Comparing Accuracy of Models Built on Subpopulation vs. Entire Population

Sudha Subramanian  
*DCS Information Systems  
E-mail:* [*igi\_sudha@yahoo.com*](mailto:igi_sudha@yahoo.com)

# Abstract with key words

**Abstract:** This paper examines how the results of combining models built on a subpopulation (MSP) compares to that of a model built on the entire population (MEP) when applied across different modeling algorithms. Candidates with similar characteristics are grouped together by applying clustering techniques on the datasets. It is seen that patterns within each subpopulation aid to enhance the accuracy of the individual models. An increase in accuracy by 1%–3% can be substantial when building predictive models. This increase in accuracy is achieved without over-fitting to the training set.

**Keywords:** Clustering, Accuracy, Increase, Subpopulation, Overfitting

# Introduction

The biggest challenge of any Data Science problem is to build a model that is simple and generalizes well with newly presented data. The purpose of this study is to show how the accuracy and fitness of a model can be enhanced by applying Clustering techniques across different modeling techniques. A comparative study was performed by applying the “Cluster-Then-Predict” methodology on models built using Linear Regression and Logistic Regression techniques. The study was done on three disparate datasets to show how 1%–3% increase in accuracy was observed on Logistic Regression models or up to 2% decrease in error was observed on model built using Linear Regression.

The paper illustrates through the results of this comparative study that the increase in accuracy (or decrease in error) is gained without overfitting the model to training data. This is supported through visualizations that illustrate the process of clustering and how this technique leverages the advantages of underlying patterns in the data.

# OBJECTIVE of the study

The main objective of this study is to show how unsupervised learning techniques could be used to identify underlying patterns in the data, and how unsupervised learning when combined with supervised learning can help build a more robust model. This study also illustrates how selection of features specific to data within each cluster could further enhance the accuracy of the overall model. Although we would expect that decision trees (supervised learning) could easily incorporate the defining features of a cluster into the first few levels of the tree, empirical evidence shows that the most common forms of decision trees do not implement this behavior.

# CLOSELY RELATED RESEARCH WORK

D2Hawkeye, Inc. (a subsidiary of Verisk Health, Inc.) is a company that develops data service solutions primarily for the healthcare industry. One of the projects, on which D2Hawkeye worked with researchers at Massachusetts Institute of Technology [1], presents an approach of grouping the subjects into buckets based on costs incurred. Their data mining approach primarily included partitioning based on members’ costs into five different bands or cost buckets. They were able to demonstrate how clustering algorithms organize objects within each bucket so that similar subjects are grouped together in a cluster and dissimilar objects belong to different clusters.

Details of this paper can be found at:

<http://people.csail.mit.edu/gjw/papers/healthcare.pdf>

Using this as the premise, a comparative study across multiple datasets was performed by applying different modeling techniques, which is described in this paper. For example, the Medicare Reimbursement dataset (one of the datasets used for the case study described in this paper) comprised of binary flags to indicate occurrence of specific illnesses during the year and the actual reimbursement costs for the current year. The goal was to build a model that would predict the reimbursement costs for the following year based on occurrences of illnesses for the individual. Since this is a linear regression problem, linear regression models were built on the entire population as well as on subpopulations (clusters). The overall performance of the models built on subpopulation was compared against that of the model built on the entire population.

This helped trigger a thought to convert this into a logistic regression problem – basically, to understand how cluster-then-predict approach would work when applied to other modeling techniques. Applying the same technique, accuracy of models built on entire population was compared with that built on subpopulations. In this case as well, overall accuracy was higher in models built on subpopulations, where clustering technique was applied. The same was extended to multiple datasets, and the findings are documented through this paper.

# increasing accuracy & fitness through clustering

The process of Clustering, is usually seen as a pre-processing step that works in conjunction with a feature extraction process. In this process, an unsupervised learning process is used to learn about the data and form smaller sets on which a supervised modeling technique is applied.

## Pre-processing / Feature Extraction

As in any model-building exercise, pre-processing is applied on the data to handle any missing values. Also, redundant data is removed from the dataset in order to reduce the dimensionality of the feature set. Correlation plots are useful to help understand correlation between variables in the dataset. New features, as identified based on the data are included if they are found to add value to the process.

Data is divided into training (70%) and testing (30%) datasets by applying random selection on the target variable, so that the target variable is distributed well in the two datasets.

In this case study, modeling is performed on the entire population and its accuracy is compared with the accuracy of models built on subpopulations (after clustering is applied). This is applied across three different datasets to understand how the accuracy of the model built on entire population (MEP) compare to that of models built on subpopulation (MSP).

