MACHINE LEARNING CHALLENGE

Unit Chair: Dr Musa Mammadov
Submission Date: 5:00PM Friday of Week 10

Table of Contents

Dataset Name: Rain in Australia

Group Name: Mon-13 (FANH) **On Campus/Cloud**: On Campus

STUDENT ID	STUDENT FULL NAME	Individual contribution*
218401269	ALEXANDER PAK YU LAI	
218241616	HARRY WILLIAM LODGE	
218459058	VIET NAM NGUYEN	
218271795	JARROD KENG YEN YONG	

* 5 - Contributed significantly, attended all meetings
4 - Partial contribution, attended all meetings
3 - Partial contribution, attended few meetings
2 - No contribution, attended few meetings
1 - No contribution, did not attended any meetings

Section 1: Brief Summary & ML Problem Formulation

Expectation:

- Bring in the main observations and conclusions that you could draw from the visualisations and analytics that was performed in group assignment 1.
- What are you going to achieve by applying Machine Learning methods on your dataset?
- What model have you decided to run on your dataset (ex: Classification/Clustering/Forecasting) and why?
- Step by step pictorial depiction of the machine learning process that you would run on your dataset. (Machine Learning Flowchart)

Start

Section 2: Results and Discussion

1. Classification Models

A. Feature Selection

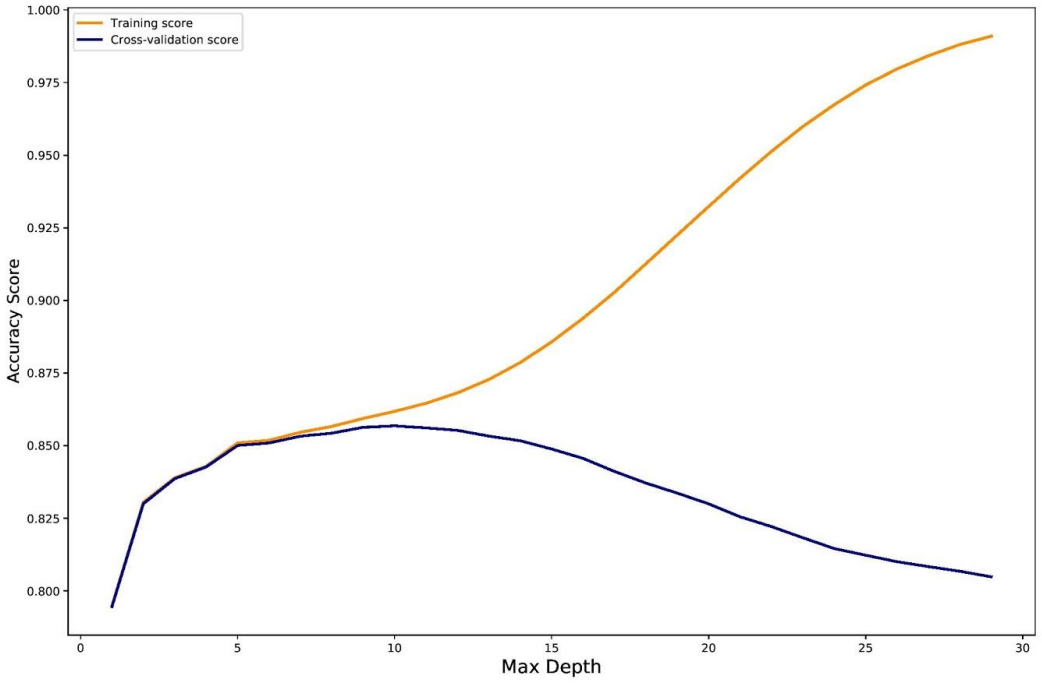
Mutual information between each feature and target (RainTomorrow) can be metrics to sort out the top features for improving model's performance and remove redundant variables. Thanks to the "sklearn" library, the mutual information of each pair between feature and "RainTomorrow" was calculated easily. As a result, the top five features with the highest mutual information scores were chosen to train and test models. They were "Humidity3pm", "Pressure9pm", "Pressure3pm", "Rainfall" and "WinGustSpeed".

