Stack Overflow Survey Analysis

Analysis of Stack Overflow survey data from multiple years to gain insights and recommendations based on various attributes.

CMPE 256- Large Scale Analytics

Spring 2019

Project Report

Professor

Dr. Gheorghi Guzun <gheorghi.guzun@sjsu.edu>

Team 14: Aztech

Archana Yadawa <ayadawa@gmail.com>

Nishtha Atrey <nishtha.atrey@sjsu.edu>

Tina Aggarwal <tina.aggarwal@sjsu.edu>

SAN JOSÉ STATE UNIVERSITY

Project Description:

Several thousands developers each year fill out a 30-min survey on the Stack Overflow website. This survey collects data about Stack Overflow users using various attributes. In this project, survey data from multiple years has been used to perform analysis and gain insights as well as recommendations with regard to multiple attributes, some of them being demographics, programming languages, coding experience, job satisfaction among developers. Post the analysis we have derived certain answers and generated recommendations.

Datasets:

We will be using survey analysis datasets from 2018-2011.
Link: https://insights.stackoverflow.com/survey

2011: 2815 rows X 65 columns
2012: 6245 rows X 74 columns
2013: 9744 rows X 129 columns
2014: 7644 rows X 124 columns
2015: 26088 rows X 222 columns
2016: 56031 rows X 71 columns
2017: 51393 rows X 154 columns
2018: 9890000 rows X 129 columns
Total: 33,534,196

Github Link:
https://github.com/tinaaggarwal/CMPE256_Project_Aztech

Data Pre-processing approaches

1. **Principal Component Analysis:** Used for feature extraction and dimensionality reduction

2. **Dummy Variables:** Convert categorical variable into dummy/indicator variables

3. **Handling Missing Values:** Normalized NaN values

4. **Numerical categorization:** Some inputted values are strings. So, we converted them into integer values by putting each input into different numeric categories.

Approaches

Approach 1: Prediction using random forest

Random forest is a supervised learning algorithm that uses many decision trees classifiers. It builds multiple decision trees and merges them together in order to obtain a more accurate and stable. An advantage of using random forest is the fact that it can be used for both classification and regression problems. Another advantage is random forest does not cause overfitting. This is because of the the many trees in the forest. For this project in particular we have used random forest classifier to train our model and predict how many students code as a hobby. We have used other factors, like the student's age, to prevent overfitting. One disadvantage of using random forest is the speed. Since many decision trees are created and used, this can slow the algorithm down. To improve the speed of this algorithm we have used SVM.

Prediction using random forest and Prediction using SVM
For the preprocessing steps, we first created a dataframe with the wanted features. Then, we did numerical categorization on the string values for our selected features. Next, we normalized NaN values.

Approach 2: Prediction using SVM

Support Vector Machine algorithm finds a hyperplane in a N-dimensional space that distinctly classifies the data points. N represents that number of features. Hyperplane that has the maximum number of margins are ideal for SVM models. We have used the same pre-processing steps as approach 1.

Approach 3: Prediction using Linear regression
Linear Regression has been used to predict the average value of Y given an X using a straight line. We have using linear regression to predict how many years of coding experience users have given their development type and age.

For the preprocessing steps, we first created a dataframe with the wanted features. Then, we did numerical categorization on the string values for our selected features. We had an additional step that uses one-hot encoding to create a matrix for the string output values. We have also used Principal Component Analysis (PCA) for feature extraction and dimensionality reduction. Both X and Y dimensions must be the same in order to find a best fit line using linear regression.

Approach 4: Recommender system KNN
We have also used KNN to create a recommender system to recommend users that match recruiters' request. To achieve this we did a simple check to see which existing node is the closest to the artificial one that was created for the features a specific recruiter is seeking. KNN graph was computed using sklearn's neighbors_graph method to generate the graph for k-neighbors for points in X.

Evaluation of Prototypes

Approach 1: Prediction using random forest

Accuracy Score = 0.821447928765002

Mean Absolute Error = 0.17855207123499806

Mean Squared Error = 0.17855207123499806

Classification Report

	precision	recall	f1-score	support
0	0.82	1.00	0.90	10609
1	0.00	0.00	0.00	2306
micro avg	0.82	0.82	0.82	12915
macro avg	0.41	0.50	0.45	12915
weighted avg	0.67	0.82	0.74	12915

