Data Science for Compensation Prediction

MIE 1624 Introduction to Data Science and Analytics

Name: Zichuan Wang
Student ID:1000474300

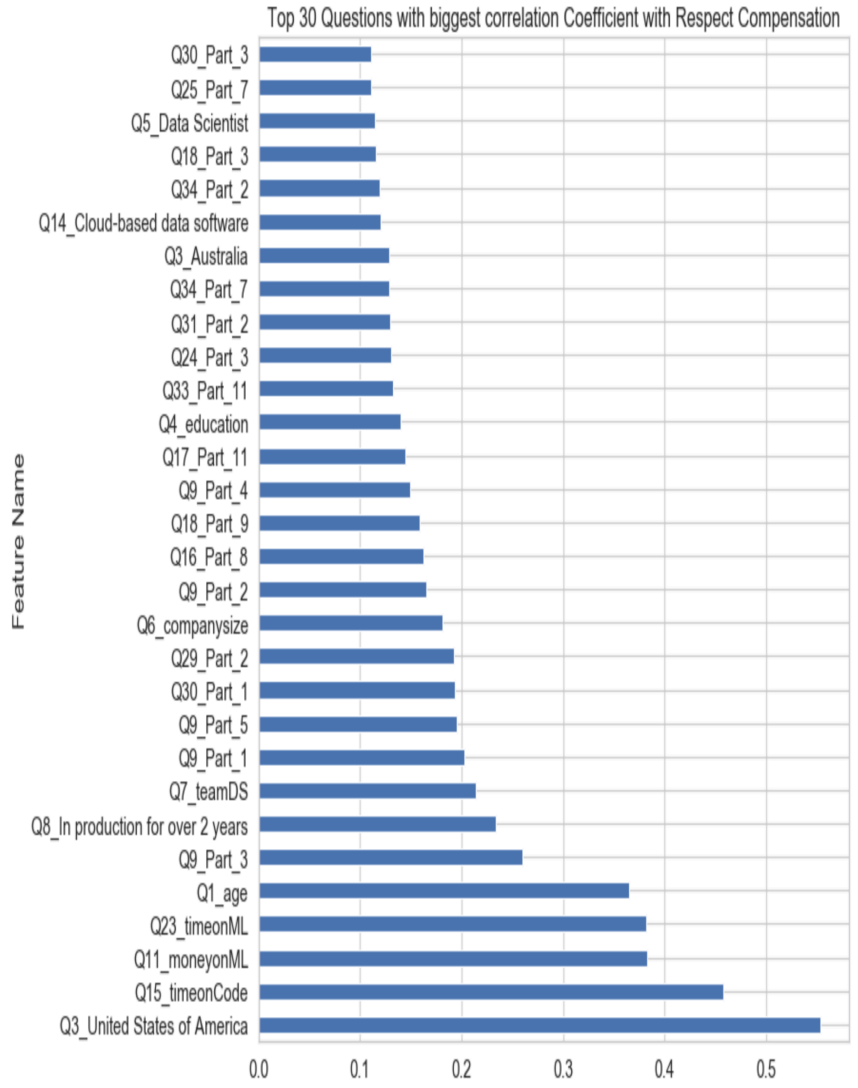
How we Interpret the Data

01

Feature Selection

Feature selection is crucial for machine learning, since it can greatly reduce the complexity without losing accuracy. There methods namely, correlation matrix, feature importance score and univariate selection, are used for feature selection.

From the graph right hand sided, whether people lives in USA is a relatively important features. The follows important features are whether this people are coding experience, money and time spent on machine learning and current age.



How we Interpret the Data

02

Model Implementation Finding

For Correlation Matrix Method, the accuracy before cross validation is 34.32%. The accuracy across all folds are shown above with a average of 34.754% and a variance of 21.757.
For Feature Importance Method, the accuracy before cross validation is 34.29%. The accuracy across all folds are shown above with a average of 34.882% and a variance of 19.076.
For Univariate Selection Method, the accuracy before cross validation is 35.89%. The accuracy across all folds are shown above with a average of 35.85% and a variance of 18.526.
Overall, Univariate Selection Method has the highest accuracy and the lowest variance, proven to be the best among all three methods.

Correlation Matrix

```
This model got an accuracy of 34.32% on the testing set
With cross validation
Fold 1: Accuracy: 27.6%
Fold 2: Accuracy: 32.72%
Fold 3: Accuracy: 31.2%
Fold 4: Accuracy: 31.36%
Fold 5: Accuracy: 31.28%
Fold 6: Accuracy: 37.6%
Fold 7: Accuracy: 33.44%
Fold 8: Accuracy: 38.591%
Fold 9: Accuracy: 42.114%
Fold 10: Accuracy: 41.633%
Average accuracy: 34.754%
Variance: 21.757
```

