**NAME: Wairiuko Samuel**

**Mini Project Report for Big Data**

**DATE:2020/05/10**

**TITTLE: Binary Classification using Logistic Regression and Random Forest in APACHE Spark (Seattle Weather 1948-2017).**

Besides coffee and technology companies, one of the things that Seattle is most well-known for is how often it rains. This is APACHE Spark project about building a binary classification model using Logistic regression and Random forest models for predicting whether it will rain on a specific day given information on the previous days of daily rainfall patterns from January 1st, 1948 to December 12, 2017. The dataset was downloaded from [www.kaggle.com](http://www.kaggle.com) and contains 5 variables and 25551 observations. The variables include the target variable “RAIN” other variables includes; DATE, PRCP, TMAX and TMIN.

1. **PROPERTIES OF DATASET**

* DATE (the date of the observation)
* PRCP (the amount of precipitation, in inches)
* TMAX (the maximum temperature for that day, in degrees Fahrenheit)
* TMIN (the minimum temperature for that day, in degrees Fahrenheit)
* RAIN (TRUE if rain was observed on that day, FALSE if it was not)

1. **OBJECTIVES**

* To build a precise model for predicting “Rain”.
* To conduct descriptive analysis, model Accuracy, recall and precision
* To report on the results and model performance comparisons and inferences/conclusions

**3.0 DATA PREPARATION, CLEANING & ANALYSIS**

The following were various steps followed in developing a pySpark model in a binary classification;

* Initialized a spark session
* Downloaded and read the dataset
* Checked for the missing values and did Descriptive analysis
* Encoded the target column with Label encoder
* Split the data into train and test sets
* Loaded the logistic regression and Random Forest Libraries
* Created Param Grid for Cross Validation and run the cross validations
* Used the test set to measure the accuracy of the model and Performance evaluation

1. **DATA ANALYSIS & SUMMARY FINDINGS**

**4.1 descriptive analysis, model Accuracy, recall and precision**

The dataset had 5 variables and 25551 observations. The model correctly identified when the rainfall was observed and when no rain was observed with 100% accuracy.

0=No rain,1=Rain

|  |  |  |  |
| --- | --- | --- | --- |
| RAIN encoded | prediction | count | accuracy |
| 0 | 0 | 2928 | 100% |
| 1 | 1 | 2173 | 100% |
|  |  |  |  |

|  |  |  |
| --- | --- | --- |
| CLASSIFICATION (Seattle Rainfall) | | |
|  | Logistic Regression | Random Forest |
| accuracy | 1 | 1 |
| recall | 1 | 1 |
| Precision | 1 | 1 |

1. **CONCLUSIONS**

* Both Random Forest and Logistic Regression classifiers were able to predict whether it will rain on a specific day given information on the previous days with 100% accuracy.
* In building the models the date column was not included because it was not a meaningful variable.
* The classifier was able to predict with 100% accuracy that there will be rain and there was rain (TP) and there were no misclassification (FP)
* The classifier was able to predict with 100% accuracy that there will be no rain and there was no rain (TN).

**5.1 RECOMMENDATIONS/INFERENCES**

* Both models produced good results in prediction and therefore the management needs to implement both for result comparison.
* Management should employ great team to implement great models. This will help to make large projects successful.
* Only meaningful variables should be included while doing the modelling this helps to improve the model accuracy.

**REFFERENCES**

* www.kaggle.com {open data set}
* APACHE Spark Python lecturer Notes and Codes.