Institute of Informatik

Statistische Methoden des Maschinellen Lernens

Agathiya Raja – 519788 [AFB 2013 Informatik]
VijayKrishna Babu - 514635 [AFB 2013 Informatik]

Technische Universität Clausthal
Angewandte Statistik
Dr. Annette Möller

23-02-2021

Table of Contents

List of Figures

List of Tables

1.Introduction

The goal of the project is to implement our skills and experience concerning machine learning and data analysis. Thereby, to enhance our knowledge and gain more experience through this project. Hence, in this report, we are displaying our results and outcomes of analysing the dataset “WeatherAustralia 2020” from the R package ‘rattle’. The outcomes shown here will be the results of combining different machine learning methods and their comparisions. In the upcoming topic, we have shown the “BoxPlots” according to the type of predictor’s potential. Box plot concept also comes under the topic “Predictor selection”. The dataset has been split into testing and training dataset. Logistic Regression, K- Nearest Neighbors, Deep Neural Network are also used for fitting the model classification. At the end of the report, Interpretation is given to conclude the outcomes of the project.

2.Exploratory and Graphical data analysis

In this section, we need to explore our data so, we start drawing graphs using box plots for the potential predictors, and we regrouped them according to the type.

1. Boxplots of regrouped variables

The variables like Cloud, Temperature, Humidity, Pressure, and Wind are grouped into subsets based on their type.

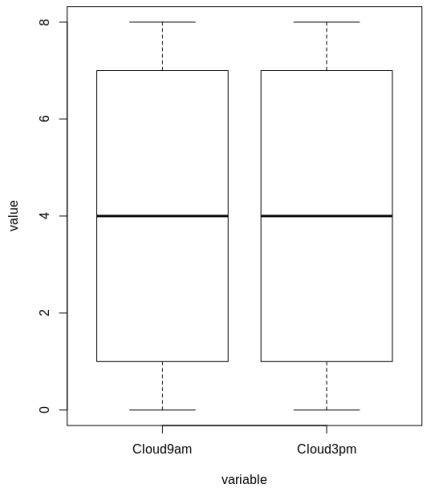


Figure 2.1: Cloud variables Box Plot.

For the variable Cloud, we grouped Cloud3pm with Cloud3am. From the given diagram, it is clear that the median is at the same value (4); hence the two box plots are perfectly matching. The box heights and Whisker levels are the same. This provides the impression that both the variables are delivering the same kind of information. Hence, we need to investigate more which will be done further on the section of Predictor selection.

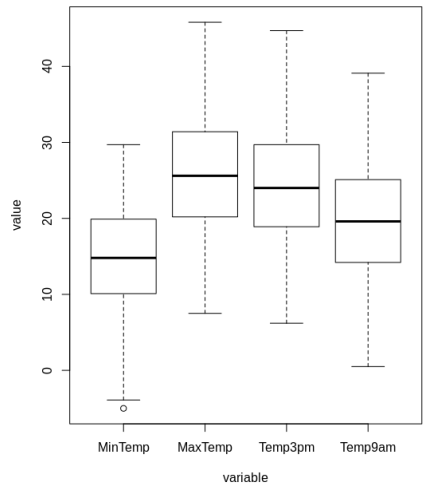


Figure 2.2: Temperature variables Box Plot.

For the variable Temperature, we grouped MaxTemp, MinTemp, Temp3am, and Temp3pm. Here, we are receiving an output with comparatively fewer overlapping with other box plots. It is the same case also with MinTemp; there is even more occasional overlapping; however, some outliers exist for this variable. Both the box plots shares the median into two equal half quantiles.

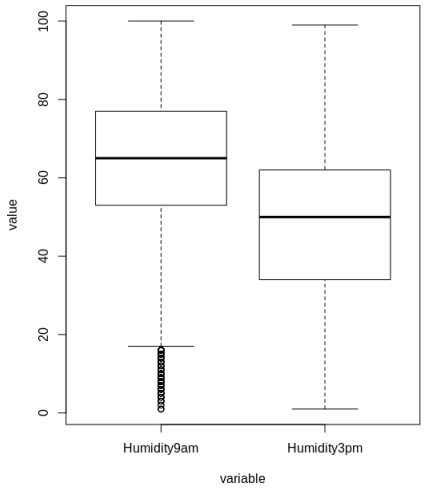


Figure 2.3: Humidty variables Box Plot.

