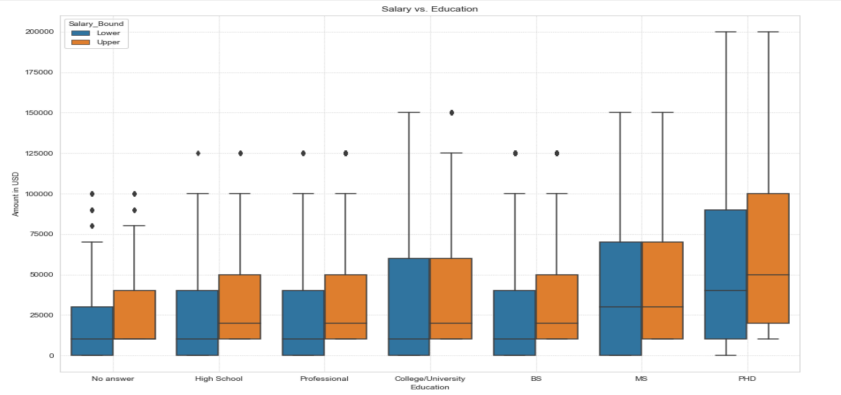
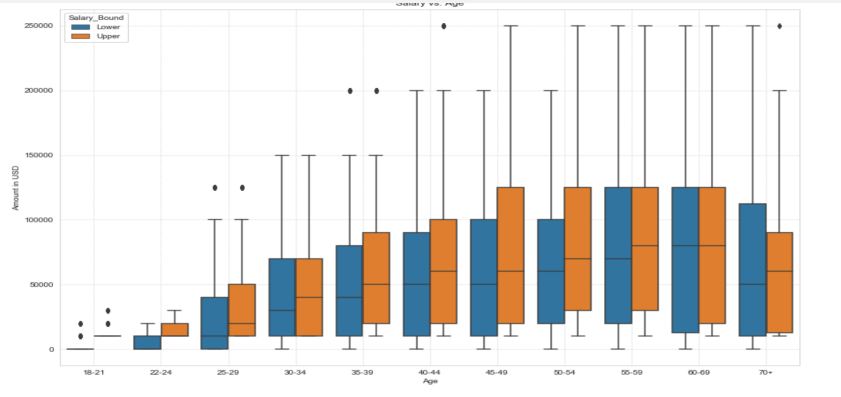
DATA SCIENCE FOR SALARY PREDICTION

Zhaohui qu 1005783127

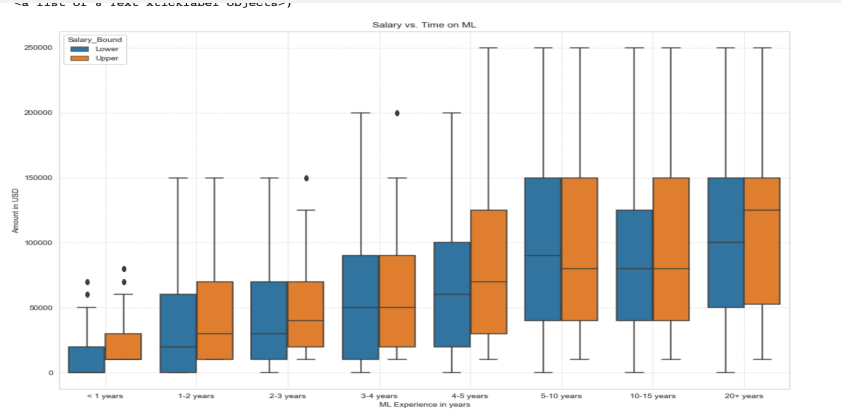
EXPLORATORY DATA ANALYSIS



If people have a higher education level, he is more likely to receive a higher salary.