## Model Built on Entire Population (MEP)

Model built on the entire population involves building a supervised model on the entire training dataset. The trained model is then applied on the testing dataset to predict the target variable. Accuracy of the model is calculated based on the number of predictions that match the actual outcome.



Figure 1 Overview of Modeling Entire Population

## Models built on Subpopulation (MSP)

Models built on subpopulation involves grouping data into appropriate clusters as identified based on the type of clustering algorithm and building supervised models on each subpopulation of the training dataset. The number of clusters to use when grouping the data into is selected based on the data or with the help of a dendrogram / elbow method approach. As a first step, data is normalized and the target variable is excluded prior to applying clustering. Data normalization is performed, so that all variables are given the same importance. For this study, clustering is applied through a k-means algorithm.

Models are built on each of these clusters using the appropriate modeling technique. The trained models are then applied on the testing datasets corresponding to each of the clusters. Accuracy is calculated for each model built on subpopulations and overall accuracy is computed by combining the results of the individual models. This hybrid approach of combining unsupervised and supervised modeling techniques has shown to result in higher accuracy compared to that of the single model built using only supervised modeling technique.



Figure 2 – Overview of Modeling Subpopulation

## Datasets Explained

For this case study, the following datasets were used to build and compare accuracy between MEP and MSP. While the Energy and Stock Returns datasets were required to predict a binary outcome (logistic regression), for the Medicare dataset the actual Medicare Reimbursement costs need to be predicted (linear regression). In addition, the Medicare dataset was converted to a format to predict if the reimbursement costs would be higher in the following year for the purpose of this case study.

Table 1 – Study repeated across separate datasets

|  |  |  |
| --- | --- | --- |
| Dataset | Dataset Explained | Independent Variables |
| Medicare Reimbursement | Individuals who are patients in the U.S. Medicare program | * Age * Binary variables, indicating if patient had diagnosis for the disorder in a particular year * Total amount of Medicare reimbursements for that year |
| Energy | Energy data for all states in the U.S. from 2000 through 2013 | * State * Year * Generation from various sources * Prices for various sectors * Sales for various sectors * Financial/Regulatory incentives * Demographic information |
| Stock Returns | Returns for company’s stock between 2000 and 2009 | * Returns for company’s stock for the first 11 months of the year |

## Summary of Results

Table 2 illustrates how the accuracy of models MEP and MSP compare across all datasets.

Table 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Dependent  Variable | Modeling Technique | Results of MEP  Single-Model Approach | Results:  MSP compared to MEP |
| Energy | Increase in Solar Energy Generation  – Yes / No | Logistic  Regression | Accuracy = 81.91% | 2.8% increase in accuracy |
| Stock Returns | Increase in Stock Price  – Yes / No | Logistic  Regression | Accuracy = 67.71% | 1.2% increase in accuracy |
| Medicare Reimbursement | Higher Costs  – Yes / No | Logistic  Regression | Accuracy = 69.83% | 3% increase in accuracy |
| Medicare Reimbursement | Reimbursement cost for next year | Linear  Regression | RMSE = 1.849 | RMSE lower by 2% |

# CASE STUDY – explained

For the purpose of this paper, the Energy dataset is used to describe the process adopted for this case study.

## Data Insights and Visualizations

As a first step, correlation between the variables in the dataset was studied through the correlation plot. The correlogram is a graph of a correlation matrix and is a way of highlighting the most correlated variables in a data table. Positive correlations are displayed as solid black circles and negative correlations as solid red circles. Color intensity and size of the circle are proportional to the correlation coefficients. The plot below includes the target variable (GenSolarBinary) and select independent variables which were found to have some form of correlation with the target variable.

