What does it take to be a Software Developer?

Inês Garcia

Abstract—

More and more professionals and young graduates are now recognizing the importance of technology education as a way to prepare for a future dominated by information technology. More and more online and face-to-face courses and other resources are emerging to provide programming skills. Thus, the professional environment of information technology becomes increasingly competitive, and it is increasingly important to know what resources to invest in to ensure a successful career as a software developer. This project aimed to create a tool that can help Bootcamps identify the best strategies to ensure their students have high employability rates.

Keywords—Bootcamp, Machine Learning, Decision Tree, Decision, Programming.

1 Introduction

With the advancement of information technology, more and more professionals, graduates and students are investing in vocational retraining courses or online resources to gain new programming skills. In this sense, there is a proliferation of resources and services, as well as professionals with skills and knowledge in various programming languages. According to the World Economic Forum[1], by 2022 many jobs will be expected to be automated and by 2022 around half of the workers will need significant re- and upskilling. While the demand for software developers is expanding, available resources and courses are increasing the supply of skilled workers. To answer to this demand, several Bootcamp companies were created, and some report that the market is now overflowing with Bootcamp graduates[2]. This in turn raises the competitiveness of the labour market. Thus, just a little knowledge in programming languages is no longer enough to secure a job as a software developer, and some behaviours are needed to ensure better chances of securing a job in the area. Those who desire a career in software development have to stand out from the competition. Bootcamps have to make sure their students are successful in finding work after the courses, since the success of their students reflects the future success of the Bootcamp in question. Taking this into account, our goal in this project was to create a product that could help Bootcamps to understand the most important actions that a student may take, in order to increase its probabilities to achieve a job as a software developer. Creating a student profile, based on previous experience, both related and unrelated to programming, their behaviours as students and additional resources used, Bootcamps can identify which are the students that may need extra help, in order to secure a job as software developer.

In this report you will find first the methodology used in the second section. The third section regards the Exploratory Data Analysis main conclusion. After that you will find the machine learning process, where we disclose the features and models used, as well as the main results of the evaluation process. Lastly we report the main conclusion from this project.

2 Methodology

For this project, we used the results from the results from the freeCodeCamp's New Coder Survey from 2016, 2017 and 2018. This results were concatenated and cleaned together. After the exploratory data analysis, the target was identi-

• *Graduated in International Relations and Master in Strategic Studies. Certified in Data Analytics.*
E-mail: ines.garcia263@gmail.com

fied and we applied supervised machine learning algorithms in order to train a model that could predict the probability of becoming a software developer (target), based on the most relevant features available. We used supervised machine learning, using the decision tree classifier.

After the initial analysis of the data available other two hypothesis arose:

1) What are the most relevant features to become a software developer?
2) Are Bootcamps relevant to get a software development job?

Framewords used:

- Numpy
- Pandas
- Seaborn
- matplotlib
- SciKit-Learn
- imblearn

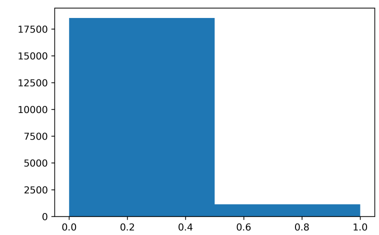
2.1 Data Cleaning and Manipulation

During the data cleaning process we started by identifying all the extra resources (Podcasts, In-person events, Online resources and Youtube channels). There was one column per resource with non-numeric entries. All the initial NaN were filled with 0, the non-numeric values were coerced to numeric, converting it to NaN, which were previously filled to 1. After this all the resources were summed, accordingly to their category.
Afterwards we dropped all the irrelevant columns, as well as the rows with NaN. In the end we got 44519 people that were not software developers at the time of the surveys and 12245 that were. As the data was very unbalanced for training our model, we under-sampled it randomly in order to have the same number of entries representing people that are and are not software developers.

3 Exploratory Data Analysis Results

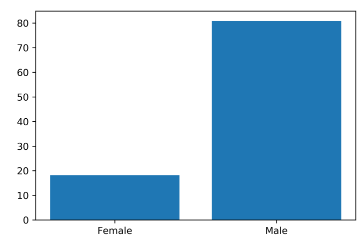
From the exploratory analysis of the data, we can first conclude that there is a small percentage of people integrating Bootcamp, only 6% of respondents.

Figure 1: People who attended a Bootcamp



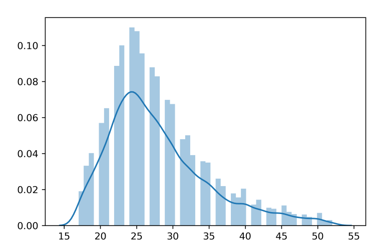
Additionally, only 18% of respondents are women, which reinforces what is found in the literature that there are still few women in the technology field.

