Predict Rain Tomorrow in Australia

1 . Overall Goals/Research Hypothesis:

Most probably one might be interested in knowing whether it’s going to be Rain Tomorrow or not. Due to various reasons, for example when to plan travel, what kind of dress to wear, whether to carry the umbrella etc. Indeed, in the today’s world, there are so many weather apps are available to do this job. But as a data science student, I am curious on how these apps predicting whether it’s going to be rain tomorrow or not.

The goal of this paper to predict whether it is going to be Rain Tomorrow or not in Australia. The parameters used to predict the outcome are location, temperature, humidity, rainfall, evaporation, pressure, cloud, rain today etc.

Research Hypothesis and challenges:

As per Kaggle dataset suggestion risk mm feature has been ignored for the analysis.

The reason this as stated below in the website discussion forum:

Since risk mm contains information about the future, and since it contains information directly about the target variable, including it would leak the future information to your model. Using it as a predictor to build a model and then testing on this dataset would give the false appearance of a high accuracy.

* 1. Unbalanced target Rain tomorrow

As the dataset is unbalanced the data has divided such a way that both training data and test data would have same ratio of Yes and No’s

* 1. Missing data

In general, missing values are replaced with either mean and median values. But after inspecting the data, and it is wide spread in different locations. The replacing values are very far from true values. Due to this, missing records are omitted though the dataset has more than 52000 records.

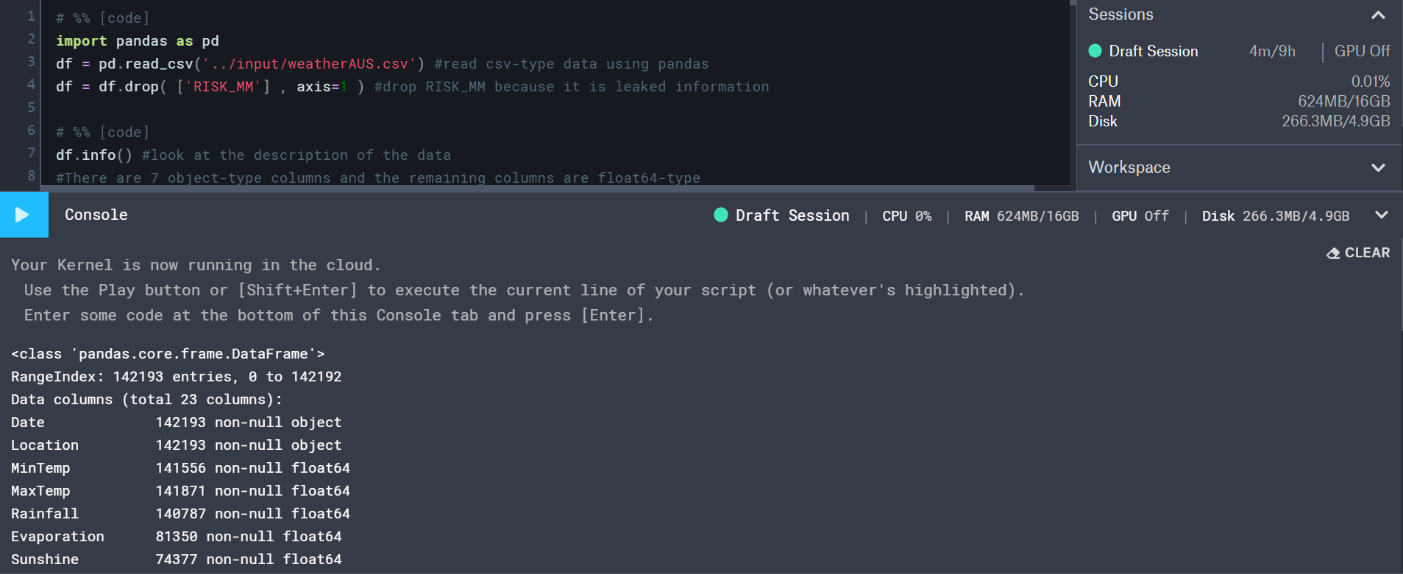
* 1. Numerical and Categorical variables

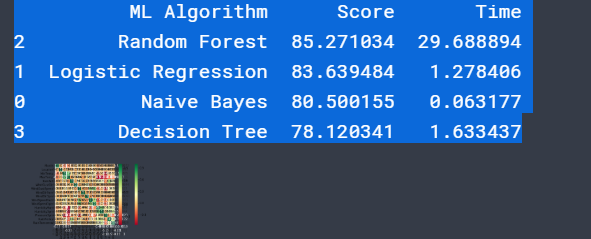
This is handled by converting text data to numeric data and by creating a dummy variable for each category

* 1. More number of features

After doing some research identified PCA methods to reduce the dimensionality and applied to the dataset. Analysis done for full model vs simplified model.

2. Previous/Related Contributions:





The resource for the previous work [1]:

The code and log file are placed in zip folder. The previous people work has helped a lot in analyzing the dataset. In resource python script has been implemented for a four different kind of algorithms Random forest, logistic regression, naive Bayes and decision tree respectively.

Some key things observed from previous work are:

1. Missing values are replaced with mean values.
2. Extracted month from the date and used for analysis.
3. Evaporation, Sunshine, Cloud9am, Cloud3pm, Risk\_MM are dropped from data.
4. Train data and test data re 75% and 25% respectively.
5. Unbalanced target issue has not been addressed
6. Random Forest has shown the best prediction compared to others but took relatively took more time compared to others.