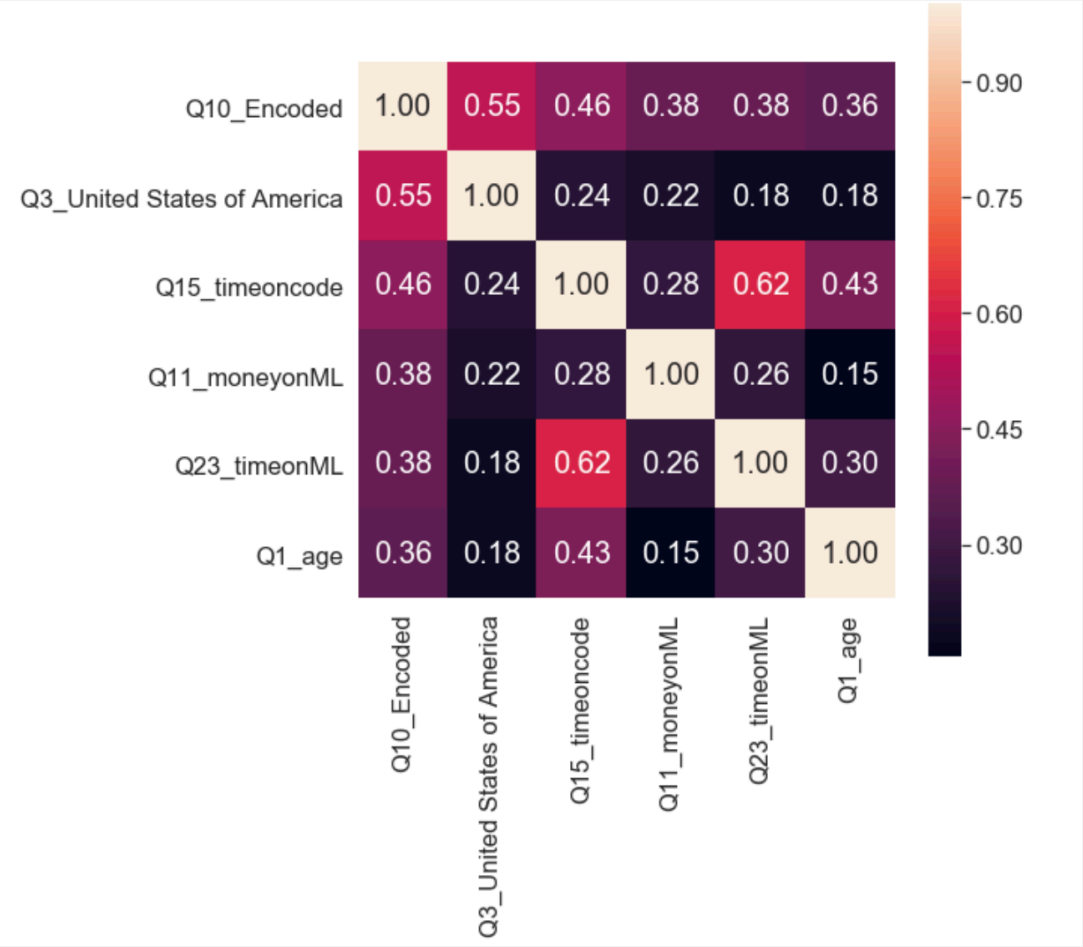


We can see that the median salary increases with growth of age and reaches peak at 60-69 years old, then drops when people are 70 years old above.



The median salary rises gradually with the time spending on machine leaning increasing, so the more time using machine learning the higher salary we could earn.

Feature Selection



TOP 1: Question 3 Are you Americans?
TOP 2: Question 15 "How long have you been writing code to analyze data (at work or at school)?"
TOP 3: Question 11 "Approximately how much money have you spent on machine learning and/or cloud computing products at your work in the past 5 years?"
TOP 4: Question 23 "For how many years have you used machine learning methods?"
TOP 5: Question 1 "What is your age (# years)?"

