Team 15 Project

Decoding Developer Realities: A Study Using Stack Overflow Survey Data

GitHub Link: https://github.com/mahesh973/DevCompPredictor-Capstone

Team Members: Mahesh Babu Kommalapati, Shivani Ashish Mundle, Sharanya Senthil

Introduction

As we are technology students looking to enter a dynamic developer industry, gaining insights into factors shaping careers is critical, especially compensation. Lack of transparency and uneven access to information on salary standards have made navigating career options challenging, especially for aspiring developers and students like us. Using this as our **motivation**, we have developed a tool for software professionals to estimate salary ranges based on their demographics, experience, and education.

The importance of our solution goes beyond just predicting salaries, but by influencing the way developers make career decisions. We aim to empower both students and seasoned professionals and provide a roadmap for informed career choices, supporting industry transparency and add to the ongoing dialogue on fair compensation in the global technology community. While existing productivity tools overlook the data-driven guidance, our application fills this gap by using the Stack Overflow developer survey. The survey provides a collective response on skills, roles, backgrounds, and salaries from professional developers which has created a rich dataset we are using for our analysis.

Our solution has a methodical flow *[Fig 1]* beginning with cleaning the survey dataset. After data cleaning, we performed feature engineering to extract the best performing features. Salary buckets are then strategically defined to enable classification models for prediction. We evaluated multiple classification models to find the best performing model for this task. Our application also provides relevant visualization for technology trends. An additional chatbot feature was added to address specific user queries.

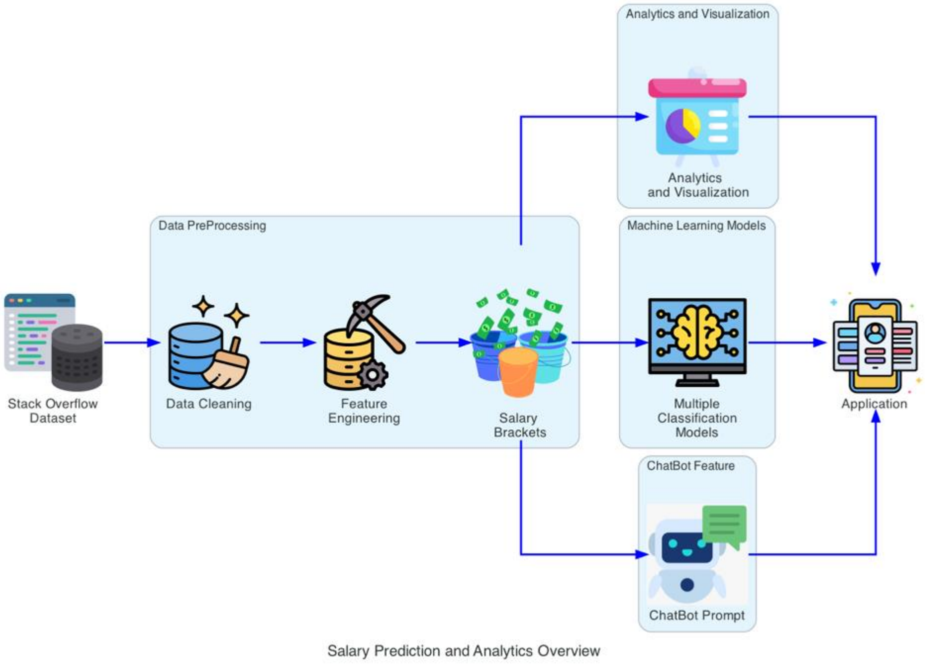


Figure 1: Application Flowchart

Methodology

Low Risk:

The low risk we highlighted in the proposal was that the Stack Overflow survey data we are using can have quality and accuracy problems. Surveys often can have issues like missing responses, people giving incorrect data, too many answer choices etc. These issues could negatively affect our machine learning model's ability to predict salary range accurately.

To mitigate this risk, we thoroughly cleaned the data by removing outlier, NA values etc. We also standardized categorical values for effective encoding. After trying out different feature combinations, we narrowed down to 7 features that yielded high accuracy in predicting salaries. Additionally, we found 18k datapoints were from the USA while the second largest Germany had only 7k *[Fig 2].* This huge imbalance across countries also posed a data accuracy problem. We mitigated this by narrowing our focus to only the USA dataset for training models. This improved accuracy with a more balanced dataset.

