Machine Learning Project: Jobs and Skill Exchange

Objectives

1. **Regression**: Predict `salary_in_usd`.
2. **Classification 1**: Predict `employment_type` (FT, PT, CT, FL).
3. **Classification 2**: Predict `experience_level` (SE, MI, EN, EX).
4. **Unsupervised Learning**: Cluster jobs and visualize patterns.

Data Preparation

- **Loading**: Loaded dataset (607 rows, no missing values).
- **EDA**: Identified imbalances (96% FT, 43% SE).
- **Cleaning**: Removed duplicates (565 rows), standardized categories, capped outliers.
- **Correlation**: Analyzed numerical and categorical relationships.
- **Transformation**: One-hot encoded categorical features (171 features), scaled numerical features, added `job_title_freq`.

Regression Results

- **Simple Ridge Regression**: R^2 = -0.0610, RMSE = 60885.80, MAE = 49266.26
- **Multiple Ridge Regression**: R^2 = 0.5705, RMSE = 38735.92, MAE = 27231.73
- **Polynomial Ridge Regression**: R^2 = 0.5655, RMSE = 38964.21, MAE = 27672.89
- **Random Forest**: R^2 = 0.5260, RMSE = 40693.44, MAE = 28669.80
- **Gradient Boosting**: R^2 = 0.5688, RMSE = 38816.25, MAE = 26858.44
- **Best Model**: Multiple Ridge Regression.
- **Note**: Random Forest underperformed due to overfitting; Gradient Boosting close but not superior.

Classification Results

- **Employment Type**:
  - Ensemble: Accuracy = 0.5398, F1-Score = 0.6799, CV F1 = 0.9505 ± 0.0003
    - Per-Class F1: CT: 0.00, FL: 0.00, FT: 0.7018, PT: 0.1667
  - Random Forest: Accuracy = 0.5133, F1-Score = 0.6569, CV F1 = 0.9460 ± 0.0054
    - Per-Class F1: CT: 0.00, FL: 0.00, FT: 0.6786, PT: 0.1333
  - **Note**: Minority classes (CT/FL) poorly predicted due to test set imbalance (109 FT, 1 CT, 1 FL, 2 PT).
- **Experience Level**:

    - Random Forest: Accuracy = 0.5929, F1-Score = 0.5950, CV F1 = 0.5784 ± 0.0332
        - Per-Class F1: EN: 0.5238, EX: 0.4615, MI: 0.5405, SE: 0.6804
    - Ensemble: Accuracy = 0.5929, F1-Score = 0.5953, CV F1 = 0.5660 ± 0.0475
        - Per-Class F1: EN: 0.4615, EX: 0.4615, MI: 0.5526, SE: 0.6939
    - **Note**: Balanced performance; pending Logistic Regression and Gradient Boosting results.
- **Best Models**: Ensemble (employment_type), Random Forest (experience_level).
- **Mitigation**: Used oversampling, class weights, RandomizedSearchCV.

Unsupervised Results

- **K-Means**: k = 4, Silhouette Score = 0.8561 (pending actual run)
- **PCA**: Explained Variance = 0.9986 (pending actual run)
- **Insights**: Clusters likely represent job types/seniority.

Recommendations

- **Regression**: Use Multiple Ridge Regression.
- **Classification**: Use Ensemble for `employment_type` (improve PT/CT/FL with adjusted oversampling); Random Forest for `experience_level`.
- **Unsupervised**: Leverage clusters for market analysis.

Visualizations

- Numerical distributions: `numerical_distributions.png`
- Salary boxplot: `salary_boxplot_capped.png`
- Categorical counts: `categorical_counts.png`
- Correlation matrix: `correlation_matrix.png`
- Confusion matrices: `cm_*.png`
- ROC curves: `roc_*.png`
- Regression predictions: `regression_predictions.png`, `rf_regression_predictions.png`
- PCA clusters: `pca_clusters.png` (pending)