For the given variable Humidity, we have grouped Humidity9am and Humidity3pm. Both the boxplots of Humidity9am and Humidity3pm show no overlapping. This is because the median of Humidity3pm lies outside the boxplot of Humidity9am, as well as the median of Humidity9am lies outside the boxplot of Humidity3pm. The outliers present in the Humidity9am stay under the minimum whisker. It is also observable that the whiskers present in Humidity3pm are more comprehensive than the Humidity9am.

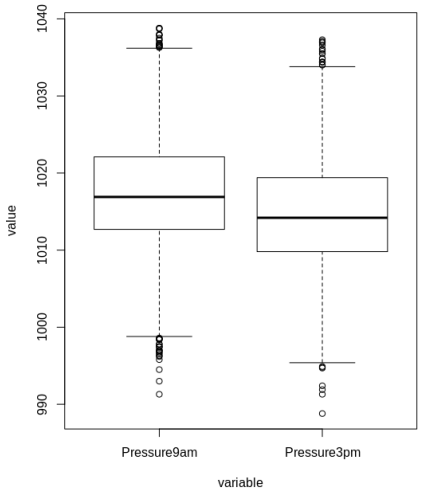


Figure 2.4: Pressure variables Box Plot.

For the variable Pressure, we have grouped Pressure3am and Pressure3pm. The output (boxplots) of the variable Pressure3am and Pressure3pm has overlapping.

This is because the median of Pressure3pm lies inside the box plot of Pressure3am and the median of Pressure3am lies inside the boxplot of Pressure3pm. Pressure3pm has comparatively fewer outliers than Pressure9am. The outliers present in both the variables are condensed above and below.

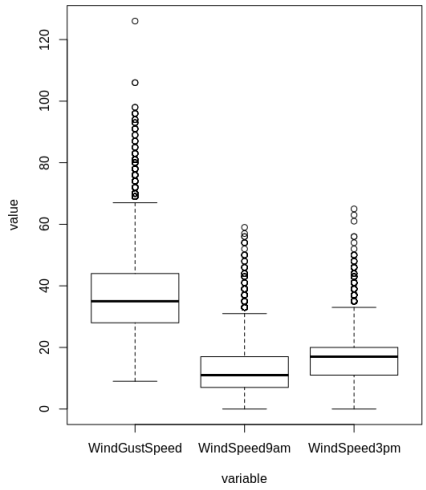


Figure 2.5: Wind variables Box Plot.

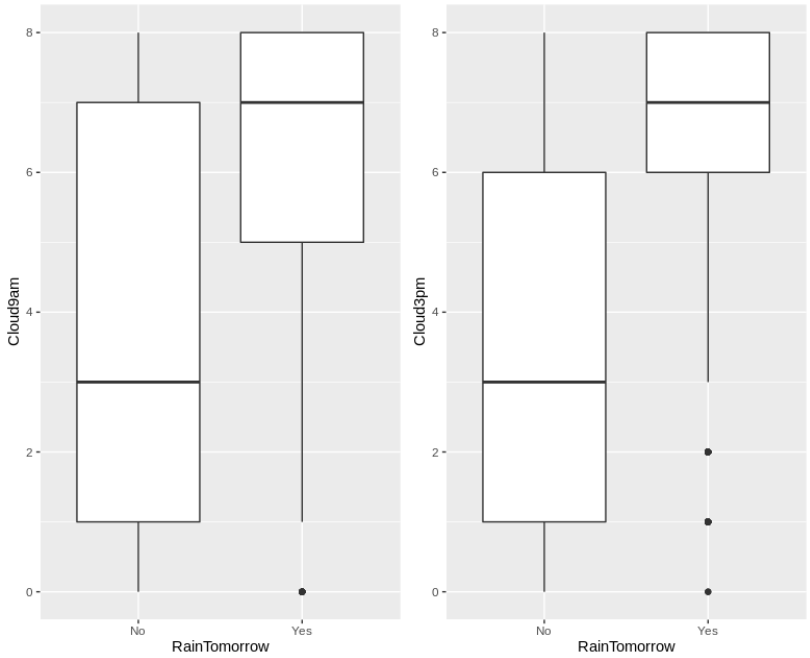
For the variable Wind, we have grouped WindGustSpeed, WindSpeed3am, and WindSpeed3pm. From the given boxplot, it is clear that there is no overlap in the first boxplot with the remaining two boxplots. It contains wider whiskers; hence it has more range. The outliers present in the given boxplots are placed above them.

