Original Case Study Narrative
Kevin Paulovici
DSC 550
May 3, 2020

Case Study: Analyze Stack Overflow Developer Survey Data

Data File

Dataset contains 2019 survey results from Stack Overflow developers.

survey_results_public.csv
survey_results_schema.csv

Data location: https://insights.stackoverflow.com/survey

Overview

This analysis will look at the 2019 survey results from Stack Overflow. Throughout this analysis we will be looking for what aspects developers have. We'll try to answer some questions such as, which gender developer tend to be, do they code for fun, is there a linear trend in years of development vs. pay. Do they participate in open source projects, and how their profession is related to programming (e.g., full time, student, hobby, etc.).

Step-by-Step Instructions

Part 1 Includes step 1 – 13

Part 2 Includes step 14 -16

Part 3 Includes step 17 -20

1. Load the data from *survey_results_public.csv* and *survey_results_schema.csv* files into a DataFrame.
    a. *survey_results_schema.csv* contains the header information, so we'll want to include this to check headers later.
2. Display the dimensions of the dataset.
    a. There are a lot of columns that are not going to be directly applicable to this analysis, we'll remove them later in Part 2.

```
The dimension of the data is: (88883, 85)
```

3. Filter data to only include United States
    a. I'm limiting the survey results to be more applicable to me. Through this analysis, I would like to see what attributes others developers have and more directly compare them to myself.
4. Display the dimensions of the file.
    a. From the two filtering criteria we set in step 3, we were able to reduce the dataset to ~ 21,000 rows – this should help us more directly focus in features we care about.

```
The dimension of the data is: (20949, 85)
```

Original Case Study Narrative
Kevin Paulovici
DSC 550
May 3, 2020

5. Display the first few rows of the new dataset and header schema.

	Respondent	MainBranch	Hobbyist	OpenSourcer	OpenSource	Employment	Country	Student
3	4	I am a developer by profession	No	Never	The quality of OSS and closed source software ...	Employed full-time	United States	No
12	13	I am a developer by profession	Yes	Less than once a month but more than once per ...	OSS is, on average, of HIGHER quality than pro...	Employed full-time	United States	No
21	22	I am a developer by profession	Yes	Less than once per year	OSS is, on average, of HIGHER quality than pro...	Employed full-time	United States	No
22	23	I am a developer by profession	Yes	Less than once per year	The quality of OSS and closed source software ...	Employed full-time	United States	No
25	26	I am a developer by profession	Yes	Less than once per year	The quality of OSS and closed source software ...	Employed full-time	United States	No

	QuestionText
Column	
Age	What is your age (in years)? If you prefer not...
Age1stCode	At what age did you write your first line of c...
BetterLife	Do you think people born today will have a bet...
BlockchainIs	Blockchain / cryptocurrency technology is prim...
BlockchainOrg	How is your organization thinking about or imp...
...	...
WorkPlan	How structured or planned is your work?
WorkRemote	How often do you work remotely?
WorkWeekHrs	On average, how many hours per week do you work?
YearsCode	Including any education, how many years have y...
YearsCodePro	How many years have you coded professionally (...

6. Replace values for select numerical features (YearsCode & YearsCodePro)
    a. Some options allow for more than 50 years or less than 1 years as an answer. For plotting, I'm lumping the less than 1 year as 0, and more than 50 with 50.
7. Convert select numerical features (YearsCode & YearsCodePro) to numerics
8. Display the summary of the data for features of interest (Age, YearsCode, YearsCodePro).
    a. Age – we can see non-realistic ages like 1 and 99, in Part 2 we'll narrow the scope of the age.
    b. This also gives an idea of the experience level of the developers.

