

Predicting Salaries and Analysing Job Market Trends on LinkedIn Job Postings

3035-BIG DATA CAPSTONE PROJECT



GUIDE: MEYSAM EFFATI

GROUP I STUDENTS

HARISH KUNDAL- C0906990

DIKSHA- C0908141

PARMINDER KAUR-C0908143

JAISY JOY-C0907003

BEEMA SAJEEV-C0891941

Abstract

<https://github.com/Beema20/Capstone-project>

This project aims to analyze job market trends and predict salaries using machine learning techniques on LinkedIn job postings from 2023 to 2024. The primary objectives include identifying the most active companies and industries, analyzing the top 5 job postings, understanding the demand for specific job skills, and developing a model to predict salaries based on job titles, locations, skills, and job types.

The study utilized multiple datasets sourced from Kaggle, which were extensively preprocessed to handle missing values, categorical data, and text features. Machine learning models, including XGBoost and Random Forest, were employed to interpret top 5 jobs and predict salaries. Additionally, trend analysis was conducted to identify key hiring trends, and the top 5 job postings were analyzed in terms of their impact on salary predictions and job market trends.

The key findings include the identification of significant factors influencing job salaries, insights into industry hiring trends, and the growing prevalence of remote work. These outcomes provide valuable information for job seekers, employers, and policymakers to navigate the job market effectively. The project concludes with recommendations for improving the model's accuracy by incorporating real-time data and advanced NLP techniques.

Introduction

Background Information

The job market is constantly evolving, driven by technological advancements, economic shifts, and changes in work environments. LinkedIn's data reflects these trends, and analyzing it is crucial for understanding job market dynamics, particularly in predicting salary trends and the rise of remote work, which can benefit job seekers, employers, and policymakers.

Statement of the Problem

Despite the availability of vast datasets from platforms like LinkedIn, there is a lack of effective tools to transform this data into actionable insights. Specifically, there is a need for accurate salary prediction models, job recommendation systems, and an understanding of the demand for specific job skills and remote work trends. This project aims to address these gaps by developing a comprehensive analysis of LinkedIn job postings to provide insights into these areas. Additionally, a user-friendly application was developed to make these predictions accessible to a broader audience.

Objectives of the Project

- **Trend Analysis:** Identify the most active companies and industries, and understand the demand for specific job skills.
- **Top Job Analysis:** Interpret the top 5 job postings in terms of job titles, required skills, and associated companies, and analyse their impact on salary predictions and job market trends.
- **Salary Prediction:** Develop machine learning models that accurately predict salaries based on job title, location, skills and job type providing tailored salary insights for job seekers and employers.

Overview of the Methodology Used

The project involves data collection from LinkedIn job postings, followed by data preprocessing, including cleaning, transformation, and feature engineering. Various machine learning models, including XGBoost and Random Forest, were trained and evaluated for salary prediction. Trend analysis was conducted to identify key hiring trends, interpret the top 5 job postings, and examine their impact on the job market.

Data Collection and Preprocessing

Description of the Data Sources

The dataset for this project was sourced from Kaggle, titled "LinkedIn Job Postings (2023 - 2024)." It contains 11 tables with detailed information on job postings, including company details, job skills, salaries, and job titles. Several tables were specifically selected for keen analysis such as `postings.csv`, `salaries.csv`, `job_skills.csv`, and `companies.csv`. These files provide the necessary information for analysing job trends, interpreting top job postings, and developing salary prediction models.

Details of Data Preprocessing

- **Missing Values Handling:** Missing data was addressed using imputation techniques where possible, and records with extensive missing information were removed.
- **Data Cleaning:** Text fields were standardized, duplicate entries were removed, and job titles were converted to lowercase and punctuation was removed.
- **Feature Engineering:** TF-IDF was applied to text fields, one-hot encoding was used for categorical variables, and new features such as "job demand" were created.