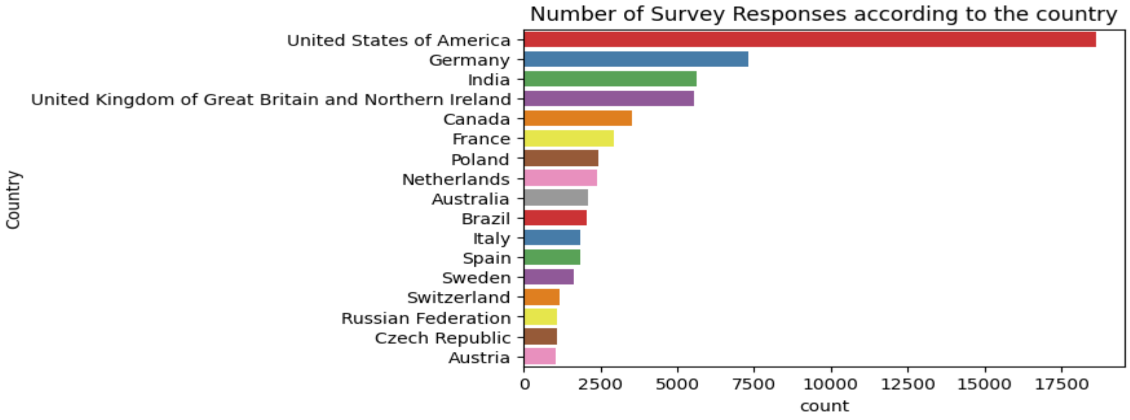
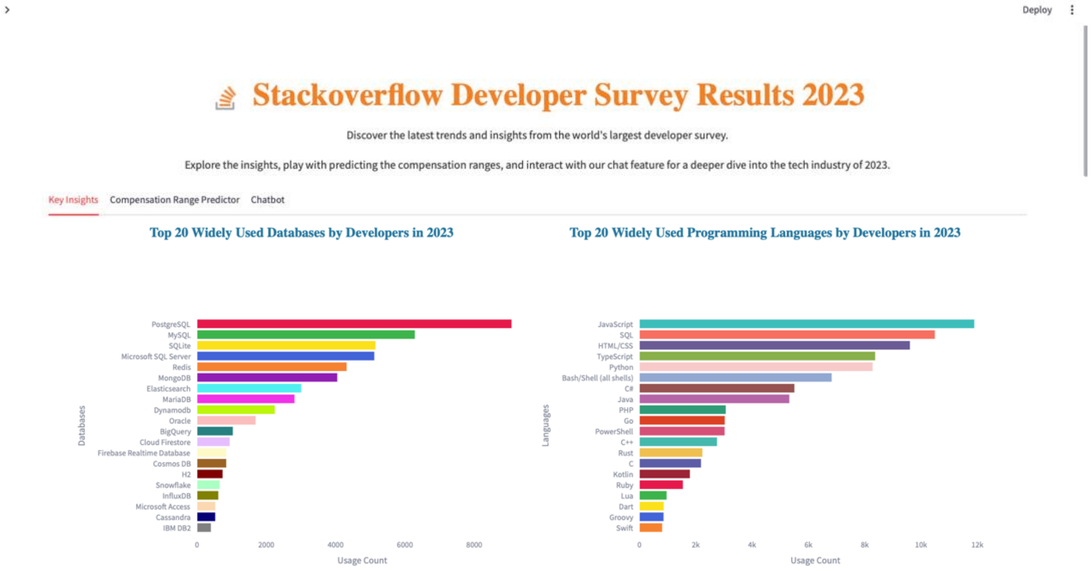


Figure 2: Data point distribution across countries

We also created some interesting data visualizations during the process that showed salary relationships with key features. *[Fig 3]* This gave additional insights for future analysis. Overall, the low-risk problem was mitigated by proper data processing and data and feature engineering.



