Capstone Project

Final Project

GitHub Link:https://github.com/akashsky1994/stack-overflow-developer-survey-analysis/

Project By:

Aakash Mishra (am11533)
Mohit Kewlani(mmk9369)
Kewal Jani (kj2062)

1. Data Set Description

First, we looked for many datasets around and took the data set about the stack overflow website survey from Kaggle. Stack overflow is the website that is used by most of the students and by many Employers. It is one of the best websites created for programmers (Everyone should agree on this, or we can do hypothesis testing).

Data Set: https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey

This Data set is a type of survey in which users are asked 130 different questions and the responses are recorded for individual user. This dataset is collected for around 100K users. The rows state the number of users, and the columns has the questions asked to the user. These questions can be found in survey_results_schema.csv file and the data in stackoverflow_data.csv file.

We were interested in this data set because it states various important features of a programmer. The data set contains information about salary, if the user used to code before his current job, weather he is satisfied with his current job, etc. The dataset will not only help us to get facts but also to get information about where we will be standing in the future.

The Challenge:

The challenge we faced with the dataset was, the data was not cleaned and there were many columns in which we had to apply cleaning and One Hot Encoding/Label Encoding and generate more columns. Also, as the datatype for certain columns were in characters and string format. we need to classify data column wise and make the data usable. We converted the data in numeric format, and it was ready to use.

After the data cleaning part, we had to handle the NA values. We handled the NA values row wise, column wise or fill forwarding, according to the requirement of the question that we were supposed to answer

We decided to answer all the interesting questions that we found and made sure that we implement all the algorithms and techniques that we learned in the Intro to Data science course for getting the answers.

1 Hypothesis Testing: Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. ... Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process.

2 Dimensionality reduction: Reduction in the dimensionality of the question asked to the user for survey. For example, which question contribute more and provide more information

3. **Clustering**: After PCA, we decided to apply Kmeans Clustering to the transformed data and find out how many types of clusters are present in our dataset, and further use that to perform classification.

4 Machine learning: Splitting the data into training and testing set and ensuring cross validation, and after that, performing each example of Regression and Classification, using different models, to compare their accuracy.

5.**Summary And EDA**: Our dataset was huge, and we got the opportunity to extract many meaningful information from the data, which we have demonstrated in this section. There were many interesting conclusions that we can draw from the plots in this section.

1. Hypothesis Testing

As our dataset is too large, we had couple of interesting Hypothesis to run.

Here are the Hypothesis Test we conducted,

1) **Null Hypothesis is Career Satisfaction remains same at all Age Level**

   We decided to perform **Man Whitney U test**, yielding a **p value of 0.09**, which is not less than 0.05, **fail to reject our Null Hypothesis**. Hence, we do not have enough evidence that Career Satisfaction remains same at all Age Level, at 5% significance level.

2) **Null Hypothesis is male and female developers are paid equally.**

   We decided to perform **independent T test**, yielding a **p value of 0.04**, which is less than 0.05, **rejecting our Null Hypothesis**. Hence, we have enough evidence that male and female developers are not paid equally, at 5% significance level.

3) Null Hypothesis is Developers who are Self Taught have no inferior complex compared to people who had traditional degree.

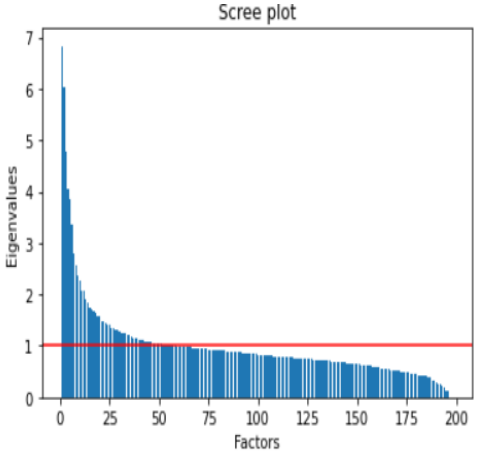
   We decided to **Chi Squared Test,** yielding a **p value of 1.4952279461561637e-20**, which is less than 0.05, **rejecting our Null Hypothesis**. Hence, we have enough evidence that Developers who are Self Taught have inferior complex compared to people who had traditional degree, at 5% significance level.

