CMPT 353 Project Report

How Major Impacts in Multiple Aspects: Salaries, Job Satisfaction, and Stack Overflow Participation

Date: Dec 2, 2019

Myungjun Lee (301250558)
SeongJin Kim (301258579)
Darren Jae Jun Yang (301183838)
Simon Fraser University

1. Introduction

As a group of computing science students, we come across a question: would our major (Computing Science) outstands in various fields compared to other majors?

Following is the list of specific fields that we are interested to compare:

1. Does years of coding affect developer's number of participation in Stack Overflow?
2. Does years of coding affect salaries of developer's salaries?
3. Does years of coding affect developer's career satisfaction?

Our essential objective is to find out if there are any differences between computer scientists, and non computing scientists currently working in any fields. We start by preparing the data, in a form that we need. Then, we can use various tests, which includes: normality test, levene test, the finally t-test. Based on the results from these tests, we can find out the differences of the means of these two, if any. To investigate further based on these results, we use classifier to see if we can build a model that can correctly predict whether it is CS major, or non CS major, given a sample of data. Depending on these results, we can finally conclude whether CS major and non CS major have any significant differences.

2. Data Preparation

In this section, we will present how we acquired the dataset, and how we transformed such dataset into a form that we can use efficiently.

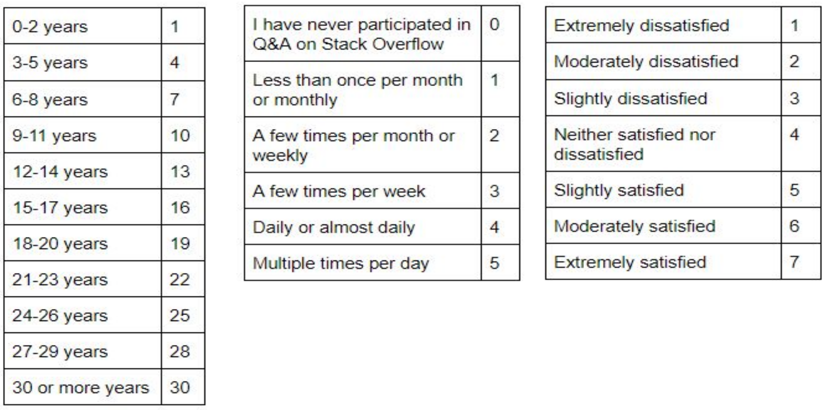
2.1 Data Acquirement

We were able to find relevant open source dataset on *Kaggle*:
https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey/
This dataset was created from a survey conducted by Stack Overflow in January 2018, asking the developer community about numerous interesting factors, which includes: Stack Overflow participation, salary, degree, major, gender, hours of sitting in front of computer, and much more. There are almost 99,000 responses in this survey, with 67,000 responses fully completed.

2.2 Data Cleaning and Transformation

From the dataset introduced above, we followed the following steps for cleaning and transforming purposes.

1) Filter out the ones that we do not need for our observations, and only retain the relevant fields: 'UndergradMajor', 'JobSatisfaction', 'StackOverflowParticipate', 'ConvertedSalary', and 'YearsCoding'.
2) Among these fields, we remove all rows that contains any invalid, or null values.
3) Some particular fields including participation and YearsCoding, the responses are in string format, such as: "Multiple times per day", and "12-14 years". We transform these data into numeric values as in *Figure 1*, left column showing pre-transformation, and right column showing post-transformation. Notice that the transformed numbers are arbitrary values that we set.



0-2 years	1
3-5 years	4
6-8 years	7
9-11 years	10
12-14 years	13
15-17 years	16
18-20 years	19
21-23 years	22
24-26 years	25
27-29 years	28
30 or more years	30

I have never participated in Q&A on Stack Overflow	0
Less than once per month or monthly	1
A few times per month or weekly	2
A few times per week	3
Daily or almost daily	4
Multiple times per day	5

Extremely dissatisfied	1
Moderately dissatisfied	2
Slightly dissatisfied	3
Neither satisfied nor dissatisfied	4
Slightly satisfied	5
Moderately satisfied	6
Extremely satisfied	7

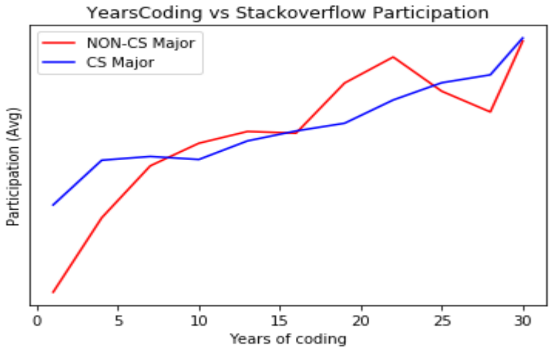
Figure 1. Transformations on YearsCoding, Participation, and Satisfaction respectively.

