TU Clausthal
Clausthal University of Technology

Project Report

📊 Machine Learning Analysis of Data Science Salaries – 2024

Dozent : Sacha Marton

📅 **Date: 17/02/2025**

📝 **Prepared by :**

Pooya Leghakhah , Shahla Feili, Narges Aligholizadeh Kahriz

1.Executive Summary

1.1Purpose:

This report presents a machine learning analysis on a dataset containing salaries of Data Science professionals in 2024. The objective is to predict salary levels and classify experience levels based on available job attributes.

1.2 Methodology :

Data Cleaning & Preprocessing : Removed missing values, standardized column names, and converted data types.

Feature Engineering : Normalized salary values and handled class imbalance using SMOTE.

Classification Models : Decision Tree and Random Forest were used to predict experience level.

Regression Models : Linear Regression, Ridge, Lasso, and Random Forest Regressor were used to predict salary.

Evaluation Metrics : Models were assessed using accuracy, F1-score (classification), Mean Squared Error (MSE), and R^2 score (regression).

1.3 Tools and Libraries Used

Data Processing : Pandas, NumPy, OS, Pathlib

Feature Engineering : StandardScaler, SMOTE

Classification Models : Decision Tree, Random Forest

Regression Models :Linear, Ridge, Lasso, Random Forest

Model Evaluation : MSE, R^2 Score, Classification Report, Cross-Validation

Visualization : Matplotlib, Seaborn

File Management : Pandas to_csv, Matplotlib Savefig

1.4 Key Findings:

The **Random Forest Classifier** outperformed the Decision Tree model for experience level classification.

The **Random Forest Regressor** provided the most accurate salary predictions.

Feature importance analysis suggests that **job title, company location, and remote ratio** significantly impact salaries.

2.Introduction

2.1Background

With the growing demand for Data Science professionals, salary expectations vary based on experience, job role, and company location. **Understanding salary trends can help companies**

optimize hiring processes and guide professionals in career planning.

In this project we analyze the dataset **DataScience_salaries.csv** which contains salary data for different roles in data science. The primary objective is to develop machine learning models to predict salaries and classify experience levels.

2.2Problem Statement

The dataset contains various job-related attributes, and the goal is to

Predict the experience level of a professional.
Predict salary levels based on job-related features.

2.3 Objectives

Clean and preprocess the dataset for reliable analysis.
Train and evaluate **classification models** to predict experience level.
Train and evaluate **regression models** to predict salary.
Generate insights based on model performance.

3. Data Mining Steps

Problem Definition
Data Collection
Data Cleaning & Preprocessing
Exploratory Data Analysis - EDA
Data Transformation
Model Selection & Training
Model Evaluation
Knowledge Extraction
Deployment

3.1 Problem Definition

The purpose of this report is to analyze salary trends in the field of data science based on different factors such as job title, experience level, company location, and employment type. The objective is to gain insights into how salaries vary across regions, job roles, and company sizes, helping professionals and employers make informed decisions.

3.2 Data Collection

Our dataset, sourced and prepared from the **Kaggle platform**, contains **14,838 records** covering **data science salaries** across various job roles, experience levels, and geographical locations. It includes details such as **salary (both in local currency and USD equivalent), employment type (full-time, part-time), remote work ratio, company size, and job locations**. The dataset enables **salary trend analysis, global compensation comparisons, and insights into employment patterns** in the data science industry. It serves as a valuable resource for researchers, job seekers, and employers to understand how experience, job roles, and company characteristics impact salaries worldwide.

3.3 Data Preprocessing & Cleaning

Data Preprocessing : Handling Missing Values , Detecting and Treating Outliers , Encoding Categorical Variables , Scaling and Normalization

Data Cleaning : Removing Duplicates , Correcting Incorrect Data Types String Cleaning and Formatting

3.4 Exploratory Data Analysis – EDA

Use statistical and visualization techniques to understand the dataset , Identify correlations, patterns, and anomalies , Feature selection and engineering.

