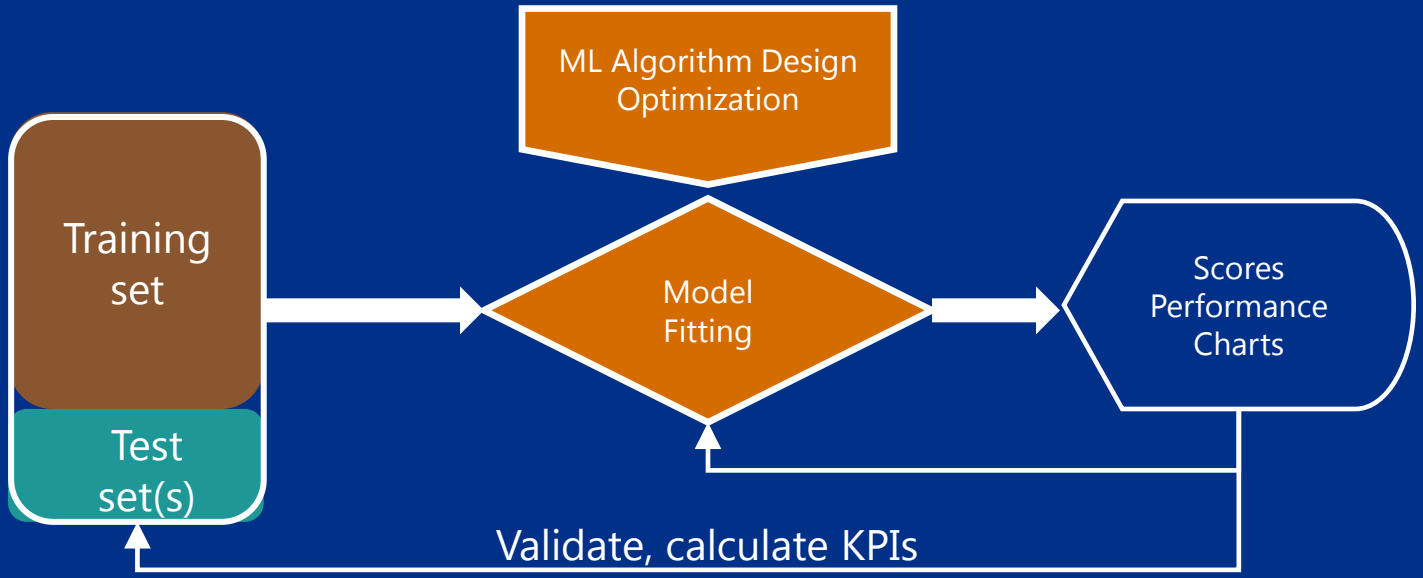
On the Practice of Predictive Modeling with Big Data: The Extra Steps that Make the Difference

Nachum Shacham
PayPal

XLDB

5/19/2015

Data Science From 10K ft

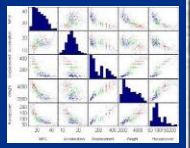


Predictive Modeling Meets Big Data: Bridging the Gap

Observational big-data

Deep, wide, distributed ,sparse, partially unknown, redundant, unreliable

Teradata



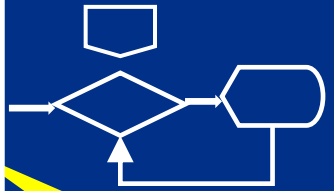
Data Sourcing

Data Exploration

Data Munging

Feature Engineering

Model(s) selection and tuning



Deploy

Validation

- Unified format (“data frame”)
- Well understood
- Statistically known and “well behaved”