Literature Review:

In addition to the study the previous work, also did some literature review to know the current practices for weather predictions as well as rainfall predictions.

The literature review carried out can be broadly classified into following categories [5]:

* 1. Missing Data
  2. Probability distributions
  3. Development of regression relationships
  4. Characteristics of climate data

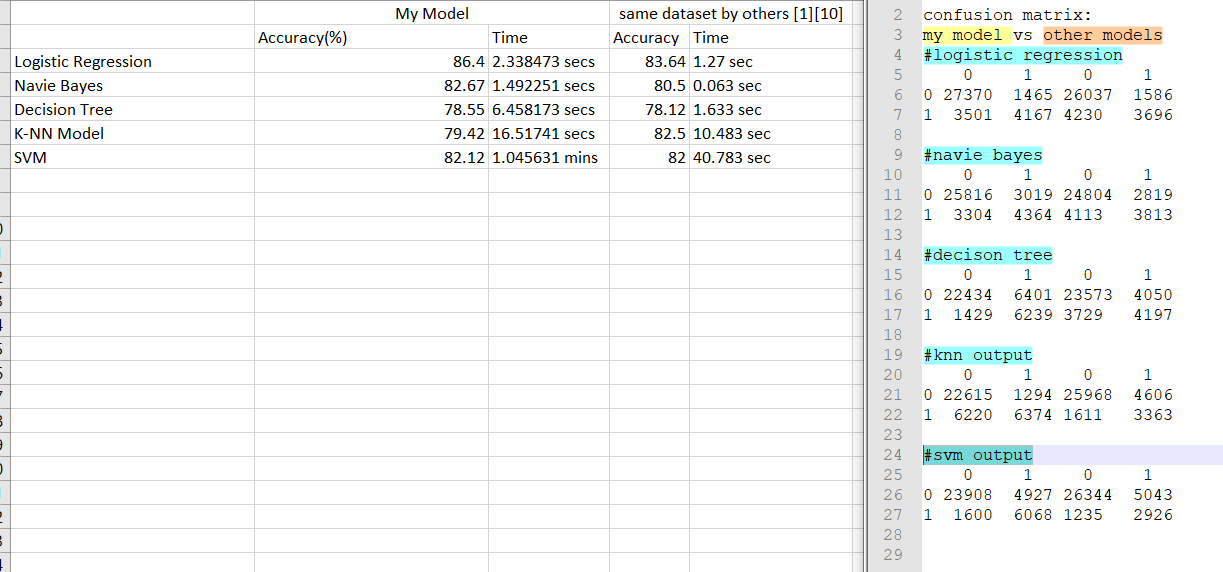
To predict the weather [2] numerical model, trends, persistent methods, climatology method and analog method are some of the techniques currently listed in literature. Numerical models are based on mathematical simulations. In this several models are tested and come up the best possible decision. Trends method focus on how the weather is currently moving. Persistence method weather today is going to be same as the weather tomorrow. Climatology method is similar to the persistence method but takes more depth look at the past to predict the future. Analog method finds the time similar to the current time in past and predict based on those results.

An active research is going on to predict the rainfall using microelectromechanical sensors [3] (MEMS) as the prediction is vital for flood disaster prediction and prevention. Also, case base expert systems, artificial neural network and a combination of both techniques can be utilized to predict the weather and rainfall [4].

As per literature review understood that Machine learning techniques adoption evolving in a rapid manner and the exponential growth can be seen in the near future. The following machine learning algorithms are found used by different users for the same data set analysis in the Kaggle website [6] and other resources:

* Random Forest
* Logistic Regression
* Naive Bayes
* Decision tree
* Linear Discriminant Analysis
* K neighbors’ classifier
* XGB classifier
* Neural Network models [7]
* Support Vector regression model [8]
* Deep Learning techniques [8]

**3. Comparison study:**



A comparison study has done for the full model and done by others. The accuracy and execution time are compared.

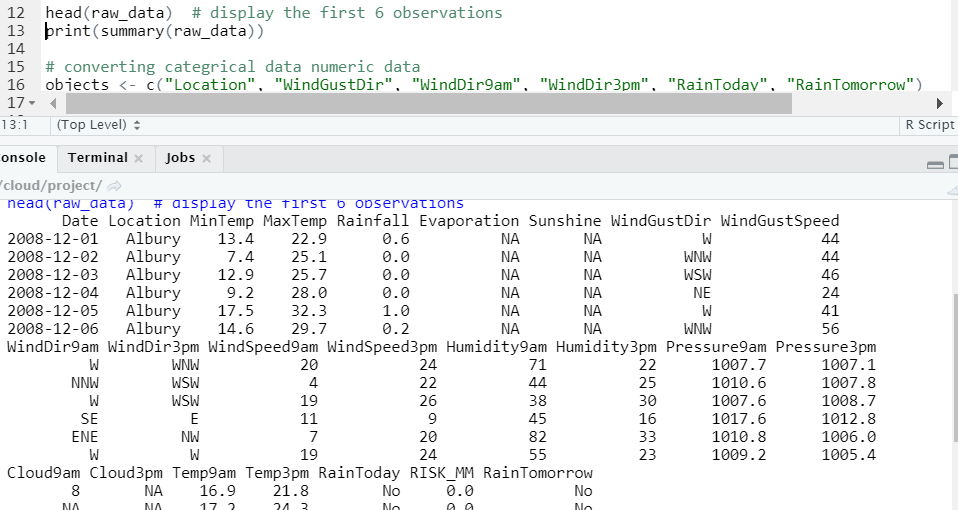
* In my model logistic regression stand out with accuracy followed by Naive Bayes, K-NN model, SVM and decision tree respectively.
* My model took a longer time when compared to others because of mainly two reasons. Those are executing scripts in different environment. My models are run R environment, while others for some model in python and others in R. And other reasons plot Roc and calculating statistics like AUC, recall, precision, sensitivity and specificity calculation time also included.
* Among all SVM model tool the relatively very long time compared to others and able to give about 82% accuracy.
* Confusion metrics are also compared with models developed by others
* In their models the NA values are replaced with the mean, but in our model, those are omitted. This decision seems to be correct as my model produced more accuracy than the others except for the K-NN model
* Some of the model selected the features using dimensionality reduction techniques.
* Decision tree model stand out less accurate in both models, which tell us it’s not a good method for this particular dataset
* In my research most of them applied logistic regression for this dataset, a few members applied the Random Forest method which produces more accurate results than logistic regression.