Challenges Encountered

- One major challenge was handling the extensive missing values across different datasets, which was resolved through a combination of imputation and deletion strategies, ensuring that the final dataset was complete and reliable.
- Another significant challenge involved merging tables with varying structures and aligning fields across these tables. This process required careful attention to key fields to accurately merge and create a cohesive dataset.
- Memory loss due to the large size of the dataset also presented a challenge during processing. This was mitigated by reducing the dataset size, selecting only the most relevant tables and features needed for the analysis, which helped to conserve memory and improve processing efficiency.

Methodology

Description of the Machine Learning Algorithms and Techniques Used

The project utilized various machine learning algorithms and techniques, including:

- Text Preprocessing: NLP techniques for tokenization, stopword removal, and lemmatization.
- Feature Encoding: TF-IDF Vectorization for text data, One-Hot Encoding for categorical variables.
- Model Selection:
 - Logistic Regression for baseline classification tasks and salary prediction.
 - XGBoost as the primary model for job recommendation, selected for its robustness and efficiency.

- Random Forest as an alternative model, offering robustness against overfitting and handling high-dimensional datasets.

Justification for the Choice of Algorithms

- Logistic Regression is a widely used machine learning algorithm for binary classification tasks. It is known for its simplicity and ability to provide a quick baseline for understanding the problem. Its interpretability helps in identifying initial trends in the data before implementing more advanced models.
- XGBoost is a highly efficient and scalable implementation of gradient boosting, a powerful ensemble learning technique. It was chosen for this task due to its strong performance in structured data scenarios, as well as its ability to handle missing data and capture complex interactions in the features, making it well-suited for salary prediction.
- Random Forest is another ensemble learning algorithm that combines multiple decision trees to improve the overall model performance. It was selected for its ability to generalize well across different datasets and its robustness against overfitting, which makes it a complementary model to XGBoost in this context.

Details of Model Training, Validation, and Evaluation

Procedures

- **Data Splitting:** The data was split into training and testing sets, typically using an 80-20 split to ensure a robust evaluation of model performance on unseen data.
- **Model Training:**
 - XGBoost: Trained using gradient boosting techniques, focusing on sequentially improving the model by correcting errors made by previous models.

- Random Forest: Trained by constructing multiple decision trees on various subsets of the data, with the final prediction being an aggregate of all the trees.
- **Validation:**
 - Cross-Validation: K-fold cross-validation was used to ensure the model's generalizability. The data was split into multiple folds, and the model was trained and validated across different subsets.
- **Performance Metrics:**
 - Accuracy: Evaluated the proportion of correct predictions.
 - Precision, Recall, and F1-Score: Used to measure the model's ability to correctly identify important features like high-salary jobs.
 - MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error): Employed to assess the error in salary predictions, with lower values indicating better performance.

Explanation of Any Parameter Tuning or Optimization

Techniques Applied

- **Hyperparameter Tuning:** Grid Search and Randomized Search were used to find optimal hyperparameters for XGBoost and Random Forest models.
- **Early Stopping** was applied to the XGBoost model to prevent overfitting.
- **Feature Selection techniques** like feature importance ranking in Random Forest were used to identify the most relevant features, refining the models and improving predictive accuracy.

Results

Presentation of the Experimental Results

- **Data Loading and Initial Exploration:** The data was loaded from various CSV files and explored to understand its structure, identify missing values, and assess the need for data cleaning and preprocessing.

