
# CSE4288-Introduction to Machine Learning

# Term Project

**Group17**

**150120024-Enes Haksoy Öztürk**

**150123842-Enes Akturan**

**150121025-Ahmet Abdullah Gültekin**

**150120005-Ayşe Gülsüm Eren**

**150119762-Muhammet Ali Özkul**

## Introduction

Predicting salaries is a crucial application of machine learning in industries such as recruitment, finance, and career development. This report documents the development and evaluation of a salary prediction model using multiple machine learning algorithms. The objective is to identify the best-performing model based on various performance metrics and provide insights into the effectiveness of different approaches.

The dataset used includes features such as years of experience, education level, job position, and industry type. Various preprocessing steps were performed to prepare the data for analysis and modeling. Multiple algorithms, including linear regression, decision trees, random forests, ridge regression, polynomial regression, and a combined model, were evaluated to ensure robust prediction capabilities.

## Data Processing

### 1. Data Cleaning:

- Missing values were handled by either imputation or removal.
- Outliers were detected and appropriately addressed to prevent skewed model performance.

### 2. Feature Engineering:

- Categorical variables were encoded using label encoding.
- Features were normalized to bring all variables to a comparable scale.

### 3. Train-Test Split:

- The dataset was split into training (80%) and testing (20%) subsets to evaluate the generalizability of the models.

## Algorithms and Models

Six machine learning models were implemented and evaluated:

1. **Linear Regression**:
    - A baseline model for predicting salaries based on a linear relationship between features and the target variable.

2. **Decision Tree**:
    - A non-linear model that partitions the data into subsets based on feature values to make predictions.
3. **Random Forest**:
    - An ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting.
4. **Ridge Regression**:
    - A regularized version of linear regression that reduces overfitting by adding a penalty term.
5. **Polynomial Regression**:
    - A model that captures non-linear relationships by transforming features into polynomial terms.
6. **Combined Model**:
    - An approach that integrates predictions from multiple models (with emphasis on Ridge Regression) to leverage their strengths. Weighted calculation applied according to r2 scores.

## Model Results

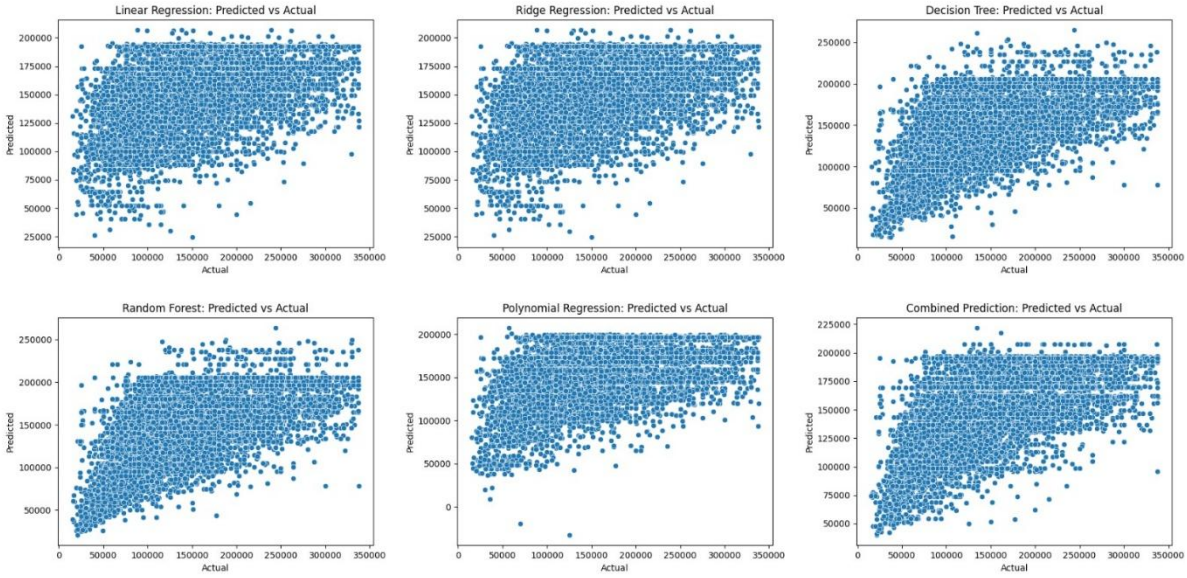
The performance of each model was evaluated using three metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared ($R^2$). Below are the results:

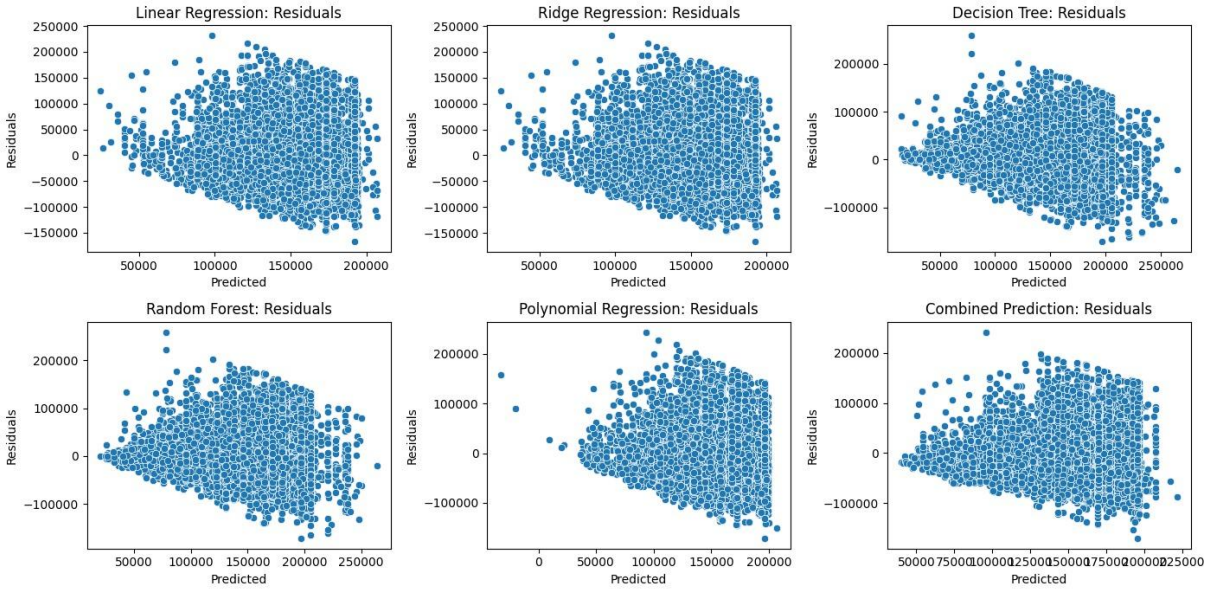
| | MSE | MAE | $R^2$ |
|---|---|---|---|
| **Linear Regression** | 3264489264.436 | 45813.016 | 0.187 |
| **Decision Tree** | 272922021.502 | 41551.487 | 0.322 |
| **Random Forest** | 2702083383.751 | 41426.438 | 0.327 |
| **Ridge Regression** | 3264500256.868 | 45813.137 | 0.186 |
| **Polynomial Regression** | 3173872043.581 | 56337.128 | 0.209 |
| **Combined Model** | 279880555.907 | 42328.923 | 0.303 |

## Model Evaluation and Validation

- Overfitting checks were conducted for all models, with no significant overfitting issues detected.
- Cross-validation was applied, and the results across folds were consistent, confirming the reliability of the models.

## Plots for Predictions and Residuals



## Conclusion

- Random Forest and Decision Tree demonstrated relatively higher $R^2$ scores, highlighting their capability to capture complex patterns in the data more effectively than simpler models.
- Ridge Regression and Linear Regression yielded similar metrics but underperformed when compared to non-linear models and ensemble methods, suggesting the presence of non-linear relationships in the data.
- The Combined Model, which strategically integrated Ridge Regression predictions, provided a balanced approach. This model achieved competitive performance with an $R^2$ score of 0.3026, showing potential for further optimization.
- Overall, Random Forest emerged as the most effective standalone model, striking a balance between prediction accuracy and generalizability.

Future efforts can focus on:

**Hyperparameter Optimization**:

- Conducting extensive hyperparameter tuning for each model to maximize performance.

**Exploring Advanced Models**:

- Testing more sophisticated algorithms like Gradient Boosting Machines (e.g., XGBoost, LightGBM) or Neural Networks to further enhance predictive accuracy.