

Project report

Objective

The objective is to predict the salary in USD (salary_in_usd) based on variables related to the job, experience level, location, and company characteristics. Two regression models were compared:

- **Linear Regression**
- **Random Forest Regressor**

The performance of the models was mainly evaluated using:

- R^2 (coefficient of determination)
- RMSE (Root Mean Squared Error)

Data and Preprocessing

The dataset includes variables such as: work_year, experience_level, employment_type, job_title, salary_currency, employee_residence, remote_ratio, company_location, company_size.

The preprocessing steps included:

- Selecting relevant variables
- Handling missing values
- Encoding categorical variables
- Normalizing the data
- Splitting the dataset (75% train / 25% test)

Methodology

Linear Regression

Tested all possible combinations of 4 explanatory variables

Evaluated using R^2 and RMSE

Random Forest

Hyperparameter optimization via GridSearchCV (n_estimators, max_depth, min_samples_split)

Evaluation using the same metrics (R^2 , RMSE)

Results

- Linear Regression
- Very poor performance
- R^2 often negative in cross-validation

- High RMSE

=> The model does not adequately explain salary variance

=> Likely non-linearity and multicollinearity

Random Forest

- Significantly better performance
- Mean $R^2 \approx 0.395$ (~39.5% of variance explained)
- Reasonable RMSE
- Best variable combination: experience_level, job_title, employee_residence, remote_ratio

=> Random Forest better captures the complex relationships in the dataset

Conclusion

Linear regression is not suitable for the studied data, likely due to non-linear relationships and imposed constraints.

The Random Forest model is the most effective and would be used for future predictions.

Potential improvements include:

- Advanced feature engineering
- Hyperparameter optimization with Bayesian search
- Enhanced cross-validation
- Testing more complex models (Boosting, Neural Networks)

Appendices

Without hyperparameter optimization and with an 80/20 split:

- Random Forest achieves approximately 40% variance explained
- Best combination: experience_level, job_title, employee_residence, company_location

Scripts provided:

pipeline.py: preprocessing

main.py: training + evaluation

Detailed results: Results.csv

Additional tests: Annexe_Results.csv