Approach 2: Prediction using SVM

Accuracy Score = 0.8212156407278358

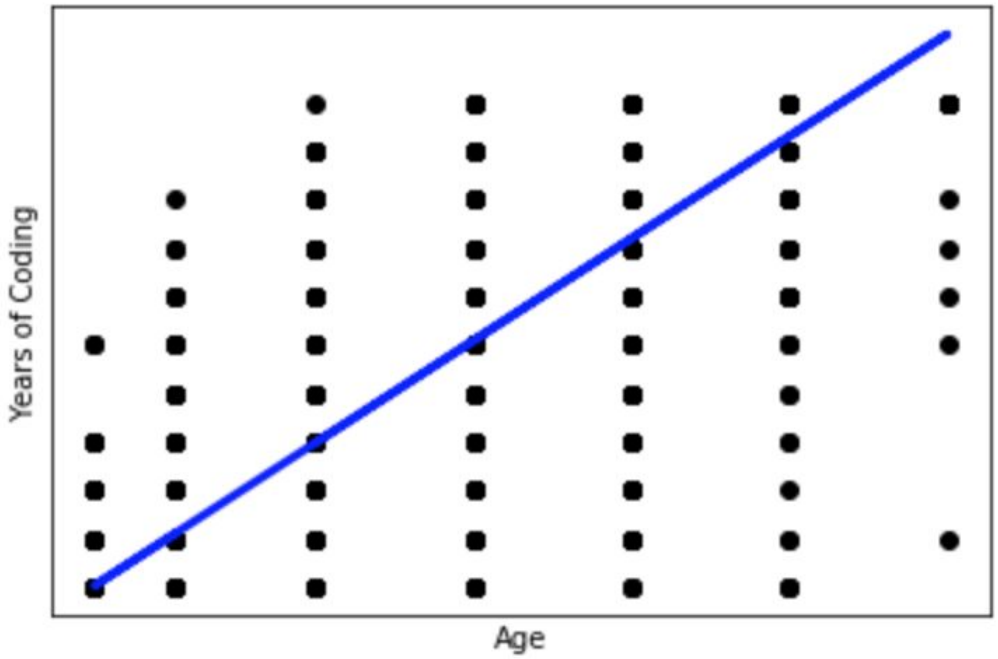
Mean Squared Error = 0.17878435927216416

Classification Report

	precision	recall	f1-score	support
0	0.82	1.00	0.90	10606
1	0.00	0.00	0.00	2309
micro avg	0.82	0.82	0.82	12915
macro avg	0.41	0.50	0.45	12915
weighted avg	0.67	0.82	0.74	12915

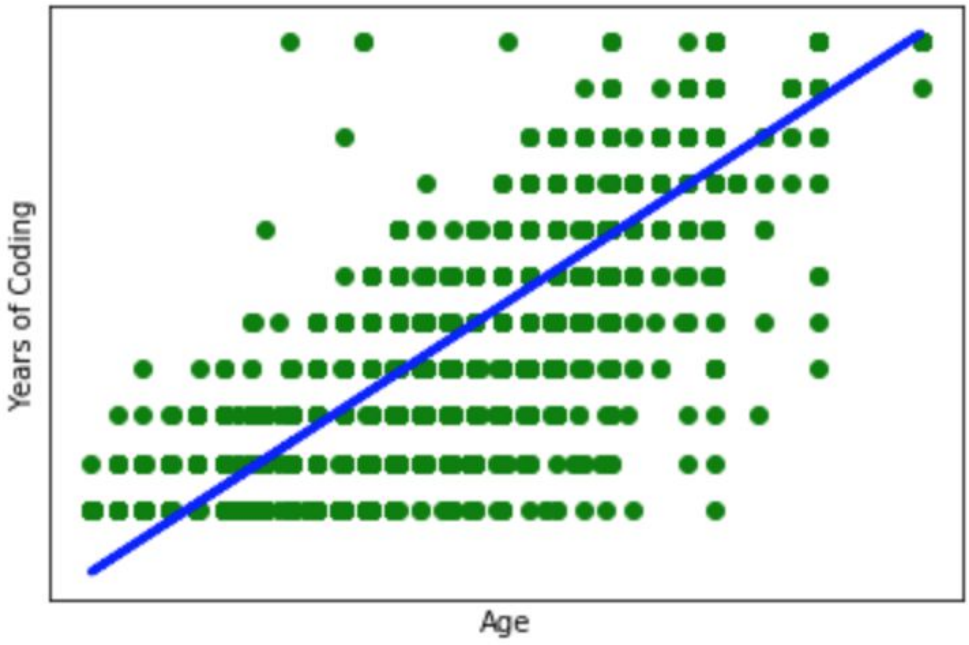
Approach 3: Prediction using Linear regression

