Predicting the possibility of raining tomorrow in Australia

Hattan Alsayigh

Abstract

The goal of this project was to use classification models to predict the possibility of rain in the next day in Australia. I worked with data provided by Kaggle and the Australia Bureau of Meteorology. leveraging geographic and categorical feature engineering along with a random forest model to achieve promising results for this binary problem.

Design

The data is provided by Kaggle and the Australia Bureau of Meteorology., and presents a binary status of Raining (Yes or no), across the country. Classifying statuses accurately via machine learning models would enable the Australian government to take action to improve operations and maintenance of water drainage and emergency response, allocate resources more quickly to needed areas.

Data

The dataset contains 142193 Observations with 23 features for each, 7 of which are categorical. A few feature highlights include Max\Min temperature, The direction of wind at specific time, and the location of areas of Australia are covered.

Algorithms

Feature Engineering

1. Converting categorical features to binary dummy variables
2. Removing the Date column
3. Fixing the unbalanced data with over sampling
4. Populate missing data with median in the numerical features
5. Removing row that contain nan values in categorical features

Models

Logistic regression, Decision Tree Classifier, k-nearest neighbours, and random forest classifiers were used before settling on random forest as the model with strongest cross-validation performance. Random forest feature importance ranking was used directly to guide the choice and order of variables to be included as the model underwent refinement.

Model Evaluation and Selection

The entire training dataset of 142193 records was split into 75/25 train vs. test,

The official metric was classification rate (accuracy); however, class weights were included to improve performance against F1 score and provide a more useful real-world application where classification of the minority class (functional needs repair) would be essential.

Final random forest 5-fold CV scores:

- Accuracy 0.8598
- F1 0.624
- precision 0.7693
- recall 0.525
- AUC 0.8899

Tools

- Numpy and Pandas and SQL for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting