Supervised Learning

Alek Francescangeli

October 11, 2022

Contents

1 The Datasets 1

2 Decision Tree 2

3 Decision Tree Pruning 2

4 Neural Network 3

5 Boosting 3

6 k-Nearest Neighbor 4

7 Conclusion 4

1 The Datasets

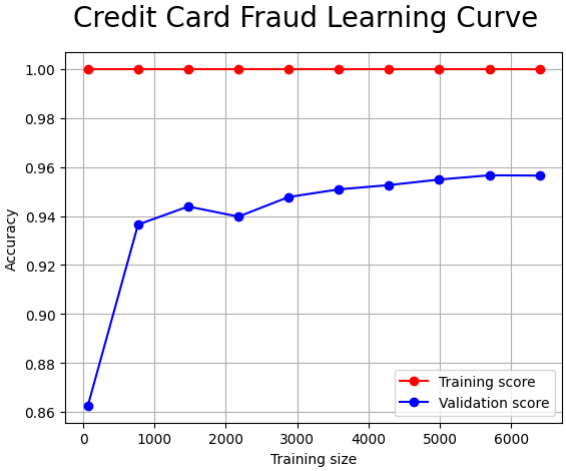
The first dataset in this paper contains data on data science job salaries for the years 2020-2022. This dataset contains approximately 2000 values collected from people in the data science field with factors such as location, remote-work ratio, and experience being included. The classification problem will tackle which factors will affect salary the most. This dataset will be split between training/testing with a 80/20 ratio. The second dataset contains data from over 200,000 credit card transactions occurring over the course of two days. From these transactions there are approximately 500 positive cases of fraud. This classification problem will attempt to identify which transactions are most likely to be fraudulent.

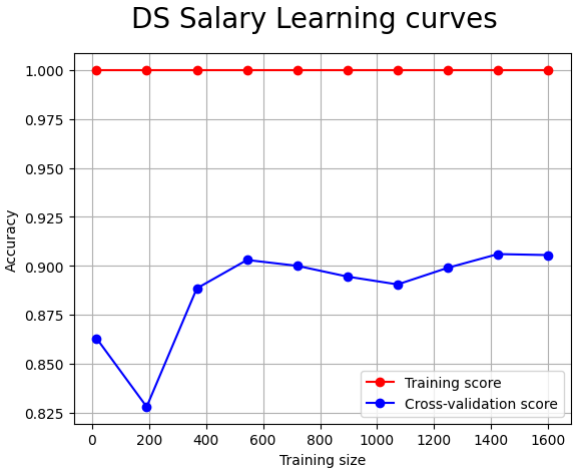
This dataset will also be split with a 80/20 ratio. These two datasets will be interesting to compare as their scope varies greatly with the credit card transaction dataset containing the credit card fraud dataset containing many more data points and details for each data point. In the case of the credit card fraud dataset I predict that the regularization parameters in particular will need to be tuned much more to account for the larger dataset and smaller differences between data points. The problem of attributes being distributed relatively equally in classifications is something that will need to be adjusted for.

All algorithms used in testing were modified from the Scikit-learn library in Python. Accuracy in these experiments will be defined as (Correct Predictions/Total Predictions).

2 Decision Tree

Using a decision tree we can see the following results for both datasets:

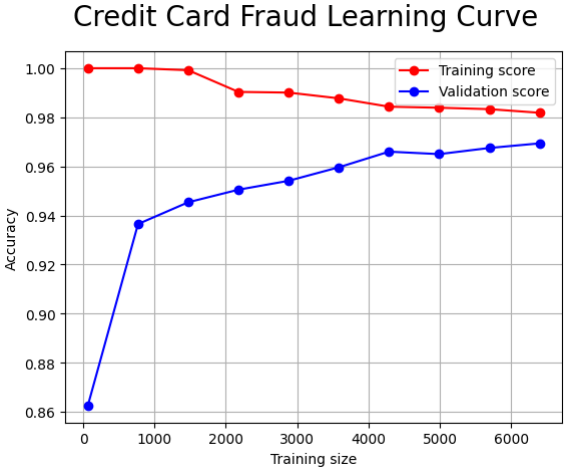


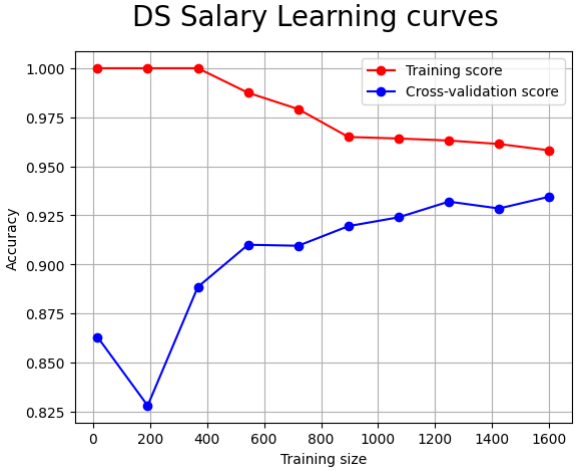


This model uses training data with accuracy=1 which will cause low bias and high variance in the model.

3 Decision Tree Pruning

Using a pruned decision tree we can see the following results for both datasets:

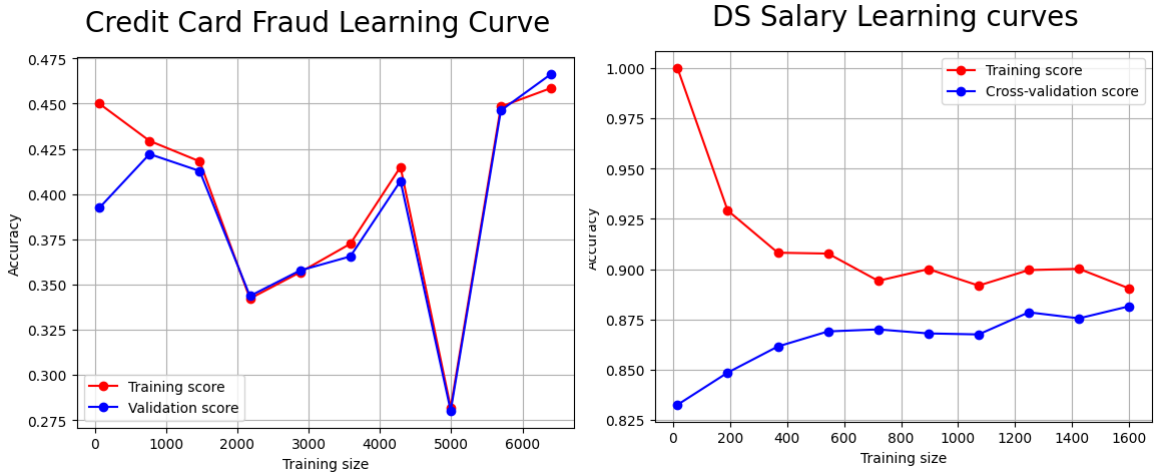




To prune the decision tree a "Minimal Cost-Complexity Pruning" algorithm was used (from the scikit-learn library). This algorithm works by pruning nodes with the smallest effective alpha. This improves the accuracy for both datasets.