3.Preparing the data

In order to perform this task, the preprocessed (filtered) data set with no missing values are used. However, this preprocessed data set has few gaps for some stations and variables, but it doesn't manipulate our test and train splitting techniques.

3.1Predictors selection

From the given Figure 2.1, it is clear that the boxplots of cloud variables are the exact match for predictors selection, but we still have to analyze more. This analysis can be done by plotting RainTomorrow vsCloud3pm and Cloud9am.



(a) model1. (b) model2.

Figure 3.1: Box Plots

There is no overlapping occurrence because the median displayed in the figure of RainTomorrow (Yes) lay outside from the boxplot of RainTomorrow (No). This provides us the impression that there is some difference between the two groups. We can observe the variation in height. E.g., the box plot of RainTomorrow (Yes) is smaller than the box plot of RainTomorrow (No). Both the scales of RainTomorrow () are the same; hence it shows that the researchers have a closer opinion on the two scales. The RainTomorrow (Yes) only contains the outliers which lie below them. Overlapping occurs in both the box plots. This overlapping describes the difference (in data points) between two groups of RainTomorrow. The opinion is concrete if the section is shorter, whereas the opinion is more variable if the section is larger. In RainTomorrow (Yes), there occurs only below outliers.

For the final discussion, we have selected Cloud3pm. There is no occurrence of overlapping in the box plots of the two classes. As per our analysis, there are more informative predictors, and we chose MaxTemp as the predictor. The reason for selecting Humidity3pm as a predictor is that there is no presence of outliers and the range (upper and lower) of the whiskers are more significant than the Humidity9am. We have also selected Pressure3pm because the outliers in it are comparatively lesser than the Pressure9am. The upper and lower range of whiskers is broader.

The reason for choosing the WindGustSpeed is that there are no overlapping occurrences with the other boxplots. The range of upper and lower whiskers is greater. We can observe two unequal sections among WindSpeed9am and WindSpeed3pm, which state the same view at some parts of the scale, whereas more variation occurs in the other part.

3.2 Train data and Test data

In the given data set "WeatherAustralia", there are some missing values for 1 year (Jan 2016- Mar 2017). We are splitting this data set in a ratio of 70 percent (Training) and 30 percent (Testing) with the use of "CreateDataPartition()". It divides data in horizontal order. We are applying all the models on same training and testing data. We also split data by date and applied KNN regression and Deep Neural Network but it was not giving good results. There were overfitting in training and validation accuracy comparison graph, and the value of loss and val-accuracy was a bit noisy. Which we can notice in fig. 3.1.

Even if we divide data by date there are some locations which do not contain any records for last 3 years.