Stackoverflow Developer Survey Results 2023

Discover the latest trends and insights from the world's largest developer survey.

Explore the insights, play with predicting the compensation ranges, and interact with our chat feature for a deeper dive into the tech industry of 2023.

Figure 3: Vizualization Dashboard from our tool

Medium Risk:

Our medium risk involved choosing a classification approach instead of a traditional regression model for the salary prediction. To mitigate this, we evaluated both approaches - implementing multivariate regression and assessing performance against multiple classification models. The regression accuracy was very low, validating our choice of classification. We systematically processed the target salary data into balanced buckets before applying classifiers. Starting with Decision Tree, Random Forest, Gradient Boost classifiers and Hybrid Classifier. Gradient Boosting & Hybrid Classifier emerged most accurate [Table 1]. We are using the Hybrid Classifier for our tool *[Fig 4].*

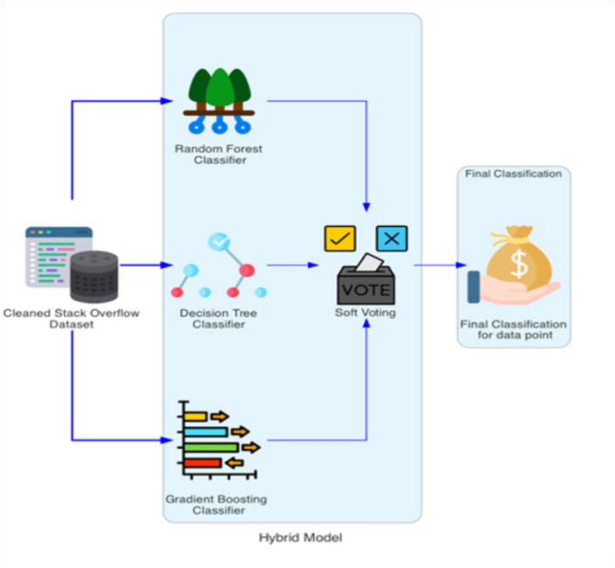
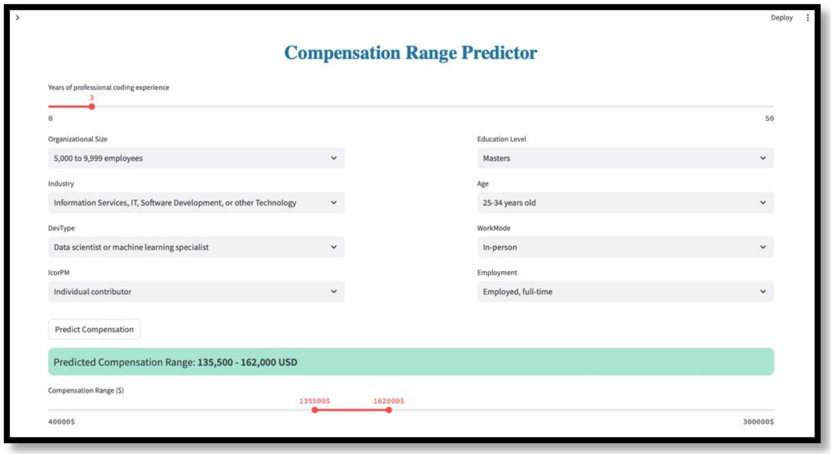


Figure 4: Hybrid Model Structure

This is likely because boosting ensemble methods can effectively handle nonlinear data like skewed salaries by iteratively learning from misclassifications. The careful evaluation of regression vs classification and controlled bucketing of salary ranges helped mitigate the medium risk, proving classification as the prudent modeling choice. [Fig 5] shows the our dashboard, where the user can enter the features and the application will show the salary range.



