Rain prediction in Australia II

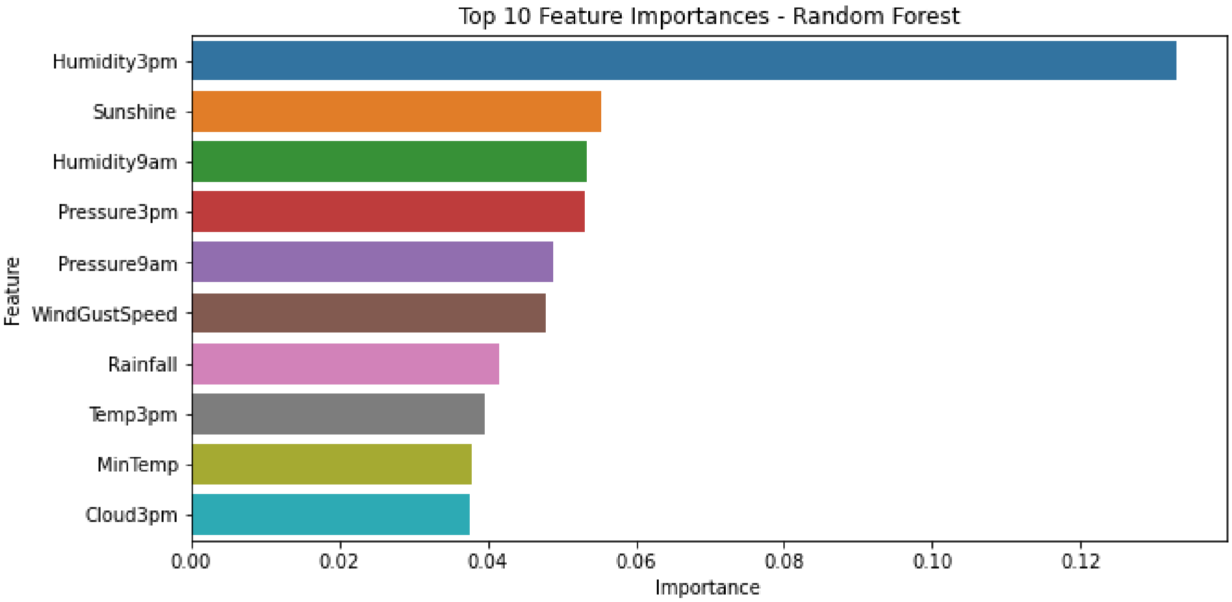
The dataset contains daily weather observations of Australian weather stations, and the goal here is to predict whether it will rain tomorrow based on the data. The target variable for prediction is “RainTomorrow”.

Analysis

In this analysis, I’ll focus on ensemble models and use voting, stacking and blending for comparing whether they improve the performance in prediction when compared to earlier models we used along with various forms of hyperparameter tuning.

I’ve picked a simple Logistic regression classifier with both L1 (Lasso) and L2 (Ridge) regularization, Random forest classifier, Gradient boosting classifier and Extreme Gradient Boosting classifiers as the base models. I did initially choose KNN model as well as one of the base models, however, currently having some convergence issues and errors that need to be fixed. Will later add it in a rerun.

The initial pre-processing of the dataset remains mostly the same as was done earlier during the use of various forms of hyperparameter tuning. Except that in the current analysis, I have added a feature importance visualisation using a Random Forest classifier model -

