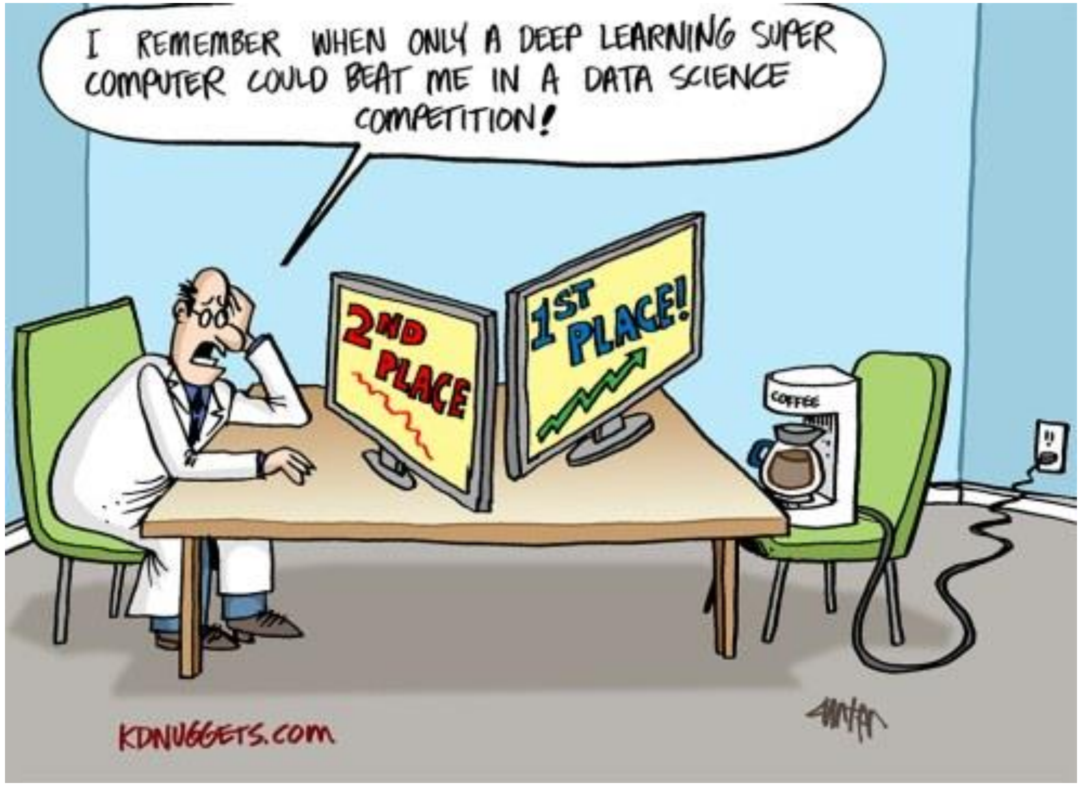
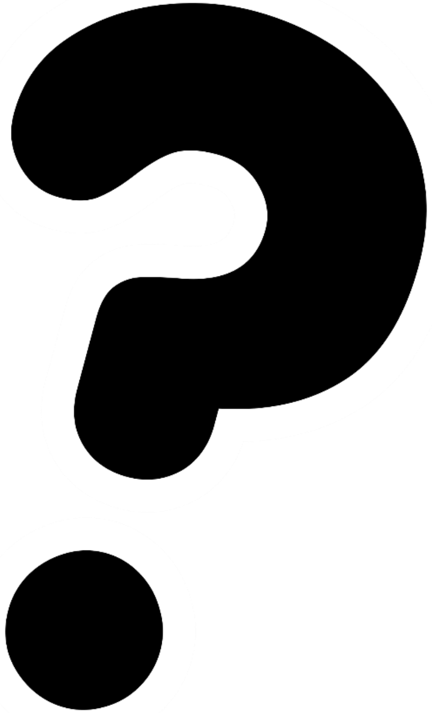
The Automation of the Feature Selection Process

