Course Project

Moon, Hyosik (test)

July 30, 2021

1 Required

- **Main objective of the analysis**
  Predict next-day rain by training classification models on the target variable RainTomorrow. We will predict the next-day rain by using three different models such as Logistic Regression, Support Vector Machine (SVM), and Random Forest with cross validation. And analyze the models accuracy and coefficients to find the impactful features.

- **Brief description of the data set**
  This dataset contains about 10 years of daily weather observations from many locations across Australia. RainTomorrow is the target variable to predict. This column is Yes if the rain for that day was 1mm or more. In data set, there are 23 columns and there are 16 floats and 7 objects columns. 1

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 145460 entries, 0 to 145459
Data columns (total 23 columns):
 #   Column         Non-Null Count   Dtype  
---  ------         --------------   -----  
 0   Date           145460 non-null  object 
 1   Location       145460 non-null  object 
 2   MinTemp        143975 non-null  float64
 3   MaxTemp        144199 non-null  float64
 4   Rainfall       142199 non-null  float64
 5   Evaporation    82670 non-null   float64
 6   Sunshine       75625 non-null   float64
 7   WindGustDir    135134 non-null  object 
 8   WindGustSpeed  135197 non-null  float64
 9   WindDir9am     134894 non-null  object 
 10  WindDir3pm     141232 non-null  object 
 11  WindSpeed9am   143693 non-null  float64
 12  WindSpeed3pm   142398 non-null  float64
 13  Humidity9am    142806 non-null  float64
 14  Humidity3pm    140953 non-null  float64
 15  Pressure9am    130395 non-null  float64
 16  Pressure3pm    130432 non-null  float64
 17  Cloud9am       89572 non-null   float64
 18  Cloud3pm       86102 non-null   float64
 19  Temp9am        143693 non-null  float64
 20  Temp3pm        141851 non-null  float64
 21  RainToday      142199 non-null  object 
 22  RainTomorrow   142193 non-null  object 
dtypes: float64(16), object(7)
memory usage: 25.5+ MB
```

Figure 1: Data set information

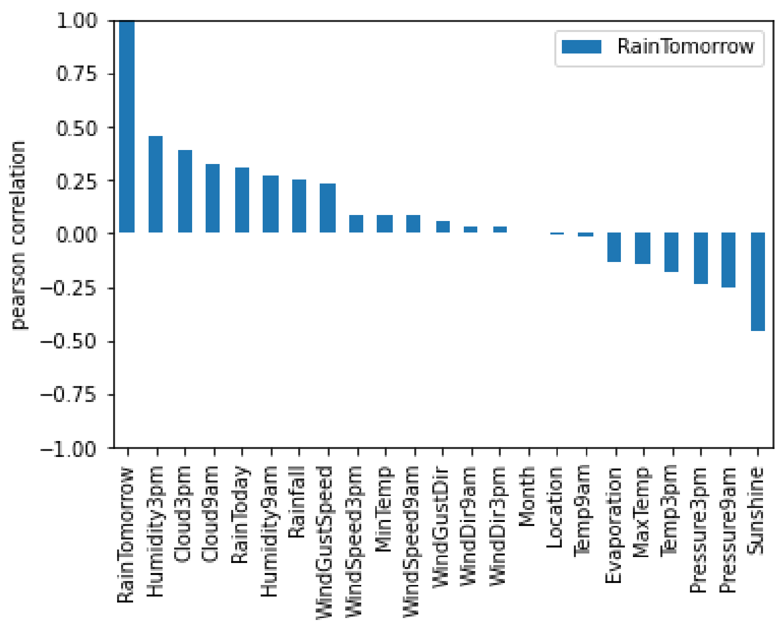
- **Brief summary of data exploration**

  1. Data cleaning. First, delete NaN data. Second, delete unused features (columns). We can delete Date column but it can be useful if we use 'month' to predict the RainTomorrow. So, we will change the Date column to month rather than delete the column.
  2. Correlation. To see the correlation, we have to change categorical variables to numeric variables. So, let's change the objects to numeric by using LabelEncoder. Find the correlation between RainTomorrow and other features (Figure 2a, 2b). We don't know what is the most impactful feature to predict Rain-

Tomorrow even though we can refer to the correlation. So, before we find the important features, we will use the all features.