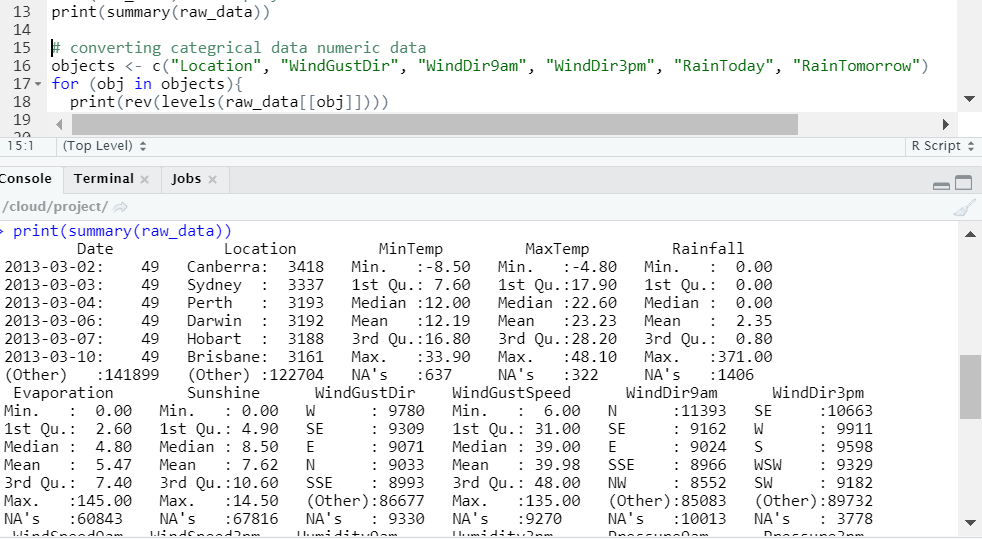
**Preprocessing activities, Features Selection / Engineering:**

Prepressing:

At first by inspect head, summary other data identified that the data has both numerical and categorical variables as well as some missing data. As stated earlier missing data records are omitted. Categorical variables are converted to numeric numbers and applied factor to those features so that model identify them as categorical variables.

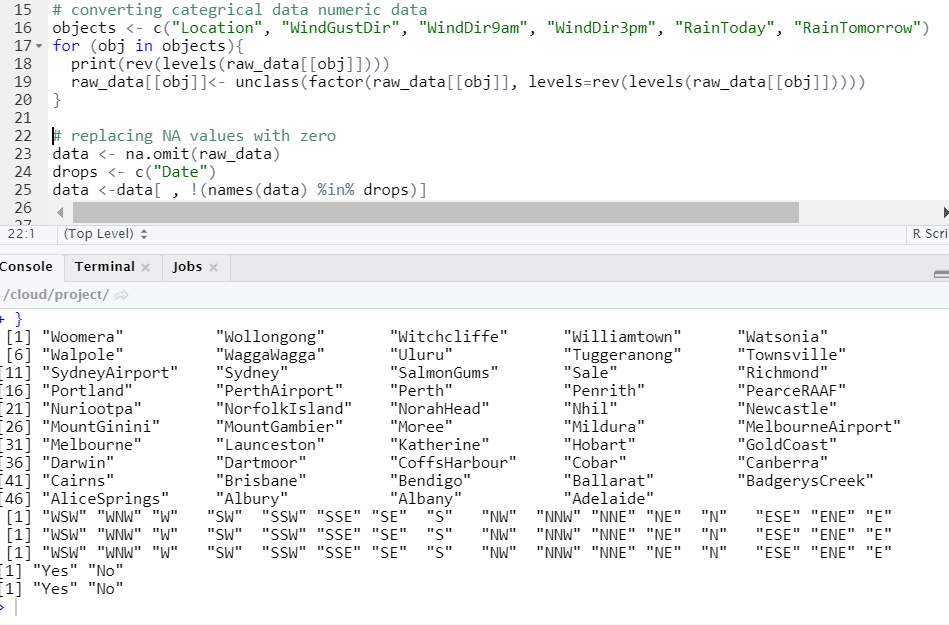
RISK\_MM is discarded as it would lead data leak.