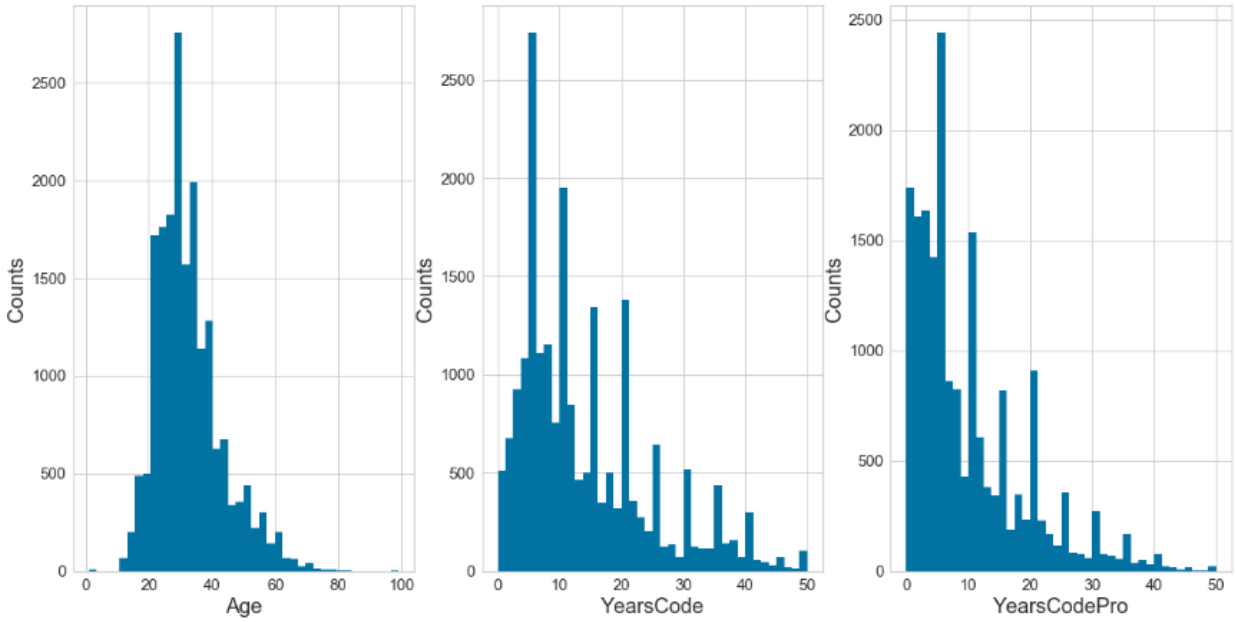
```
count    18864.000000
mean        32.753281
std         10.495166
min          1.000000
25%         25.000000
50%         31.000000
75%         38.000000
max         99.000000
Name: Age, dtype: float64
```

```
count    20790.000000
mean        13.970996
std         10.481683
min          0.000000
25%          6.000000
50%         10.000000
75%         20.000000
max         50.000000
Name: YearsCode, dtype: float64
```

```
count    18359.000000
mean         9.915845
std          9.002617
min          0.000000
25%          3.000000
50%          7.000000
75%         14.000000
max         50.000000
Name: YearsCodePro, dtype: float64
```

Original Case Study Narrative
Kevin Paulovici
DSC 550
May 3, 2020

9. Plot histogram for features (Age, YearsCode & YearsCodePro).
    a. We can see by limiting the age range we will still be cover the majority of participants.

