Exploratory Data Analysis

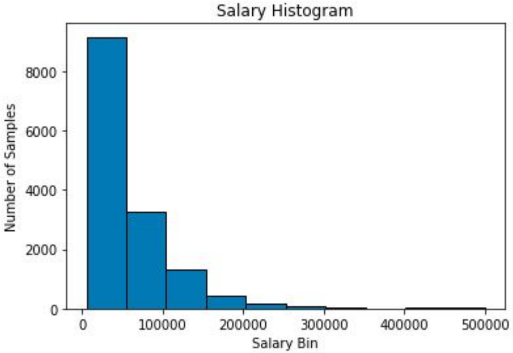

figure1:distribution of yearly income

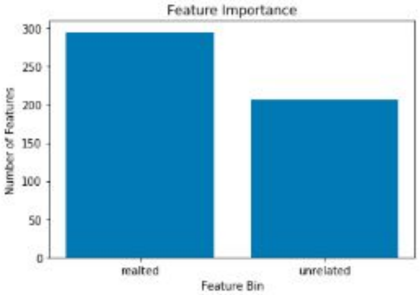

figure2:feature importance

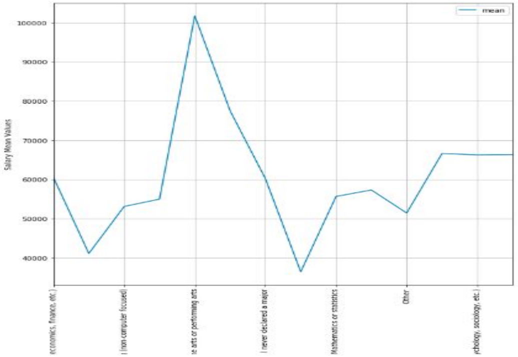

figure3:undergraduate major and salary

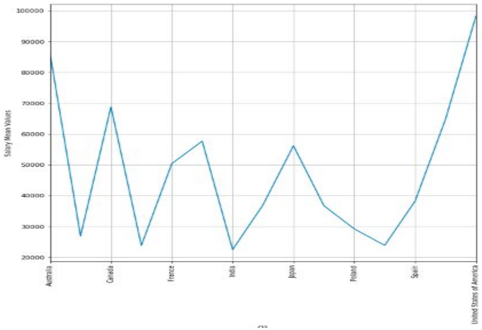

figure4:highest degree and salary

each one of them represents the distribution of yearly income, the relation between the undergraduate major and salary, the relation between highest degree and salary.

Trends in the Data:

1.From the first graph, we can see the distribution of salary among different value groups. The majority of income is below 100,000 dollars and the number decreases as the income increases.

2.The second graph tells us the relation between the undergraduate major and yearly income, from which we can see that surprisingly the fine art graduates earn the most among other majors.

3.The last graph shows the relation between data scientist's residence country and their wage, as we know the economic level differs a lot from country to country thus this feature can tells us which countries are likely to provide higher salary for data scientists.

model feature importance

1.To rank the features by their importance, we use the rfe algorithm to rank the features by running a linear regression algorithm on the raw dataset, the reason why we choose linear regression is because it is more simple than logistic regression and easier to converge. After that, we extract the outcomes and seperate the results into related and unrelated according to their rank(all features ranked 1 are considered as related, while the rest are all unrelated). Finally, we plot the above graph to visualize the importance and we can see that there are around 300 features considered real while around 250 features are not.

2.I implemented two feature selection methods at the same time, the PCA and RFE algorithms. RFE algorithm provided by sklearn allows me to rank all the features according to their importance to the model and after ranking all the features, I selected those features that are ranked as 1 and then ran PCA to reduce the dimension of the features, this step is particularly important because too many features may leads to overfitting for the reason that the noises in the dataset are also included, on the other hand too few features may cause problems for logistic regression to converge, thus by tuning the parameter for pca learner, I finally chose 250 feature which gurantee both convergence and not overfitted.

model results and visualizations

1. Classify the data using one-vers all:I splited the training data into three classes to save some run time, the label variable 'label_Y_Range1' contains all data label that are smaller than 55000,'label_Y_Range2' has all data label that are bigger than 55000 but smaller than 137500, finally we simply substract the possibility of label_Y_Range2 from 1 to get the possibility of class 3 which are labels that are bigger than 137500
2. Bias-Variance Trade-off in Hyperparameter Tuning:the more our model fits to its training data, the lower bias it would have but it gets more sensitive to the fluctuations in input. The bias of the model can be reflected by its accuracy, the higher tha accuracy is the less bias it would have and the higher standard deviation it has the higher variance it would have and they are contradictory concepts.
3. solver can have significant influence on accuracy, should use grid search to find the optimal solver

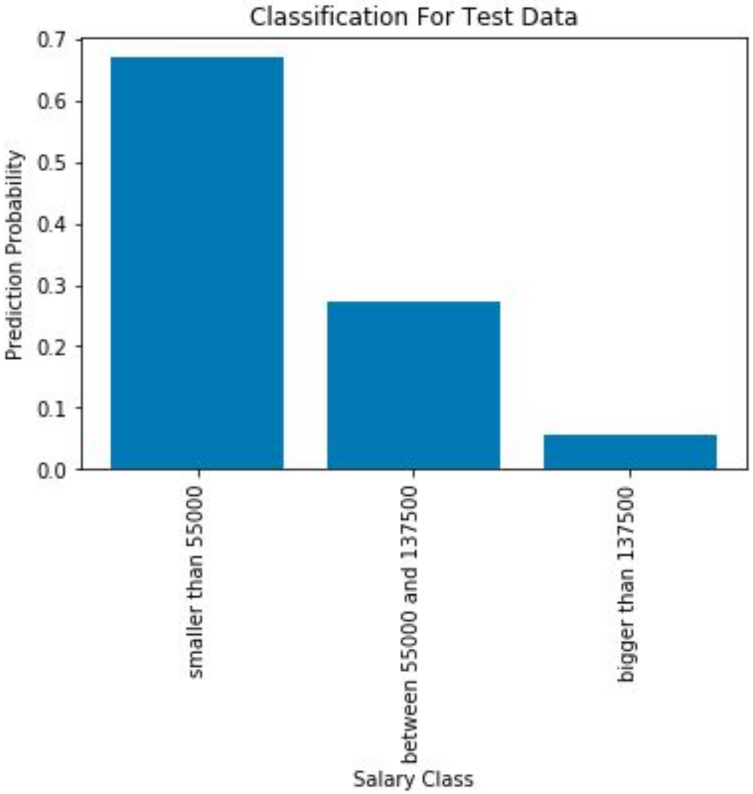

fig5: model predict result

From the classification graph, we can clearly see the distribution of classes probability and this classification problem can also be solved by one-hot encoding label, to solve multi-classification problem and I think in this case it can provide a more precise outcome, and easier to implement. Because, the training data set has 18 distinct values making it complicated to separate into binary classes.