Predictive Modeling Techniques in Marketing

by Michael Vainder

VP Modeling

March 28 , 2019

ENVIRONICS ANALYTICS

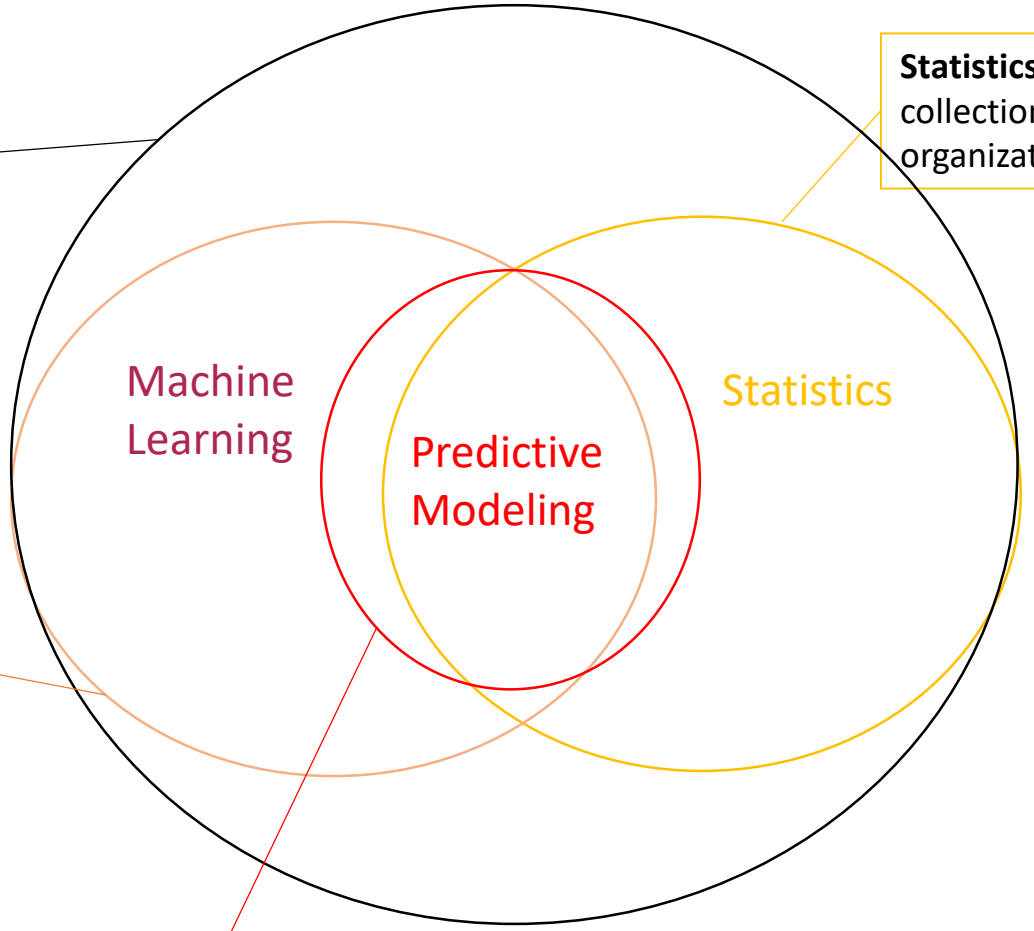
Main bullets

- Predictive modeling
- Marketing applications
- Predictive modeling process
- Overview of techniques and some challenges

Venn Diagram

Data Mining is a process of discovering new insights using Statistics and Machine Learning

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data



Machine Learning is a data analysis and knowledge extraction using computational techniques

Predictive Modeling is a process that use statistical and machine learning techniques to predict new or future observations

Common Predictive Modeling Techniques

- Regression
- Discriminant Analysis
- Decision Trees
- k-Nearest Neighbours (k-NN)
- Naïve Bayes
- Support Vector Machines (SVM)
- Neural Network

Some Marketing Activities addressed by Predictive Modeling

- Customer Acquisition
- Cross/Up selling
- Churn management
- Lifetime Value prediction
- Fundraising
- Multichannel customer management
- Price optimization
- Assortment planning
- Risk management
- Site selection

ENVIRONICS ANALYTICS

Customer Acquisition

- Questions to answer: Who is our customer? Where to get a new customers?
- Solution: Build a model that will locate potential customers in the market for the product or service.
- Input: List of client's current customers with appended customers information (demographic and behavioural data).
- Output: List of potential new customers and the corresponding scores.
- Techniques: Profiling, logistic regression, decision tree, nearest neighbours algorithm and other data mining techniques.

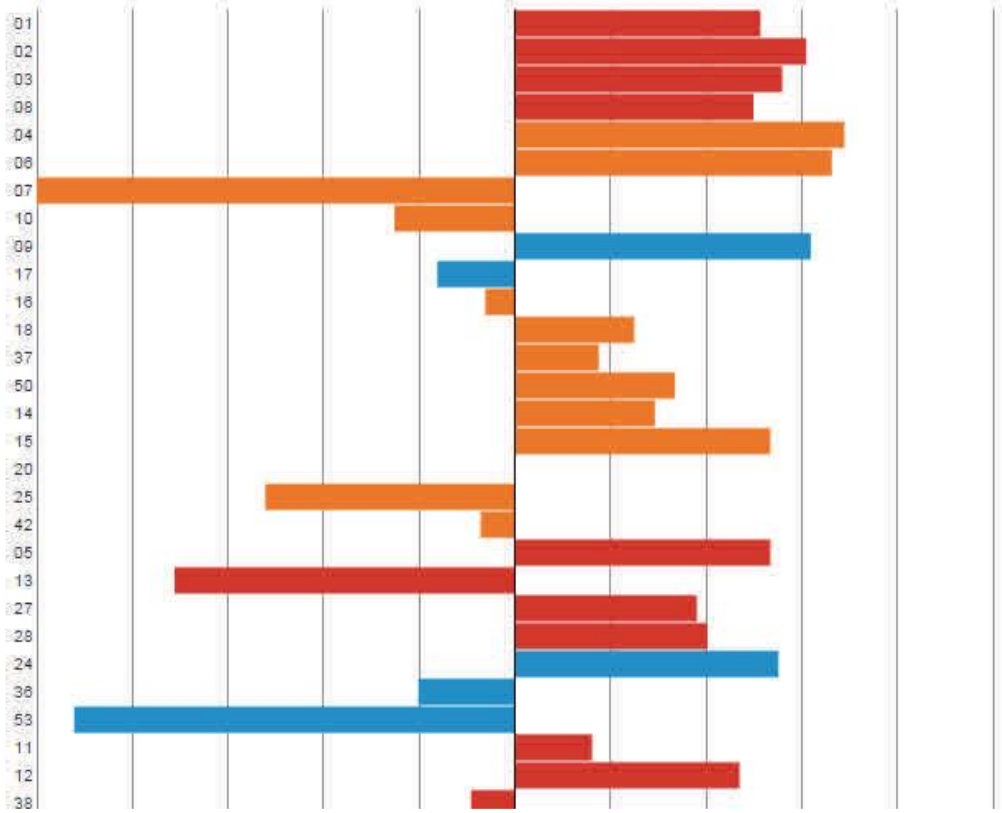
Customer Acquisition: Segmentation Based Targeting Method

