Insights from the stack overflow developer survey 2020

Lei Shen

<https://github.com/idjuven/Stacksurvry2020>

SHENL024@hotmail.com



**Introduction of Stack Overflow developer survey 2020**

The annual Stack Overflow developer Survey is one of the largest surveys for people who code for living all over the world. The aim of this year’s survey is the diversity and about 65000 developers participated this survey. The survey contains about 60 questions, with the scope from demographics, coding skills, compensations etc.

In this post, I am trying to answer three questions: 1. Which programming language is the most popular for current and future developers? 2. Is there any relationship between the Age, working hours and compensation? 3. Can we build a linear regression model to predict the salary?

**Data Preprocessing and visualization**

I preprocessed the data and obtained information such as number of rows and columns in the data frame, set of columns with 0 missing value and set of columns with lot of missing values. To find the professional distribution of developers, I plotted the distribution in figure 1.

As we can see, almost 75% of the survey participants are “I am a developer by profession”. About 13% of the survey participants are “students who are learning to code”. About 10% of the survey participants are “not a primary developer, but code occasionally for the work”. About 3% of the survey participants are not developer but code as a hobby.

Regarding the educational background, more than 40% of the participants have bachelor’s degree. 20% of the participants have master’s degree and 10% of the participants have some college education but without earning a degree.

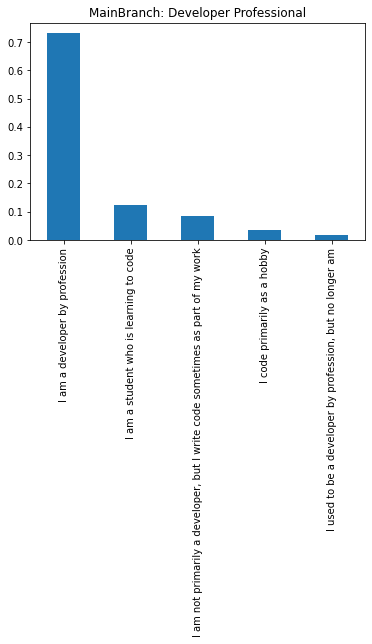


Figure 1: MainBranch: Professionals of developers who took the survey.

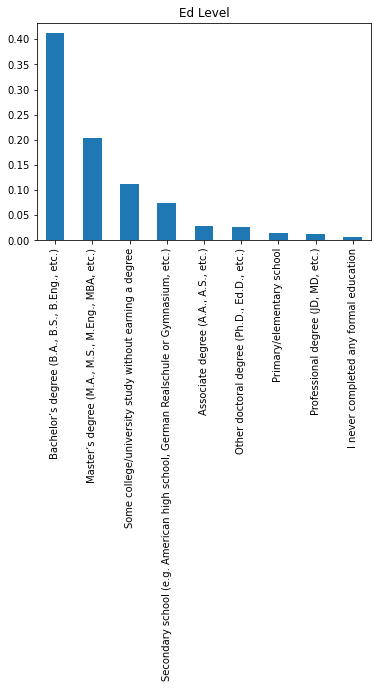


Figure 2: Education background of developer who participate in the survey

**Question 1: most favorite programming language (now and future)**

Table one and two list top 5 languages are current in use and desire for the next year respectively. From the table, we can see that HTML/CSS, Javascript, SQL, C#, PHP are among the top 5 programming languages the developers are using for their daily work.

On the other hand, the most favorite programming language for developers in the coming years are python and Rust. They are not in the top 5 popular languages developers are using. This shows the trends in the future. Therefore, students who are still in colleges may pay attention to this survey and work hard to sharp their skills in these two languages: python and Rust. These two are the most demanding skills for the job market in the next few years.

|  |  |
| --- | --- |
| HTML/CSS; JavaScript; PHP; SQL | 1073 |
| HTML/CSS; JavaScript | 900 |
| C#; HTML/CSS; JavaScript; SQL | 856 |
| C#; HTML/CSS; JavaScript; SQL; TypeScript | 754 |

Table 1: Top 5 languages developers are currently using

|  |  |
| --- | --- |
| Python | 1152 |
| Rust | 528 |
| HTML/CSS; JavaScript; TypeScript | 499 |
| C# | 461 |
| Go | 412 |

Table2: Top 5 languages desired for next year

**Question2: any relation between age and salary, working hours and salary?**

I also listed all the numerical columns and try to see if there is any correlation between them using the heatmap plot from seaborn. These columns include “Respondent”, “Age”, “CompTotal”, “CovertedComp”, and “WorkWeekHrs”. As we can see from figure3, there is no apparent relations as indicated by the numbers on the heatmap.

One may have the conclusion that the older the developer is, the higher salary he or she is getting paid since typically more experienced developers earn more salary. One may also conclude that the longer you work every week, the more you will make. According to this year’s survey, I do not see any necessary relationship between salary and age, salary and working hours.

According to the heat map, the correlation coefficient between salary and age is 0.11, so there might be some relation between salary and age. However, the correlation coefficient between salary and working hours is only 0.03. Therefore, their relation is not clear. The longer time you work does not mean that you will get paid more.

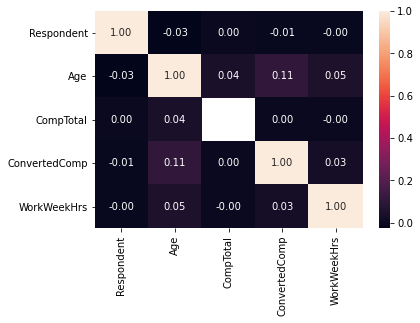


Figure 3: Heat map plot of relationship between different columns in the data frame: Respondent, Age, CompTotal, ConvertdComp and WorkWeekHrs.

**Question3: linear regression model for salary prediction**

One main interesting question is that if we can build a machine learning model that can be used to predict the salary for people with different age, weekly working hours. Since there are many missing values in the columns of “Age”, “WorkWeekHrs”, “CompTotal” and “ConvertComp”, the first thing to do is to drop these NAN values in these columns.

I set X as “[['Age','ConvertedComp','WorkWeekHrs']]”, y as “df\_dropna['CompTotal']”. I then constructed the linear regression model. The model performance was evaluated by the R2 and mean square errors (Table 3).

|  |  |  |
| --- | --- | --- |
| Model Performance | R2 | Mean Square Errors |
|  | -2.0844157589456897e+76 | 1.784543165147552e+91 |

**Conclusions**

With the power of data science, I did a research on the 2020 stack overflow developer survey. There are a lot of information in the survey, I only focus on three questions:

1. **What are the most favorite programming languages for the developers and what do they recommend for the future?**

**HTML/CSS, Javascript, SQL, C#, PHP** are among the top 5 programming languages the developers are using for their daily work.

The most favorite programming language for developers in the coming years are **python and Rust.**

1. **Is there any relationship between Age and salary? Is there any relationship between weekly working hours and salary?**

Unfortunately, I do not find any direct relation between age and salary, weekly working hours and salary. Older people do not necessarily earn more than younger developers. People spend more time in working every week do not get paid more than developers who work less hours per week.

1. **Can we build a machine learning model to predict the salary?**

We successfully constructed a linear regression model to predict the salary. The performance is fairly good.