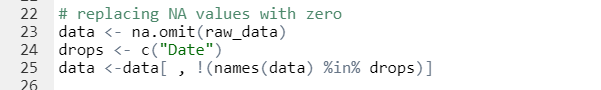


Inspection of head and summary shows us that

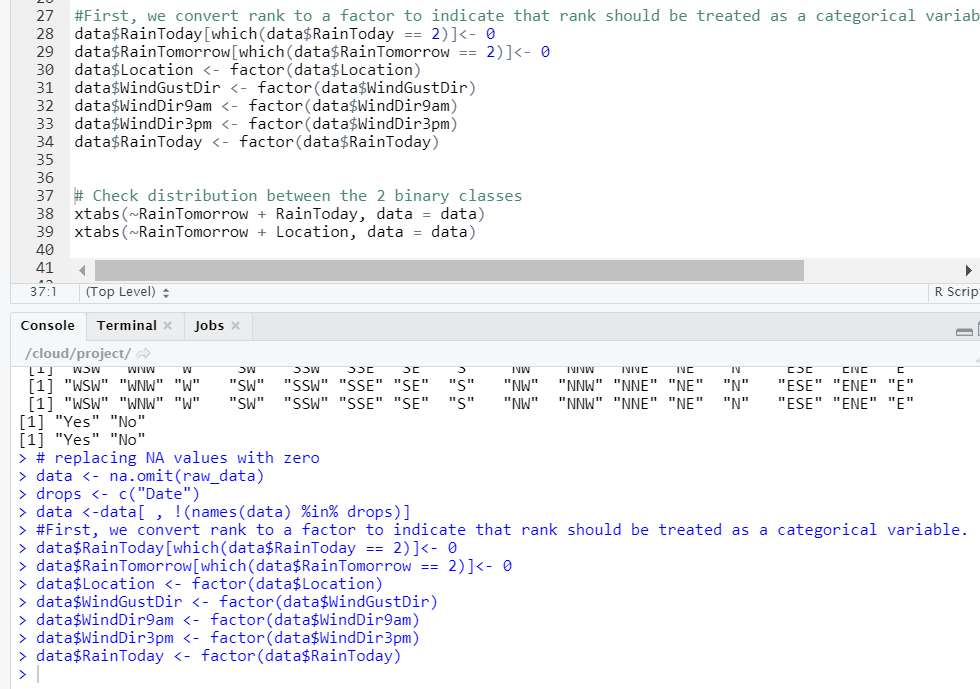
* dataset has both categorical features and numerical features
* some values as NA



So categorical objects are converted to Numerical. The respective numbers set to the object values can be seen in the console output.