ENVIRONICS ANALYTICS

Profile - Customers

Sample Customers - CustomerCount vs Canada

SG	SESI	Name	Count	%	Base Count	Base %	% Pen	Index
U1	1	Cosmopolitan Elite	197	1.01	72,436	0.49	0.27	205
U1	2	Urbane Villagers	444	2.29	131,113	0.90	0.34	255
U1	3	Arts & Affluence	344	1.77	114,703	0.78	0.30	226
U1	8	Boomerang City	871	4.49	330,383	2.26	0.26	199
S1	4	Suburban Success	515	2.65	121,107	0.83	0.43	320
S1	6	Kids & Careers	1,373	7.07	350,202	2.40	0.39	295
S1	7	Nouveaux Riches	0	0.00	108,647	0.74	0.00	0
S1	10	Emptying Nests	146	0.75	146,851	1.00	0.10	75
E1	9	Satellite Burbs	1,254	6.46	362,040	2.48	0.35	261
E1	17	Exurban Wonderland	221	1.14	198,051	1.35	0.11	84
S2	16	Pets & PCs	534	2.75	429,400	2.94	0.12	94
S2	18	Management Material	297	1.53	168,135	1.15	0.18	133
S2	37	Trucks & Trades	491	2.53	304,593	2.08	0.16	121
S2	50	Suburban Scramble	507	2.61	253,819	1.74	0.20	150
S3	14	Diversity Heights	366	1.89	195,463	1.34	0.19	141
S3	15	Heritage Hubs	747	3.85	263,527	1.80	0.28	214
S3	20	South Asian Achievers	150	0.77	113,040	0.77	0.13	100
S3	25	South Asian Society	86	0.44	136,353	0.93	0.06	48
S3	42	Home Sweet Rows	295	1.52	239,849	1.64	0.12	93
U2	5	Asian Sophisticates	456	2.35	160,538	1.10	0.28	214
U2	13	Asian Avenues	73	0.38	188,840	1.29	0.04	29
U2	27	Diverse City	406	2.09	190,285	1.30	0.21	161
U2	28	Metro Multiculturals	575	2.96	259,943	1.78	0.22	167
E2	24	Fresh Air Families	1,004	5.17	340,784	2.33	0.29	222
E2	36	Exurban Homesteaders	175	0.90	165,041	1.13	0.11	80
E2	53	Outdoor Originals	17	0.09	155,893	1.07	0.01	8
U3	11	Urban Digerati	425	2.19	268,579	1.84	0.16	119
U3	12	Street Scenes	595	3.06	239,016	1.64	0.25	188
U3	38	Grads & Pads	232	1.20	192,884	1.32	0.12	91



ENVIRONICS ANALYTICS

Customer Acquisition: Modeling Based Targeting Method

The target variable is buy or not buy. Potential predictors are demographic and behavioural variables appended to the client's customer database

Build the model using logistic regression, decision tree, nearest neighbour algorithm or other data mining techniques.

Apply the model to the list of new potential customers (can be list from the third party) and score them. After that sort by score and based on lift provided by the model target best segments.

Price Optimization

- Questions to answer: How our retailing pricing strategy can be customized for different brands, stores, regions, markets? What factors are most important in our retailing pricing strategy?
- Solution: Develop the model that will predict price sensitivity - on the basis of price/demand information and available customer demographics and competitive information.
- Input: Product price and volume information. Competitors price information (if available).
- Output: Factors that explain variation in price and their magnitude of impact.
- Techniques: Regression, decision tree, time series, data mining algorithms.

Price Optimization using regression approach

Client has requested to help to understand how changes in price, promotions and TV Ads affect sales volume for two given banners

The dataset represents a maximum of 160 weekly observations on banner level variables such as:

- All Sales: Unit Volume, Ton Volume, $;
- Promotion Sales: Unit Volume, Ton Volume, $;
- Ads indicator

Model:

$\log(\text{Units}) = a_0 + a_1 * \log(\text{Price}) + a_2 * \log(\text{Price cut ratio}) + a_3 * \text{Ad}$

Price Optimization using regression approach. Results

	Elasticity	Promotion coefficient	Ads coefficient
Banner A	-1.0898**	2.7339**	0.0838
Banner B	-1.7652***	0.3780	0.4191***

* Significant at 0.05 level; ** Significant at 0.01 level; *** Significant at 0.001 level;

- Results show that Product for A and B banners is elastic (price elasticity is less than -1). For A banner a 1% in price increase will result in 1.1% decrease in sales. For B banner a 1% price increase will result in 1.76% decrease in sales.
- The promotion coefficient represents a price cut. For A banner increase in 1% cut price increase sales by 2.7 % in sales. For B banner increase in 1% cut price increase sales only by 0.38% in sales. For B banner this parameter is not statistically significant
- Results shows that Ads can increase sales for banners. Even if the results indicate that A banner sales are ads-dependent, this parameter is not statistically significant.

ENVIRONICS ANALYTICS

Cross – Selling

- Questions to answer: What products should my firm cross-sell and to which customers?
- Solution: Build a model that will predict the products the customer is most likely to buy with other products as the next purchase .
- Input: Customer behavioural database with available customer demographic information.
- Output: Probability to buy a given product or list of recommended products the customers are likely to buy next.
- Techniques: Regression, decision tree, collaborative filtering, other data mining techniques.

Cross – Selling. Regression approach

A telecom client has requested to help in identifying current customers who has no home phone but has high chance to get this product.

The target variable is customer has or not has home phone. Potential predictors are customer usage and billing information, tenure and EA demographic variables appended to customer database using postal code as key variable.

Model:

$$\text{Probability to buy home phone} = \frac{1}{1+e^{-XB}}$$

Where X – Predictors, B – coefficients for predictors

Cross – Selling. Collaborative filtering approach

1 User Attributes
(Geo-Demographic, Lifestyle data)

	Attrib1	Attrib2	Attrib3	Attrib4	Attrib5
User1	1	10	5	1	1
User2	0	21	1	2	3
User3	2	44	1		1
User4	1	18	0	8	2
User5	0	16	2	3	4

2 User similarity matrix

	User1	User2	User3	User4	User5
User1		4.8	5. 2	11.1	6.9
User2	4.8		4	8.7	4.4
User3	5.2	4		9.1	4.8
User4	11.2	8.7	9.1		7.1
User5	6.9	4.4	4.8	7.1	

3 List of closest (similar) users

	1st place	2nd place
For User1	User2	User3
For User4	User5	User2

4 Purchase history

	Item1	Item2	Item3	Item4	Item5
User1		1	1		1
User2			1		
User3	1		1		1
User4	1	1		1	
User5		1	2	1	

5 Recommendation list based on User CF model

For User1: {Item3} + {Item1,Item3,Item5} = {Item1,Item3,Item5}

Up-Selling

- Questions to answer: How much more can a firm sell to current customers?
- Solution: Build a model that will predict up-sell potential for current customers.
- Input: Customer behavioural database with available customer demographic information.
- Output: List of current customers with score/probability to sell more.
- Techniques: Logistic regression, decision tree, neural network and other data mining algorithms, survival analysis.

Churn Management

- Questions to answer: How can we control customer churn?
- Solution: Build the model that can identify which customers are likely to churn and why they might churn. That will help client focus on proactive churn management and prevent customers from churning.
- Input: Client's database of current and churned customers with demographic and behavioural information observed for several time periods (months, years).
- Output: List of customers with assigned score/probability of ceasing buying/donating. Attributes that predict the churn and their impact on churn.
- Techniques: Logistic regression, decision tree, time series, survival analysis.

Predictive modeling process

- Data preparation (60-70%)
  - Identify target variable
  - Identify potential predictors
  - Compile available data sources
  - Pre-processing (clean up, conducting exploratory data analysis)
- Creating development and testing samples
- Modeling (35-25%)
  - Select variables
  - Select modeling technique
  - Estimate model
  - Evaluate model
- Deployment (5%)

Data preparation

Some questions to be answered during the data preparation process

- Are there timing considerations?
- What level of detail should the datasets be aggregated to before merging (temporal, spatial, other)?
- Is there a unique key that can be used to merge or compile available datasets
- What should be done with missing values in potential predictors?

