Weather Forecasting using Logistic Regression

Norma Quiroz

GitHub
Kaggle

Introduction

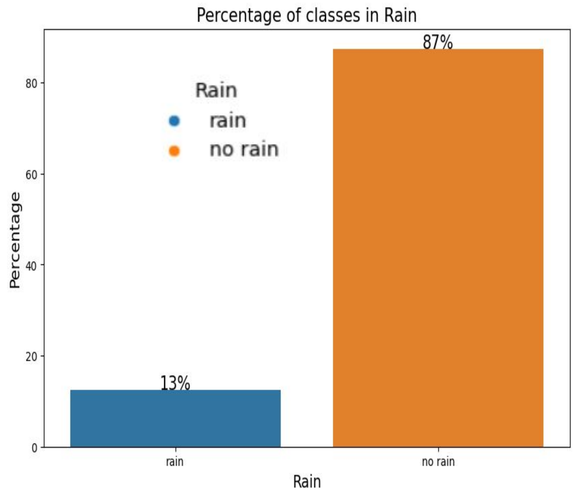
The Weather Forecast Dataset at Kaggle consists of 2500 observations with six weather conditions: Temperature, Humidity, Wind Speed, Cloud Cover, and Pressure.

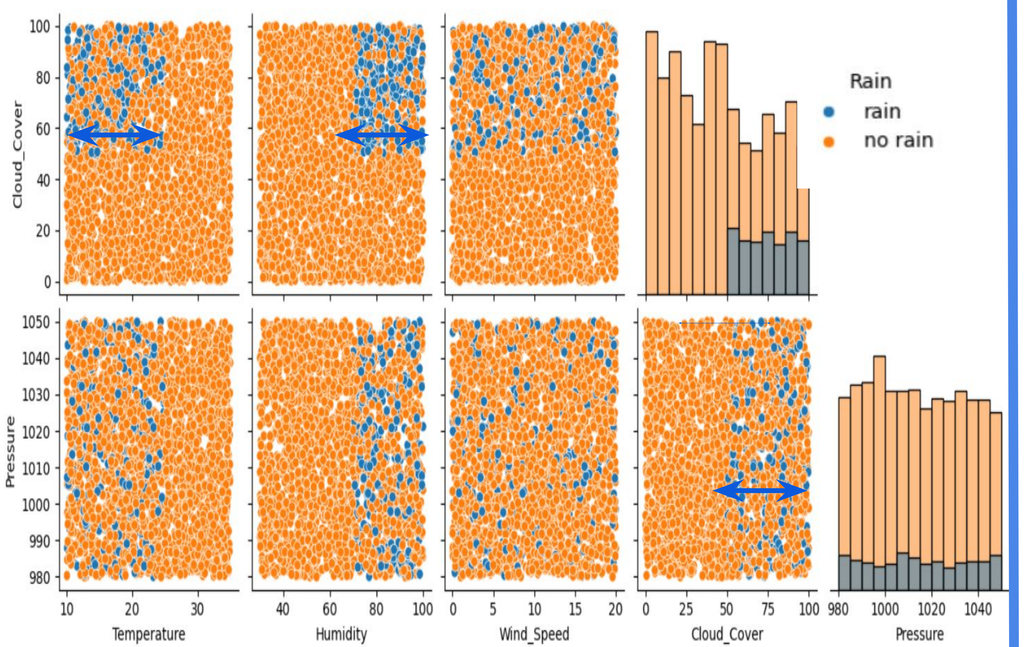
Objective: predict the rainfall based on these weather conditions by applying logistic regression.

Data Exploration

- ❑ Rain is a categorical variable with two classes.
- ❑ The rest of the variables are continuous.
- ❑ No missing values, no outliers.
- ❑ No correlation between variables.
- ❑ Scatter plots show the different range of values for the *Rain* classes in *Temperature*, *Humidity* and *Cloud Cover*.

- ❑ The *Rain* variable is slightly imbalanced.





