A healthy comparison between Logistic regression and Random forest techniques to predict the Rain for tomorrow

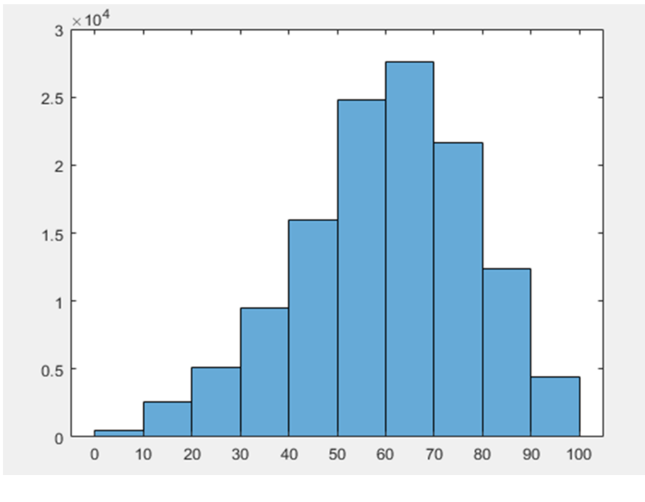
Vivek Kanna Jayaprakash | 210033252

1.0 Brief Description and motivation:

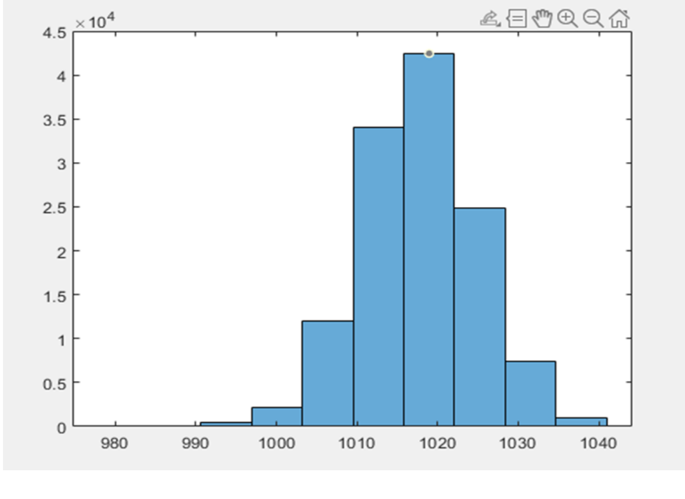
"Will it rain tomorrow? " is one of the most asked question asked by most of us, everyday on every single area across the face of the world. In this project we attempt to predict the rain for tomorrow classification problem using Logistic regression and Random Forest techniques and compare them later using Parislab @ UCLA model [3].There has been many Machine learning classification models [2]predicting the rain based on various attributes of temperature, windflaw, direction, cloud position etc. but since there has been very few attempts to predict the same using the average humidity and pressure recorded on the said day we set the goal of predicting the rain using the two attributes alone

2.0: Initial analysis of data and basic statistics

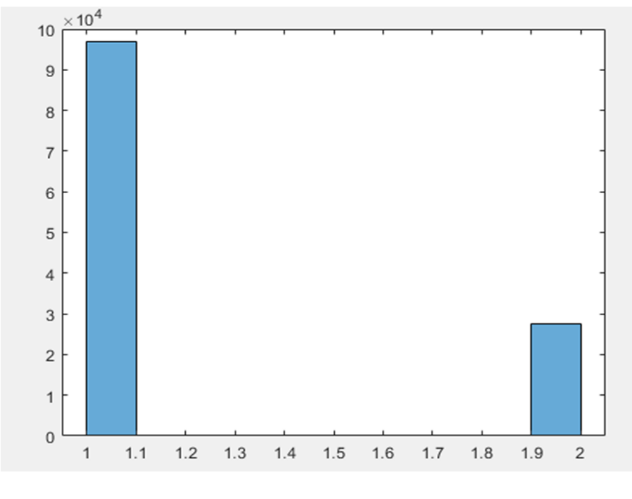
- Raw dataset Rain in Australia (2008-2017): by Joe Young from Kaggle.
- As we are attempting to predict the weather for tomorrow using relative humidity and pressure alone, we take averages of the same per day and drop the other values.
- The null values constitutes only to 10 % of the entire dataset, we drop the same.
- The Cleaned data consists of 124412 rows, 2 features and 1 target variable.
- The data is not standardized
- Class imbalance exists in the ratio of 75:25 of the target variable,
- Visual representation of the data indicates the data is normal with very less skey and the data set is not co-related to each other.
- Initial analysis suggests that relative humidity contributes more to the rain than pressure.



