The Future of AI Careers: Emerging Roles, Skill Demands, and Salary Trends

Samuel Fandino[*]

Department of Computational Mathematics, Science and Engineering

Michigan State University, East Lansing, MI 48824

(Dated: November 1, 2025)

Abstract

The rapid expansion of artificial intelligence (AI) is reshaping the global job market, leading to the emergence of new roles, evolving skill requirements, and shifting salary patterns. This project explores a global dataset of AI-related job postings to uncover trends in employment, skill demand, and compensation across different industries and regions. Through data preprocessing, exploratory analysis, and baseline modeling, the project aims to understand how factors such as education, experience, and job type influence salaries and opportunities in the AI field. Early findings suggest meaningful relationships between experience levels, job locations, and compensation, as well as a growing demand for professionals with specialized AI expertise. By analyzing these dynamics, the project seeks to provide valuable insights for both job seekers and organizations—helping people align their skill development with market needs and helping companies plan for future talent demands.

BACKGROUND AND MOTIVATION

The AI and machine learning (ML) job market is rapidly evolving, with new roles, skill requirements, and salary trends constantly emerging. Understanding these shifts is important for both job seekers and organizations: professionals need to know which skills are in demand to advance their careers, while companies need insight into workforce trends to plan recruitment and training strategies effectively. Without a clear picture of the changing landscape, people risk investing in skills that are becoming obsolete, and organizations may face talent shortages or misaligned hiring priorities.

Many analyzes have explored the trends of the AI workforce through surveys or job postings, revealing general patterns in skill demand and salary levels. However, these studies tend to be descriptive rather than predictive and rarely apply machine learning systematically to detect hidden patterns, clusters of emerging roles, or factors influencing compensation. This gap presents an opportunity to use large-scale, global datasets alongside ML techniques to gain more detailed and actionable insights.

The goal of this project is to uncover trends in AI-related roles, skills, and salaries in different regions and industries. employing machine learning methods such as clustering, unsupervised learning, and predictive modeling, the project aims to estimate both salary ranges and identify emerging groups of roles. The insights generated can help professionals develop relevant skills, help organizations plan talent strategies, and provide a data-driven understanding of the evolving AI job market.

DATA DESCRIPTION

Data Origins

The dataset used in this project is titled “Global AI Job Market Salary Trends 2025”, created and uploaded by Bisma Sajjad on Kaggle. (kaggle.com) Unlike many scraped or survey-based collections, this data set was synthetically generated to realistically simulate the patterns observed in the global AI and machine learning job market. It was designed to reflect key variables such as salary (in USD and local currency), years of experience, education requirements, job titles, company size, and regional variation.

The synthetic data was generated using statistical modeling and sampling techniques to approximate real distributions found in public labor reports and AI workforce studies. Its purpose is to enable safe and reproducible analysis of employment and compensation trends in the AI sector without privacy or licensing concerns.

Dataset Characteristics

- Number of samples (rows): 15,000
- Number of features (columns): 20
- Data types: Numerical and Categorical
- Target variable: Initially, the target variable is the salary in USD

Data Quality Analysis

Missing Values

The dataset is well-organized and does not contain missing values in its original form. However, missing data was deliberately introduced completely at random to simulate a more realistic scenario, reflecting challenges encountered in practical machine learning applications. Descriptive statistics show a wide salary range and considerable variability in experience and remote work ratios. Outliers exist in salary-related attributes, which aligns with the global scope of the data.

Class Balance

No single category overwhelmingly dominates any feature, indicating that the dataset does not suffer from major class imbalance issues.

Statistical Summary

The following figures exhibit some of the dataset properties such as the correlation between numerical features (FIG. 1) and the average salary by remote work ratio and experience level (FIG. 2).