Pre-processing

- Checking for duplicate records
- Obtain descriptive statistics for target and potential predictor variables
- Plotting each potential predictor variable versus the target variable to see if any relationship exists.
- Running correlation analysis between target and potential predictors and also between potential predictors.
- Performing variable transformations (for continuous variables), such as log/power/inverse/ for continuous variables or (re)grouping or ‘binning’ of categorical variables.

Creating training and testing samples

- Splitting (sampling without replacement)
- Cross validation
- Bootstrap

Why do we do this? One reason is because attempting to test the performance of any model on the same set of data on which the model was built will produce overoptimistic results

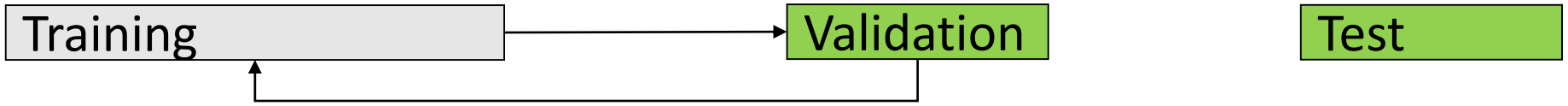
Overfitting

Goal is to build the model that fits the data well … but a model is said to be ‘over-fit’ when it produce very accurate results on a given (training) dataset but has poor accuracy when it is used to predict within a new dataset or sample

What can cause overfitting?

- Outliers
- Non-significant variables (variables that have some idiosyncrasies of the specific training set and do not represent the general population)
- Bias in the training data (e.g. small training sample size)
- Complex model (e.g. too many predictors)

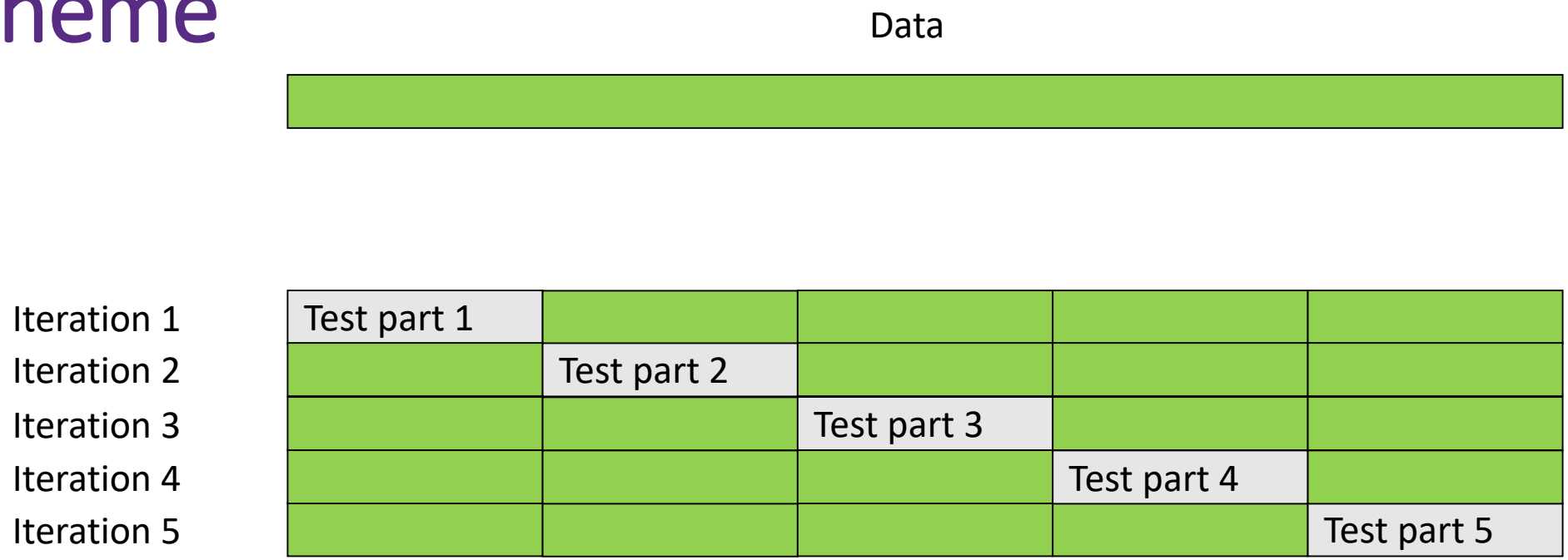
Validation & Evaluation: splitting schema



Validation & Evaluation: Splitting schema

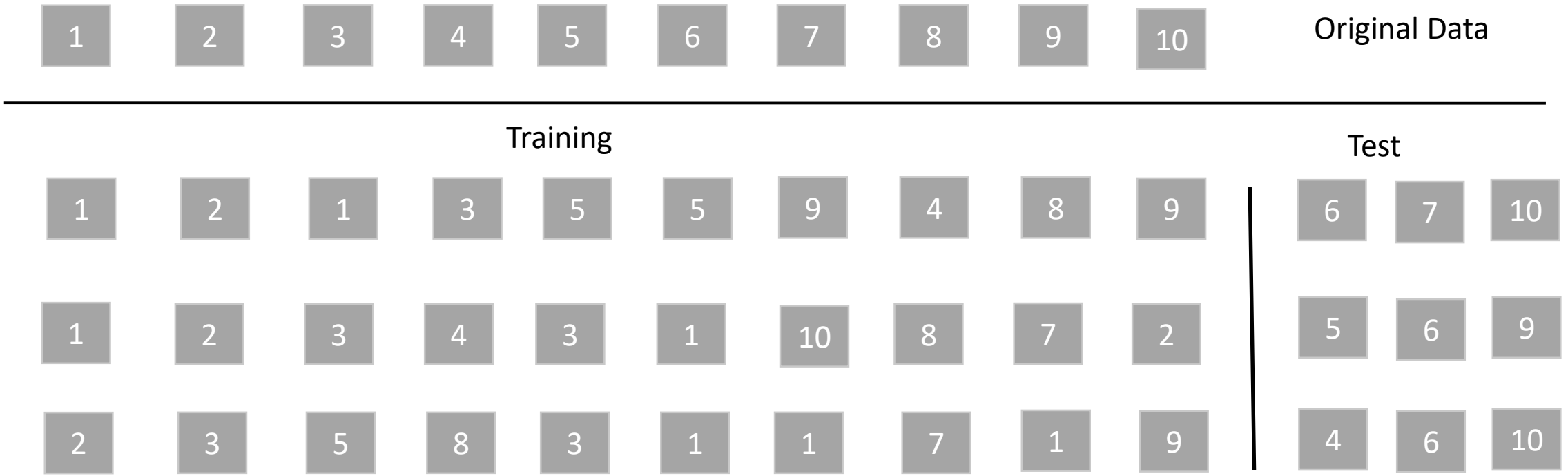
How should we split the data into a training, validation and test sets?
Common rules: 60/20/20 , 50/25/25

Validation & Evaluation: multi-folder cross validation scheme



Validation & Evaluation: bootstrap scheme

- A bootstrap procedure involves repeated resampling with replacement. We take many random samples with replacement from the dataset, and for each of these samples, we compute a performance measure.



Do we have enough data for modeling?

