Software Developer Salary Estimator

Elena Gialamas, 04.11.2024

DESCRIBE THE PROBLEM

Problem Statement and Scope

The goal of this project is to develop a machine learning model to estimate the average salary of a Software Developer. This prediction is based on features likely to influence earnings, such as experience, education, and country of employment. The project aims to centralize information on developer salaries, reducing the need for users to research across multiple sources.

Business Objective

The primary goal is to provide salary predictions that support career planning. Current salary estimators for specific occupations often rely on limited user-reported data, leading to biased outputs and overlooking various influential factors. This project aims to offer a comprehensive salary overview for individuals exploring a career in software development.

Metrics

Success will be measured by the model’s ability to predict salaries within an acceptable range of error. The primary metric is Root Mean Squared Error (RMSE), which penalizes outliers more severely, helping control prediction variance. Complete accuracy isn't the goal; an average estimate is often more useful for career planning. For example, the test set showed an RMSE error of approximately €13.000, indicating an acceptable margin of error and, therefore, success within the intended scope.

Data

Data Source and Structure

This project uses data from the annual Stack Overflow Developer Survey, which had approximately 65,000 participants. The survey covers various topics, including coding practices, technologies, tools, AI, and workplace experience. The dataset is structured in CSV format.

Data Setup

The survey includes over a hundred questions, but most are irrelevant to this model. Initial work with the dataset focused on understanding the survey structure, specific questions, and naming conventions.

Ethical Considerations

The Stack Overflow survey data is anonymized and voluntarily provided, so privacy concerns are minimal. However, the survey population may not fully represent all demographics and regions, introducing potential bias. Dishonest answers in the survey might also be a reason of concern. Transparency around data structure and collection methods is therefore important.

Data Preparation

Since the survey was voluntary, many entries have missing values. Salary data, in particular, is often incomplete due to its sensitive nature. As salary is the target value for prediction, imputing these missing values could introduce significant bias, so rows with missing target values were dropped. This drastically reduced the amount of available data. Additionally there are issues like differing currencies that needed to be addressed.

User experience was also a key consideration. Many categorical features had a wide variety of labels, so some were grouped to simplify user interaction. Though this might have led to reduced accuracy. On top of that Imputing, Encoding and Standardization/Scaling were necessary depending on the model.

Modeling

Model Selection and Baseline Performance

Three models were tested: Linear Regression, Decision Tree Regressor, and Random Forest Regressor. RMSE results were used to compare their performance.

Evaluation and Iteration

Pipelines were used to integrate preprocessing and training steps, allowing simultaneous hyperparameter tuning for scaling, imputation, and model training. Additionally, the pipeline could be used for prediction in the application. This reduced the amount of work needed for user imput, since preprocessing was automatically taken care of.

Deployment

Deployment Plan

The model will be deployed as a web application, allowing users to input their details and receive salary predictions directly. For this purpose Streamlit Community Cloud will be used.

Monitoring, Maintenance and Improvements

Since Stack Overflow updates its survey annually, the training process would need to be redone each year. This is especially important as the survey questions are updated to reflect current topics. Data analysis will therefore play a role in identifying new influencing factors.

While the model performs reasonably well, it is currently trained on only one dataset, and only a few models were tested. Improving performance and reliability will require a more iterative approach, including testing additional models and increasing computation. Relying solely on one data source may also introduce bias, which could be mitigated by diversifying data sources.

Resources

Dataset: https://survey.stackoverflow.co/2024/professional-developers/ General Tutorial: https://www.youtube.com/watch?v=xl0N7tHiwlw

Scikit Learn: https://scikit-learn.org/stable/index.html Streamlit Documentation: https://docs.streamlit.io/get-started Python: https://www.python.org/doc/ pandas: https://pandas.pydata.org/docs/

Missing Values function: https://towardsdatascience.com/cleaning-missing-values-in-a-pandas-dataframe-a88b3d1a66bf

Imputation theory: https://www.youtube.com/watch?v=m_qKhnaYZlc Imputation tutorial: https://www.youtube.com/watch?v=KWrZ59nLLSg

Gernal pipelines: https://www.youtube.com/watch?v=xIqX1dqcNbY Cross val with pipelines: https://www.youtube.com/watch?v=f_xB7kbZR_g

Encoding: https://www.youtube.com/watch?v=irHhDMbw3xo

currency convertion source: https://wechselkurse-euro.de/

ChatGPT: https://chatgpt.com/