Data processing

Features X **Target y**

	Temperature	Humidity	Wind_Speed	Cloud_Cover	Pressure	Rain
0	23.720338	89.592641	7.335604	50.501694	1032.378759	rain
1	27.879734	46.489704	5.952484	4.990053	992.614190	no rain
2	25.069084	83.072843	1.371992	14.855784	1007.231620	no rain

Logistic regression predictor

Considerations:

- ❖ In the context of weather, false negatives can be more costly than false positives. Hence, the best metrics to consider are recall and f1-score.
- ❖ The data set is a bit imbalanced.

A weighted logistic regression is suitable for these considerations.

The model predicting the probability of rain is given by the function

$$P(y_i = 1|X_i) = \frac{1}{1 + e^{3.76 + 2.1\,T - 3.01\,H - 0.06\,WS - 2.65\,CC + 0.02\,P}}$$

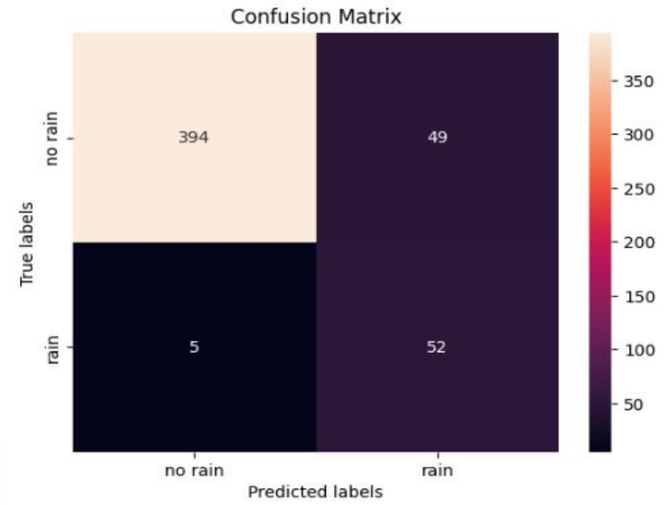
with coefficients

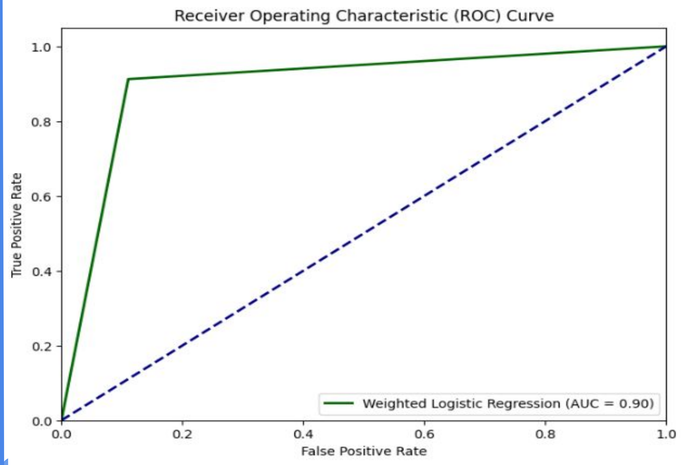
T	H	WS	CC	P
Temperature	Humidity	Wind_Speed	Cloud_Cover	Pressure
-2.102817	3.011339	0.061277	2.649962	-0.024458

Metrics

	precision	recall	f1-score
0	0.99	0.89	0.94
1	0.51	0.91	0.66
accuracy			0.89
macro avg	0.75	0.90	0.80
weighted avg	0.93	0.89	0.90

Model Evaluation





Conclusions

1. The weighted logistic model reduces the number of false negatives.
2. The model has an acceptable accuracy, but recall is much better.
3. Humidity is the most influential variable in the model due to its highest coefficient.
4. Cloud Cover and Temperature contribute significantly but Temperature does in the opposite direction. Wind Speed and Pressure contribute less.
5. The area under the ROC-AUC curve is 0.9 which shows the good performance of the model.