4 Neural Network

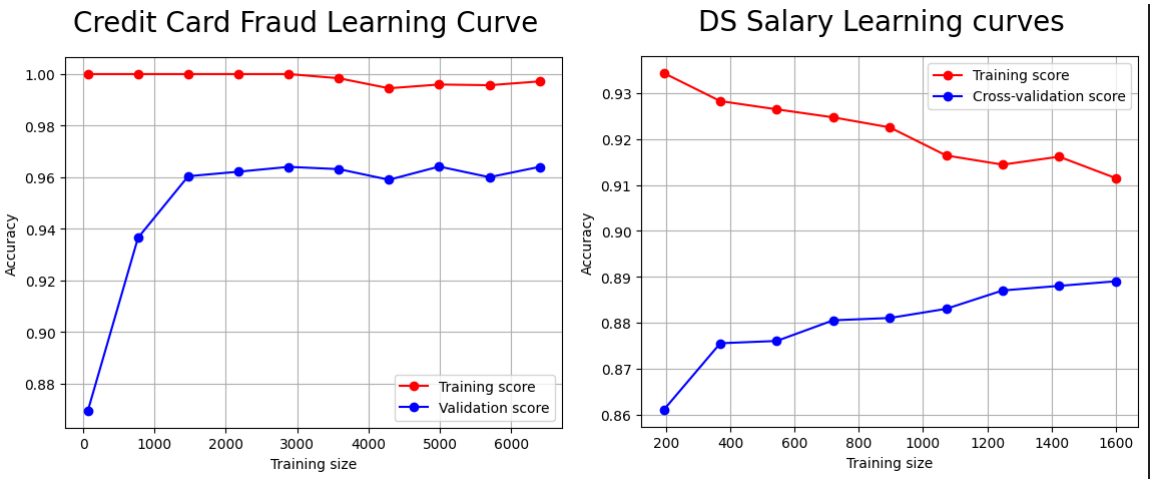
Using a neural network method we can see the following results for both datasets:



The neural network for both datasets was done with both datasets with 4 layers and 30 units for each level. This was not able to be effective for the credit card fraud learning curve. I believe this is caused by the size of the dataset. I believe that this method could be tuned to give better results along with allowing more time for the neural network to be trained.

5 Boosting

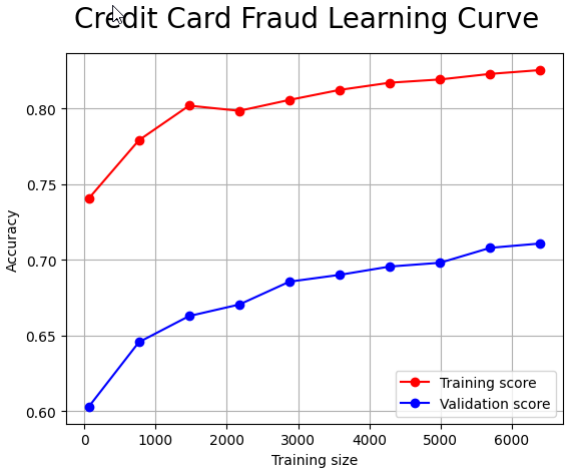
Using a boosting method on our decision tree we can see the following results for both datasets:

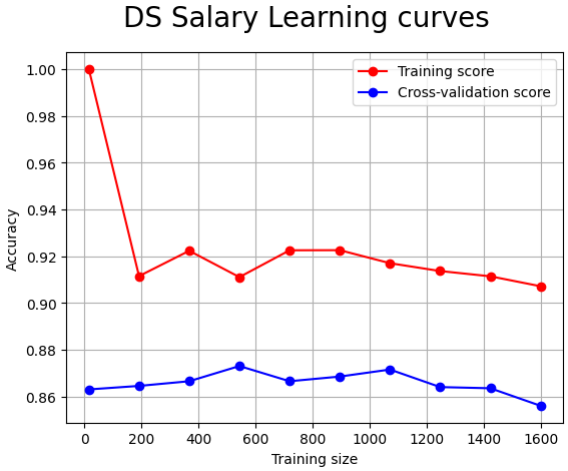


The boosting method used for these dataset was "Gradient Tree Boosting". This method was chosen as a way to try to minimize errors in the previous decision tree. In both cases this method was not as effective as the pruned decision tree experiment.

6 k-Nearest Neighbor

Using the K-nearest neighbor method we can see the following result for both datasets:





After some tuning the k-nearest neighbor the best results came with 6 neighbors and a leaf-size of 25.

7 Conclusion

After going through these experiments the best model for the Data Scientist salary dataset seems to be the pruned decision tree with an accuracy of 0.93 and the best model for the Credit Card fraud dataset is also the pruned decision tree with an accuracy of 0.97. The only method that came close was the boosting method with a difference in accuracy of less than .001 for the Credit Card fraud dataset.