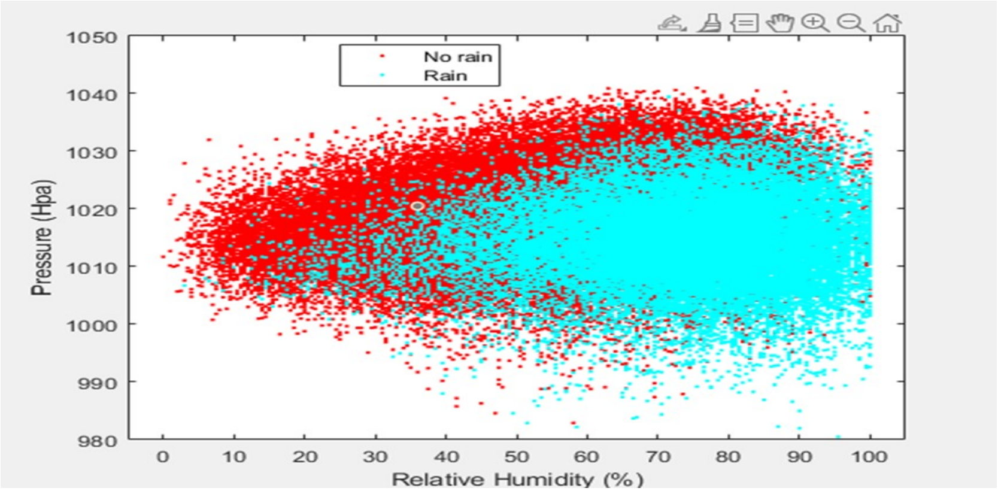
Humidity (non standardized)



Pressure (Non Standardized)



Target variable imbalanced



Rain or no rain with respect Relative Humidity Vs Pressure

3.0 Brief Summary of Logistic regression and Random forest. With the pros and cons

3.1 Logistic Regression

Logistic regression is a supervised machine learning technique used in classification to predict the outcomes of the dependent variable based on the input given in the training set. The target variable should be binary in nature. It is used in many cases which attempts to tells us iof a binary event/multinomial is occurring or not. (Yes or no)[6]

3.1.1 : Pros

- Very easy to train , test and implement than any other models.
- Can easily be upgraded to fit in new sets of data
- Provides informative and well calibrated output results which helps us interpret any hidden layers of our data.

3.1.2: Cons

- Linearity assumption between the dependent and independent variables in the major limitation
- Doesn't attempt to work with complex features due to the simple mechanism,
- Always required a large dataset.

3.2 Random Forest classifier

Random forest is a flexible, easy to use supervised machine learning technique used in classification and regression tasks. A RF model consists of many decision tress which predict their outcomes separately . The more trees, more the accuracy of the model. RF classification techniques is an ensemble method that while training various decision tress using methods of bootstrapping & bagging.[6]

3.2.1 : Pros

- Can handle non linearities and high dimensional in the given data
- Works for both classification and regression problem and also for categorical and non categorical variables.
- Missing values is not much of a concern

3.2.2 : Cons

- Takes a long time to train
- Difficult to extrapolate as small changes in the new data can cause huge changes to the algorithm.
- Random forest name is trademarked and requires a license to use commercially.

