Appendix :

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

Vivek Kanna Jayaprakash | 210033252 | Machine learning| MSc Data Science | City, University of London

1.0 : Glossary: [1], [2]

Accuracy : Classification model in which the predictions were found correct.

batch normalization: In the hidden layer of the activation function, the input and output gets normalised.

Bagging: Ensemble technique that fits many models on different subsets of the training data set.

Bias: A prejudice or favouritism towards a group of anything.

binary classification: A type of classification that outputs one of two mutually exclusive sub classes.

Boosting: ML technique that iteratively combines simple and not accurate classifiers into a high accuracy classifier.

Bucketing: Converting continuous feature to multiple binary features

categorical data: Features with discrete set of values

Class: A set of enumerated target value of a label

classification model: ML model for distinguishing two or more discrete classes.

class-imbalanced dataset: A binary classification issue where the labels for the classes have significantly different frequencies.

confusion matrix: An NxN table to summarize the success of a classification model.

DataFrame: A type of data for representing datasets in pandas

data set or dataset: A collection of examples

Down sampling: Reducing the amount of info to train a model efficiently.

Ensemble: Merging predictions of multiple models of overall different structure

false negative (FN): An eg which the model mistakenly predicts -ve class.

false positive (FP): An eg which the model mistakenly predicts +ve class

feature: An input variable for making predictions.

feature engineering: Determination process in which features might be useful in training a model then converting to raw data from the files in the log and other sources.

fine tuning: A secondary optimisation performance to adjust the parameters of an already trained model to fit a new issue.

generalization: Model’s ability to make correct prediction on new and unseen data

Gradient**:** Partial derivatise vector respecting all independent variables

Hyperparameter: Points one tweaks during successive runs of training.

Hyperplane: A boundary that separates a space into two halves or spaces.

Label: The answer or result portion in an example.

linear model A= Model which assigns a feature to a weight respectively to make predictions

logistic regression: A classification model that used a sigmoid to convert linear model into a value of 0 and 1 and 1

model: The concept of representing a ML system has learned from the training data.

Noise: A factor which disturbs a signal in a dataset.

Normalization: process of converting an actual range of values into a standard range.

Objective: A metric that is going to be optimised by the algorithm.

Optimizer: An implementation of the gradient descent algorithm.

Outliers: Distinct values form the group.

Overfitting: Model creation that matches the training data very close that training data failed to make exact predictions.

Pandas: Column oriented DA API in ML

Parameter: Classification model metric.

Pre-processing: Data for processing before the process of training.

random forest: Ensemble approach to find the decision tree which fits the training data best by creating many DT’s and determining the average.

Recall: Classification model metric

Scaling: Feature engineering practise to tame a feature’s range of values.

test set: Subset in a dataset that used to test the model after the training phase.

Training set: Subset in a dataset used to train a model.

True negative (TN): Correctly predicted negative class.

true positive (TP): Correctly predicted positive class.

Underfitting: A model’s poor predictive ability because of the less complexity captured in the training set.

Under sampling: Removing examples form the mail class in an imbalanced dataset.

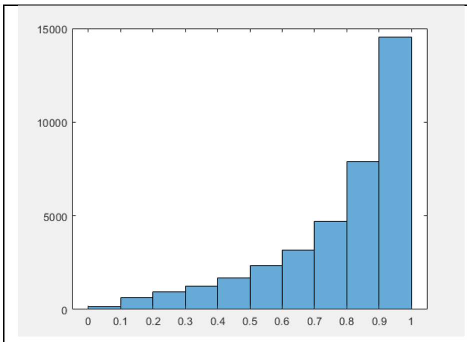
Validation: A process to evaluate the ML model quality.

