Devepoler Salary: Prediction

1. Problem statement

In the global software industry, determining appropriate salary expectations for software-related positions is a complex process, influenced by factors such as country, experience, and level of education. Job seekers often struggle to assess their market value, while employers face challenges in offering competitive salaries that attract and retain top talent. As a result, there is a need for a reliable and user-friendly tool that can provide salary predictions tailored to these key variables.

The application empowers job seekers to better understand their worth in the job market and negotiate appropriate compensation, while also assisting employers in benchmarking salary offerings against industry standards.

2. Solution approach

The development of an intuitive web app to predict the developer's salary based on country, education level, and experience. The prediction will be a reference parameter for a better understanding of software market conditions. The following steps are applied:

a. Real-data collection.
b. Data cleaning and formatting.
c. Outlier handling
d. Machine learning development.
e. Test web app deployment.

3. Stakeholders and Benefits

a. Programmers/Software Developers: They can use the app to get an estimate of their potential earning capacity based on their current or planned education and experience level. It can guide their career development, salary negotiations, or job search strategies.
b. Employers: Tech companies, startups, or any employer in need of programmers can use the tool to understand the average salaries for specific roles based on the required experience and education. It could inform their budgeting, compensation strategies, and recruitment processes.
c. Recruitment Agencies: These agencies can use the app to guide their clients (both companies and job-seekers) about potential compensation for different roles, thereby improving their service effectiveness.

4. Data

The real data of software developers' salaries was obtained from StackOverflow Survey of 2021, where the data analysis of the raw data can be accessed.

5. Tools

Data visualization, Data cleaning, Data processing, Sklearn, UX, Streamlit.

6. Results

The user inputs are job vacancy country, education level and years of experience. The web app can be accessed here: Developers Salary App. The complete code of the project can be accessed here: GitHub Project.

6.1. Data cleaning and formatting

The raw data of salary programmers was obtained from StackOverflow Survey, covering over 80000 instances according to 48 features, such as education level, experience, country, technologies, employment, state, learn code methodology, age, and gender, among others. The completion of many of the survey fields was not mandatory. For this reason, raw data presented a significant amount of missing data.

For this example, only job vacancy country, education level, and years of experience were considered. The instances with missing values were removed taking into account the large amount of data.

The original dataset contains 166 categories for the "Country" variable. An excessive number of categorical values with few instances could generate a machine-learning model with overfitting. Thus, the model focuses on exception and shows less capacity for generalizing different values for prediction. Regarding this purpose, the categories with less than 400 instances were grouped into the "Other" category. In this way, 18 countries remained.

The education level ordinal category also presented a high number of values, which were grouped into the following categories: "Master's degree", "Bachelor's degree", "Postgrad", and "Less than Bachelor's degree".

6.2. Outlier Handling

Programmer salaries values can cover a broad range. Fig 1 shows the boxplot of salaries as function of countries.