4.0: Hypothesis statement :

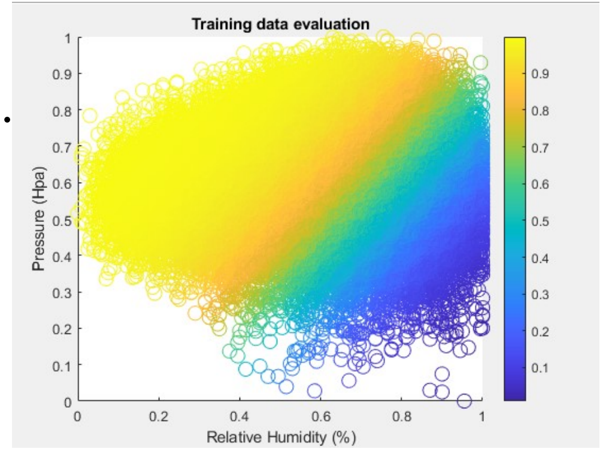
- Both the models is assumed to perform significantly better than a random guess to predict the rain based on the relative humidity and pressure data.[1]
- Random forest model will take significantly longer time to train than the logistic regression especially with the huge amount of data,
- According to the Parislab@UCLA model[3], the accuracy for the logistic regression model is found to be at 72% and we aim top keep that as the target for our model. But since our data is significantly larger we may have certain variations.
- Since we are considering only two sperate metrices of Humidity and pressure the linear regression model is assumed to fare better compared to random forest due to high linear behavior showcased by the dataset.
- Hyperparameter tuning would not be necessary in the random forest model due to the less complexity of the dataset.[4]

5.0 : Choice of Training :

- The cleaned data set is spit into a ratio of 70:30 for training and testing respectively. The test data is completely unknown during the process of the training.
- Mnrfit a multinominal linear regression model is used in this case to train a binary data set as per the paper referenced and it is used to train for the LR phase and we chose not to have any validation sets as hyperparameter training is not done.[4]
- We are doing random permutation of the index to shuffle the index data completely randomly to get a randomized dataset for training.
- Ensemble bagging method is used with 100 trees in random forest.

5.1: Evaluation methodology :

- The unseen test set is run on both the model to predict if the test set is running perfectly so we can achieve a greater level of confidence within the model and is to run to predict if the accuracy is fine so we get a good proper evaluated model.
- The precision recall and F1 score is calculated to identify the training set is performed well during the evaluation.