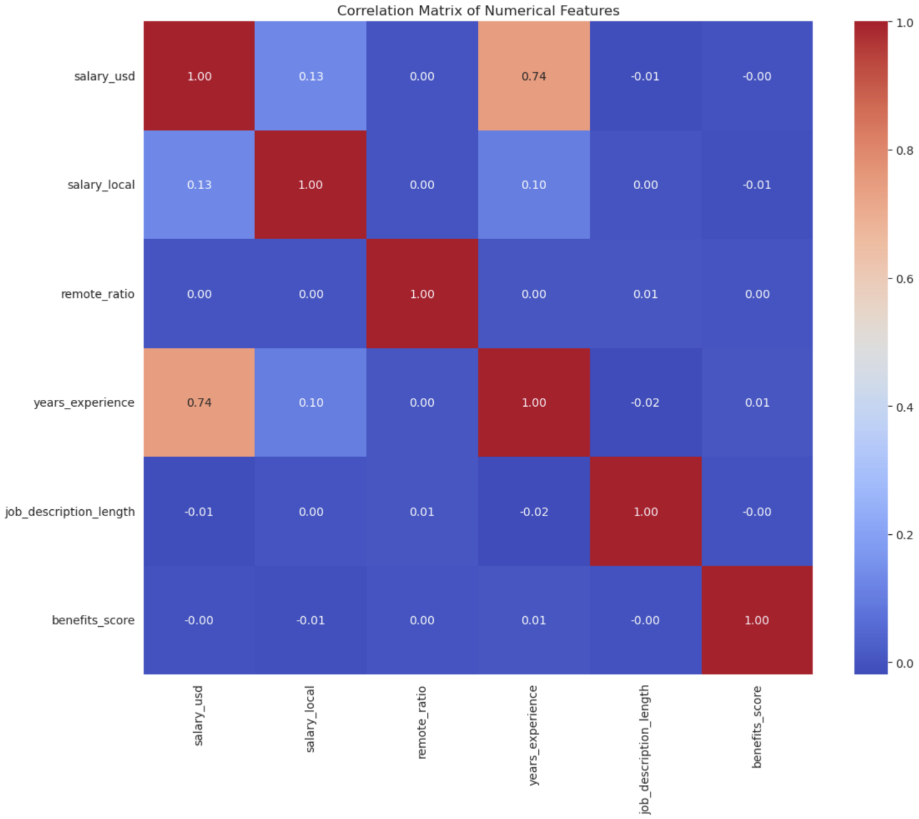

FIG. 1: Correlation Matrix of Numerical Features

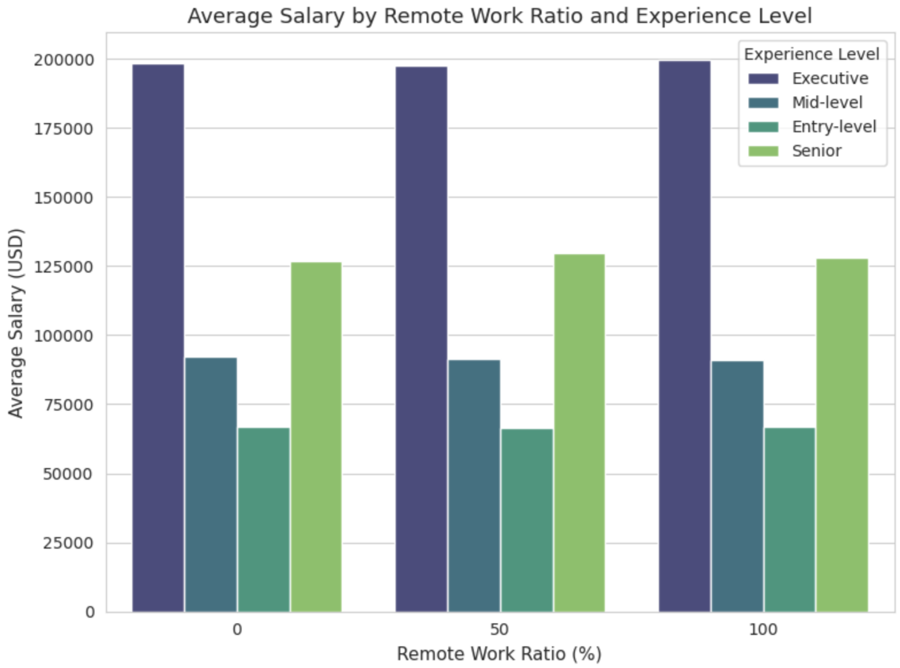

FIG. 2: Average Salary by Remote Work Ratio and Experience Level

PREPROCESSING

Before implementing any machine learning models, the dataset was carefully preprocessed to guarantee consistency, interpretability, and suitability for analysis. A structured workflow was followed, which included data splitting, feature engineering, and various scaling, transformation, and encoding techniques. These processes ensured that the data was properly cleaned, categorical variables were transformed into numerical representations, and feature ranges were standardized—ultimately preparing the dataset for both clustering and regression-based modeling.