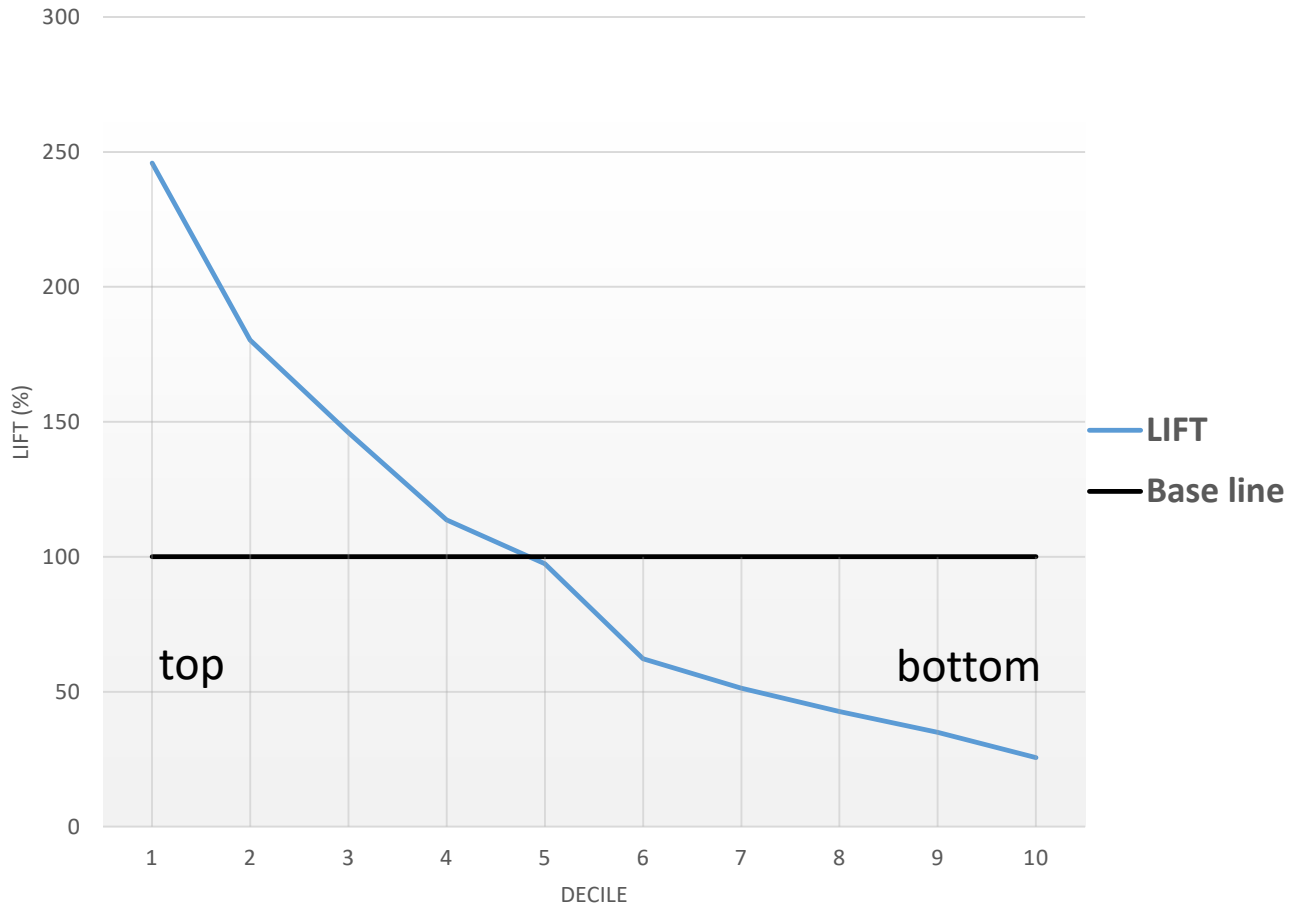
- How large should the sample be?
- How many events (for binary outcome) in a sample do we need to build a model?
  - Min 10 cases with events/outcomes per predictor variable

- Sample size

$$N=10*k/p,$$

Where k = number of predictors, p = proportion of events/outcomes

Performance evaluation metrics: ranking metrics



Lift Chart. Lift is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model.

Metrics for numeric prediction

$$Mean\ Absolute\ Error = \frac{|\widehat{Y_1} - Y_1| + \cdots + |\widehat{Y_n} - Yn|}{n}$$

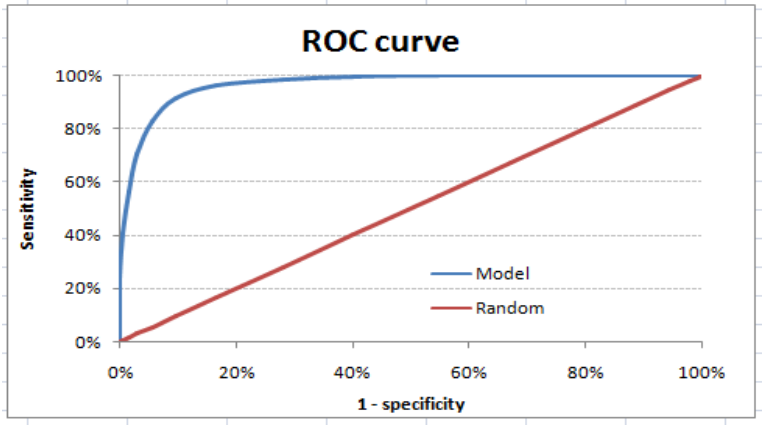
$$Root\ Mean\ Squared\ Error = \sqrt{\frac{(\hat{Y}_1 - Y_1)^2 + \cdots + (\hat{Y}_n - Yn)^2}{n}}$$

$$Relative\ Squared\ Error = \frac{(\hat{Y}_1 - Y_1)^2 + \cdots + (\hat{Y}_n - Yn)^2}{(Y_1 - \bar{Y})^2 + \cdots + (Y_n - \bar{Y})^2}$$

where $\widehat{Y}$ is predicted value, Y is actual value, $\bar{Y}$ is mean of actual values

Classification model evaluation metrics

		Predicted	
		+	-
Actual	+	TP	FN
	-	FP	TN



Accuracy= $\frac{TP+TN}{TP+FN+FP+TN}$ Recall= $\frac{TP}{TP+FN}$ Precision = $\frac{TP}{TP+FP}$ F1 score= $\frac{2*(Recall*Precision)}{Recall+Precision}$

Sensitivity = Recall Specificity= $\frac{TN}{TN+FP}$

Multi Class:

Average Recall=1/K* $\sum_{i=1}^{K}\frac{TP_i}{TP_i+FNi}$ Average Precision=1/K* $\sum_{i=1}^{K}\frac{TP_i}{TP_i+FPi}$ Average F1 score=1/K$\sum F1_i$ score

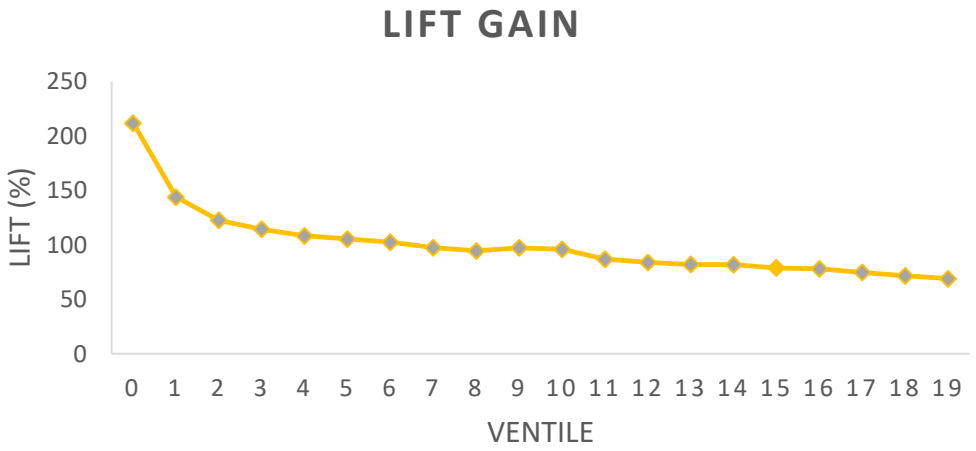
ENVIRONICS ANALYTICS

What is better?

