Big Data (AES 690) Homework 3

Mitchell Dodson

February 22, 2024
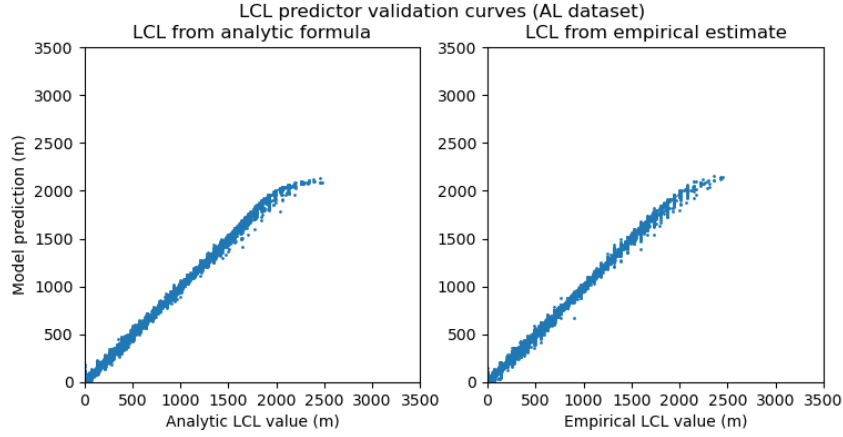
Predictor trained on Alabama ASOS data

Figure 1: Feedforward network validation curves for emulating analytic LCL calculations (left) and empirical LCL estimates (right). Data used for evaluation includes training and validation data from Alabama ASOS stations in July, 2023. $R^2 = 99.82\%$ for the analytic predictions and $R^2 = 99.81\%$ for the empirical predictions. Mean absolute error magnitudes were $14.23\,\mathrm{m}$ and $14.32\,\mathrm{m}$, respectively.

Figure 1 shows the validation curves of a 7-layer feed-forward predictor with 14,002 parameters. The model receives surface measurements of the temperature, relative humidity, wind speed, and mean pressure as inputs, then generates predictions for the LCL height in meters as calculated by the formula from (Romps, 2017), and an empirical estimate, respectively. In both cases, the predictions generally trend well, but start to underestimate the LCL height in the upper extremes of values represented by the dataset. This may be caused by bias in the training sample distribution, which over-represents surface conditions near saturation (and subsequently having lower LCL heights).
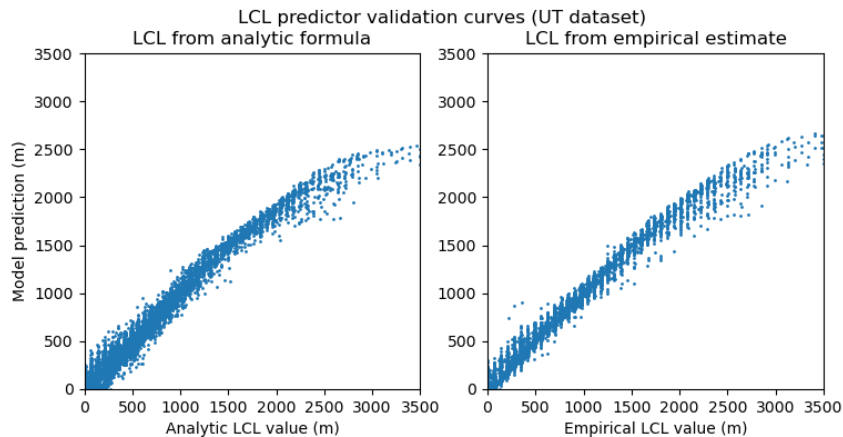

Figure 2: Validation curves for the same model as above, evaluated on unseen ASOS data from Utah in March 2023. $R^2 = 96.64\%$ for the analytic predictions and $R^2 = 97.21\%$ for the empirical predictions. Mean absolute errors were $67.1\,\mathrm{m}$ and $55.1\,\mathrm{m}$, respectively.

As the point distributions in Figure 2 indicate, the same model struggled to generalize to the more arid conditions from the March 2023 ASOS data from Utah stations. Due to the dry conditions, the unseen dataset contains much higher LCL heights than did the data used to train the model. As such, the highest LCL heights predicted by the model are only about $2,500\,\mathrm{m}$ (the upper limit of training data), while the corresponding actual values reach heights of more than $3,000\,\mathrm{m}$. Even LCL values within the normal range of the training data have prediction error magnitude around four times higher than that of Alabama stations.

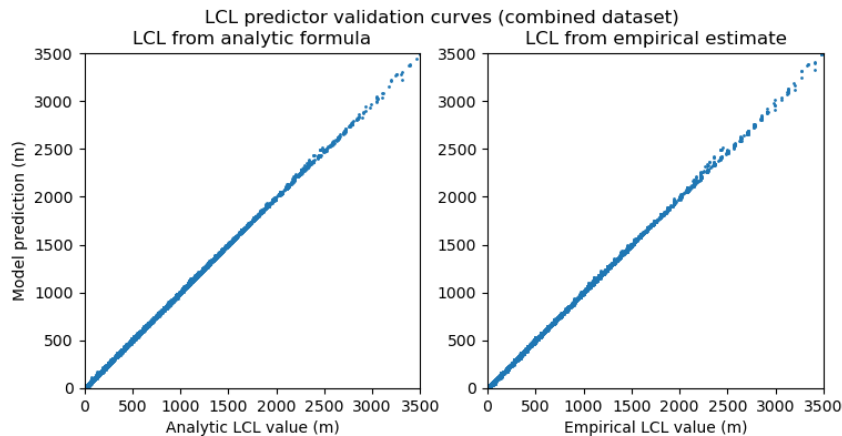## Predictor trained on combined ASOS data

Figure 3: Validation curves for a feedforward model trained on ASOS data from Alabama as well as Utah, evaluated on its training and validation data. $R^2 = 99.983\%$ for the analytic predictions and $R^2 = 99.973\%$ for the empirical predictions. Mean absolute errors were $5.46\,\mathrm{m}$ and $6.77\,\mathrm{m}$, respectively.

The next model is an 11-layer feedforward model with 164,050 parameters, trained on the same four input variables including data from both Alabama and Utah. After training on only 60% of the data (the other 40% going towards validation), the new model achieved considerably lower error rates than previous iterations when evaluated on the full dataset. Although it isn't surprising that a much larger model would make a better prediction, the 14,002 parameter model from the first section outperformed much larger models trained on the same data. With this in mind, I suspect that fitting the model on the additional data reduced the bias of the training set, and enabled the model to discover more general representation of the relationship between surface conditions and LCL altitude.

It's worth noting that the model architectures used in this report were selected by randomly searching plausible hyperparameter configurations forming a solution space with size 207,360. The hyperparameters subject to variation include model type, layer and node count, learning rate, feature masking amount, activation function, and dropout rates. With a sample size of 128 models, the most performant had at least 6 layers, a wide variety of layer widths, no dropout, no masking, logistic activation functions, and learning rate ceilings between 1×10^{-4} and 1×10^{-6}.

My code for this assignment is available on github at:
`https://github.com/Mitchell-D/aes690hw3`

Models were developed with the tracktrain workflow:
`https://github.com/Mitchell-D/tracktrain`