COMP 450 GROUP PROJECT

Automated Developer Salary Prediction Using 2025 Stack Overflow Survey Data and OCR-Enabled CV Parsing

Zehra Mert 042201058
Onat Sarıbıyık 042101097
Zeynep Yılmaz 042101088

1. Motivation

Determining fair compensation for software developers is crucial and challenging in today's job market. Depending on the region, technologies, experience levels and educational backgrounds salaries may vary significantly and employers/HR personnel may be inclined towards limited or biased data while developers struggle to pitch for the appropriate compensation. Especially as of today, 43% of companies now leverage AI in the recruitment process[1]. Existing AI solutions often rely on manual self reporting or historical data which may result in failure to detect the post-2024 market shift. In addition, these systems may lack the ability to provide estimations directly from an individual's curriculum vitae (CV).

Stack Overflow holds a developer survey annually which provides insights on the state of software development [2]. In 2025 they included a new focus on AI agent tools, LLMs and community platforms, which offered a large up-to-date dataset [3]. Leveraging this dataset in our ML engine which will work on extracted data via Optical Character Recognition can result in a robust salary prediction model. This multimodal predictor can address a key HR pain point identified in reviews of AI resume parsers [4][5], which saves them from the torment of manual data entry and provides personalized salary estimates from an uploaded resume.

The motivation of bridging the gap between survey based insights and real developer candidates pushed us to develop a system that can also empower developers to make informed career decisions and help hiring professionals in implementing more transparent and fair compensation practices therefore addressing a key problem in the constantly evolving AI augmented workplace [6][7].

2. Literature Survey

One of the main themes in recent literature is the growing interest in using sophisticated computational techniques for labor market analysis, especially salary estimation. In today's highly competitive job market, accurate salary prediction is a critical advantage that provides insights to individuals when considering career transitions [7]. Several studies have explored salary prediction using structured datasets or text-based data from resumes and job postings.

Chen and Li [8] proposed a machine learning model to predict salaries based on applicant resumes,with an emphasis on skill-based features. Similarly, Akay *et al.* [9] used regression-based methods for the same purpose.Ji *et al.* [10] proposed a recent work that improved the salary prediction accuracy by introducing a disentangled composition effect neural model, while Malaiarasan *et al.* [11] and Xu [12] presented generalized machine learning

frameworks for salary estimation.Saraswathi and Akhila [7] further emphasized the use of ML to forecast future salaries based on professional features.

Sarhan et al. [13] presented a CV content recognition framework that makes use of Tesseract-OCR and YOLOv8 for resume processing and comprehension. A methodology for job domain prediction and resume parsing was presented by Mittal et al. [14]. In addition to these, Gunjal et al. [15] examined new AI-based resume parsing techniques. Jiang et al. [16] suggested a multi-modal pre-training model for effective resume comprehension, while a Turkish study by Saatçı et al. [17] used natural language processing (NLP) techniques for automated resume screening.

In summary, most prior research focuses either on **structured survey data** or **textual job descriptions**. Few studies combine **OCR-based CV parsing** with structured **developer survey features**, making this project both innovative and relevant for current AI-driven HR analytics.

3. Proposed Methods

Step 1: Gathering and Cleaning Data

- The 2025 Stack Overflow Developer Survey dataset should be downloaded and preprocessed.
- Features like Country, Experience, Primary Language, Education Level, and Employment Type are extracted.
- Collect anonymized resume samples (simulated or from Kaggle).
- Use OCR (EasyOCR) to extract text from PDF resumes.
- Employ natural language processing (NLP) techniques to extract information about education, years of experience, and skills.

Step 2 – Feature Engineering

- Combine features extracted by OCR and attributes from the survey data.
- Transform categorical into sequential.
- Scale continuous features from 0 to 1 (for example by using the MinMaxScaler).
- Create derived columns as example Skill Match Index and Tech Stack Popularity Score.

Step 3 – Model Development

- Train multiple regression and ensemble models:
    - Linear Regression, Ridge, Lasso
    - Random Forest Regressor
    - Gradient Boosting, XGBoost, and LightGBM
    - Neural Network (MLP Regressor for non-linear relationships)
- Evaluate using **RMSE**, **MAE**, and **R^2** metrics.
- Perform cross-validation and hyperparameter tuning with **GridSearchCV**.

Step 4 – Interpretability & Visualization

- For model interpretability, it's best to use SHAP or LIME.
- See the importance of features (skills, location, experience etc).
- Build an app (example – Streamlit / Flask) to predict a salary from uploaded CV.