Sometimes traditional statistical measures don't really give a feel for how successful the model is

- $R^2 = 0.15$
- Relative squared error = 85%
- But excellent lift curve - as it goes down nicely with each ventile

Actual $	Predicted $	Rank
300	480	1
210	360	2
150	175	3
130	150	4
110	120	5
90	46	6
60	45	7
50	43	8
43	36	9
10	20	10



Example: Annual donations

Variable selection methods

- **Wrapper** methods utilize the given predictive algorithm to find the best subset of variables using predictive accuracy as performance evaluation measure
- **Filter** methods estimate the usefulness of each variable for the prediction process according to various metrics, evaluating variables one at a time. They are independent from predictive algorithm and applied as a step prior to model selection.
- **Principal Component Analysis** create multiple new variables from correlated groups of predictors. Those new variables exhibit little or no correlation between them (orthogonal) — making them much more useful in modeling they may be representative of the different components of underlying information making up the original variables.

Magnitude of correlation

A

Less than 0.1	insubstantial
[0.1-0.3)	small
[0.3-0.5)	moderate
Greater than 0.5	large

B

Less than 0.1	insubstantial
[0.1-0.3)	small
[0.3-0.5)	moderate
[0.5-0.75)	large
[0.75-0.95)	very large
[0.95-1]	perfect

Regression type

- Linear regression (continuous target variable) : oldest and widely used type of regression technique. It can help to understand and predict customers/consumers behavior, business trends and factors influencing the business decisions. Disadvantages: very sensitive to outliers and multicollinearity between predictors.
- Logistic regression (binary target variable) : often used in marketing for customers/consumers scoring. Disadvantages: sensitive to outliers and multicollinearity between predictors.
- Poisson regression (count target variable) : is appropriate when we examine a phenomenon of very rare events. It predicts the number of events that occur in a specific time period. Disadvantages: makes very restrictive assumption that variance=mean value.
- Negative binomial regression (usually for over-dispersed count outcome variables)
- Ridge regression (continuous target with multicollinearity effect among predictors) : A more robust version of linear regression, putting constraints on regression coefficients to make them much more natural, less subject to over-fitting. Disadvantages: Cannot perform variable selection.
- Lasso regression (continuous target):Similar to ridge regression, but automatically performs variable reduction. For correlated independent variables , Ridge regression has better prediction power than LASSO.

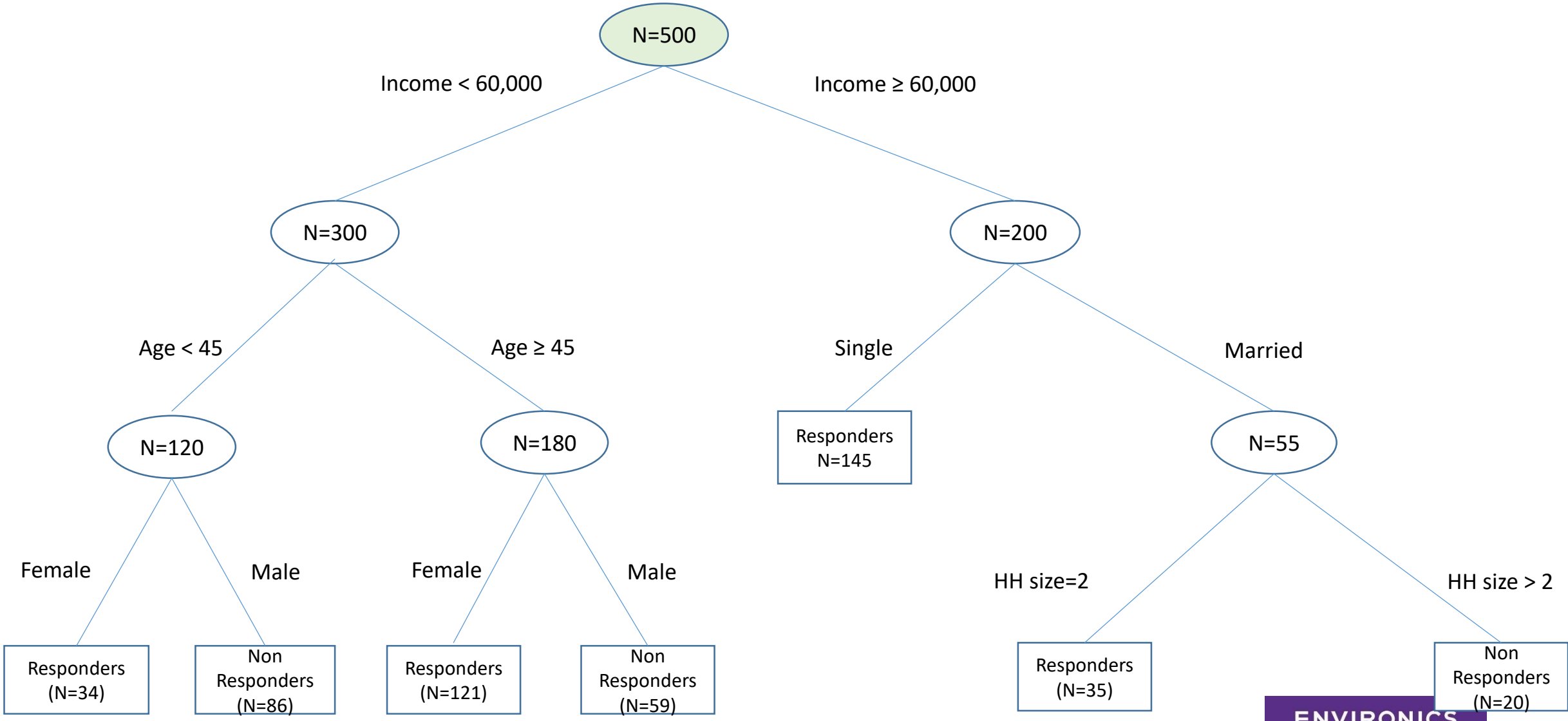
Multicollinearity

- **Multicollinearity** arises when two variables that measure the same thing or similar things (e.g., weight and BMI) are both included in a multiple regression model. **Multicollinearity** results in unstable coefficient estimates which makes it very difficult to assess the effect of independent variables on dependent variables.

Decision Tree

- A predictive modeling technique that uses a set of rules applied to calculate a target variable
- Can be used for classification (categorical variables) or regression (continuous variables) applications
- Rules are developed using software available in many statistics packages
- Different algorithms are used to determine the “best” split at a node in the tree

Decision Tree example



Decision Tree type

- CHAID (Chi-square Automatic Interaction Detector)
- CART (Classification And Regression Trees)
- C4.5 tree
- Random Forest: An ensemble classifier using many decision tree models

Decision Tree: advantages & disadvantages

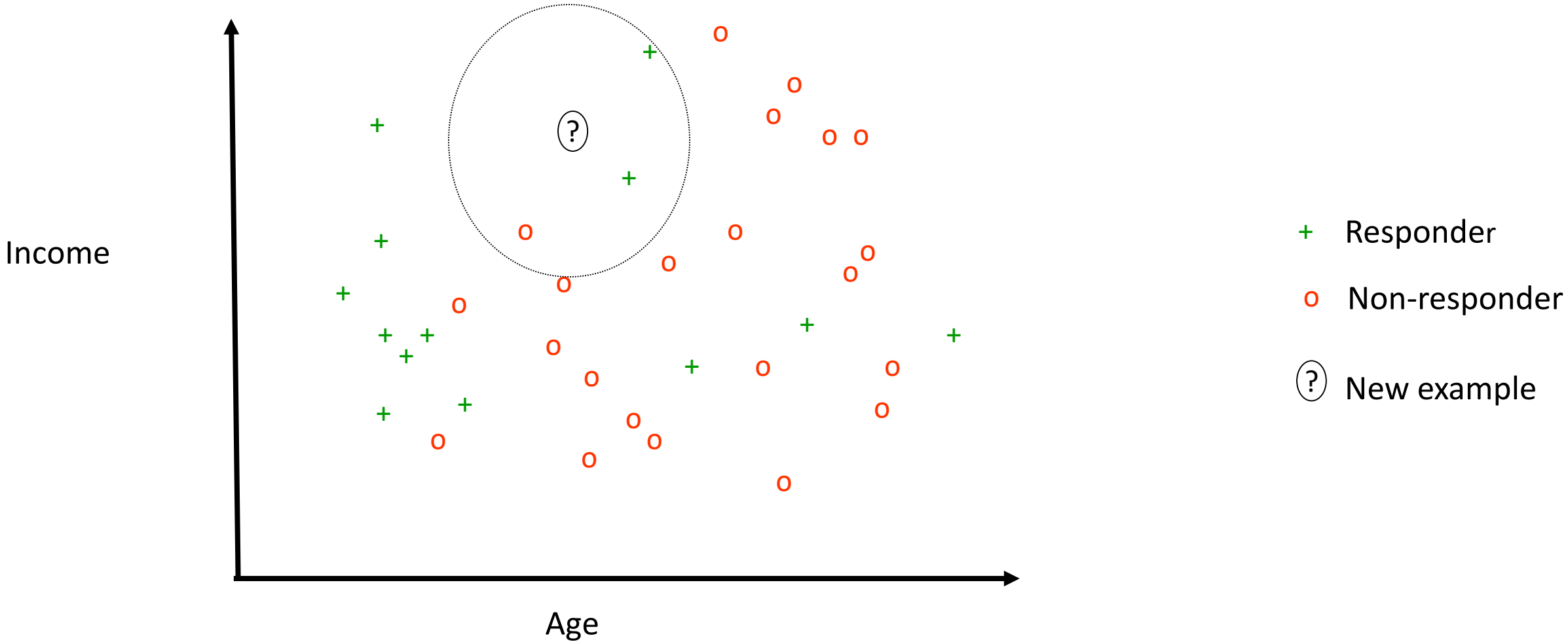
Advantages:

- Easy to interpret the decision rules
- Robust with regard to outliers in training data
- Classification is fast once rules are developed
- Variable selection & reduction is automatic
- Does not require the assumptions of statistical models

Disadvantages:

- Not possible to predict beyond the minimum and maximum limits of the response variable in the training data
- Tends to overfit training data which can give poor results when applied to the full data set or a new dataset
- As trees do not make any assumptions about the data structure, they usually require large samples

K-Nearest Neighbours (K=3)



New example's target value based on nearest neighbours' target values.

K-Nearest Neighbours: advantages & disadvantages

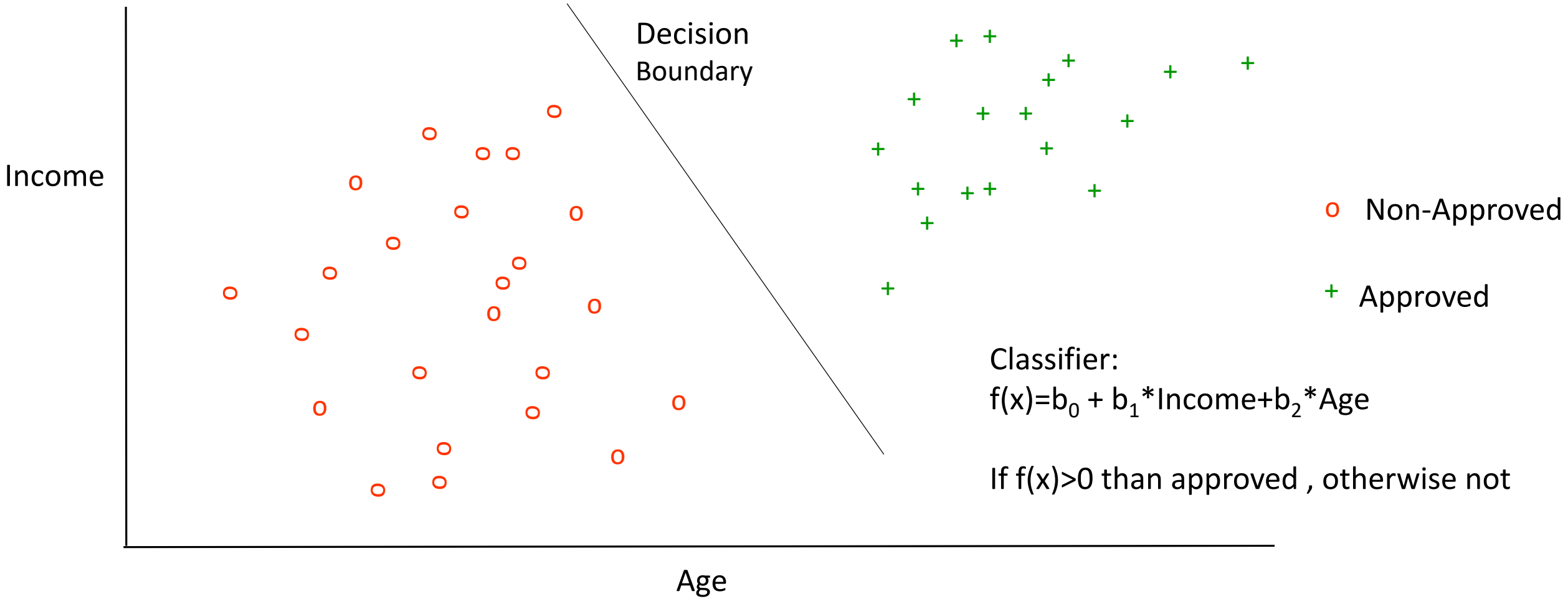
Advantages:

- Easy to understand
- Simple calculations

Disadvantages:

- Costly in terms of processing time

Linear Discriminant Analysis.



Linear Discriminant Analysis classifies objects in two or more groups according to *linear combination* of features