B. Naive Bayes

Before the dataset fit to the Naive Bayes model, "Yes" and "No" in the target variable were assumed as 1 and 0 respectively. To avoid overfitting, the dataset with top five features above needed to be split where test and training set were in a ratio of one to four. After the model experienced the training process, the train-set and test-set accuracy scores were similar at 0.81. In addition, the precision and recall for class 0 were higher than class 1, which interpreted that the model predicted class 0 more correctly. Overall, the accuracy score was high, so it could be one of the suitable models to determine whether it will rain tomorrow or not.

C. Decision Tree

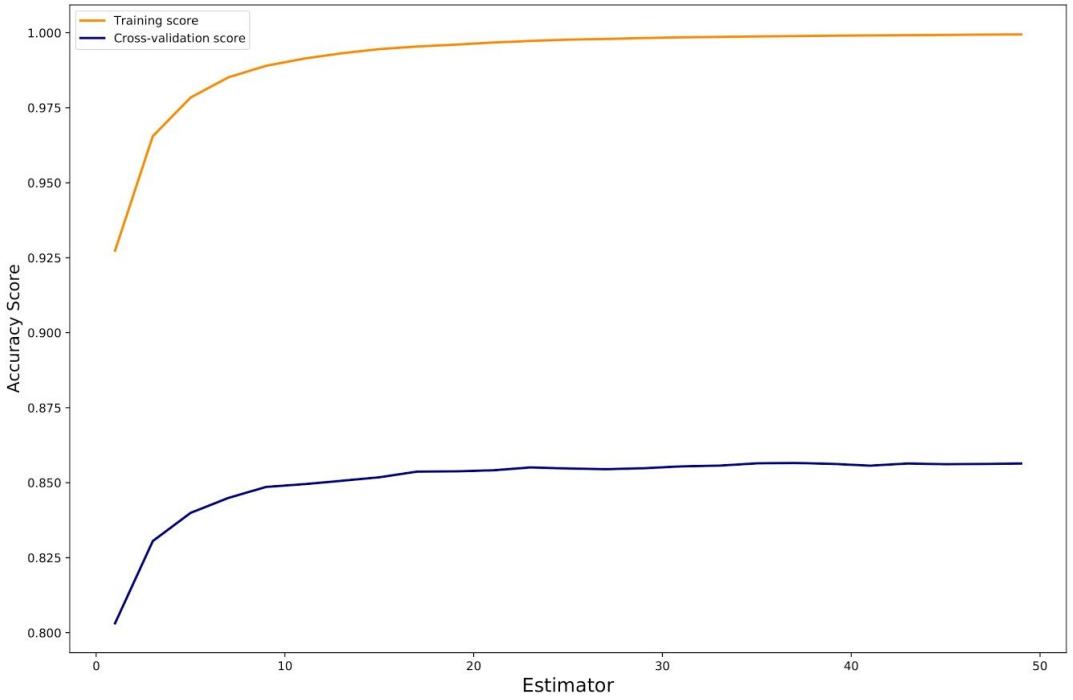
The Decision Tree model is able to process the categorical target, so encoding the target is unnecessary. However, since the numbers in most of the top five features were quite big, the dataset set with these features were applied "MinMaxScaler" technique. According to Hale's article (2019), "MinMaxScaler" reduces the range of each feature down to between 0 and 1 by taking the smallest value out from each value in the feature, and dividing the difference between the biggest and smallest in the feature. The dataset was split in the same strategy used in the Naive Bayes model. Since the Decision Tree model accepts the value of maximum depth as a hyperparameter, the most proper maximum depth can help the model not be overfit or underfit. Therefore, applying 10-fold cross validation on the training set with a range of maximum depth from 1 to 30 resulted in the best fit hyperparameter of 5 in depth.



The result of accuracy scores after the model was trained with value of 5 in depth was 0.85. In addition, the precisions of class “No” and “Yes” were 0.88 and 0.64, and the recalls of them were 0.91 and 0.57 in that order, so these scores illustrated the better prediction on class “No” than “Yes”. Because of high accuracy scores and no overfitting as well as underfitting, the Decision Tree can predict accurately the status of Australian weather tomorrow.

D. Random Forest

The Random Forest is quite similar to the Decision Tree, but the big difference is that it entails many individual decision trees, one of which is called as ensemble, and the best decision tree is selected as the final model (Yiu 2019). The beginning steps the model underwent are the same as the Decision Tree, such as “MinMaxScaler” standardization technique and data splitting with a ratio of 1 to 4. The most time-consuming part is to find the most suitable number of ensembles. This search was completed by 10-fold cross validation on the training set with the number of ensembles varying from 1 to 50.