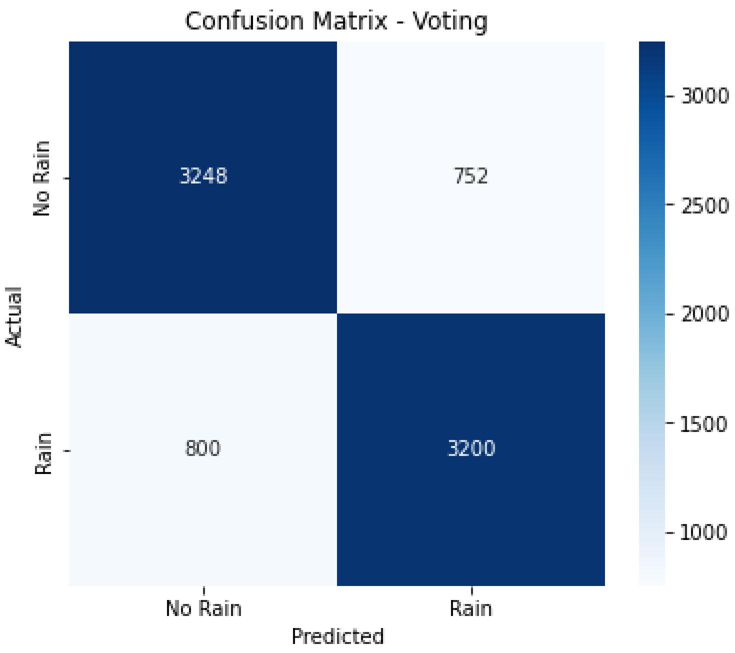


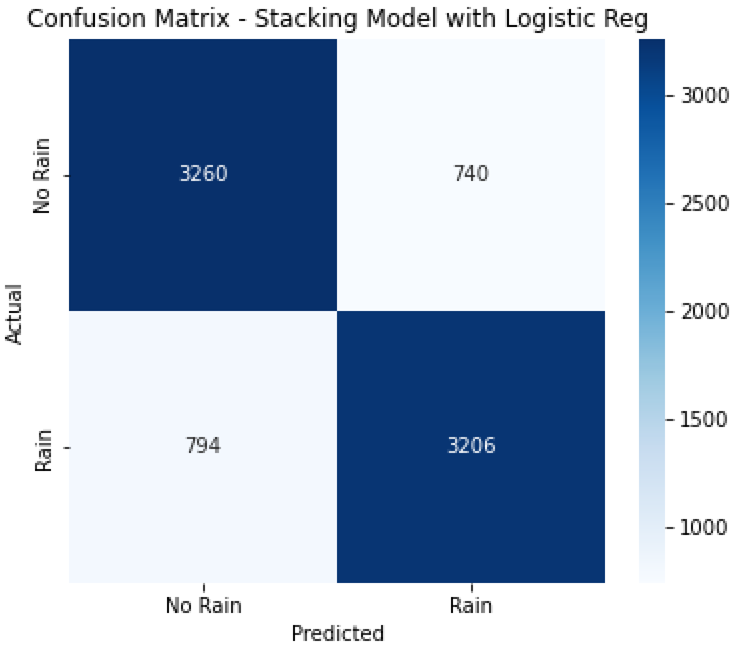
Something additional I’d like to add to this project in a future rerun would be some form of feature engineering using the Wind direction columns and some of the additional columns shown above to extract more information that could help in predicting better.

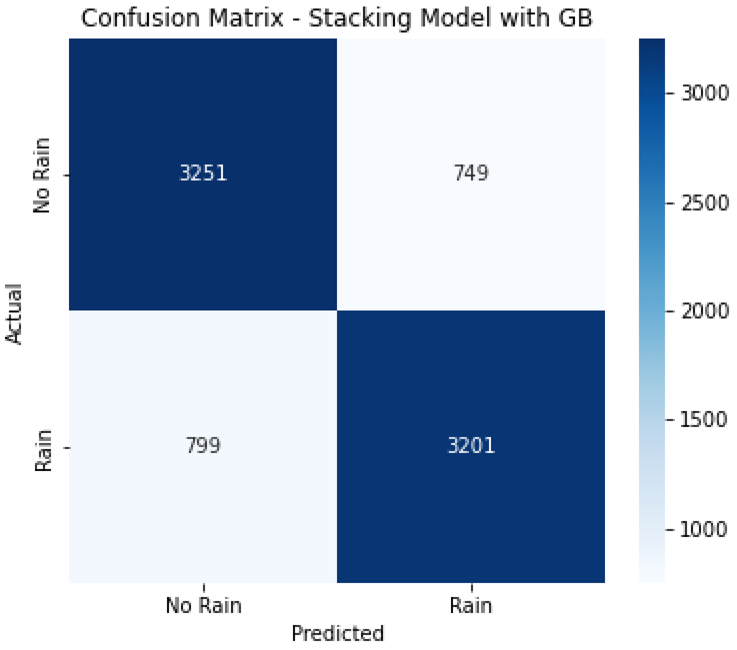
Model	Accuracy
Voting	80.6%
Stacking with Logistic regression as final estimator	80.83%
Stacking with Gradient Boosting as final estimator	80.65%
Blending with Logistic regression as final estimator	80.64%
Blending with Gradient Boosting as final estimator	80.44%