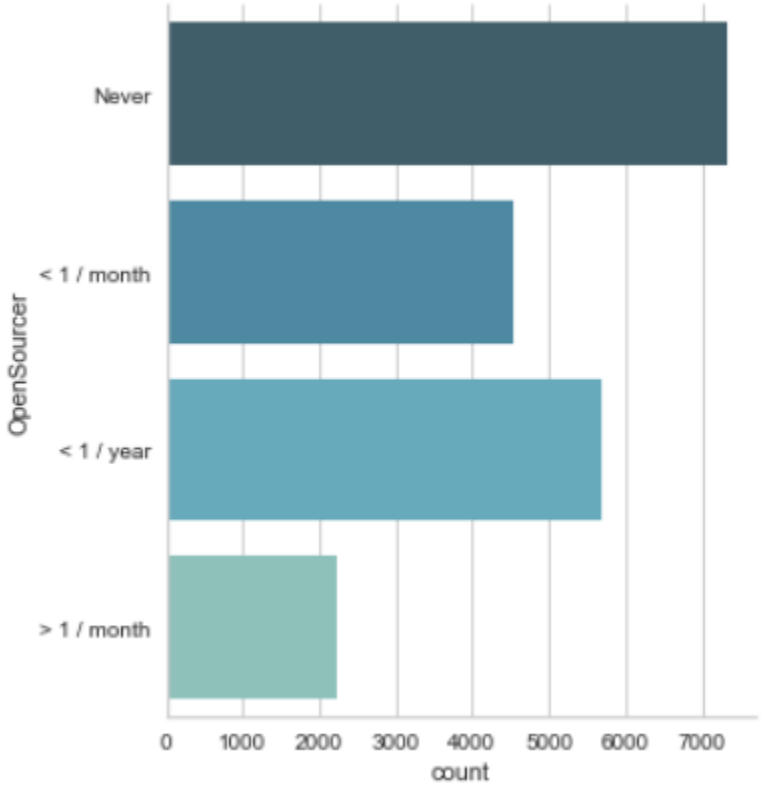


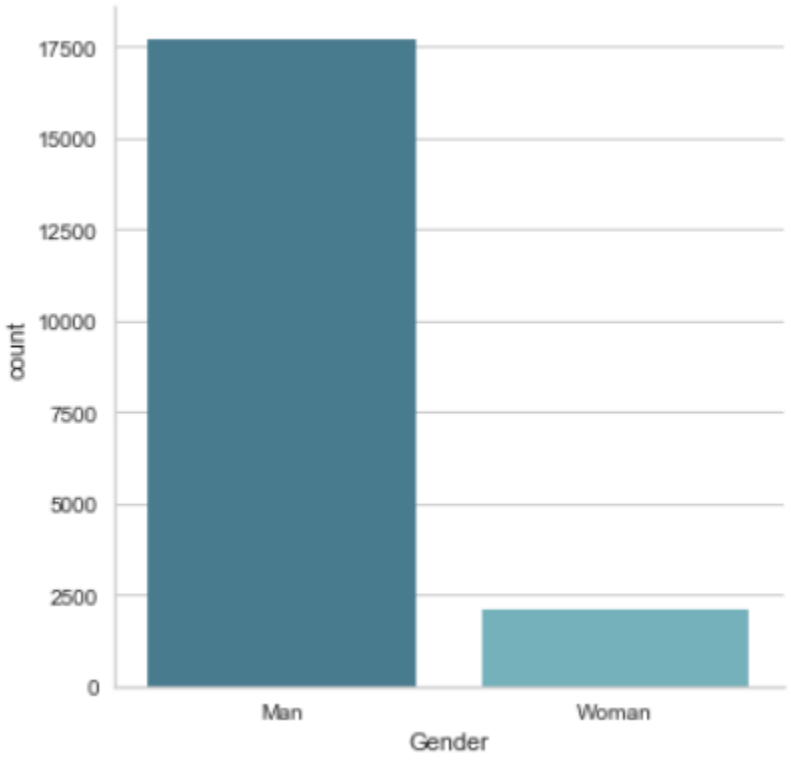
10. Replace values for OpenSourcer, the original survey has long text, to display better on the plot this should be converted to shorthand. Also replace text for Employment for similar reason.
11. Filter genders to Man & Woman; other genders are relatively small.

```
The dimension of the data is: (19792, 85)
```

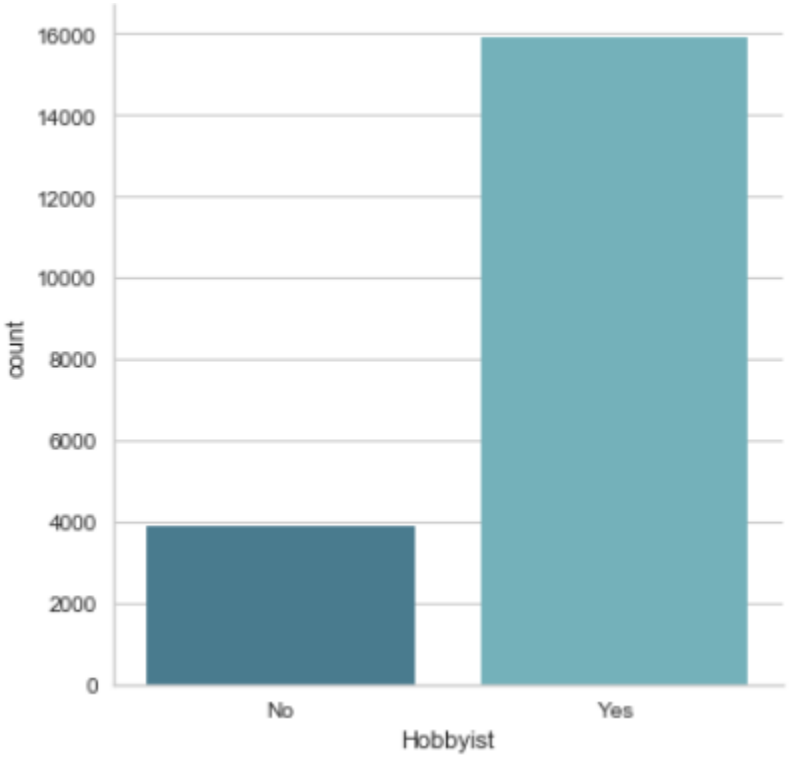
12. Make bar chart for other features of interest.
    a. This data enables us to better understand the distribution of various features. We can latter use these as filters and compare to salary. Additionally, we can use them for machine learning to see which are most important when considering salary.
    b. Open Source providers – the majority of developers rarely is at contribute.
    c. Gender – Overwhelming male dominated, this was expected since STEM fields follow the same trend. It will definitely be interesting to compare salary's here when holding older features constant.
    d. Hobbyist – the majority of developers program for fun. It is surprising so few contribute regularly to open source programs when the majority like to program for fun.
    e. Students – only a small sample are part-time or full-time students.
    f. MgrIdiot- This feature is for the confidence developers have in their managers. It looks like most have somewhat or are very confident in their managers.
    g. Employment status – most respondents are full-time employed. Unemployed respondents will be removed since they skew the data on salary. I'm curious to see how independent compares to full-time salarys.