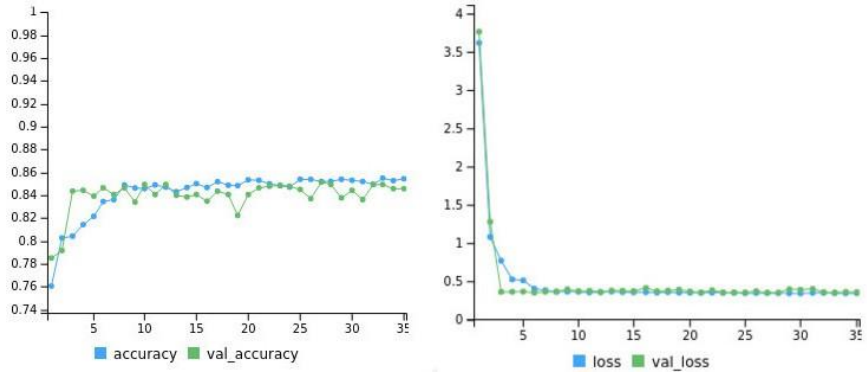


Figure 3.2: Overfitting and Noise

4. Comparison of Classification Methods

4.1 Fit classification model

We are using Logistic Regression, k-Nearest Neighbours and Deep Neural Networks as employ classification methods for binary response RainTomorrow.

4.1.1 Logistic Regression

4.1.1.1 ANOVA and models comparaison

We prepare 3 models and we compare them using ANOVA. In the 1st model, we remove Rain- Today, Date, Cloud3pm, and Evaporation. In the 2nd model, we remove Date, Evaporation, Cloud3pm. In the 3rd model, we use all the 10 predictors. We conclude that the 3rd model is the best and that is resumed in the following table and explained afterward:

Table 4.1: models characteristics.

Model	RD	AIC	Fisher Score	DF	Acc.(%)
null	7204	——	——	6830	——
model1	4437	4465	6	6817	86.20
model2	4418	4448	6	6816	86.24
model3	4383	4419	6	6813	86.28

As we can firstly notice, comparing the null model (with intercept only) to the other 3 models held a significant reduction of the residual deviance (noted as RD from 7204 to a mean of almost 4413) with a loss of 13 Degree of Freedom (noted as DF in the table). By adding RainToday to the set of predictors of the original model we get slightly better results in model 2, and this can be observed in the difference in the residual deviance, which is reduced in the larger model by 19 with a loss of 1 degree of freedom, we can also notice that the AIC value was also reduced, in this case by 17 which is slightly better than the original model. The third model is an extension of the second model, it includes all chosen predictors in the predictors selection phase (please consult the previous section...); this last one was performing best in terms of the residual deviance rate and AIC value, we had a reduction of RD=35 and AIC=29 in comparison to model 2, RD = 54 and AIC = 46 in comparison to model 1. We calculated the classification accuracy by taking the mean of the number of correctly classified instances by our models compared to the real response. The accuracy values are almost the same (slightly differentiating), but it will be useful for further investigations and overall assessments of all used classification methods in this project.

4.1.1.2 Fitting of the models

This part was about predicting the likelihood of rain to occur the next day, meaning convert the categorical variable to a continuous one. Our task is a binary classification task, thus, we took the threshold of 0.5 to classify either it will rain

tomorrow or not [please consult subsection ANOVA for more details about the use predictors].

4.1.2 k-Nearest Neighbors

We are using KNN regression for different 3 models which has different predictors. We are comparing results of all three models and filtering best model from that.

Table 4.2: K-NN Regression Outcomes.

Metrics	Model 1	Model 2	Model 3
K	21	25	25
Accuracy	79.21%	86.80%	87.04%
Brier Score	0.1484	0.1009	0.0968
ROC	0.7348	0.8812	0.8933

In the above table, we can see there are 3 different models and four different criteria to check assess classification performance. From these criteria, we can say which model is best for our dataset. The criteria are Number of K, Accuracy, Brier Score, and ROC-Curve. For e.g if we see the value of accuracy the model 3 has the highest accuracy 87.04% comparing to other 2 models, and it is same for the Brier Score and ROC-Curve model 3 has a better value compare to other models.