Ronen Meiri & Jacob Zahavi

Automated Data Science





http://www.kdnuggets.com/2016/03/automated-data-science.html

Outline

- The feature selection problem
- Objective of study
- Review of feature selection methods used in our study:
  - Statistical methods: stepwise regression
  - Stochastic search methods: simulated annealing
  - Feature reduction methods: principal component analysis
- Databases and set up of study
- Results
- Conclusions

Characteristic of the analysis problem

- Many observations, in the order of millions or more
- Tens to hundreds (or thousands) of potential (Curse of Dimensionality)
- Many are redundant
- Other are “noisy”
- Some are irrelevant
- Rare events

Feature Selection

- The process of selecting an "optimal" subset of features
- Objectives
  - Distinguish informative and predictive features from coincidental features
  - Improve prediction accuracy (best fit)
  - Reduce bias (no over fitting)

Objective of Study

- Seeking the “best” feature selection in linear regression
- Testing:
  - Statistical methods (StepWise Regression)
  - Stochastic search methods (Simulated Annealing)
  - Feature reduction methods (Principal Component Analysis and Radial Basis Functions)
- Objective: Maximize the Goal function in the validation dataset.

DMWAY

StepWise Regression (SWR)

- Process introduces and eliminates predictors based on F-distribution
  - Significance level for entering variables: F-to-Enter
  - Significance level for removing variables: F-to-Remove
  - Where F-to-Enter < F-to-Remove
- Default values in most packages (per-comparison):
  - F-to-Enter = 5%
  - F-to-Remove = 10%
- Two alternative approaches have been devised to control the level of significance:
  - Bonferroni correction: $\alpha^* = \alpha/K$
  - False Discovery Rate (FDR): $\alpha_m^* = m\alpha/K$

DMWAY

Simulated Annealing (SA)

- Simulates the annealing process coming from condensed matter physics
- A solid is heated in a heat bath
- At sufficiently high temperature, the solid is liquefied
- By **slowly** cooling down the temperature, system attains a thermal equilibrium and system re-arrange in a lower-energy state
- As temperature goes to zero, system converges to its ground level state (minimum energy)
- **Converges to global optimum – or best set of features**

Feature Reduction Methods

- Also known as feature extraction methods
- Representing information hidden in original variables by fewer features
- Principal Component Analysis (PCA)
  - Variable reduction method
  - Seeks to remove multicollinearity by using a weighted sum of the original predictors to create new features which are uncorrelated (orthogonal)
- Radial Basis Functions (RBF)
  - Kernel function usually Gaussian distribution on the density of observations
  - A neural network (NN) type model

Datasets

Name	# Obs.	# Resp.	# Pred.	Response
Non Prof.	99,200	27,208	307	binary
Specialty	106,284	5,758	380	Continuous
Gift	101,284	9,707	104	Counter

Each was split randomly into a training and validation datasets

Evaluation Metrics

- Number of predictors in final model
- in training dataset
- in validation dataset
- ratio
- Gini coefficient – the area between the predictive model curve and the random (null) model. Sometimes multiplied by two to render a measure between 0-1.
- M-L: The maximum lift, where $Lift = \frac{\%\, responce\ in\ PM\ selection}{\%\, responce\ in\ all\ data}$