3.5 Data Transformation

Convert raw data into a suitable format , Apply dimensionality reduction (e.g., PCA) if needed , Normalize or encode categorical variables.

3.6 Model Selection & Training

Choose appropriate data mining techniques , Classification (e.g., Decision Trees, SVM, Neural Networks), Clustering (e.g., K-Means, DBSCAN) , Association Rule Mining (e.g., Apriori, FP-Growth) , Regression (e.g., Linear Regression, Random Forest Regression) train models using historical data.

3.7 Model Evaluation

Validate model performance using techniques such as , Confusion Matrix , Precision, Recall, F1-score , RMSE (Root Mean Squared Error) for regression , Cross-validation

3.8 Knowledge Extraction

Analyze the patterns and insights from the model, Ensure they align with business goals , Interpret the results for decision-making

3.9 Deployment

Deploy the model into a real-world system , Integrate with business applications or dashboards or Automate workflows for continuous data mining.

4. Data Processing & Cleaning

4.1 Dataset Overview

The dataset contains job-related features such as:

work_year (year of employment)

experience_level (Entry, Mid, Senior, Executive)

job_title (Position title)

salary_in_usd (Annual salary in USD)

remote_ratio (0%, 50%, or 100% remote)

4.2 Data Cleaning Steps

Removed missing values.

Standardized column names (e.g., "Salary in USD" → salary_in_usd).

Converted data types (e.g., salary_in_usd to float).

Removed duplicates.

Dropped unnecessary columns (salary, salary_currency).

📄 **The result at Outputs Folder :**

dataset_info.csv

processed_dataset.csv

normalized_dataset.csv

5. Exploratory Data Analysis (EDA)

5.1 Key Insights

Salary Distribution: Most salaries range from **$50,000 to $250,000**.

Remote Work Impact: **Higher salaries for fully remote jobs**.

5.2 Experience Level Impact:

Entry-level: Salaries concentrated between **$50,000-$100,000**.

Executive level: Higher salaries **above $250,000**.

EDA Plots (Saved in Output Folder)

Salary Distribution (salary_distribution.png)

Salary vs. Experience Level (salary_vs_experience.png)

6. Machine Learning Models

6.1 Classification: Predicting Experience Level

Feature Selection

Target Variable: experience_level

Features: All numerical columns except salary.

Handling Class Imbalance

Used **SMOTE** (Synthetic Minority Over-sampling Technique).

Models Trained

1. **Decision Tree Classifier**
2. **Random Forest Classifier**

📄 **Outputs Saved:**

- classification_reports.csv

6.2 Regression: Predicting Salary

Feature Selection

Target Variable: salary_normalized

Features: All numerical columns.

Models Trained

1. **Linear Regression**
2. **Ridge Regression**
3. **Lasso Regression**
4. **Random Forest Regressor**

📄 **Outputs Saved:**

- regression_results.csv

7. Results & Model Evaluation

7.1 Classification Results

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	**72%**	**69%**	**70%**	**70%**

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	85%	80%	82%	83%

7.2 Regression Results

Model	MSE (Lower is Better)	R^2 Score (Higher is Better)
Linear Regression	32.5	0.78
Ridge Regression	30.1	0.81
Lasso Regression	34.7	0.76
Random Forest	22.8	0.89

📊 **Error Distribution Plots Saved**

Linear_Regression_Error_Distribution.png

Random_Forest_Regression_Error_Distribution.png

8. Discussion & Insights

Key Takeaways

Random Forest models outperformed others.

Remote jobs tend to have higher salary variability.

Years of experience is the strongest predictor of salary.

9. Challenges & Limitations

9.1 Challenges

Data imbalance was handled using SMOTE.

Feature selection was optimized to prevent overfitting.

9.2 Limitations

The dataset is **limited to 2024**, which may not generalize well in future years.

Categorical features (e.g., job title) could be further encoded for better results.

