LEARNING ABOUT VACCINE PROCUREMENT SCHEDULES

NICHOLAS J. MORRIS

Rochester Institute of Technology NY, USA

DATA DICTIONARY

Data Overview:

This data table represents how an affordable and profitable solution schedule for global vaccination changes in the face of budget uncertainty. The solution schedule is represented by Bundle_Name, Markets, MarketID, Selling_Qty, Selling_Price_Low, Selling_Price_High. Budget uncertainty is represented by the interaction of Price_Drop and Bundle_Impact. Information regarding the nature of the global vaccine market is given by Supply_Capacity, Production_Cost, Bundle_Demand, MARR, Birth_Cohort, Markets.

Data Table Description:

Variable	**Definition**	**Domain**
fileID	The experiment ID, The ABP run number.	Integers: [1, 62400]
Scenario	Indicates a distinct state of Budget Uncertainty.	Integers: [1,1248]
Replication	Indicates the repeated instance of a Scenario.	Integers: [1,50]
Bundle	Represents a distinct Vaccine.	Integers: [1,52]
Bundle_Name	The name of the Vaccine, The set of Antigens in the Vaccine.	Strings
BundleImpacted	Indicates if a Bundle was affected by Budget Uncertainty during a Replication.	Binary: 0 = No, 1 = Yes
Produce_Bundle	Indicates if a Bundle should be produced in a fileID instance.	Binary: 0 = No, 1 = Yes
Markets	Indicates the Global Market Structure.	Categorical: 2 Markets, 4 Markets, 8 Markets, 12 Markets

MarketID	Represents a distinct Market, Markets are ranked by decreasing Income levels.	Integers: [1,12]
MarketImpacted	Indicates if a Market was affected by Budget Uncertainty during a Scenario.	Binary: 0 = No, 1 = Yes
Selling_Qty	Indicates the total units of a Bundle sold to a Market in a fileID instance.	Integers
Supply_Capacity	Indicates the Global production capacity for a Bundle.	Integers
Production_Cost	Indicates the Global production cost to recuperate if a Bundle is produced.	Integers, USD
Bundle_Demand	Indicates the maximum number of times a Bundle can be used to satisfy dosage demand of Antigen(s) for a single child.	Integers
Birth_Cohort	The total number of children, The consumer demand.	Integers
Selling_Price_Low	Indicates the lowest price a Bundle should be sold for a Market in a fileID instance.	Reals, USD
Selling_Price_High	Indicates the highest price a Bundle should be sold for a Market in a fileID instance.	Reals, USD
Reservation_Price	Indicates the most a Market is willing to pay for a Bundle in a fileID instance.	Reals, USD
Surplus_Low	Indicates the savings for a Market by purchasing a low priced Bundle in a fileID instance: Selling_Qty * (Selling_Price_Low - Reservation_Price).	Reals, USD

Surplus_High	Indicates the savings for a Market by purchasing a high priced Bundle in a fileID instance: Selling_Qty * (Selling_Price_High - Reservation_Price).	Reals, USD
Revenue_Low	Indicates the revenue earned from a Market purchasing a low priced Bundle in a fileID instance: Selling_Qty * Selling_Price_Low.	Reals, USD
Revenue_High	Indicates the revenue earned from a Market purchasing a high priced Bundle in a fileID instance: Selling_Qty * Selling_Price_High.	Reals, USD
Price_Drop	Indicates how much a Market's Reservation Price for any Bundle could decrease by, during a Scenario.	Categorical: 1%-12%, 13%-26%, 27%-40%,
Bundle_Impact	Indicates how many of the Bundle's are at risk of a reduction in a Market's Reservation Price, during a Scenario.	Categorical: 1%-20%, 21%-40%, 41%-60%, 100%
MARR	The Minimum Annual Rate of Return that must be satisfied to warrant the usage of the Global production capacity.	Categorical: 5%, 10%, 15%, 20%

Finding a Solution Schedule:

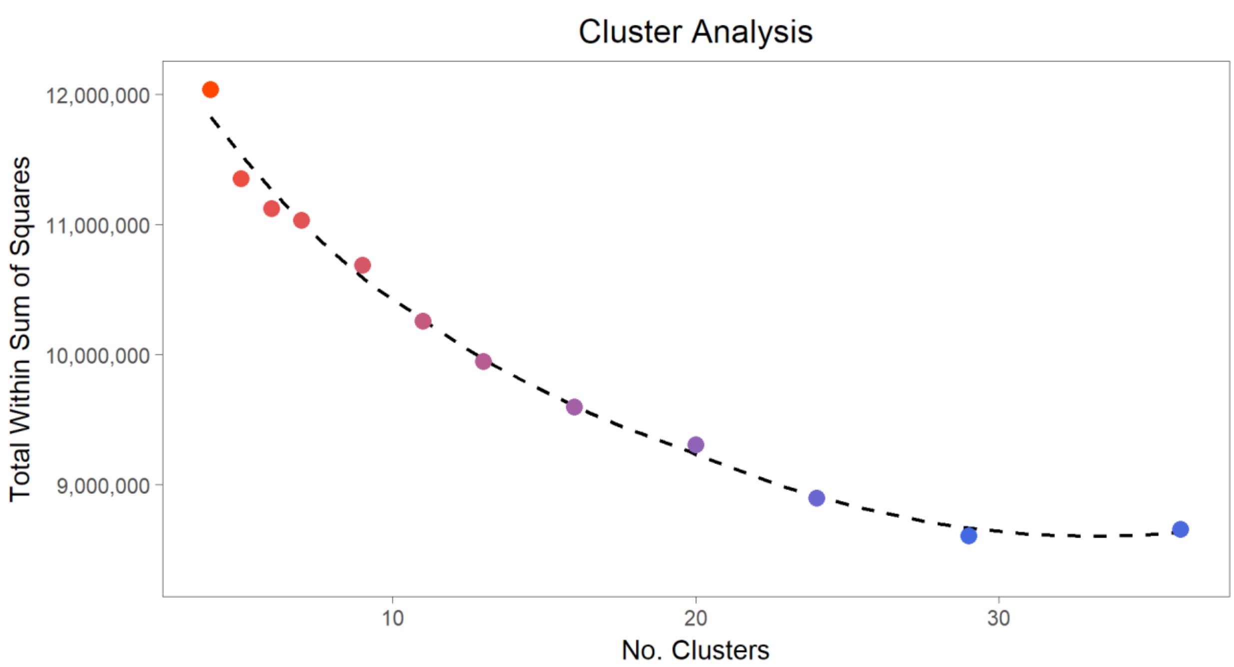
If you want to look for a single solution schedule, then you would just filter the data table based on fileID. For example, fileID = 1 represents the first solution schedule, which happens to be a 2 Market solution. Also, fileID = 62400 represents the last solution schedule, which happens to be a 12 Market solution. Consider looking at the following columns to interpret a solution schedule: Bundle_Name, Markets, MarketID, Selling_Qty, Selling_Price_Low, Selling_Price_High.

METHODOLOGY

The structure of the data consists of 136,800 rows and 156 columns. Each row represents the procurement schedule for a single market, and the columns represent schedule features. The schedule features include a low selling price, a high selling price, and a selling quantity for 52 vaccines.

The schedule features differ from one another in magnitude and units of measure, but each feature takes on real number values. So, each of the columns are normalized to a mean of 0 and a standard deviation of 1 to allow for fair comparison between these incommensurable numeric features. The kmeans algorithm is run on this data set multiple times with logarithmically spaced cluster sizes between 4 and 36 to determine a minimum number of clusters is that keeps total within sum of squares relatively low.

The figure below shows the results of the kmeans trials. This plot is used to find the point at which additional clusters give a smaller marginal rate of return. This point can be determined by looking for the "elbow" of the curve, the point at which the curve bends most. The chosen number of clusters to move forward with is 22 because before this point there is a decent improvement in total within sum of squares, and after this point there is minimal improvement.

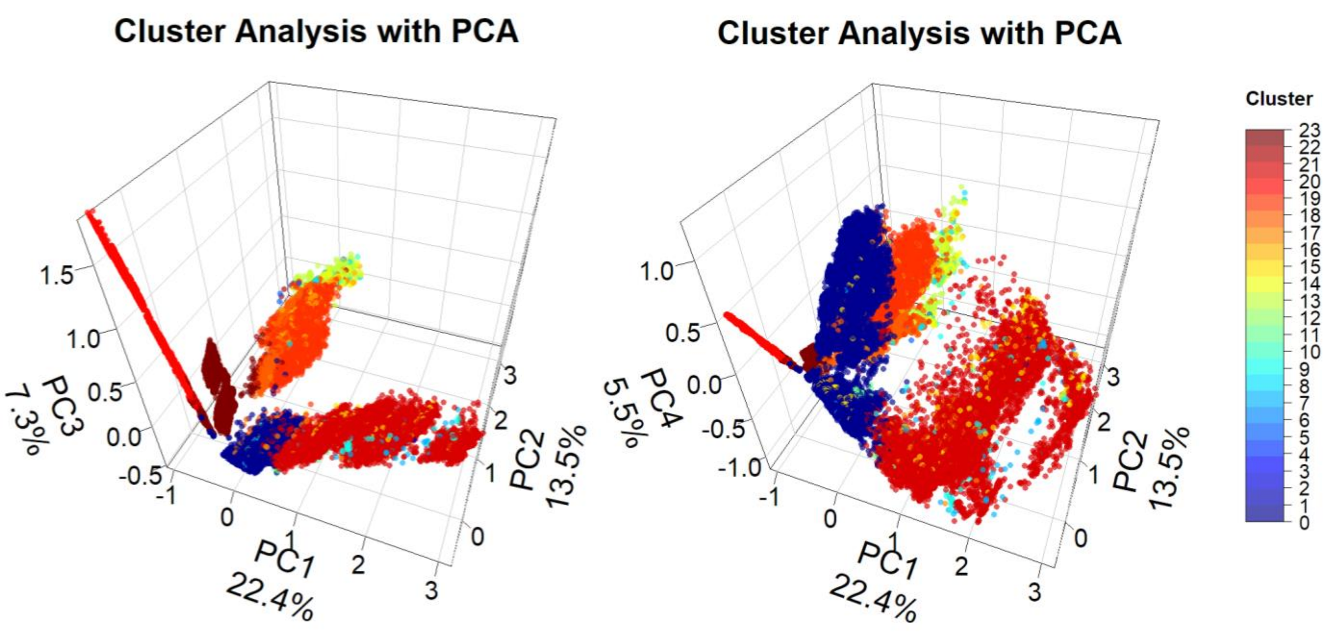


Principal component analysis is a technique that can be used to reduce the total number of features in a data set to a smaller set of orthogonal features that capture some of the variance of each original feature. This is useful to visualize the expected heterogeneous clusters from a kmeans cluster analysis within a small orthogonal feature space representative of the original data.

When principal component analysis was applied to the original data of 156 features, the first 5 components capture a total of 51.5% of the variance in the 156 features. The performance of each component is shown below. The eigenvalues represent the length of each component, where a larger eigenvalue means that the spread of the data is wider across the axis of that component.

Measure	PC1	PC2	PC3	PC4	PC5
Standard deviation	0.728	0.565	0.415	0.360	0.263
Proportion of Variance	0.224	0.135	0.073	0.055	0.029
Cumulative Proportion	0.224	0.359	0.431	0.486	0.515
Eigenvalue	0.530	0.319	0.172	0.130	0.069

Cubic plots below show the relationship between the clusters and the four principal components that capture the most variance of the original 156 schedule features. The coloring of the clusters demonstrates good coverage of the data and decent cluster segregation. The separation of the clusters shown by this plot is limited by the availability of colors to qualitatively distinguish 24 clusters from each other in a single feature space. The separation of clusters is also limited by the total explained variance of the principal components which is less than 50% in both plots. The plot on the left shows 3 distinct axes within the feature space of the first three principal components. These three axes may represent the three features of low pricing, high pricing, and selling quantity for 52 vaccines. The plot on the right resembles a box shape, where the faces of the box which may represent schedule difference across the 52 vaccines, or even the budget uncertainty scenarios that force solution schedules to vary for the same market.



The 6 clusters below represent 98% of the schedules that are meant to be clustered. So, 18 clusters were removed for capturing only 2% of the noise in the feature space. The table below represents the expected values of surplus, debt, profit, and value for all solution schedules in each cluster. Surplus is the total savings a market has by purchasing vaccines below their reservation price at low price points. Debt is the total loss a market has incurred by purchasing vaccines above their reservation price at low price points. Profit is the total amount of money earned after achieving a minimum return on each of vaccine that was purchased at high price points. So, negative values of profit indicate that the total revenue from sales were less than the total minimum return required for producing vaccines. Value is an equilibrium measure independent of the price points. Value can be computed using low or high prices as demonstrated below:

Value = Surplus(Low Prices) – Debt(Low Prices) + Profit(Low Prices)
= Surplus(High Prices) – Debt(High Prices) + Profit(High Prices)

Cluster	Size	Surplus	Debt	Profit	Value
23	37,777	$2,642,937,310	$1,066,091	$1,134,584,211	$2,600,901,563
20	10,692	$98,399,982	$1,177,294	-$115,354,753	-$52,339,094
18	4,809	$2,957,783,684	$6,023,049	$1,861,859,554	$3,093,458,093
21	14,867	$3,801,214,368	$1,011,955	$1,334,780,711	$3,720,909,707
19	10,473	$2,698,533,955	$10,506,159	$2,240,198,900	$3,055,727,473
0	55,699	$743,405,255	$303,354	-$30,743,056	$591,276,395

The table below represents the expected values of risk, GNIpc, infant mortality, and annual births for all markets in each cluster. This table characterizes which kind of countries are assigned to each cluster. For example, Cluster 20 represents micro-states like Monaco that have high levels of income with low population levels. This would explain why Cluster 20 schedules aren't profitable for the manufacturers because these micro-states don't have the demand volume to warrant a minimum return on the manufacturer's investment in of vaccines. Cluster 20 doesn't represent all solution schedules for all markets, so the manufactures can still make a return on their investment by leveraging the remaining schedules from other profitable clusters.

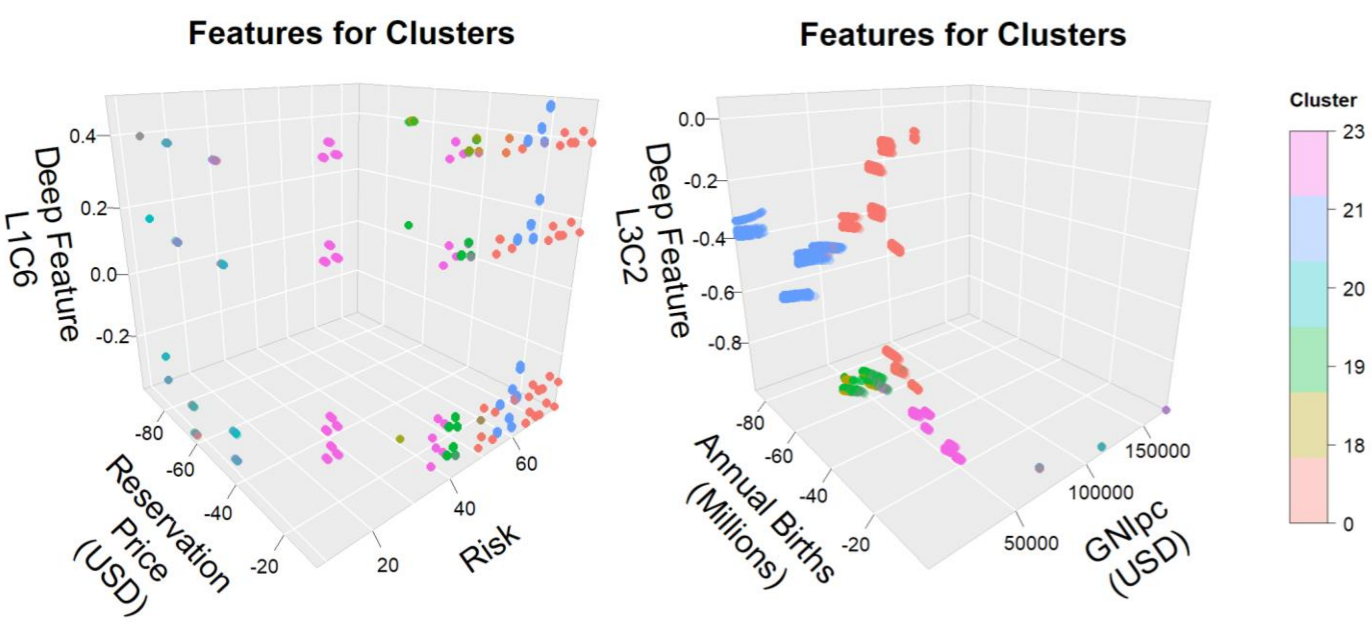
Cluster	Size	Risk	GNIpc	Birth Mortality	Annual Births
23	37,777	32.3	$26,898	1.08%	8,844,205
20	10,692	12.2	$131,916	0.32%	194,004
18	4,809	44.6	$6,266	1.93%	24,292,331
21	14,867	59.1	$1,857	4.11%	49,220,244
19	10,473	43.9	$5,819	1.89%	26,162,525
0	55,699	59.5	$9,846	4.19%	10,166,749

The table below shows the coverage rate for each antigen that is required by national immunization schedules. Cluster 21 and Cluster 0 have difficulty securing the total required doses of HepB. These clusters are show by the previous table to be high-risk low-income markets with larger demand levels, which explains why it is difficult to fully immunize the children in these markets.

Cluster	Size	DTP	HepB	Hib	IPV	MMR	V
23	37,777	100%	100%	100%	100%	100%	100%
20	10,692	100%	100%	100%	100%	100%	100%
18	4,809	100%	100%	100%	100%	100%	100%
21	14,867	100%	76.68%	99.97%	100%	99.92%	99.98%
19	10,473	100%	100%	100%	100%	100%	100%
0	55,699	99.99%	59.37%	99.15%	100%	95.87%	99.89%

Describe some of the top features using the 3d plots below.

Note: the order of variable importance is subject to change when reproducing results (stochastic gradient decent of h2o autoencoder) but the magnitude of importance for each variable doesn't change significantly. The autoencoder's deep features are subject to change, but the final autoencoder performance doesn't change significantly.



Describe the chosen features using the table below.

Rank	Feature	Relative Importance	Scaled Importance	Percentage
1	Birth_Mortality(2.68,4.43]	1,902,812	100.0%	4.11%
2	GNIpc(3.64e+03,7.24e+03]	1,310,517	68.9%	2.83%
3	Birth_Cohort(155,6.2e+05]	1,300,845	68.4%	2.81%
4	Country_Risk(8.06,14.3]	1,224,960	64.4%	2.64%
5	Reservation_Price(10.5,12.7]	1,095,177	57.6%	2.36%
---	---	---	---	---
46	Birth_Cohort(1.53e+07,2.39e+07]	369,736	19.4%	0.80%
47	Birth_Mortality(0.659,1.5]	368,990	19.4%	0.80%
48	Reservation_Price(8.82,9.04]	361,790	19.0%	0.78%
49	Country_Risk(50.3,54.4]	352,839	18.5%	0.76%
50	Birth_Cohort(2.39e+07,4.09e+07]	348,536	18.3%	0.75%
---	---	---	---	---
146	DF.L3.C5(0.887,0.891]	28,390	1.5%	0.06%
147	DF.L2.C3(-0.585,-0.575]	27,512	1.4%	0.06%
148	DF.L3.C5(0.891,0.931]	26,793	1.4%	0.06%
149	DF.L1.C6(0.155,0.173]	24,152	1.3%	0.05%
150	DF.L2.C3(-0.482,-0.475]	20,214	1.1%	0.04%

Describe how data is split up into a training, validation, and testing data sets (and what each data set means)

Describe how the training, scoring, and early stopping works in h2o (use h2o documentation)

Describe how a random grid search is used to do hyperparameter tuning of supervised learning models.

Show the grid search parameters for GLM, RF, GB, and NNET

Hyperparameter	Values
lambda	(1, 0.5, 0.1, 0.01, 0.001, 0.0001, 0.00001, 0)
alpha	(0, 0.5, 1)
normalize	(TRUE, FALSE)
intercept	(TRUE, FALSE)

Hyperparameter	Values
ntrees	(50, 250, 500)
min_rows	(1, 5, 11)
max_depth	(10, 20, 40)
min_split_improvement	(0, 1e-5)

Hyperparameter	Values
epochs	(10, 100, 1000),
hidden	(S, M, L, {S, S}, {M, M}, {L, L}, {S, S, S}, {M, M, M}, {L, L, L})
activation	("RectifierWithDropout", "TanhWithDropout")
input_dropout_ratio	(0, 0.15)
l1	(0, 1e-5)
l2	(0, 1e-5)
rho	(0.9, 0.95, 0.99, 0.999)
epsilon	(1e-10, 1e-8)

Hyperparameter	Values
ntrees	(50, 250, 500)
learn_rate	(0.025, 0.05, 0.1)
max_depth	(5, 10, 20)
min_rows	(1, 5, 11)
sample_rate	(0.7, 1)
col_sample_rate	(0.7, 1)
min_split_improvement	(0, 1e-5)

Describe how the super learner works.

Describe cluster predictive performance (of unseen data) using the table below.

Model	Log Loss	Kappa	Macro Accuracy	Micro Accuracy
Regression	0.1990	0.9041	0.9296	0.9765
Random Forest	0.1776	0.9043	0.9298	0.9766
Gradient Boosting	0.1878	0.9043	0.9298	0.9766
Neural Network	0.2605	0.8895	0.9177	0.9726
Super Learner	0.1791	0.9043	0.9298	0.9766

REFERENCES

[1] Awasthi, Pranjal. Supervised Clustering. papers.nips.cc/paper/4115-supervised-clustering.pdf.

[2] Finley, Thomas. “Supervised Clustering with Support Vector Machines.” Proceedings of the 22nd International Conference on Machine Learning, engr.case.edu/ray_soumya/mlrg/supervised_clustering_finley_joachims_icml05.pdf.

[3] Pappuswamy, Umarani. A Supervised Clustering Method for Text Classification. Learning Research and Development Center, University of Pittsburgh, www.public.asu.edu/~kvanlehn/Stringent/PDF/05CICL_UP_DB_PWJ_KVL.pdf.

[4] Bair, Eric. “Semi-Supervised Clustering Methods.” NIH Public Access, www.ncbi.nlm.nih.gov/pmc/articles/PMC3979639/pdf/nihms502381.pdf.

High performance machine learning environment:

http://docs.h2o.ai/

https://cran.r-project.org/web/packages/h2o/h2o.pdf

Super Learner - Dissertation & Library:

http://digitalassets.lib.berkeley.edu/etd/ucb/text/Polley_berkeley_0028E_10767.pdf

https://cran.r-project.org/web/packages/SuperLearner/SuperLearner.pdf

Macro-Economic Data:

https://data.worldbank.org/