Data Science Salary Prediction Project

Uta Nishii and Kyllan Wunder

Project Overview

This project aims to predict data science salaries based on various features such as job title, location, and experience level. We used a dataset from Kaggle to train and evaluate our models.

Feature Engineering

In our final data pipeline, we applied several targeted feature engineering steps to transform the raw salary data into a format optimal for modeling:

1. **Job Title Standardization and Grouping**:
   - Developed a custom mapping function to consolidate similar job titles into broad categories (e.g., grouping various forms of "Data Scientist", "Data Engineer", and "Machine Learning Engineer").
   - Applied one-hot encoding to the cleaned job titles (stored in `job_title_clean`) to capture categorical nuances.
2. **Cumulative Experience Encoding**:
   - Transformed the ordinal `experience_level` feature into cumulative boolean indicators (e.g., `experience_level_EN`, `experience_level_MI`, `experience_level_SE`, and `experience_level_EX`) to reflect career progression from entry-level to expert.
3. **Employment Type and Company Size Encoding**:
   - One-hot encoded categorical features such as `employment_type` and `company_size` to enable their use in modeling.
4. **Location-Based Feature Engineering**:
   - Created a new binary feature `same_country` to indicate whether the employee's residence matches the company's location, capturing regional salary variations.
5. **Additional Categorical Encoding**:
   - Applied label encoding to remaining categorical features (e.g., `employee_residence` and `company_location`) to convert them into numerical representations.

These transformations produced a well-structured, encoded dataset, which was then saved as `encoded_data.csv` for further model training and evaluation.

Model Exploration

We experimented with a variety of regression models to predict salaries (in USD) using our engineered features. Below is a summary of our findings:

1. Decision Tree Regressor

- **Performance:**
  - Mean Squared Error (MSE): ~2.59e9
  - R-squared (R^2): ~0.28
  - Mean Absolute Error (MAE): ~36,936
- **Key Observations:**
  - Top features included Employee Residence, Job Title Data Analyst, and Senior Experience Level.

2. Linear Regression

- **Performance:**
  - Mean Squared Error (MSE): ~2.28e9
  - R-squared (R^2): ~0.37
  - Mean Absolute Error (MAE): ~37,523
- **Key Observations:**
  - Significant coefficients were observed for features such as Expert Experience Level, AI Scientist Job Title, and Research Scientist Job Title.

3. Random Forest Regressor

- **Performance:**
  - Mean Squared Error (MSE): ~2.15e9
  - R-squared (R^2): ~0.41
  - Mean Absolute Error (MAE): ~35,351
- **Key Observations:**
  - Important features included Employee Residence, Senior Experience Level, and Job Title Data Analyst.

4. **Gradient Boosting Regressor**

- **Initial Results:**
  - Mean Squared Error (MSE): ~2.02e9
  - R-squared (R^2): ~0.44
  - Mean Absolute Error (MAE): ~34,964
- **After Hyperparameter Tuning:**
  - Tuned parameters (e.g., $learning\_rate = 0.2$, $max\_depth = 6$, $n\_estimators = 50$, $min\_samples\_split = 20$, $min\_samples\_leaf = 1$) yielded comparable performance with Mean Squared Error (MSE) ~2.10e9, R-squared (R^2) ~0.42, and Mean Absolute Error (MAE) ~35,314.
- **Key Observations:**
  - Feature importance consistently highlighted Employee Residence, Senior Experience Level, and Job Title Data Analyst.

5. **Support Vector Regression (SVR)**

- **Performance:**
  - Mean Squared Error (MSE): ~3.61e9
  - R-squared (R^2): ~0.001
  - Mean Absolute Error (MAE): ~47,589
- **Key Observations:**
  - SVR underperformed compared to the other models, suggesting it may not be well-suited for this dataset without significant tuning.

6. **XGBoost Regressor**

- **Initial Results:**
  - Mean Squared Error (MSE): ~2.19e9
  - R-squared (R^2): ~0.40
  - Mean Absolute Error (MAE): ~35,730
- **After Hyperparameter Tuning:**
  - Tuning with various parameter grids improved performance. One configuration achieved Mean Squared Error (MSE) ~2.07e9, R-squared (R^2) ~0.44, and Mean Absolute Error (MAE) ~35,000.
- **Key Observations:**
  - Top features for XGBoost included Job Title Data Analyst, Employee Residence, and Senior Experience Level.

Summary and Observations: - Ensemble methods (Gradient Boosting and XGBoost) showed superior predictive power over simpler models. - Hyperparameter tuning via grid search contributed to performance improvements by refining key parameters. - Features such as Employee Residence and aspects of job title and experience levels were consistently important across models, underscoring their relevance in predicting data science salaries.

Final Model

Our final selected model is an XGBoost regressor configured with the following parameters: `{python} colsample_bytree: 0.8 learning_rate: 0.1 max_depth: 5 min_child_weight: 1 n_estimators: 75 subsample: 0.8`

This model achieved the following performance metrics on the test set:

- **Mean Squared Error (MSE):** 2.02e9
- **R-squared (R^2):** ~0.44
- **Mean Absolute Error (MAE):** ~34,731.91

Feature importance analysis showed that key predictors included features such as Senior Experience Level, Data Analyst Job Title, Employee Residence, Expert Experience Level, and Company Location.

Bootstrapped Prediction Intervals

To assess the uncertainty in our predictions, we performed bootstrap resampling ($n = 200$) on the training data to generate 95% prediction intervals for each test point. Below is a sample table showing the median predictions along with the lower and upper bounds of the 95% confidence intervals for 10 test points:

Median Prediction	Lower Bound	Upper Bound
164981.39	160134.59	170268.42
217269.36	157434.42	253222.79
131459.73	123213.11	140212.94
180258.56	169015.77	193376.33
37127.55	8867.46	63867.34
69534.13	50980.96	90188.41
131459.73	123213.11	140212.94
102085.18	87544.72	120380.70
73648.38	67299.25	79629.63
97997.09	79318.91	116376.80

These intervals provide a measure of the reliability of our predictions. Overall, the final model demonstrates robust performance and offers useful prediction intervals for decision-making.