	RainTomorrow
RainTomorrow	1.000000
Humidity3pm	0.455358
Cloud3pm	0.388574
Cloud9am	0.323972
RainToday	0.309098
Humidity9am	0.271033
Rainfall	0.254342
WindGustSpeed	0.233158
WindSpeed3pm	0.088862
MinTemp	0.087428
WindSpeed9am	0.083904
WindGustDir	0.061751
WindDir9am	0.035992
WindDir3pm	0.032203
Month	0.001046
Location	-0.005100
Temp9am	-0.018179
Evaporation	-0.130002
MaxTemp	-0.147467
Temp3pm	-0.183586
Pressure3pm	-0.230418
Pressure9am	-0.254816
Sunshine	-0.453407

(a) Correlation



(b) Correlation bar plot

3. Scaling. We also need to scale the features to use the Logistic Regression and SVM. So, we will use MinMaxScaler.

	count	mean	std	min	25%	50%	75%	max
Location	56420.0	0.505153	0.292049	0.0	0.280000	0.520000	0.760000	1.0
MinTemp	56420.0	0.529259	0.168417	0.0	0.401575	0.522310	0.658793	1.0
MaxTemp	56420.0	0.457255	0.158424	0.0	0.331818	0.450000	0.581818	1.0
Rainfall	56420.0	0.010332	0.034020	0.0	0.000000	0.000000	0.002910	1.0
Evaporation	56420.0	0.067773	0.045521	0.0	0.034483	0.061576	0.091133	1.0
Sunshine	56420.0	0.533491	0.259183	0.0	0.344828	0.593103	0.737931	1.0
WindGustDir	56420.0	0.499036	0.319487	0.0	0.200000	0.533333	0.800000	1.0
WindGustSpeed	56420.0	0.277194	0.115959	0.0	0.191304	0.260870	0.339130	1.0
WindDir9am	56420.0	0.474862	0.310722	0.0	0.200000	0.466667	0.733333	1.0
WindDir3pm	56420.0	0.504962	0.314113	0.0	0.200000	0.533333	0.800000	1.0
WindSpeed9am	56420.0	0.210265	0.127954	0.0	0.107692	0.200000	0.276923	1.0
WindSpeed3pm	56420.0	0.240362	0.115002	0.0	0.148649	0.229730	0.324324	1.0
Humidity9am	56420.0	0.658741	0.185133	0.0	0.550000	0.670000	0.790000	1.0
Humidity3pm	56420.0	0.496020	0.201970	0.0	0.350000	0.500000	0.630000	1.0
Pressure9am	56420.0	0.613347	0.115348	0.0	0.537563	0.612688	0.689482	1.0
Pressure3pm	56420.0	0.609961	0.111179	0.0	0.533981	0.608414	0.684466	1.0
Cloud9am	56420.0	0.530213	0.349645	0.0	0.125000	0.625000	0.875000	1.0
Cloud3pm	56420.0	0.480724	0.294139	0.0	0.222222	0.555556	0.777778	1.0
Temp9am	56420.0	0.471445	0.163790	0.0	0.344140	0.461347	0.598504	1.0
Temp3pm	56420.0	0.448357	0.161239	0.0	0.323113	0.441038	0.570755	1.0
RainToday	56420.0	0.220879	0.414843	0.0	0.000000	0.000000	0.000000	1.0
RainTomorrow	56420.0	0.220259	0.414425	0.0	0.000000	0.000000	0.000000	1.0
Month	56420.0	0.493183	0.313762	0.0	0.181818	0.454545	0.727273	1.0

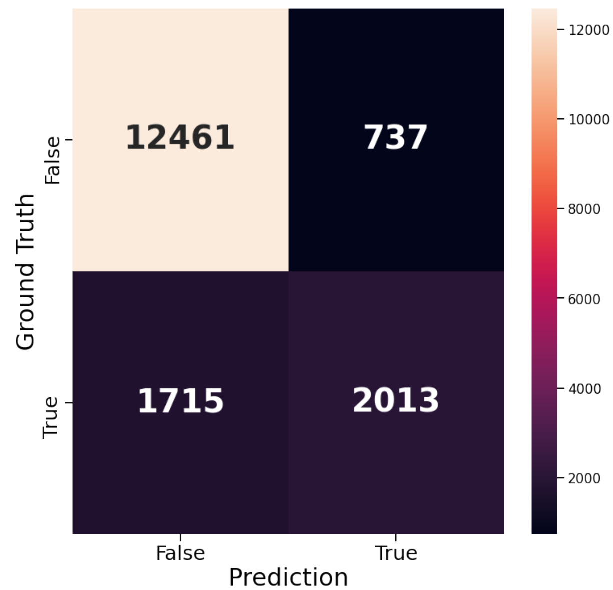
Figure 2: MinMaxScaling

- **Summary of training at least three linear regression models** I implemented three different prediction models which are Linear Regression, Support Vector Machine, and Random Forest. And in order to regularize the models I also used cross validation.

  1. LinearRegressionCV. The weighted f1-score and accuracy of the model were 0.846, 0.855 respectively. I trained the model by using stratified shuffle split because the target was skewed to 0 (No rain) (Figure 3a, 3b).