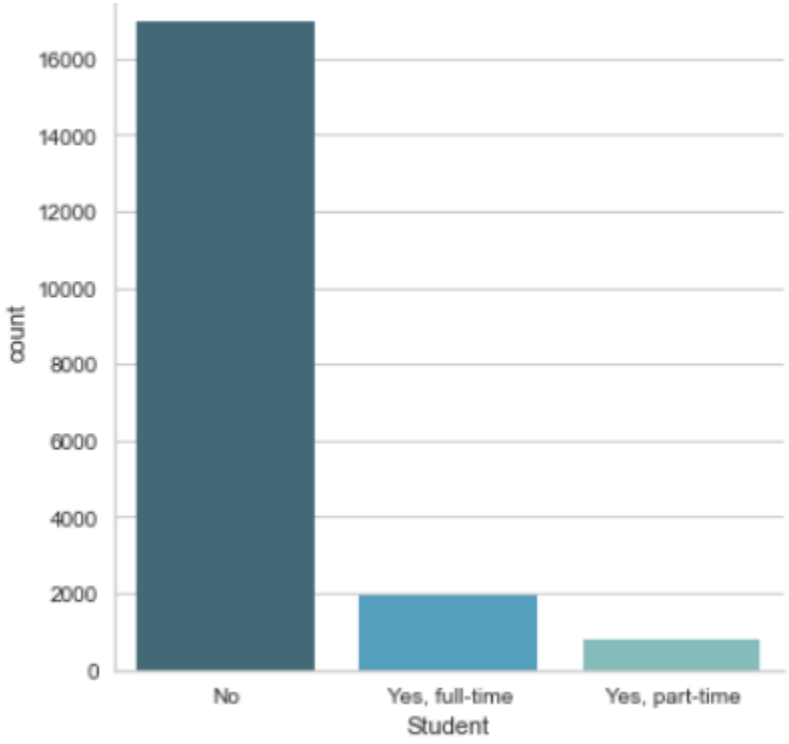
Original Case Study Narrative
Kevin Paulovici
DSC 550
May 3, 2020