Set Up of Study

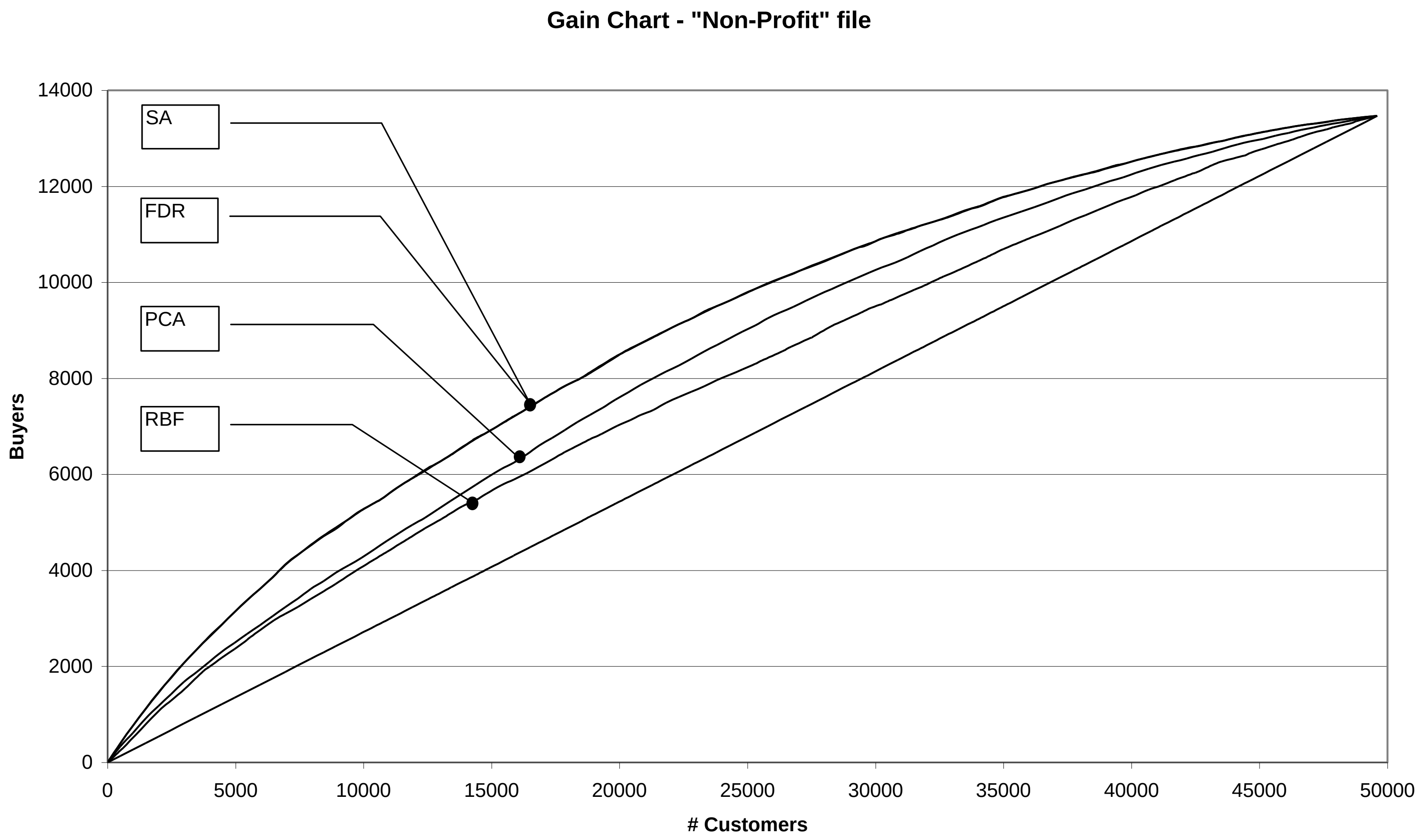
Approach: comparing the best performing model

