Initial preprocessing

- Perform log-transformation on Evaporation, if using. Use one of Evaporation and LogEvaporation in the model, but not both, and see which one performs better
- May want to use Month (categorical/factor variable)

Chosen model

Choose Model 4 as the final model.

Evaluated performance on the test set using models/model4/evaluate_model.py (output in results-evaluate.txt).

- Accuracy: 85.53%
- Precision: 77.35%
- Recall: 47.97%

The models

Model 1

- Don't use Sunshine, Evaporation, Cloud9am, or Cloud3pm since these features have a lot of missing data in the training set
- Use RainToday categorical variable instead of Rainfall
- For remaining numerical features, impute missing data
- For remaining categorical features, use dummy indicator variables for each category; ignore (don't remove) if data missing
- Used 500 trees

Performance (files from models/model1): Ran build_feature.py to generate training and test data. Ran train_test_model.py to select a random forest classifier with 500 trees (output in results-train-test.txt).

5-fold cross-validation results:

- Accuracy: mean 0.8556 (85.56%), standard deviation 0.0008279
- Precision: mean 0.7715 (77.15%), standard deviation 0.004204
- Recall: mean 0.4816 (48.16%), standard deviation 0.008902

Model 2

- Use all or some of Sunshine, (Log)Evaporation, Cloud9am, Cloud3pm. Impute missing data
- Use RainToday categorical variable
- For remaining numerical features, impute missing data

- For remaining categorical features, use dummy indicator variables for each category; ignore (don't remove) if data missing
- Used 500 trees

Performance (files from models/model2): Ran build_feature.py to generate training and test data. Ran train_test_model.py to select a random forest classifier with 500 trees (output in results-train-test.txt).

5-fold cross-validation results:

- Accuracy: mean 0.8577 (85.77%), standard deviation 0.002935
- Precision: mean 0.7729 (77.29%), standard deviation 0.008674
- Recall: mean 0.4910 (49.10%), standard deviation 0.008869

Model 3

- Choose Model 1 (chosen from above using to cross-validation), but don't use months at all

Performance (files from models/model3): Ran build_feature.py to generate training and test data. Ran train_test_model.py to select a random forest classifier with 500 trees (output in results-train-test.txt).

5-fold cross-validation results:

- Accuracy: mean 0.8537 (85.37%), standard deviation 0.002018
- Precision: mean 0.7672 (76.72%), standard deviation 0.006683
- Recall: mean 0.4815 (48.16%), standard deviation 0.007381

Model 4

- Choose Model 1 (chosen from above using cross-validation), but use Rainfall instead of RainToday

Performance (files from models/model4): Ran build_feature.py to generate training and test data. Ran train_test_model.py to select a random forest classifier with 500 trees (output in results-train-test.txt).

5-fold cross-validation results:

- Accuracy: mean 0.8564 (85.64%), standard deviation 0.001447
- Precision: mean 0.7751 (77.51%), standard deviation 0.005532
- Recall: mean 0.4866 (48.66%), standard deviation 0.008417