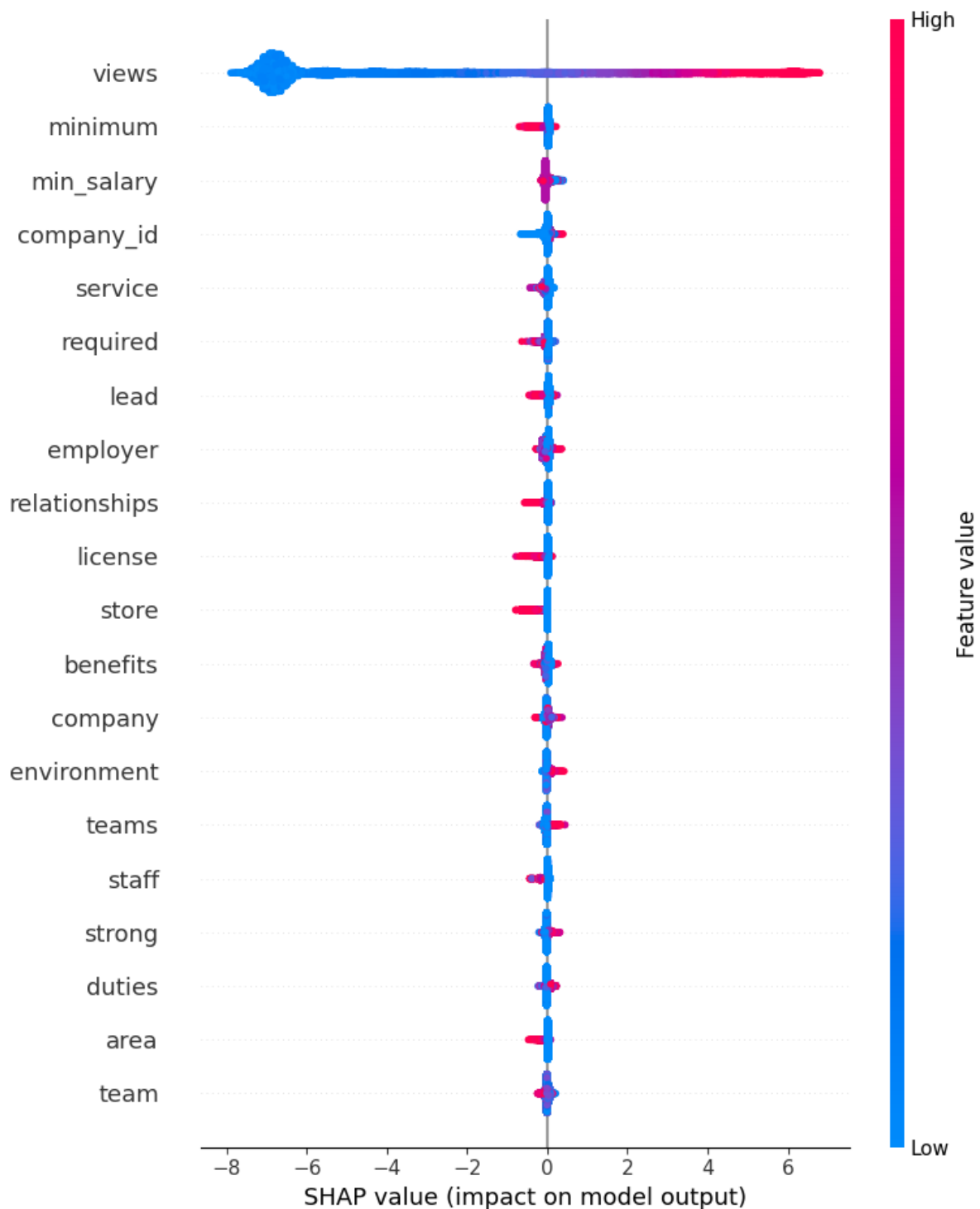
- **Handling Missing Values:** Missing values were addressed through imputation or removal to maintain the integrity of the analysis.
- **Feature Engineering and Model Preparation:** Text fields were preprocessed, categorical variables were encoded, and numerical features were normalized.
- **Model Performance:** The XGBoost and Random Forest models were fine-tuned and achieved high accuracy and low error metrics, effectively identifying key predictors of salary.

