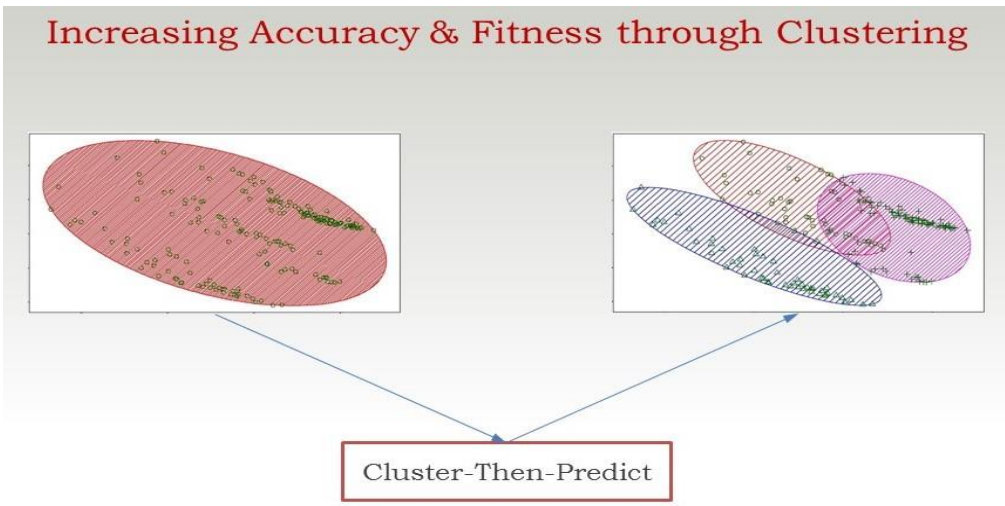
Increase Accuracy through Clustering

Sudha Subramanian
September 2016

Case Study: Increasing Accuracy and Fitness through Clustering

Comparative Study across 3 different Datasets



CHALLENGE - How best to make the model a good fit and generalize well on newly presented data, without overfitting to training data?

- Problem of overfitting - models the training data too well (negatively impacts performance of the model on new data)
- Problem of underfittng - neither model the training data nor generalize to new data

Comparative Study - Advantages of Clustering

- Modeling done on entire population; then clustering was applied and modeling done on subpopulation using same set of independent variables
- 1-3% increase in accuracy observed across all datasets, without changing the set of independent variables used to build the model

Three different datasets used for this exercise

1. Energy data for all states in the US from 2000 to 2013; includes State, Year, generation from various sources, prices for various sectors, sales, financial / regulatory incentives etc.; to predict if there will be increase in Solar Energy Generation
2. Stock Returns for company's stocks between 2000 and 2009 for first 11 months of the year; to predict if there will be increase or not in 12th month
3. Medicare reimbursement costs in 2008, with binary variables indicating if patient had diagnosis for the disorder in the year; predict costs for following year based on reimbursements in the previous year

Results:

Case Study - Summary

Entire Population (M_{EP}) vs. Subpopulation (M_{SP})

Dataset	Dependent Variable	Modeling Technique	M_{EP} Results Single-Model Approach	Results: M_{SP} compared to M_{EP}
Energy	Increase in Solar Energy Generation – Yes / No	GLM	Accuracy = 81.91%	2% increase in accuracy
Stock Returns	Increase in Stock Price – Yes / No	GLM	Accuracy = 67.71%	1.2% increase in accuracy
Medicare Reimbursement	Higher Costs – Yes / No	GLM	Accuracy = 69.83%	2.7% increase in accuracy
Medicare Reimbursement	Reimbursement cost for next year	LM	RMSE = 1.849	RMSE lower by 2%