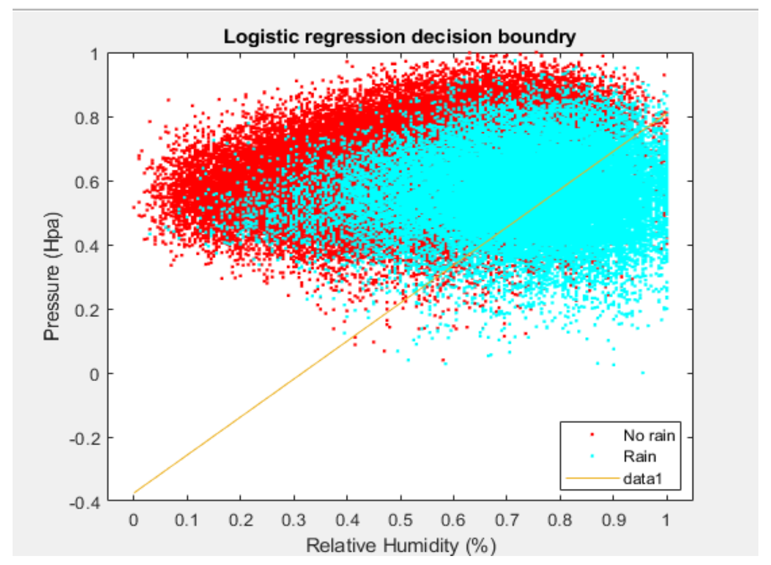


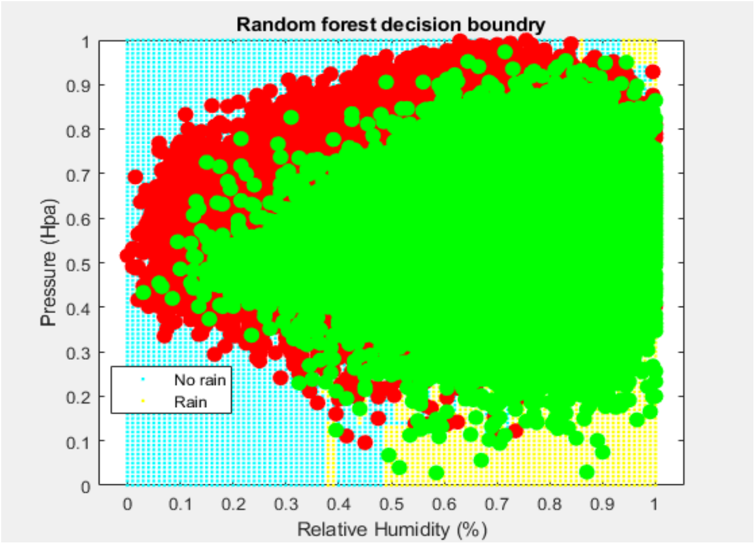
6.0 ; Choice of parameters and experimental results

Logistic regression : mnr fit a multinomial linear regression model is used for a binary classification problem in this case as per the reference attached gives a standardized result with respect to the classification problem. The Beta model generated produces a intercepts and beta coefficient for feature 1 and beta coefficient for feature 2. LR trains the value of the coefficients of intercepts beta 1 and beta 2 to generate a sigmoid which predicts the Propensity and predictably of belonging to one class or not and to identify the probability we take the sigmoid of Z. A decision boundary or a equation of hyperplane is delivered which indicated z=0.This decision boundary splits the data set into two parts weather it will rain or not. Threshold was set to the cut off pint in probability.>0.5 to experiment exactly with the exact decision boundary split. Based on this it is understood that the less pressure and more humidity contributes rain and vice versa..[3]

Random Forest : A bag model is used in passing over the MATLAB fit ensemble method to fit the custom tree. This method is ideal as it is easily used to evaluate the maximum number of splits, predicter variable and other said variables for better hyperparameter tuning. The number of trees selected was 100 and the maximum number of leaf splits were 2344, leaf size was 1.Tuning hyperparameters using the fit ensemble function did not improve the model accuracy. It was dropped from the model. Mainly due to less accuracy .

.