4) Now that we have transformed the fields into numerical format, we proceed to the data splitting process. As our intention is to find out the differences between Computing Science (CS) major, and non-CS major in various aspects, we simply divide the data into two data frames using UndergradMajor: CS_major, and non_CS_major.

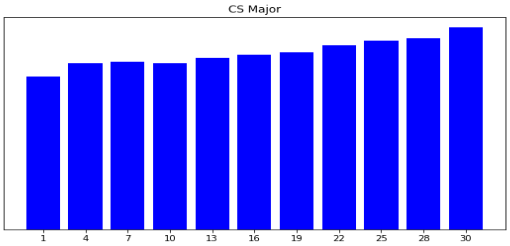
3. Data Analysis

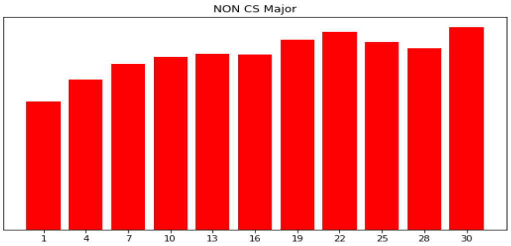
In this section, we are going to take a look at three different analysis in detail and explain similarities/differences between CS and Non-CS major.

Approach 1 - Stack Overflow Participation



As the diagram shown above, non-CS Major participates in Stack Overflow less than CS major students in the early years of coding. In the end, which is 30 or more years of coding, participation in Stack Overflow are around the same.



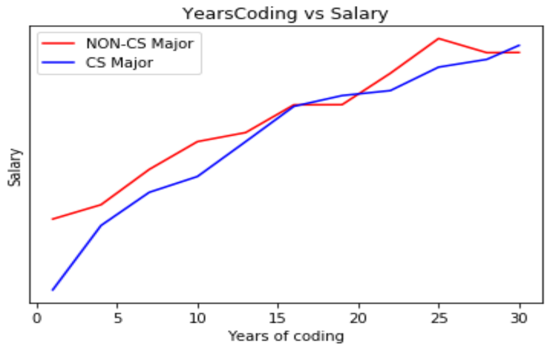


Before we obtain any p-values for different types of test, first we observed our datasets with the bar graph. Even by just looking at the bar graph shown above, the datasets look fairly normally distributed. CS major seems more normally distributed than non-cs major. Overall, both datasets look normally distributed. And we can confirm this by taking various tests as below.

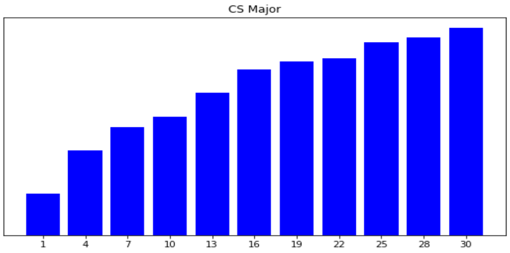
CS Major normaltest p-value	**0.919**
NON-CS Major normaltest p-value	**0.178**
Levene test p-value	**0.388**
T-test p-value	**0.742**

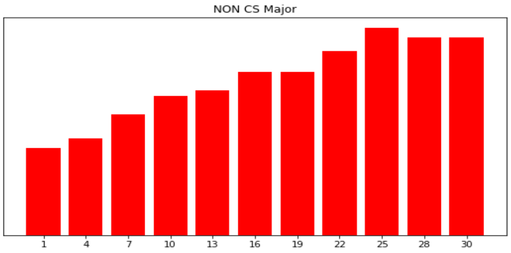
It shows that both datasets are normal since a normal test p-value is greater than 0.05 where p-value of cs major is much greater than non-cs majors. And we conducted levene test between two datasets and we have got p-value of 0.388 which is greater than 0.05, meaning they have equal variance. Since we figured out that both datasets are normal and have equal variance, we took a T-test between two data sets and we got p-value of 0.742 which is greater than 0.05, meaning it is very likely that both datasets have the same means. By analyzing line graph, bar graph, and various tests, it seemed quite evident that there is not much difference in developer's Stack Overflow participation whether developers have CS background or not.