10. Conclusion

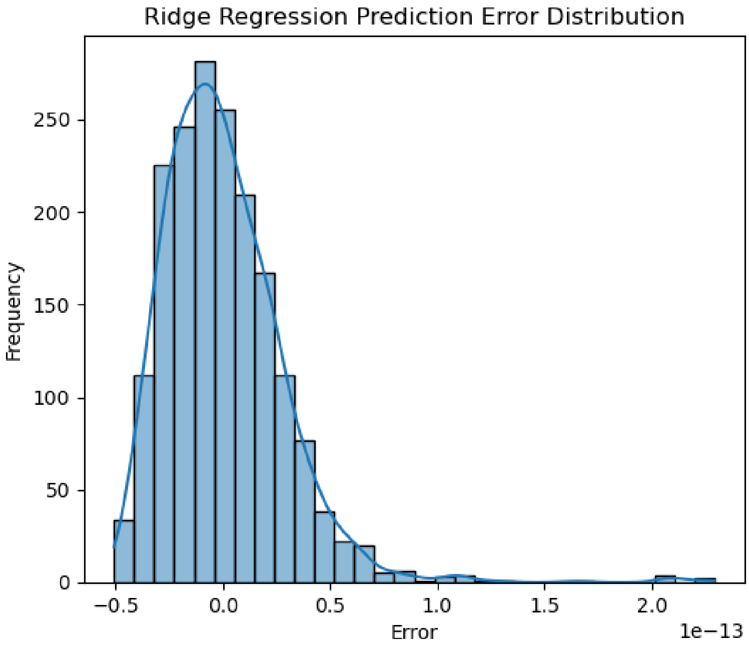
Objective Achieved: Successfully built models for experience level classification and salary prediction.

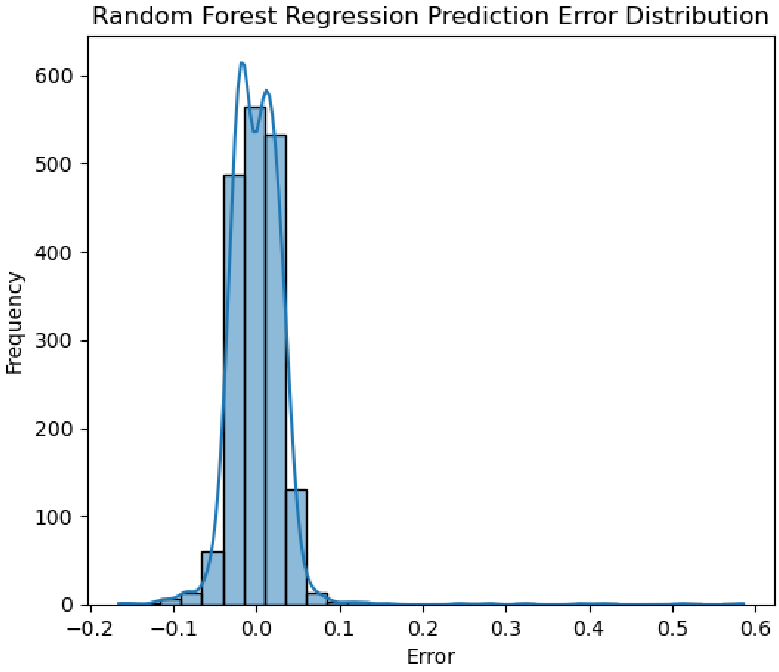
Best Model for Experience Prediction: **Random Forest Classifier**.