Data Splitting

The dataset was divided into training and testing subsets to ensure unbiased model evaluation. A random splitting strategy was used, assigning 80 percent of the data to the training set and 20 percent to the testing set. This approach is suitable because the dataset is synthetic and non-temporal, meaning there is no time dependency between samples. Additionally, since the dataset is generally balanced between job categories, stratification was not necessary.

Feature Engineering

Feature engineering was a crucial step in refining the dataset for modeling and exploratory analyses. Several irrelevant or redundant variables were removed, including application date, salary local, and salary currency, since all compensation values were standardized in U.S. dollars (salary usd). The posting date field was converted into a proper datetime format, enabling potential time-based analyses such as studying hiring trends over different months or years. Features like job description length were also excluded, as preliminary analysis suggested limited predictive power. Beyond cleaning, potential feature creation and extraction were explored to enhance model performance. For instance, clustering techniques such as K-Means may later be applied to group job postings into similar role families based on shared skill sets, required experience, or industry sector. Polynomial transformations could also be introduced to capture nonlinear relationships between experience level and salary. These transformations aim to increase the expressiveness of the models and uncover hidden structures within the data.

Scaling, Transformation, and Encoding

Numerical variables such as salary usd and years experience were scaled using StandardScaler to ensure consistent feature ranges and prevent model bias toward higher-magnitude variables. Categorical features including experience level, employment type, and company size were encoded using one-hot encoding, converting them into binary vectors suitable for regression and clustering algorithms. Missing values, which were synthetically introduced to better simulate real-world conditions, were handled using SimpleImputer. Mean imputation was applied for numerical variables, while categorical variables were imputed with the mode. These transformations ensured data consistency, minimized bias, and allowed the models to process all samples without information loss.

MACHINE LEARNING TASK AND OBJECTIVE

This section focuses on the machine learning aspect of the project.

Why Machine Learning?

Machine learning is essential for analyzing complex and dynamic labor markets such as the global AI and technology job sector. Traditional analytical methods often rely on simple statistical summaries or manual categorization, which fail to capture nonlinear patterns and evolving relationships among roles, skills, and compensation. With thousands of job postings containing both numerical and categorical variables, human-driven analysis would be inefficient and prone to bias.

Task Type

This project primarily involves supervised learning and unsupervised learning approaches:

- **Supervised Learning:**
  - Regression: The regression task focuses on predicting salary levels based on various job attributes such as experience level, required skills, company size, and location. This helps estimate fair compensation for given profiles and identify factors most strongly associated with pay.
- **Unsupervised Learning:**
  - Clustering: Unsupervised methods like K-Means will later be explored to identify natural groupings among job postings—revealing emerging role families or skill clusters across the AI job market.

MODELS

Describe the machine learning models you will compare. You need at least three models in increasing order of complexity.

Model Selection

Model 1: [Linear Regression]

Linear Regression was selected as the baseline model due to its simplicity and interpretability. It assumes a linear relationship between the predictors (e.g., experience level, company size, employment type) and the target variable (salary usd). Despite its limitations in capturing nonlinear trends, Linear Regression provides an essential foundation for understanding the directional impact of each feature on salary outcomes.

The coefficients derived from the model help quantify how much salary is expected to increase or decrease with a unit change in a given feature while holding others constant. For instance, we can observe how transitioning from a mid-level to a senior-level position impacts salary, or how company size correlates with compensation.

Model 2: Random Forest Regressor

To overcome the linearity limitation, a Random Forest Regressor was introduced as an intermediate-level model. Random Forest is an ensemble learning method that constructs multiple decision trees using bootstrapped samples of the dataset and averages their predictions to reduce variance and improve accuracy.

This approach is particularly effective for the AI job market dataset because relationships between salary and factors such as experience, location, and employment type are often nonlinear and interactive. For example, the salary gap between entry-level and senior positions may vary dramatically across different company sizes or regions — patterns a linear model cannot easily capture.

