UPDATE 3 - Fairness Assessment and Mitigation in Car Insurance Claim Prediction

1. Introduction & Ethical Challenge

- **Goal:** Develop and evaluate a model to predict insurance claim likelihood (CLAIM_FLAG) based on customer data (car_insurance_claim.csv).
- **Ethical Challenge:** The main ethical challenge is preventing unfair bias in predictions based on sensitive attributes. Such bias can lead to unequal treatment or access. This analysis focuses on mitigating disparities across GENDER, MSTATUS, EDUCATION, OCCUPATION, URBANICITY, and PARENT1.
- **Objective:** This project trains a baseline Logistic Regression model to predict CLAIM_FLAG and evaluates its fairness using Fairlearn, with a focus on Equalized Odds—ensuring similar true and false positive rates across groups. To reduce disparities, ExponentiatedGradient is applied with an EqualizedOdds constraint for each sensitive attribute. The impact of mitigation is assessed through both fairness improvements and model performance.

2. Data & Preprocessing Summary: The `car_insurance_claim.csv` dataset (~10k records) was cleaned by removing ~2,645 rows with missing values and standardizing currency fields, leaving ~7,657 records. Categorical features were one-hot encoded. Sensitive features like GENDER, MSTATUS, PARENT1, EDUCATION, URBANICITY, OCCUPATION, along with ID and CLM_AMT, were excluded from training but kept for fairness evaluation. The target variable was `CLAIM_FLAG` (0 = No Claim, 1 = Claim).

3. Methodology

- **Baseline Model:** The baseline predictive model chosen was Logistic Regression, selected for its interpretability and initial predictive performance.
- **Data Handling:** The data was split into 80% training and 20% testing sets with stratification to maintain class balance. To address the class imbalance (~27% positive cases), SMOTE was applied to the training data. Features were then standardized using `StandardScaler`, fitted on the SMOTE-processed training set and applied to both sets.
- **Fairness Assessment:** Fairness was evaluated using Fairlearn's `MetricFrame`, which assessed the baseline model's test set predictions across sensitive attribute groups. Key metrics included Accuracy, Precision, Recall, F1-score, Selection Rate, and False Positive Rate. This analysis revealed several notable disparities, including:
- **GENDER**: Recall difference of 18.0 points (z_F: 64.0%, M: 46.0%), selection rate difference of 7.4 points, and FPR difference of 3.0 points.
- **MSTATUS**: Recall disparity of 21.9 points (z_No: 67.7%, Yes: 45.8%), selection rate difference of 19.2 points, and FPR difference of 13.3 points.
- **EDUCATION**: There was a significant recall disparity of up to 36.2 points between groups.
- **OCCUPATION**: Extreme recall disparity of up to 60.0 points and selection rate differences up to 43.7 points.
- **URBANICITY**: Recall disparity of 44.1 points and selection rate difference of 32.1 points.
- **PARENT1**: Recall disparity of 35.0 points, selection rate difference of 37.1 points.

Disparities were quantified using the differences between maximum and minimum group metric values, highlighting areas requiring fairness mitigation.

- **Mitigation:** Mitigation used Fairlearn's ExponentiatedGradient with an EqualizedOdds constraint, applied separately to each sensitive attribute. Due to integration issues, SMOTE was excluded, and a simplified pipeline with StandardScaler and Logistic Regression was used. Models were trained on imbalanced data and evaluated with MetricFrame to assess fairness improvements.

4. Results

- **Baseline Model Performance & Fairness:** The initial Logistic Regression model achieved an overall Accuracy of ~76.1% and an F1-score (for claims) of ~38.3% on the test set.
    - **Fairness Assessment Summary:** The baseline model showed significant fairness disparities across all assessed attributes including:
        - **Recall (Equal Opportunity):** Max gaps of ~60pp (OCCUPATION), ~44pp (URBANICITY), ~36pp (EDUCATION), ~35pp (PARENT1)
        - **FPR (Equalized Odds):** Gaps up to ~33pp (OCCUPATION), ~29pp (PARENT1), ~22pp (URBANICITY).
        - **Selection Rate (Demographic Parity):** Differences of ~44pp (OCCUPATION), ~37pp (PARENT1), ~32pp (URBANICITY).
- **Mitigation Impact (ExponentiatedGradient with EqualizedOdds):**
    - **Fairness Improvement:** Applying mitigation individually generally reduced disparities targeted by Equalized Odds.
        - **Recall (TPR) Gaps:** Reduced by ~44pp for URBANICITY (to 0.2pp), ~28pp for PARENT1 (to 7.4pp), and ~19pp for MSTATUS (to 2.9pp). Larger gaps remained for OCCUPATION (40.5pp) and EDUCATION (20.8pp).
        - **FPR Gaps:** Dropped below 6pp post-mitigation, with ~1.0pp (MSTATUS), ~1.1pp (URBANICITY), ~2.1pp (GENDER), ~4.3pp (EDUCATION), ~4.6pp (PARENT1), and ~5.6pp (OCCUPATION).

Performance Impact: These fairness improvements were achieved with minimal negative impact on overall performance. Accuracy generally stayed similar (~75-78%), and F1 score (for claims) often slightly improved post-mitigation.

5. Standard: We recommend deploying individually mitigated models for attributes like URBANICITY, MSTATUS, and PARENT1, where fairness improved without harming performance. Fairness means not just excluding sensitive data, but also correcting its impact on outcomes. For features like OCCUPATION and EDUCATION, where disparities remained high, we suggest further review and stronger justification before use. Customers deserve transparency—especially when models rely on factors beyond their control.

6. Conclusion: The baseline Logistic Regression model showed significant fairness gaps, which were notably reduced—especially FPR disparities—using ExponentiatedGradient without major performance loss, particularly for binary attributes like URBANICITY, PARENT1, and MSTATUS. We recommend deploying individually mitigated models for these features and suggest future work explore multi-attribute mitigation and address persistent gaps in OCCUPATION and EDUCATION.