Research on the Applicability of Weather Forecast Model——Based on Logistic Regression and Decision Tree

Feiyang Deng

Department of statistics and management, Shanghai university of finance and economics,shanghai,200433,China

Abstract. With the continuous development of meteorological technology, weather forecast has become an indispensable part of human in agriculture and life. Regardless of whether it is a climate forecast on rainfall in agriculture or a weather forecast on mobile phones or TV, a variety of climate forecasts bring great convenience to people's production and life. For ordinary people, the prediction of whether it will rain the next day seems to be more practical. Many people determine the clothing and whether to bring an umbrella to go out for the next day by checking the weather forecast on mobile phones and TVs. It can be said that the prediction of rain directly affects people's living habits. On such basis, this article studies the question of whether it will rain tomorrow in life. Based on about 140,000 meteorological data from the Australian Meteorological Bureau, this article studies and analyzes the influential effect of whether it will rain tomorrow by establishing a logistic regression and decision tree model, and sets up a prediction model to predict whether it will rain tomorrow.

1. Introduction
In recent years, many scholars have made in-depth research on the topic of climate prediction, especially for the prediction of rainfall in the agricultural industry. Many models have been established, such as the time series ARIMA model and etc. The Guangzhou Institute of Tropical Marine Meteorology has developed a short-term climate prediction model for the South China region and simulated the form of climate prediction through weather maps. Some scholars have proposed a method that combines factor analysis and stepwise regression, which has reached a good long-term prediction level and can predict the planting time a long time in advance. At the same time, some scholars have combined principal component analysis with other methods such as SVM, and established a rainfall prediction model after reducing the dimensionality of the forecast factors. These models usually have good prediction effects in agricultural production applications, which proves to have certain practical value.

Although these studies have improved people's ability to predict rainfall to some extent, there is still room for further research. For example, most of the existing models are based on data statistics over several years or even decades, and mainly study the long-term change pattern of rainfall in agricultural production, instead of the question of whether rain will occur on a certain day in urban life like “Will it rain tomorrow?” mentioned in the article. In the long run, even if the rainfall shows a clear trend, it is still difficult to make a judgment about whether it will rain on a certain day based on this trend. In addition, the current prediction model can only predict the rainfall trend in a large area, but the daily weather is actually different for residents of different cities in this range. Such a model is difficult to judge the difference in rainfall in different cities.
2. Data and design of the model

2.1. Data collection and variable selection
In order to study the factors that affect whether it will rain tomorrow, the weatherAUS dataset downloaded from the kaggle website is used in this article. The data set contains the weather observation data recorded by the Australian Meteorological Agency every day since 2012. There are a total of 140,000 items, which include meteorological data such as the recording location, daily minimum temperature, maximum temperature, rainfall and target variables: will it rain tomorrow? Australia has a moderate territorial area, and the weather in each region has certain integrity and differences, so Australia is suitable as the data for model research. On the one hand, the model can be practical in a relatively large range, on the other hand, it will not lose accuracy in a relatively small area. Through a statistical analysis of the weatherAUS weather data obtained on the Kaggle website, this article screens out the influencing factors such as the recording location, the daily minimum temperature, the maximum temperature, and rainfall. Based on logistic regression and decision tree model, this paper aims to explore the factors influencing whether it will rain tomorrow.

Table 1 Introduction of part of Variables.

| ATTRIBUTE       | INSTRUCTION                                      | TYPE     |
|-----------------|--------------------------------------------------|----------|
| DATE            | Observation date                                 | Character type |
| LOCATION        | The common name of the weather station location  | Character type |
| MINTEMP         | Minimum temperature(°C)                         | Numerical type |
| MAXTEMP         | Maximum temperature(°C)                         | Numerical type |
| RAINFALL        | Rainfall recorded that day(mm)                  | Numerical type |
| EVAPORATION     | Evaporation amount from 0:00 am to 9:00 am(mm)   | Numerical type |
| RAINTOMORROW    | Will it rain tomorrow?                           | Boolean type |

2.2. Research model
The question of whether it rains studied in this paper is essentially a binary classification problem, and its variables include both numerical and categorical variables. In order to study the influencing mechanism of the dependent variable, it is necessary to study the importance of influence of each independent variable on the dependent variable. Logistic regression is good at linear analysis, able to grasp global laws and provide prediction probabilities, while decision tree is good at exploring internal laws, and can give rules for each step of classification. As a result, it is suitable for using the logistic regression to establish the regression model and adopting the decision tree to set up the classification model.