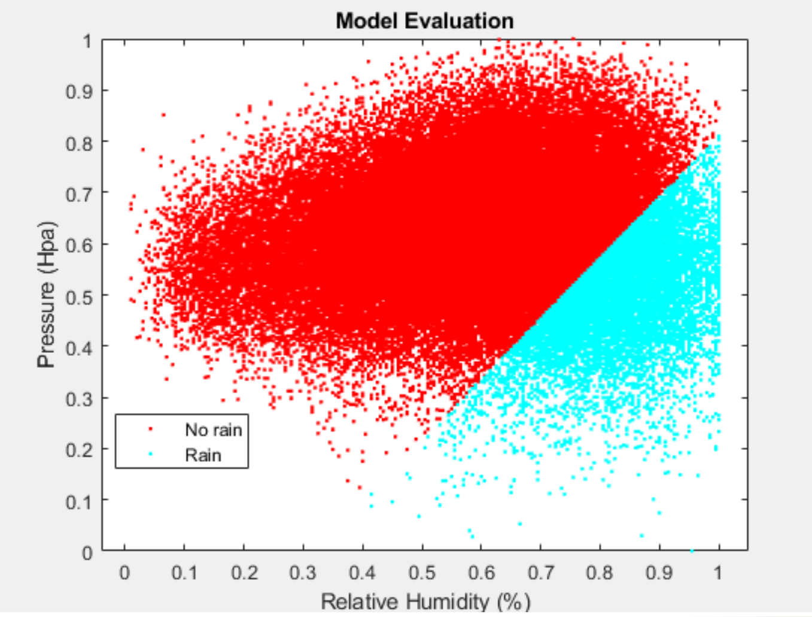


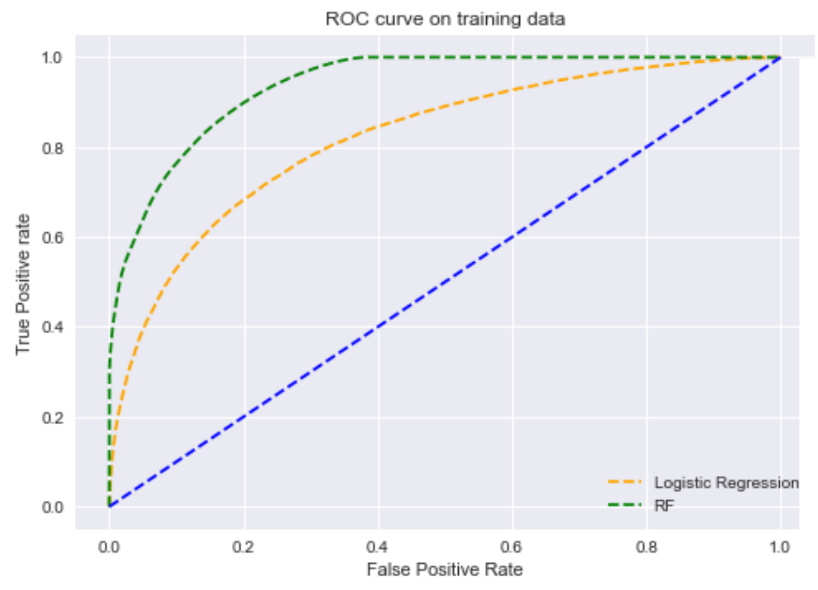


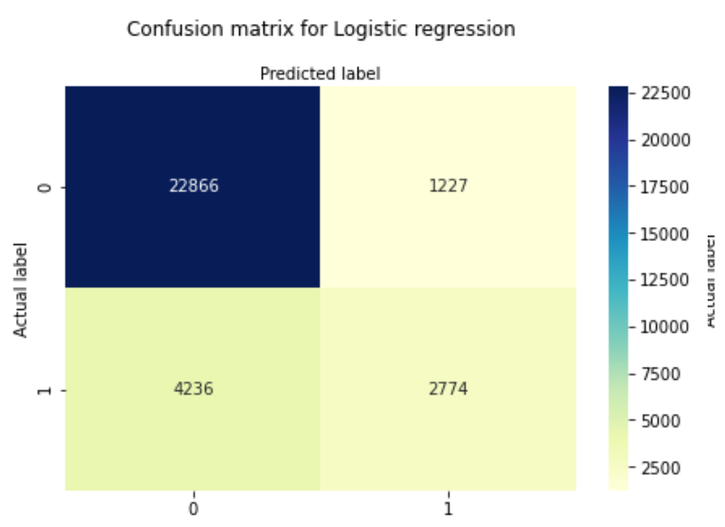
Models	Accuracy	Precision	Recall	F1 Score
Logistic regression	82.5	0.6795	0.3949	0.5035
Random Forest	76.22	0.5333	0.428	0.4634

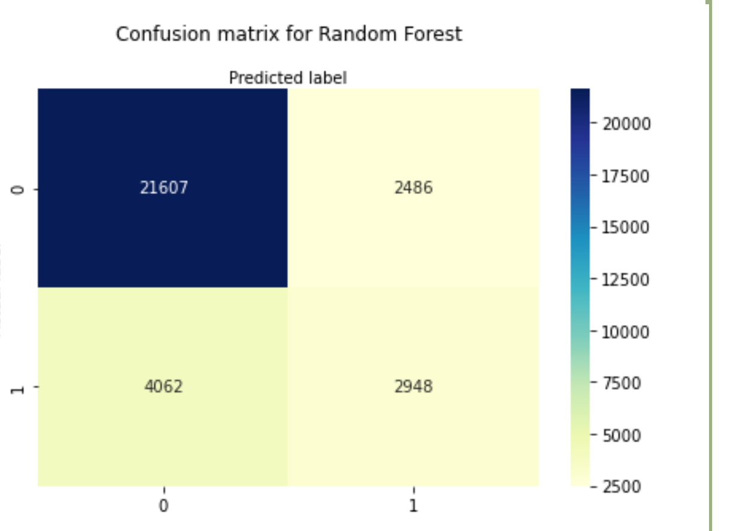
7.0; Analysis and critical evaluation of results

