Making out trends in computer science using the Stack Overflow survey

Marwa Ebesh
Matrikelnummer 6161671
marwa.ebesh@uni-tuebingen.de

Abstract

Every year, Stack Overflow conducts a massive survey of people on the site, covering all sorts of information like programming languages, salary, code style and various other information. We are planning to analyse the Stack Overflow survey to make out trends in computer science and compare groups of different nationality/age/gender/education level.

1 Introduction

Since 2011, Stack Overflow asks developers about their favourite technologies, job satisfaction, and age of first code, as well as how much they earn, and level of education. The year 2019 represented a large group of respondents: nearly 90,000 developers took the 20-minute survey that year. As that Stackoverflow is the world's largest and most trusted community of software developers, We want to empower developers by providing them with rich information about themselves. And these precious information can be used to educate employers about who developers are and what they need. The aim of the study is to gain insight into the careers of programmers. Of particular focus are the factors that influence a developer's salary and the tools used. I address the overall goal through 4 specific questions:

1.1 Working hypothesis

1. What are the most popular languages and tools among developers?
2. How are programmer's salaries effected by years of programming experience?
3. Is it possible to predict whether a programmer is looking for a new job or not?
4. Is job satisfaction related to other features recorded in the survey – like salaries?

1.1.1 Assumptions

- Stack overflow's data is representative of the community of data scientists
- All salaries listed are in USD (see their methodology for information on how they converted local currencies used by respondents to U.S. dollars)
- Years spent programming is a good indicator of professional experience as opposed to years spent coding as a job

2 Description of the data

Data is directly taken from StackOverflow and licensed under the ODbL license. The data were downloaded from the platform Stack overflow. Different data sets were suggested and I picked the most suitable for the analysis. The data set for the annual survey has a high usability and is well suited for analyses.

- Data source: `https://insights.stackoverflow.com/survey`
- Size: 77,3 MB
- Rows: 88883
- Columns: 85

2.1 Brief Description for most important Columns

surveyresultsschema.csv lists each survey question in long form. There are a total of 85 different Questions. The questions used for analysis are the columns of our dataset, the most important ones are:

- Country: Stack Overflow serves the international community, and the survey received responses from almost every country on Earth.
- Edlevel: Which of the following best describe your education level?
- ConvertedComp: What is your current annual base salary, before taxes, and excluding bonuses, grants, or other compensation converted to annual USD salaries using the exchange rate on 2019-02-01, assuming 12 working months and 50 working weeks.
- YearsCode: Including any education, how many years have you been coding?
- HaveWorkedLanguage: Which of the following languages have you done extensive development work in over the past year?
- JobSat: How satisfied are you with your current job?
- JobSeek: Which of the following best describes your current job-seeking status?
- Age: What is your age (in years)

2.2 Data type

int64, object and float64

2.3 Missed values

The stack overflow survey data set is great to work with as it has many data points (respondents). Unfortunately, it also comes with a lot of missing data points and it is not clear why exactly each of these data points are missing. For instance, looking at the data regarding languages that developers have worked with, it is not clear if NaN represents :

1. a null value because respondents were lazy and did not fill this in, or
2. if respondents did not work with any languages in the past year Because of this ambiguity, we have elected to sometimes drop data entries with null and sometimes (due to lack of entries) to fill it with mean values. Of the 88883 surveyed, there is a considerable amount of missing data points for most of the columns, However, the most affected columns (with 35 percent missed values) are: ['CompTotal', 'ConvertedComp', 'CodeRevHrs', 'BlockchainOrg'].

3 Description of data preparation

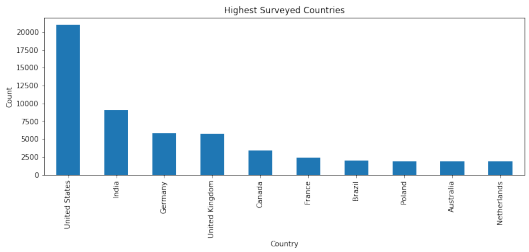
The data is relatively clean in its native form, however, there were two important preprocessing tasks:

1. Convert 2 categories in YearsCode: '50 or more years' and 'less than a year' from its current form to a numeric value arbitrarily to 50 and 0.
2. For analysis, Since the job satisfaction question has categorical values, I decided to map these into numerical values with a scalar from 1(very bad) to 4(very good).

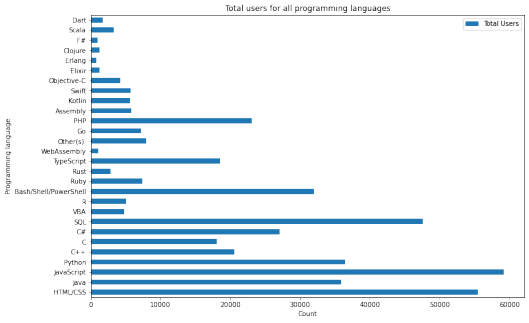
Since the first question should be answered by creating a machine learning predictor for the data, and the second question we will analyze the correlation between job satisfaction and salary, we need to prepare the data to be used properly. Therefore, we converted numerical NaNs to the mean of the column and converted categorical binary values for the first one. Moreover, we deleted all NaN entries for the second one.