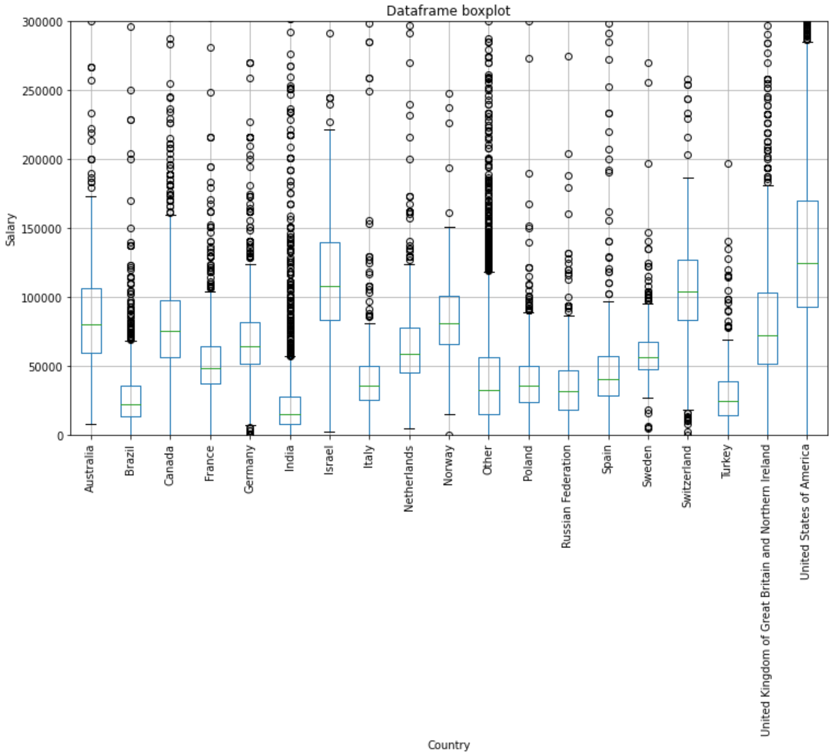


Fig 1. Programmer salaries by countries.

As shown in Fig 1, raw data present a significant number of outliers, especially in the upper quartile. The salaries higher than 270000 and lower than 15000 were removed to build a more representative machine learning model.

6.3. Machine learning model

The cleaned data consists of a combination of numeric and categorical values. LabelEncoder processing methodology (Scikit-Learn library) was applied to transform categorical data into numerical data.

The data was split: 80% training, and 20% testing, random seed: 42 (model reproducibility). Linear Regression, DecisionTreeRegressor, and RandomForestRegressor algorithms were tested. The performance metric selected was Mean Squared Error (MSE). It is a measure of the average squared difference between the actual values (ground truth) and the predicted values generated by the model. Table 1 shows the different algorithms tested with their metrics.

Table 1. Machine learning algorithms tested.

Algorithm	Description	MSE [$]
Linear Regression	The model assumes a linear relationship between the input features and the output, and it aims to minimize the sum of the squared differences between the actual and predicted values (least squares method).	$42882
DecisionTreeRegressor	The tree is built by repeatedly splitting the dataset based on the feature that leads to the largest reduction in the residual sum of squares. Decision trees are able to handle non-linear relationships, missing data, and categorical features. However, they can be prone to overfitting.	$32954
RandomForestRegressor	It is an ensemble learning method that combines multiple DecisionTreeRegressors to improve the overall performance and reduce overfitting.	$32880

According to the metrics values obtained, The RandomForestRegressor shows the best performance fitting the data (lowest MSE). The RandomForestRegressor hyperparameter optimization was carried out using GridSearch methodology.

Hyperparameter tuning of a RandomForestRegressor using Grid Search involves systematically searching through a predefined set of hyperparameters to find the best combination that minimizes the MSE. The hyperparameters tunned included: trees estimators (n_estimators), maximum depth of the trees (max_depth), the minimum number of samples required to split a node (min_samples_split), and a minimum number of samples required at each leaf node (min_samples_leaf).

The pipeline, including data processing and best RandomForestRegressor, was saved through binary protocol using Pickle library, to be deployed in the web app. The code of data cleaning, processing, and machine learning testing and tunning can be accessed here.

6.4. Model deployment

The model was deployed into a test web app using the Streamlit framework. The app can be accessed here: Salary Developers. Fig 2 shows the average programmer salary by country. The USA exhibits the highest average salary per year ($125000), followed by Israel ($112000), and Switzerland ($104000). Similarly, it is important to highlight that in these three countries, the cost of living is more expensive than in the rest of the countries, leading to higher salaries. In agreement with this observation, Brazil, India, and Turkey (countries with the lowest economic cost of living), present the lowest programmer salaries (around $32000/year).

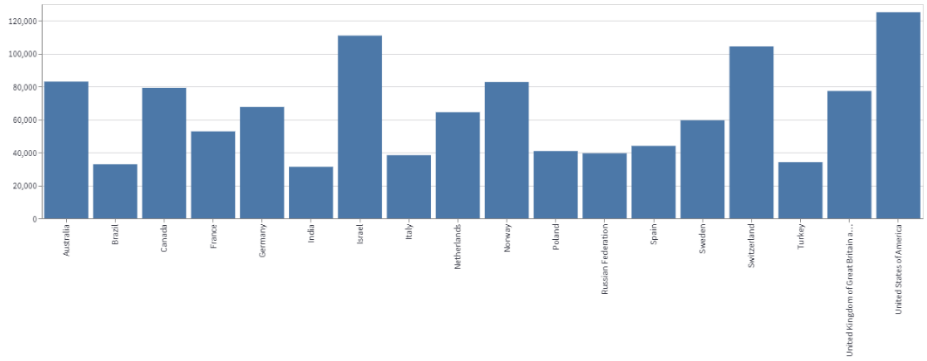


Fig 2. Average developer salaries by countries (web app cation).