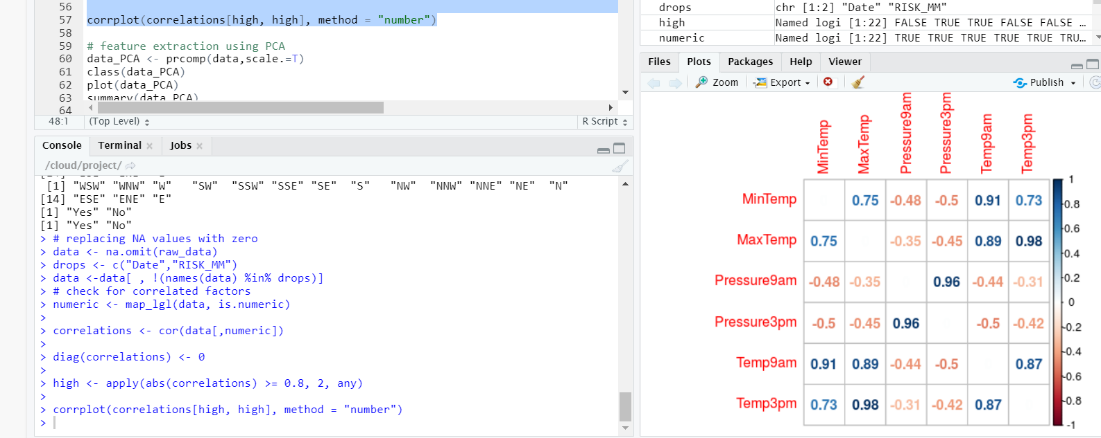


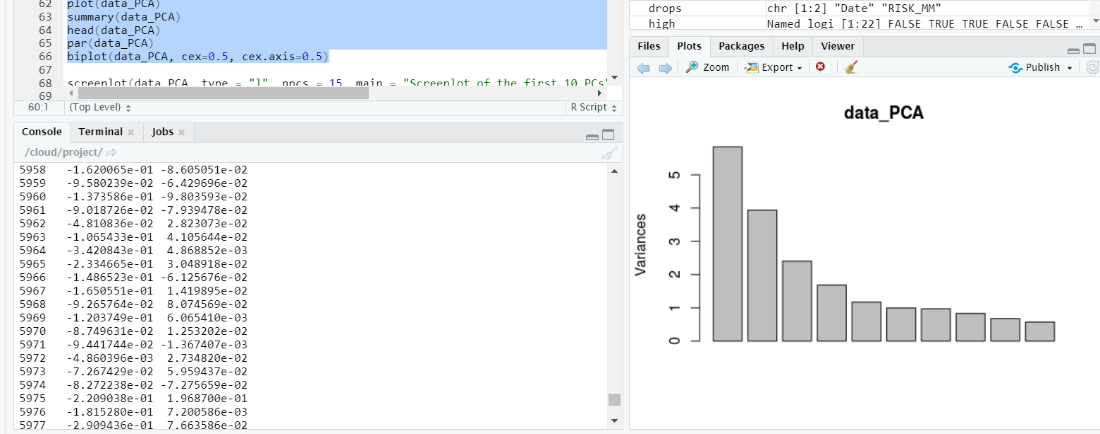
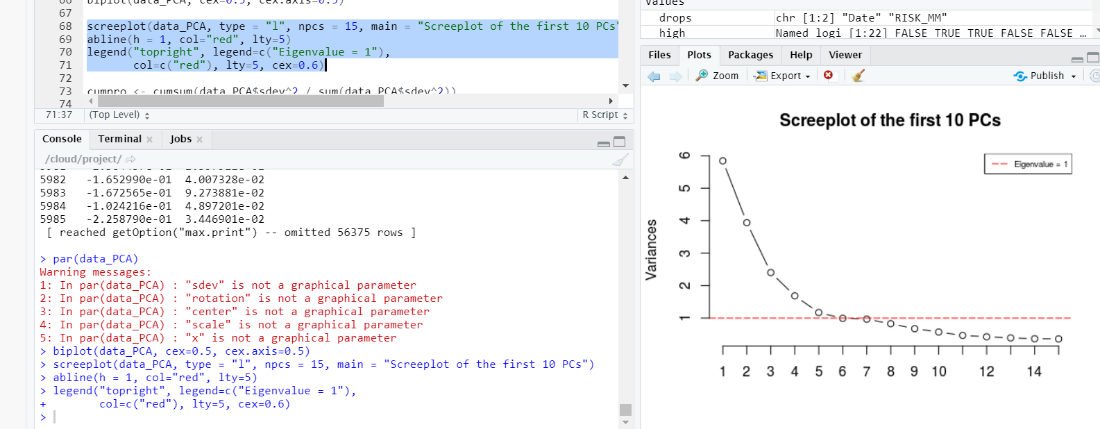
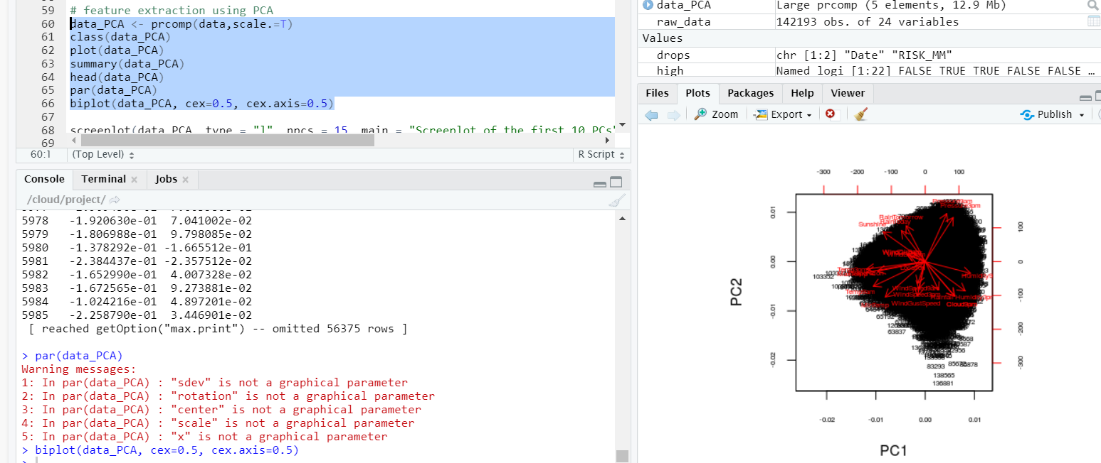
And with the above code omitted the NA values and remove the date column form the dataset. Now the dataset has 23 features.

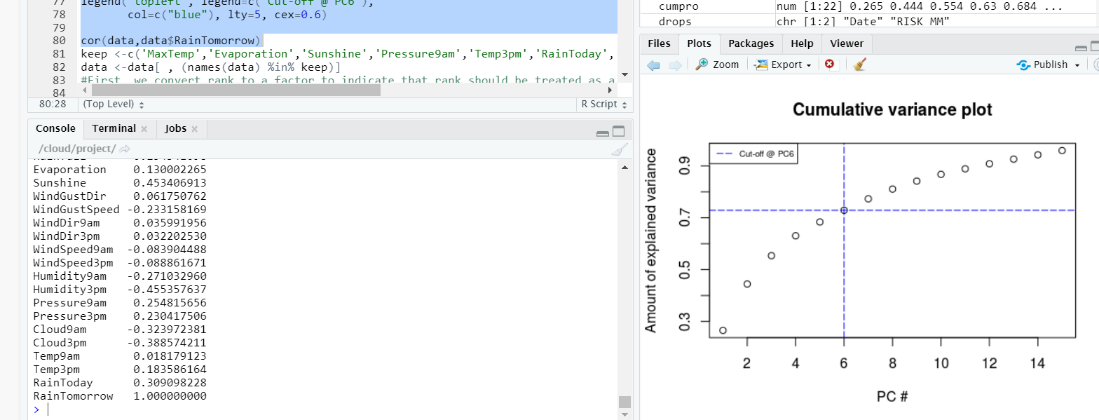


* RainToday, RainTomorrow feature values are replaced yes with 1 and No with 0.
* Factorized the categorized variable so that regression model identifies them

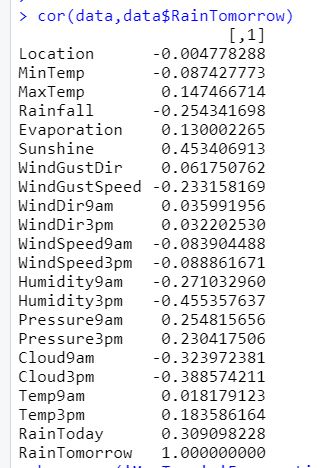
Feature selection:







As our model has total 22 features, but with considering categorical dummy variable it would go around 84 variables. For this to select the features PCA method is applied.