Feature Importance

```
This model got an accuracy of 34.29% on the testing set
With cross validation
Fold 1: Accuracy: 28.8%
Fold 2: Accuracy: 32.72%
Fold 3: Accuracy: 30.96%
Fold 4: Accuracy: 31.84%
Fold 5: Accuracy: 32.08%
Fold 6: Accuracy: 35.92%
Fold 7: Accuracy: 33.68%
Fold 8: Accuracy: 39.311%
Fold 9: Accuracy: 41.873%
Fold 10: Accuracy: 41.633%
Average accuracy: 34.882%
Variance: 19.076
```

Univariate Selection

```
This model got an accuracy of 35.89% on the testing set
With cross validation
Fold 1: Accuracy: 29.04%
Fold 2: Accuracy: 34.88%
Fold 3: Accuracy: 32.0%
Fold 4: Accuracy: 32.64%
Fold 5: Accuracy: 32.8%
Fold 6: Accuracy: 37.52%
Fold 7: Accuracy: 35.04%
Fold 8: Accuracy: 40.032%
Fold 9: Accuracy: 42.434%
Fold 10: Accuracy: 42.114%
Average accuracy: 35.85%
Variance: 18.526
```

03

Visualization

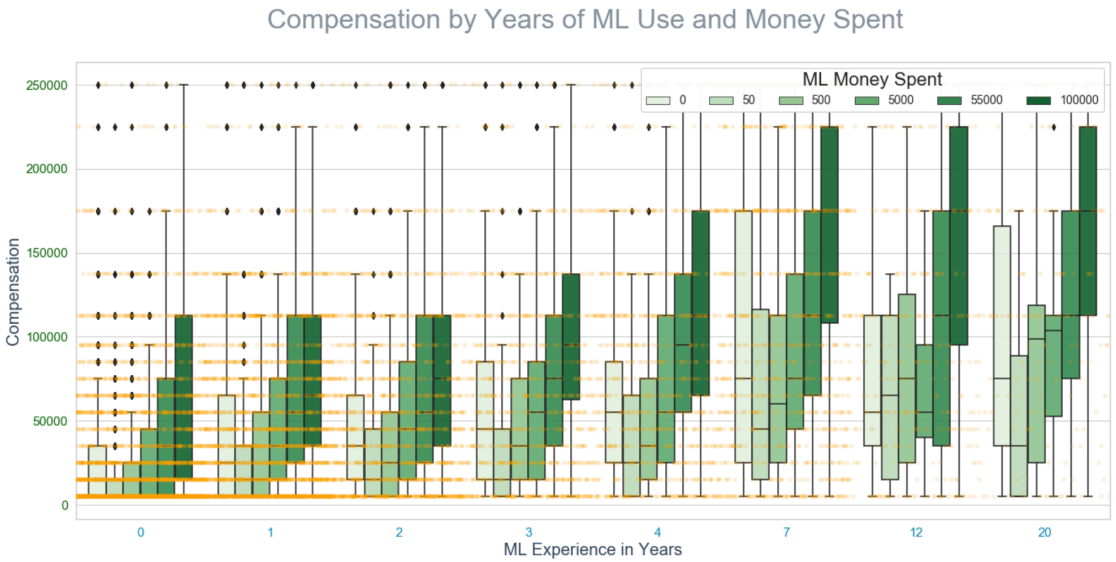
ML money spent: According to the above figure, the respondent who spent moderate amount of money on machine learning has the lowest income. Respondent who spent the most amount of money has the highest average income. Respondent who barely spent any money has medium income.

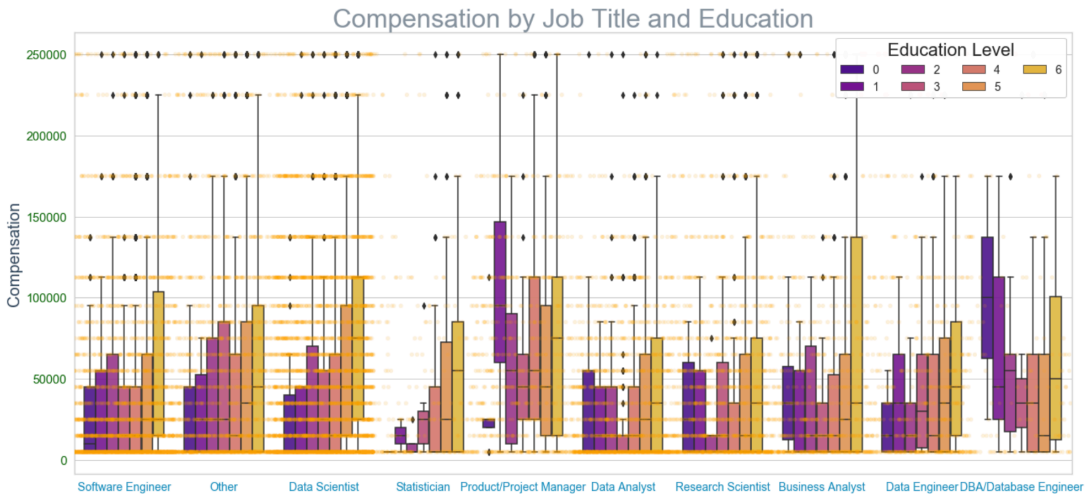
ML experience: generally, respondent with more experience has higher income.