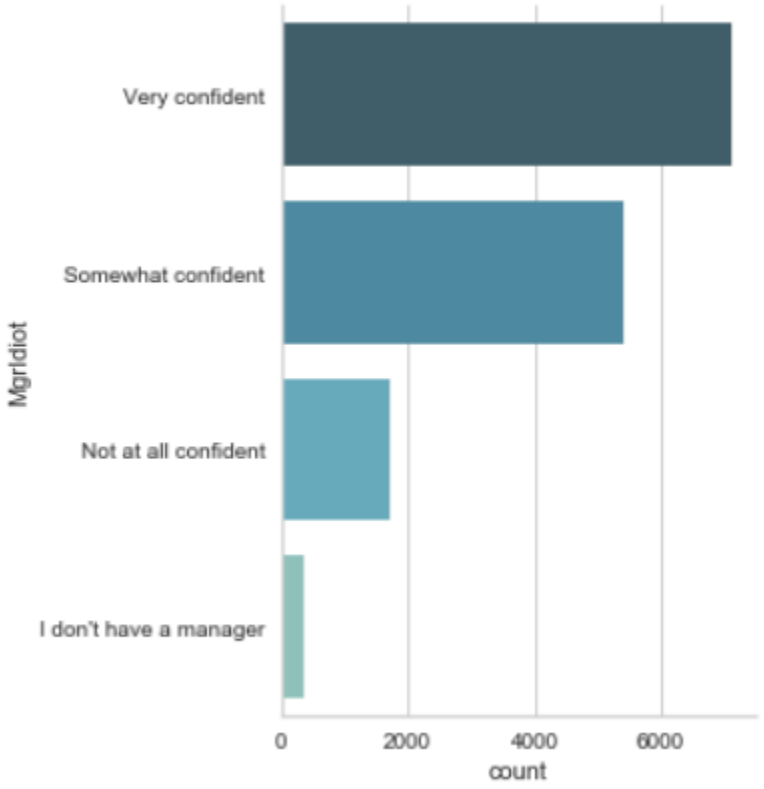


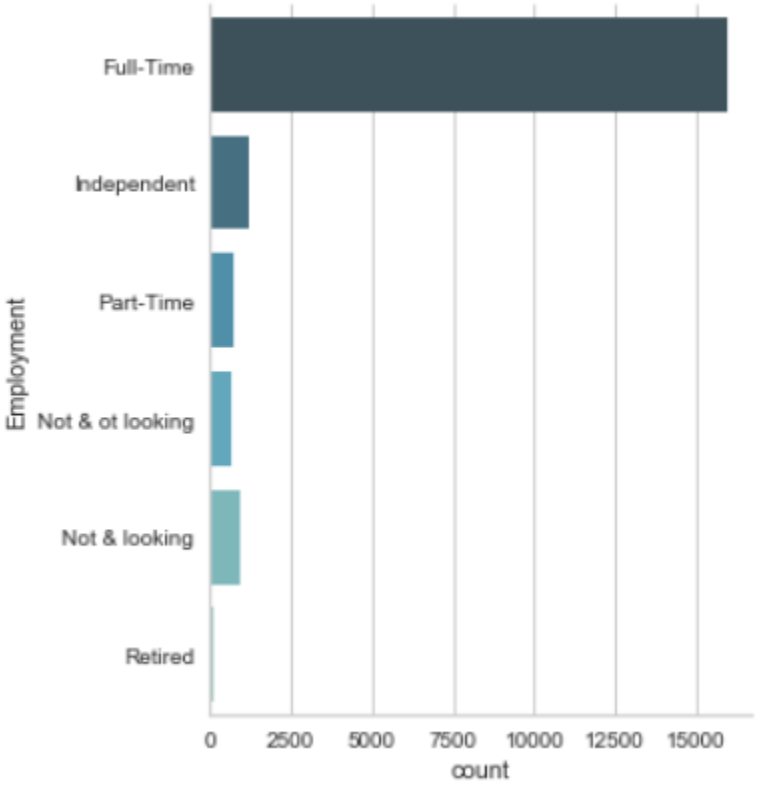
Original Case Study Narrative
Kevin Paulovici
DSC 550
May 3, 2020





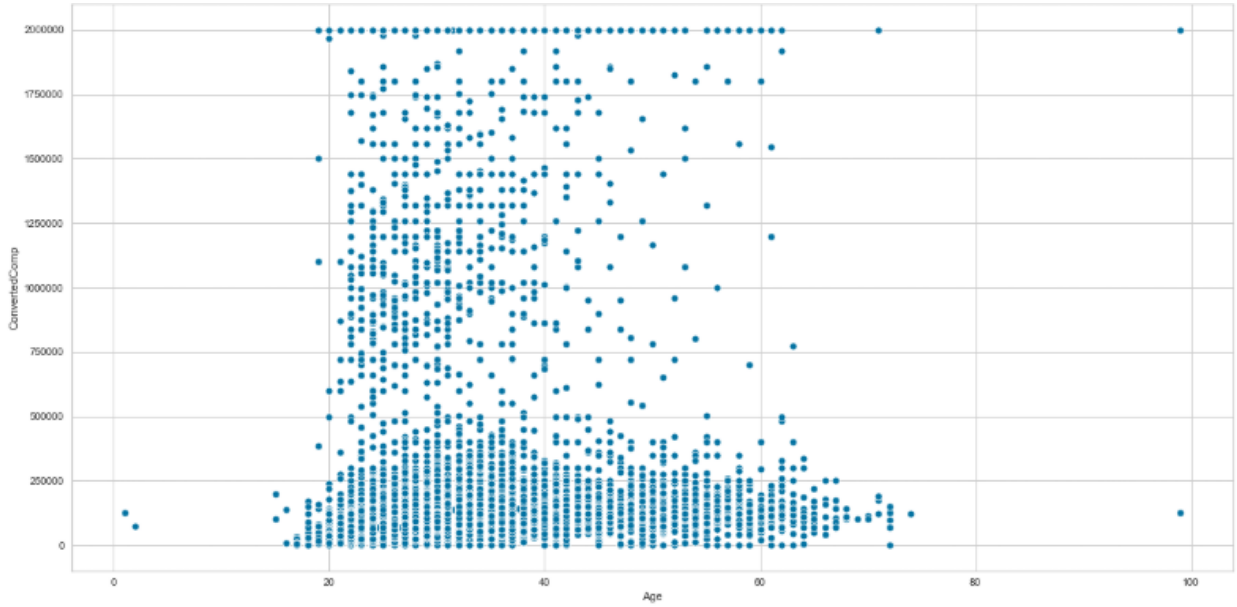
Original Case Study Narrative
Kevin Paulovici
DSC 550
May 3, 2020