In each class of models

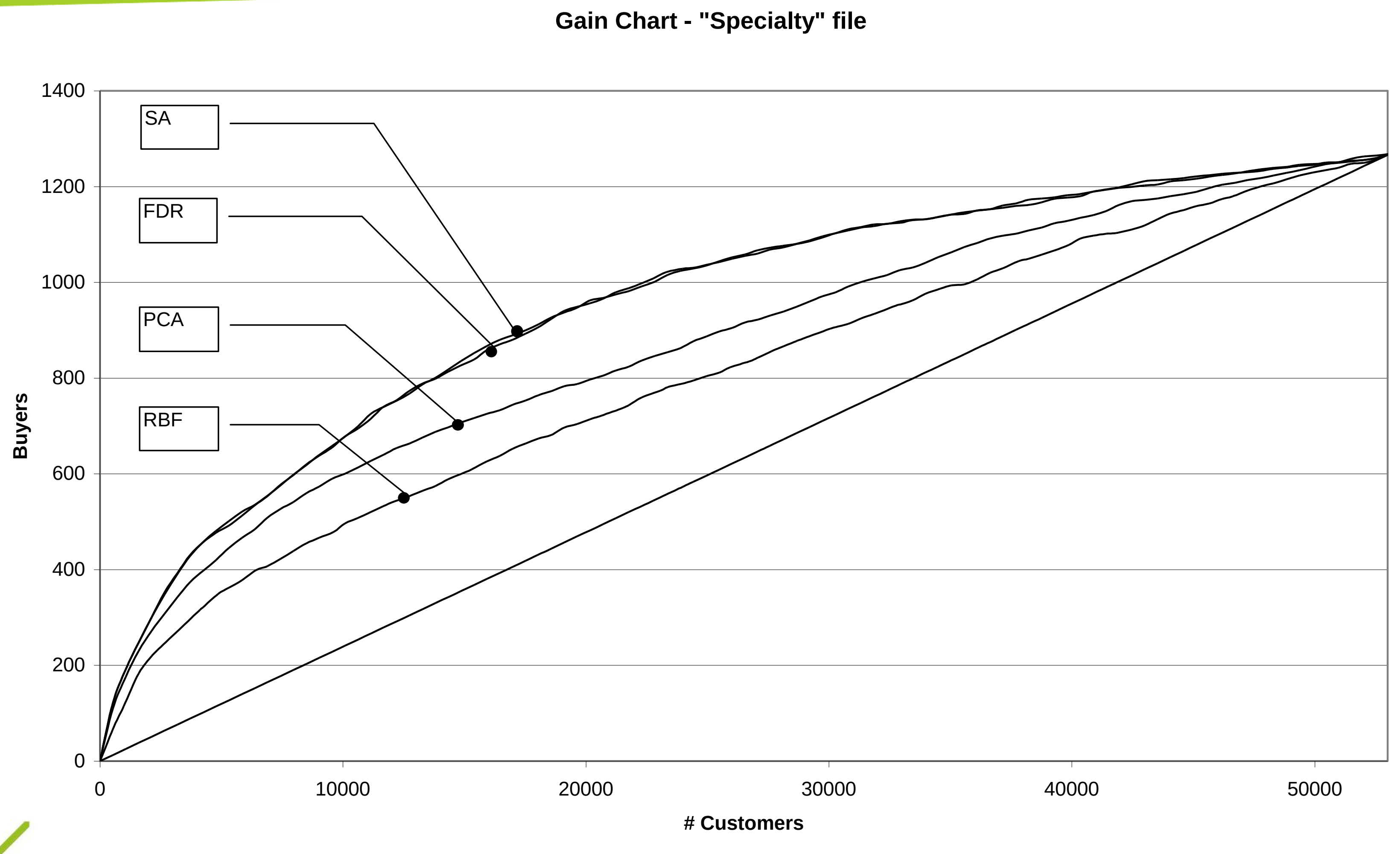
- SWR was calibrated based on FDR
- SA configuration was selected from among 60 parameter combinations (5 objective criteria, 4 cooling rates and 3 confidence intervals)
- For PCA, we sought the optimal number of PC's in the range 1-50
- For RBF, we sought the optimal number of radial bases in the range 2-51 (1-50 DF)