	lr_l1
precision	0.846639
recall	0.855134
fscore	0.846787
accuracy	0.855134
auc	0.742063

(a) Evaluation metrics for LRCV

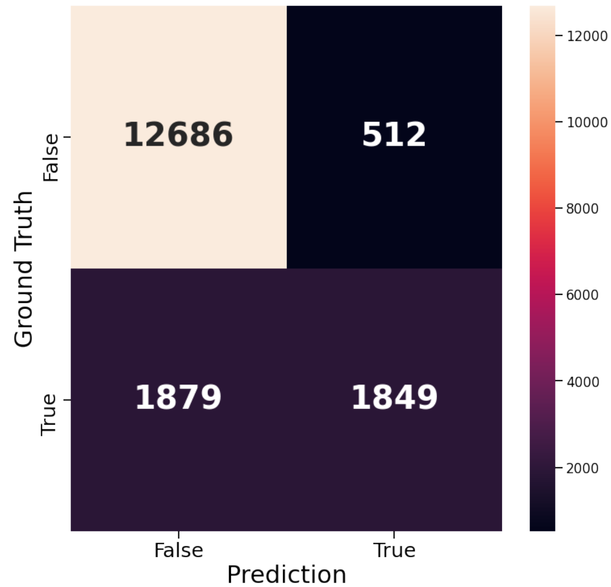


(b) Confusion metrics

- Support Vector Machine. The weighted f1-score and accuracy of the model were 0.846, 0.858 respectively. As a kernel function, I used rbf (Radial Basis Function) (Figure 3c, 3d).

	SVC_Gaussian
precision	0.851643
recall	0.858738
fscore	0.846359
accuracy	0.858738
auc	0.728591

(c) Evaluation metrics for SVM_Gaussian

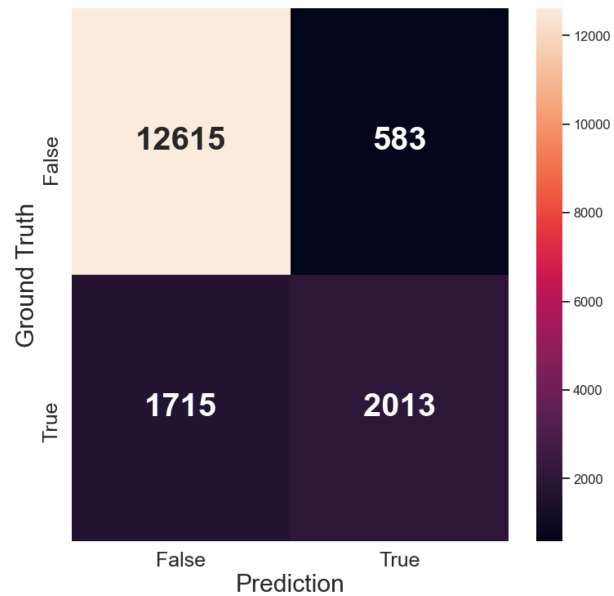


(d) Confusion metrics

- Random Forest. The weighted f1-score and accuracy of the model were 0.854, 0.864 respectively, and to find the best number of trees, out-of-error scores were used. And then the model use the 400 trees to predic the RainTomorrow (Figure 3e, 3f, 3).

	Random_Forest
precision	0.857217
recall	0.864233
fscore	0.854873
accuracy	0.864233
auc	0.747897

(e) Evaluation metrics for Random Forest



(f) Confusion metrics

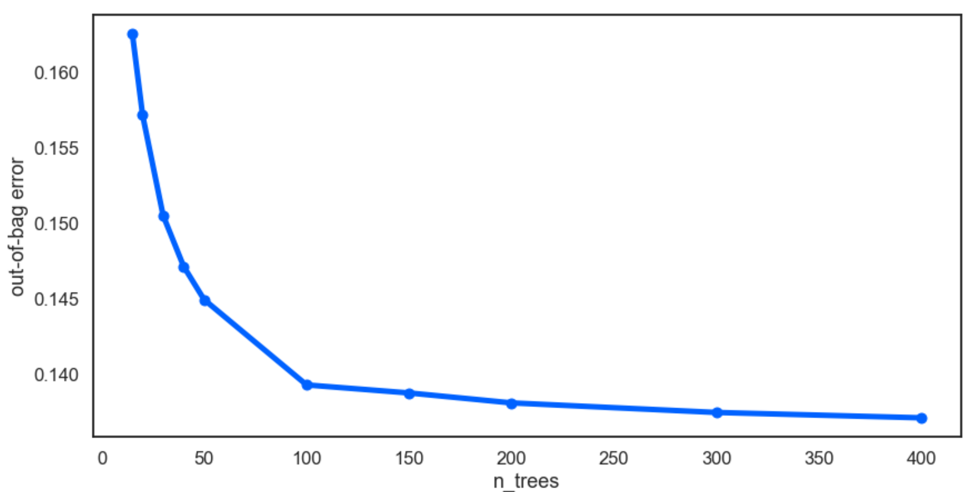


Figure 3: Out of error scores

- **Explanation of your final regressions model** Overall, all models showed the similar f1-score and accuracy. The best one was Random Forest which has the fscore 0.854873 and accuracy 0.864233. The ROC, Precision_Recall curve accuracies aren't quite ideal. This is because the RainTomorrow is unbalanced. For example, 'No (rain)' takes about 77%, 'Yes (rain)' takes about 23%. As a result the model predict the RainTomorrow relatively close to 'No (rain)' even though the ground truth is 'Yes (rain)'. It leads to relatively low ROC and Precision_Recall curve accuracies (Figure 4).