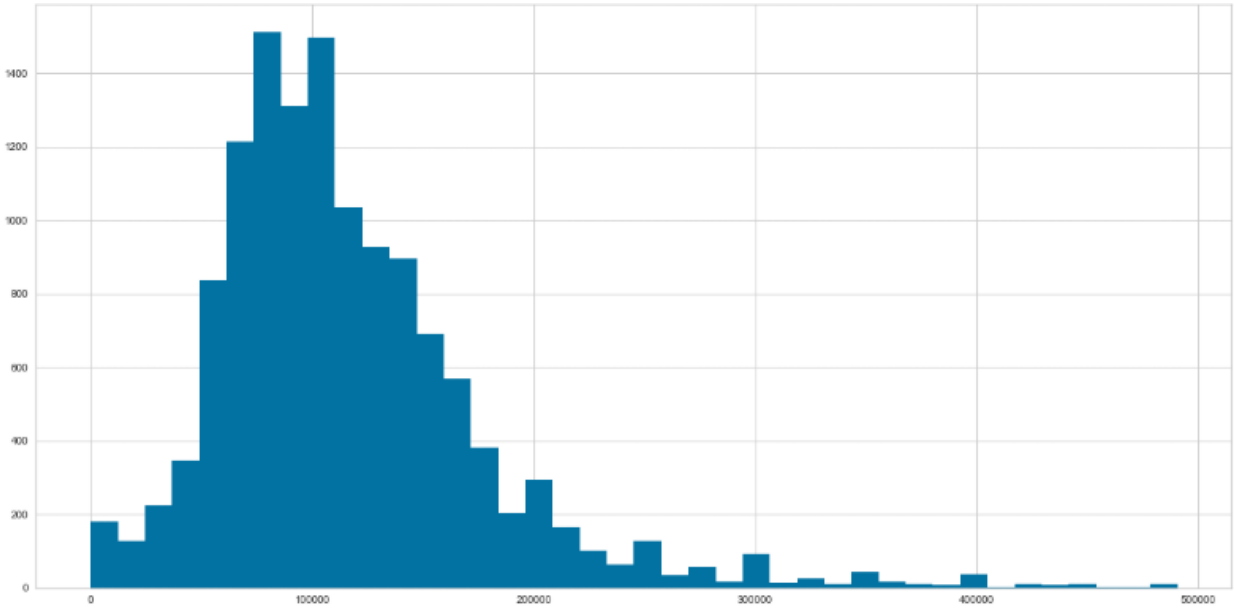




Original Case Study Narrative
Kevin Paulovici
DSC 550
May 3, 2020

13. Scatter plot /Histogram of ConvertedComp (Salary) vs. Age
    a. We see there is a wide range of salary values (yearly).
    b. It doesn't seem like a typical developer as much as the survey results show. We'll only consider salary values of under 500K in step 14.





```
count     13127.000000
mean     116373.503466
std       62200.536300
min           0.000000
25%       76000.000000
50%      105000.000000
75%      140000.000000
max      490600.000000
Name: ConvertedComp, dtype: float64
```

14. Filter additional data based on what we've seen from the graph analysis above.
    a. Step 3 removed all countries except the United States.
    b. Step 11 removed any gender that was not Man or Woman.
    c. Remove anyone that is not currently employed, retired folks or people not currently working won't provide sufficient data relative to salary.
    d. Restrict age range to 18 – 65. This is the core working demongraphic.
    e. Restrict ConvertedComp (Salary) to under 500K.
15. Remove features not of interest.
    a. Since there are 85 features, it is easier to show the features that were included
    b. Features included relate to describe the type of people included and/or related to salary.
    c. May revisit this step for Part 3 tuning.

	Column	Question Text
1	MainBranch	Which of the following options best describes ...
2	Hobbyist	Do you code as a hobby?
3	OpenSourcer	How often do you contribute to open source?
5	Employment	Which of the following best describes your cur...
13	YearsCode	Including any education, how many years have y...
15	YearsCodePro	How many years have you coded professionally (...
31	ConvertedComp	Salary converted to annual USD salaries using ...
32	WorkWeekHrs	On average, how many hours per week do you work?
77	Age	What is your age (in years)? If you prefer not...
78	Gender	Which of the following do you currently identi...

16. Drop na values
    a. Now that all filters and feature reduction have been applied we can remove any missing values.

```
The dimension of the data is: (12486, 11)
```

Original Case Study Narrative
Kevin Paulovici
DSC 550
May 3, 2020

17. Create target column – above or below average Salary
    a. Using the ConvertedComp (yearly salary) we can set a new column to determine if the value is above or below the average – this will be the target for later steps

```
count          12486
unique             2
top        Below Avg
freq            7307
Name: Above_Below_Avg_Sal, dtype: object
```