2.2.1. Logistics regression
Logistic regression is a generalized linear model, it is suitable for the condition that the dependent variable y only has two possible values. In other words, the distribution of the dependent variable is the Bernoulli distribution (or two-point distribution). Usually, 1 and 0 can be used to represent the two possible outcomes of the dependent variable. Since the value of the independent variable does not belong to [0,1], the independent variable undergoes the following logit transformation to make its value range in [0,1], which is called the logistic regression model.

$$f(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

$$E(y_i) = \frac{1}{1 + \exp\left(-\left(\beta_0 + \sum_{j=0}^{g} \beta_j x_{ij}\right)\right)}, \quad i = 1, 2, \ldots, n$$

Since y is 0-1 type Bernoulli random variable, its probability distribution is:
Therefore, the likelihood function of $y_1, y_2, ..., y_n$ is:

$$P(Y = y) = \prod_{i=1}^{n} P(Y_i = y_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}$$

Take the logarithm, it is:

$$L(y, \beta) = \sum_{i=1}^{n} y_i \left( \beta + \sum_{j=1}^{p} \beta_j x_{ij} \right) - \ln \left( 1 + \exp \left( \beta + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)$$

The $\hat{\beta}$ that makes $\ln (y, \beta)$ reach the maximum value is the maximum likelihood estimate of $\beta$.

### 2.2.2. Decision tree

Decision tree is a widely used classification model. It can extract a tree-type classification model from the given training samples. The decision tree is composed of nodes and directed edges. There are two types of nodes, internal nodes and leaf nodes. Each internal node in the tree records which attribute is used for classification, each branch represents the output of a judgment result, and each leaf node represents the result after the final classification. One internal node represents a feature or an attribute; one leaf node represents a classification.

The decision tree learning algorithm is usually a recursive selection of the optimal feature, which is based on segmenting the training data. This is the best classification process for each sub-data set, which corresponds to the division of the feature space and the construction of the decision tree.

Entropy of the decision tree: Entropy is a concept borrowed from information theory quantifying randomness or disorder. The set with high entropy value is quite diversified, and the decision tree expects to find a segmentation that can reduce the entropy value, so as to eventually increase the isotropy in the group.

For the given data segmentation $S$, the constant value $c$ represents the level of the category, $p_i$ represents the proportion of the case in the level $i$ falling into the category

$$Entropy(S) = \sum_{i=1}^{K} p_i \log_2 (p_i)$$

The decision tree calculates the homogeneity changes caused by the division of each possible feature, and this calculation becomes the information gain. For feature $F$, the calculation method of information gain is the entropy value of the data partition ($S_1$) before the division minus the entropy value of the data partition ($S_2$) generated by the division.

The Gini index of decision tree: The Gini index is an index used to select the optimal feature when the classic decision tree CART is used to classify problems. The classification tree uses the Gini index to select the optimal feature and determines the optimal binary cutoff point of the feature at the same time.

Assume there are $K$ categories, the probability of the sample point in the $k$ category is $p_k$:

$$Gini(p) = \sum_{k=1}^{K} p_k (1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

As to the binary problem, if the probability of the sample point belonging to the first category is $p$, the Gini index is:

$$Gini(p) = 2p(1-p)$$

The more messy the categories contained in the population, the greater the GINI index. Calculate the Gini index of all features and the cut points of the features when generating the classification tree. The one with the smallest Gini index is the optimal feature and its optimal cut point.

Decision tree’s pruning: The decision tree is a complex tree generated by taking full account of all data points, so there may be overfitting. Therefore, the decision tree should be pruned to reduce the
complexity of the tree and the probability of overfitting. Decision tree pruning refers to deleting all the child nodes of a subtree and using the root node as a new leaf node. There are two pruning schemes: pruning first and pruning later.

3. Empirical analysis

3.1. Pretreatment of the dataset

For independent variables, we looked at the data set and found that the number of missing values of the two variables Evaporation and Sunshine reached about half of the number of observations in the data set. Too many missing values are not conducive to analysis, so I decided to delete these two variables, as well as the two variables that are not related to the rain and the amount of rain on the next day. The missing values of the remaining numeric variables are complemented by the mean, and the missing values of the remaining sub-type variables are filled with the "Missing" type, so as to study whether the missing variables have an impact on the prediction.

For the dependent variable RainTomorrow, there are 31,877 "YES" categories and 110,316 "NO" categories, showing a certain imbalance. Our data set has a total of 140,000 observations, which belongs to the large sample category, so we have reason to use the Bayesian method to estimate the a priori probability of rain tomorrow by the sample is 0.225, and the a priori probability of not rain tomorrow is 0.775.

Next, in order to build the model, we randomly divide the data set into a training set and a test set at a 7:3 ratio.

