BINARY CLASSFICIATION ON CLUSTERED DATA

by

Kakoli Chowdhury, B.Tech, University of Rajasthan, India, 2011

A Major Research Paper

presented to Ryerson University

in partial fulfilment of the requirements for the degree of

Master of Science

In the Program of

Data Science and Analytics

Toronto, Ontario, Canada, 2018

© Kakoli Chowdhury 2018

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PAPER (MRP)

I hereby declare that I am the sole author of this Major Research Paper. This is a true copy of the MRP, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

Kakoli Chowdhury

BINARY CLASSFICIATION PREDICTION ON CLUSTERED DATA

Kakoli Chowdhury

Master of Science 2018

Data Science and Analytics

Ryerson University

**ABSTRACT**

Land-Mobile radio systems support many vital communication functions supporting government and private operators, some related to public safety and mission critical functions. The models produced will help in understanding the usage patterns at different time periods to predict occupancy and demand by different channels across the spectrum.

CRC (Canadian Research Corporation) is providing Layer 1 data sampled every three milliseconds. This data is further explored and processed under this MRP. Sub-setting of data is conducted based on clustering and descriptive statistical analysis designed to differentiate between channels exhibiting different occupancy % patterns.

Applying algorithms on clustered data is expected to show distinct behaviors that are further utilized to find the best prediction model for spectrum availability.

**Keywords:**

Spectrum; Channel Usage; Occupancy

Table of Contents

[a. Structured Abstract 6](#_Toc521535764)

[b. Introduction 1](#_Toc521535765)

[c. Background and Literature Review 1](#_Toc521535766)

[Exploratory Analysis 1](#_Toc521535767)

[Literature Review 3](#_Toc521535768)

[d. Methodology 4](#_Toc521535769)

[Data Cleanup 5](#_Toc521535770)

[Clustering 5](#_Toc521535771)

[Selection of Response Variable 6](#_Toc521535772)

[Choice of Factors 6](#_Toc521535773)

[Experimental Design 7](#_Toc521535774)

[e. Results and Discussions 8](#_Toc521535775)

[Dataset Preparation 8](#_Toc521535776)

[Selection of the Threshold- Baseline 9](#_Toc521535777)

[Actual Experiments 10](#_Toc521535778)

[Conclusions and Future Work 11](#_Toc521535779)

[References 11](#_Toc521535780)

**List of Figures**

Figure 1: Attributes availability in layer 3 data 2

Figure 2: Histogram of either occupied or unoccupied channels 2

Figure 3: Histogram of Unoccupied channels 2

Figure 4: Outcome for unsupervised clustering 9

Figure 5: Sensitivity vs 1-specificity curve for Naive Bayes 9

Figure 6: Sensitivity vs 1-specificity curve for Logistic Regression 9

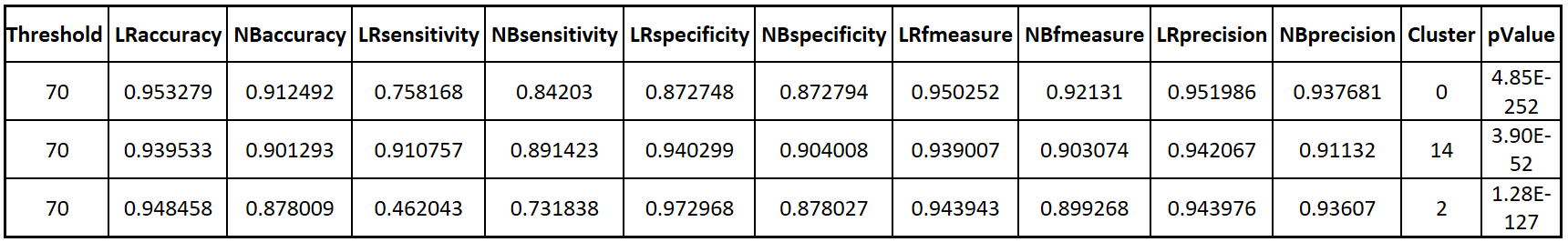
Figure 7: Threshold vs F-measure between two prediction algorithms 10

Figure 8: Threshold vs Accuracy between two prediction algorithms 10

Figure 9: Results comparison between baseline and clustered data sets 10

# a. Structured Abstract

1. **Background**: Canadian Radio Corporation (CRC) provides Layer 3 data that provides channels and their occupancy. There is a need to explore the area of dynamic spectrum allocation through predictive modeling. Existing work on application of predictive classification algorithms on clustered data is very limited.
2. **Aim**: This MRP aims at studying the behavior of linear binary classification algorithm on the clustered data, taking into consideration the distinct behavior shown by each clusters and their effect on occupancy prediction.
3. **Methodology**:
   1. **Data cleanup**: task involves removing the channels that do not contain enough data to become useful in prediction model. 7000+ channels were reduced to approximately 2300 after cleanup.
   2. **Clustering**: divided the channels into three separate clusters using unsupervised K-means clustering method. The outcome of the cluster resulted in Occupied, Partial, either occupied or unoccupied.
   3. **Data labelling**: for the experiment has been done manually. Occupancy percent more than threshold is given label 1 otherwise 0.
   4. **Finding the threshold**: running the first experiment to identify the threshold where maximum accuracy is achieved using Logistic regression and Naïve Bayes algorithm. Graph of accuracy vs threshold is plotted to find the right threshold.
   5. **Setting the baseline**: on the selected threshold, experiment is ran on random 300 channels out of 2300. The data used is on Layer 3 data. K-fold is value taken 10 as well as number of iterations were set to 10. This experiment is ran on data without clustering.
   6. **Experiments**: same steps are ran on data set for cluster 1 and 4 and a second experiment on cluster 2 keeping all other parameters same as baseline. Results are noted down and compared.
4. **Results**: shows the comparison of the experiment outcomes between different clusters. For each prediction algorithm, accuracy, sensitivity, specificity, f-measure and precision has been calculated and compared.



1. **Conclusions:** following conclusions can be drawn from the results of the experiment.
2. Logistic regression performed consistently better in terms of accuracy as compared to Naïve Bayes with or without clustering.
3. Across different data sets, accuracy of an occupancy prediction within a given model didn’t change much.
4. It can also be concluded that clustering did not add any specific benefits in binary classification prediction when compared with non-clustered data

# b. Introduction

The purpose of this document is to capture the methodology and findings observed when preprocessed clustered data is passed through binary linear classification prediction models. This document explores the preprocessing methods, data filtration techniques, setting up and running the experiments. Some previous work done in this area has also been explored. Any existing work helps further establishes the validity of the results. Controllable and uncontrollable factors through the experimentation process have also been discussed. Toward the end, detail results along with drawn conclusions will be presented.

# c. Background and Literature Review

## Exploratory Analysis

**Channel Occupancy Analysis**

The data analysis is for conducting Channel Occupancy analysis across different frequencies at different time periods. Using the channel occupancy data, the task is to come up with sub data sets with channels that are fully occupied, partially occupied, either occupied or unoccupied.

**Data Set**

The data set consists of the following columns, StartTime, EndTime, Channel, Power\_dbm, SNR, and Occupancy. The first task was to check the quality of data set.

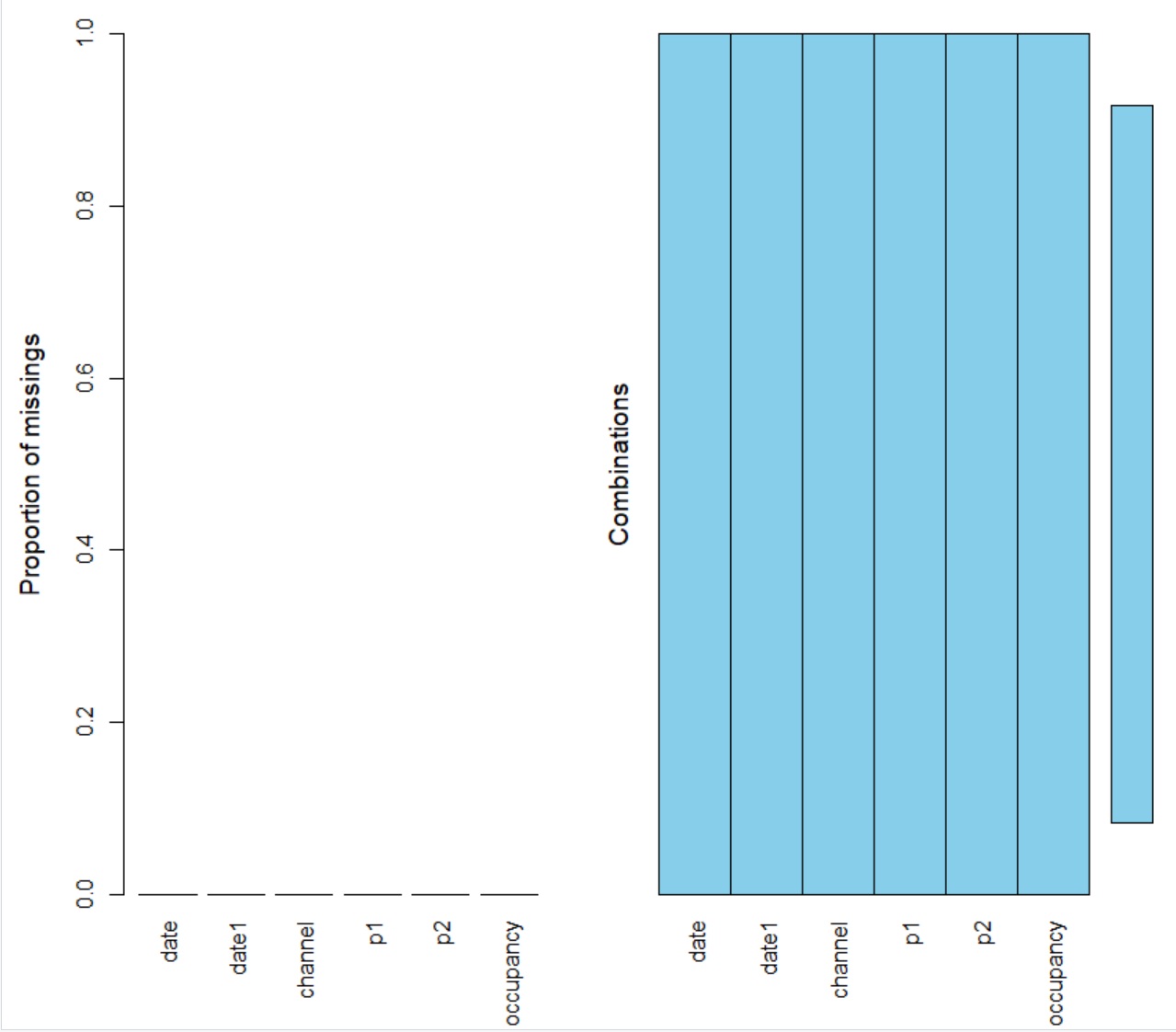


Figure 1: Attributes availability in layer 3 data

The dataset looks good. The next step is to visualize the data by plotting histograms.

**Sample Data Visualization**

Take a sample of the dataset and plot histograms to visualize data to get a base understanding of different categories.

|  |  |
| --- | --- |
| Figure 2: Histogram of either occupied or unoccupied channels | Figure 3: Histogram of Unoccupied channels |

**Descriptive Statistics**

For further understanding of the data, descriptive statistics analysis was performed. The following descriptive statistics will be generated

* Mean
* Median
* Mode
* Variance

Through this we obtain centralization, variation and distribution of channel occupancy data. When these stats were applied on layer 3 data, data points were reduced to one per channel. On this dataset, some exploration for the clustering was performed. However, not as great were obtained in the first iteration. To improve upon the results, these data points were further processed to find out following additional attributes:

1. mean\_median: median subtracted from mean
2. mode\_mean: mean subtracted from mode
3. median\_mode: mode subtracted from median

Further to it, clustering was performed on the new dataset which is explained in more detail in the next section. Comparison of the dataset based on different sets or variations of descriptive statistics and graph analysis was performed to observe any patterns.

**Clustering**

Cluster analysis is performed using k-means clustering for channel grouping. Once data is clustered profiling and interpreting of the segments is conducted to ensure robustness of clustering method and the input variables in the dataset. The outcome of the clusters resulted in three different clusters namely- partial, unoccupied, either occupied or unoccupied. Details of the clustering has been discussed in methodology and experiment design.

Applying descriptive statistics and clustering helped in analysing the large dataset and segment channels based on their occupancy.

## Literature Review

In article by Broos et. Al, the authors investigate the problem of classifying the sources of galactic x-rays burst into one of the four classes by assigning individual scores using Naïve Bayes is used to classifier [1]. Attributes and their properties that contributes to the classification includes individual source X-ray, visual, infrared and project position that are known to clusters. Some of these properties, for e.g. infrared J-band flux shows bimodal distribution that originates from stars at average distance. The paper uses the method to maximize the likelihood of witnessed data with reference to its association with distinct classes. Likelihood approach clarifies and quantifies stated notions regarding the classification. The article further assigns likelihood and prior to each source property to calculate posteriors.

In the article by Srivastava et. Al, the authors describes generative model of data that combines multiple modalities. The outcome represents a unified joint density information view by fusing input multi-modalities together [2]. In the given scenario, logistic regression is used for classification of the mixture of these representations. Logistic regression is used to obtain MAP information for all these representations views classifications.

In the article by Penet et. Al, the authors explores the concept and definition of violence and its detection using multimodal information integrated with time based information. [3] The paper focusses on two score based structure learning algorithms: naïve Bayesian and forest augmented Naïve Bayes. It concludes that late fusion of modalities perform better than early fusion

In this article Tor D. Tosteson and Bernard Rosner discusses conditional logistic regression models for clustered binary data [4]. Data is collected through a study of familial aggregation and sleep disorders. The author showed distribution of clustered members follows conditional logistic regression model proposed by Qu et al. (1987). This strategy helped identifying individuals with known disease status and then determine the status and other characteristics of remaining family members.

**Conclusion**

The purpose of this review was to find earlier researches and implementation done in the areas of predicting outcomes using linear modeling algorithms on unimodal, bimodal and multi-modal distributions of data. Also, it was evident that not much has been explored in terms of applying binary linear classification from the clustered data. Some of these research papers have touched upon linear models for the predictions but no significant outcome has been documented so far.

Through the medium of this MRP I propose to extend the research in areas of implementing naive Bayes and logistic regression on different distribution of the dataset as well as different clusters and identify their behaviors and commonalities.

# d. Methodology

This section lays out in detail the methodology taken to design and conduct the machine learning experiment to predict the occupancy for channels. There are multiple parts in this section where individual approaches to data cleanup, selecting response variable and experiment setup and design has been discussed.

## Data Cleanup

The very first step while designing the experiment is to perform analysis on channels to find out which channels offer more consistent data for across different months. This would ensure channels with less data are eliminated as they do not affect the prediction algorithm by creating imbalance.

The method used to eliminate such channels follow these steps:

1. For a given channel
   1. Subset layer 3 data
   2. Create index on the Start Time field
   3. Find missing index values between the min(Start Time) and max(Start Time)
   4. Calculate % of missing data:

(Number of data points) / (Number of missing + reported data points) \* 100

1. Once this activity is completed for all ~7900 channels, take the 75th percentile and eliminate all the channels below this percentile.

Outcome: the steps above lead to elimination of 2/3rd of the channels reduced the channels to approx. 2400. These channels contain consistent number of data points and we will move forward with remaining experiment taking only these channels. These 2400 channels formed the basis of the rest of the experimentation data.

## Clustering

On the cleaned dataset, the outcome of descriptive statistics were used to build unsupervised clusters. Following four attributes were used to identify clusters:

1. var\_channel: variance of the channel
2. mean\_median: median subtracted from mean
3. mode\_mean: mean subtracted from mode
4. median\_mode: mode subtracted from median

kmeans in-built function in R was used to perform clustering and following parametric values were used:

1. Centers: Value of k, which is number of clusters, were taken as 4.
2. iter.max: the maximum number of iterations where convergence took place. In our experiment, value was taken as 10.
3. Nstart: number of random set chosen to be at 4

Unsupervised clustering was implemented on approximately 2400 channels from layer 3 data and the result was a group of four clusters were identified:

1. Either occupied or unoccupied: cluster 1
2. Partial: cluster 2
3. Unoccupied: cluster 3
4. Either occupied or unoccupied: cluster 4

**Explanation of each cluster:**

1. **Either occupied or unoccupied**: these are the clusters where most of the data lies either towards the low occupancy, for e.g. less than 20% or towards the higher side with more than 80% occupancy. Some channels will fall between 20% and 80% as well.
2. **Partial**: this cluster represents all the channels were occupancy is similar to normal distribution.
3. **Unoccupied**: clusters have occupancy percent at 20% or below. Although they are meaningful for quick filtration of the data, they do not serve any purpose in the prediction algorithms due to lack of data availability.

## Selection of Response Variable

The problem at hand requires us to predict the binary occupancy of the channel i.e. given a percent of channel’s occupancy, if it falls below a threshold we call it unoccupied otherwise occupied. Thus, the response variable is the ‘Occupancy percent’.

The next step in the process is to annotate data by creating labels for the dataset. We need to label the data manually since there is no labeling already available. In the beginning, right threshold is unknown. In order to find an optimum value, threshold will be set to a range of value from 0 to 100 inclusive with each iteration raised by 10 than the previous one. For e.g.

1. Starting from one extreme (zero) and labeling all values as 1
2. Other extreme being (100) and label all values as 0.
3. Other values taken: 10, 20, 30, 40, 50, 60, 70, 80, 90

In between values will be adjusted and curve will be designed to find the optimum threshold through manual inspection.

## Choice of Factors

In the experiment, few controlled and uncontrolled factors are known. Below is an outline for them:

**Controllable factors**

1. Algorithms: Naïve Bayes Classifier and Logistic Regression will be used as prediction algorithm. The choice of these algorithms have been discussed previously in the literature review.
2. Hyper parameters: Another controllable factor is the threshold value used to label class data into binary occupied (1) and unoccupied (0). This value can be controlled through experiments and the most optimum value can be used to conduct the experiments.

**Uncontrollable factors**

1. Missing data: there is missing data due to lack of reading or equipment failure or any other unknown reasons. Missing data cannot be inferred and will be removed from the dataset.
2. Noise in the data: there may be some noise in the data introduced due to some unknown factors while measuring the data points. However, finding such scenarios are not in the scope of this project.

## Experimental Design

1. One of the basic principle followed in the experiment design is the randomness of experiment. For e.g. dividing the data into train and test is done without prior knowledge or any built-in rules.
2. To effectively conduct experiment, the cleaned dataset will be divided into train, test and validation set. The purpose of each of the dataset is as follows:
   1. Train: the actual dataset on which training will take place
   2. Test: will be used to test the model in an unbiased manner. This is used once training of the model is fully completed.
   3. Validation: gives an estimate of model while in training. This set can be used to tune the model further.
3. K-fold: to implement the concept of train, test and validation dataset k-fold method will be implemented. In this methodology, dataset will be divided into K-equal validations sets. Value of k will be chosen as 10 for the purpose of this experiment. Each K-fold will go through ten iterations and the results will be the average of the results from each iteration.
4. Setting the experiment to find threshold with different configurations:
   1. Threshold value: for the class label is one of the configuration that will be found out through manual inspection of the curve between threshold vs accuracy as well as threshold vs f-measure. (Refer to Figure 7 and 8)
   2. Since there is a strong correlation between power\_dbm and SNR, these parameters were dropped from the prediction model. The resultant data contains only Start Time, End Time and Occupancy label. To make the prediction more meaningful, temporal occupancy was introduced in the dataset, which is explained in more detail in the next step.
   3. Temporal occupancy: since the available data is time range based, the given data points could be made more meaningful if previous hours are also considered while predicting the occupancy for current hour. Additional data will be processed in our main data set to include previous six hours of data. Whenever there is no reading available, that data point will be treated as NA. All data rows with NAs will be removed from the final data set.
5. Once threshold value is known, three separate data structures will be created and the prediction algorithms will be executed as follows:
   1. First dataset is a list of three hundred channels picked randomly out of 2400 channels that have more than 75th percentile of data. This dataset will serve as the baseline.
   2. Second dataset is a list of all channels that belong to either occupied or unoccupied cluster (named 1 and 4 in our experiment)
   3. Third dataset is a list of all channels that belongs to occupied cluster (named 2 in our experiment)

Each dataset goes through the list of respective channels on the selected through threshold. Datasets will be divided into train, test and validation sets with K-fold value as 10. For each run, a confusion matrix is created and certain performance measures are noted as below:

1. Accuracy
2. Precision
3. F-measure
4. Sensitivity
5. Specificity

These parameters are measured and stored in a csv for both algorithms and results are compared in the end.

# e. Results and Discussions

## Dataset Preparation

As part of data preparation, descriptive statistic experiments were ran on the entire dataset of 7000+ channels to find out inherent properties of the channels. The chosen method to uncover commonalities in the channels was unsupervised clustering. Clustering was ran on all 7000+ channels available in the Layer 3 dataset for the month of October.

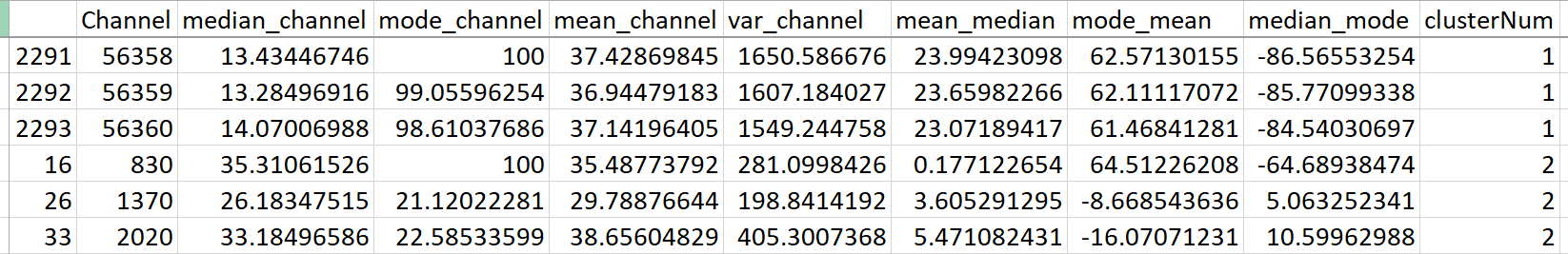


Figure 4: Outcome for unsupervised clustering

**The outcome of the cluster is as follows:**

1. Cluster 1: either case (occupied or unoccupied)
2. Cluster 2: partial case
3. Cluster 3: unoccupied case
4. Cluster 4: either case (occupied or unoccupied)

## Selection of the Threshold- Baseline

A baseline experiment was conducted on the random 300 channels selected from a list of 2300 channels. These 2300 channels contained data more than 75th percentile of the channels. Logistic Regression and Naïve Bayes prediction models were ran on the data and outcome of the baseline experiment are as follows:

1. Curves for Logistic Regression and Naïve Bayes

|  |  |
| --- | --- |
| Figure 5: Sensitivity vs 1-specificity curve for Naive Bayes | Figure 6: Sensitivity vs 1-specificity curve for Logistic Regression |

1. F-measure and Accuracy v/s Threshold for Logistic Regression and Naïve Bayes

|  |  |
| --- | --- |
| Figure 7: Threshold vs F-measure between two prediction algorithms | Figure 8: Threshold vs Accuracy between two prediction algorithms |

Based on the experiments above, threshold was selected to be ‘70’. All selected clustered data frames will be ran on threshold value of 70.

