Stack Overflow: Developer Survey Analysis

Team 14 - Aztech:

Archana Yadawa
Nishtha Atrey
Tina Aggarwal

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.

Data Pre-processing

Data Pre-processing

Principal Component Analysis (PCA)

Used for feature extraction and dimensionality reduction

Dummy Variables

Convert categorical variable into dummy/indicator variables

Data Pre-processing

Missing Value Handling

Normalized NaN values by replacing them with mean, median or dropping them as needed for the required analysis.

Numerical Categorization

Some inputted values are strings. So, they were converted into integer values by putting each input into different numeric categories.

Approaches

Approach 1: Prediction using Random Forest algorithm

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.

Approach 2: Prediction using Support Vector Machine (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

Approach 4: Recommender System using K-Nearest Neighbor algorithm (KNN)

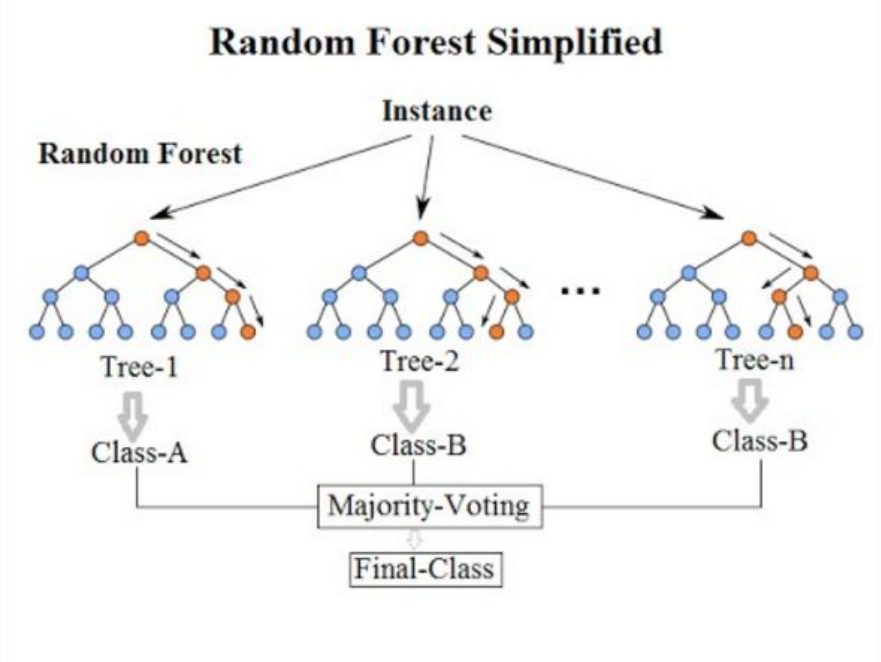
We have also used KNN to create a recommender system to recommend users that match recruiters' request.

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.

Prototype Evaluation

Approach Evaluation



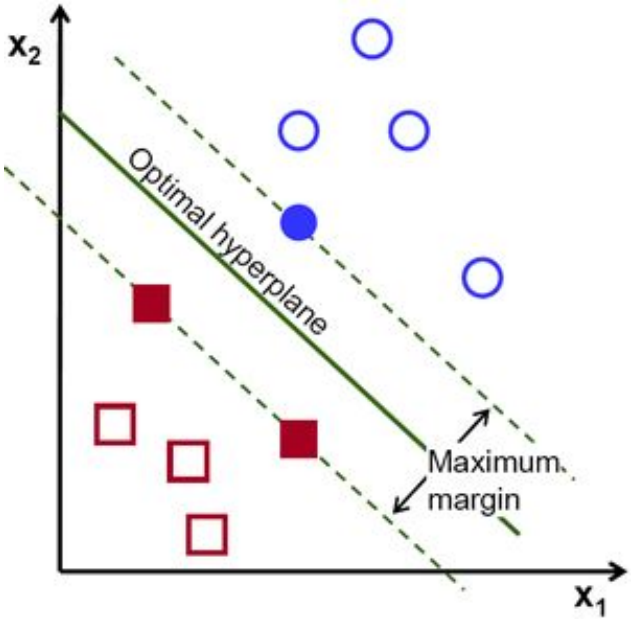
Approach 1: Prediction using random forest