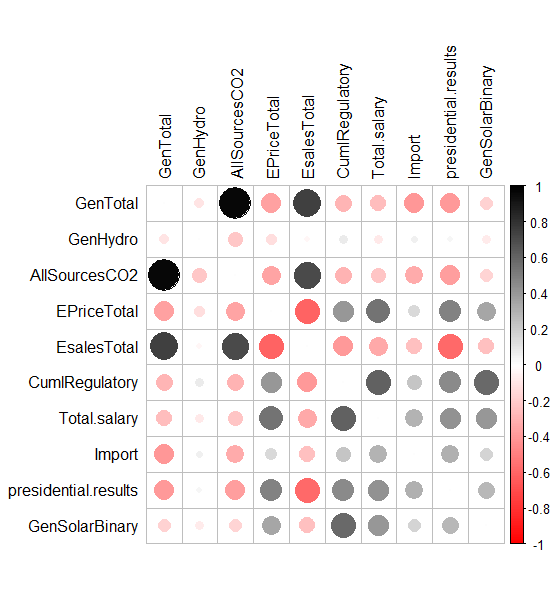


Figure 3 Correlation Plot

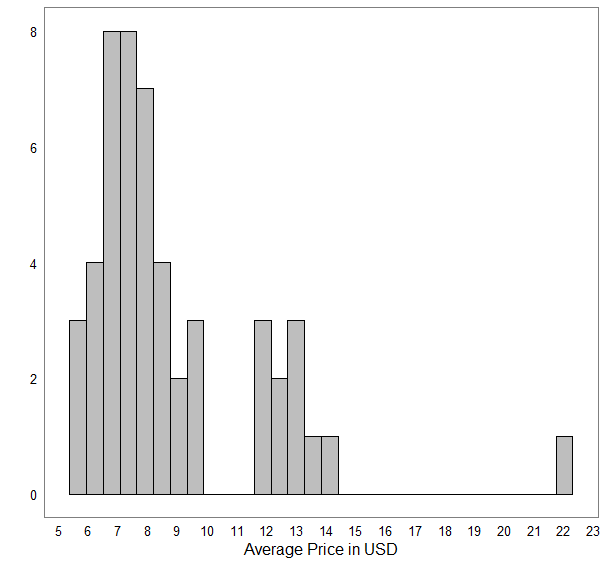
As a next step, a series of visualizations were done on the data to gain useful insights into the data. As an example, the following plot illustrates the different price buckets across all the states.

Figure - Histogram of Average Price across States

## Feature Selection

For the purpose of modeling, independent features were selected based on the visualizations and correlation plots as part of the data mining phase. These primarily include:

* Total Price
* Incentives (Financial & Regulatory)
* Party (Presidential Results)
* Annual Wages Per Capita
* Whether the State was an importer

## K-Means Clustering

For this study, a k-means algorithm was used for clustering, which requires that the value of “*k*” (number of clusters) be provided. The value of “*k*” is to be chosen such that the total intra-cluster variation is minimized and such that inter-cluster variation is differentiable.

The optimal number of clusters was assessed using the elbow method, which involves observing a set of the possible number of clusters relative to the minimality of the within-cluster sum of squares value. Authors Trupti M. Kodinariya and Prashant R. Makwana [2] describe the process of identifying the optimal number of clusters using elbow method in their paper “Review on determining number of Cluster in k-means Clustering”.

The following R code was used to plot the within-cluster sum of square (*wss*) for different values of *k* (number of clusters).

|  |
| --- |
| # Compute and plot wss for k = 2 to k = 10  k.max <- 10 # Maximum number of clusters  data <- train.norm  wss <- sapply(1:k.max,  function(k){kmeans(data, k, nstart=10 )$tot.withinss})  plot(1:k.max, wss,  type="b", pch = 19, frame = FALSE,  xlab="Number of clusters K",  ylab="Total within-clusters sum of squares") |

This produces a plot, as shown below, which helps identify the optimal number of clusters as *k*=3, as indicated by the red dotted line. For values of *k≥*4, the change in total within the sum of square value is minimal. Also, the mean values of the features show a reasonable level of variation when *k*=3.

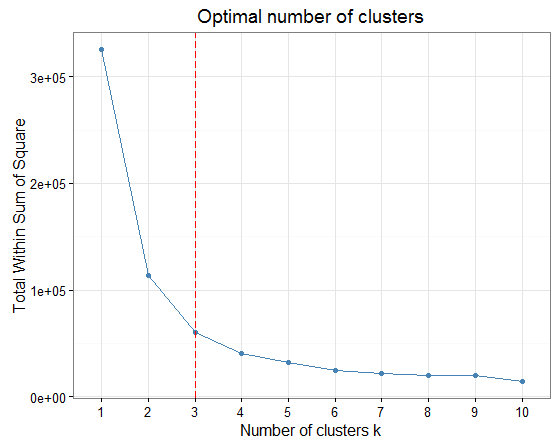


Figure 5 - Elbow Method (Optimal Clusters)

The steps involved in applying k-means clustering:

1. Derive the optimal number of clusters (*k*) through the elbow method and through manual verification of variation in mean values for different values of *k*
2. Normalize the data, so that all variables are given same importance
3. Exclude target variable from the set of variables based on which clustering is done, so that there is no bias in how the clusters are formed. Including the target variable defeats the purpose of performing the clustering
4. Associate the data points to the respective clusters in training and testing datasets

Using *k*=3, k-means clustering was applied on the normalized dataset and as seen in the table below, the mean values of the key variables in the three clusters show reasonable amount of variation.