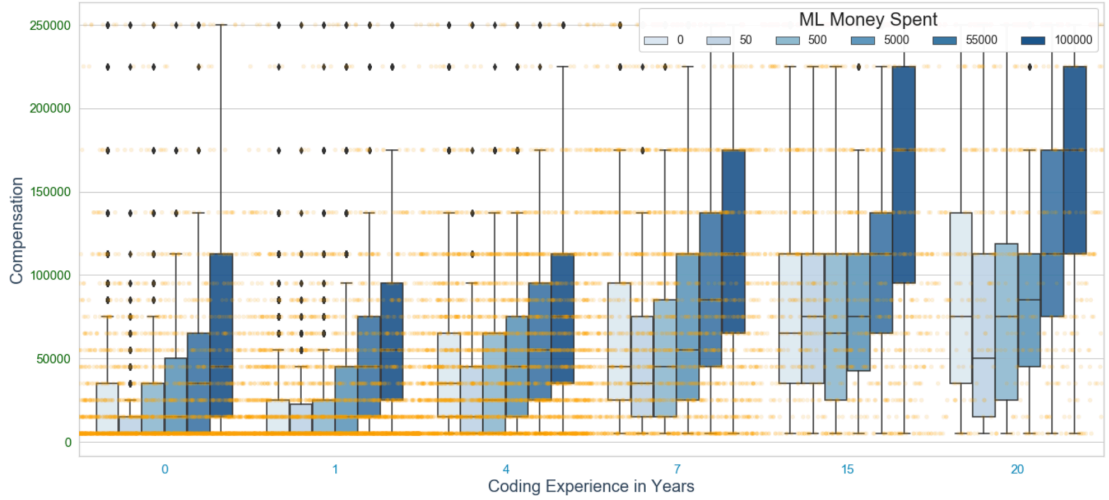
Coding Experience: has similar effect as machine learning experience.

Company Size / Data Science Team size: both has positive correlation with compensation, but the effect is not as strong as related experience.

Most of the jobs' salary increase as a function of education level. But program manager and database engineer with the lowest level of education have the highest average compensation. One possible explanation is that interpersonal skills play a bigger role than academic excellence in those kind of jobs.







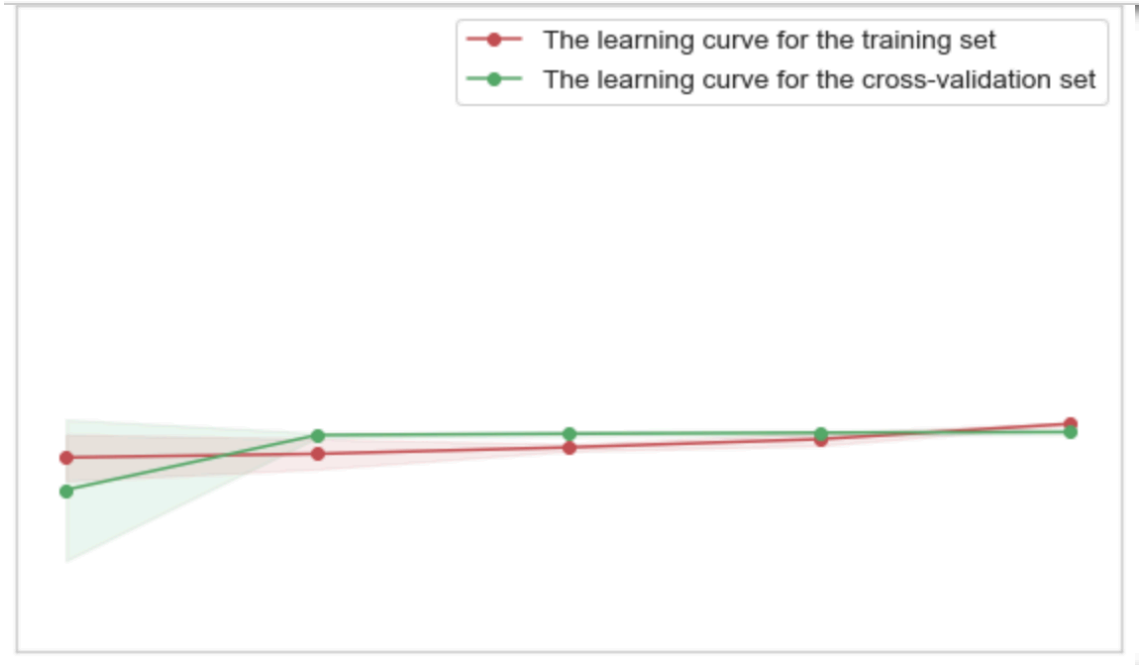
04

Bias & Variance Trade-off

With small C value, the training score is close to the testing score. This indicates that the model has high bias. Also both training and testing score are low accuracy score. This indicates that the model has high bias.

With large C value, the gap between training and testing score becomes smaller and smaller as number of training data increases. This indicates that the model has low variance and high bias.

C = 0.01



C = 100