criteria were applied:

1. **Equal Accuracy**: The accuracy was calculated separately for male and female customers to check if the model performs equally well across both groups.
2. **Demographic Parity**: We calculated the proportion of positive predictions for males and females. A large disparity in these proportions could indicate gender bias.
3. **Equal Opportunity**: We calculated the **recall** (true positive rate) for both

males and females to assess whether the model is equally likely to identify positive outcomes for each group.

These metrics were analysed to determine whether gender bias exists in the model's predictions.

2.5 The Expandable AI with LIME

Local Interpretable Model-agnostic Explanations (LIME) is a technique used to explain the predictions of machine learning models. It works by approximating a complex model locally with a simpler, interpretable model to help users understand why a specific prediction was made. In this study, LIME is applied to explain the predictions of a Support Vector Machine (SVM) model. By using LIME, we can gain insights into the factors influencing the SVM's decisions, making the model more transparent and interpretable for better understanding and trust in its predictions.

3. Findings

Based on the evaluation of the fairness criteria, we found that the model exhibits some bias towards the male group:

3.1 Accuracy:

Accuracy for Male**:** 77.12%

Accuracy for Female**:** 85.25%

This indicates that the model performs better in predicting outcomes for female customers. The model performed better with female group as compared to the male. So, this shows that the model is biased with male. So, if the female customer will go to get the health insurance, then model most likely say “YES” to her and chances becomes high to get the insurance.

3.2 Demographic Parity:

Demographic Parity for Male: 65.25%

Demographic Parity for Female: 59.02%

The demographic parity metric shows the proportion of positive predictions, means who accept the offer. Here, male has 65.25% and female has 59.02% which suggesting that the model is more likely to predict a positive outcome for male customers. The male customers are most likely to accept the offer to have health insurance. Although, the difference is not much between each gender but still, there is difference, and the suggestion is bit higher for male group.

3.3 Equal Opportunity (Recall):

Recall for Male: 93.10%

Recall for Female: 95.00%

The recall metric shows how well the model is at identifying actual positives (those who accepted the offer) for each gender. Here, the recall for males (0.93) and females (0.95) showed a relatively small difference, indicating that the model is fairly accurate in identifying positive cases for both groups, though females had a slightly higher recall. The difference in recall between males and females is quite small (93.10% for males vs. 95.00% for females). This suggests that the model is balanced in identifying positive cases for both genders. The small difference might not point to a major issue of bias, but any difference should be closely monitored.

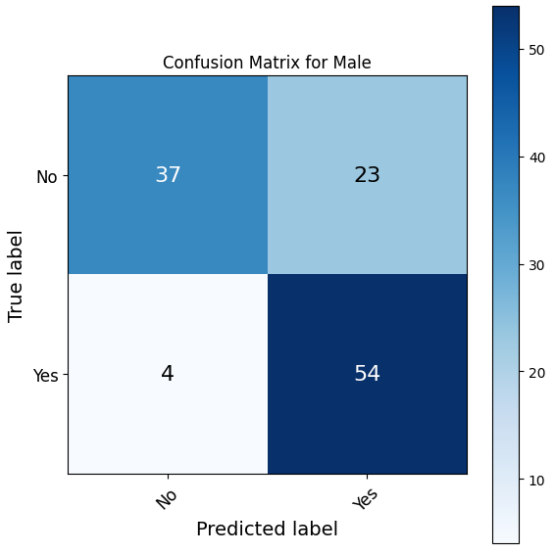


Fig 3: Confusion Matrix for Male

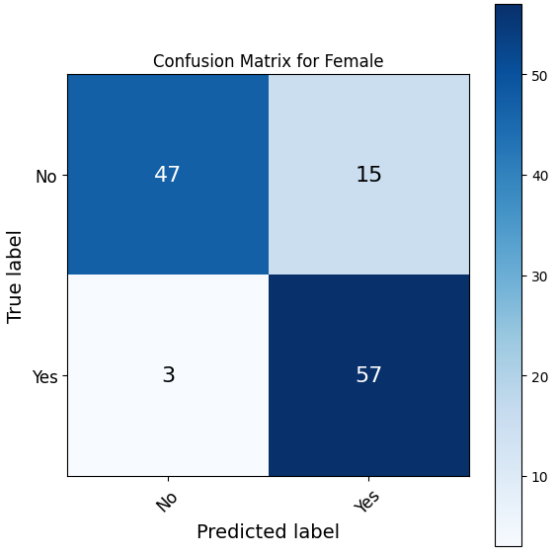


Fig 4: Confusion Matrix for Female

Category	True Positive (TP)	False Negative (FN)	False Positive (FP)	True Negative (TN)
Overall	111	7	38	84
Male	54	4	23	37
Female	57	3	15	47

Table 1: Confusion Matrix Comparison (numerical values)

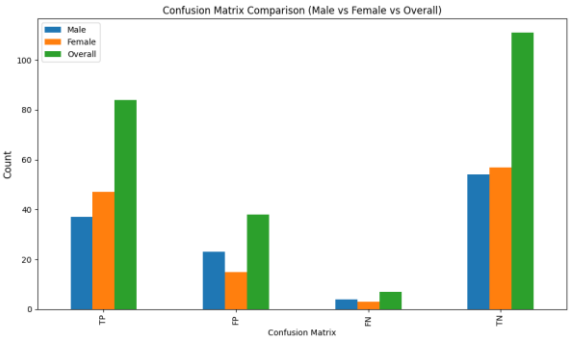


Fig 5: Bar-Graph Comparing Confusion Matrix Results

3.4 Limitations of Fairness Criteria

While the applied fairness metrics offer valuable insights, there are limitations:

- Accuracy: In imbalanced datasets, accuracy can be misleading. It might show high performance even when the model is biased towards predicting the majority class.
- Demographic Parity: This metric does not account for differences in the base rates between groups, meaning it may not always be the best indicator of fairness.
- Equal Opportunity: While recall is important, it overlooks other factors such as precision, meaning that a model with a high recall could still make a disproportionate number of false positives.

