Machine Learning Engineer Nanodegree

Capstone Proposal

Chao-Wei Lo
June 1st, 2019

Domain Background

Stack Overflow is a question and answer site for programmers. As of January 2019, Stack Overflow has over 10 million registered users and it exceeded 16 million questions in mid 2018 [1]. Every year they will do a survey to ask the developer community about everything from their favorite technologies to their job preferences. In 2018, there are over 100,000 developers took this survey [2]. These developers came from all over the world, and the survey contained many interesting topics.

In my country (Taiwan), the software industry doesn't develop very well, especially compare to the hardware industry. So the culture of our software development is usually more like traditional hardware company, many companies don't encourage their employee to share their knowledge, in some cases, you might get in trouble if you share too many/detail information to the public. When I saw the second question in this survey was "Do you contribute to open source projects?", I was very curious about how other people would react to this question. Why some people would like to contribute to open source? Maybe we could find out some interesting facts in the survey result.

Problem Statement

In the capstone project, I am trying to find the most significant features that might determine a developer would like to contribute open source project or not. For example, a junior developer might have more time and passions to join open source projects to do more practices, gain more experiences and earn the reputation from the open source community. But a senior developer might need to do more stuffs not relating to coding or spend more time to take care his/her family instead of the open source project. So the age might be an important feature to impact an open source community. If we can find more specific features that might impact the willing to join an open source project, then maybe we can have more ideas to encourage people to contribute open source.

Datasets and Inputs

I used the dataset provided by Stack Overflow [2]. This survey included many different topics such as basic information, work, education, career, technology, culture and etc. So there are total 129 columns. Here I am focus on "OpenSource" topic. So I will use the "OpenSource" as the output value. The other 128 columns would become the input values. The dataset has provided the detail schema for all columns, I listed the detail as below.

The dataset has total 98857 records. For the output "OpenSource" column, thres is no missing value. And the data distribution of the "OpenSource" column is:

```
Yes: 43087 (43.59%)
No: 55770 (56.41%)
```

Although it's not perfect balanced data (50%-50%), we can still treat this distribution as some kind of balanced data.

Input

- Respondent: Randomized respondent ID number
- Hobby: Do you code as a hobby? Yes/NO
- Country: Current reside country
- Student: Current student status
- Employment: Current employment status
- FormalEducation: Formal education level
- UndergradMajor: College major
- CompanySize: Approximately how many people are employed by the company or organization you work for?
- DevType: Which of the following describe you? Please select all that apply.
- YearsCoding: Including any education, for how many years have you been coding?
- YearsCodingProf: For how many years have you coded professionally (as a part of your work)?
- JobSatisfaction: How satisfied are you with your current job? If you work more than one job, please answer regarding the one you spend the most hours on.
- CareerSatisfaction: Overall, how satisfied are you with your career thus far?
- HopeFiveYears: Which of the following best describes what you hope to be doing in five years?
- JobSearchStatus: Which of the following best describes your current job-seeking status?
- LastNewJob: When was the last time that you took a job with a new employer?
- AssessJob1 ~ AssessJob10: Imagine that you are assessing a potential job opportunity. Please rank the following aspects of the job opportunity in order of importance (by dragging the choices up and down), where 1 is the most important and 10 is the least important.
- AssessBenefits1 ~ AssessBenefits11: Now, imagine you are assessing a job's benefits package. Please rank the following aspects of a job's benefits package from most to least important to you (by dragging the choices up and down), where 1 is most important and 11 is least important.
- JobContactPriorities1 ~ JobContactPriorities5: Imagine that a company wanted to contact you about a job that is a good fit for you. Please rank your preference in how you are contacted (by dragging the choices up and down), where 1 is the most preferred and 5 is the least preferred.
- JobEmailPriorities1 ~ JobEmailPriorities7: Imagine that same company decided to contact you through email. Please rank the following items by how important it is to include them in the message (by dragging the choices up and down), where 1 is the most important and 7 is the least important.
- UpdateCV: The main reason that you update CV
- Currency: The currency you use
- Salary: Current gross salary (before taxes and deductions), Please enter a whole number in the box below, without any punctuation. If you are paid hourly, please estimate an equivalent weekly, monthly, or yearly salary. If you prefer not to answer, please leave the box empty.
- SalaryType: Salary weekly, monthly, or yearly
- ConvertedSalary: Salary converted to annual USD salaries using the exchange rate on 2018-01-18, assuming 12 working months and 50 working weeks.
- CurrencySymbol: Three digit currency abbreviation.