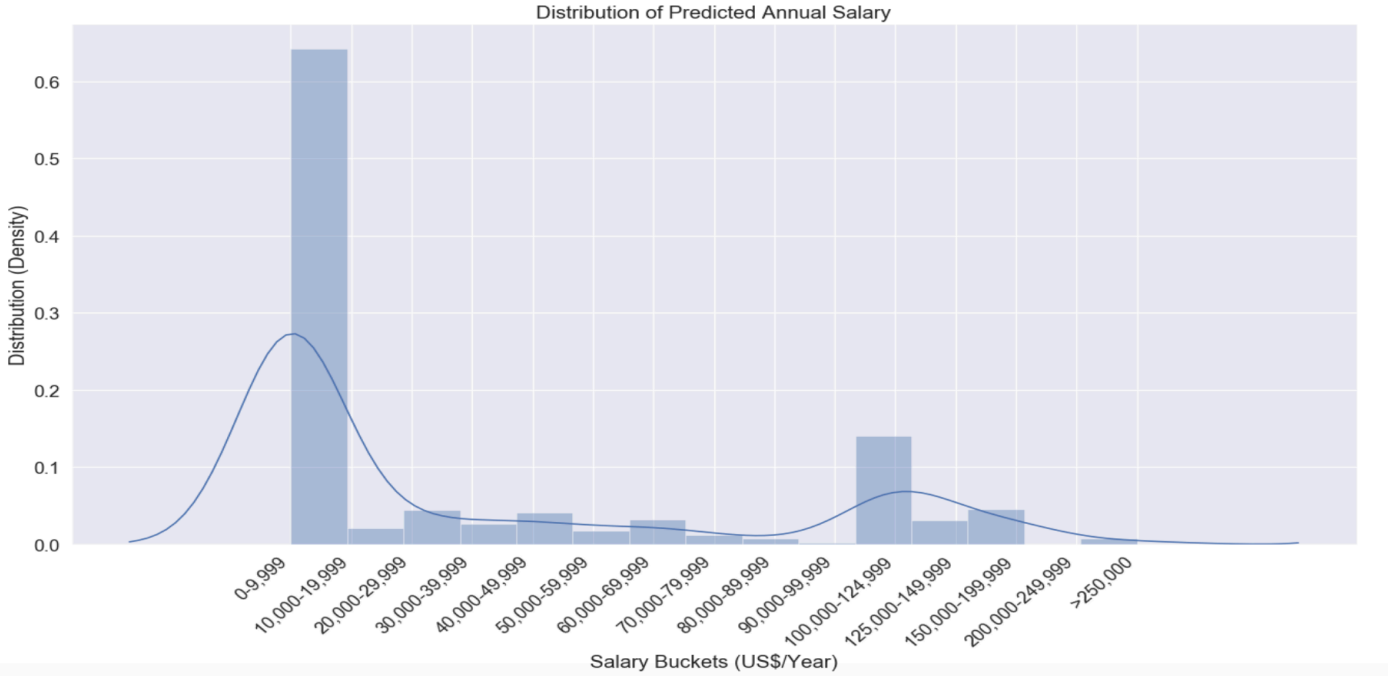
Model implementation

- The Test Set vs. The Training Set

We use Logistic Regression model for classification, our ideal parameters are:
Slover = ' saga'
Penalty = 'L1'
C = 0.3

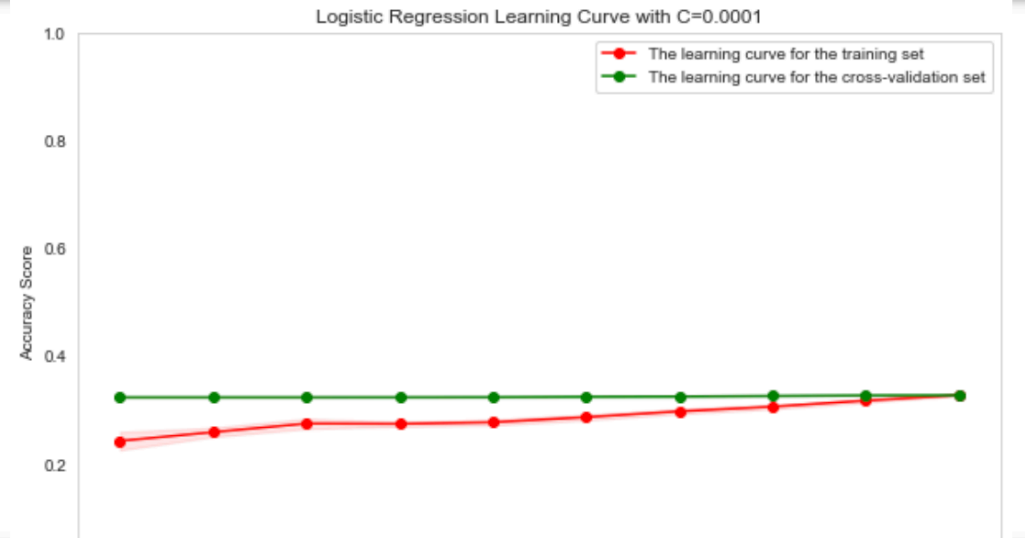
Dataset	Accuracy	Precision	Recall	F1
Training Set	39.2%	33.2%	39.2%	0.315
Test Set	37.1%	29.1%	37.1%	0.289