Without optimization:
Coefficients = 0.6329504
Mean Squared Error = 27.13
Variance Score = 0.54



With Optimization by adding Features to fit the data better:
R-Squared value: 0.81006743759681

Using PCA:
Coefficients = 0.57236587
Mean Squared Error = 10.17
Variance Score = 0.78



Results Derived Post Analysis and Related Charts

1. Career Satisfaction Levels

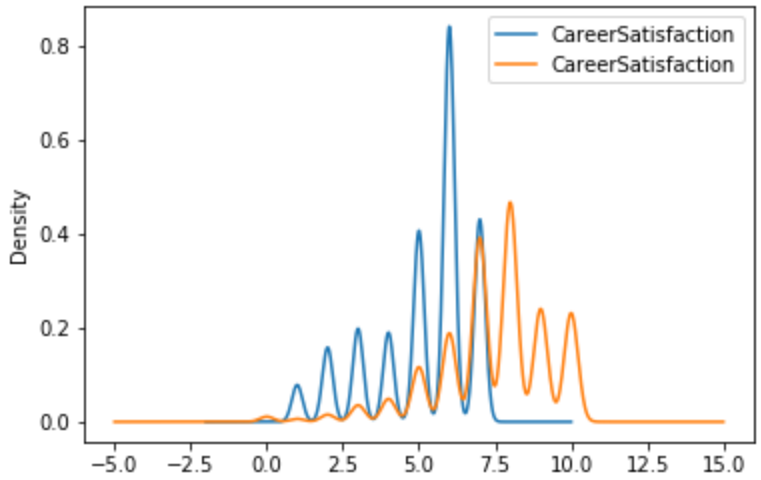


Figure 1: The above plot signifies that overall Career Satisfaction increased among people from 2017 to 2018

2. Gender comparison against DS and Non DS candidates over the course of 2 years (2017-2018)

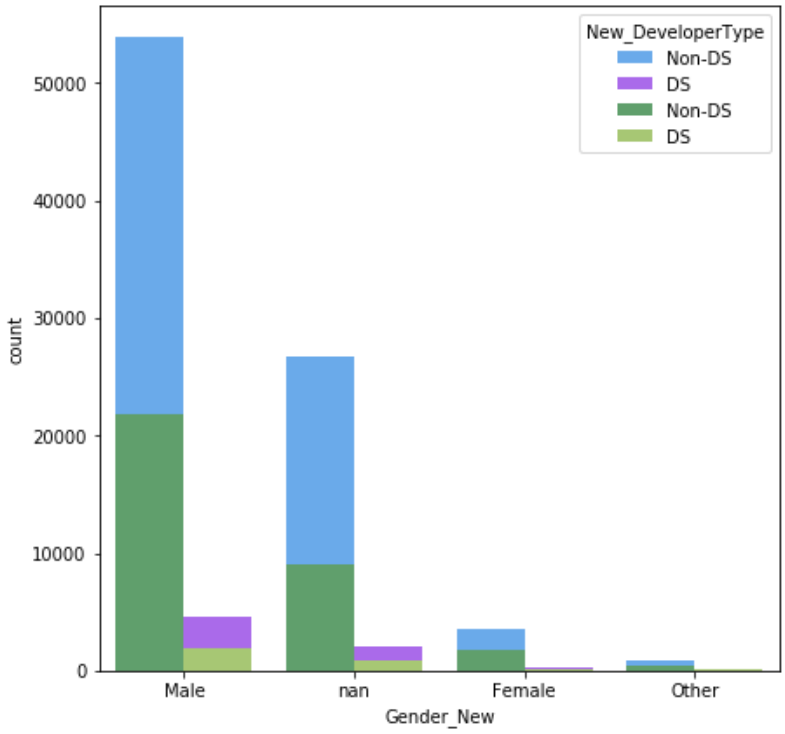


Figure 2: The above countplot signifies that there were far more Non-DS males in 2018 as compared to Non-DS males in 2017. Also the overall Count of developers almost doubled between 2017 and 2018