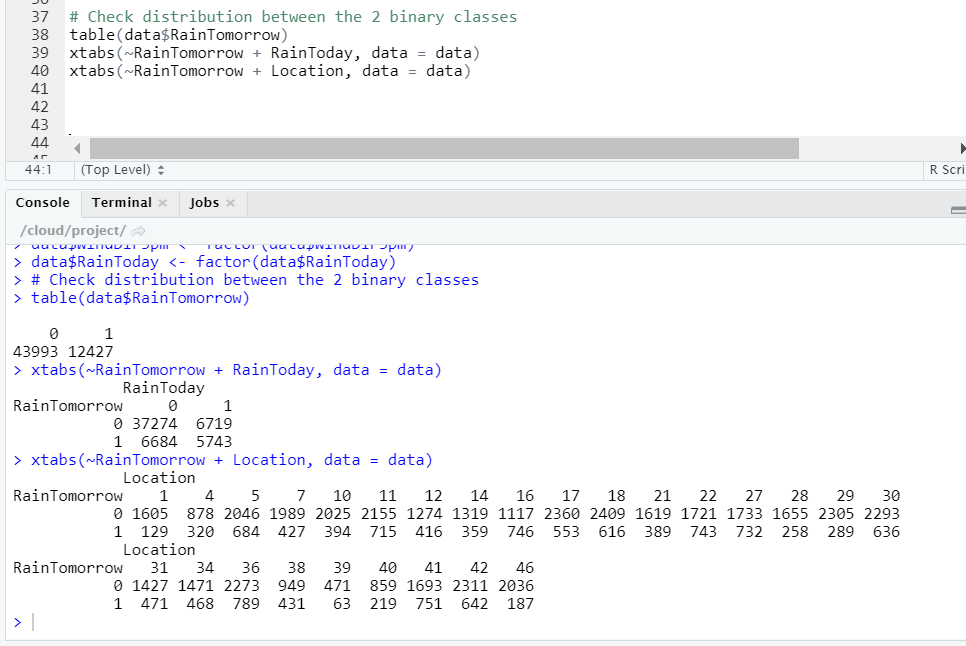
The results obtained from PCA analysis are presented above. From this analysis come to know 6 features are the most important features. Later applied cor(data,data$RainTomorrow) function to identify the topmost features. 

The dependent features are:

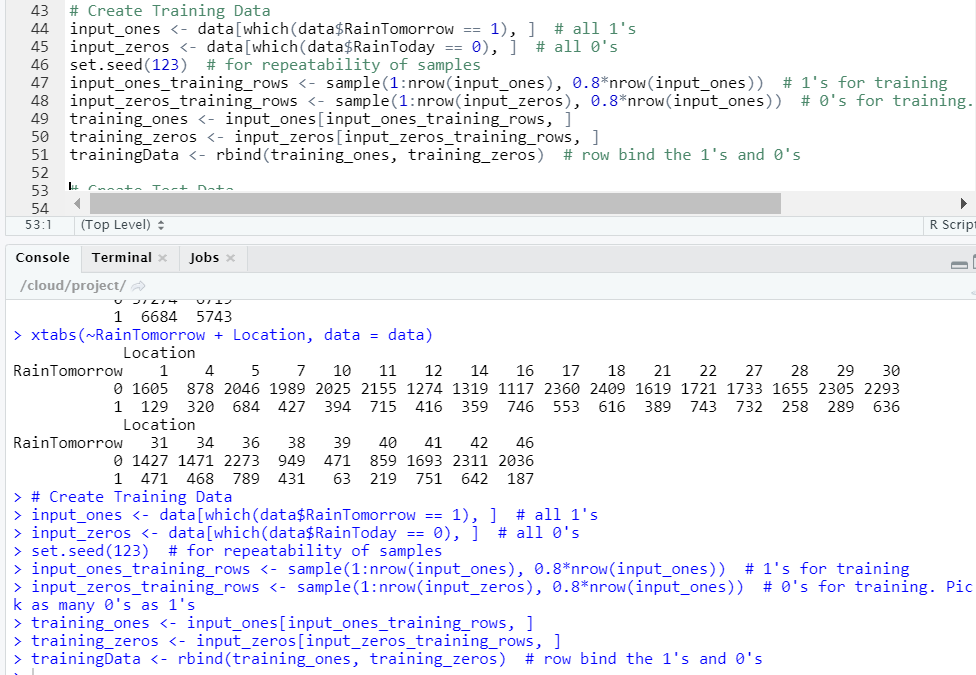
1. Max Temp
2. Evaporation
3. Sunshine
4. Pressure9am
5. Temp3pm
6. Rain Today