18. Convert categorical data to numbers (MainBranch, Hobbyist, OpenSourcer, Employment, Gender)

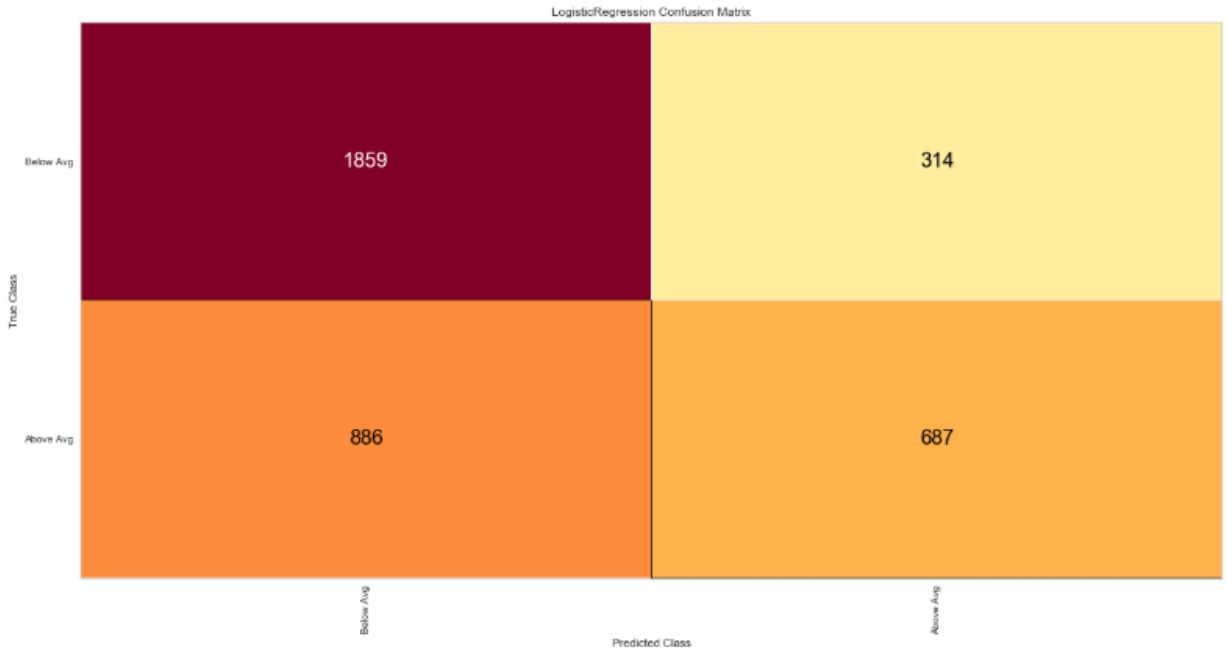
```
    MainBranch_I am a developer by profession  \
3                                           1
12                                          1
21                                          1
22                                          1
25                                          1

    MainBranch_I am not primarily a developer, but I write code sometimes as part of my work  \
3                                                   0
12                                                  0
21                                                  0
22                                                  0
25                                                  0

    Hobbyist_No  Hobbyist_Yes  OpenSourcer_< 1 / month  \
3             1             0                        0
12            0             1                        1
21            0             1                        0
22            0             1                        0
25            0             1                        0

    OpenSourcer_< 1 / year  OpenSourcer_> 1 / month  OpenSourcer_Never  \
3                        0                        0                  1
12                       0                        0                  0
21                       1                        0                  0
22                       1                        0                  0
25                       1                        0                  0

    Employment_Full-Time  Employment_Independent  Employment_Part-Time  \
3                      1                       0                     0
12                     1                       0                     0
21                     1                       0                     0
22                     1                       0                     0
25                     1                       0                     0

    Gender_Man  Gender_Woman
3            1             0
12           1             0
21           1             0
22           1             0
25           1             0
```

Original Case Study Narrative
Kevin Paulovici
DSC 550
May 3, 2020

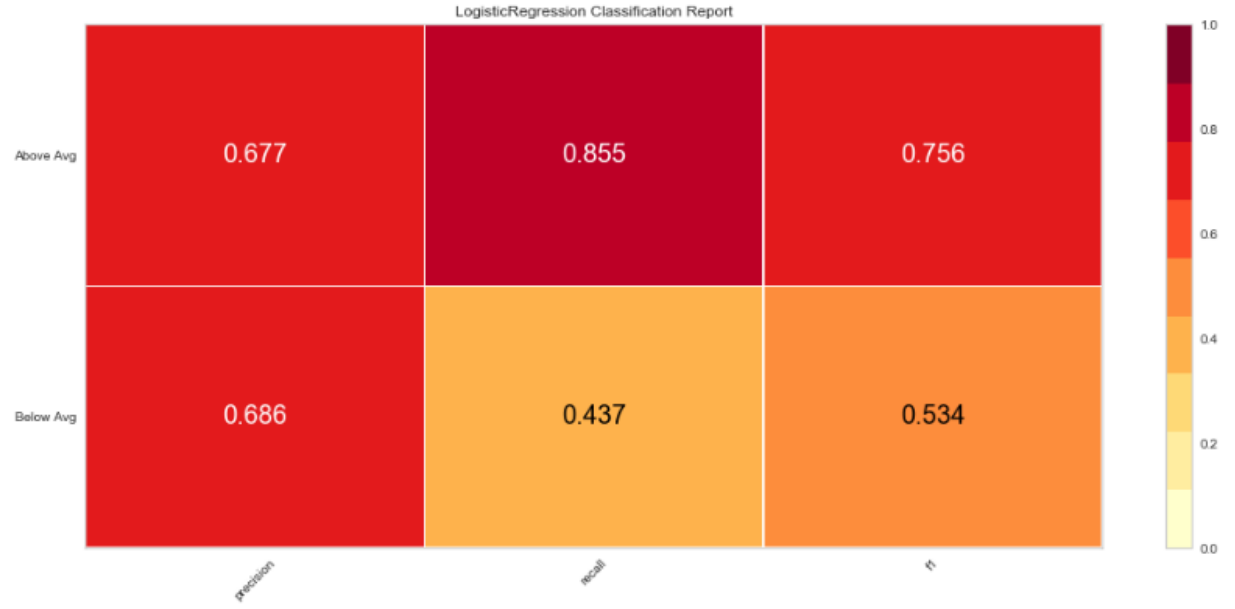
19. Training – split data into training and testing datasets
    a. As part of the tuning process I should try to even the distribution of below/above avg salary to help with the accuracy of the model.