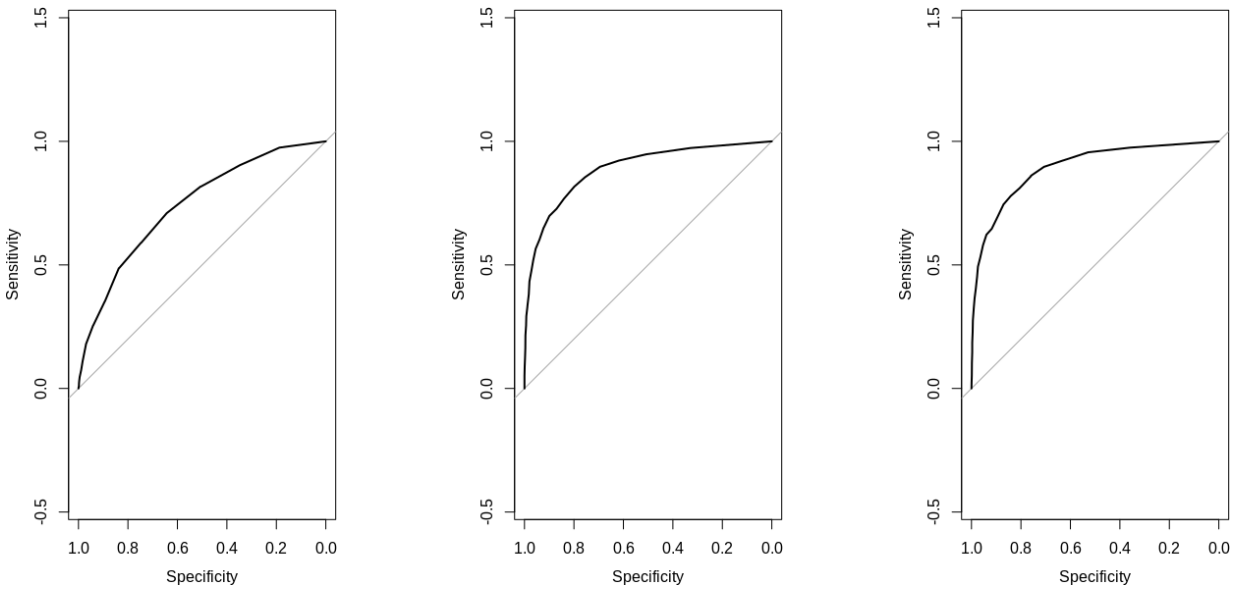
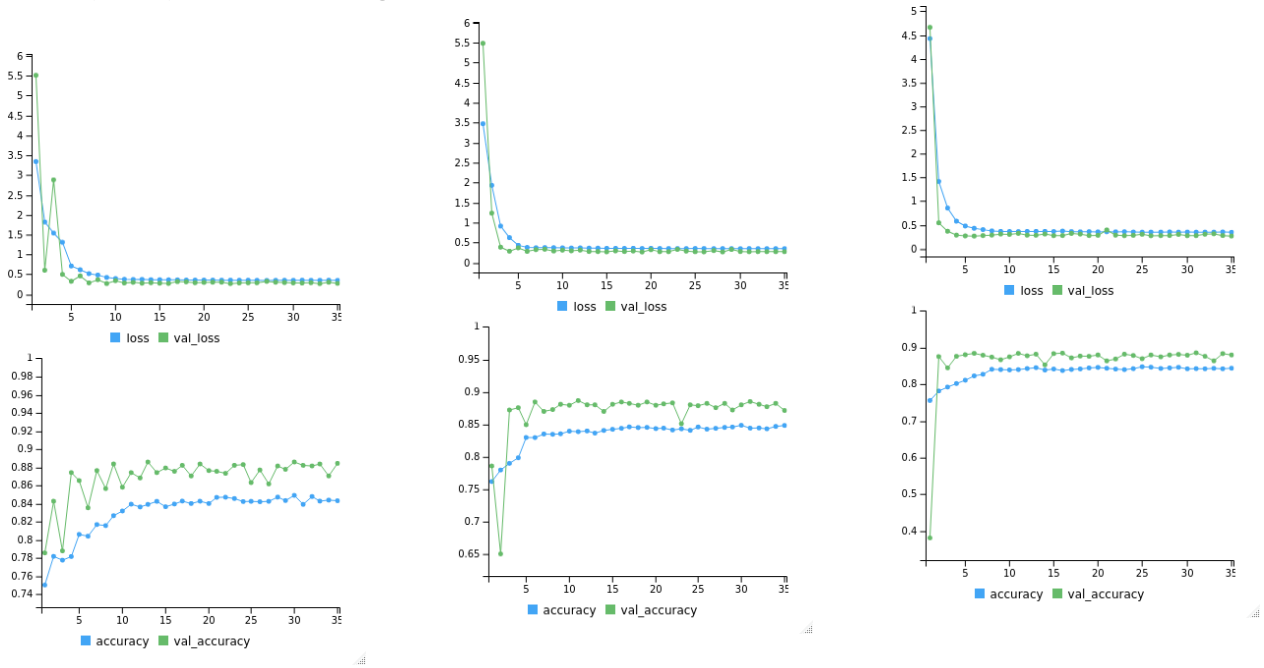