Table 3 – Average values across clusters

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cluster | Financial | Regulatory | Salary | Results | Import |
| #1 | 11.993976 | 7.765060 | 20.90381 | 0.9939759 | 0.4698795 |
| #2 | 66.447761 | 17.447761 | 25.25605 | 0.8208955 | 0.5522388 |
| #3 | 8.835938 | 3.878906 | 18.09670 | 0.0000000 | 0.2070312 |

## Cluster Plots

After clustering was applied on the normalized dataset, training and testing datasets were identified across all three clusters. The training clusters were used to build the model, based on which predictions were done against the testing data corresponding to each cluster. Cluster plot [3] provides a visual representation of the clusters formed through the process of clustering across training and testing datasets. These plots provide a visual understanding of distribution of the data on the entire population and the subpopulations through the process of clustering.

Since our dataset is multi-dimensional, it was necessary to reduce dimensionality so that the clusters could be understood better. To understand distribution of the data points across training and testing datasets, a bivariate cluster plot using principal components was used to visualize the clusters. The ellipses indicate the clusters and are visually distinguished through color and based on how the data points are depicted.

The following shows the distribution of the data points in the ‘TRAINING’ dataset:

|  |  |
| --- | --- |
| Entire Population (TRAINING DATA): | Clustering Applied (TRAINING DATA): |
|  |  |

Figure 6 - Training Data Points

The following shows the distribution of the data points in the ‘TESTING’ dataset:

|  |  |
| --- | --- |
| Entire Population (TESTING DATA): | Clustering Applied (TESTING DATA): |
|  |  |

Figure 7 - Testing Data Points

As seen in the figure above, distribution of data in these clusters are similar between the training and testing datasets. This indicates that the models built on subpopulation will be able to perform well within each cluster, by leveraging the features specific to that population.

## Modeling – Entire Population & Subpopulation

The target variable for this dataset is binary and so, this is a Logistic Regression problem. Using the engineered features and through iteratively performing the modeling exercise, the best fit model was built on the Entire Population using the training dataset. The testing dataset was used to evaluate and score the model’s performance. The evaluation was made by identifying the count of predictions that match the actual.

After performing this on the Entire Population (MEP), the same exercise of modeling and predicting against the test dataset was repeated on the Subpopulations and the models thus built are referred to as MSP. Essentially, multiple models (three in this case) were built on each set of training and testing datasets specific to each cluster. Same set of features were used for modeling both MEP and MSP.

Subpopulation models were built on the clusters using the respective training datasets. Given that “predtest1”, “predtest2” and “predtest3” are predictions across the three clusters “test1”, “test2” and “test3”, the overall accuracy of MEP was calculated as:

|  |
| --- |
| # Combine all predictions / outcomes to get overall accuracy  AllPredictions = c(predtest1, predtest2, predtest3)  AllOutcomes = c(test1$GenSolarBinary, test2$GenSolarBinary, test3$GenSolarBinary)  table(AllOutcomes, AllPredictions >= 0.5) |

Overall accuracy of the model built on Subpopulations (MSP) was calculated by combining the predictions across all three models and identifying the count of predictions in the result set that match the actual. Comparing the accuracy of MEP and MSP showed that accuracy of MSP was higher by 1.4%.

Table - Accuracy of MEP and MSP

|  |  |  |
| --- | --- | --- |
| Modeling | Features used for modeling | Accuracy |
| MEP (Entire Population) | GenHydro, GenSolar, CumlFinancial, CumlRegulatory, Total.Salary, Import | 81.9% |
| MSP (Subpopulation) | GenHydro, GenSolar, CumlFinancial, CumlRegulatory, Total.Salary, Import | 83.33% |

In addition, the models built on Subpopulation were tweaked by adjusting the set of features. Since the clusters are smaller and are groups of data points with similarity in features, they are essentially treated as independent models and appropriate feature set may be chosen for each subpopulation. By adopting this technique, accuracy of MSP was higher than that of MEP by 2.4%.