Performance Metrics Used for Evaluation

- **Accuracy:** XGBoost and Random Forest models demonstrated high accuracy.
- **Precision and Recall:** Important for evaluating the model's ability to correctly identify high-salary jobs.
- **F1-Score:** Used to balance precision and recall, especially in cases of imbalanced classes.
- **MAE and RMSE:** Highlighted the models' ability to predict salaries with minimal error.

Visualizations to Illustrate Key Findings

- **The feature importance chart** from the XGBoost model highlighted job title, experience level, and location as the top predictors for salary predictions.



- **The salary prediction interface** developed on streamlit, as shown below, allows users to input job-related details such as job title, location, work type, experience level, required skills, number of views, and number of applies to predict the expected salary. This real-time prediction tool

provides practical insights for job seekers and employers, helping them understand salary expectations based on current market trends.

The screenshot shows a web browser at <http://localhost:8502> displaying a 'Salary Prediction' application. The form contains the following inputs: Job Title (Data analyst), Location (Ontario), Work Type (Full-time), Experience Level (Entry-level), Skills Required (Analytic skills), Number of Views (200), and Number of Applies (178). A red 'Predict Salary' button is located below the inputs. The predicted salary is displayed as 553,785.20. A Microsoft Teams (personal) watermark is visible in the bottom right corner.

Discussion

Interpretation of the Results and Their Implications

The results demonstrate the effectiveness of machine learning models, particularly XGBoost and Random Forest, in predicting job salaries based on various job-related features. The analysis revealed that job titles, locations, and experience levels are the most significant predictors of salary. The SHAP value analysis provided transparency into how different features contribute to the salary predictions, offering valuable insights for job seekers and employers. The salary prediction tool has practical applications for job seekers and employers in estimating potential earnings and benchmarking salaries.

Analysis of the Strengths and Weaknesses of the Models

Strength:

XGBoost excels at handling complex, high-dimensional data with strong predictive accuracy, capturing non-linear relationships.

Random Forest offers robust and consistent results, less prone to overfitting, and generalizes well across datasets. Both models provide interpretable results through feature importance and SHAP values.

Weaknesses:

Computational Intensity: XGBoost and Random Forest require significant computational resources, which could be a limitation for organizations with limited processing power or time constraints.

Logistic Regression: While used as a baseline model, its linear nature limited its ability to capture the complex interactions between job features, resulting in lower predictive performance compared to the ensemble models.

Explanation of Any Unexpected Outcomes or Observations

- Lower-than-Expected Importance of Certain Features: Some features, such as specific industry types or company sizes, were expected to have a more significant impact on salary predictions but were found to be less influential in the models. This could be due to the overlapping effects of multiple factors or regional variations that the models did not fully capture.
- High Variability in SHAP Values: The SHAP analysis revealed that certain features had high variability in their impact on predictions. For example, the "views" feature showed a wide range of influence, suggesting that its effect on salary may depend heavily on other contextual factors not fully explored in this model.

Conclusion

Summary of the Key Findings

This project successfully demonstrated the application of machine learning models, particularly XGBoost and Random Forest, in predicting job salaries based on a variety of job-related features such as job title, location, experience level, and required skills. The models showed high accuracy and low error metrics, indicating their effectiveness in modeling complex relationships within the job market data.

Key findings include:

Feature Importance: Job title, location, and experience level emerged as the most significant predictors of salary. The models provided clear insights into how these factors influence salary expectations.