Energy Dataset

Load Energy Dataset:

```
energy = read.csv("energy.csv")
```

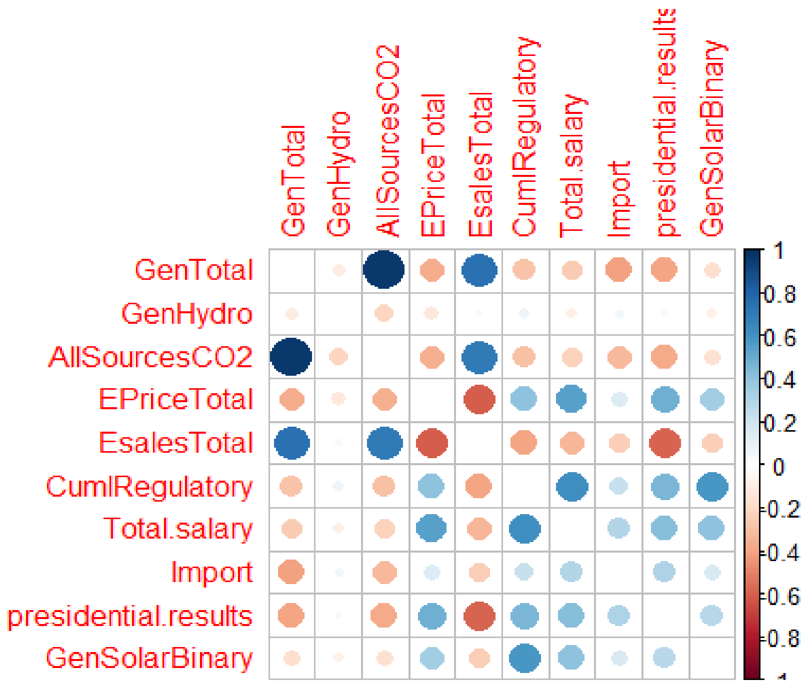
Target Variable: GenSolarBinary (whether there will be increase in Solar Power Generation or not)

Independent Variables:

- Information for all 50 states from 2000 to 2013
- Values normalized by population of the State for the year
- Generation Information
- Price across different sectors (Residential, Commercial, Industrial)
- Incentives (Financial & Regulatory)
- Emission Information
- Annual Wages, Presidential Results, Importer of Energy

Correlation Plot

Shows correlation of the Independent Variables to the Outcome/Target Variable
(Plot of selected set of Independent Variables)



Get Average Cost for Generation & Price by State

```
AvgPriceByState = filter(energy, !(STATE=="AK" | STATE=="HI")) %>%
  group_by(STATE) %>%
  summarise(AvgGenTotal=mean(GenTotal), AvgPriceTotal=mean(EPriceTotal)) %>%
  arrange(STATE)
```

Load US Map:

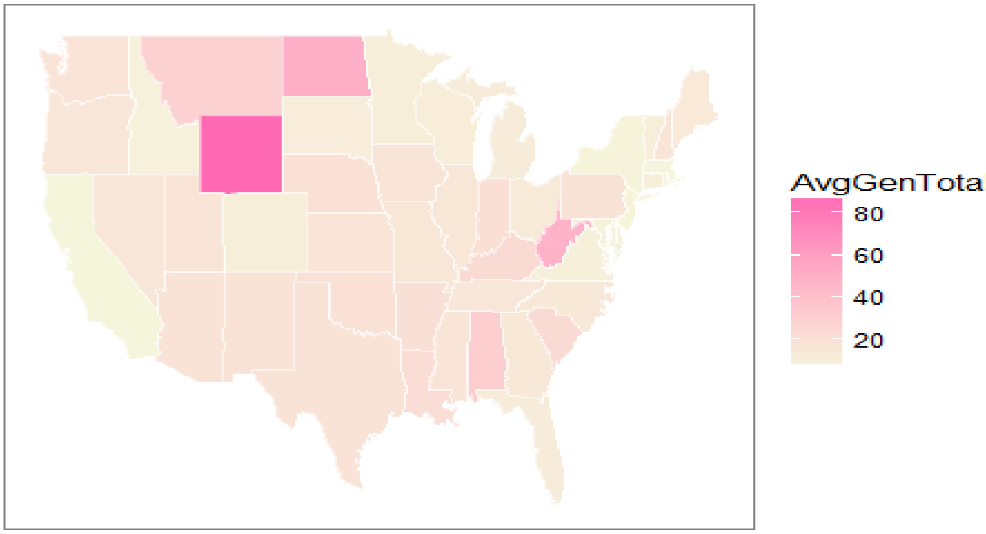
```
us.dat <- map_data("state")
```

Merge datasets for plotting average price in US Map

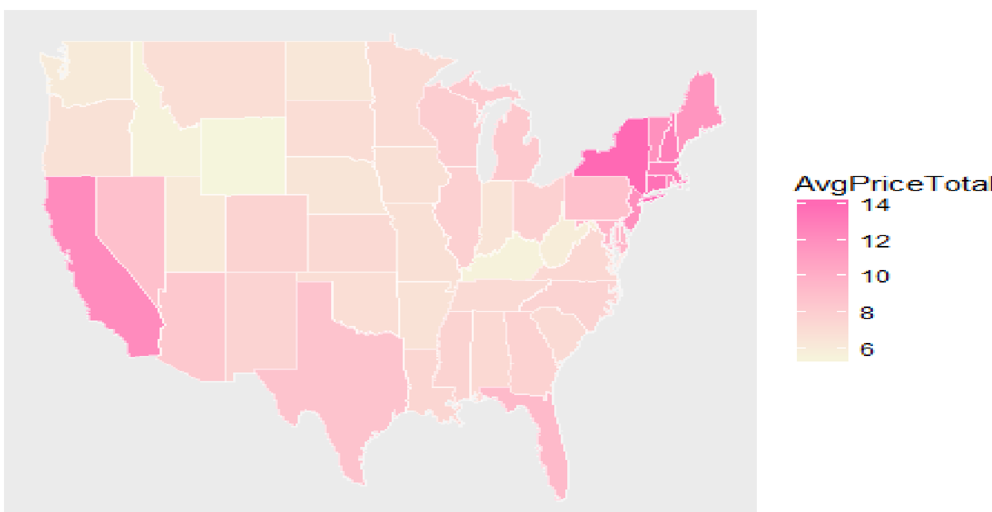
```
EnergyMap = merge(USMap, AvgPriceByState, by.x="State", by.y="STATE")
```

ENERGY DATASET: VISUALIZATIONS

PLOT: Average Cost for Generation by State

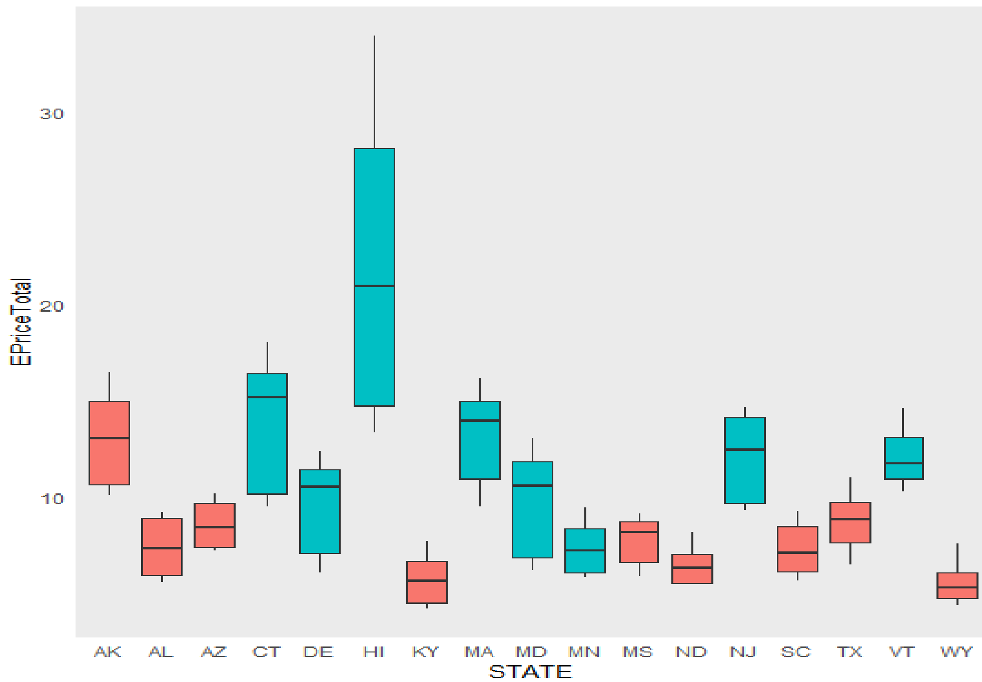


PLOT: Average Energy Price by State



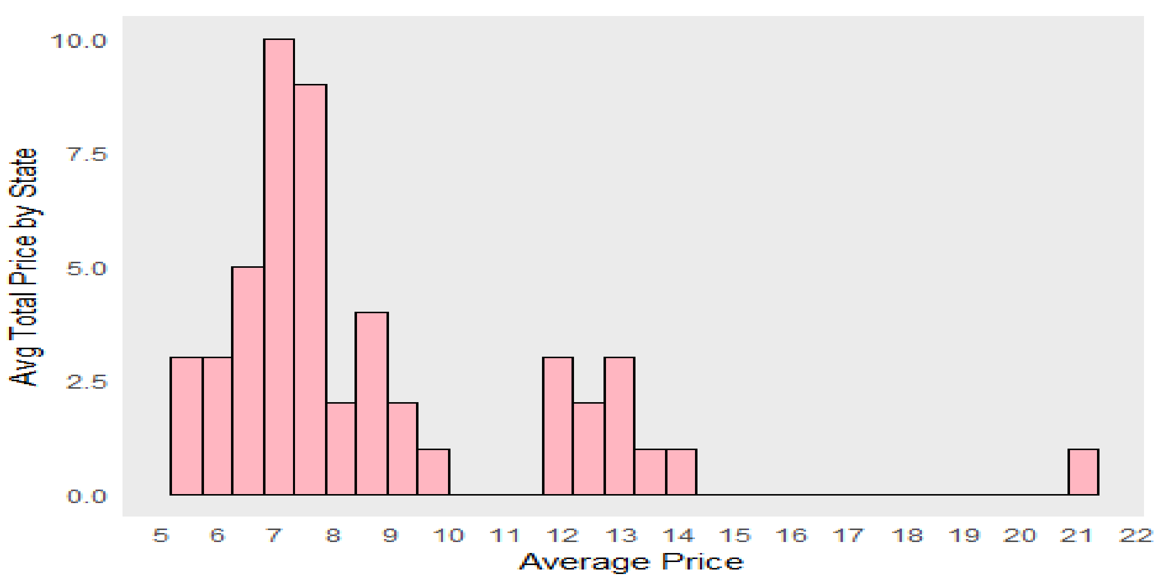
BOX PLOT: Price by State

- Shows the average energy costs across different States
- These are color coded by the party (Republican / Democrat); select States only shown
- Cost is highest in HI and lowest in WY



Histogram: Average Price

Shows the price ranges across all States ($5-$9; $11-$15; >$20)

