Tips and Tricks for Customer Segmentation

July 2015

Salford Systems

Outline

- What is customer segmentation?
- The **German credit problem**
- CART - Classification and Regression Trees
  - Costs/Priors
  - Size of segmentation
- Optimization
  - TreeNet - stochastic gradient boosting
  - Automation techniques
- Case studies
  - Insurance policy renewals
  - Health club memberships
  - Other applications
- Q & A

Customer Segmentation

- **Definition:** dividing a database into distinct groups of individuals who share common characteristics
- Need to identify key differentiators:
  - Demographic (age, race, gender)
  - Geographic (home location, work location)
  - Psychographic (social class, personality)
  - Behavioral (spending, price sensitivity)
- Benefits:
  - Apply target-specific marketing strategies
  - Prevent exhausting resources to target every customer/group
  - Generate maximum profit from each customer/group
  - Adjust marketing strategies over time when customer behavior changes
- Two approaches:
  - Pre-segmentation (most common): simply dividing the database
  - Post-segmentation: determining how a database was divided

The German Credit Problem

- Database on UCI Machine Learning Repository
  - https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)
- Predict if a customer is a "good" (1) or "bad" (2) credit risk
- The dataset contains 1,000 records, 20 characteristics/predictors
- Key tasks:
  - Define distinct segments of customers based on their attributes
    - Extract "rules" for each group
  - Decrease false "good" customers rate
    - It is worse to classify a "bad" customer as "good", then to classify a "good" customer as "bad"
  - Optimize our segmentation with advanced data mining techniques

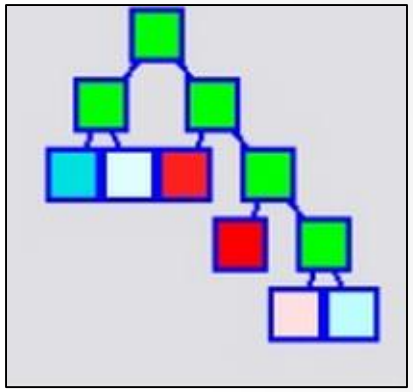
Sample Data

Residence	Property	Age	Other installment	Housing	Credits	People	Telephone	Foreign	Type
2	other	28	none	own	2	1	none	yes	2
1	other	25	none	rent	1	1	none	yes	2
4	society saving agreement	24	none	rent	1	1	none	yes	2
1	other	22	none	own	1	1	yes	yes	1
4	other	60	none	own	2	1	none	yes	2
4	other	28	none	rent	1	1	none	yes	1
2	other	32	none	own	1	1	none	yes	2
4	society saving agreement	53	none	own	2	1	none	yes	1
3	other	25	bank	own	3	1	none	yes	1
2	np property	44	none	for free	1	1	yes	yes	2
2	other	31	none	own	1	2	yes	yes	1
4	other	48	none	own	3	1	yes	yes	1

Other variables include: Credit history, credit amount, savings account
Note: mix of categorical and numeric variables

Tip 1: Use CART

- **C**lassification **a**nd **R**egression **T**rees
  - Separates relevant from irrelevant predictors
  - Yields simply, easy to understand results
  - Doesn't require variable transformations
  - Impervious to outliers and missing values