Moreover, Random Forest inherently handles both numerical and categorical data well.

Model 3: Gradient Boosting Regressor

The Gradient Boosting Regressor serves as the most advanced model in this study, built to maximize predictive performance through iterative learning. In contrast to Random Forests, which train multiple trees independently, Gradient Boosting constructs trees sequentially—each new tree focuses on minimizing the remaining errors from the previous ones. This step-by-step refinement enables the model to uncover complex relationships and subtle feature interactions that simpler models might fail to detect.

Regularization and Hyperparameter Tuning

To improve model generalization and prevent overfitting, regularization techniques and careful hyperparameter tuning were applied across all models. For the linear regression baseline, L2 regularization (Ridge regression) was considered to penalize large coefficients and reduce model variance. For the Random Forest model, hyperparameters such as the number of trees, maximum tree depth, and minimum samples per leaf were tuned using grid search to balance model complexity and predictive performance.

For the Gradient Boosting Regressor, key hyperparameters including learning rate, number of estimators, and maximum depth of each tree were systematically adjusted.

EVALUATION FRAMEWORK

To evaluate the performance of our machine learning models in predicting salaries in the AI job market, we use both primary and secondary metrics that quantify prediction accuracy and model reliability.

Primary Metric

The primary metric for this regression task is the Root Mean Squared Error (RMSE). RMSE measures the average magnitude of prediction errors and penalizes larger errors more heavily. It is particularly useful in salary prediction because large deviations from the true salary are more critical than small deviations. Lower RMSE values indicate better predictive performance and a closer fit to the observed data.

Secondary Metrics

In addition to RMSE, we use the Mean Absolute Error (MAE) as a secondary metric. MAE provides a complementary view of prediction accuracy by averaging the absolute differences between predicted and actual values. Unlike RMSE, MAE treats all errors equally, giving a straightforward measure of average prediction error in the same units as the target variable (USD).

Metric Definitions

The mathematical definitions of the metrics used are as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

where y_i is the true salary for the i-th job posting, $\hat{y}_i$ is the predicted salary, and n is the total number of samples. These metrics together provide a robust assessment of model performance for salary prediction.

Train/Validation/Test Split Strategy

To evaluate model performance and ensure generalization, the dataset was divided into three subsets: training (70%), validation (15%), and testing (15%). The split was performed using random sampling to preserve the overall distribution of salaries, experience levels, and job categories. Since the data are not time-dependent, a random split was preferred over a time-based or stratified approach. The validation set was used for hyperparameter tuning, while the test set provided an unbiased assessment of final model performance.

Baseline Comparisons

As a baseline, a simple Linear Regression model without regularization or feature expansion was implemented. This model serves as a benchmark to evaluate whether more complex models, such as Random Forests and Gradient Boosting Regressors, offer significant improvements in predictive performance. The baseline helps quantify the added value of non-linear models and advanced feature interactions in predicting salaries based on job-related attributes.

Success Criteria

The success of the machine learning models will be measured by their ability to minimize prediction error while maintaining interpretability. Specifically, a model will be considered successful if it achieves:

- A lower RMSE and MAE compared to the baseline model.
- Stable performance across training, validation, and testing sets, indicating minimal overfitting.
- Meaningful feature importance or interpretability, allowing insights into which factors most influence AI-related salaries.

Meeting these criteria demonstrates that the model not only performs well quantitatively but also provides practical insights into the structure of the global AI job market.

TIMELINE AND MILESTONES

Project Gantt Chart

Narrative Summary

The project timeline spans from November 1 to December 8, 2025, following five main phases: data preparation, model development, evaluation, interpretation, and report finalization.

Parallel activities such as exploratory data analysis and feature engineering occur early to accelerate model readiness. Random Forest and Gradient Boosting development partially overlap to utilize experimentation time efficiently. Hyperparameter tuning and comparative evaluation follow, leading into feature interpretation and visualization tasks.