## Actual Experiments

Once baseline threshold was obtained, three further experiments were conducted. One of each data frame as mentioned in #5 Experimental Design. Results of those experiments are indicated in the table below:

1. Cluster 0 indicates the results when ran on random channels taken regardless of any cluster
2. Cluster 14: indicates the results when ran on the channels taken from cluster 1 and 4
3. Cluster 2: indicates the results when ran on the channels taken from cluster 2

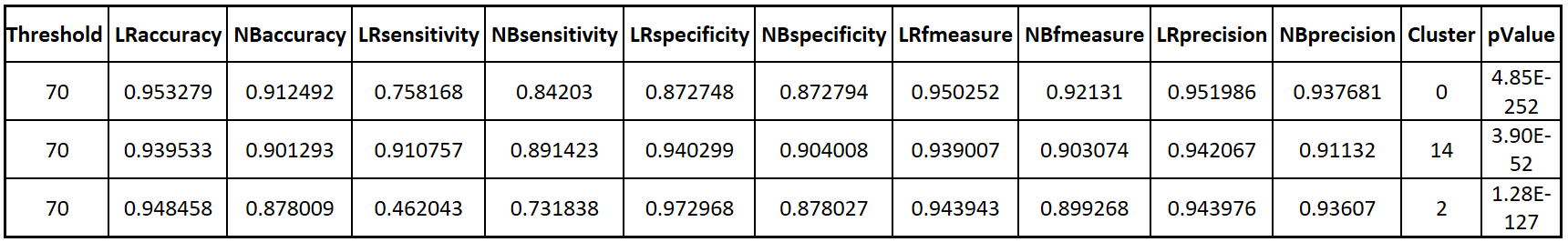


Figure 9: Results comparison between baseline and clustered data sets

**Attributes in the results table:**

1. Threshold: Selection of the threshold value is 70.
2. LRaccuracy: value of the accuracy with Logistic Regression
3. NBaccuracy: value of the accuracy with Naïve Bayes
4. LRsensitivity: value of the sensitivity with Logistic Regression
5. NBsensitivity: value of the sensitivity with Naïve Bayes
6. LRspecificity: value of the specificity with Logistic Regression
7. NBspecificity: value of the specificity with Naïve Bayes
8. LRfmeasure: value of the f-measure with Logistic Regression
9. NBfmeasure: value of the f-measure with Naïve Bayes
10. LRprecision: value of the precision with Logistic Regression
11. NBprecision: value of the precision with Naïve Bayes
12. Cluster: indicates the cluster
13. pValue: indicates the t-test independent value when accuracies of Logistic Regression and Naïve Bayes are calculated

# Conclusions and Future Work

1. In general, between two prediction models Logistic regression performed consistently better in terms of accuracy as compared to Naïve Bayes in all three experiments. p-Value is the t-test independent calculated for accuracy between logistic regression and Naïve Bayes. P-Value is positive closer to zero therefore null hypothesis rejection is valid.
2. Across different data sets with and without clustering, accuracy of an occupancy prediction within a given model didn’t change much across different datasets.
3. Naïve Bayes performed better in terms of sensitivity whereas Logistic Regression performed better for specificity. Partial Cluster 2 has achieved best specificity.
4. Logistic regression performed better than Naïve Bayes in while measuring f-measure and precision. However, all the clusters showed worse results when compared to non-clustered data for f-measure and precision.

Overall, it can be concluded that clustering did not add any specific benefits in binary classification prediction when compared with non-clustered data. The overhead of running the data through clustering does not gain us any significant results as compared to non-clustered data. It is only adding computational complexity.

One thing that can be improved in the future could be automated labeling of the data. Rather than manual labeling which could sometimes become rigid and have overhead, automated labeling may result in better prediction outcomes. This area could be explored further.

# References

[1] Broos, P. S., Getman, K. V., Povich, M. S., Townsley, L. K., Feigelson, E. D., & Garmire, G. P. (2011). A naive Bayes source classifier for X-ray sources. *The Astrophysical Journal Supplement Series*, *194*(1), 4.

[2] Srivastava, N., & Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems* (pp. 2222-2230).

[3] Penet, C., Demarty, C. H., Gravier, G., & Gros, P. (2012, March). Multimodal information fusion and temporal integration for violence detection in movies. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on* (pp. 2393-2396). IEEE.

[4] Tosteson, T., Rosner, B., & Redline, S. (1991). Logistic Regression for Clustered Binary Data in Proband Studies with Application to Familial Aggregation of Sleep Disorders. Biometrics, 47(4), 1257-1265. doi:10.2307/2532384