Table - Accuracy of MSP (different features)

|  |  |  |
| --- | --- | --- |
| Modeling (MSP) | Features used for modeling | Overall Accuracy |
| MSP-1 (Cluster #1) | GenHydro, GenSolar, CumlFinancial, CumlRegulatory, Total.Salary, Import | 84.76% |
| MSP-2 (Cluster #2) | EPriceTotal, CumlFinancial, CumlRegulatory |
| MSP-3 (Cluster #3) | GenHydro, GenSolar, CumlFinancial, CumlRegulatory, Total.Salary, Import |

## Confusion Matrix

A Confusion Matrix is a representation of actual and predicted classifications as a result of modeling, especially for binary classification problem. Performance of the model is evaluated using the data in the matrix. Confusion Matrix helps understand the number of outcomes as:

* Correct Outcomes (True Positives & True Negatives)
* Incorrect Outcomes (False Positives & False Negatives)

A fourfold plot [4] was used to display the 2 x 2 confusion matrix to help provide a visual understanding of the accuracy of the models built.

|  |  |
| --- | --- |
|  |  |

Figure 8 Confusion Matrix: Entire Population vs. Subpopulation

## Validating the Case Study

Validation was performed by applying the approach to datasets with varying percentage of training/test data.

Results of these runs are captured in the table below:

Table 6 - Validation Results

|  |  |  |  |
| --- | --- | --- | --- |
| Sample % | Accuracy | ↑ Accuracy | Cluster Plot |
| Training = 50%  Testing = 50% | MEP – 86%  MSP – 87.7% | 1.7% |  |
| Training = 60%  Testing = 40% | MEP – 84.3%  MSP – 85.7% | 1.4% |  |
| Training = 70%  Testing = 30% | MEP – 81.9%  MSP – 83.3% | 1.4% |  |

As seen in the table above, an increase in accuracy in models built on Subpopulation is observed across different sample populations chosen between training and testing datasets.

The above results were obtained when the same set of independent features were used to train both MEP and MSP. When the features were adjusted in the individual models specific to each cluster, a further increase in overall accuracy was observed.

# CONCLUSION

This paper presents a hybrid approach of combining unsupervised and supervised learning techniques to achieve an increase in accuracy of the Model, without overfitting to the training data. As seen, the unsupervised learning process helps identify inherent patterns in the data and then apply supervised learning techniques to build the predictive models.

Similar results were observed for the Stock Returns and Medicare Reimbursement datasets. With the Medicare Reimbursement dataset, a Linear Regression model was built as the actual reimbursement costs needed to be predicted. In this case as well, RMSE (Root Mean Squared Error) was lower by almost 2%. For the modified version of the Medicare Reimbursement dataset where the outcome/target variable was converted to a binary value, an increase in accuracy by almost 2.7% was observed.

We would expect that decision trees could easily incorporate the defining features of a cluster into the first levels of the tree. However, empirical evidence shows that most common forms of decision trees do not implement this behavior. Outside of deep learning methods and advanced neural networks, most statistical models have limited adaptability to population nuance without over-fitting.

Through this technique, it was possible to realize a 1%–3% increase in accuracy, that is achieved without overfitting to the training data. This process has shown that using Linear / Logistic Regression in conjunction with clustering makes it a much more powerful method, as it is able to leverage the variance in the data. This study shows that the “cluster-then-predict” approach of modeling across subpopulations outperforms the baseline approach of modeling on the entire population across multiple datasets and various modeling techniques.

# References

## References

1. Dimitris Bertsimas, Margret V. Bjarnadottir, Michael A. Kane, J. Christian Kryder, Rudra Pandey,Santosh Vempala, Grant Wang et al. 2008, Algorithmic Prediction of Health-Care Costs, Informs
2. Trupti M. Kodinariya & Prashant R. Makwana, 2013, Review on determining number of Cluster in K-Means Clustering, International Journal of Advance Research in Computer Science and Management Studies
3. MClust Package:

Chris Fraley, Adrian E. Raftery, T. Brendan Murphy, and Luca Scrucca (2012) mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation Technical Report No. 597, Department of Statistics, University of Washington

Chris Fraley and Adrian E. Raftery (2002) Model-based Clustering, Discriminant Analysis and Density Estimation Journal of the American Statistical Association 97:611-631

1. Friendly, M. (1994), A fourfold display for 2 by 2 by k tables. Technical Report 217, York University, Psychology Department, <http://www.math.yorku.ca/SCS/Papers/4fold/4fold.ps.gz>.
2. ggmap Package:

D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL: <http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf>

## Datasets

1. <https://knoema.com/atlas/topics/Energy/datasets>
2. <https://www.cms.gov/Research-Statistics-Data-and-Systems/Research-Statistics-Data-and-Systems.html>
3. <http://www.eia.gov/state/seds/seds-data-complete.cfm#CompleteDataFile>
4. <http://www.infochimps.com/>