- We are attempting to predict tomorrow rain based on two features of relative humidity and pressure alone but in general metrological sense it actually depends on various other aspects of temperature, wind direction, cloud content etc. hence there may be actually mis repetency with the actual rain data in real world.
- The low recall and high precision of our model denotes that our classifier is very conservative and picky and in the attempt to make the correct prediction, it misses a lot of true rain days. It is similar to the literature we took basis from [3]
- Logistic regression shown significantly more detailed results than the random forest as the it is a highly linear model in general due to high linearity in the data.[5]
- Class unbalancing problem is a huge issue an issue in the model as 75:25 ratio with the target variable causing a imbalanced dataset.
- The train and test scores prediction for both the models vary a lot vary a lot mainly due to the amount of data present in the test set.
- The training time taken for the RF model is higher than the LR model which is as per the literature expectancy.
- Both models didn't perform well compared to Paris@lab data as we are not able to predict a lot a true_rain shown in the confusion matrix
- ROC curve indicates that the RF as a model performs better with the high area under the curve, indicates its prediction may be consistent
- The error in RF is lower mainly due the boosting MATLAB function used which is denoted by the lower precision recorded.









8.1 : Lessons learned:

1.Class imbalance in targets a huge concern especially for the training of models for getting accurate results.

2.Random forest is highly advance modelling technique and it requires a higher GPU computer. MATLAB online is not the perfect tool for the same.

3.Hyperparamet tuning should be attempted again efficiently to increase the accuracy of the model

8.2 : Future work : 1.Cross validation of the training set is a technique which can be implemented. 2.Python can be used a s good alternative tool for MATLAB because of the coding simplicity 3. Generalize sampling techniques like SMOTE an be utilized to make the model train well

4.Other modelling algorithms can be train and tested to find the best model.5.glmfit for the binary logistic regression model should be used.

8.0: References :

[1]Jarmulska, B. (2020). https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2408~aa6b05aed7.en.pdf. In: *Random forest versus logit models: which offers better early warning of fiscal stress?* Europe: European Central bank, pp.22–24, [2]

[2]Nguyen, G. (2016). *Evaluating statistical and machine learning methods to predict risk of in-hospital child mortality in Uganda*. MSc Thesis. pp.6–9.

[3]PARISLAB@UCLA (2020). *Training a Logistic Regression Classification Model with Matlab – Machine Learning for Engineers*. [online] www.youtube.com. Available at: https://www.youtube.com/watch?v=q3VGCbXiZm4.

[4]Ruiz-Gazen, A. and Villa, N. (2013). *STORMS PREDICTION: LOGISTIC REGRESSION VS RANDOM FOREST FOR UNBALANCED DATA*. MSc Dissertation. pp.2–4, 6.

[5]Schonlau, M. and Zou, R.Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal: Promoting communications on statistics and Stata*, 20(1), pp.3–29.

[6]Geng, M. (2006). *A COMPARISON OF LOGISTIC REGRESSION TO RANDOM FORESTS FOR EXPLORING DIFFERENCES IN RISK FACTORS ASSOCIATED WITH STAGE AT DIAGNOSIS BETWEEN BLACK AND WHITE COLON CANCER PATIENTS*. [MSc Dissertation] pp.21–24. Available at: http://d-scholarship.pitt.edu/7034/1/realfinalplus_ETD2006.pdf