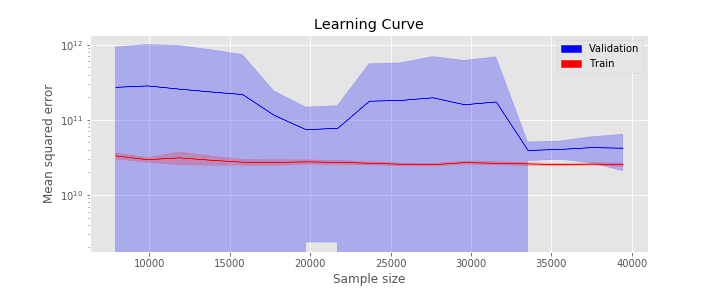
**Learning Curve:**

We examine the learning curve of the model to examine whether the model suffers from immediate over- or underfitting problems and whether these would be remedied by collecting more data (Raschka 2017:196).

The following plot illustrates the learning curve of our model:



The plotted curved represent the average performance of the sets, while the width of the plotted curves expresses the 95%-confidence interval of the performance.

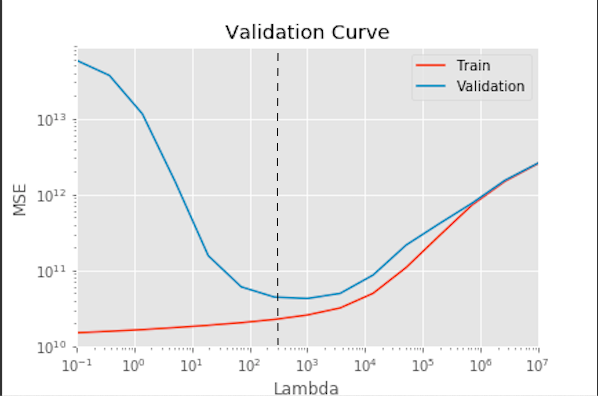
Interpreting the graph could appear quite daunting, as the performances on the validation set are very inconsistent and varies a lot. Though inconsistent, it seems as if the gap is decreasing between the performances of the training- and validation sets as the sample size increases.

At around 33.000 observations the validation curve drastically decreases and both curves flatten out. Though the confidence interval of the validation performance is still noteworthy it also improves dramatically when the sample gets sufficiently high. Just before 40.000 observation, the confidence interval of the validation curve actually drops below the train curve, which should not be possible. We have not found a meaningful explanation aside from noise or chance (<https://jakevdp.github.io/PythonDataScienceHandbook/05.03-hyperparameters-and-model-validation.html>)

All in all, the learning curve would indicate that the performance of the model will not benefit from additional training data.

**Validation Curve:**

We examine the validation curve of the model to see whether we have found a good bias-variance-trade-off with our optimized hyperparameter. We plot the average train and cross-validation performance for different values of



The dotted line approximately represents the optimized value of lambda from our previously conducted k-fold cross-validation.

Firstly, we are somewhat satisfied with the average performance of the model. With an MSE of approximately and a MAE of DKK 43,630, we find that the model performs quite well on unseen data. – Skal måske stå et andet sted.

The curve suggest that we have found the optimal degree of regularization with our hyperparameter. Had we chosen a smaller value of , the model would have had too high variance and thus been overfit. This is deduced by how poorly the validation predictions performs at smaller lambdas, as opposed to the training performance. On the contrary, had we chosen a greater value of , the model would have been too biased. This is illustrated by how poorly the training data, as well as the validation data, performs at higher values of . In that case, we would have underfit our model.

Nonetheless, it is apparent that our model performs better on the training data, which indicates that it retains a degree of overfitting. For the abovementioned reasons, it would not be meaningful to minimize this performance gab by adjusting the hyperparameter. Other options for decreasing the variance of the model, could be to include more training data, but as illustrated by the learning curve, this would not seem to make for more less variance. Lastly, as overfitting is often caused by a sizeable number of features, omission would be a reasonable mean. As we have already conducted feature selection on the basis of correlations, we did not find a feasible way to conduct further selection. This could be a subject for further studies, to investigate whether this could improve the performance and evaluation of the model.

In conclusion, it would seem that we have found the hyperparameter which yields the optimal bias-variance-trade-off. We can thus conclude that we have optimized our model under the given circumstances.

**Conclusion**

In this assignment, we have used machine learning to train a model to predict valuation prices of real estate in Denmark. We did this by scraping Boliga.dk for all of their active houses for sale and combining this with municipal data on income, education, schools and urbanization. Additionally, we also scraped data from hvorlangterder.dk which calculates the distance from an address, to different every-day commodities such as hospitals, schools and shopping.   
We have explained how we carried out our gathering of data, how we have structured the data and how we cleaned it.

We performed descriptive statistics on the data to gain a better acquaintance with the data.

We conducted feature selection by assessing the correlation between the different features and the target variable, which in our assignment has been the valuation price.

We trained both a Lasso- and a Ridge regression model and compared their performances with the performance of a standard linear regression. Though k-fold cross-validation, we estimated the optimized hyperparameters for both regularization models. We found that the Ridge regression performed best on average, with the performance measure being the smallest mean square error of the model.

We have assessed to which degree we have managed to find a good bias-variance-balance for our model and concluded that our model suffers from a degree of overfitting. This was done by examining the validation curve of our final model.

Finally, we have discussed the results of our analysis. We have critically assessed the boundaries of our research and drawn up potential possibilities for improvement.