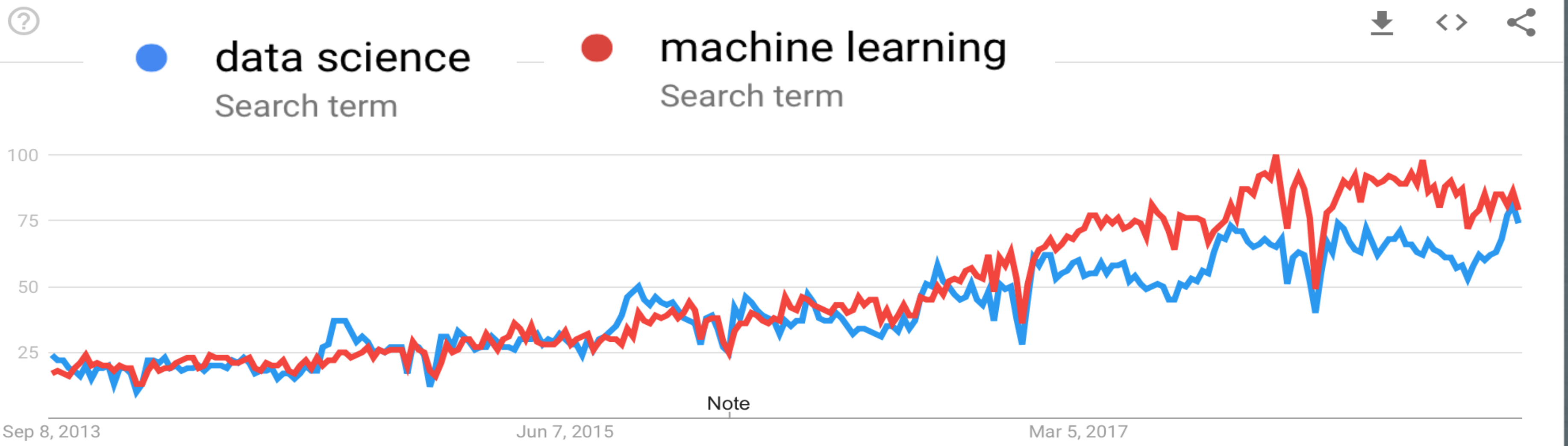
Doing the Data Science Dance

Dean Abbott
Abbott Analytics, SmarterHQ
KNIME Fall Summit 2018

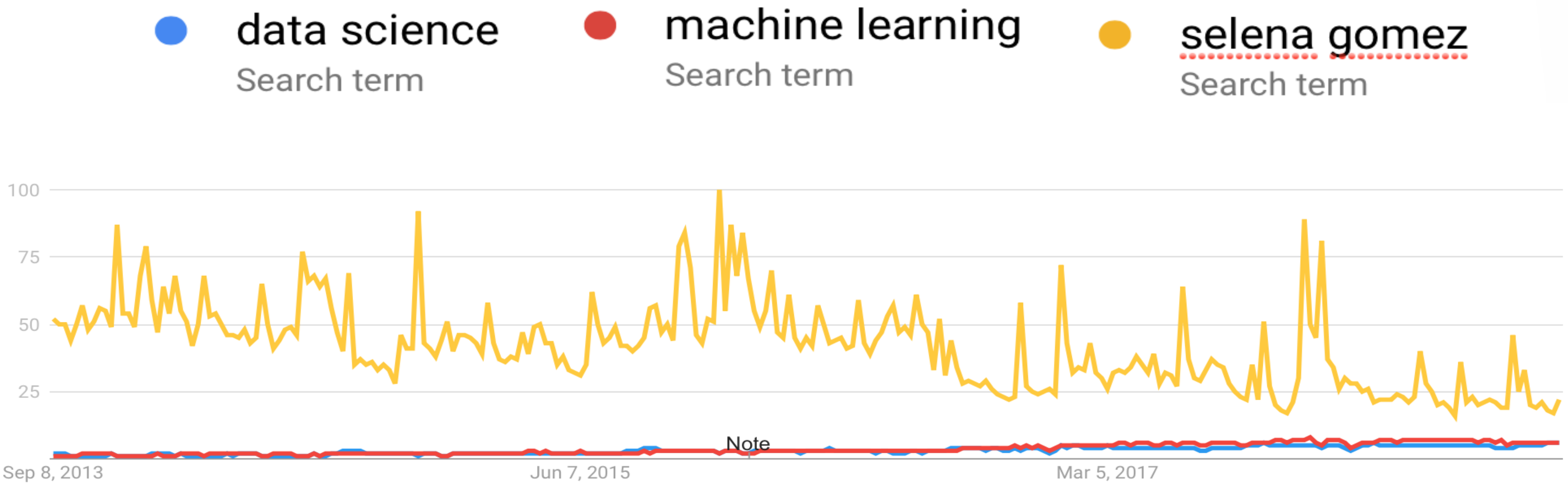
Email: dean@abbottanalytics.com
Twitter: @deanabb

Data Science vs. Other Labels

Google Trends

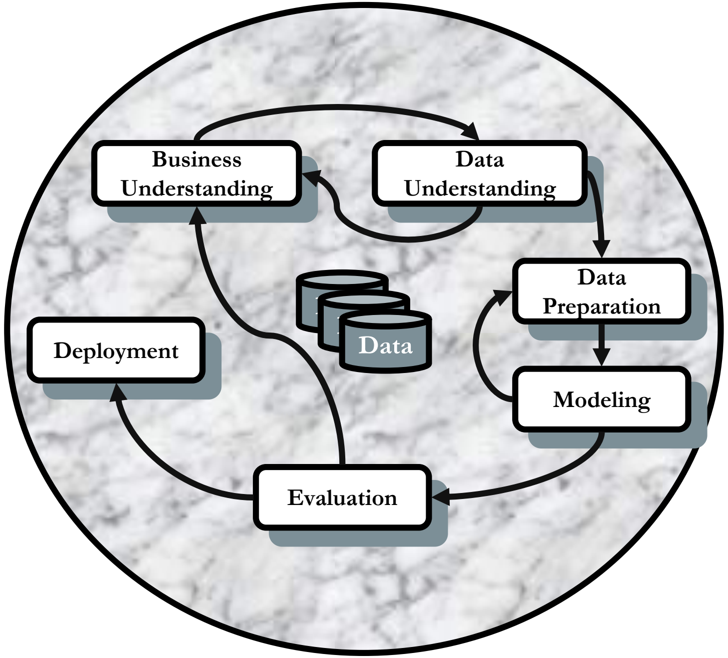


Google Trends

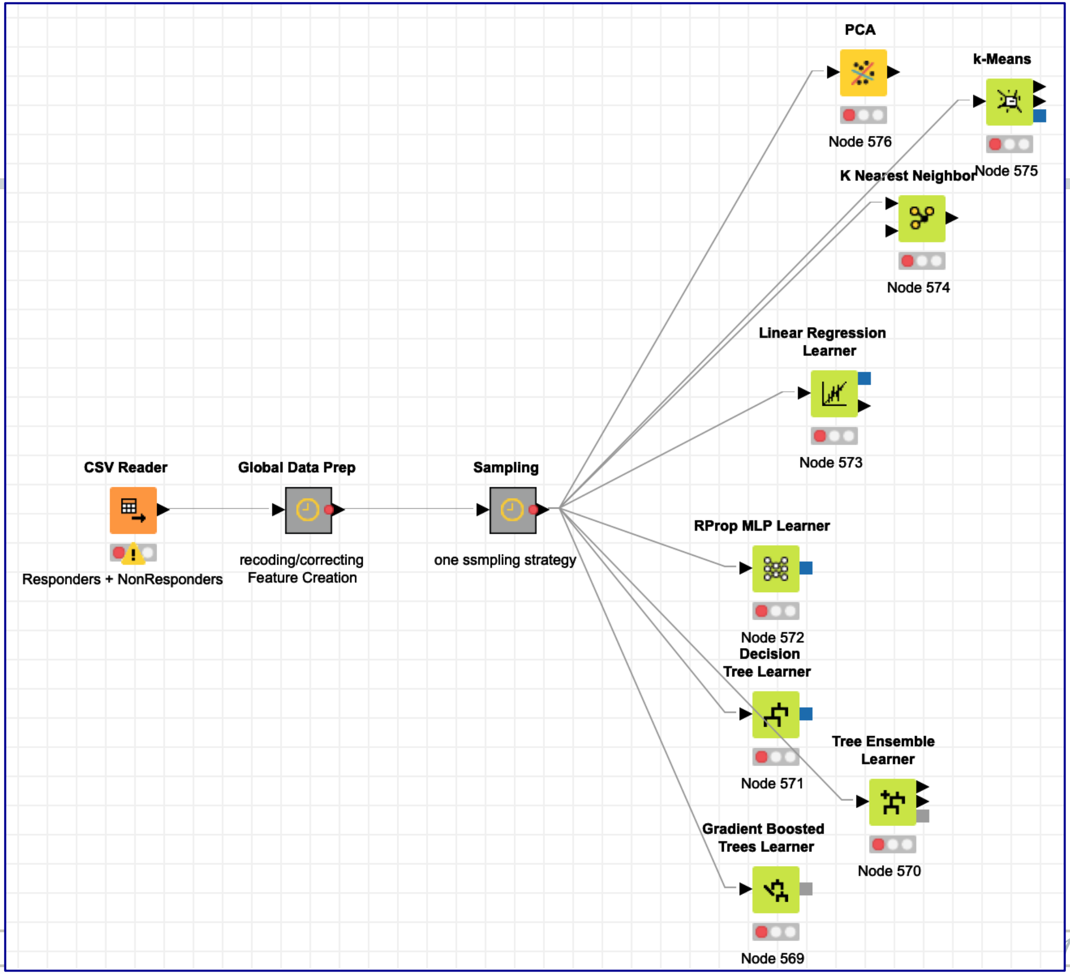


What do Predictive Modelers do? The CRISP-DM Process Model

- CRoss-Industry Standard Process Model for Data Mining
- Describes Components of Complete Data Mining Cycle from the Project Manager's Perspective
- Shows Iterative Nature of Data Mining



What we Want to Do!



How The Citizen Data Scientist Will Democratize Big Data

Published on April 6, 2016

How The Citizen Data Scientist Will Democratize Big Data
Published on April 6, 2016

Retailer Sears, for example, recently empowered 400 staff from its business intelligence (BI) operations to carry out advanced, Big Data driven customer segmentation – work which would previously have been carried out by specialist Big Data analysts, probably with PhDs.

Is it a Recipe?

What's wrong with my cake? 10 common baking problems fixed!

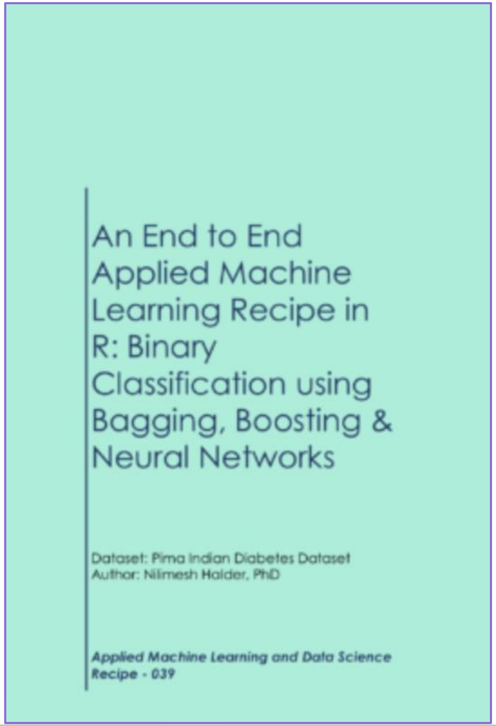
Jessica Dady

March 30, 2018 6:00 am

10 common baking problems fixed!

1. My cake didn't rise
2. My cake is greasy
3. My cake is stuck in the tin
4. My cake is burnt
5. My cake is raw
6. My cake mix has split
7. My cake is too dry
8. My cake has sunk in the middle
9. My cake has risen unevenly
10. My cake has shrunk

Is it a Recipe?



Can we apply a recipe to machine learning and data science modeling processes?

Good Set of Data Prep Steps!

Open for Innovation KNIME

Seven Techniques for Dimensionality Reduction

Missing Values, Low Variance Filter, High Correlation Filter, PCA, Random Forests, Backward Feature Elimination, and Forward Feature Construction

Rosaria Silipo Rosaria.Silipo@knime.com
Iris Adae Iris.Adae@uni-konstanz.de
Aaron Hart Aaron.Hart@knime.com
Michael Berthold Michael.Berthold@uni-konstanz.de

1. High number of missing values
2. Low variance
3. High correlation with other data columns
4. Principal Component Analysis (PCA)
5. First cuts in random forest trees
6. Backward feature elimination
7. Forward feature construction

https://www.knime.org/files/knime_seventechniquesdatadimreduction.pdf

Data Preparation Dependencies

Neural Newtorks
Linear Regression*
Logistic Regression
K Nearest Neighbor*
PCA*
Nearest Mean*
Kohonen Self-Organizing Maps*
Support Vector Machines
Radial Basis Function Networks
Discriminant Analysis

- Fill missing values
- Explode categorical variables
- *Outliers and scale very influential
- Sometimes automatic in software; beware of how!

Decision Trees
Naïve Bayes
Rule Induction
Association Rules

- Categoricals are fine
- Numeric data must be binned (except some decision trees)
- Outliers don't matter
- Missing values a category

Why Are Outliers a Problem? Squares…

Linear Regression: Mean Squared Error

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^{\boxed{2}}$$

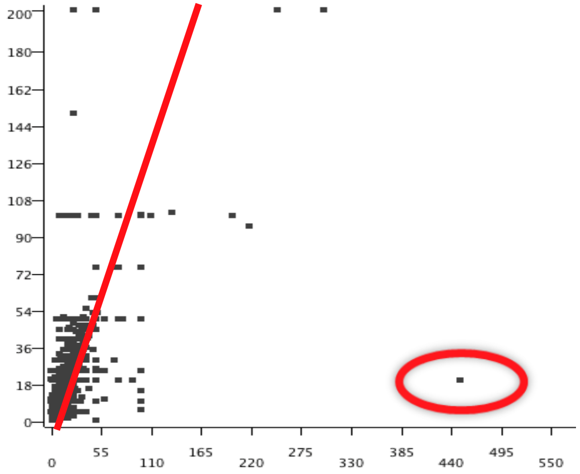
https://en.wikipedia.org/wiki/Mean_squared_error

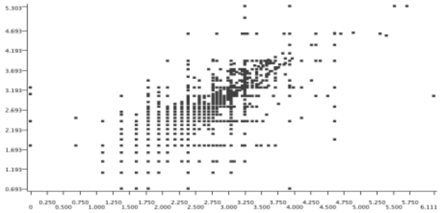
K-Means Clustering

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^{\boxed{2}}}$$

https://en.wikipedia.org/wiki/Euclidean_distance

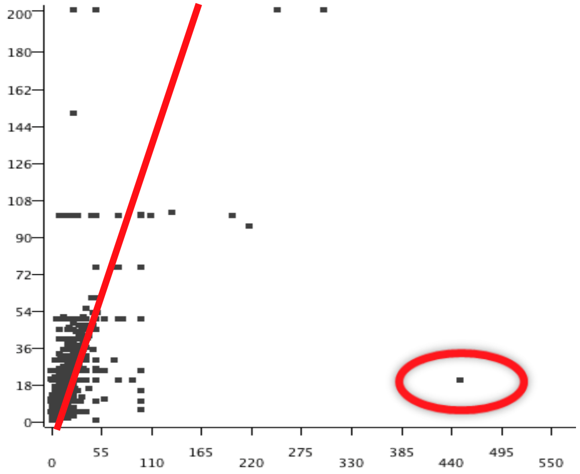
Effect of Outliers on Correlations (and Regression)

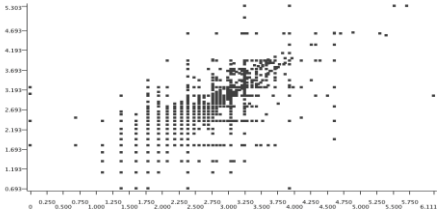




- 4,843 records

Effect of Outliers on Correlations (and Regression)

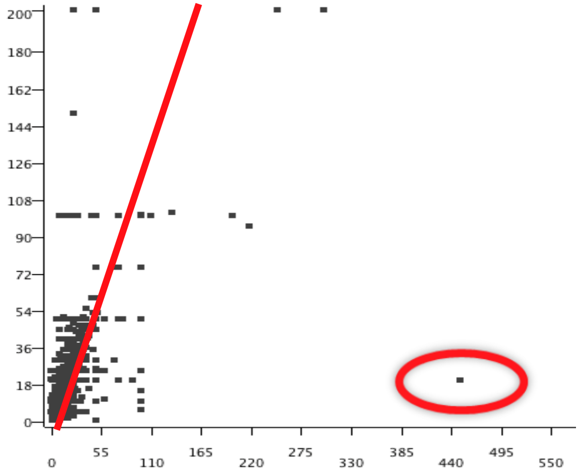


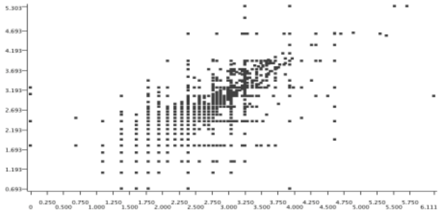


- 4,843 records

correlations	LASTGIFT	TARGET_D	LASTGIFT_log10	TARGET_D_log10
LASTGIFT	1	0.645	0.747	0.552
TARGET_D	0.645	1	0.641	0.847
LASTGIFT_log10	0.747	0.641	1	0.750
TARGET_D_log10	0.552	0.847	0.750	1

Effect of Outliers on Correlations (and Regression)





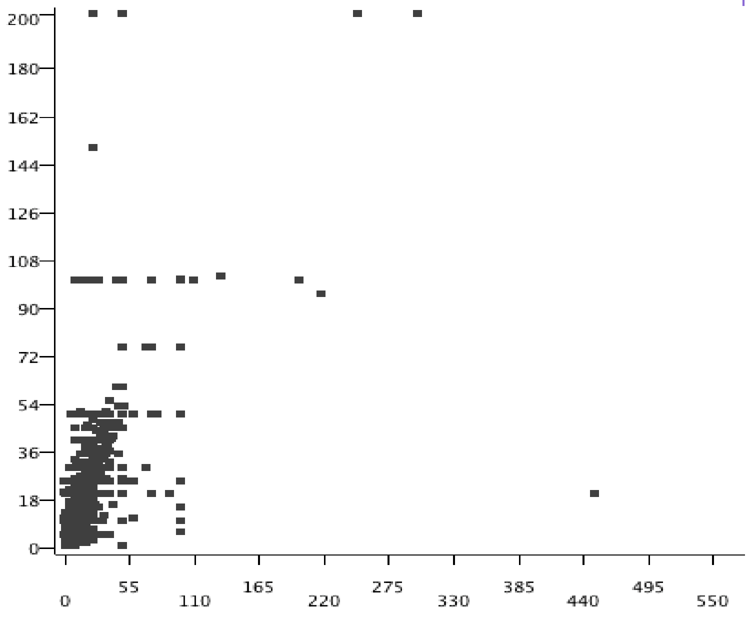
- 4,843 records

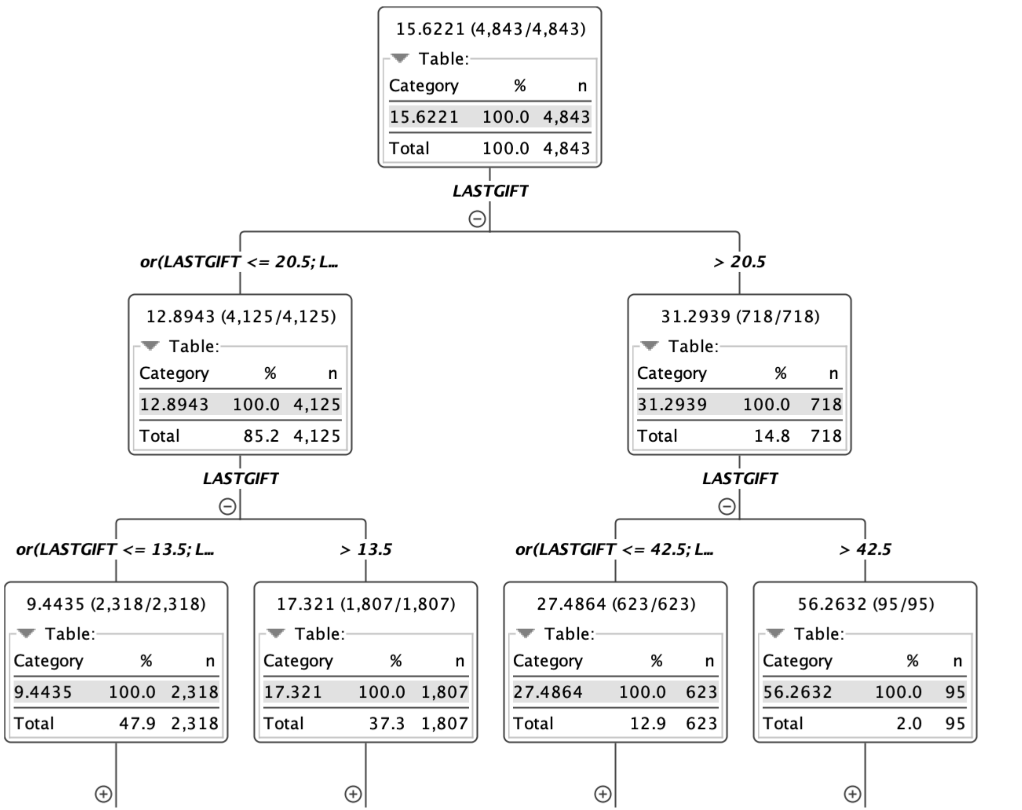
correlations	LASTGIFT	TARGET_D	LASTGIFT_log10	TARGET_D_log10
LASTGIFT	1	0.645	0.747	0.552
TARGET_D	0.645	1	0.641	0.847
LASTGIFT_log10	0.747	0.641	1	0.750
TARGET_D_log10	0.552	0.847	0.750	1

remove one outlier	LASTGIFT	TARGET_D	LASTGIFT_log10	TARGET_D_log10
LASTGIFT	1	0.725	0.799	0.617
TARGET_D	0.725	1	0.643	0.847
LASTGIFT_log10	0.799	0.643	1	0.752
TARGET_D_log10	0.617	0.847	0.752	1

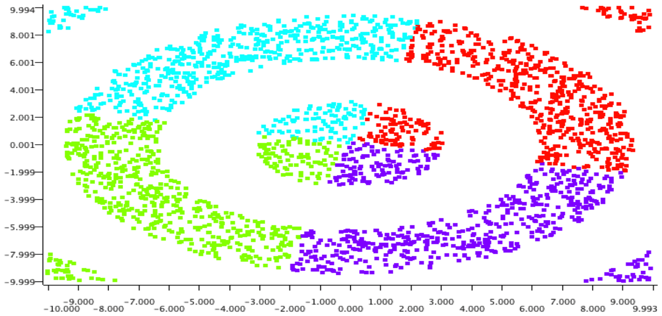
Corresponds to R^2 increase from 0.42 to 0.53

Decision Trees Can Handle it

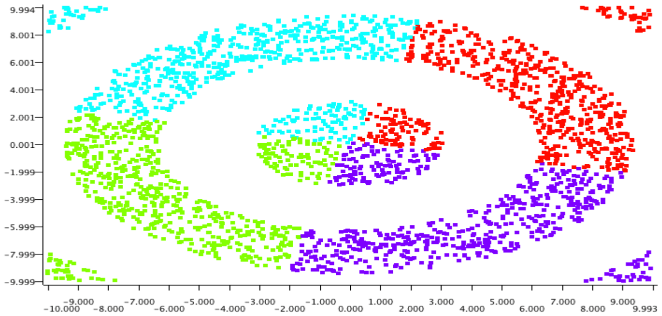


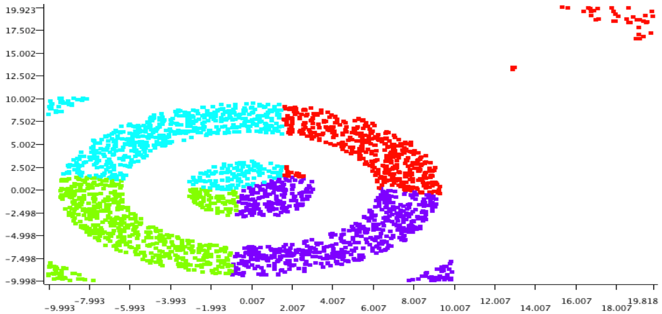


Effect of Distance on Clusters

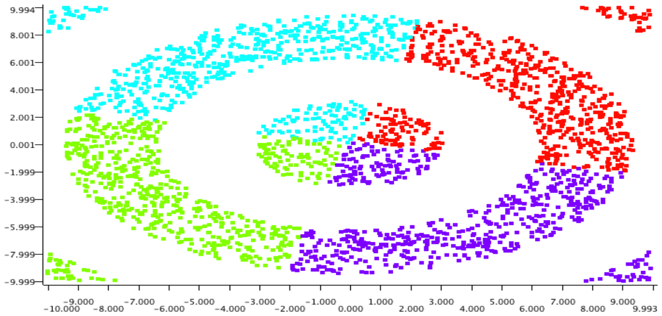


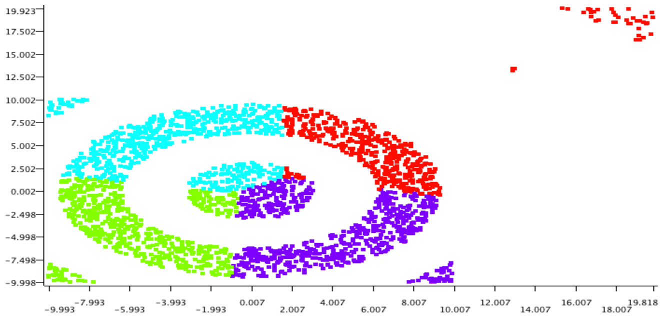
Effect of Distance on Clusters

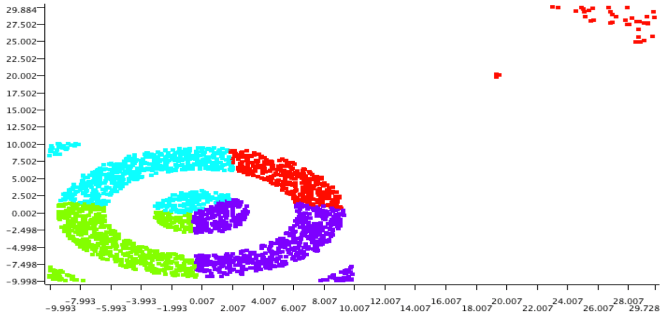




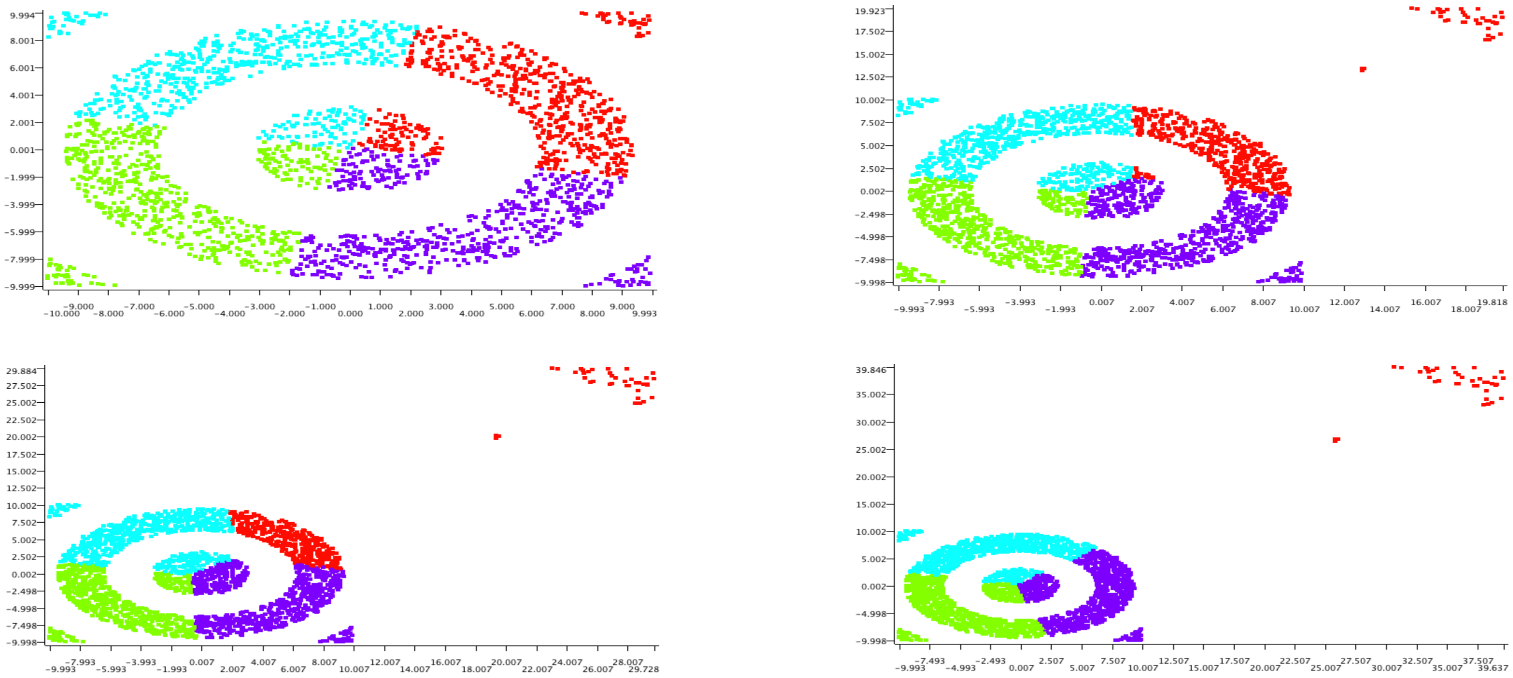
Effect of Distance on Clusters

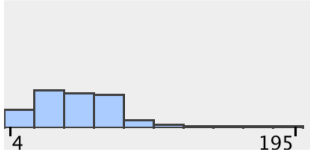
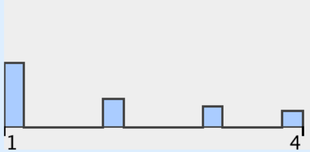




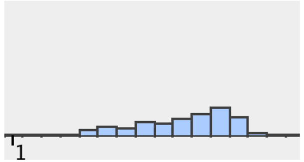
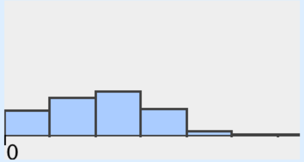
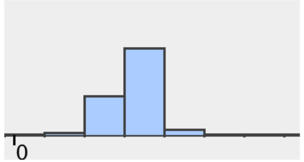


Effect of Distance on Clusters



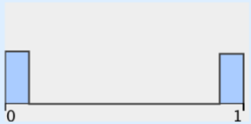
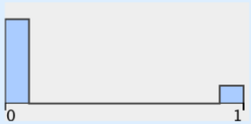
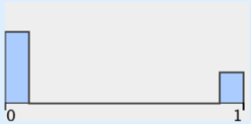
Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. −∞	Histogram
NUMPROM	4	46.9733	?	195	22.9704	0.4376	0.0185	0	0	0	
NGIFTALL	1	9.602	?	237	8.5543	2.0787	11.4809	0	0	0	
LASTGIFT	0.0	17.3124	?	1,000	13.9566	16.2866	728.4362	0	0	0	
FISTDATE	0.0	9,135.6516	?	9,603	320.394	-0.7834	12.9627	0	0	0	
RFA_2F	1	1.9101	?	4	1.0727	0.7855	-0.7734	0	0	0	

Log transform the heavily skewed fields

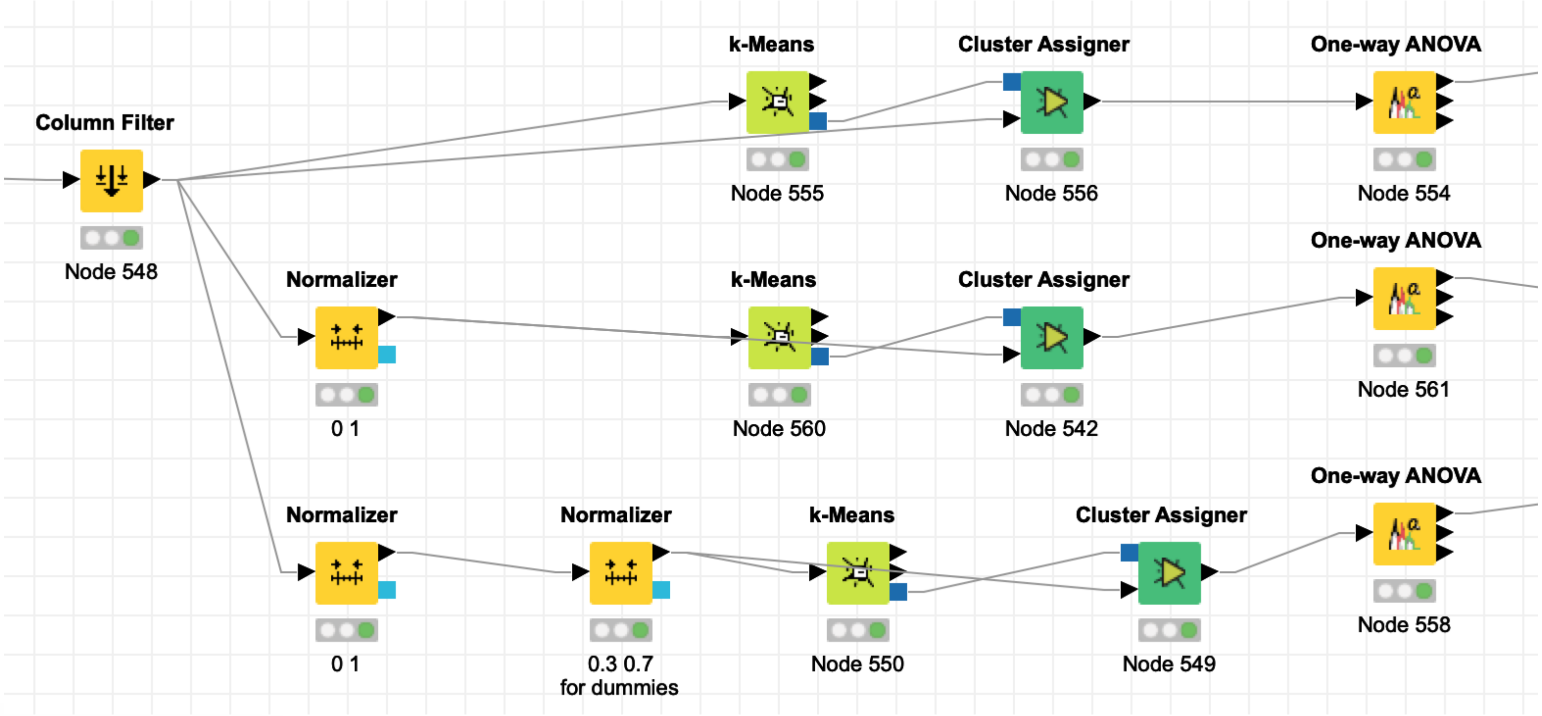
Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. −∞	Histogram
NUMPROM_log10	0.699	1.6225	?	2.2923	0.2389	-0.5638	-0.5334	0	0	0	
NGIFTALL_log10	0.301	0.8956	?	2.3766	0.3447	-0.0742	-0.7723	0	0	0	
LASTGIFT_log10	0.0	1.199	?	3.0004	0.2354	-0.4802	3.5736	0	0	0	

Dummy Vars

Note: stdev are Typically 0.5

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
D_RFA_2A	0.0	0.0777	?	1	0.2677	3.1555	7.9573	0	0	0	
F_RFA_2A	0.0	0.4922	?	1	0.4999	0.0311	-1.9991	0	0	0	
G_RFA_2A	0.0	0.2033	?	1	0.4025	1.4745	0.1742	0	0	0	
DOMAIN3	0.0	0.1756	?	1	0.3805	1.7053	0.908	0	0	0	
DOMAIN2	0.0	0.4825	?	1	0.4997	0.0699	-1.9952	0	0	0	
DOMAIN1	0.0	0.2987	?	1	0.4577	0.8797	-1.2261	0	0	0	
DOMAIN4	0.0	0.0189	?	1	0.1362	7.0647	47.911	0	0	0	

Try K-Means with Different Normalization Approaches



K Means Clustering: Magnitude and Dummy Bias

Measurements are F Statistic

Variable	Type	Natural	Scalled [0,1]	Scaled [0, 1]; dummies [0.3, 0.7]
FISTDATE	continuous	415,191.15	873.90	862.42
LASTGIFT_log10	continuous	502.33	17,134.27	8,936.27
NGIFTALL_log10	continuous	38,724.24	3,148.09	3,718.02
NUMPROM_log10	continuous	77,773.14	845.03	1,331.08
D_RFA_2A	dummy	355.94	Infinity	6,341.91
DOMAIN1	dummy	51.50	239,491.96	20,391.53
DOMAIN2	dummy	16.15	54,942.39	13,003.09
DOMAIN3	dummy	12.47	155,098.25	4,580.00
DOMAIN4	dummy	6.56	270.42	148.01
F_RFA_2A	dummy	801.02	33,172.69	78,485.65
G_RFA_2A	dummy	81.61	93,041.59	18,953.72
RFA_2F	ordinal	453.53	6,909.78	62,559.28
	Avg Continuous	133,047.71	5,500.32	3,711.95
	Avg Dummy	189.32	96,002.88	20,271.99
	Avg Ordinal	453.53	6,909.78	62,559.28

PCA: Natural Units

Natural Units	1. eigenvector	2. eigenvector	3. eigenvector	4. eigenvector	5. eigenvector	6. eigenvector	7. eigenvector	8. eigenvector	9. eigenvector	10. eigenvector	11. eigenvector
eigenvalue	1.254	0.380	0.308	0.211	0.153	0.082	0.046	0.027	0.016	0.005	0.005
RFA_2F	-0.952	-0.031	-0.124	0.062	0.206	0.175	0.022	-0.001	-0.015	-0.010	-0.028
D_RFA_2A	-0.123	0.003	0.009	-0.032	-0.100	-0.446	-0.861	0.005	-0.181	-0.011	-0.020
F_RFA_2A	0.194	0.088	-0.750	0.089	0.093	0.467	-0.336	-0.002	0.209	0.018	0.043
G_RFA_2A	0.059	-0.097	0.605	0.016	0.326	0.439	-0.361	-0.001	0.432	0.028	0.062
DOMAIN3	-0.009	-0.138	-0.079	-0.781	0.101	0.028	-0.001	-0.381	-0.003	0.429	-0.159
DOMAIN2	-0.022	0.768	0.117	0.282	0.005	-0.006	-0.004	-0.341	-0.005	0.421	-0.153
DOMAIN1	0.032	-0.610	-0.036	0.534	-0.095	-0.039	0.001	-0.354	-0.005	0.425	-0.151
DOMAIN4	0.001	-0.008	-0.005	-0.015	0.008	-0.001	0.003	0.783	-0.003	0.583	-0.216
NUMPROM_log10	-0.049	0.005	0.094	-0.056	-0.510	0.299	-0.070	-0.003	0.024	-0.265	-0.748
NGIFTALL_log10	-0.144	0.014	0.101	-0.090	-0.721	0.279	-0.053	0.000	0.006	0.216	0.560
LASTGIFT_log10	0.117	-0.027	0.126	0.031	0.185	0.436	-0.085	0.001	-0.858	0.007	0.032

PCA: Scaled Units

Scaled Units [0,1]	1. eigenvector	2. eigenvector	3. eigenvector	4. eigenvector	5. eigenvector	6. eigenvector	7. eigenvector	8. eigenvector	9. eigenvector	10. eigenvector	11. eigenvector
eigenvalue	0.381	0.333	0.218	0.186	0.054	0.046	0.036	0.027	0.005	0.002	0.002
RFA_2F	-0.057	0.256	-0.283	-0.569	0.622	-0.369	0.001	-0.003	0.000	0.045	-0.047
D_RFA_2A	-0.035	0.147	-0.175	-0.316	0.096	0.898	0.164	0.006	0.003	0.062	-0.004
F_RFA_2A	0.019	-0.841	0.036	0.033	0.436	0.059	0.302	-0.004	-0.002	-0.062	0.018
G_RFA_2A	0.062	0.437	0.306	0.557	0.536	0.084	0.296	-0.003	-0.003	-0.141	0.018
DOMAIN3	0.136	-0.004	-0.706	0.360	0.020	-0.001	-0.012	-0.380	-0.457	-0.001	-0.005
DOMAIN2	-0.770	-0.008	0.291	-0.069	0.006	0.010	-0.016	-0.341	-0.448	-0.001	-0.003
DOMAIN1	0.615	0.010	0.446	-0.304	-0.018	0.010	-0.013	-0.354	-0.451	-0.003	-0.001
DOMAIN4	0.009	-0.002	-0.014	0.008	0.006	-0.001	-0.010	0.783	-0.621	-0.003	-0.006
NUMPROM_log10	-0.012	0.066	-0.035	-0.061	-0.246	-0.142	0.645	-0.006	-0.015	0.128	-0.691
NGIFTALL_log10	-0.023	0.096	-0.084	-0.146	-0.246	-0.158	0.613	-0.004	-0.023	-0.128	0.695
LASTGIFT_log10	0.015	-0.008	0.065	0.118	0.071	-0.026	0.048	0.000	-0.006	0.968	0.189

abbott analytics

PCA: Scaled and Dummy Scaling

Scaled Units [0,1]; Dummies [0.3,0.7]	1. eigenvector	2. eigenvector	3. eigenvector	4. eigenvector	5. eigenvector	6. eigenvector	7. eigenvector	8. eigenvector	9. eigenvector	10. eigenvector	11. eigenvector
eigenvalue	0.148	0.061	0.050	0.039	0.033	0.013	0.007	0.004	0.002	0.002	0.001
RFA_2F	-0.907	-0.044	-0.256	0.204	-0.006	0.253	0.025	0.001	0.023	-0.048	0.000
D_RFA_2A	-0.147	0.000	-0.006	-0.010	0.018	-0.460	-0.861	-0.005	0.154	-0.027	-0.004
F_RFA_2A	0.251	0.092	-0.676	-0.256	-0.117	0.497	-0.336	0.002	-0.168	0.070	0.001
G_RFA_2A	0.060	-0.095	0.534	0.364	0.083	0.536	-0.370	0.001	-0.359	0.094	0.004
DOMAIN3	-0.011	-0.139	-0.060	-0.177	0.768	0.036	-0.002	0.381	0.004	0.003	0.457
DOMAIN2	-0.032	0.767	0.090	0.131	-0.260	-0.003	-0.004	0.341	0.005	0.007	0.448
DOMAIN1	0.045	-0.609	-0.031	0.066	-0.538	-0.046	0.001	0.354	0.004	0.012	0.451
DOMAIN4	0.001	-0.008	-0.005	0.000	0.016	0.001	0.002	-0.783	0.004	0.004	0.622
NUMPROM_log10	-0.124	0.010	0.310	-0.584	-0.129	0.209	-0.065	0.004	0.013	-0.695	0.016
NGIFTALL_log10	-0.238	0.016	0.273	-0.599	-0.119	0.041	-0.005	-0.003	-0.024	0.702	-0.002
LASTGIFT_log10	0.114	-0.020	0.097	0.086	0.008	0.375	-0.063	-0.001	0.904	0.084	-0.007

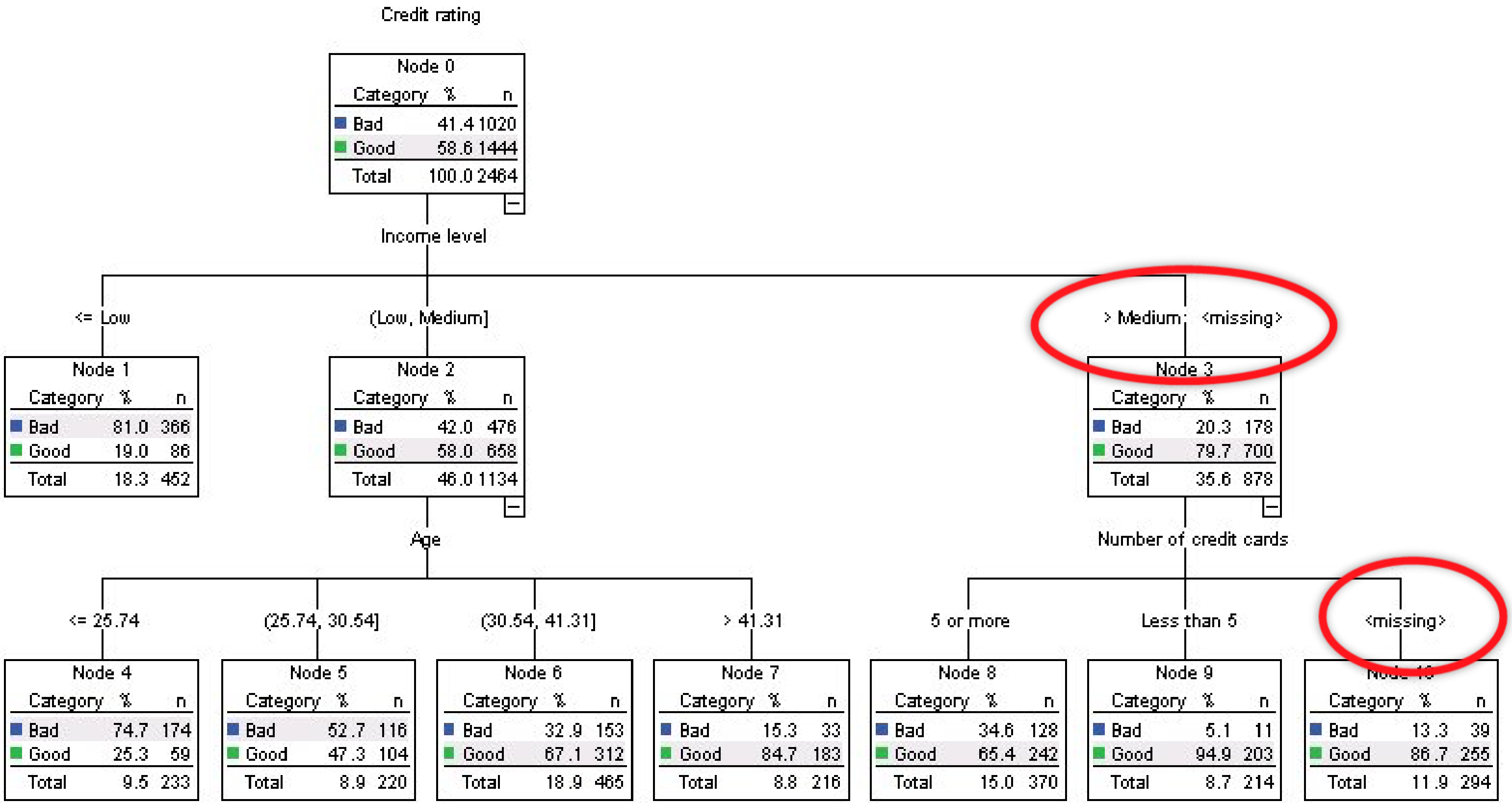
Scaled Units [0,1]	1. eigenvector	2. eigenvector	3. eigenvector	4. eigenvector	5. eigenvector	6. eigenvector	7. eigenvector	8. eigenvector	9. eigenvector	10. eigenvector	11. eigenvector
eigenvalue	0.381	0.333	0.218	0.186	0.054	0.046	0.036	0.027	0.005	0.002	0.002
RFA_2F	-0.057	0.256	-0.283	-0.569	0.622	-0.369	0.001	-0.003	0.000	0.045	-0.047
D_RFA_2A	-0.035	0.147	-0.175	-0.316	0.096	0.898	0.164	0.006	0.003	0.062	-0.004
F_RFA_2A	0.019	-0.841	0.036	0.033	0.436	0.059	0.302	-0.004	-0.002	-0.062	0.018
G_RFA_2A	0.062	0.437	0.306	0.557	0.536	0.084	0.296	-0.003	-0.003	-0.141	0.018
DOMAIN3	0.136	-0.004	-0.706	0.360	0.020	-0.001	-0.012	-0.380	-0.457	-0.001	-0.005
DOMAIN2	-0.770	-0.008	0.291	-0.069	0.006	0.010	-0.016	-0.341	-0.448	-0.001	-0.003
DOMAIN1	0.615	0.010	0.446	-0.304	-0.018	0.010	-0.013	-0.354	-0.451	-0.003	-0.001
DOMAIN4	0.009	-0.002	-0.014	0.008	0.006	-0.001	-0.010	0.783	-0.621	-0.003	-0.006
NUMPROM_log10	-0.012	0.066	-0.035	-0.061	-0.246	-0.142	0.645	-0.006	-0.015	0.128	-0.691
NGIFTALL_log10	-0.023	0.096	-0.084	-0.146	-0.246	-0.158	0.613	-0.004	-0.023	-0.128	0.695
LASTGIFT_log10	0.015	-0.008	0.065	0.118	0.071	-0.026	0.048	0.000	-0.006	0.968	0.189

Scaled Units [0,1]; Dummies [0.3,0.7]	1. eigenvector	2. eigenvector	3. eigenvector	4. eigenvector	5. eigenvector	6. eigenvector	7. eigenvector	8. eigenvector	9. eigenvector	10. eigenvector	11. eigenvector
eigenvalue	0.148	0.061	0.050	0.039	0.033	0.013	0.007	0.004	0.002	0.002	0.001
RFA_2F	-0.907	-0.044	-0.256	0.204	-0.006	0.253	0.025	0.001	0.023	-0.048	0.000
D_RFA_2A	-0.147	0.000	-0.006	-0.010	0.018	-0.460	-0.861	-0.005	0.154	-0.027	-0.004
F_RFA_2A	0.251	0.092	-0.676	-0.256	-0.117	0.497	-0.336	0.002	-0.168	0.070	0.001
G_RFA_2A	0.060	-0.095	0.534	0.364	0.083	0.536	-0.370	0.001	-0.359	0.094	0.004
DOMAIN3	-0.011	-0.139	-0.060	-0.177	0.768	0.036	-0.002	0.381	0.004	0.003	0.457
DOMAIN2	-0.032	0.767	0.090	0.131	-0.260	-0.003	-0.004	0.341	0.005	0.007	0.448
DOMAIN1	0.045	-0.609	-0.031	0.066	-0.538	-0.046	0.001	0.354	0.004	0.012	0.451
DOMAIN4	0.001	-0.008	-0.005	0.000	0.016	0.001	0.002	-0.783	0.004	0.004	0.622
NUMPROM_log10	-0.124	0.010	0.310	-0.584	-0.129	0.209	-0.065	0.004	0.013	-0.695	0.016
NGIFTALL_log10	-0.238	0.016	0.273	-0.599	-0.119	0.041	-0.005	-0.003	-0.024	0.702	-0.002
LASTGIFT_log10	0.114	-0.020	0.097	0.086	0.008	0.375	-0.063	-0.001	0.904	0.084	-0.007

Missing Value Imputation

- Delete the record (row), or delete the field (column)
- Replace with a constant
- Replace missing value with mean, median, or distribution
- Replace missing with random self-substitution
- Surrogate Splits (CART)
- Make missing a category
  - Simple for "rule-based" algorithms; Turn continuous into categorical for numeric algorithms
- Replace with the missing value with an estimate
  - Select value from another field having high correlation with variable containing missing values
  - Build a model with variable containing missing values as output, and other variables without missing values as an input

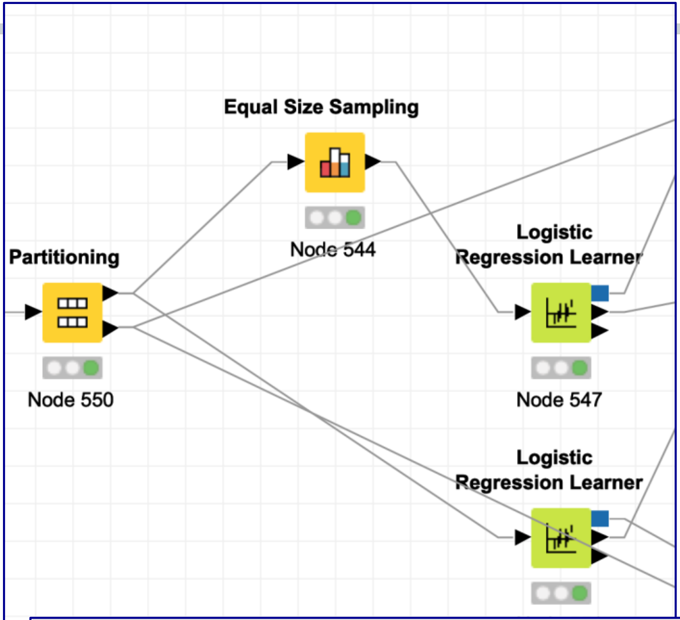
CHAID Trees: Missing Values are Just Another Category



Summary

Data Preparation Step	Linear Regression	K-NN	K-Means Clustering	PCA	Neural Networks	Decision Trees
Fill Missing Values	Y	Y	Y	Y	Y	ळ
Correlation Filtering	Y	Y	Y			
De-Skew (log, box-cox)	Y	Y	Y	Y		
Mitigate Outliers	Y	Y	Y	Y	ळ	ळ
Remove Magnitude Bias (Scale)	Y	Y	Y	Y	ळ	
Remove Categorical "Dummy" Bias	Y	Y	Y	Y		
Mitigate Categorical Cardinality Bias	छ़घ़	छ़घ़	छ़घ़	छ़घ़	छ़घ़	Y

Stratify or Not to Stratify…
That is the Question!?

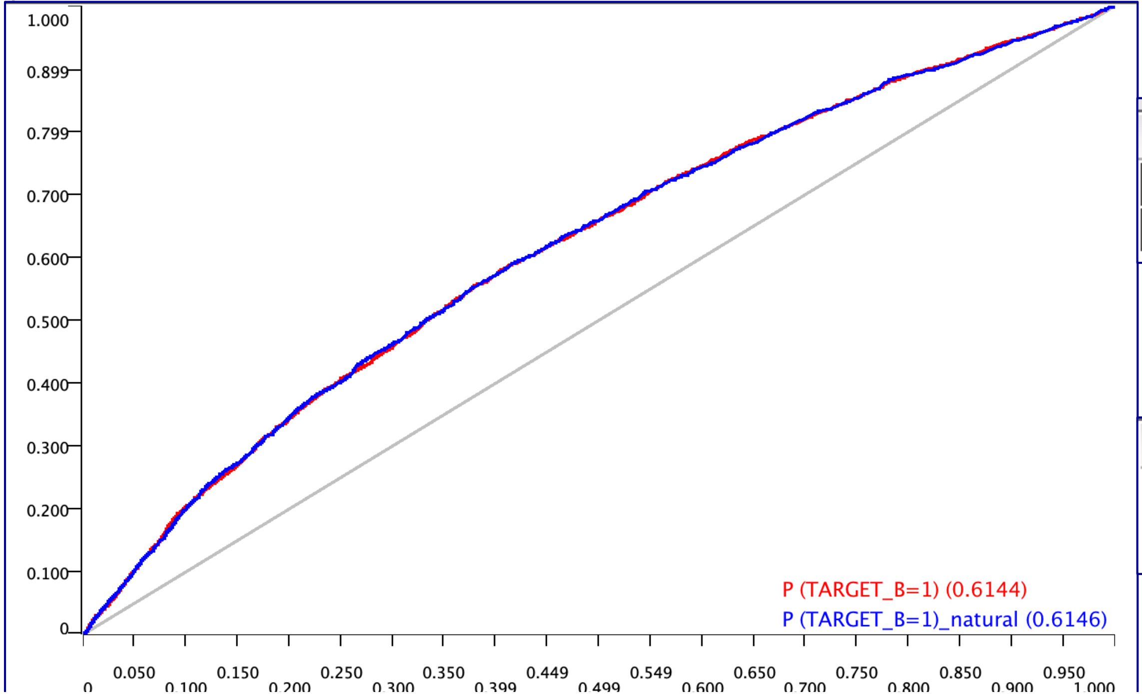


5.1% TARGET_B = 1:
unbalanced data

TARGET_B \ Prediction (TARGET_B)	1	0
1	0	2418
0	0	45288

Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Recall	D Precision
1	0	0	45288	2418	0	?
0	45288	2418	0	0	1	0.949
Overall	?	?	?	?	?	?