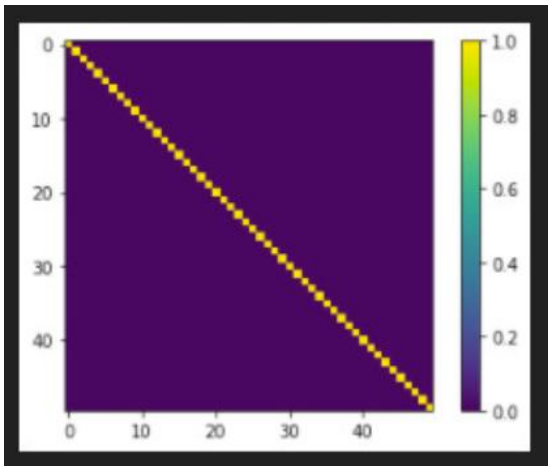
2. Dimensionality reduction

Is dimensionality reduction a good choice for prediction?

Dimensionality Reduction. We had a data set of arounds 100K users and We thought that if would be great if we are able to reduce the dimension of the data by removing the data that are corelated to each other and keep those data that contribute to the Dataset.

There were certain questions which had non-numeric data, so it was not possible to apply PCA. For this problem we converted all the columns to numeric values after converting the data we applied Z-score to the data set so that the data gets normalized. After the data was normalized, we fetched the eigen values and applied the kaiser criteria to get the columns which were contributing more to the data set. As a result, from a set of 200 columns we detected 49 columns that had eigenValues>1.as seen in the plot below, hence contributing in a significant manner. We plotted the correlation matrix, between the questions and got the result as below which shows that there is no correlation among the data (as seen in the plot below).





The index of the columns which we found contributed most to the data were the following.

```
([152,  35,   3,  30,  51,  49,  23, 153, 192,  22,  23,  13,  73,
  169,  26, 154,  26, 134,  98,  97,   8, 134,  27,  63,  92,  11,
  148,   5, 126,  60,  19, 179, 190,  12,  11,   5, 129,  58,  12,
  183, 191,   6,  58, 125, 133,   6,   7])
```

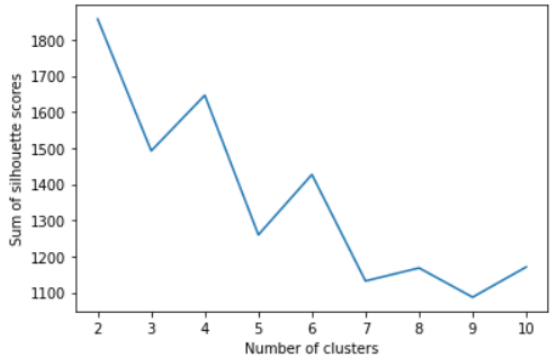
Later we use this column, to perform Kmeans clustering and with the help of silhouette method, we find the optimal clusters for our dataset.