Linear Discriminant Analysis. advantages & disadvantages

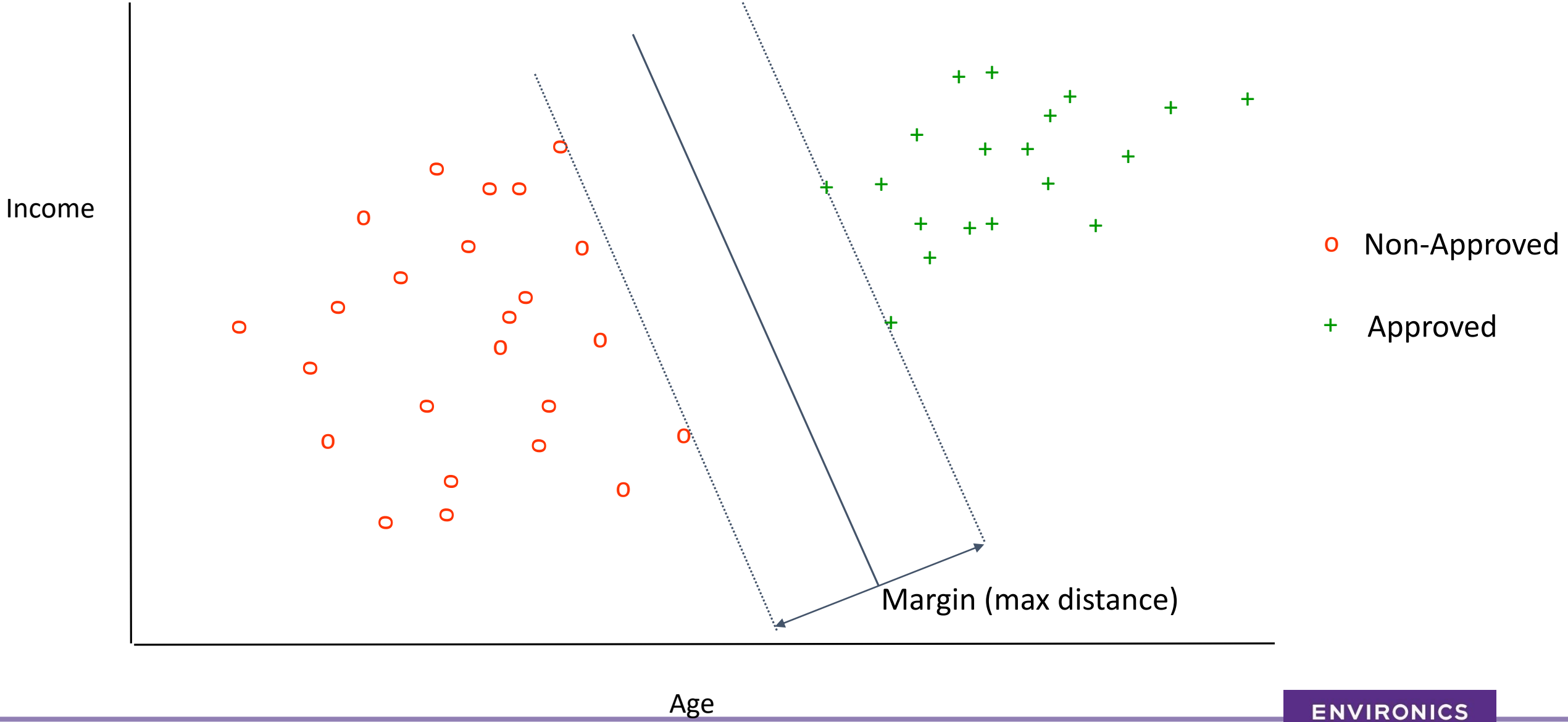
Advantages:

- Simple
- Fast

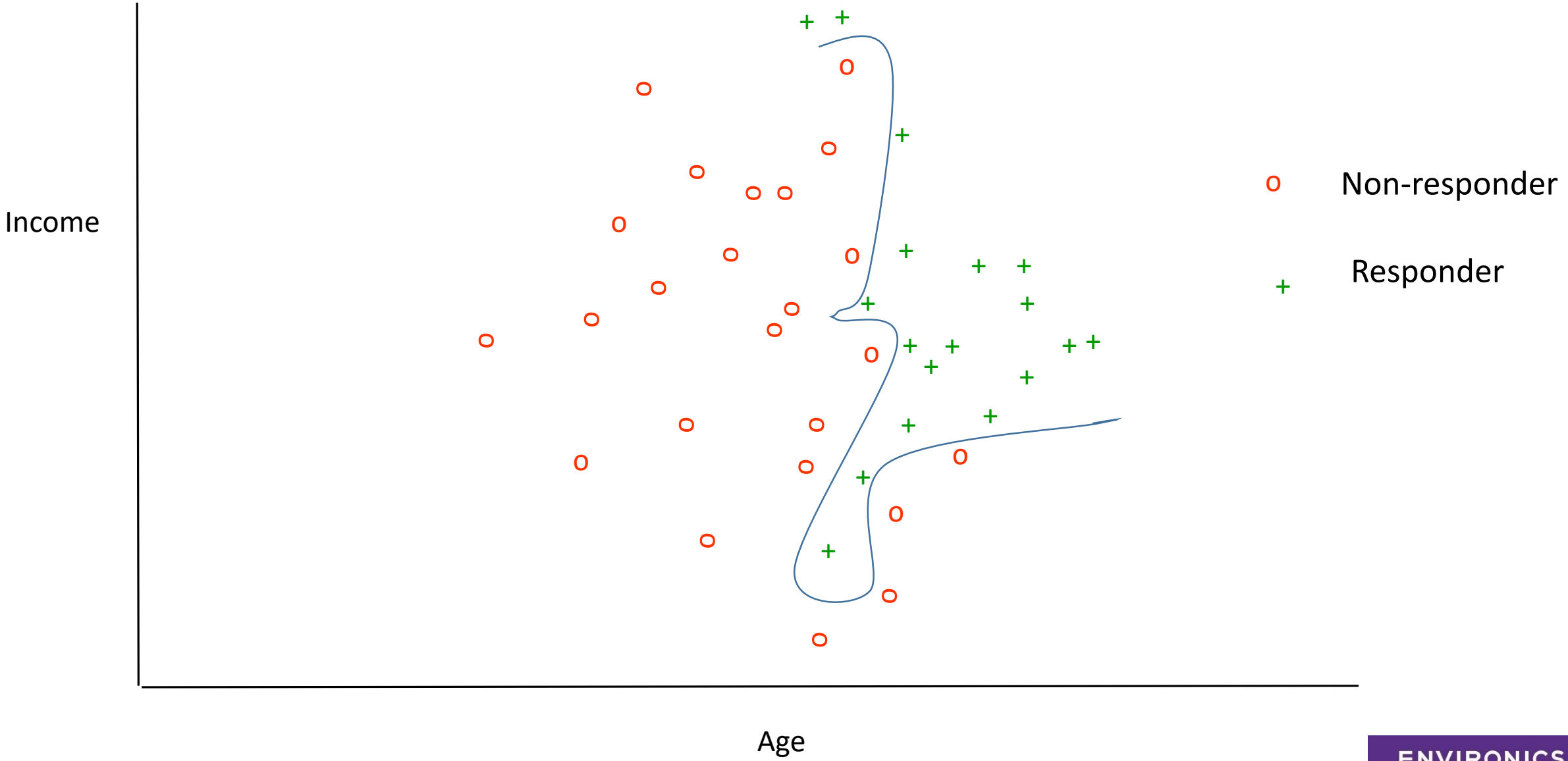
Disadvantages:

- Accuracy not as good as the newest predictive techniques

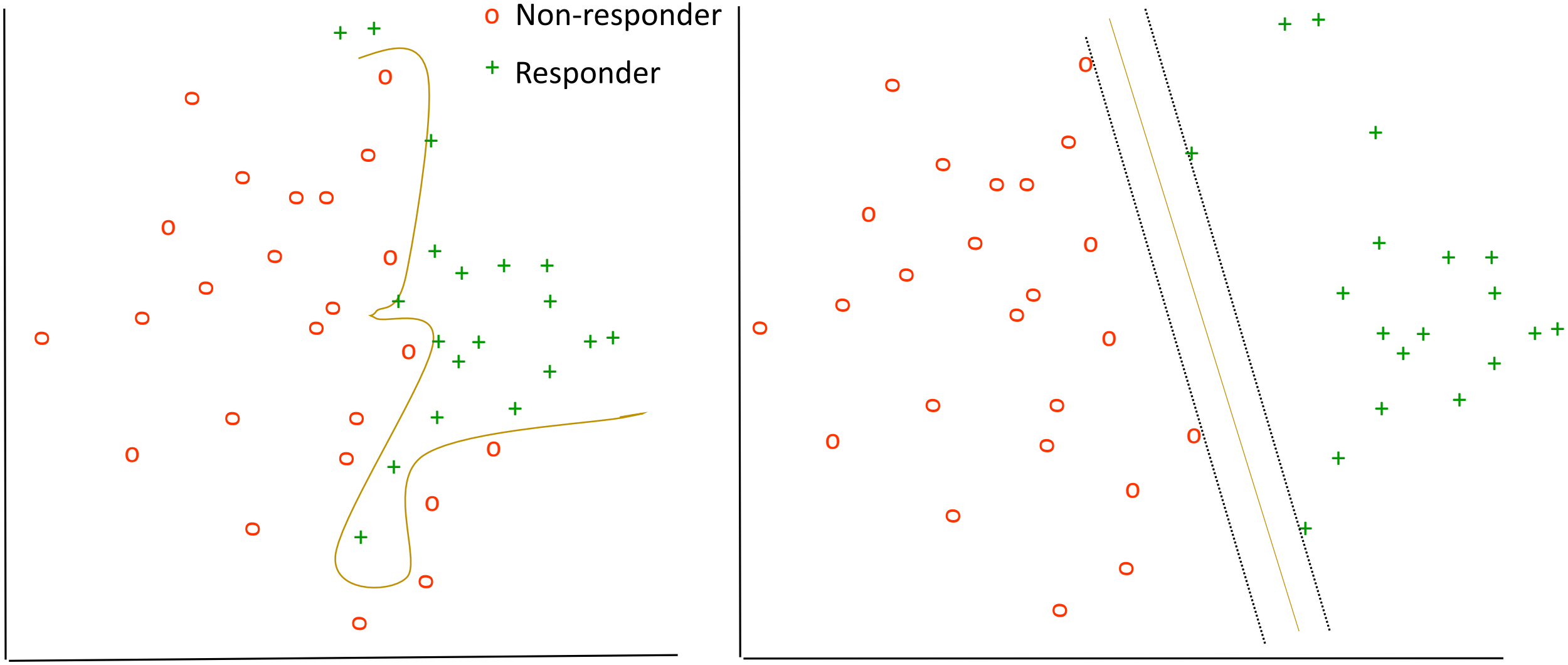
Support Vector Machines (SVM). Simple separation