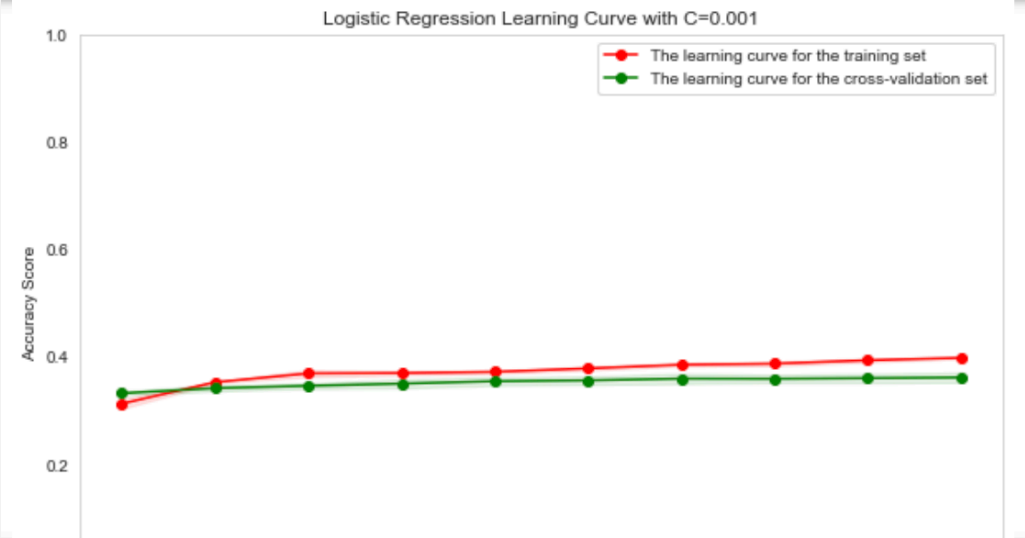
- Distribution of Predicted Annual Salary

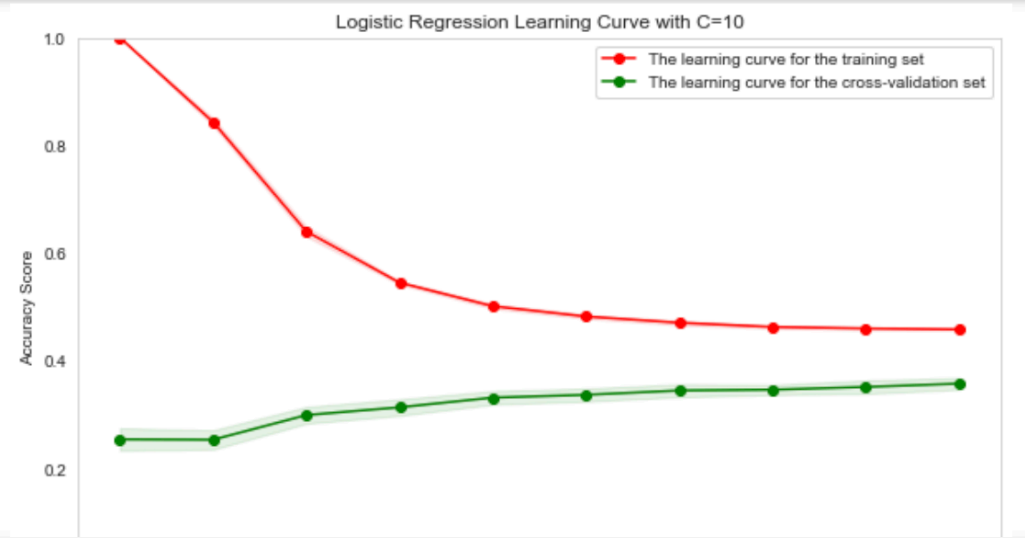


Why does not the logistic regression model fit our dataset well? It is possible that only small part of salary buckets could be separated linearly. From the distribution of predicted annual salary buckets in S6.2, we can see that 0-9,999 and 100,000-124,999 have a higher probability than other buckets and the left buckets have a even probability except for 90,000-99,999 and 200,000-249,999 that have a about zero probability. Therefore, 0-9,999 and other buckets can be separated linearly, but the other buckets are possible to be mixed up.

Bias-Variance trade off







When C is very small(C=0.0001), from the learning curve we can see that both training score and cross-validation score are low when training set size is close to 0. While the size of training dataset is increasing, the cross-validation score does not change, but the training score increases gradually and finally catches up the cross-validation score. However, both scores are low, so we have a high bias and low variance in this case.

While with the C value going larger and larger, from the learning curve we can see that the training score gradually improves and test score slightly improves, however, the gap between the training score and test score become larger and larger, therefore we have a low bias and high variance compared with smaller C.