3.2. Logistic regression

3.2.1. Model building

The maximum number of iterations is set to 100 by establishing a model through the training set, and part of the results are as follows.

| Estimate  | Std.Error | z value | Pr(>|z|) |
|-----------|-----------|---------|---------|
| (Intercept) | 52.0033056 | 1.9463975 | 26.718 | <2.00E-16 *** |
| LocationAlbany | -0.9470982 | 0.1072968 | -8.827 | <2.00E-16 *** |
| LocationAlbury | -0.3750377 | 0.0999073 | -3.754 | 0.000174 *** |
| LocationAliceSprings | -0.7173017 | 0.1254012 | -5.72 | 1.06E-08 *** |
| LocationBadgerysCreek | -0.432138 | 0.095264 | -4.536 | 5.73E-06 *** |
| LocationBallarat | -1.3209001 | 0.0986195 | -13.394 | <2.00E-16 *** |
| LocationBendigo | -0.6340318 | 0.1008554 | -6.287 | 3.25E-10 *** |
| RainTodayNo | -1.5623077 | 0.0789678 | -19.784 | <2.00E-16 *** |
| RainTodayYes | -1.0820242 | 0.0808727 | -13.379 | <2.00E-16 *** |

It can be seen that indicators such as area, wind speed, humidity, air pressure, cloud cover, temperature, and whether or not it was raining the previous day have a very significant impact on the rain, but the impact of wind direction on the rain is not very significant.

3.2.2. Model evaluation

The accuracy of the model on the training set and the test set reached 87.46% and 84.95%, respectively, indicating that the fitting effect of the model is very good, and no fitting problems have occurred. The confusion matrix in the test set is as follows.

| Predicted | Actual |
|-----------|--------|
| 31328     | 1817   |
In order to further verify the effect of the model, we draw the ROC curve of logistic regression in the test set as follows.

![Fig. 1. Logistic Regression ROC Curve.](image)

The AUC value reaches 0.871, indicating that the model has a rather good effect.

3.2.3. Summary

Although the fitting effect of the logistic regression model is good, it can be seen from the coefficient table that only a small number of variables have no significant effect on whether it will rain tomorrow. We can speculate that many variables greatly affect whether or not it will rain tomorrow, but this cannot explain which variables have played a key or even decisive role. Under such circumstances, logistic regression may no longer be suitable for exploring the influencing factors of whether it will rain tomorrow, so we further use the decision tree model.

3.3. Decision tree

3.3.1. Model building

Use the training set in the above preparation to build a decision tree model, use the 10-fold cross-validation method, set the minimum sample number of leaf nodes to 30, set the a priori probabilities to 0.775 and 0.225, respectively, and use information entropy for modeling. The tree structure is as follows.

![Fig. 2. Structure of the Decision Tree](image)

As can be seen from the tree structure diagram, there are fewer nodes in the tree structure. Unlike the results of logistic regression, only a few variables such as humidity, wind speed, and location play a decisive role in determining whether it will rain tomorrow.

At the same time, we can see from the tree structure that the greater the humidity at 3 pm, the easier it is to be judged to rain tomorrow; the greater the strongest wind speed, the easier it is to be judged to
rain tomorrow; the more humid and windy weather, the easier it is to rain rain. This is also consistent with our common sense, indicating that the results of the model have a certain degree of credibility. The most important variables in the decision tree model are Humidity3pm, Cloud3pm, Temp3pm, Humidity9am, MaxTemp, WindGustSpeed.

3.3.2. Model evaluation

The prediction results of the decision tree model on the training set and the test set have 83.24% and 83.32% accuracy, respectively, which is slightly lower than the logistic regression model. The fitting effect is better, and there is no overfitting phenomenon. The confusion matrix in the test set is as follows.

Table 4 Confusion Matrix of Decision Tree.

| Predicted | 31646 | 1499 |
|-----------|-------|------|
| Actual    | 5665  | 3848 |

In order to further verify the effect of the model, we draw the ROC curve of the decision tree in the test set as follows.

Fig.3. Decision Tree ROC Curve.

The area of AUC is 0.726, indicating that the effect of the model is slightly inferior to that of logistic regression.

4. Conclusion

The prediction accuracy of logistic regression and decision tree model is not much different, but the ROC area of logistic regression is slightly higher. In logistic regression, indicators such as area, wind speed, humidity, air pressure, cloud cover, temperature, and whether or not it rained the previous day have a very significant impact on the rain, while the decision tree focuses on the significant effects of humidity and wind speed, indicating that the logistic regression model has a better grasp of the overall data, and the decision tree has a better exploration of the local laws of the data. Therefore, it is more appropriate to use logistic regression in the actual application of large amounts of data.

References

[1] Yan Jinghua, Performance Analysis of Monthly Precipitation Prediction of Short-term Climate Model in South China [J], Guangdong Meteorology, 1998 (s2).

[2] Zheng Xiaohua, Li Ke, Qu Zhenjiang, Research on Precipitation Field Forecast from April to October Based on Land Surface Processes [C] //Collected Papers of the Weather Forecasting and Impact Assessment Technology Branch of the 2007 Annual Meeting of the Chinese Meteorological Society, 2007.
[3] Nongjifu, Application of Principal Component Analysis and Support Vector Machine in Regional Precipitation Forecast[J], Practice and Understanding of Mathematics (22): 91-96.