Support Vector Machines. Non trivial example



Support Vector Machines. Non trivial example (2)



ENVIRONICS ANALYTICS

Support Vector Machines: advantages & disadvantages

Advantages:

- Less sensitive to overfitting. Small changes to data cannot greatly affect accuracy

Disadvantages:

- Difficult to identify variable importance (contribution)
- Computationally expensive

Naïve Bayes approach

Assume a client provided us with his customer database. For each customer we know:

- Age (Age Cohort)
- If customer lives in DA with high proportion of HHs with University degree or higher level (Education)
- If customer lives in DA with is high proportion of HHs with kids (Kids)
- If customer buys a product A (Buyer=Yes)

Our client wants to know if a new customer comes (with attributes $35 \leq Age < 45, Education = Univ\ degree\ and\ Kids = Yes$) will he buy product A?

Using Bayes' rule we can calculate:

$$P(Buyer = Yes | 35 \leq Age < 45, Education = Uplus, Kids = Yes) = \frac{P(35 \leq Age < 45, Education = Uplus, Kids = Yes | Buyer = Yes) * P(Buyer = Yes)}{P(35 \leq Age < 45, Education = Uplus, Kids = Yes)}$$

and

$$P(Buyer = No | 35 \leq Age < 45, Education = Uplus, Kids = Yes) = \frac{P(35 \leq Age < 45, Education = Uplus, Kids = Yes | Buyer = Yes) * P(Buyer = No)}{P(35 \leq Age < 45, Education = Uplus, Kids = Yes)}$$

Naïve Bayes approach (2)

Classifier:

$$f = \frac{P(35 \le Age < 45, Education = Uplus, Kids = Yes | Buyer = Yes) * P(Buyer = Yes)}{P(35 \le Age < 45, Education = Uplus, Kids = Yes | Buyer = No) * P(Buyer = Yes)}$$

If $f \geq 1$ then classifier predicts “Buyer”, otherwise – “Not Buyer”

Or using common assumption of **conditional independency** of predictors

$$f = \frac{P(35 \le Age < 45 | Buyer = Yes) * P(Education = Uplus | Buyer = Yes) * P(Kids = Yes | Buyer = Yes) * P(Buyer = Yes)}{P(35 \le Age < 45 | Buyer = No) * P(Education = Uplus | Buyer = No) * P(Kids = Yes | Buyer = Yes) * P(Buyer = No)}$$

Naïve Bayes Approach: advantages & disadvantages

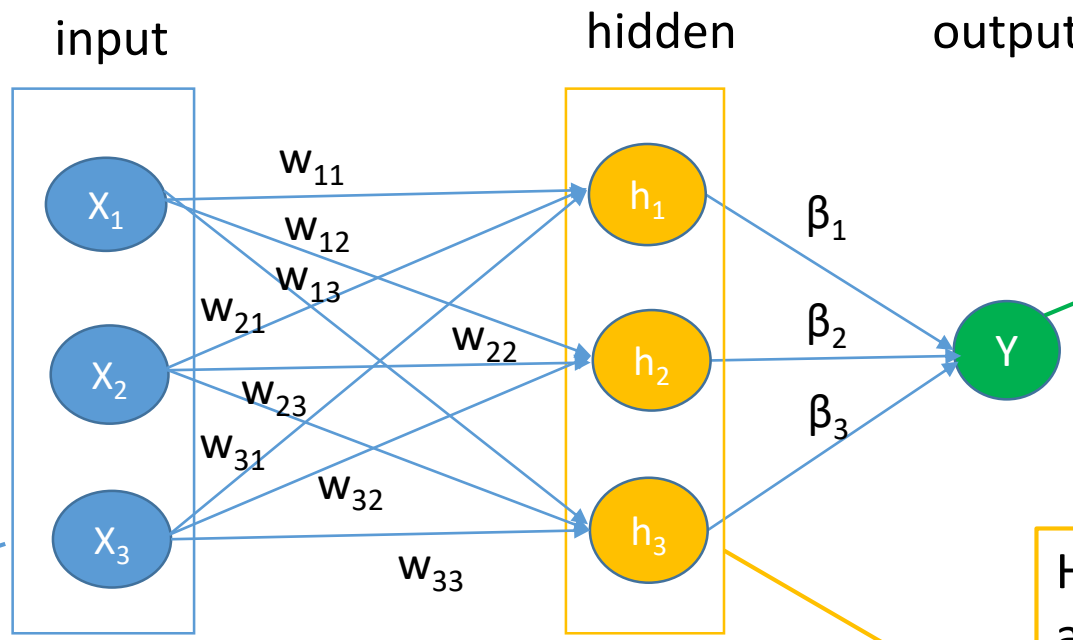
Advantages:

- Very simple to understand
- Very efficient in terms of storage space and computational time
- Simple approach to dealing with incremental problems – e.g. bidding where previous evidence was available in advance

Disadvantages:

- Very strong assumption is common – conditional independency
- Requires that continuous variables be put into classes (i.e. discretization)

Neural Network



Input Predictors

Output is linear (weighted) combination of hidden variables $\mathbf{Y=\beta_0 + \sum \beta_k h_k}$

Hidden (unobserved) variables are transformation of linear (weighted) combinations of some or all original predictors

$\mathbf{h_i=g(w_{0i} + \sum w_{ij} x_j)}$

Neural Network: advantages & disadvantages

Advantages:

- Ability to solve problems that do not have an algorithmic solution or the available solution is too complex to be found
- When an element of the neural network fails, it can continue without any problem by their parallel nature

Disadvantages:

- Black-box. Very difficult to explain which variables drive the explanatory process and how results were calculated.
- Prone to overfitting
- A big neural network requires large processing times

Conclusion

“All models are wrong but some are useful"

George Box

References

- Robert C. Blattberg, etc. (2008). Database Marketing
- Jacob Cohen, etc. (2003). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences
- Foster Provost & Tom Fawcett (2013). Data Science for Business
- Rand R. Wilcox (2010). Fundamentals of Modern Statistical Methods
- Alex Guazzelli (2012). Predicting the future, Part 2: Predictive modeling techniques
- David W. Hosmer, Stanley Lemeshow (2000) Applied Logistic Regression. Sec. Edit.
- Frank. E Harrell (2010) Regression Modeling Strategies
- Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996) A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology 49:1373-1379.

Questions?