Approach 2 - Salary



Developers with both CS- and Non-CS- major increase their salaries as they gain more experience of coding. The diagram shows that CS-major salary is mostly lower than Non-CS. In addition a salary of developers with CS backgrounds are steadily increasing while non-cs developer's salary sometimes stays the same (from 15 to 19 years of coding) or even drops (from 25 years to so on). In the end, it seems there is no much difference in salaries between cs-background and non-cs backgrounds.



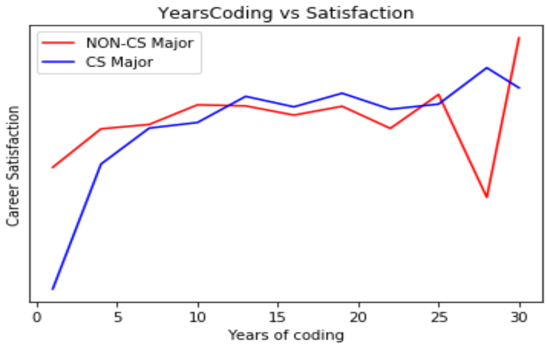


If we look at the bar graphs above, compared to Stack Overflow Participation's bar graph, the datasets seemed less normal. It makes sense because in the line graph, we analyzed that the salary increases as developers get more experience of coding regardless of their background. Therefore the bar graph is more left-skewed. We took various tests to understand our datasets better.

CS Major normaltest p-value	**0.475**
NON-CS Major normaltest p-value	**0.573**
Levene test p-value	**0.576**
T-test p-value	**0.557**

It shows that both datasets are normal since p-value is greater than 0.05. It shows that both datasets are normal since a normal test p-value is greater than 0.05. Also we conducted Levene test and p-value was 0.576 which is greater than 0.05, meaning they have equal variances. Since they are normal and have equal variance, we could have conducted T-Test of two datasets and p-value found out to be greater than 0.05 which means the means of both datasets are equal.

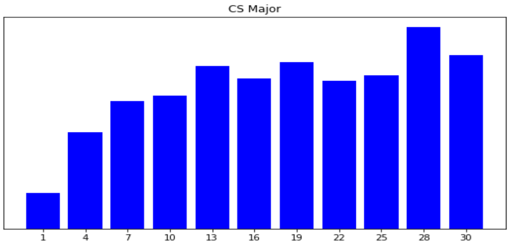
Approach 3 - Career Satisfaction

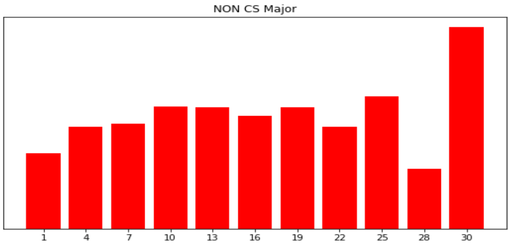


Graphs above shows drastic difference in career satisfaction between developer with cs- and non-cs major in beginning. People with CS-major background has very low career satisfaction during early years of coding while people with non-cs major has relatively high career satisfaction. Developer with cs-major shows less career satisfaction than non-cs until around 13 years of coding. However, developers with CS-background consistently becomes more satisfied with their careers while career satisfaction of developers with non-cs major background experiences up-and-down in their satisfaction. that eventually they become less satisfied than people with CS background. According to a diagram, satisfaction of developers with non-cs background drops drastically after 25 years of coding but there is a high chance that it is because of a faulty data.

CS Major normaltest p-value	**6.72e-05**
NON-CS Major normaltest p-value	**0.406**
Levene test p-value	**0.741**
T-test p-value	**0.811**

CS major dataset is definitely not normal because the p-value is smaller than 0.05. By looking at line graph, we know that data is left-skewed because career-satisfaction of developer with CS major background. Therefore we have transformed data and replot the diagram again.