Figure 2: Distribution by gender



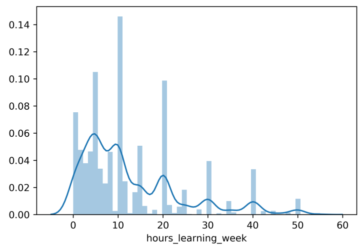
Regarding ages, we can observe that the majority of survey respondents are between 20 and 30 years old. The average age of the sample is 28 years, however, most respondents are about 25 years old. We conclude that young people within these ages are much less resistant to technology and are more likely to be new to programming skills than older workers with more years of experience.

Figure 3: People who attended a Bootcamp



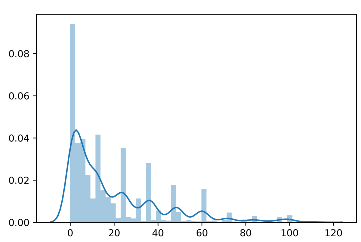
On average, respondents to this survey spend about 12 hours learning programming, whether through online resources, podcasts, YouTube videos, or other resources. Most respondents spend about 10 hours learning to program, while others spend about 20 hours.

Figure 4: Amount of hours that students invest in learning programming skills



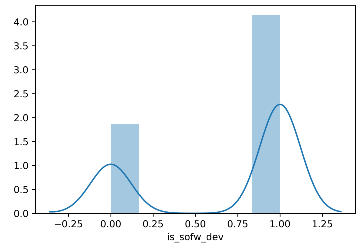
In our sample, we also concluded that the vast majority do not have much professional programming experience. About 47% of all respondents have less than 12 months of professional programming experience, 22% have less than 3 months of programming experience.

Figure 5: Progamming experience (in months)



Finally, and considering that we are talking about the importance of Bootcamps in ensuring students' success in getting a job as software developers, we looked at how many of the respondents who had participated in a Bootcamp got a job as software developers. As we can see, about 69% of those who participated in a Bootcamp managed to get a job as software developers.

Figure 6: People who enrolled in a Bootcamp and by status (software developers or not)



4 Machine Learning

The machine learning model used was the Decision Tree Classifier. This was the model that we though it would better fit our data and the result we wanted to achieve. To fit the model, serveral features were dropped. The final dataset, already balanced, had only the following features:

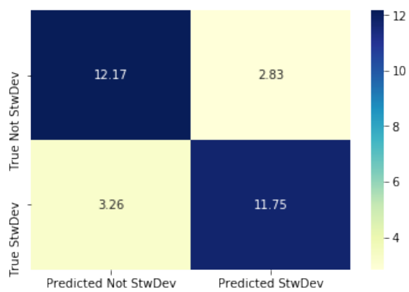
1) Number of hours invested in learning code, per week (hours_learning_week);
2) For how many months the person has been programming (months_programming);
3) If the person attended a Bootcamp or not (attended_bootcamp);
4) How much the person has expended in coding material and resourcs (coding_expenses);
5) Age (age);
6) Current work status (employment);
7) How many YouTube channels the person has used to learn code (youtube);
8) How many online resources the person has used to learn code (online_resources);
9) How many inperson event (job, hiring and tech fairs and events) the person has been in (inperson_events);
10) How many podcasts the person has found most relevant to learn code (podcasts);
11) If the person is currently working as a Software Developer (is_sofw_dev) - TARGET;

The information regarding gender was dropped as well, since there was an imbalance related to gender, were the male gender was the most prominent. The machine learning model was, then based on 10 features that can be easily answered by students in Bootcamps.

4.1 Machine Learning main Results

After fitting the model to our data set we the went to the evaluation of the accuracy, precision and recall. As per the accuracy, we got a value of 80.23% for the Train set and 79.73% for the Test set. This indicates that the model is not overfitting the Training set.

Figure 7: Confusion Matrix



Furthermore we got a Precision score of 81.15%, and a Recall score of 78.88%. This means that, out of 100, around 81 software developers were accurately predicted, and out of 100, around 79 people that were not software developers were accurately predicted.

Although both the accuracy, precision and recall do not have extremely elevated values, we believe that this is mostly due to the nature of the data in itself, that is self declarative. This kind of data is more prone to mistakes, errors and mismatches with reality.

5 Conclusions

The model created enables students and Bootcamps to have a sense of the probabilities of successfully becoming a software developer, based on individual learning habits and past experience. The goal was to create a model that could help target students with more difficulties in integrating the Tech market, so that Bootcamps could help and create the best policies and materials to guarantee that their students have the best outcome possible. After creating this model, several conclusion were possible to draw several conclusions, based on the learning model:

- As it would be expected, coding is the most important feature to guarantee a job as software developer.
- Previous work experience the second most important feature to guarantee a job as a software developer.
- For those with three months of coding experience or less and previous or current employment, Bootcamps appear as very important to guarantee a job as a software developer.

As we can see, this can be a very important tool for Bootcamps, but also for students to understand how to increase the probabilities of becoming a software developer in the furture.

References

[1] World Economic Forum, *The Future of Jobs Report 2018*. Available at: <https://www.weforum.org/reports/the-future-of-jobs-report-2018>.

[2] Marcel Degas, *Muse*. Available at: <https://medium.com/@marceldegas/san-francisco-bootcamp-bubble-cee59e48bf3e.64f229tc9>.