Figure 4.1: K-NN Regression ROC-CURVES

As observed in the above figure of ROC-Curves we can see models 3 and 2 have better results than model 1. All the models have True Positive Rate but ROC-Curve of model 3 and model 2 is close to 1.0. The value of ROC-Curve model 3 is 0.8933 and model 2 is 0.8812 which are really good values whether for model 2 it is 0.7348 which is not good compared to other models.

4.1.3 Deep Neural Networks

For this task we made use of the "Keras" package with R, the algorithm was accelerated with an enabled CUDA GPU. The data partitioning is playing a major role, and directly affecting the training/validation and testing processes, and we couldn't do much by changing the parameters of the created classifiers. Our approach was to go from a less dense (light) model to a dense one by changing the number of nodes (units) we are taking.



(a) model1. (b) model2. (c) model3.

Figure 4.2: Accuracy/Loss for training/validation for 3 different models.

All models were formed of 3 layers; the first layer has a "ReLu" activation function and with an input shape of 8 (8 used predictors), the hidden layer was also activated by the "ReLu" function; finally, the output layer was created by 2 nodes because we

are having a binary classification problem (2 possible outputs), and it was activated with "softmax", which is a natural choice since our task is a classification problem and it requires a non-linear activation function. As we noticed some signs of overfitting in model 2 [figure 2] we tried to randomly drop out 30% of the input from the hidden layer and 10% from the output layer and that's for model 3. As for the compilation of our models, we used the "categorical_crossentropy" and that is also for the categorical nature of our response variable "RainTomorrow". We fitted the models using 35 epochs and batch size of 5, and with a validation split of 20%. The following figures show the performance of different classifiers:

As we can see, model 1 has a noisy validation loss rate and also noisy validation accuracy which is often a sign of overfitting. Although in model 2, the validation loss is less noisy than model 1, we can observe an overfitting behavior almost immediately at the beginning of the training (epoch 2 or 3) and also along the training process and that is noticeable in the curve.

(b). In curve (c) representing model3, we can notice that both the loss and the accuracy values are more stable along with the training/validation phase, this last one was chosen to represent the MLP method we used. As a conclusion, and by comparing the performance of different deep neural networks classifiers, we can assume that we couldn't get much difference between them in terms of the accuracy/loss values, so we tried to select the best model from this category based on a comparison of the graphical representations, meaning, by looking for signs of overfitting then avoid that specific model, also we calculated some metrics (Accuracy, Loss, Brier Score, and AUC) for the final assessment and comparison of the different classifiers.

4.2 Assessment of classification performance

For the assess classification performance we are using accuracy, Brier score, and ROC curves. We are investigating the overall performance aggregated over all test cases (dates and stations).

4.3 Interpretation

The comparison of best model of all three classification methods are in below table 5.1.

Table 4.3: Final Values.

Metrics	Logistic Regression	K-NN Regression	Deep Neural Network
Accuracy	86.28%	87.04%	85.78%
Brier Score	0.0996	0.0968	0.1005
ROC	0.8935	0.8933	0.8879

As observed in the above table we can observe that Logistic Regression and K-NN Regression gives the better results compare to Deep Neural Network. Which means for this dataset "WeatherAustralia" and Response "RainTomorrow" Logistic Regression and K-NN Regression are better compare to other method.