Data Science with Big Data: Quantitative change → Process adjustment

Data Exploration: Knowing What's in the Data and Fitting the Pieces Together

Understaning Data Contents & meaning

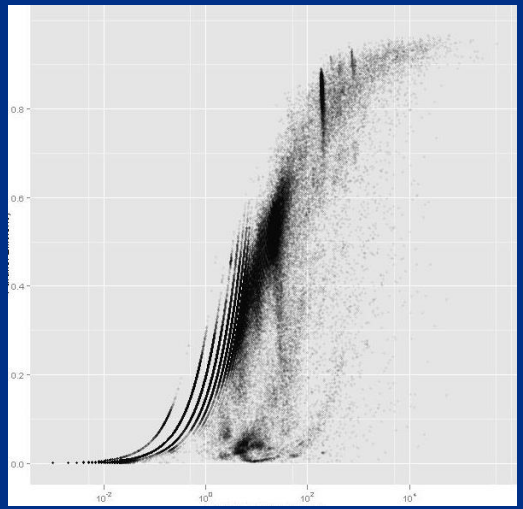
Distributions and outliers

Null values: codes, quantities

Language

Join keys

Units



Big data visualization

Correlations

Data Munging: Reshaping the Data → Decisions, Decisions

categorical variable

Dummy variables

Sub groups

Irrelevant/redundant values

Transform

Discard

Nullify

Missing values

Unify null symbols

Impute / discard

Unify values
Money, time

Remove outliers

Skewed distributions

Transform

Standardize

Join keys do not match

Reformat keys

Ignore rows

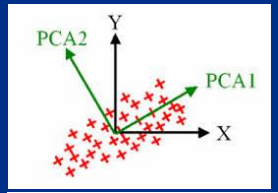
Feature Engineering

Transform raw data to better represent the problem to the model

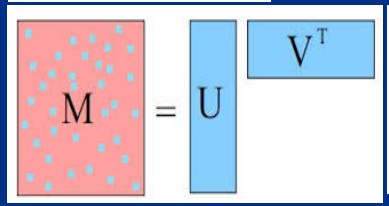
Manual Feature Manipulation

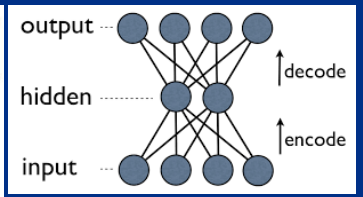
- Dummy variables
- Text integration
- Sub/super sampling (unbalanced sets)

Eye Color	X1	X2
Brown	1	0
Blue	0	1
Green	0	0

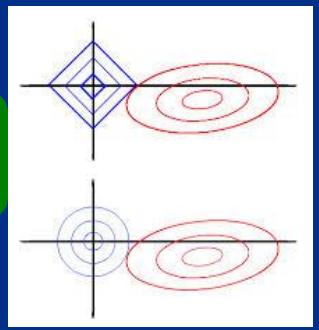


Feature learning algorithms





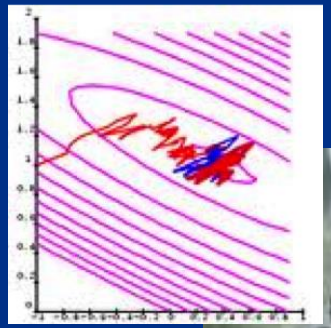
Integrated in model fitting



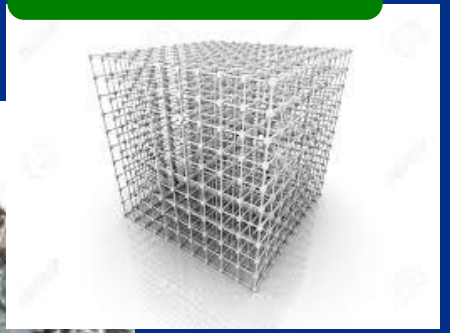
Better features → Simpler models, better results

Model Tuning

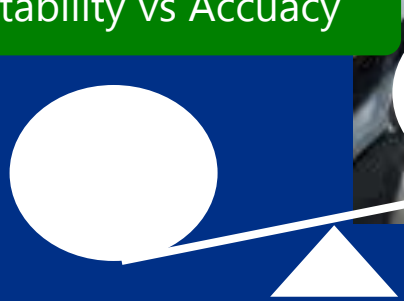
Learning Rate



Parameter Grid



Interpretability vs Accuacy



Ensemble: degree, type

Model Evaluation

- Who should judge the value of the results?
  - Evaluation criteria
- What are the cost/value of FP, TP, FN, TP?
  - FP: PR cost
  - FN: Financial loss
  - TP: Intended gain
- Continuous Model evaluation in production

Big Data Machine Learning in A Cloud

Rich library of ML algorithms

Auto detect feature format

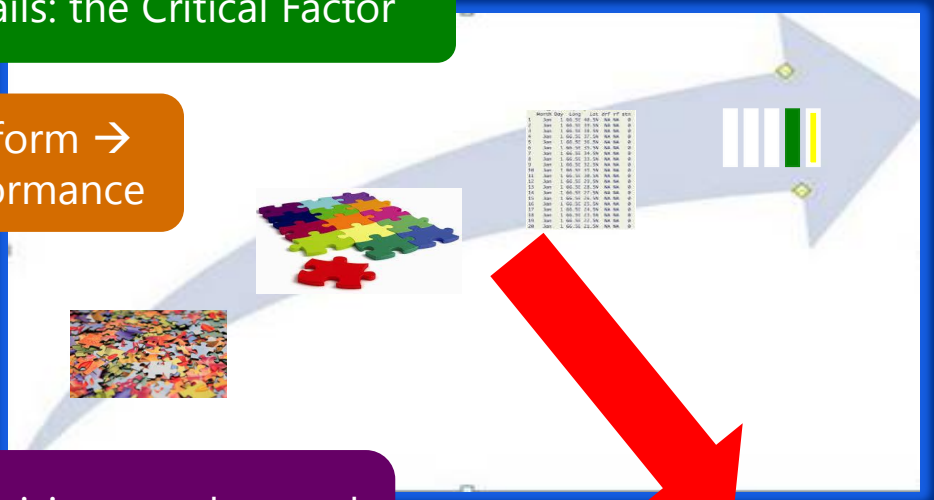
Interoperate with ordinary R: split work for complete munging

Runs through parameter grid

Summary: The “Extra” Steps Make The Difference

Details: the Critical Factor

Platform → Performance



Prediction

Decisions: understand the impact

Delayed

Incomplete

Irrelevant

THANK YOU