Comparing Logistic Regression with and without Equal Size Sampling



No Stratified Sampling

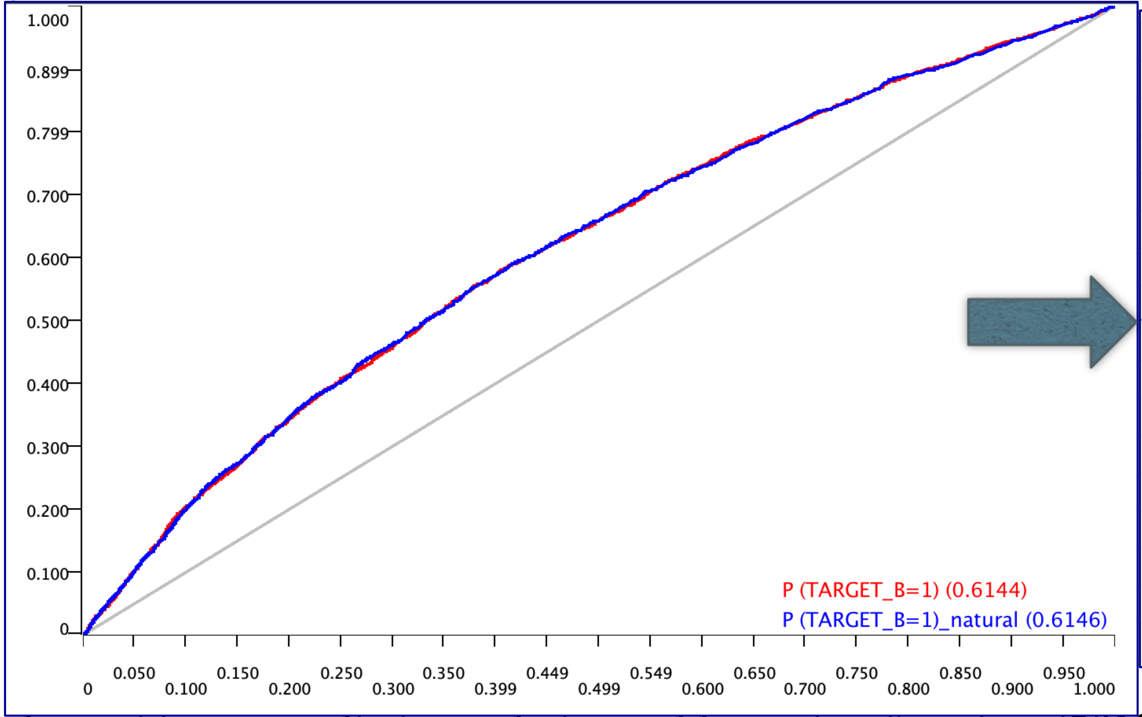
Row ID	1	0
1	0	2418
0	0	45288

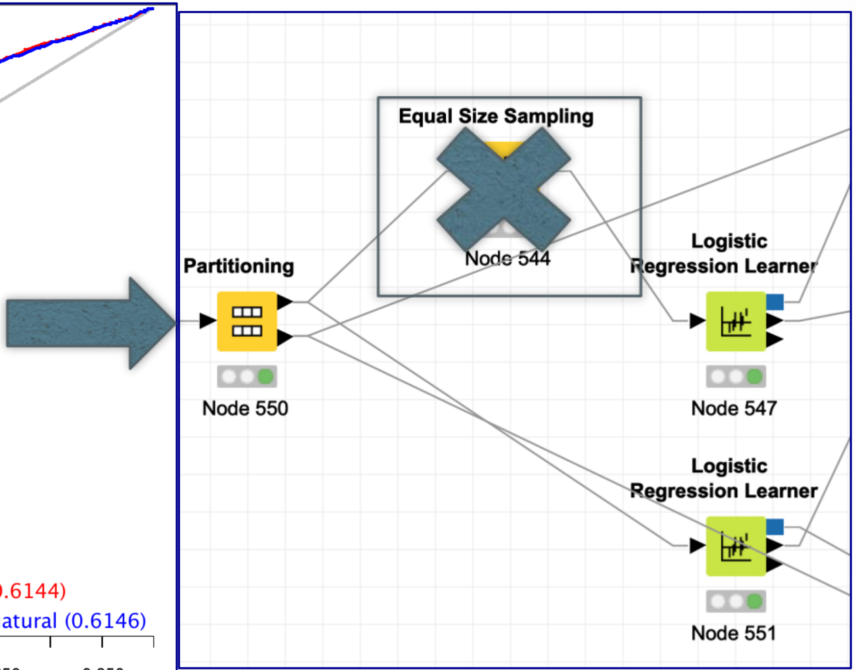
Equal Sampling

Row ID	1	0
1	1421	997
0	18725	26563

https://www.predictiveanalyticsworld.com/sanfrancisco/2013/pdf/Day2_1050_Abbott.pdf

Don't Need to Stratify With Many Algorithms





https://www.predictiveanalyticsworld.com/sanfrancisco/2013/pdf/Day2_1050_Abbott.pdf

Know the Algorithm when Developing Sampling Strategy

Variable	Stratified: Coeff.	Std. Err.	P>|z|	Natural (orig): Coeff._natural	Std. Err._natural	P>|z|_natural	coeff diff	coeff compare
RFA_2F	-0.133532984	0.0338	0.000	-0.1563345	0.024	0.000	0.023	within SE
D_RFA_2A	-0.163727182	0.1210	0.176	-0.0934212	0.079	0.237	0.070	within SE
F_RFA_2A	0.038231571	0.0884	0.665	0.0357819	0.062	0.565	0.002	within SE
G_RFA_2A	0.316663027	0.1267	0.012	0.2779701	0.091	0.002	0.039	within SE
DOMAIN2	-0.068966948	0.0767	0.369	-0.1169964	0.056	0.036	0.048	within SE
DOMAIN1	-0.266408264	0.0837	0.001	-0.2845323	0.060	0.000	0.018	within SE
NGIFTALL_log 10	-0.46212497	0.0998	0.000	-0.4444304	0.072	0.000	0.018	within SE
LASTGIFT_log 10	0.062766545	0.2044	0.759	0.1813683	0.141	0.199	0.119	within SE
Constant	0.695770991	0.2785	0.012	3.5393926	0.194	0.000	2.844	outside SE

Input Variable Interactions

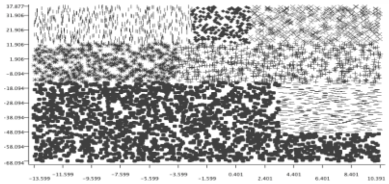
- Algorithms are mixed on interactions in theory
  - Linear Regression, Logistic Regression, kNN, kMeans clustering, PCA…. are **main effect models**

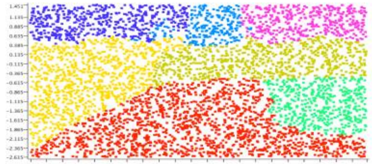
- Decision trees are greedy searchers
  - Built to find interactions
  - But, only if they can be found in sequence (one at a time, stepwise)

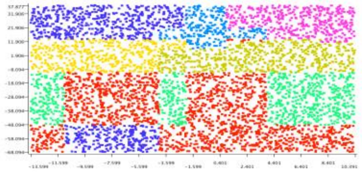
- Neural Networks find interactions well (XOR)

- Naïve Bayes find *intersections*, not interactions

- Algorithms don't always identify interactions well or well-enough in practice

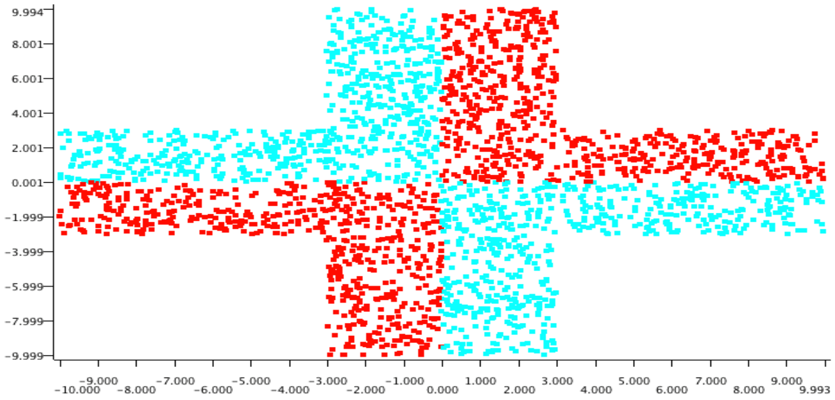




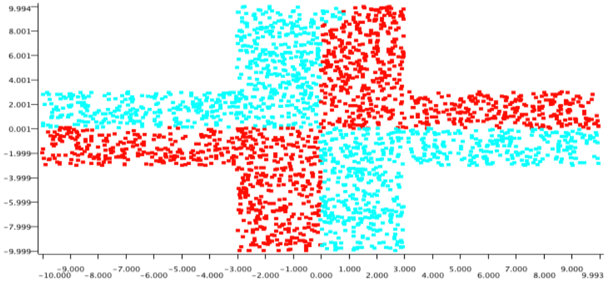


Simple Interaction Function

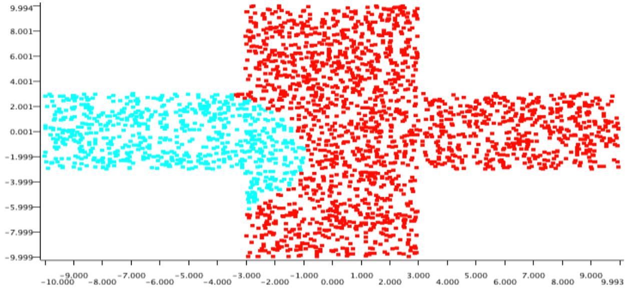
- Two uniform variables: x and y
- 2,564 records
- if (x*y > 0) return ("1");
    - else return("0");