Weight: Coefficient for a feature in a linear model**.**

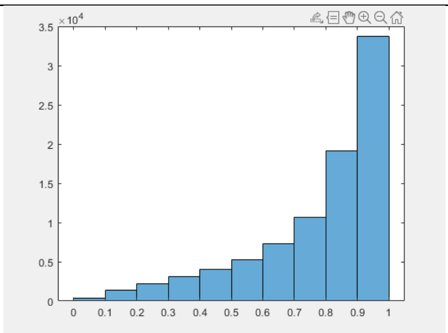
2.0 : Intermediate results:

Attempted to split the dataset as per the reference with the non standardised dataset, which produced various errors In the training including high level of overfitting and bias to the target variable.

High variance between test and train data

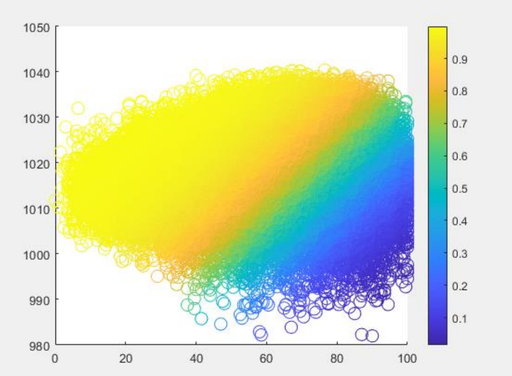


Train data

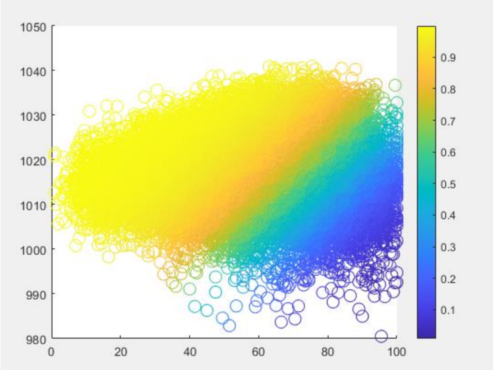


Test data

For the training evaluation : the class imbalance and before standardisation actus very poorly hence it doesn’t train particularly well. Hence normalised data is required for better evaluation



Training evaluation



Test evaluation

Hyperparameter tuning:

The hyperparameter tuning didn’t produce any significant increase in accuracy owing many to the non cross validated events in the data.

3.0 : Implementation details :

Data pre processing: The pre-processing is executed through Python using pandas and NumPy library mail for splitting the dataset and calculating mean, median and dropping the various features in the dataset.

Data feature engineering : mainly used for various EDA activities like creating dummies, moving e target variable to the end used in python, A range of lambda commands used

For data normalisation MATLAB is used and for creating a pure nice balanced dataset

Model selection: The process of choosing logistic regression and random forest

The idea to predict rain was the main aim here and the major paper found was by Parislab @ UCLA used multinominal logistic regression for their dataset. Hence the same approach was used to predict for the same.

Random forest was chosen mainly for it fit ensembling feature where one could predict linearity within the model.

For model building MATLAB is as discussed, used

Logistic regression : mnr fit the multinomial linear regression model is used mainly to implement the choices as required. Sigmoid function is used and basic beta functions are used to determine the accuracy of the model[3]

Random forest: It is developed in MATLAB as well Using fit ensemble methods and confusion matrix is used for the true positive and true negative features.

A precision and recall is used to determine if the findings are accurate.

The poster is done using MS publisher using MS pain snipping tool to snip an clip the pictures from MATLAB and the appendix is done using MS word.

The use cases of the said rain tomorrow predicted model is mainly in the meteorological industry where the usage of the same is at the peak and it can be determined for various type of metrological findings.

4.0 : References :

[1] *Machine Learning Glossary* (2018) *Google.com*. Available at: https://developers.google.com/machine-learning/glossary.

[2] Kodratoff, Y. (1988) *Introduction to machine learning*. Translated by S. Thorp. London, England: Routledge.

[3] Kleinbaum, D. G. and Klein, M. (2010) *Logistic regression: A self-learning text*. 3rd ed. New York, NY: Springer