Weeks 14–15 (Dec 1–8) are dedicated to report writing, figure refinement, and final presentation preparation, with a buffer period built in from Dec 4–7 to handle unexpected issues such as model instability, missing plots, or computational delays. The schedule aligns with course deadlines: presentations occur during Week 15 (Dec 2–4) and the final report with code submission is due December 8, 2025. This flow is summarized in a Gantt Chart (FIG. 3)

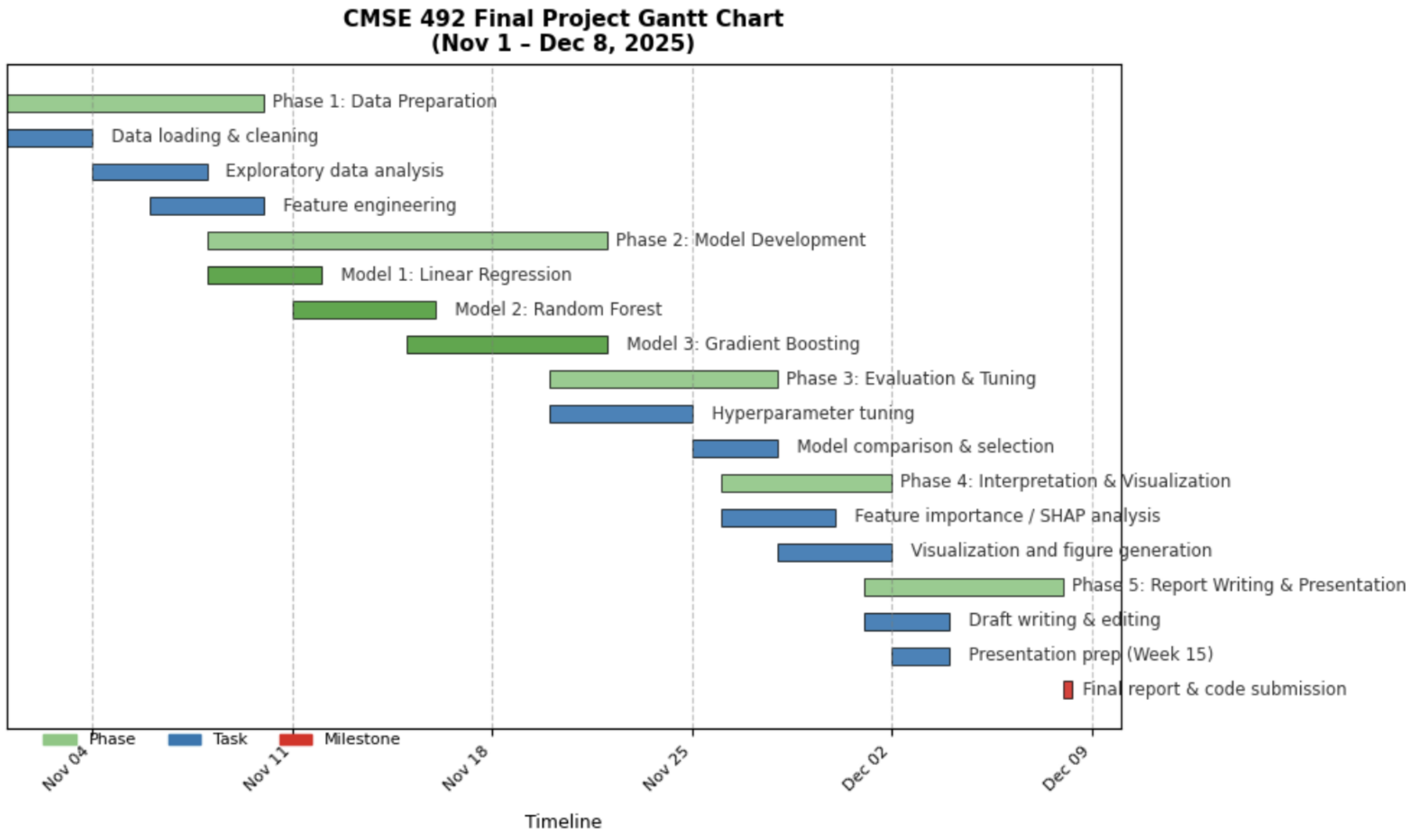

FIG. 3: Gantt chart showing project timeline, dependencies, and key milestones through December 8, 2025.

Code Availability

The complete code for this project is available at: https://github.com/fandinos/cmse492_project

* fandinos@msu.edu