- CommunicationTools: The tools you use to communicate, coordinate, or share knowledge with your coworkers
- TimeFullyProductive: Suppose a new developer with four years of experience, including direct experience working with your company's main technical stack, joined your team tomorrow. All other things being equal, how long would you expect it to take before they were fully productive and contributing at a typical level to your main code base?
- EducationTypes: Which of the following types of non-degree education have you used or participated in? Please select all that apply.
- SelfTaughtTypes: You indicated that you had taught yourself a programming technology without taking a course. What resources did you use to do that?
- TimeAfterBootcamp: You indicated previously that you went through a developer training program or bootcamp. How long did it take you to get a full-time job as a developer after graduating?
- HackathonReasons: You indicated previously that you had participated in an online coding competition or hackathon. Which of the following best describe your reasons for doing so?
- AgreeDisagree1: To what extent do you agree or disagree with each of the following statements? I feel a sense of kinship or connection to other developers
- AgreeDisagree2: To what extent do you agree or disagree with each of the following statements? I think of myself as competing with my peers
- AgreeDisagree3: To what extent do you agree or disagree with each of the following statements? I'm not as good at programming as most of my peers
- LanguageWorkedWith: Which of the following programming, scripting, and markup languages have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the language and want to continue to do so, please check both boxes in that row.)
- LanguageDesireNextYear: Which of the following programming, scripting, and markup languages have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the language and want to continue to do so, please check both boxes in that row.)
- DatabaseWorkedWith: Which of the following database environments have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the database and want to continue to do so, please check both boxes in that row.)
- DatabaseDesireNextYear: Which of the following database environments have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the database and want to continue to do so, please check both boxes in that row.)
- PlatformWorkedWith: Which of the following platforms have you done extensive development work for over the past year? (If you both developed for the platform and want to continue to do so, please check both boxes in that row.)
- PlatformDesireNextYear: Which of the following platforms have you done extensive development work for over the past year? (If you both developed for the platform and want to continue to do so, please check both boxes in that row.)
- FrameworkWorkedWith: Which of the following libraries, frameworks, and tools have you done extensive development work in over the past year, and which do you want to work in over the next year?
- FrameworkDesireNextYear: Which of the following libraries, frameworks, and tools have you done extensive development work in over the past year, and which do you want to work in over the next year?

- IDE: Which development environment(s) do you use regularly? Please check all that apply.
- OperatingSystem: What is the primary operating system in which you work?
- NumberMonitors: How many monitors are set up at your workstation?
- Methodology: Which of the following methodologies do you have experience working in?
- VersionControl: What version control systems do you use regularly? Please select all that apply.
- CheckInCode: Over the last year, how often have you checked-in or committed code?
- AdBlocker: Do you have ad-blocking software installed on any computers you use regularly?
- AdBlockerDisable: In the past month, have you disabled your ad blocker for any reason, even temporarily or for a specific website?
- AdBlockerReasons: What are the reasons that you have disabled your ad blocker in the past month? Please select all that apply.
- AdsAgreeDisagree1: To what extent do you agree or disagree with the following statements: Online advertising can be valuable when it is relevant to me
- AdsAgreeDisagree2: To what extent do you agree or disagree with the following statements: I enjoy seeing online updates from companies that I like
- AdsAgreeDisagree3: To what extent do you agree or disagree with the following statements: I fundamentally dislike the concept of advertising
- AdsActions: Which of the following actions have you taken in the past month? Please select all that apply.
- AdsPriorities1 ~ AdsPriorities7: Please rank the following advertising qualities in order of their importance to you (by dragging the choices up and down), where 1 is the most important, and 7 is the least important.
- AIDangerous: What do you think is the most dangerous aspect of increasingly advanced AI technology?
- AIInteresting: What do you think is the most exciting aspect of increasingly advanced AI technology?
- AIResponsible: Whose responsibility is it, primarily, to consider the ramifications of increasingly advanced AI technology?
- AIFuture: Overall, what's your take on the future of artificial intelligence?
- EthicsChoice: Imagine that you were asked to write code for a purpose or product that you consider extremely unethical. Do you write the code anyway?
- EthicsReport: Do you report or otherwise call out the unethical code in question?
- EthicsResponsible: Who do you believe is ultimately most responsible for code that accomplishes something unethical?
- EthicalImplications: Do you believe that you have an obligation to consider the ethical implications of the code that you write?
- StackOverflowRecommend: How likely is it that you would recommend Stack Overflow overall to a friend or colleague? Where 0 is not likely at all and 10 is very likely.
- StackOverflowVisit: How frequently would you say you visit Stack Overflow?
- StackOverflowHasAccount: Do you have a Stack Overflow account?
- StackOverflowParticipate: How frequently would you say you participate in Q&A on Stack Overflow? By participate we mean ask, answer, vote for, or comment on questions.
- StackOverflowJobs: Have you ever used or visited Stack Overflow Jobs?
- StackOverflowDevStory: Do you have an up-to-date Developer Story on Stack Overflow?
- StackOverflowJobsRecommend: How likely is it that you would recommend Stack Overflow Jobs to a friend or colleague? Where 0 is not likely at all and 10 is very likely.
- StackOverflowConsiderMember: Do you consider yourself a member of the Stack Overflow community?

- HypotheticalTools1 ~ HypotheticalTools5: Please rate your interest in participating in each of the following hypothetical tools on Stack Overflow, where 1 is not at all interested and 5 is extremely interested.
- WakeTime: On days when you work, what time do you typically wake up?
- HoursComputer: On a typical day, how much time do you spend on a desktop or laptop computer?
- HoursOutside: On a typical day, how much time do you spend outside?
- SkipMeals: In a typical week, how many times do you skip a meal in order to be more productive?
- ErgonomicDevices: What ergonomic furniture or devices do you use on a regular basis? Please select all that apply.
- Exercise: In a typical week, how many times do you exercise?
- Gender: Which of the following do you currently identify as? Please select all that apply. If you prefer not to answer, you may leave this question blank.
- SexualOrientation: Which of the following do you currently identify as? Please select all that apply. If you prefer not to answer, you may leave this question blank.
- EducationParents: What is the highest level of education received by either of your parents? If you prefer not to answer, you may leave this question blank.
- RaceEthnicity: Which of the following do you identify as? Please check all that apply. If you prefer not to answer, you may leave this question blank.
- Age: What is your age? If you prefer not to answer, you may leave this question blank.
- Dependents: Do you have any children or other dependents that you care for? If you prefer not to answer, you may leave this question blank.
- MilitaryUS: Are you currently serving or have you ever served in the U.S. Military?
- SurveyTooLong: How do you feel about the length of the survey that you just completed?
- SurveyEasy: How easy or difficult was this survey to complete?

Output

- OpenSource: Do you contribute to open source projects? Yes/No

Solution Statement

According the survey data, we will try to predict a person would like to contribute open source or not? So it's a supervised binary classification problem. Since the survey has many multiple choice, optional and similar questions, the dataset is messy instead of tidy [3], so the first thing to do is the data cleaning and feature selection. In the beginning, we need to drop some irrelevant features such as "Respondent", "MilitaryUS", "SurveyTooLong" and "SurveyEasy". After basic feature selection, we could apply one-hot encoding for the remaining features, but for the multiple-choice question, we have to apply customized one hot encoding for each answer because the calculation of the answer combination would be complex and meaningless. Then I will quickly try some different classification algorithms to find the best one. Finally we can leverage GridSearch method to tune the model parameters to get a better performance.

Benchmark Model

The project will select the decision trees as the benchmark model of the classification problem. The decision trees can perform prediction with the minimum effort of data pre-processing. It's performance doesn't affect by the value scale, missing or outlier values, and it's very intuitive and easy to interpret and explain [4]. As a benchmark model, we want the model to generate a consistent/reproducible result, so we can assign a

constant to the random_state and keep the other parameters as default value. Then we can validate the decision trees model with many different classification metrics.

Evaluation Metrics

For classification problem, confusion matrix is a specific table layout that allows visualization of the performance of an algorithm [5].

	Predicted Positive	**Predicted Negative**
Actual Positive	(TP) True Positive	(FN) False Negative
Actual Negative	(FP) False Positive	(TN) True Negative

Since the data distribution of "OepnSource" is similar to balanced data, here we select the "Accuracy" and "F1 score" as the classification threshold metrics [6].