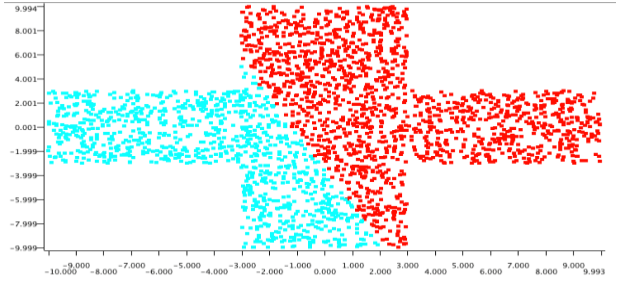
Four Classifiers



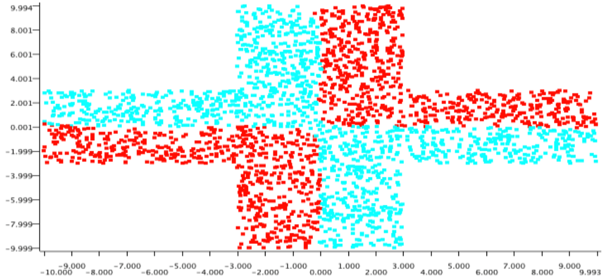
Decision Tree, min Leaf node 50 records



Naïve Bayes

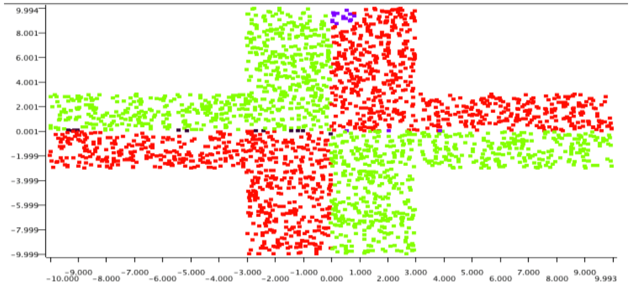


Logistic Regression



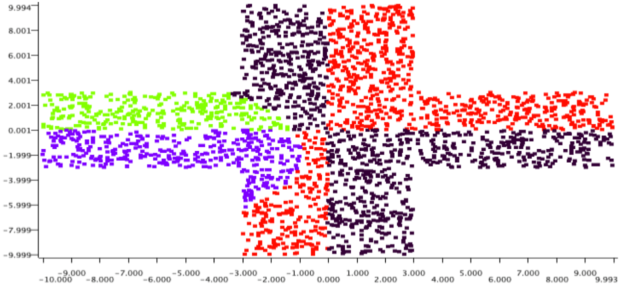
Rprop Neural Net, 300 epochs

Errors

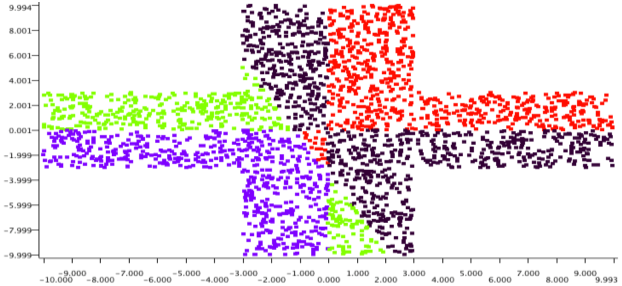


True correct
False incorrect
False correct
True incorrect

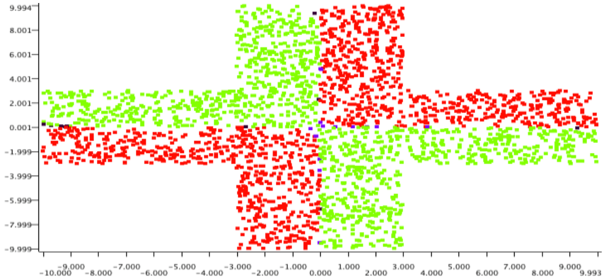
Decision Tree, min Leaf node 50 records



Naïve Bayes



Logistic Regression



Rprop Neural Net, 300 epochs

Don't Build Interactions Manually*

- Too many…too many

Table 4-16: Number of Two-Way Interaction Combinations

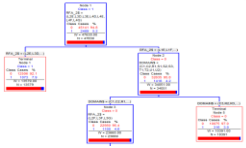
NUMBER OF VARIABLES	NUMBER OF POSSIBLE TWO-WAY INTERACTIONS
5	10
10	45
50	1,225
100	4,950
500	124,750
1000	499,500

- So what do you do?

* Except for those you know about

Automatic Interaction Detection

- Trees: build 2-level trees
  - Pros: works with continuous and categoricals
  - Cons: greedy, only finds one solution at a time (Battery)



- Association rules: build 2-antecedent rules
  - Pros: exhaustive
  - Cons: only works with categoricals

A	B	C	D	E	F	G
row ID	Support	Confidence	Lift	Consequent	Split Value 1	Split Value 2
rule0	0.01004613	34.1176471	3.47151569	DaysToNextPurchase_le_60	AssetCount_51-100	PurchaseFlag_eq_true
rule8	0.01004613	5.30164534	4.8576988	DaysToNextPurchase_4-7	ChannelEngagement_8000-20000	DaysSinceLastPurchase_31-60
rule18	0.01004613	3.17982456	1.55197428	DaysToNextPurchase_le_7	AverageDaysBetweenVisits_31-60	DaysSinceLastPurchase_null
rule22	0.01004613	3.17982456	1.55197428	DaysToNextPurchase_le_7	PriorPurchase_eq_false	AverageDaysBetweenVisits_31-60
rule26	0.01004613	8.49194729	4.14465751	DaysToNextPurchase_le_7	ChannelEngagement_8000-20000	DaysSinceLastPurchase_ge_91
rule30	0.01004613	16.3380282	4.47697953	DaysToNextPurchase_le_14	AssetCount_51-100	DaysSinceLastPurchase_null
rule34	0.01004613	16.3380282	4.47697953	DaysToNextPurchase_le_14	PriorPurchase_eq_false	AssetCount_51-100
rule38	0.01004613	3.9862543	1.09232147	DaysToNextPurchase_le_14	DaysSinceLastPurchase_ge_91	DaysSinceLastVisit_61-90
rule42	0.01004613	7.03030303	1.0945618	DaysToNextPurchase_le_30	VisitQuality_1-1000	AverageDaysBetweenVisits_le_1
rule46	0.01004613	13.4570766	2.09515889	DaysToNextPurchase_le_30	AssetCount_11-20	DaysSinceLastVisit_8-14
rule50	0.01004613	9.78077572	0.99520686	DaysToNextPurchase_le_60	ChannelEngagement_1000-3000	AssetCount_6-10

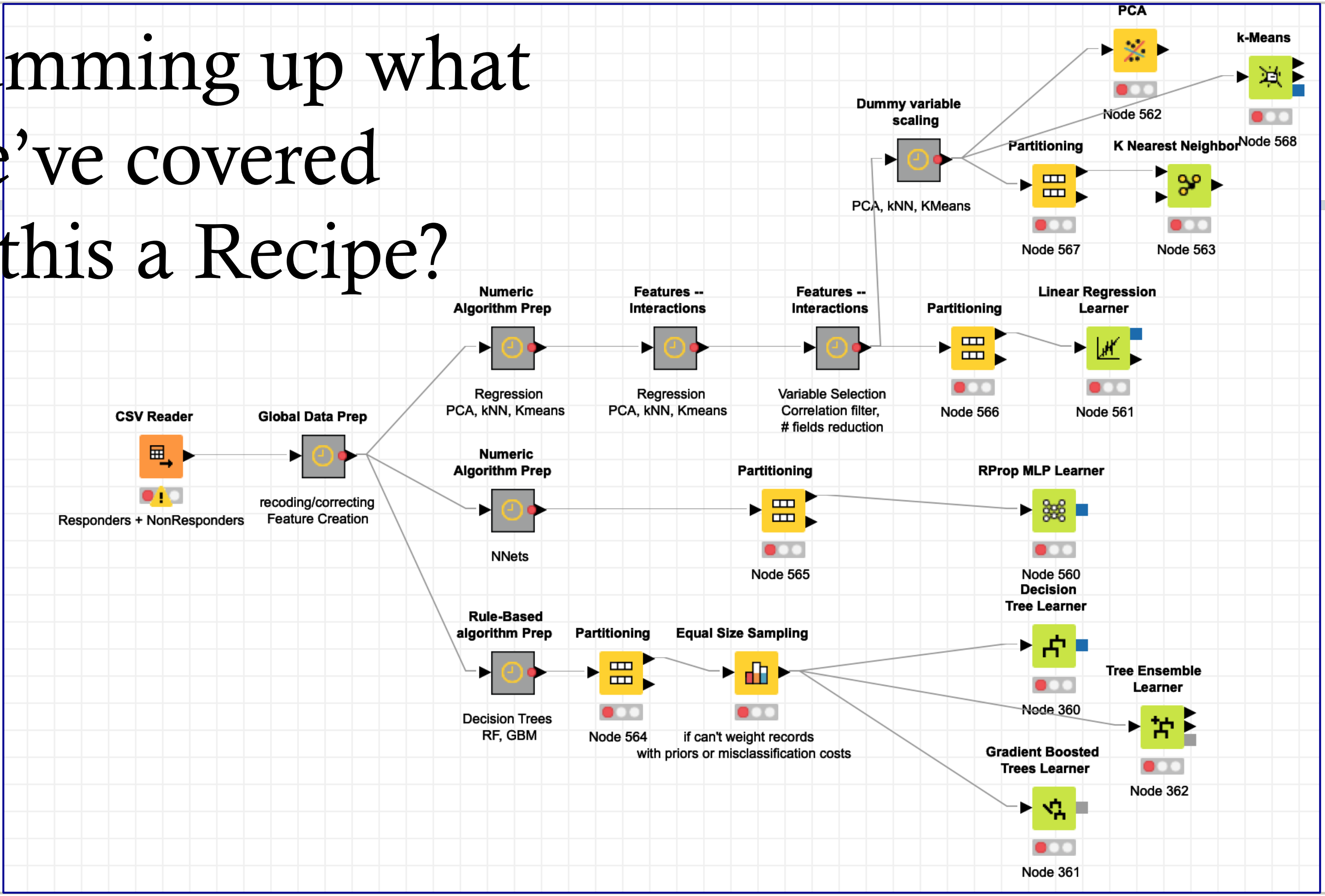
- Use the linear/logistic regression algorithm itself, loop over all 2-way interactions
  - Pros: context is the model you may want to use, easy to do in R, Matlab, Python, SAS (coding)
  - Cons: slow, have to code, what to do with dummies

Statistics on Logistic Regression

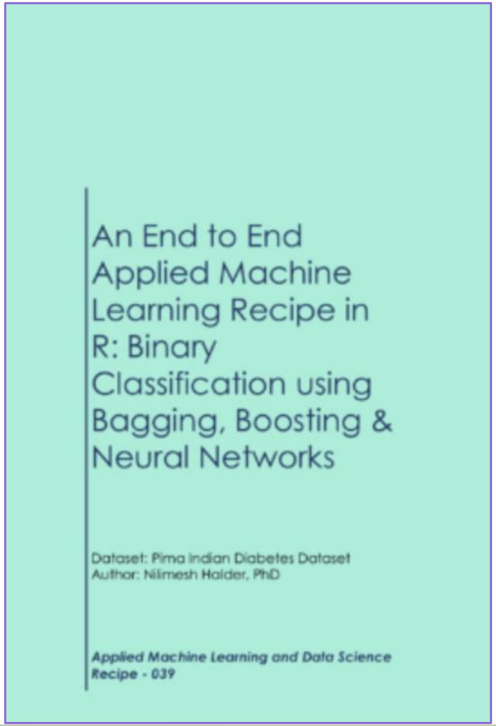
Logit	Variable	Coeff.	Std. Err.	z-score	P>|z|
1	NGIFTALL	0.0239	0.0034	7.0132	2.33E-12
	LASTGIFT	-0.0093	0.0028	-3.3185	0.0009
	Constant	-0.0985	0.0694	-1.419	0.1559

Log-likelihood = -3,322.7813
Number of iterations = 8

Summing up what we've covered
Is this a Recipe?



Is it a Recipe?....YES!


An End to End Applied Machine Learning Recipe in R: Binary Classification using Bagging, Boosting & Neural Networks

Dataset: Pima Indian Diabetes Dataset
Author: Nilmesh Halder, PhD

Applied Machine Learning and Data Science Recipe - 039

Can we apply a recipe to machine learning and data science modeling processes?

Conclusions

- Know what the algorithms can do (and not do!) before deciding on data preparation
  - When are data shapes and data ranges important?
- It's not hard….just requires some thought
- Once you know what to do, you have your recipe!