Rain Tomorrow is the target feature

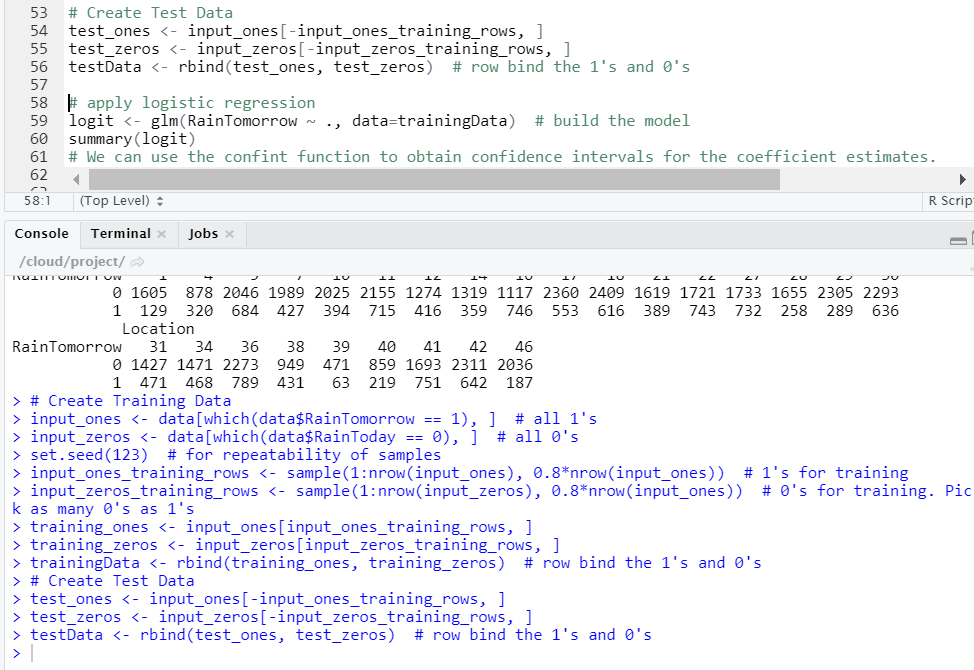
Engineering:



* Observed the distribution of yes and no with RainToday and Location features
* Also checked the no of 0 and 1 in RainTomorrow features. We can see clearly we have more zero’s than 1. The class is biased.



* Set seed makes the dataset always contains same random numbers generated
* Using sample function from catools library data has been split into 80% training data and 20% test data
* As the class biased data has been split into such way that both training data and test data have same percentage of yes and no’s
* In the above code 80% data from which has values 1 and 80% data which has zero values taken as training data.



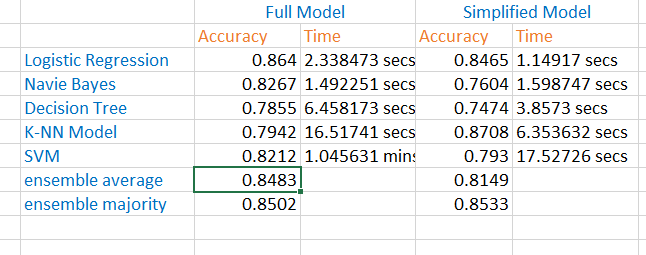
* Similar logic is applied to get the test data

**5. Training Methods:**

Logistic regression, naïve Bayes, decision tree, K-NN model and SVM models are used to train the data.

Also applied average and majority vote ensemble methods to the predicted results obtained from above models.

For the average methods calculated the average and value above 0.5 are taken as 1 and below are 0(NO). For the majority voting top 3 best performance are used and selected1 if any two of them are predicted as 1.



Accuracy and time taken for the full model and simplified model are shown above.

The key thing, I think separate from others using several models as well as using ensemble methods and calculating the metrics which are most interested to customers.

Interesting things:

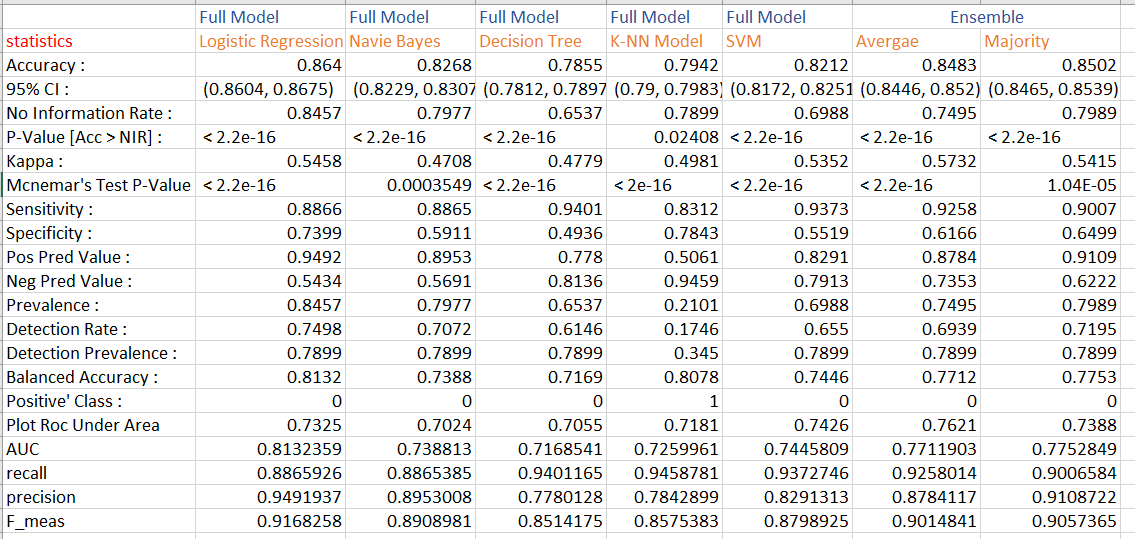
* Naïve Bayes took longer time for simplified model
* K-NN model performed amazingly well in simplified model with outstanding accuracy of 87.08%
* Ensemble average performed well in full model compared to simplified model
* Ensemble majority performance is almost same in both cases
* Logistic regression performed top in full model