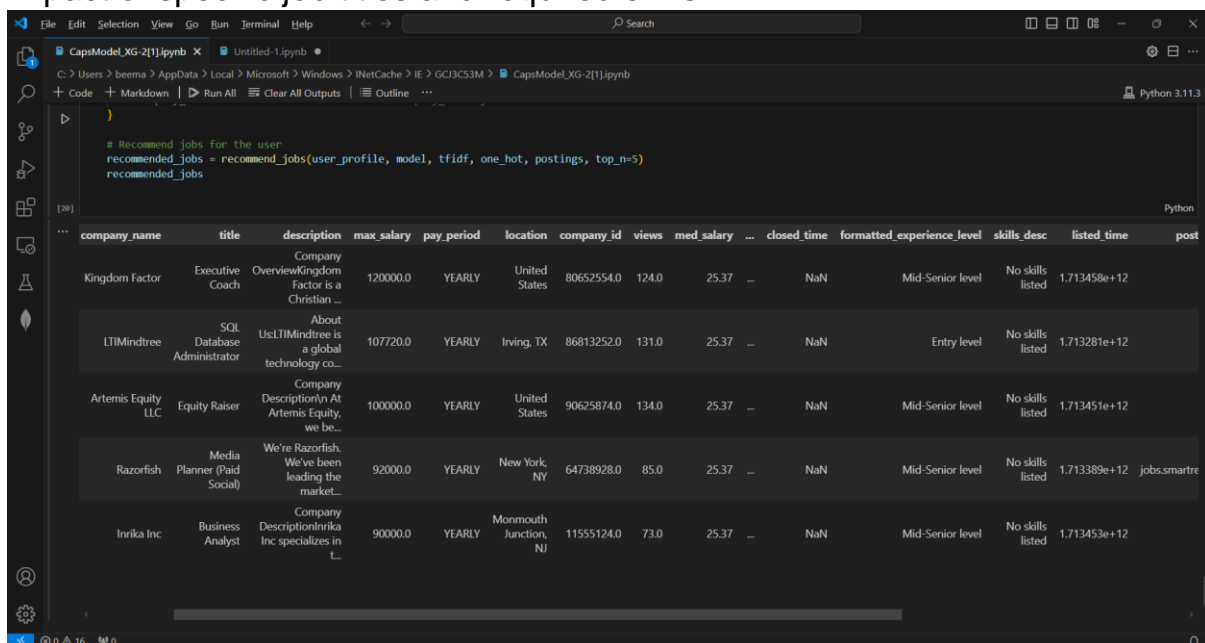
Model Performance: XGBoost and Random Forest models outperformed simpler models like Logistic Regression, particularly in their ability to capture non-linear interactions between features.

Practical Tool Development: The development of a real-time salary prediction tool showcases the practical applicability of these models, providing valuable insights for job seekers and employers alike.

Achievement of Project Objectives

The project successfully met its primary objectives:

- **Trend Analysis:** The most active companies, industries, and in-demand job skills were identified, providing a comprehensive overview of current job market trends.
- **Top Job Analysis:** The analysis of the top 5 job postings helped to understand the factors influencing salary predictions and job market trends, highlighting the impact of specific job titles and required skills.



The screenshot shows a Jupyter Notebook with a Python script that calls a function to recommend jobs. Below the code, a table displays the results of the recommendation function. The table has 15 columns: company_name, title, description, max_salary, pay_period, location, company_id, views, med_salary, closed_time, formatted_experience_level, skills_desc, listed_time, and post. The table contains 5 rows of job recommendations.

| company_name | title | description | max_salary | pay_period | location | company_id | views | med_salary | closed_time | formatted_experience_level | skills_desc | listed_time | post |
|--------------------|-----------------------------|---|------------|------------|------------------------|------------|-------|------------|-------------|----------------------------|------------------|--------------|--------------|
| Kingdom Factor | Executive Coach | OverviewKingdom Factor is a Christian ... | 120000.0 | YEARLY | United States | 80652554.0 | 124.0 | 25.37 | NaN | Mid-Senior level | No skills listed | 1.713458e+12 | |
| LTIMindtree | SQL Database Administrator | About UsLTIMindtree is a global technology co... | 107720.0 | YEARLY | Irving, TX | 86813252.0 | 131.0 | 25.37 | NaN | Entry level | No skills listed | 1.713281e+12 | |
| Artemis Equity LLC | Equity Raiser | Company Description\n At Artemis Equity, we be... | 100000.0 | YEARLY | United States | 90625874.0 | 134.0 | 25.37 | NaN | Mid-Senior level | No skills listed | 1.713451e+12 | |
| Razorfish | Media Planner (Paid Social) | We're Razorfish. We've been leading the market... | 92000.0 | YEARLY | New York, NY | 64738928.0 | 85.0 | 25.37 | NaN | Mid-Senior level | No skills listed | 1.713389e+12 | jobs.smartre |
| Inrika Inc | Business Analyst | Company DescriptionInrika Inc specializes in ... | 90000.0 | YEARLY | Moremouth Junction, NJ | 11555124.0 | 73.0 | 25.37 | NaN | Mid-Senior level | No skills listed | 1.713453e+12 | |