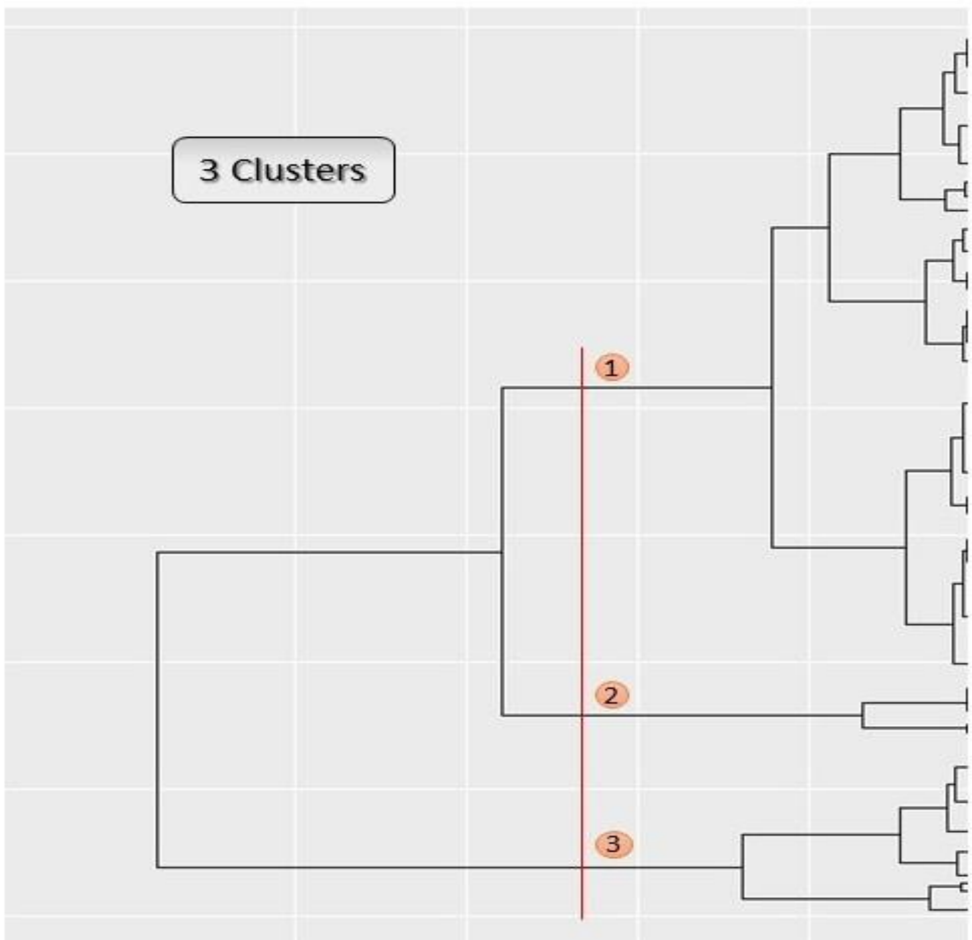


Through visualization and correlation plot, Independent Variables were identified for Clustering. These include:
* Total Price
* Incentives (Financial & Regulatory)
* Party (Presidential Results)
* Annual wages per capita
* State was importer or not

CLUSTERING Applied

- K-means clustering chosen for the dataset; k=3 based on visualization of the data and also based on Dendrogram
- Data Normalized, so that all variables are given same importance
- Target variable excluded from the set of variables based on which clustering is done

Dendrogram Plot: To identify the number of clusters

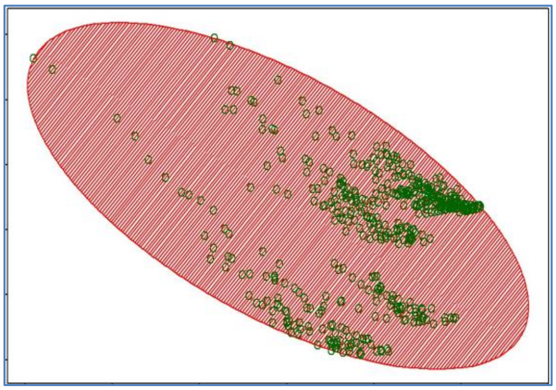


CLUSTER PLOTS

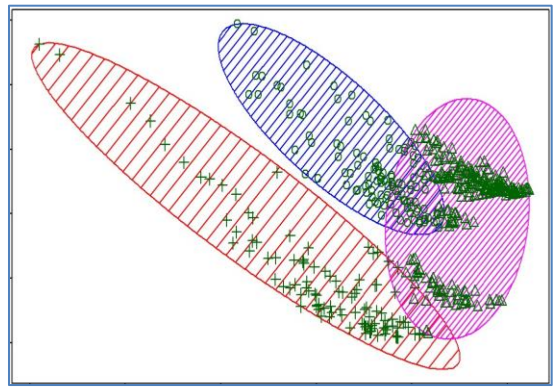
Cluster plot shows the distribution of the entire population. This creates a bivariate plot visualizing the clusters, using principal components.

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

Entire Population (TRAINING DATA):

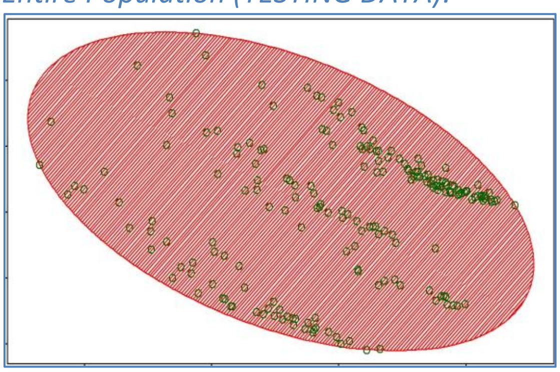


Clustering Applied (TRAINING DATA):