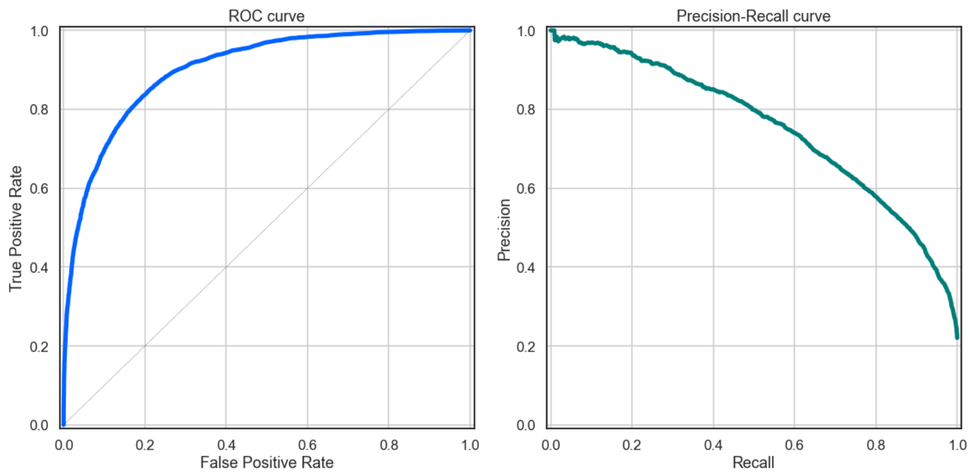


Figure 4: ROC, Pricision-Recall curves

- **Summary Key Findings and Insights** The fact that the accuracies among the models are similir means that the reliability of the models is also high. And the three most important features to predic the RainTomorrow are 'Hymidity3pm, Sunshine, and Pressure3pm'. These features are similir to the features that had high correlation values (Figrue 5). As a result, we can conclude that 'Hymidity3pm, Sunshine, and Pressure3pm' are the most important factores to predict the tomorrow's rain and with more features we can predict the tomorrows rain with 86% accuracy.

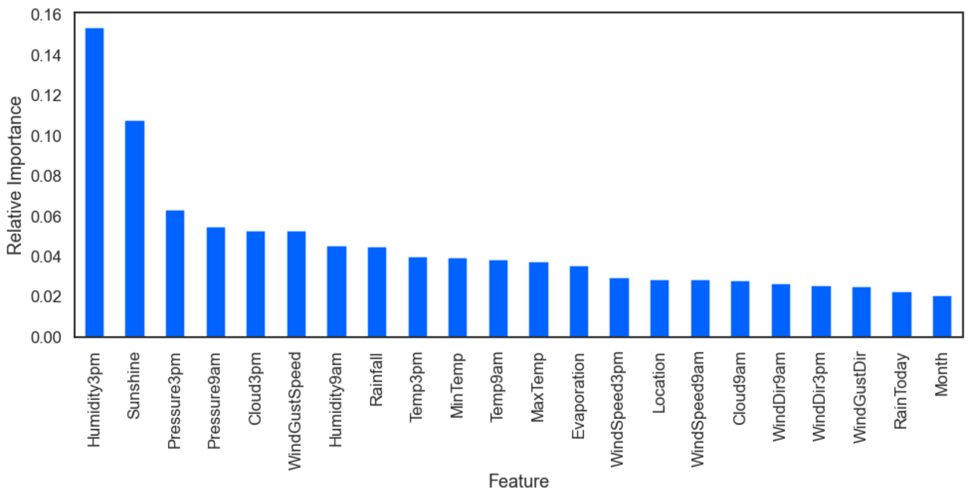


Figure 5: Feature importance

- **Suggestions for next steps** If the data were balanced, then we might predict more accuracte data. So, manipulating the unbalanced data will be a good-next step to imporve the model.