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 to support career planning. Current salary estimators for specific occupations often rely on limited user-reported data, which can lead to biased outputs and ignore key 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 heavily, 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 success within the project’s 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, though 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 could also be a 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 missing salary data could introduce bias, so rows with missing target values were dropped. This significantly reduced the available data. Additionally, there were issues like differing currencies that needed to be addressed.

User experience was a key consideration. Many categorical features had a wide variety of labels, so some were grouped to simplify user input. Though this might have led to reduced accuracy, it improved usability. Other steps, such as imputation, encoding, and standardization, were applied as needed 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 predictions in the application, simplifying the workflow by automating preprocessing.

Computation

A grid search was used on the entire pipeline. That included five fold cross validation. On top of that

Deployment

Deployment Plan

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

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 currently relies on only one dataset, and only a few models were tested. Improving performance and reliability will require more iterative testing and additional data sources to reduce bias.

Challenges During the Project

Data Cleaning

Data cleaning was one of the main challenges. Both target and feature columns had missing values, so much of the data had to be removed, reducing the dataset size significantly. Stratifying the data split to ensure equal representation across countries created further issues, as some other features were still unevenly represented. To avoid data leakage, training and test sets were preprocessed separately, but this revealed additional missing values in certain columns after the split, affecting model performance.

A Simple Imputer was used for categorical features for compatibility with the pipeline, but it risked adding bias. A more advanced imputer, such as an Iterative Imputer, could have provided more accurate results by predicting missing values using machine learning. This was not included due to two concerns. One, reliability, as it was an experimental feature during the creating of this project. Two, computation time.

Computation

The project required long computation due to the use of grid search and five-fold cross-validation across the entire pipeline. While simpler models, such as Linear Regression and Decision Trees, could complete training within a couple of hours, the Random Forest model took several days to fit. These long computation times limited the models and hyperparameter combinations that could be tested.

Conclusion

More time for testing could allow for a more thorough search for the best model. Careful data cleaning is critical, as well. Early research on optimal preprocessing methods could improve results. More advanced imputation methods and additional computational power would likely enhance the model’s performance and accuracy.

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/