GENDER EQUALITY

Problem Statement

A Machine Learning model to analyze and predict gender pay gaps across industries, job roles, and regions. The goal of a project focused on analyzing and predicting gender pay gaps is driven by social, economic, and organizational imperatives aimed at promoting workplace equity and informed decision-making.

1. Project Overview

The objective of this project is to analyze and predict gender pay gaps across various industries, job roles, and regions. By leveraging machine learning, we aim to identify key factors that influence salary differences and quantify the gender pay gap. This analysis could support decisions in workforce planning and equal pay initiatives.

2. Dataset

The dataset used in this project is synthetic and includes the following columns:

- **Gender**: Gender of the employee (Male/Female).
- **Industry**: Industry type (e.g., Technology, Finance, Healthcare).
- **Job Role**: Specific role within the industry (e.g., Data Scientist, Financial Analyst).
- **Region**: Geographical region of employment (e.g., North America, Europe).
- **Years of Experience**: Total years of experience in the job role.
- **Education Level**: Highest educational qualification (e.g., Bachelor, Master, PhD).
- **Salary**: Annual salary in dollars (target variable).

3. Methodology

The project was conducted in the following steps:

3.1 Data Exploration

Using descriptive statistics and visualizations, we explored the dataset's structure, checking for missing values and distributions. Key observations included:

- **Gender Distribution**: Balanced representation across gender.
- **Salary Distribution by Gender**: Boxplots indicated significant differences between median salaries for different genders in certain industries, suggesting a potential gender pay gap.
- **Industry-Specific Pay Gaps**: Industry-based salary distributions varied, highlighting differences in pay structure across industries.

3.2 Data Preprocessing

Several preprocessing steps were applied to prepare the dataset for modeling:

1. **Encoding Categorical Variables**: Categorical columns (e.g., gender, industry) were encoded into numerical values using LabelEncoder.
2. **Feature Scaling**: Continuous features like "Years of Experience" and "Salary" were standardized to improve model performance.
3. **Train-Test Split**: The data was divided into training and testing sets in an 80-20 split.

3.3 Model Selection and Training

The XGBoost regressor was selected as the predictive model due to its efficiency and high performance in regression tasks. The model was trained with the following hyperparameters:

- **Objective**: reg:squarederror
- **Estimators**: 100
- **Learning Rate**: 0.1
- **Max Depth**: 6

3.4 Model Evaluation

The model's performance was evaluated using common regression metrics:

- **Mean Absolute Error (MAE)**: Measures average absolute errors.
- **Mean Squared Error (MSE)**: Penalizes larger errors more heavily than MAE.
- **Root Mean Squared Error (RMSE)**: Square root of MSE, providing error in original units.

4. Results

The XGBoost model's performance on the test set was as follows:

- **MAE**: ~\sim~ 0.20 (scaled values)
- **MSE**: ~\sim~ 0.06 (scaled values)
- **RMSE**: ~\sim~ 0.24 (scaled values)

These metrics indicate a reasonably accurate model, with relatively low error rates.

4.1 Feature Importance Analysis

The feature importance plot showed that:

- **Job Role** and **Years of Experience** were the most important factors, implying that role-specific experience contributes significantly to salary predictions.
- **Industry** and **Region** also played critical roles, with considerable variation in salary attributed to these factors.
- **Gender** had a less pronounced, though still noticeable, impact, confirming the existence of a measurable gender-based influence on salary.

4. Conclusion

This project successfully applied machine learning to analyze and predict gender pay gaps. Insights from the feature importance plot and salary distributions suggest that job role, experience, industry, and region are influential factors, while gender also affects pay disparities, although to a lesser extent.

6. Future Work

To enhance the model and findings, future work could:

1. Integrate real-world data for more accurate, generalizable insights.
2. Use advanced model tuning and ensemble techniques to improve predictive performance.
3. Expand the dataset to include additional features (e.g., company size, job level) that could affect pay gaps.