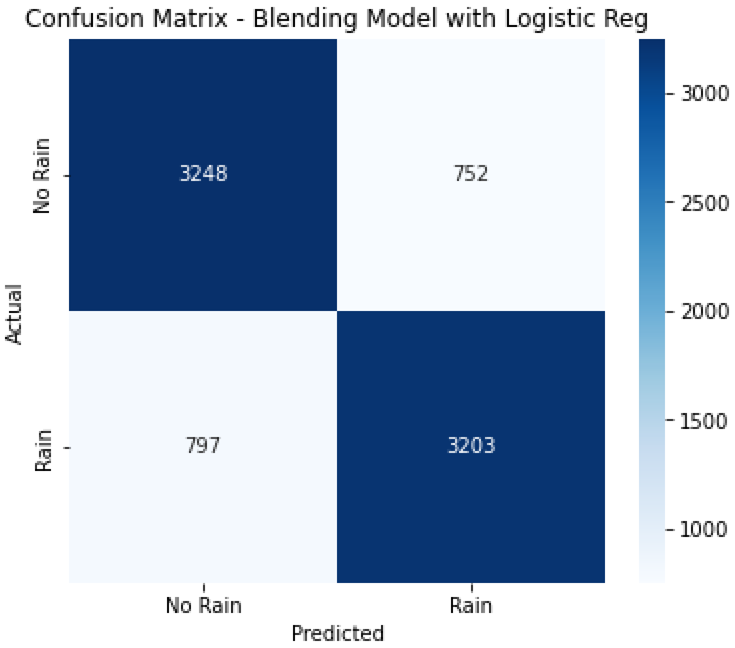
From the above table, we can see the accuracy scores across the various combinations of models are more or less identical with the Stacking model with Logistic regression model as final estimator performing marginally better than the rest.

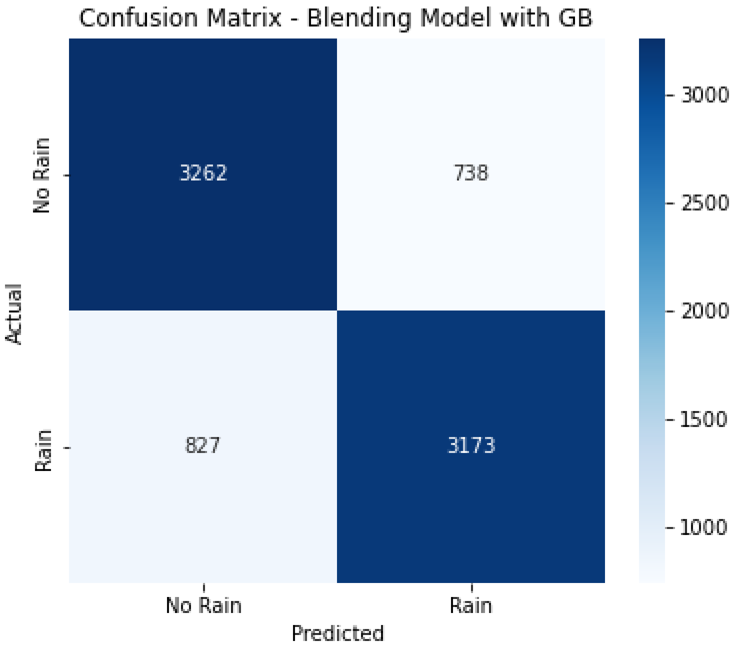
Finally, I have again consolidated the list of models with voting, stacking and blending and plotted the confusion matrix for each model as can be seen in the plots below-




Confusion Matrix - Stacking Model with Logistic Reg


Confusion Matrix - Stacking Model with GB


Confusion Matrix - Blending Model with Logistic Reg


Confusion Matrix - Blending Model with GB

Model	True Positives	True Negatives	False Positives	False Negatives
Voting Classifier	3200	3248	752	800
Stacking Model with Logistic Regression	3206	3260	740	794
Stacking Model with Gradient Boosting	3201	3251	749	799
Blending Model with Logistic Regression	3203	3248	752	797
Blending Model with Gradient Boosting	3173	3262	738	827

The voting classifier model performs well, but it has slightly more false negatives (missed rain predictions). The Stacking model with Logistic regression model as final estimator has the highest true positives and the lowest false negatives as well. The Stacking model with gradient boosting model as final estimator has in comparison both less true positives and true negatives, while having more false positives and false negatives.

The Blending model with Logistic regression model as final estimator is almost identical to the Voting Classifier, but with slightly fewer false negatives. The Blending model with Gradient Boosting as final estimator performs similarly but has the highest false negative count compared to the other combinations. It also has the highest true negatives.

.