4 Data Analyses process

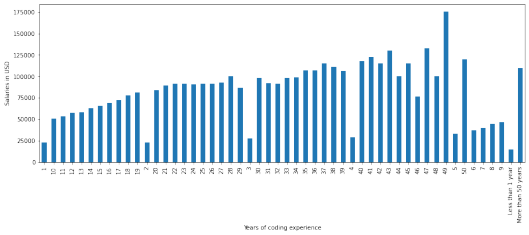
I started by finding out what are the highest Surveyed Countries.



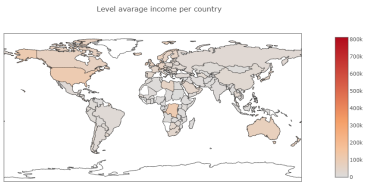
Then I tried to discover which programming languages have been used in that year.



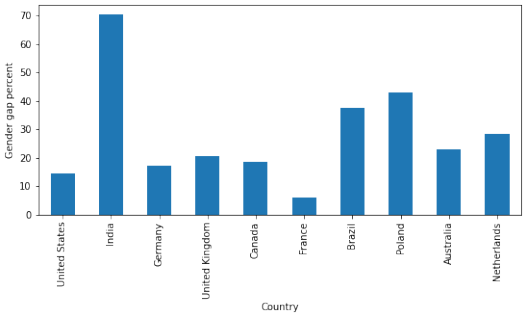
It was also interesting to figure out the relationship between salaries and years of coding experience. Interestingly, the stack overflow survey has elected to find out the amount of years developers have been coding / coding on jobs using 1 year bins.



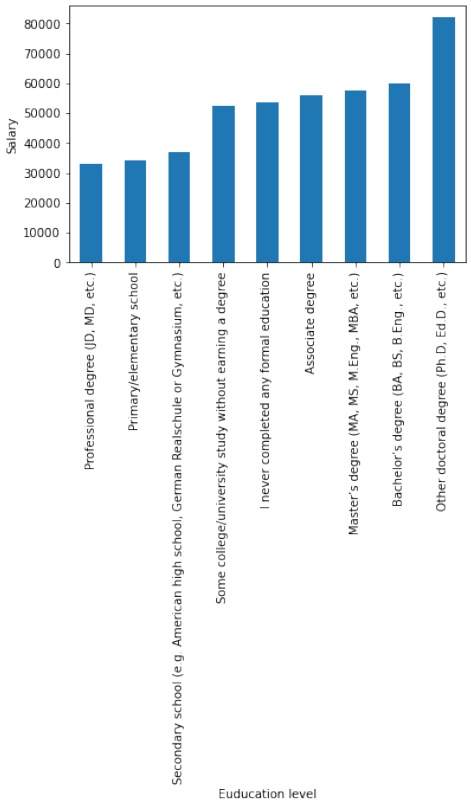
The following map shows the level of average income per country.



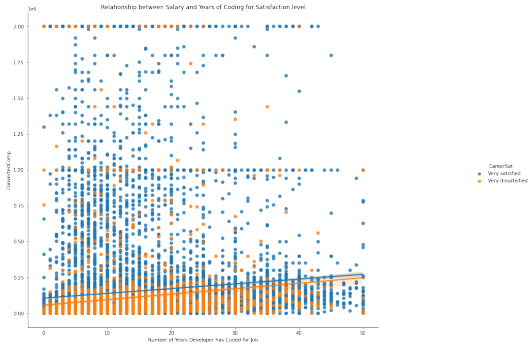
Sadly, gender pay gap is present in developer industry.



As expected, education level has an effect on salaries.



The most interesting: Relationship between Salary and years of coding for satisfaction level Also, on the top of the plot, there appear to be a number of outliers.



My research shows that there is no correlation between making more money and a higher job satisfaction

4.1 Machine learning method

Is it possible to predict whether a programmer is looking for a new job or not? For this we will use three different methods:

- Logistics Regression
- Native Bayes Classification (BernoulliNB)
- Support vector machine (SVC)

4.2 Quality criteria

We would like to achieve a score of at least 0.75

5 Application of different algorithms

5.1 Logistic Regression

Logistic regression is a very robust classification method which is able to predict the probability of a dependent binary target variable.

5.2 Bayes Classification

The Bayes classification with Bernoulli is a statistical classification method which is designed for binary/Boolean features. In contrast to Bayes, Bernoulli refers to the principle of decisions under risk. This method was added that it is well suited here.

5.3 Support Vector Machine

This algorithm divides data by means of an auxiliary vector in such a way into classes, that these have a distance to each other as large as possible. This algorithm is versatile due to its generalizability, and well suited for a binary classification method.

Table 1: Score table

Algorithm	Test-score	Cross-Validation 3 scores
Logistic regression	0.74	0.73 0.73 0.73
BernoulliBN	0.74	0.73 0.73 0.73
SVC	0.74	- - -

6 Lesson learned

Only the time factor prevented me from getting better results. The project has rounded what we have learned. I always tend to choose interesting (to me) relationships between variables that are complicated. In this case, I was trying to use regression to show a linear relationship between job satisfaction and other features, but honestly it's hard to prove because it's a complicated reality to model. so modeling it with a linear regression model is only sort of okay. The division of the y-values for JobSeek between True and False was not equal (73 percent - 26 percent) which might lead to a worse score since the algorithms can be affected by the size of the class, they will be preferred based on size.

- A common misconception about developers is that they've all been programming since childhood. In fact, we see a wide range of experience levels. Among professional developers.
- There are numerous useful questions in the data set and a plethora of insight questions beyond what is listed in the goals section. Further analysis insight questions can made in other projects.