The above graph shows no intersection between two lines of training and cross-validation. Similarly, they intended to run parallel to each other from 1 to possibly infinity. Therefore, the number of ensembles of 3 could be the best one to reduce overfitting as much as possible. Accordingly, the accuracy scores of the model on the training set and test were 0.96 and 0.83 respectively. Also, the precision and recalls stated that the model predicted better on class “No” than “Yes" since 0.88 and 0.64 were the precisions of class "No” and "Yes", and 0.91 and 0.57 were the recalls of these classes respectively. Overall, because of the big difference between train-set and test-set accuracy scores, the Random Forest model can return overfitting and less accurate prediction and the usage of this model should be considered carefully.

E. K-Nearest Neighbors

Start

F. XGBoost

Start

2. Regression Models

A. Linear Regression

Since the Linear Regression model is only able to manipulate on continuous variables, all of categorical variables were dropped out of the dataset. Furthermore, the entire dataset was standardized to the range from 0 to 1 by “MinMaxScaler” provided by “Sklearn” library. Before the model was fit to the dataset, the correlation coefficient of each feature and target was calculated to check whether the model had the ability of fitting this dataset. As a result, the highest correlation coefficient was only around 0.37. The consequence of giving the model a trial to fit this dataset was the low accuracy score, only approximately 0.21. In conclusion, the model was extremely underfitting, so it is not able to predict tomorrow Australian weather.

B. Logistic Regression

Logistic Regression can be responsible for both numerical and categorical features, none of features would be removed out of the dataset. To avoid unnecessary loss and time-consuming process, the entire dataset was standardized into the range of 0 to 1, and split into the training set and test set in a ratio of 4 to 1. As a result, the accuracy score gained after training the model was nearly 0.85. Besides, the metrics report stated that the model predicted perfectly on class “No” instead of “Yes” since the figure for class “No” was higher “Yes” in both precision and recall.

	precision	recall	f1-score
No	0.87	0.95	0.91
Yes	0.73	0.50	0.59

Briefly, the Logistic Regression performed well in predicting whether it will rain tomorrow in Australia.

3. Neural Network

The neural networks were rather simple in their construction and showed a high quality of results from the testing that was done. The way we created them was by first preparing the data with as many possible variables and then after testing we were able to remove the variables we suspected had no effect and could use this method to prove this. It was clear from the testing that the assumptions with wind direction were justified. The calculations done by this method was one of the most accurate with the best iteration having a weighted average having a value of 1.0 when using ‘riskmm’ and 0.85 without.

	precision	recall	f1-score	support
No	0.86	0.95	0.90	27524
Yes	0.74	0.46	0.57	8025
accuracy			0.84	35549
macro avg	0.80	0.71	0.74	35549
weighted avg	0.83	0.84	0.83	35549

Base(contains wind direction, no risk mm)

	precision	recall	f1-score	support
No	0.87	0.94	0.90	27651
Yes	0.70	0.51	0.59	7898
accuracy			0.84	35549
macro avg	0.79	0.72	0.75	35549
weighted avg	0.83	0.84	0.83	35549

Contains no wind direction or risk mm

	precision	recall	f1-score	support
No	1.00	1.00	1.00	27606
Yes	0.99	1.00	1.00	7943
accuracy			1.00	35549
macro avg	1.00	1.00	1.00	35549
weighted avg	1.00	1.00	1.00	35549

Contains both wind direction and risk mm

4. Suggestions

Start

Expectation:

- Performing your ML tasks (feature selection, data classification or clustering, finding the set of best (top) features).
- Reporting your performance evaluation metrics.

Section 3: Conclusions

Expectation:

- Explain what you could achieve by running ML models on your dataset and was it helpful to solve your problem?
- Any suggestion on improving your model and achieving better results?

Section 4: References

Hale, J 2019, *Scale, Standardize, or Normalize with Scikit-Learn*, Towards Data Science, Medium, retrieved 16 May 2020, <https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02>.

Yiu, T 2019, *Understanding Random Forest*, Towards Data Science, Medium, retrieved 16 May 2020, <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>.

Young, J 2017, *Rain in Australia*, Kaggle, retrieved 18 April 2020, <https://www.kaggle.com/jsphyg/weather-dataset-rattle-package>.