3. Questionnaire and response analysis

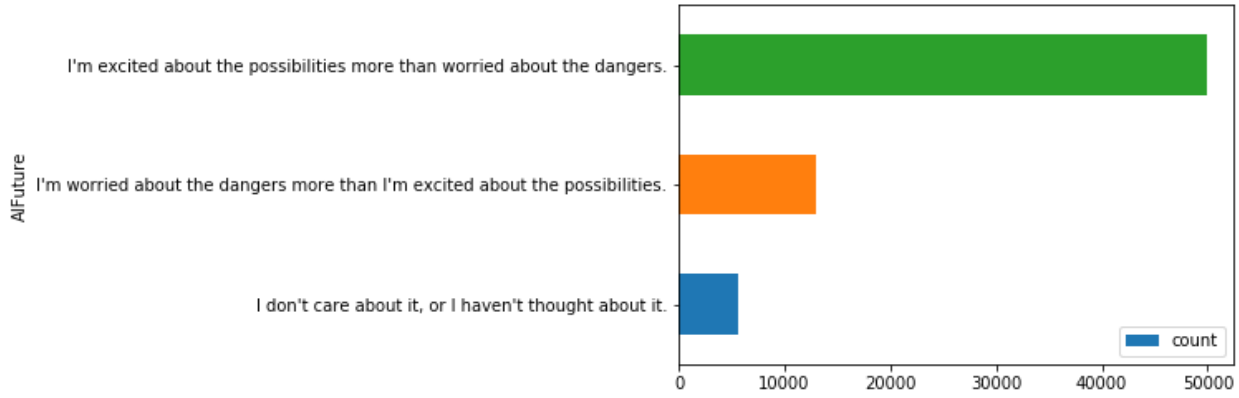


Figure 3: This chart shows the top responses of the question “Do you think AI is the future”, that was provided by the questionnaire

4. Questionnaire and response analysis

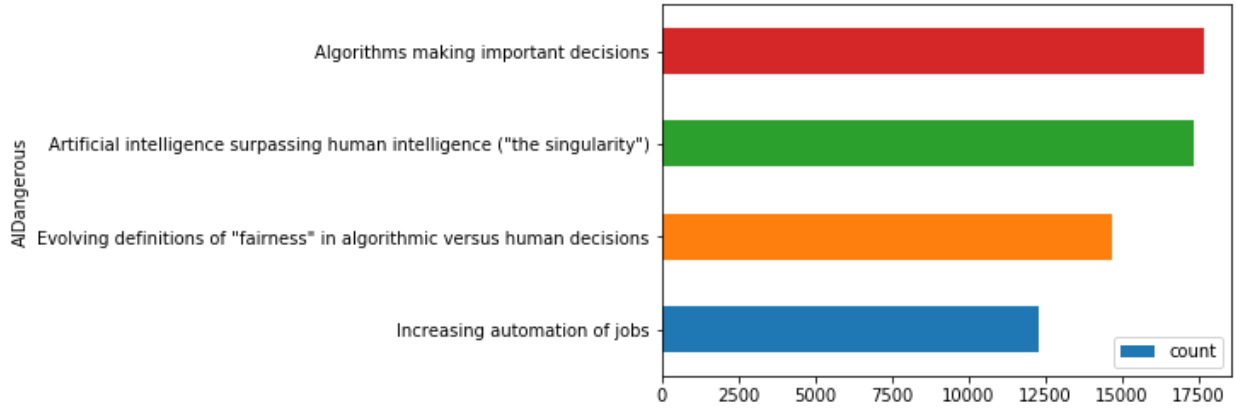


Figure 4: This chart shows the top responses of the question “Do you think AI is dangerous”, that was provided by the questionnaire

5. Questionnaire and response analysis

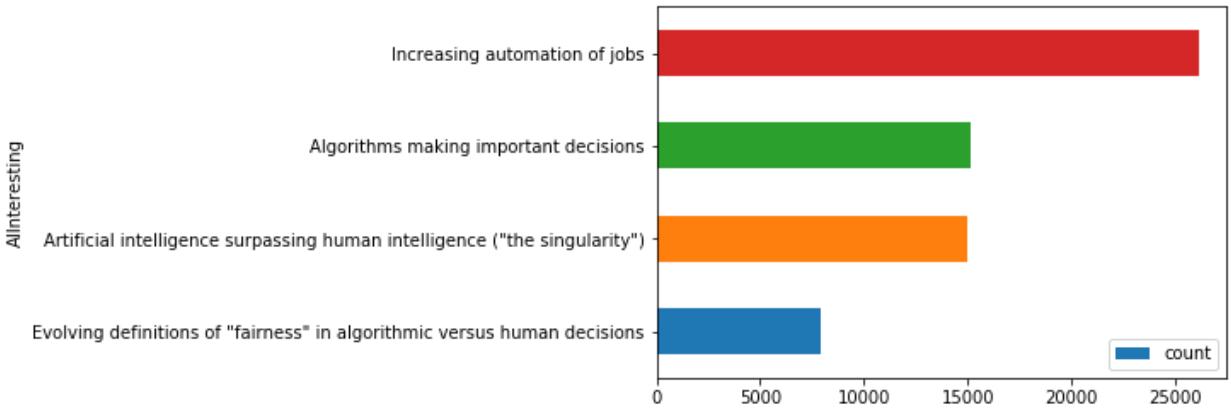


Figure 5: This chart shows the top responses of the question “Do you think AI is interesting”, that was provided by the questionnaire