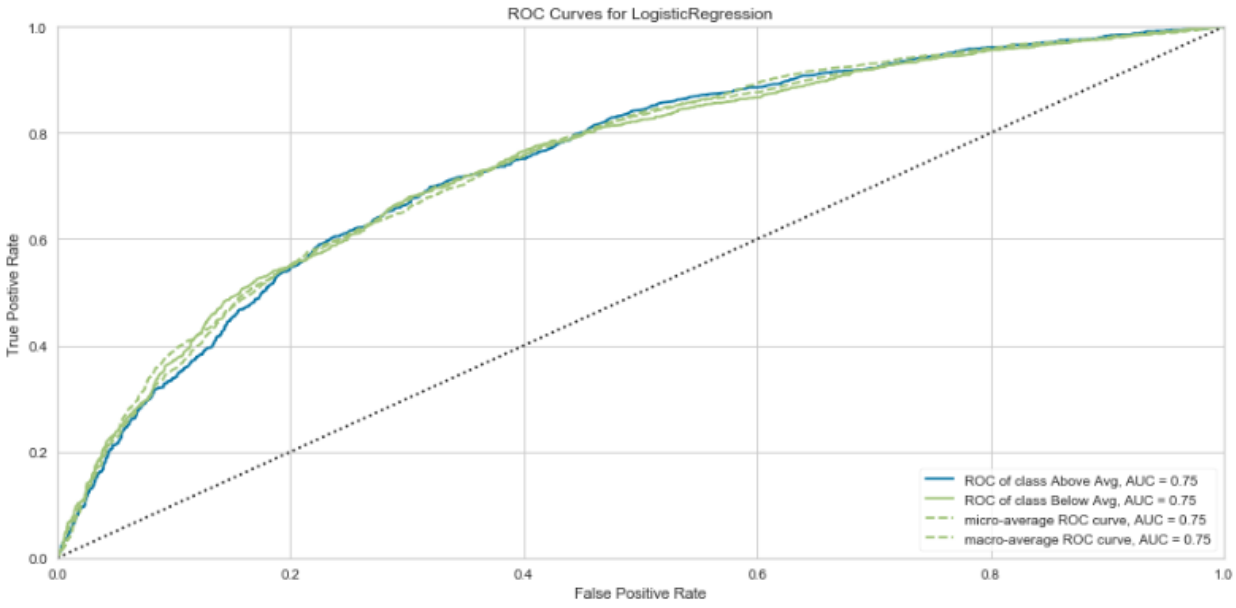
```
No. of samples in training set:  8740
No. of samples in validation set: 3746


No. of Above / Below Avg in the training set:
Below Avg    5134
Above Avg    3606
Name: Above_Below_Avg_Sal, dtype: int64


No. of Above / Below Avg in the validation set:
Below Avg    2173
Above Avg    1573
Name: Above_Below_Avg_Sal, dtype: int64
```

20. Evaluation – The goal is to predict if the salary is above/below the average salary based on the features selected using logistic regression
    a. Metrics:
        i. Confusion Matrix - ~68%
        ii. Precision, Recall & F1 score – fairly low-medium level of scores.
        iii. ROC curve – All values above the dotted line but the model needs some improvements.



Original Case Study Narrative
Kevin Paulovici
DSC 550
May 3, 2020





Future Considerations

The confusion matrix show an accuracy of ~68%. This seems pretty low, with some parameter tuning or some refinement with the filtering considerations, this can probably be improved. Additionally, target was broken into above or below average salary. This is probably too coarse given the spread of salaries reported. Maybe additional classifications could be considered. With all the filtering and parameters refined, additional scatter/line plots should be added to demonstrate trends and expectation of salary.