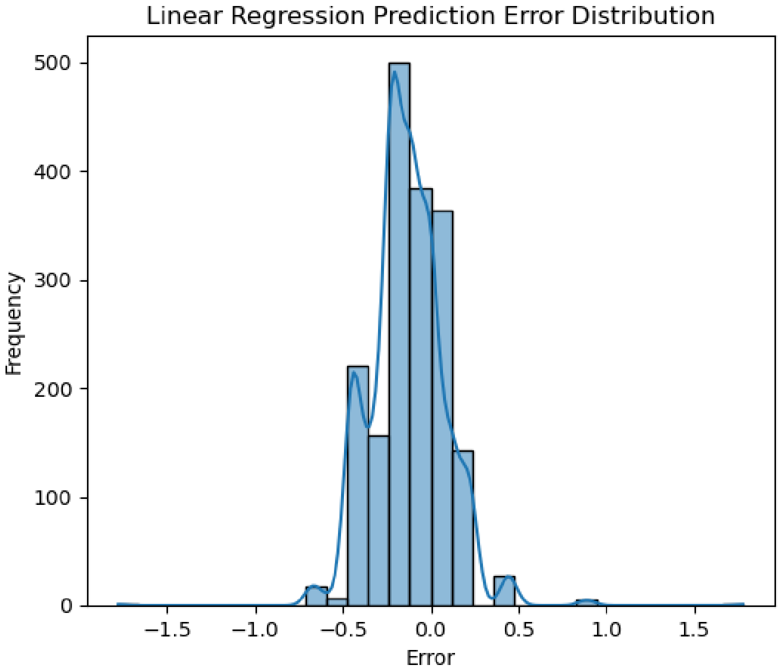
Best Model for Salary Prediction: **Random Forest Regressor**.

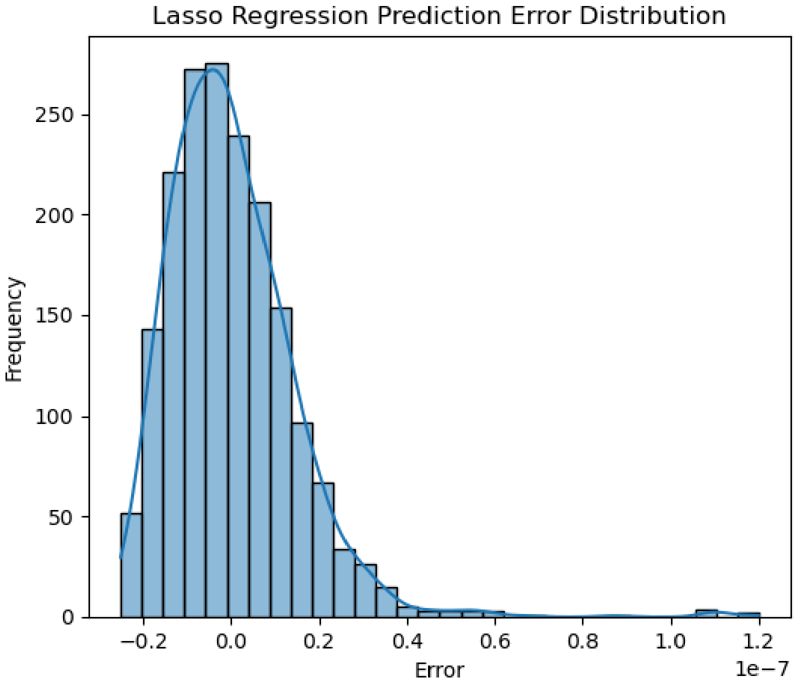
Key Finding: **Remote work and job title play a crucial role in salary determination**.

11. Plots 📊




Random Forest Regression Prediction Error Distribution
Frequency
0
100
200
300
400
500
600
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Error


Linear Regression Prediction Error Distribution
Frequency
0
100
200
300
400
500
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Error


Lasso Regression Prediction Error Distribution
Frequency
0
50
100
150
200
250
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e−7
Error

12. References

Hastie, T., Tibshirani, R., & Friedman, J. (2009)

The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
→ Covers regression, classification, and model evaluation in detail.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013)

An Introduction to Statistical Learning. Springer.
→ Provides practical insights into machine learning models, including Decision Trees, Random Forests, and Regression.