Compensation Range Predictor

Years of professional coding experience: 3 (0 – 50)

Organizational Size: 5,000 to 9,999 employees

Education Level: Masters

Industry: Information Services, IT, Software Development, or other Technology

Age: 25-34 years old

DevType: Data scientist or machine learning specialist

WorkMode: In-person

IcorPM: Individual contributor

Employment: Employed, full-time

Predict Compensation

Predicted Compensation Range: **135,500 - 162,000 USD**

Compensation Range ($): 135500$ – 162000$ (40000$ – 300000$)

Figure 5: Salary Range Predictor Tool

Model	Training Accuracy	Testing Accuracy	AUC
Decision Tree Classifier	0.90	0.78	0.91
Random Forest Classifier	0.91	0.75	0.94
Gradient Boosting Classifier	0.95	0.84	0.98
Hybrid Classifier	0.95	0.84	0.96

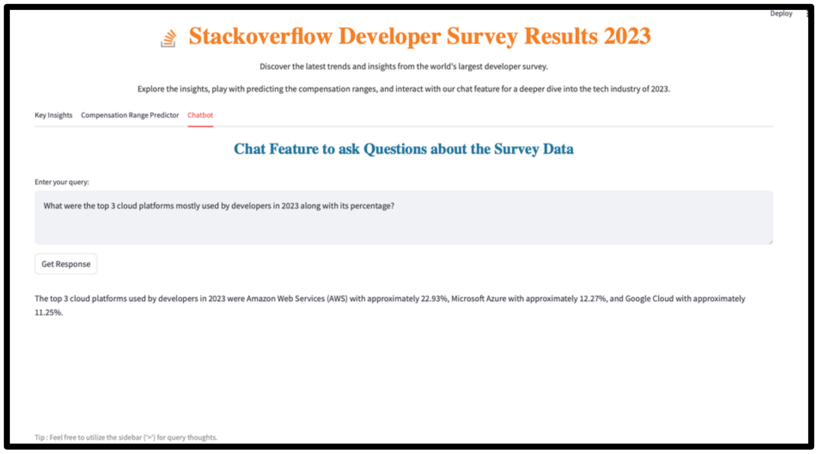
Table 1: Classification Model Comparison

High Risk:

Our project faced significant challenges due to the volatile nature of salary data, which fluctuated due to inflation and global economic trends, especially across different countries. To tackle this, we proposed a robust solution: a pipeline that re-trained the model each time new survey data was released. This strategy aimed to maintain the model's performance and relevance over time. While we created a script to combine the data from previous years, we were not able to mitigate this risk completely due to the time constraints. We could not create a pipeline as we prioritized the creation of our application over the pipeline.

We also recognized a slight imbalance in the data, which could affect the model's performance. To address this, we combined data from previous years' surveys to ensure a balanced representation, thereby enhancing prediction accuracy.

Another major challenge was the development of a chatbot [Fig 6] that could handle specific salary and feature-related queries. As this was an ambitious extra feature, we first evaluated some basic open source chatbot options. This helped us understand how feasible it would be to add this complex functionality within the project timeline and resources.



Stackoverflow Developer Survey Results 2023

Discover the latest trends and insights from the world's largest developer survey.

Explore the insights, play with predicting the compensation ranges, and interact with our chat feature for a deeper dive into the tech industry of 2023.

Chat Feature to ask Questions about the Survey Data

Enter your query:

What were the top 3 cloud platforms mostly used by developers in 2023 along with its percentage?

The top 3 cloud platforms used by developers in 2023 were Amazon Web Services (AWS) with approximately 22.93%, Microsoft Azure with approximately 12.27%, and Google Cloud with approximately 11.25%.

Tip : Feel free to utilize the sidebar ('>') for query thoughts.

Figure 6: Chatbot Feature from Our Tool

Conclusion

Through this project, we used the Stack Overflow's developer survey data to create a **salary range prediction tool** for software professionals. Our data processing and modelling decisions were driven by focusing on the large USA dataset from the years 2023 and 2022. Our current application allows querying salary ranges by inputting developer attributes. It also visualizes technology trends in the technology community. We also provided a basic chatbot functionality to answer any question related to the survey data. In summary, the tool provides personalized and equitable compensation insights to inform career decisions. Going forward, we intend to expand capability for more countries, enhance chatbots and as more survey data gets released over time, we will retrain the model to keep salary predictions relevant eventually building a pipeline. To conclude, we transformed raw survey data into an empowering career guidance application that will enable data-backed technology career choices.

References:

Data Source: https://insights.stackoverflow.com/survey

https://survey.stackoverflow.co/2023/