4. Expected Results

As a result of this project we expect to deliver a high accuracy predictive model for developer salaries and we anticipate it to be reliable. The primary success metric will be the R^2 score which we aim for a score greater than 0.80. This score indicates that the model explains over 80% of the salaries when tested on the unseen data. Secondary metrics will be the Root Mean Square Error which is expected to be less than $8,000 to provide a clear average range for the prediction error.

The provided flowchart (Figure 1) explains the general structure of our system. The model also will reveal some insights on key drivers of salary by yielding a clear data driven hierarchy of factors that influence a greater salary. We aim to utilize some interpretability tools like SHAP which will quantify and visualize the impact of technical skills, experience/ education as well as geographic and employment factors.

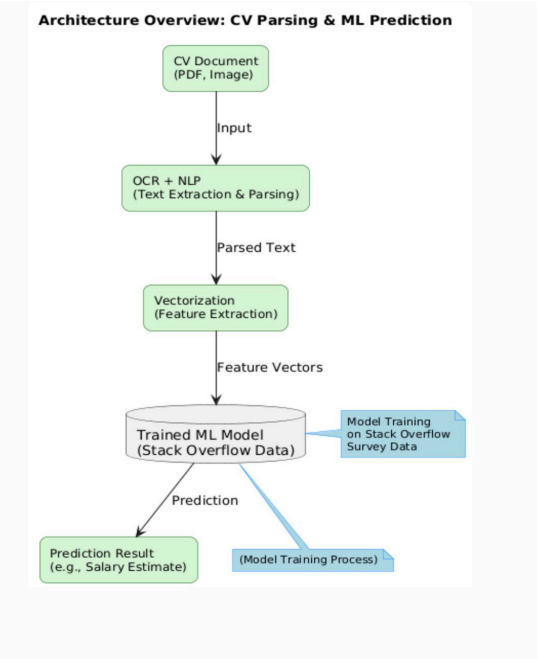

Figure 1: Architecture Overview

5. References

[1] SHRM, "The Role of AI in HR Continues to Expand," 2025.

[2] Stack Overflow Developer Survey Results, 2023. [Online]. Available: https://survey.stackoverflow.co

[3] McKinsey & Company, "Superagency in the Workplace: Empowering People to Unlock AI's Full Potential," 2025.

[4] ApyHub, "Top 5 Free AI Resume Parsers for Recruiters," 2025.

[5] Airparser, "Top 5 CV and Resume Parsers in 2025," 2025.

[6] AIHR, "11 HR Trends for 2026: Shaping What's Next," 2025.

[7] M. Saraswathi and J. Akhila, "Using Machine Learning to Determine Your Future Salary," International Journal of Soft Computing and Engineering (IJSCE), vol. 13, no. 2, 2023.

[8] Y. Chen and X. Li, "Salary Prediction Based on the Resumes of the Candidates," SHS Web of Conferences, vol. 170, p. 03013, 2023.

[9] M. F. Akay, B. Düzgün, and C. Ulus, "Development of Salary Prediction Models for the Information Technology Industry," Journal of Data Science & Modern Techniques, vol. 2, p. 102, 2025.

[10] Y. Ji, Y. Sun, and H. Zhu, “Enhancing Job Salary Prediction with Disentangled Composition Effect Modeling: A Neural Prototyping Approach,” arXiv preprint, 2025.

[11] S. Malaiarasan, M. A. Riyaz, and M. Appadurai, “Salary Prediction Using Machine Learning,” International Journal of Scientific Research & Engineering Development, vol. 8, no. 2, 2025.

[12] M. Xu, “Salary Prediction Using Machine Learning,” Scholarly Review Journal, Issue 13, Summer 2025.

[13] A. M. Sarhan, H. A. Ali, M. Wagdi, B. Ali, A. Adel, and R. Osama, “CV Content Recognition and Organization Framework based on YOLOv8 and Tesseract-OCR Deep Learning Models,” ResearchGate, 2024.

[14] V. Mittal, P. Mehta, D. Relan, and G. Gabrani, “Methodology for Resume Parsing and Job Domain Prediction,” Journal of Statistics & Management Systems, 2020.

[15] M. B. Gunjal, T. P. Thorat, K. S. Muttha, V. C. Shete, and P. D. Sagar, “A Review Paper on Resume Parser Using AI,” International Journal of Innovative Research in Technology (IJIRT), vol. 11, no. 8, 2025.

[16] F. Jiang et al., “Towards Efficient Resume Understanding: A Multi-Granularity Multi-Modal Pre-Training Approach,” arXiv preprint, 2024.

[17] M. Saatçı, R. Kaya, and R. Ünlü, “Resume Screening with Natural Language Processing (NLP),” Alphanumeric Journal, vol. 12, no. 2, 2024.