We have used exponent (by 16) in order to transform. After transformation, the bar graphs look as above. This bar graph seemed to align with the line graph. Therefore we conducted the various test again.

CS Major normaltest p-value	**0.0245**
NON-CS Major normaltest p-value	**0.0248**
Levene test p-value	**0.998**
T-test p-value	**0.932**

We still find that datasets are not normal since p-values are less than 0.05. P-value for normal tests came out to be always less than 0.05 for both CS and Non-CS major but p-value became closer to 0.05. In addition, according to the Central Limit Theorem, if the size of dataset is more than 40, then we may assume data are normal. Then, T-Test of two datasets is conducted and p-value found out to be greater than 0.05 which means the means of both datasets are equal.

4. Classification

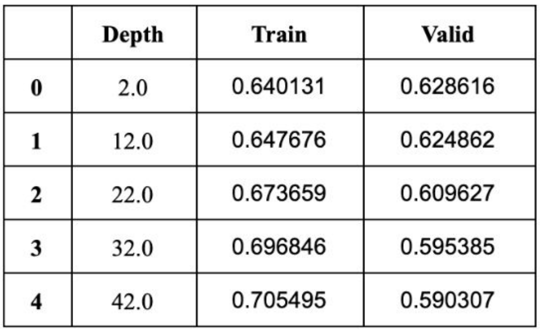
To investigate further, we used classifiers to see if the trained models can predict the major correctly. We start by labelling the feature — 0 is denoted as non CS major, and 1 for CS major. In order to train and test the specific values from the original data, we used salaries. We did not incorporate Stack Overflow participation and career satisfaction as these use arbitrary values that we set during the data transformation step. The classifiers that we use in this step are: GaussianNB, and DecisionTreeClassifier. We also used MinMaxScaler function in each pipeline, as the difference between the minimum and maximum value of salaries may be huge.

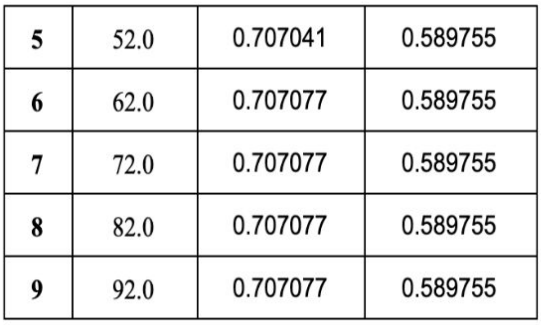
Starting with the **GaussianNB**, below is the model score that we trained and tested.

Train	Test
0.6388060800117773	0.6322587767719143

While only looking at the train and model score, which does not differ a lot, we can say that it is trained well, without any overfitting. However, model with 63% of accuracy is not considered as a good model, which misses a lot of true positives.

The next classifier we used is the **DecisionTreeClassifier**. For this classifier, we tried various values for the parameter — depth. Adjusting this variable was important to optimize the trade-off between train and test scores. Below is the model score with different depth.





	Depth	Train	Valid
0	2.0	0.640131	0.628616
1	12.0	0.647676	0.624862
2	22.0	0.673659	0.609627
3	32.0	0.696846	0.595385
4	42.0	0.705495	0.590307
5	52.0	0.707041	0.589755
6	62.0	0.707077	0.589755
7	72.0	0.707077	0.589755
8	82.0	0.707077	0.589755
9	92.0	0.707077	0.589755

As observed from the table above, the number of depth may increase the training performance by up to 70%. However, this leads to the overfitting as the highest score for the validation is less than 60%. The most optimal choice of the parameter is when depth number is set to 12 — both training and testing accuracy is around 62%.

Both results from GaussianNB and DecisionTreeClassifier tells us that the best accuracy the model can predict is around 62%. This is not good enough to conclude that the trained model can confidently predict the major, with given data. Therefore, based on the results from this part, we can conclude that CS and non CS have no significant differences.