Fig 3 shows the average programmer salary in all countries by experience (years). From an overall perspective, the base annual salary is around $50000. The annual salary increases with years of experience. The general trend is that experienced programmers achieve higher job positions with better salaries. A pronounced decrease is observed above 47 years of experience. Also, variations in annual salary are observed for more than 35 years. Only 2.64% of all data represent programmers with more than 35 years of experience. In this way, the variation in the general trend can be attributed to the lack of data for ages between 35 and 50 years old and the broad standard deviation in the cited range of age [1].

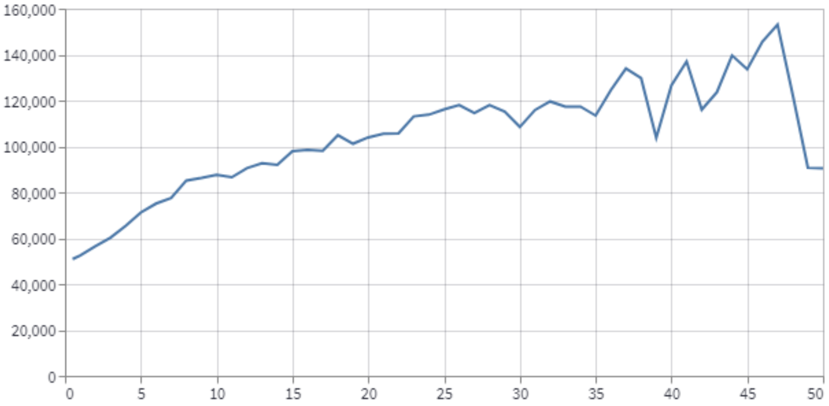
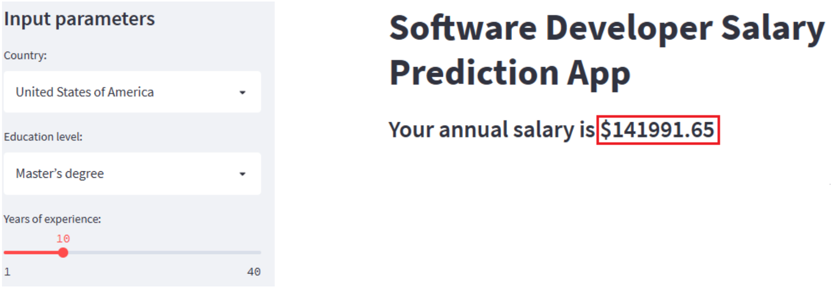


Fig 3. Programmer salary by years of experience (web app caption)

Fig 4 shows a caption of web app for programmer salary prediction. The user input parameter are Country, Education Level, and years of experience, given the salary prediction output based on the deployment of pipeline and machine learning model.



Input parameters

Country: United States of America

Education level: Master's degree

Years of experience: 10

Software Developer Salary Prediction App

Your annual salary is $141991.65

Fig 4. Programmer salary web app (caption).

7. Conclusions

a) Web app to predict and analyze program salary based on multiple features. Extra functionalities and features for the prediction can be incorporated.

b) Data cleaning and processing techniques were applied to reduce and transform the categorical features. Also, outliers handling was performed.

c) The top three countries with higher annual salaries were: the USA ($125000), Israel ($112000), and Switzerland ($104000). The high salaries in these countries correspond with an expensive cost of living.

d) RandomForestRegressor (ensemble model) algorithm exhibited the best metric performance (MSE) between all models. The tunning hyperparameters of the selected algorithm were performed using the Grid Search methodology.

e) A pipeline with data processing (including label encoding) and machine learning input and output was performed. The deployment was carried out using the Streamlit Framework server.

8. Further works

a) Incorporate extra statistical functionalities into the web app. Exploratory Data analysis can be found in the raw data source.

b) Analyze the correlation between programmer salary with the cost of living by country

c) Test different RNN architectures model prediction.

9. References

[1] Stack overflow Survey 2021