During phase 1 of this project, we studied variable significant in predicting the *RainTomorrow* variable. After imputing missing data and checking correlation, we determined 8 variables to be the most significant predictors: *Humidity9am*, *Humidity3pm*, *Cloud9am*, *Cloud3pm*, *WindGustSpeed*, *Temp3pm*, *HumidityDiff*, and *TempDiff*. Including the response variable, *RainTomorrow*, in our list we created a new dataset. We then split the dataset with 70% of the data forming a training set and 30% going into a testing set. We used 3 different model-building techniques to determine which would be most accurate: classification trees, random forests, and neural networks. We also studied the importance of our selected variables in the random forest and neural network models.

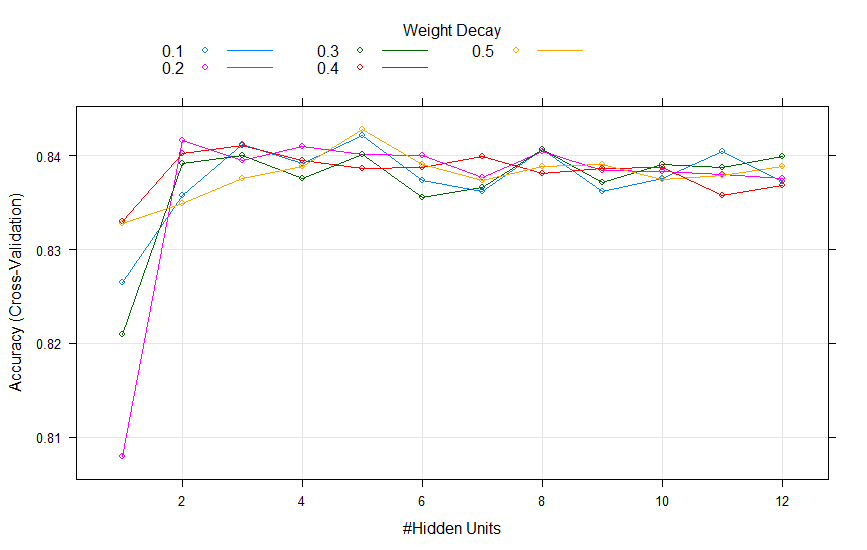
|  |  |  |  |
| --- | --- | --- | --- |
| Random Forest | | Neural Network | |
| Humidity3pm | 100.000 | TempDiff | 100.000 |
| Cloud3pm | 37.660 | Cloud3pm | 43.650 |
| Humidity9am | 27.291 | Temp3pm | 31.457 |
| WindGustSpeed | 22.028 | Cloud9am | 23.114 |
| HumidityDiff | 17.898 | WindGustSpeed | 10.147 |
| TempDiff | 9.556 | Humidity9am | 5.048 |
| Cloud9am | 6.945 | HumidityDiff | 4.091 |
| Temp3pm | 0.000 | Humidity3pm | 0.000 |

As we can see above, humidity appears to be one of the strongest predictors in our random forest model. However, the neural network places more importance on temperature difference, with humidity having much lower importance in the model. In both models, cloud cover seems to hold significant importance while wind gust speed seems moderately low.

Our first classification tree was built without specifying the complexity parameter in order to explore which value may work best. The optimal cp value for this tree ended up being about 0.221, which we then used to build a new tree. Predicting *RainTomorrow* on the training set with the pruned tree resulted in about 0.828 accuracy. Compared to the naïve model accuracy of 0.78, and considering a p-value of nearly 0, this model appeared to be a good quality model. We then made predictions on the testing dataset, which resulted in an accuracy of about 0.824. With both predictions showing greater accuracy than the naïve model, the classification tree may be viable for predicting *RainTomorrow*.

Improving on the classification tree method, we then built a random forest with 10 k-fold cross-validation and 50 trees. Developing predictions on the training set with our random forest resulted in an accuracy of 0.9991, almost 100%. While the p-value is ideal at virtually 0, the fact that our model had near perfect predictions raises concerns. In the real world, perfectly accurate models don’t exist. We then tested predicting on the testing set, which resulted in an accuracy of only 0.84 with a similarly significant p-value. This indicated that our random forest was overfit to the training set and would not be a strong model for future predictions.

The final step involved creating a neural network. We once again used 10 k-fold cross-validation to develop two networks, a basic model with specified parameters and an optimal network. The basic network was created with a size of 12 and decay of 0.1. Making predictions on the training set with our basic model showed an accuracy of about 0.841, while predictions on the testing set were around 0.844. We then used RStudio to search for the optimal size from 1-12 and decay between 0.1-0.5. A model with size 5 and decay 0.5 was determined to have the best accuracy at around 0.843, as seen below.



Building predictions again with the testing set calculated an accuracy around 0.8401, while predicting on the testing set resulted in about 0.8431 accuracy, each with a significant p-value near 0. Thus, we can conclude that among the 3 modeling techniques used, a neural network seems to provide the highest-quality model for the provided data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Training Set** | | **Testing Set** | |
| **Accuracy** | **p-value** | **Accuracy** | **p-value** |
| Classification Tree (pruned to 0.221) | .8276 | 2.2e-16 | .8235 | 2.2e-16 |
| Random Forest | .998 | 2.2e-16 | .838 | 2.2e-16 |
| Neural Network (Basic) | .841 | 2.2e-16 | .8441 | 2.2e-16 |
| Neural Network (fitted) | .8401 | 2.2e-16 | .8431 | 2.2e-16 |
| *Naïve accuracy = 0.78* | | | | |

This model can feasibly be used in real-world applications as well; data on humidity, temperature, and cloud cover at certain times can be fed into the model for a relatively accurate prediction on whether rain will be present tomorrow. However, as the model is used its quality may decline over time. At some point, a new model will need to be developed based on new data.