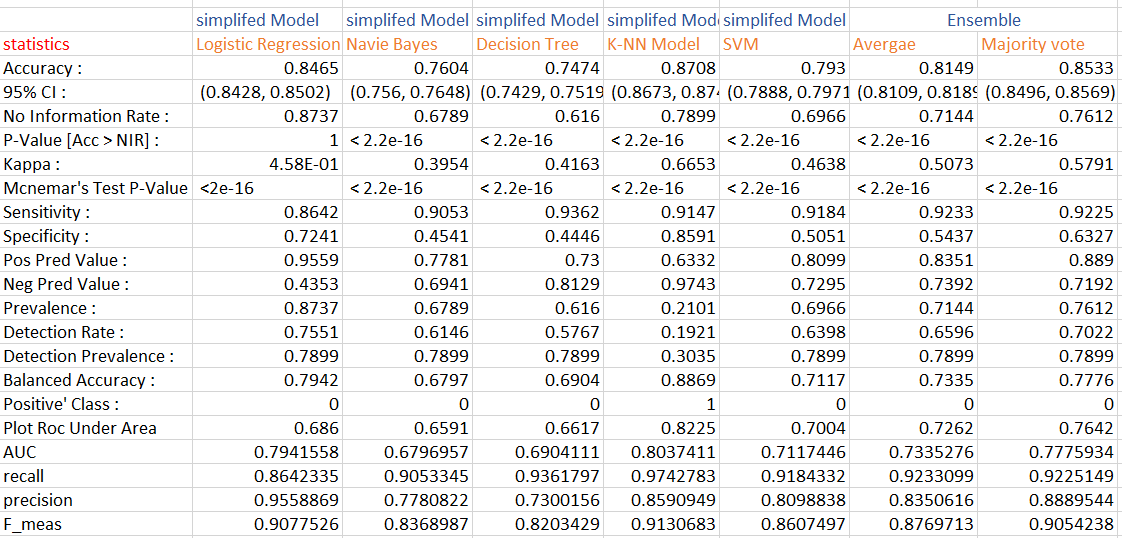
**6.Accuracy metrics reporting, charts, Model Execution Time:**

Accuracy and time completion are shown in training methods.

The ROC plots for each model are included in zip file in full model and simplified model file respectively. The value and under ROC curve is also captured to statistics table.

Below are some other accuracy metrics:



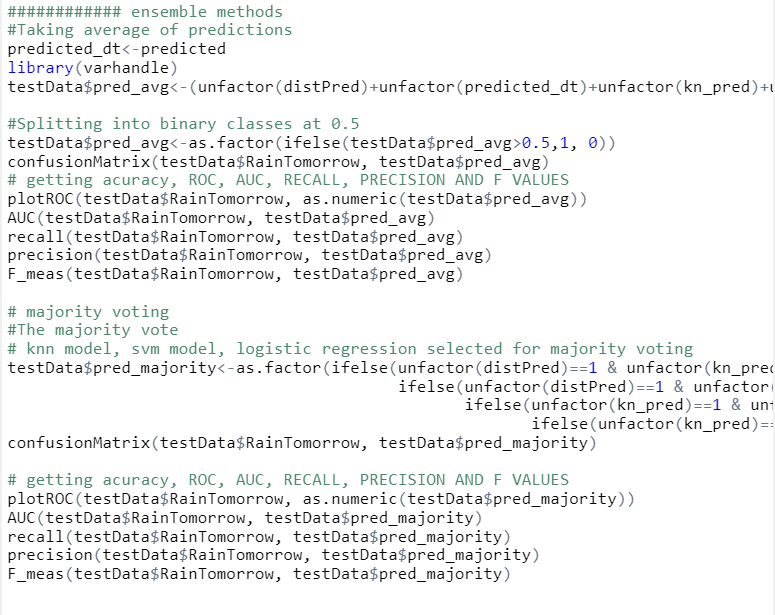


**7.Use of ensemble methods:**

The ROC plots for each model are included in zip file in full model and simplified model file respectively. The value and under ROC curve is also captured to statistics table.

The results are combined with the above models.

The below code is used to get the ensemble results



Summary:

The most important features are analyzed using PCA method and those are Max Temp, Evaporation, Sunshine, Pressure9am, Temp3pm, Rain Today, Rain Tomorrow

As overall 5 models and 2 ensemble methods are used for the analysis of data. Those models are logistic regression, decision tree, naive Bayes, K-NN model, SVM, ensemble average and ensemble majority vote.

R programming is used for the analysis.

References:

1. <https://www.kaggle.com/tiepmh/is-it-raining-tomorrow>
2. <http://users.wpi.edu/~rmclark/FolderofFiles/LITERATUREREVIEW.pdf>
3. <https://www.researchgate.net/publication/326014915_Research_on_Real-Time_Local_Rainfall_Prediction_Based_on_MEMS_Sensors>
4. <https://user.ceng.metu.edu.tr/~e132252/KE_Survey.pdf>
5. <https://www.academia.edu/22332243/CHAPTER_2_LITERATURE_REVIEW>
6. <https://www.kaggle.com/jsphyg/weather-dataset-rattle-package/kernels?sortBy=hotness&group=everyone&pageSize=20&datasetId=6012&language=R>
7. <https://stackabuse.com/using-machine-learning-to-predict-the-weather-part-1/>
8. <https://www.researchgate.net/publication/222361742_Rainfall_forecasting_by_technological_machine_learning_models>
9. <https://togaware.com/onepager/>
10. <https://www.kaggle.com/momenon/weather-prediction-in-r-forked>