These limitations underscore the importance of using multiple fairness metrics to gain a comprehensive understanding of bias.

3.5 LIME Result

In this study the instance 14 is used to check how the SVM model will affect the prediction value. The actual value for this instance is yes (1) but the model predicted the wrong value (0).

```
Chosen Instance 14:
Actual Value: 1
Predicted Value: 0
Prediction: Incorrect
```

Fig 6: Chosen instance '14' for LIME evaluation.

This wrong prediction is influenced by my feature's value for the instance 14.

	Feature	Importance
0	-0.58 < Previously_Insured <= 1.72	-0.391695
1	Vehicle_Age_> 2 Years > -0.28	0.070883
2	Policy_Sales_Channel > 0.83	-0.067925
3	Vintage > 0.86	0.049405
4	Annual_Premium > 0.49	-0.035142
5	Driving_License > 0.00	0.031340
6	Vehicle_Damage_Yes > 0.62	0.031340
7	Age > 0.60	0.014474
8	-1.00 < Gender_Male <= 1.00	-0.005506
9	Region_Code > 0.52	-0.003157

Fig 7: LIME output for instance '14' and effect of various features.

In the image above, the table shows the importance of 10 features that influence the SVM model's prediction for a given instance. Features like "Previously_Insured" and "Policy_Sales_Channel" have negative importance values, indicating they pushed the model toward predicting "0" (incorrect prediction). In contrast, features such as "Vehicle_Age_> 2 Years" and "Vintage" have positive values, meaning they influenced the

model toward predicting "1" (correct prediction). The magnitude of each feature's importance indicates how much it contributed to the decision: larger values (positive or negative) suggest stronger influence.

The expression "-0.58 < Previously_Insured <= 1.72" refers to the range of values for the "Previously_Insured" feature that the model considered when making the prediction. This range helps us understand how the specific value of "Previously_Insured" within these bounds affects the model's output. For example, values closer to -0.58 or 1.72 suggest a stronger impact on the model's prediction.

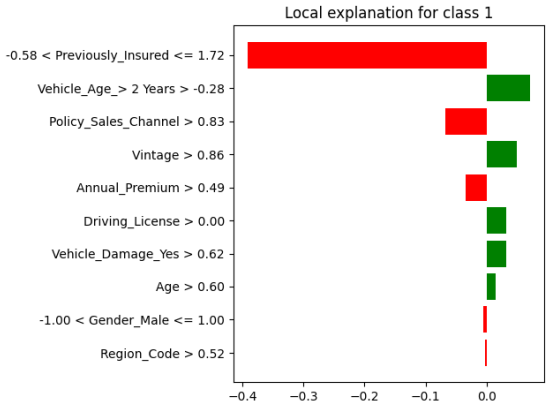


Fig 8: A bar-graph representation of LIME output.

This graph is for better visualization of the LIME result. The accompanying of this bar graph provides a visual representation of these feature importances. Each bar represents a feature's contribution to the prediction, with the length of the bar showing its magnitude. Negative bars (RED) push the model toward predicting "0," while positive bars (GREEN) indicate a push toward predicting "1." This graphical representation helps visualize the impact of each feature, making it easier to understand which ones are most influential in the model's decision-making process.

4. Conclusion

This study highlights the potential for gender bias in machine learning models used in health insurance cross-sell prediction. Despite the high overall accuracy and reasonable recall rates for both male and female customers, the model demonstrated discrepancies in terms of accuracy and demographic parity. The accuracy was higher for female customers, while males were more likely to receive positive predictions.

Incorporating Local Interpretable Model-agnostic Explanations (LIME) provided valuable insights into how specific features influenced the model's predictions. By explaining individual predictions, LIME helped identify which factors contributed to the discrepancies observed in gender-related predictions, offering a clearer understanding of the model's decision-making process.

While fairness metrics like equal accuracy, demographic parity, and equal opportunity provided insights into the model's behavior, they also have limitations. Future work should explore additional fairness criteria, consider techniques like bias mitigation, and further integrate interpretability methods like LIME to address these imbalances.

AI and ML models must be continuously assessed for fairness to ensure that automated decision-making processes do not inadvertently harm or disadvantage certain groups. By adopting a multifaceted approach to fairness evaluation, including model interpretability and fairness metrics, we can improve AI models and their impact on society.

5. References

1) Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. Available online.

2) Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

3) Holstein, K., Wortman Vaughan, J., Wallach, H., Daumé III, H., & Kiciman, E. (2019). Improving fairness in machine learning systems: What do industry practitioners need to know? Proceedings of

the 2019 CHI Conference on Human Factors in Computing Systems, 1-16.

4) Kleinberg, J., Levy, K., & O'Neil, M. (2018). Discrimination in Online Ad Delivery. Communications of the ACM, 61(6), 18-21.

Figures

Fig 1: Confusion Matrix

Fig 2: Confusion Matrix for Overall Model

Fig 3: Confusion Matrix for Male

Fig 4: Confusion Matrix for Female

Fig 5: Confusion Matrix Comparison (Male vs Female vs Overall)

Fig 6: Chosen instance '14' for LIME evaluation.

Fig 7: LIME output for instance '14' and effect of various features.

Fig 8: A bar-graph representation of LIME output.

Table 1: Confusion Matrix Comparison (numerical values)