DMWAY

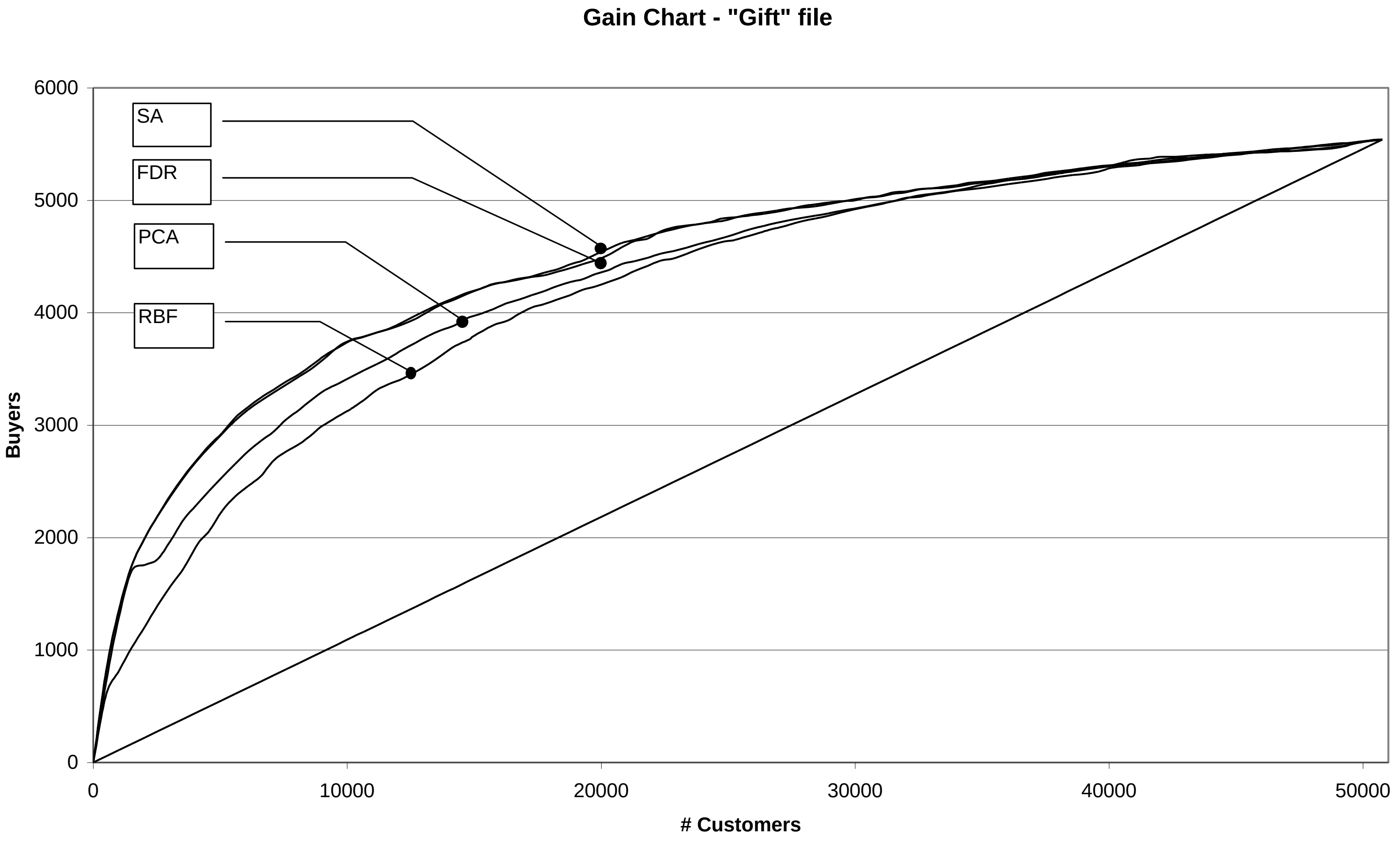
Gains Charts – Non Profit



Gains Charts- Specialty



Gains Charts - Gift



Conclusions – SWR and SA

- SWR yield comparable results to SA in all cases
- Both capture almost the same predictors in the final model

	Non-Profit	Specialty	Gift
Both	25	27	29
Only SWR	3	8	2
Only SA	-	1	4

- Solution is likely to be close to a global, if not the global, optimum
- Hypothesis – marketing data are “well behaved”
- Optimal solution to feature selection, if not unique, lie on the same plateau
- As a result, even the “greedy” SWR algorithm is capable of finding this solution
- Conclusion was further verified using simulated studies

Summary

The winner is FDR

- Feature Selection is probably dominated by few “Strong” predictors
- FDR able to optimize the balance between FP and FN
- We have data with “complex” structures, but probably most business data do not have it
- Further research is required to generalize the results of this study

DMWAY