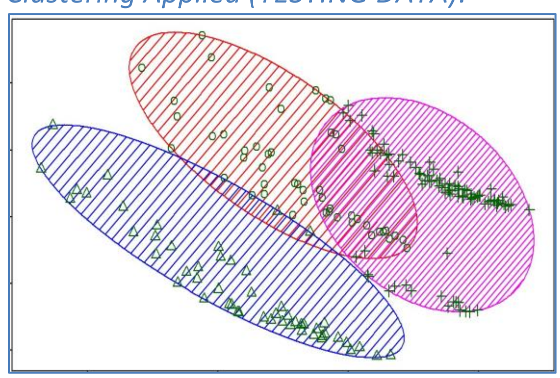


The following shows the distribution of the data points in the 'TESTING' dataset:

Entire Population (TESTING DATA):



Clustering Applied (TESTING DATA):



MODELING (Logistic Regression)

Entire Population

- Significant features identified
- GLM (Generalized Linear Model) applied

Subpopulation

- Models built on individual Clusters
- GLM Models built using adjusted set of features
- Overall Accuracy calculated based on accuracy for each cluster

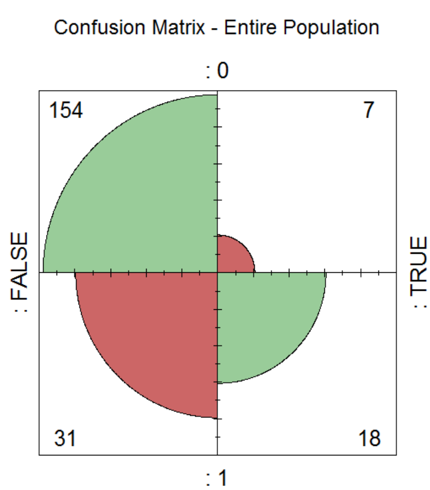
Confusion Matrix helps understand the number of outcomes as:

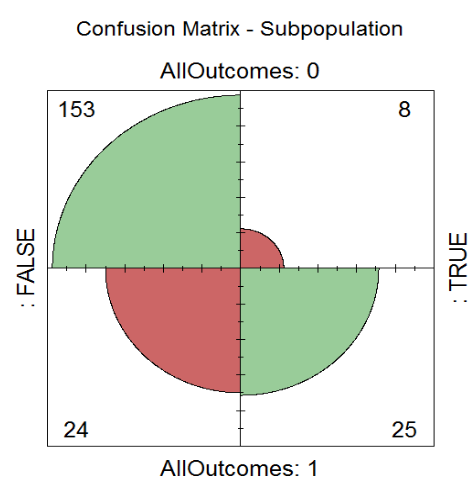
- Correct Outcomes (True Positives & True Negatives)
- Incorrect Outcomes (False Positives & False Negatives)

Confusion matrix is derived by applying the 'table' function on the actual outcomes vs. predicted outcomes. The Accuracy is calculated as:

$$\boxed{\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Outcomes}} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}}$$

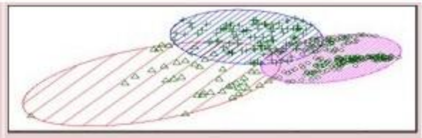
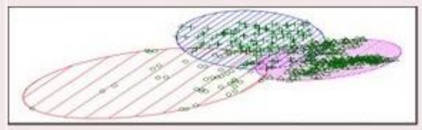
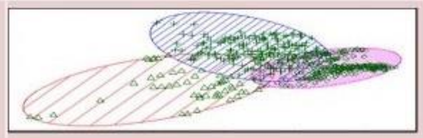
Confusion Matrix: Entire Population vs. Subpopulation





Validating the Case Study

Validation was done by applying the approach on datasets with varying percentage of training / test data.
Results of each of these runs are captured as shown below:

Sample %	Accuracy	↑ Accuracy	Cluster Plot
Training=50% Test=50%	M_{EP} – 86% M_{SP} – 87.7%	1.7%	
Training=60% Test=40%	M_{EP} – 84.3% M_{SP} – 85.7%	1.4%	
Training=70% Test=30%	M_{EP} – 81.9% M_{SP} – 84.3%	2.4%	

Summarizing the Case Study

- Clustering is an unsupervised way of identifying inherent patterns in the data and grouping them

- We would expect decision trees could easily incorporate the defining features of a cluster into the first levels of the tree

- 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, we are able to realize 1-3% increase in accuracy, that could make a big difference

- Within each cluster, any modeling technique can be applied. For example, one cluster can be modeled as GLM, while another one could be a RF model