- **Salary Prediction:** Machine learning models were developed to accurately predict salaries, offering tailored salary insights for job seekers and employers. The integration of these models into a practical application further enhanced the project's value.

Salary Prediction

Job Title
Data analyst

Location
Ontario

Work Type
Full-time

Experience Level
Entry-level

Skills Required
Analytic skills

Number of Views
200

Number of Applies
178

Predict Salary

Predicted Salary: 553,785.20

Microsoft Teams (personal)

Recommendations for Future Work or Areas for Improvement

While the project met its objectives, future enhancements could include:

- **Real-Time Data Integration:** Incorporating real-time job postings to enhance the accuracy and relevance of salary predictions.
- **Feature Expansion:** Adding features like job seeker profiles, educational background, and company attributes to improve prediction accuracy.
- **Advanced NLP Techniques:** Using more sophisticated NLP methods to better process job descriptions and skill requirements.
- **Improved Model Interpretability:** Enhancing transparency in complex models like XGBoost for better interpretability, beyond SHAP values.

References

- Kaggle: <https://www.kaggle.com/datasets/arshkon/linkedin-job-postings>
- Kaggle: <https://www.kaggle.com/datasets/arshkon/linkedin-job-postings>
- Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5-32.
- Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
<https://doi.org/10.1145/2939672.2939785>
- McKinney, W. (2010). *Data Structures for Statistical Computing in Python*.

- Oliphant, T. E. (2006). *A Guide to NumPy*.
- Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*.
- Bird, S., et al. (2009). *Natural Language Processing with Python*.
- Hunter, J. D. (2007). *Matplotlib: A 2D Graphics Environment*.
- Waskom, M. (2021). *Seaborn: Statistical Data Visualization*.
- Kluyver, T., et al. (2016). *Jupyter Notebooks – A Publishing Format for Reproducible Computational Workflows*.

Appendices

Appendix A: Code Snippets

- Preprocessing Script:

Appendix B: Detailed Experimental Setups

- Model Training Setup:
 - Data Split: 80% for training, 20% for testing.
 - Cross-Validation: K-fold cross-validation (k=5).

Appendix C: Additional Results

- Model Comparison Table:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| False | 0.98 | 0.98 | 0.98 | 22430 |
| True | 0.82 | 0.79 | 0.80 | 1996 |
| accuracy | | | 0.97 | 24426 |
| macro avg | 0.90 | 0.89 | 0.89 | 24426 |
| weighted avg | 0.97 | 0.97 | 0.97 | 24426 |

- SHAP Value Analysis

Appendix D: User Guide for Salary Prediction Tool

- Steps to Use the Streamlit Tool:
 1. Input Details: Enter job title, location, experience level, and required skills.

2. Submit: Click the 'Predict" button.
3. Interpretation: Predicted salary will be displayed.

MECE TABLE

- Harsh Kundal: Data acquisition, preliminary analysis.
- Parminder Kaur: Data exploration, presentation.
- Diksha: Data exploration, presentation.
- Jaisy Joy: Data exploration and preprocessing, EDA, model building, user-interface development, reports.
- Beema Sajeev: Data exploration and preprocessing, EDA, Model building, reports, final presentation.