Accuracy Score = 0.821447928765002

Mean Absolute Error = 0.17855207123499806

Mean Squared Error = 0.17855207123499806

Approach Evaluation

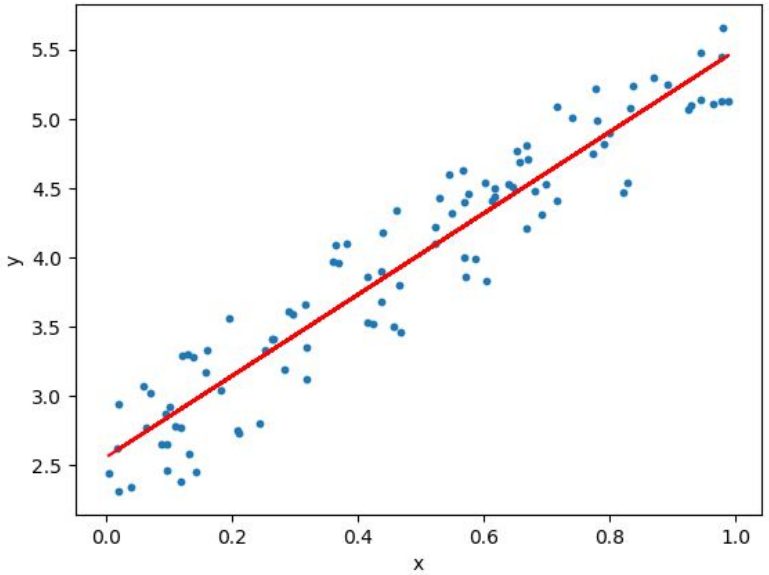


Approach 2: Prediction using SVM

Accuracy Score = 0.8212156407278358

Mean Squared Error = 0.17878435927216416

Approach Evaluation



Approach 3: Prediction using Linear regression

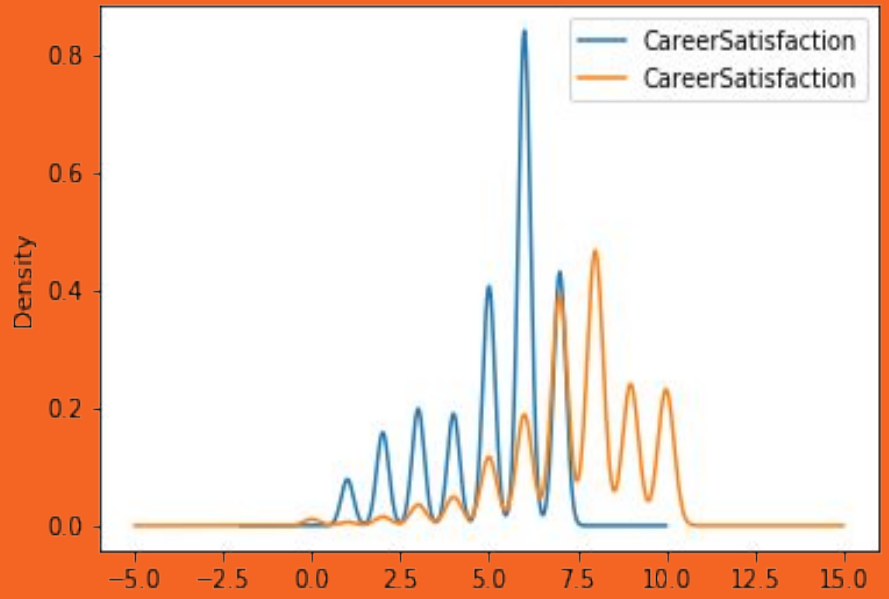
-Without PCA:
Coefficients = 0.6329504
Mean Squared Error = 27.13
Variance Score = 0.54

-With PCA
Coefficients = 0.57236587
Mean Squared Error = 10.17
Variance Score = 0.78

Answers derived post dataset analysis

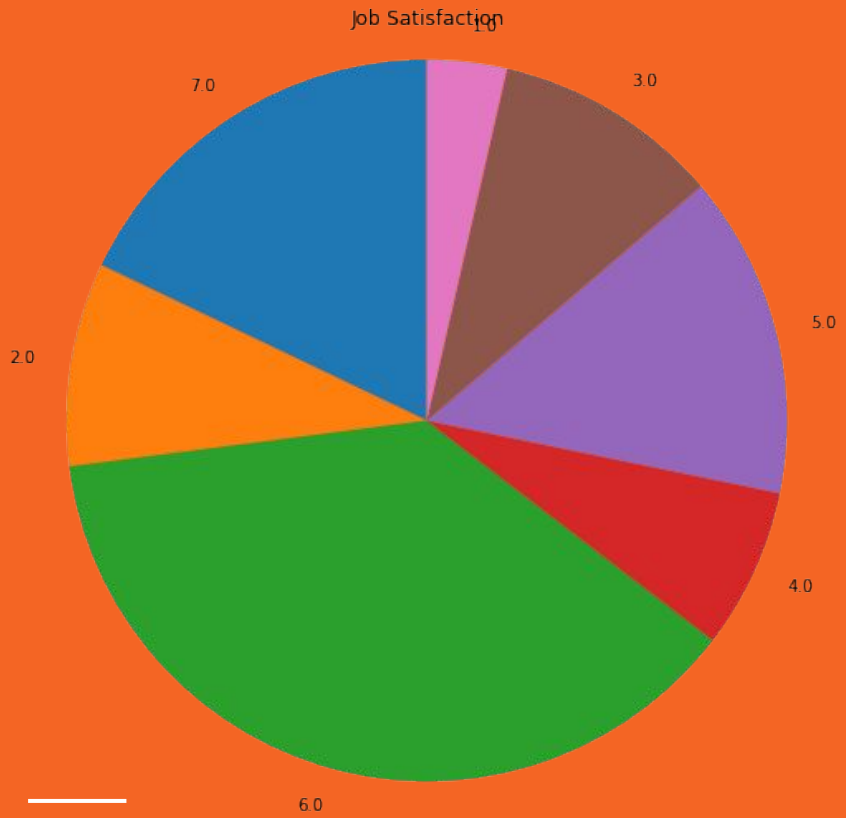
Career Satisfaction Levels

The achieved plot signifies that overall Career Satisfaction increased among people from 2017 to 2018



Job Satisfaction Levels

The pie chart shows the job satisfaction levels among all users, ratings range from 1 to 7



What developers think about AI?

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

~ Thank you ~