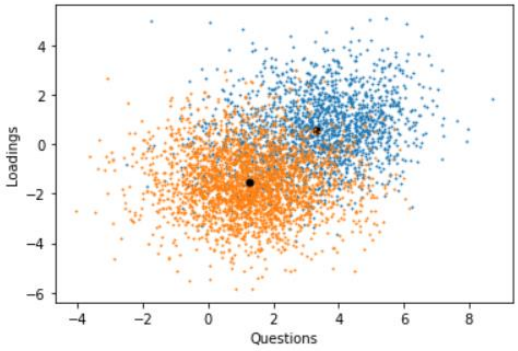
NYU

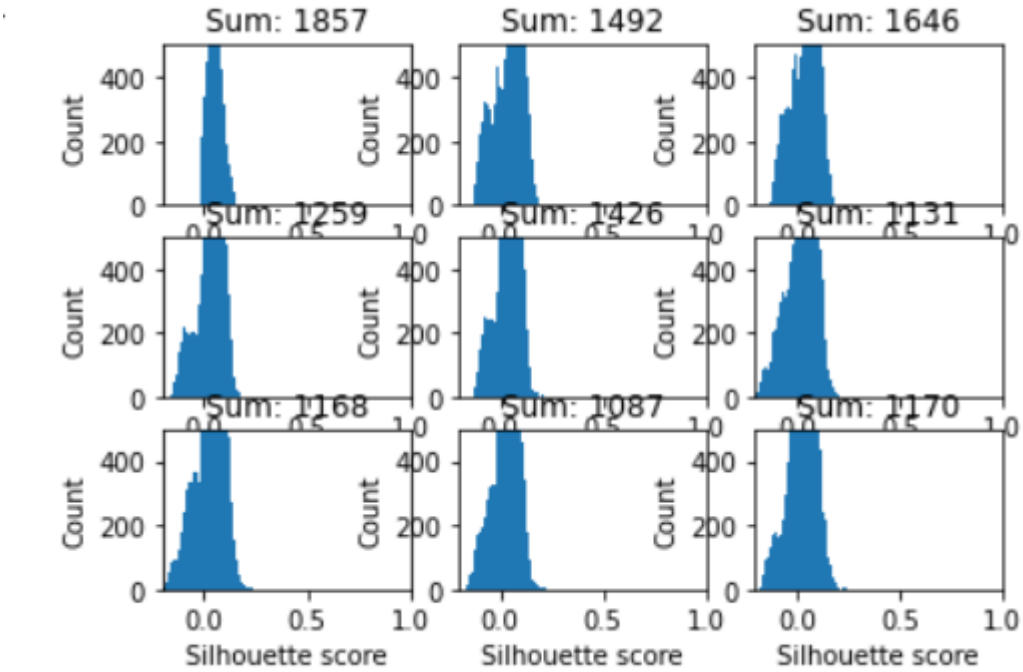
3.Clustering

After getting optimum features from the PCA we got column that shows 22% of the variance. And we decided to apply k-means clustering to visualize the data.

We thought that K-means would be a good model to predict the cluster in which the person belongs. We trained the model by taking all the features. To obtain the clusters we ran silhouette score. We found out that the optimum cluster we will need is 2 as shown in figure.







The result for the clusters we got were the following `0 : 24201 1 : 13067. This implies that 24201 people were in one cluster and 13607 were in another. Making sure our dataset is not imbalanced. Hence, making it great to run Classification and Regression Models.`

NYU

4.Prediction (Machine Learning)

Classification

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modelling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

Question: prediction of the Job Satisfaction based on all the transformed features (which model is accurate)?

To predict the Job Satisfaction. We trained dataset using transformed independent variables and kept the converted salary as dependent variable. We divided the dataset in 50:50 ratio. 50% data for training and 50% data for testing, ensuring cross validation. We applied different models to predict the Job Satisfaction and found the accuracy of models as follow.

We applied Random Forest classification, SVR and Logistic Regression. Here are their accuracies.

SVR:
linear 0.909305570462595
rbf 0.9046903509713428

Logistic Regression:
0.9068369646882044

Random forest:
0.8592894708597187

Regression

Regression models predict a value of the Y variable given known values of the X variables. Prediction within the range of values in the dataset used for model-fitting is known informally as interpolation. Prediction outside this range of the data is known as extrapolation. Here, we are trying to predict Salary of a developer using all the features of dataset.
Here are all the Regressor Models we performed,

Question: prediction of the Salary based on all the features of dataset (which model is accurate)?

To predict the Salary. We trained dataset using all independent variables of dataset and kept the converted salary as dependent variable. We divided the dataset in 50:50 ratio. 50% data for training and 50% data for testing, ensuring cross validation. We applied different models to predict the Salary Prediction and found the accuracy of models as follow.

We applied Random Forest Regressor, Ridge and Lasso Regression. Here are their accuracies.

Ridge Regression:
```
testing error 4.879455374538774e-12
```

Lasso: Accuracy:
```
0.9999999804995144
```

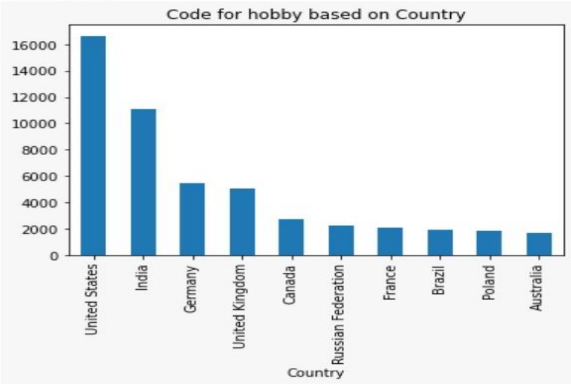
Random forest:

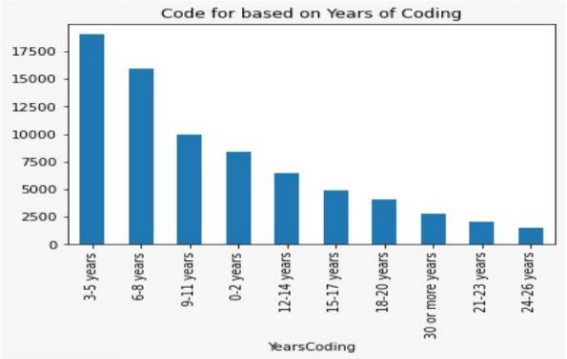
```
0.9999897089623649
```

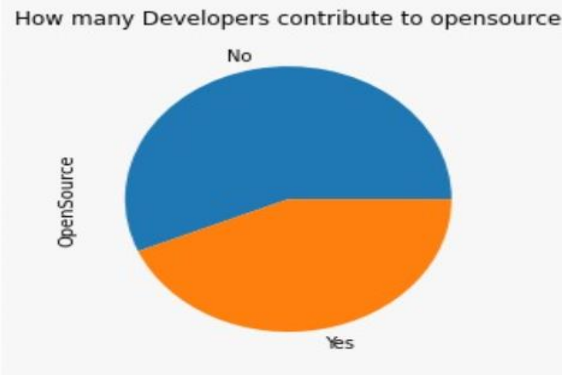
5.Summary and EDA.

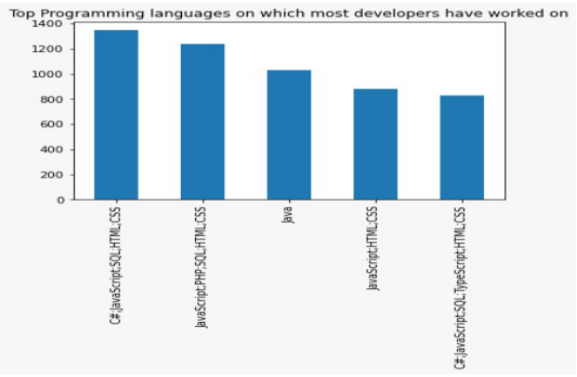
This project we tried to Implement all the aspects of the course and we can say that we were happy with the results. The most challenging part was to clean the data as the data as not cleaned and convert the data to a usable numeric format. We got some atoning results from Hypothesis testing. We also found that PCA is a good example to reduce the data, but it is better to predict the data by including other features as well because there are indirectly corelated to each other and boost the prediction score drastically compared to the features used by PCA. Also, we can decide whether we should use PCA or not depending on the type of model we will be requiring. We also tried to show both Classification and Regression model prediction.

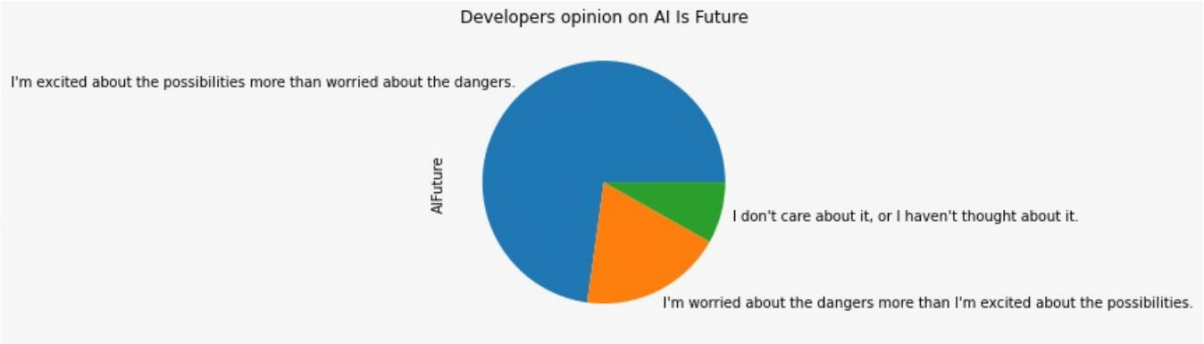
EDA


Code for hobby based on Country
16000
14000
12000
10000
8000
6000
4000
2000
0
United States
India
Germany
United Kingdom
Canada
Russian Federation
France
Brazil
Poland
Australia
Country


Code for based on Years of Coding
17500
15000
12500
10000
7500
5000
2500
0
3-5 years
6-8 years
9-11 years
0-2 years
12-14 years
15-17 years
18-20 years
30 or more years
21-23 years
24-26 years
YearsCoding


How many Developers contribute to opensource
No
OpenSource
Yes


Top Programming languages on which most developers have worked on
1400
1200
1000
800
600
400
200
0
C#;JavaScript;SQL;HTML;CSS
JavaScript;PHP;SQL;HTML;CSS
Java
JavaScript;HTML;CSS
C#;JavaScript;SQL;TypeScript;HTML;CSS


Developers opinion on AI Is Future
I'm excited about the possibilities more than worried about the dangers.
AIFuture
I don't care about it, or I haven't thought about it.
I'm worried about the dangers more than I'm excited about the possibilities.