Analysis of U.S. Software Engineering Jobs Using KDD Methodology

Kelly Nguyen*

[1]Charles Davidson School of Engineering,
San Jose State University, San Jose, California

Abstract

This research paper presents a comprehensive analysis of software engineering jobs in the United States. The paper follows the Knowledge Discovery in Databases (KDD) methodology, covering key phases such as data selection, data cleaning, data preprocessing, data transformation, data mining, pattern evaluation, and knowledge representation. Various machine learning models were built to predict job ratings, and their performances were evaluated. The paper concludes with actionable insights and recommendations for future research.

Introduction

The landscape of software engineering jobs in the United States is both dynamic and ever-evolving. Rapid technological advancements, the emergence of new programming languages, and changes in software development paradigms contribute to the complex and shifting nature of this landscape. Understanding it is crucial for various stakeholders, including job seekers looking for opportunities, employers wanting to attract the best talent, and policymakers aiming to support the technology sector effectively.

To provide a structured and comprehensive overview of this landscape, this paper focuses on analyzing software engineering job postings. We dive deep into key attributes such as:

- Job title: The designation and role advertised.
- Company: The organization offering the job, including its size and reputation.
- Type of employment: Whether the job is full-time, part-time, or contractual.
- Salary: The compensation package, including any bonuses or benefits.
- Ratings: Employee reviews and ratings of the company.

The analysis is conducted following the Knowledge Discovery in Databases (KDD) methodology. KDD serves as a widely accepted, systematic framework for data science and analytics projects. The methodology is broken down into the following key stages:

1. **Data Selection**: The process of defining and collecting the data to be analyzed.
2. **Data Cleaning**: Removing noise and irrelevant data to improve the quality of the dataset.
3. **Data Preprocessing**: Preparing the data for mining, including feature engineering and data transformation.
4. **Data Transformation**: Converting the data into a suitable format or structure for analysis.
5. **Data Mining**: Applying data mining algorithms to discover patterns in the prepared dataset.
6. **Pattern Evaluation and Interpretation**: Assessing and understanding the discovered patterns.
7. **Knowledge Representation**: Presenting the findings in a comprehensible and usable format.

This paper goes through each of these stages in detail to offer a thorough analysis of software engineering job postings in the United States.

Data Source The dataset used in this analysis comprises 58,433 job postings, each with 29 attributes. These attributes include job title, company name, job type, location, salary, and ratings, among others. The dataset provides a comprehensive overview of the software engineering job market in the U.S.

Methodology

Data Selection

Data selection forms the cornerstone of any analytical study. The dataset used in this research is both expansive and detailed, featuring 58,433 job postings across various states in the U.S. Each posting is characterized by 29 attributes, including but not limited to, job title, company name, type of employment, location, salary range, and employer ratings. This dataset offers a panoramic view of the job market, laying a strong foundation for an in-depth analysis.

*Corresponding author: kelly.nguyen01@sjsu.edu
Received: September 21, 2023, **Published:** September 21, 2023

Data Cleaning

Data quality is paramount for reliable analytical outcomes. The data cleaning phase aimed to streamline the dataset by removing irrelevant columns and handling missing or inconsistent data. For instance, columns that did not contribute to the research objectives were eliminated.

Data Preprocessing

Data preprocessing is the unsung hero of data analytics, often making the difference between mediocre and outstanding results. This phase involved several steps designed to prepare the data for effective mining. Categorical features were converted into numerical form through label encoding, making them machine-readable. Feature engineering brought additional layers of information into the analysis. For instance, an 'allows_remote' feature was introduced to indicate remote job opportunities. Data normalization was also carried out to level the playing field among different scales of data.

Data Mining

The data mining phase is the heart of the analytical process. Three different types of regression models were employed to predict job ratings. These included Linear Regression, Ridge Regression, and Lasso Regression. The models were meticulously trained on a randomly selected subset comprising 80% of the total dataset. The remaining 20% served as the test set for performance evaluation.

Model Evaluation

Model evaluation is crucial for gauging the effectiveness of the data mining process. Several metrics were used for this purpose, including Mean Squared Error (MSE), Mean Absolute Error (MAE), and the R2 Score. While the models showed promise, they also highlighted room for improvement, particularly in their R2 scores and error metrics. These findings serve as a valuable guide for future research and model optimization.

Pattern Evaluation and Interpretation

This phase provided a post-analysis evaluation of the discovered patterns and models. The results indicate that while the models are a step in the right direction, there's substantial scope for enhancement. This could include using more advanced machine learning algorithms or incorporating additional features that capture the nuances of the job market more effectively.

Summary and Findings

The paper undertook a rigorous analysis of the U.S. software engineering job market using a structured KDD methodology. While the predictive models showed promise in estimating job ratings, their performance metrics indicated room for improvement. The extensive data cleaning and preprocessing phases ensured data quality, and the feature engineering efforts added valuable dimensions to the analysis. Overall, the study sheds light on the multifaceted aspects of the software engineering job market and sets the groundwork for future research.

Future Work

The study opens several avenues for future research. Firstly, the predictive models can be enhanced by incorporating advanced machine learning techniques like ensemble methods and deep learning algorithms. Secondly, a more granular analysis can be performed by focusing on specific sectors within software engineering, such as machine learning engineering or data engineering. Additionally, a temporal analysis to understand job market trends over time could be of significant value. Lastly, a comparative study involving software engineering job markets from different countries may provide global insights into the field.

In conclusion, this research paper offers a comprehensive and insightful analysis of the software engineering job market in the United States. Utilizing the robust KDD methodology, it brings several layers of the job market into focus. While the predictive models used in the study showed potential, they also revealed areas for improvement. The paper rounds off by offering a set of actionable insights and recommendations, pointing the way for future academic and industry-specific research.

References

[1] Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J., *Knowledge Discovery in Databases: An Overview*, 1992, AI Magazine, 13(3), 57-70.

[2] James, G., Witten, D., Hastie, T., & Tibshirani, R., *An Introduction to Statistical Learning*, 2013, Springer.

[3] Kaggle, *Software Engineering Jobs Dataset*, n.d., Retrieved from `https://www.kaggle.com/`.