5. Conclusion

We conclude that in terms of salaries and participation, there is no significant difference between developer with CS background and non-CS background. Salaries and Stack Overflow participation are tended to increase as developers get more experienced in coding whether they have CS or non-cs backgrounds. However there is some difference in job satisfaction. Developers with CS background have very low career satisfaction during early years of coding experience but steadily increases while developers with non-CS background has relatively high career satisfaction but experiences up-and-down as time goes. To further observe such finding we use classification Based on the classifiers' model score the test performances are poor, considering they are around 50-60 percent. We conclude that it is hard to classify or differentiate between cs and non cs. Certainly we had limitations. A quantity data we had is not fully quantity that it was difficult to come out with very precise result. For example, for years of coding experience column, instead of value '12', '13', or '14', they are categorized as range such as '12-14 years'. Therefore we had to augment some fields which is not most ideal situation. If we had more time, we could have found a better dataset or include/eliminate parts of datasets which would make our result more clear and precise

6. Project Experience Summary

Steven Lee

The number of tasks I've completed in this project is the data preparations including the augmentations, different types of tests that aims to conclude whether the mean of CS and Non CS is same or not, then finally different types of classifiers. For the data preparation phase, I first specified our features — what we want and what we need. Then based on these features, I removed redundant and empty data that could become a hindrance in the future steps. I also augmented some fields' values, such as from "12-14 years" to "13" for precisor numeric comparison purposes. Moreover, I tried different types of tests to compare the means — ultimately used normality, levene, and t-test to see whether the p-value is qualified to conclude that the means are the same. Throughout these tests, I visualized the results by plotting various graphs, and saw intuitive similarities and differences between the two datasets. For the classifier, although I tried many other classifiers such as SVM, however based on the model score, GaussianNB and DecisionTreeClassifier were the optimal choices that fits under our interests. Before I begin, I labelled the values in binary form — 0 denoting non cs majors, and 1 denoting cs majors, for training purposes. After numerous trials of different fields, even the optimized scores were around 50-60 percent. At this point, I was able to conclude that that majors (CS and Non CS) do not have any differences in Salary, Career participation, etc throughout their career.

SeongJin Kim

In this project, I have completed data preparation to find significant differences in some aspects between their majors (CS vs non-CS) based on coding experiences. In the data preparation phase, data sets were downloaded as a csv file from 'Stackoverflow 2018 Developer Survey' and imported as a pandas dataframe for use in the project. Several unnecessary columns were removed from the dataset to maximize efficiency in terms of space and time. In addition, I have implemented the functions of drawing several graphs in approach 2 using the 'matplot.pyplot' library to visualize the salary and normality differences between computer science majors and non-computer science majors. These graphs help both readers and developers to have better and clear understanding on the relationship between the two groups (CS vs non-CS). During the project, I detected some code redundancies because all three approaches do similar things. Therefore, I implemented functions such as draw_line_graph(), draw_histography(), print_p_values() to refactor some code. By implementing these functions, all approaches can use the same functions with different parameters based on their target data. In the classification phase, we have used several classification techniques like GaussianNB(), DecisionTreeClassifier(), and etc.. To visualize the scores, I implemented a feature that shows the table of DecisionTreeClassifier() optimized sores based on their depth by adding one simple for-loop. The table shows that the optimized scores were about 0.5 to 0.6 which is not sufficient. Therefore, we were able to conclude that there are no relationships between participants' majors and salaries throughout their careers.

Darren Yang

First I found the data from Kaggle and verified possible research we can conduct with the professor through email. And then, I downloaded csv files, I started to prepare for the data. First I had to eliminate some unnecessary part of the data and clean because datasets are too large with unnecessary columns I did not need. My part was to analyze the relationship between career-satisfaction and years of coding experience and find difference between developers with cs major background and developers with non-cs major. First, I augmented some values where each value is a range (e.g. '4-6' years to 5) so that we can graph more easily. Testing normality test requires more step for career satisfaction. Unlike the other two analysis, p-value of normality test for datasets of relationship between career satisfaction and years of coding was less than 0.05 which is not what we expected. Definitely I could see that the dataset was left-skewed. Therefore, I tried to transform the data by x -> e^x or x -> x^n where I have tried different value for n in order to make p-value 0.05 as close as possible. But since our dataset is large, we could have assumed that data is normal. After each of our group came out with the graph, we found redundant code so we refactored the code. Using DecisionTreeClassifier() for classification technique where the score came out to be around 0.6. As how the graph and the score from DecisionTreeClassifier() shows, there is no effect on career satisfaction, salary, and Stack Overflow participation with years of coding.