6. Gender analysis of DS and Non DS

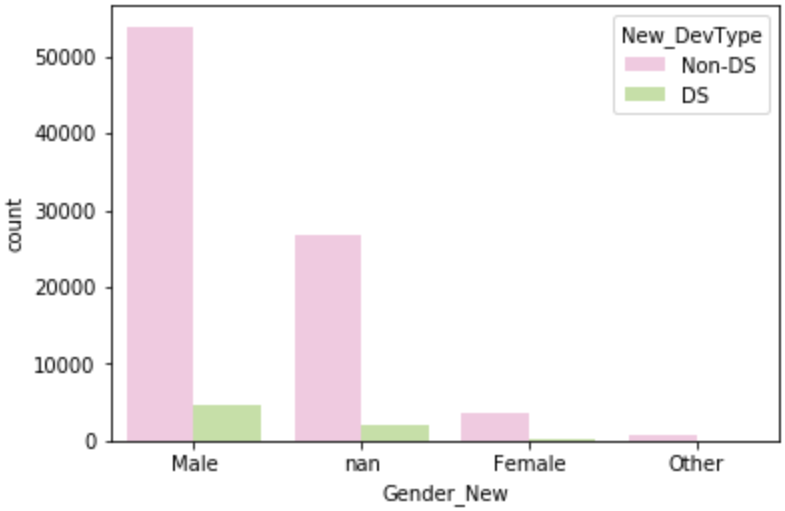


Figure 6: The above bar chart shows the comparison of Non Data Scientists and Data Scientists based on their gender

7. Analysis of job satisfaction of users

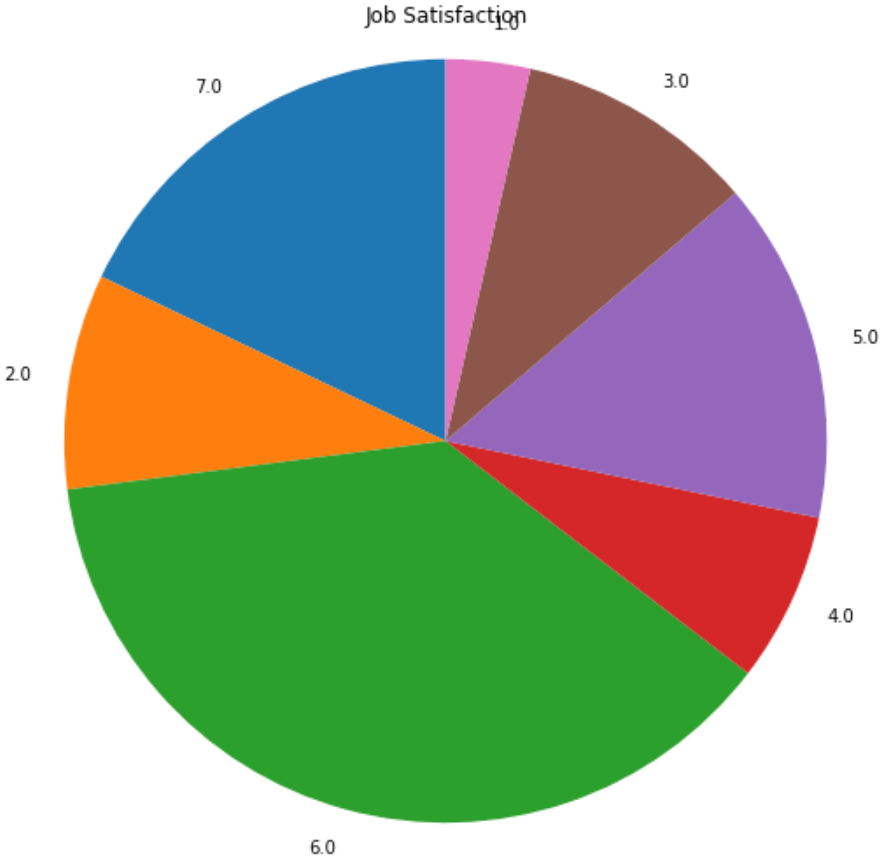


Figure 7: The above pie chart shows the job satisfaction rating for all users