- Accuracy: Accuracy measures how often the classifier makes the correct prediction. It's the ratio of the number of correct predictions to the total number of predictions (the number of test data points). The formula is :

```
Accuracy = (TP + TN)/ (TP + FP + TN + FN)
```

- F1 score: The F1 score is the harmonic average of the precision and recall. It has no preference between precision and recall. The formula is :

```
F1 score = 2 * (Precision * Recall)/(Precision + Recall)
```

    - Precision: Precision and recall have the similar formula. The different part is "Precision" cares about "(TP)True Positive and (FP)False Positive" but "Recall" cares about "(TP)True Positive and (FN)False Negative". Take spam email filter as example, we would like to choose "Precision" as the evaluation metrics because we can accept to check some spam mails in the mailbox (Spam mail filter is wrong and it's false negative.) but we don't want to miss any important mails (Spam mail filter is wrong and it's false positive.)

```
Precision = TP / (TP + FP)
```

    - Recall: Take health check as another example, we would like to choose "Recall" as the evaluation metrics, because we can accept to do more checks in the hospital (Health check is wrong and it's false negative.) but we can't afford to suffer from an serious illness without any alert (Health check is wrong and it's false positive.)

```
Recall = TP / (TP + FN)
```

Project Design

The project has following stages:

- Development environment preparation: Install & test the libraries. The development environment of my office doesn't allow internet connection, so I must learn how to offline install all necessary libraries.
- Data acquisition: Import the dataset of stack overflow survey 2018. This dataset has only one csv file need to import.
- Data preprocessing
    - Data Exploration: Detail check all columns, values and relations. Here we can leverage many plot tools to visualize and investigate the data. After we understood the feature meaning and correlation, we can determine which features could be dropped.
    - Cleaning null values (missing data): There were many NaN in the dataset. For single/multiple choice question, I will replace the NaN with "0". For continuous data features such as "Salary", I will replace the NaN with mean.
    - Outlier detection: Outlier might skew the result. For continuous data such as "Salary", we will use Tukey's Method for identifying outliers [8]: An outlier step is calculated as 1.5 times the interquartile range (IQR). A data point with a feature that is beyond an outlier step outside of the IQR for that feature is considered abnormal. Then we will drop the outlier data points.
    - Separate the features and labels: We need to separate the "OpenSource" feature as output label datas, the other features will become the input dataset.
    - Encoding data: Most features in the dataset are categorical data, we will use one-hot encoding to transform them for better performance.
    - Feature scaling: For numerical data such as "Salary", we use min-max scaler to normalize it's value for better performance.
    - Sample selection: We will split the dataset into training and testing set. We will fix the random_state to get the consistent result.
- Data modeling
    - Benchmark model selection: We select the decision trees as benchmark model. We will fix the random_state to get the consistent result.
    - Model selection: We choose following models from sklearn to challenge benchmark model. We will fix the random_state if the model is available.
        - tree: DecisionTreeClassifier (Benchmark model)
        - naive_bayes: GaussianNB
        - svm: SVM
        - linear_model: LogisticRegression, SGDClassifier
        - ensemble: AdaBoostClassifier
        - neighbors: KNeighborsClassifier
- Execution
    - Model training and tuning: We will use cross validation with grid search technique to mitigate the selection bias and overfitting issue. The training dataset would be split to cross-validation (training vs. validation) dataset. After we test all the models we selected, we can collect the performance metrics of all models.

    - Performance evaluation: We compare the metrics of all models to the benchmark model. Sometimes the result is very clear so we can find the best model, but sometimes we can't. In such bad cases, we might need to go further back to check everything and retry several times. If we still didn't get lucky to resolve the problem, we should have a conclusion to explain why the model doesn't work well.

Reference

- [1] Stack Overflow Wiki - https://en.wikipedia.org/wiki/Stack_Overflow
- [2] Stack Overflow 2018 Developer Survey - https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey
- [3] Hadley Wickham, Tidy Data - http://vita.had.co.nz/papers/tidy-data.pdf
- [4] 4 key advantages of using decision trees for predictive analytics - http://www.simafore.com/blog/bid/62333/4-key-advantages-of-using-decision-trees-for-predictive-analytics
- [5] Confusion matrix - https://en.wikipedia.org/wiki/Confusion_matrix
- [6] Rich Caruana , Alexandru Niculescu-Mizil, An empirical comparison of supervised learning algorithms, Proceedings of the 23rd international conference on Machine learning, p.161-168, June 25-29, 2006, Pittsburgh, Pennsylvania - https://www.cs.cornell.edu/~caruana/ctp/ct.papers/caruana.icml06.pdf
- [7] Preparing and Architecting for Machine Learning - https://www.gartner.com/binaries/content/assets/events/keywords/catalyst/catus8/preparing_and_architecting_for_machine_learning.pdf
- [8] Lab 5: Testing Our Way to Outliers - https://www.stat.cmu.edu/~cshalizi/statcomp/13/labs/05/lab-05.pdf