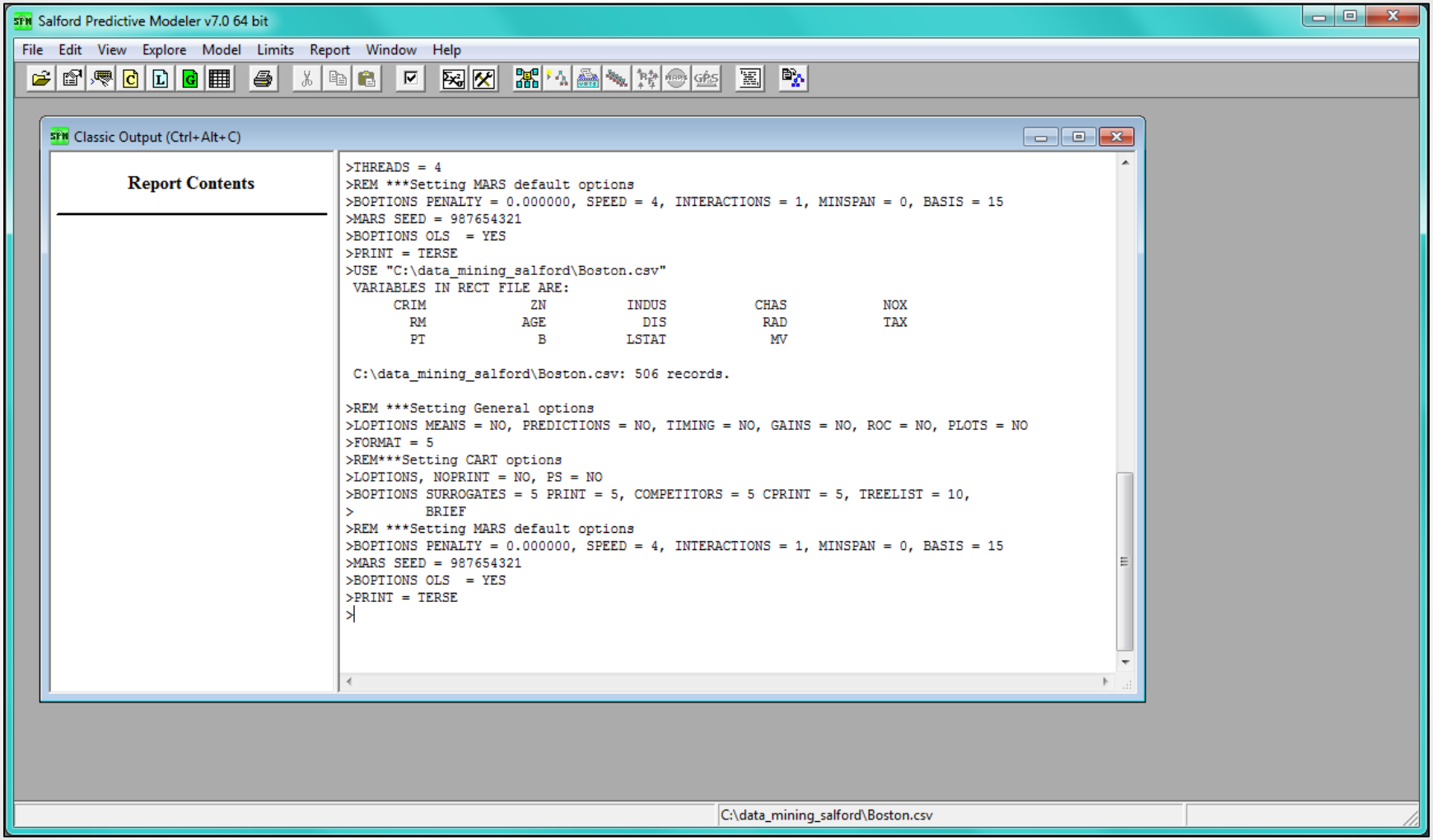


- Fastest, most versatile predictive modeling algorithm available to analysts
- Provides the foundation to modern data mining techniques such as bagging and boosting
- Data partitioning in this fashion creates distinct groups of records without needing to use any clustering methods

Tip 1: Use CART

- Other commonly used methods are
  - Neural Networks
  - K Nearest Neighbors
  - Logistic Regression
  - Discriminant Analysis
- A paper comparing classification methods on this dataset described CART as
  - Easiest to use
  - Having the most sophisticated and automatic pruning method
  - Producing the smallest trees
  - Best in terms of accuracy
- However, CART does have disadvantages
  - Sharp decision boundaries
  - Evolves around strongest effects
  - Difficulty capturing global linear patterns

Building a CART Model



Tip 2: Determine Priors and Costs

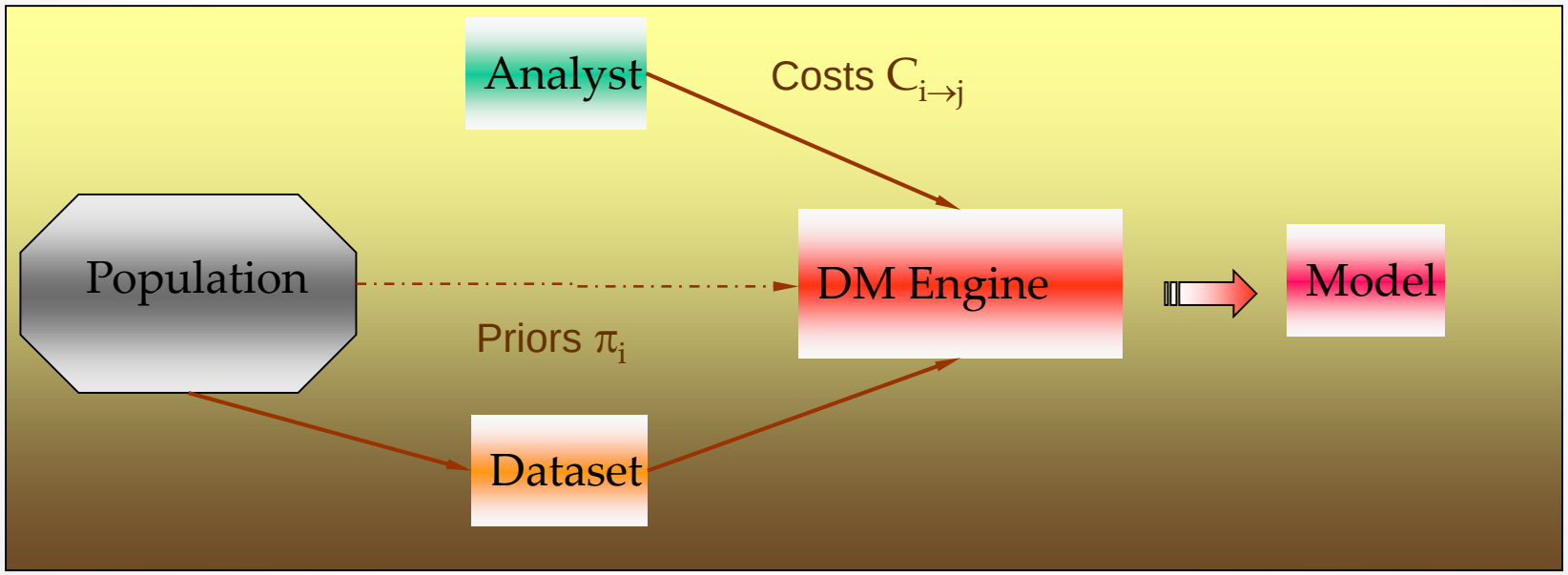
- In the German credit problem, identifying customers of one class ("bad" - 2) is more important to us than identifying a good credit risk
- When one class is of more interest (typically the case), you should adjust CART control parameters to reflect this
- For example, the owners of the database tell us that classifying a "bad" customer as "good" is ***5*** times worse than the other way around
  - This is a "cost", implemented in a cost matrix:

	Predicted = 1	Predicted = 2
Actual = 1	0	1
Actual = 2	5	0

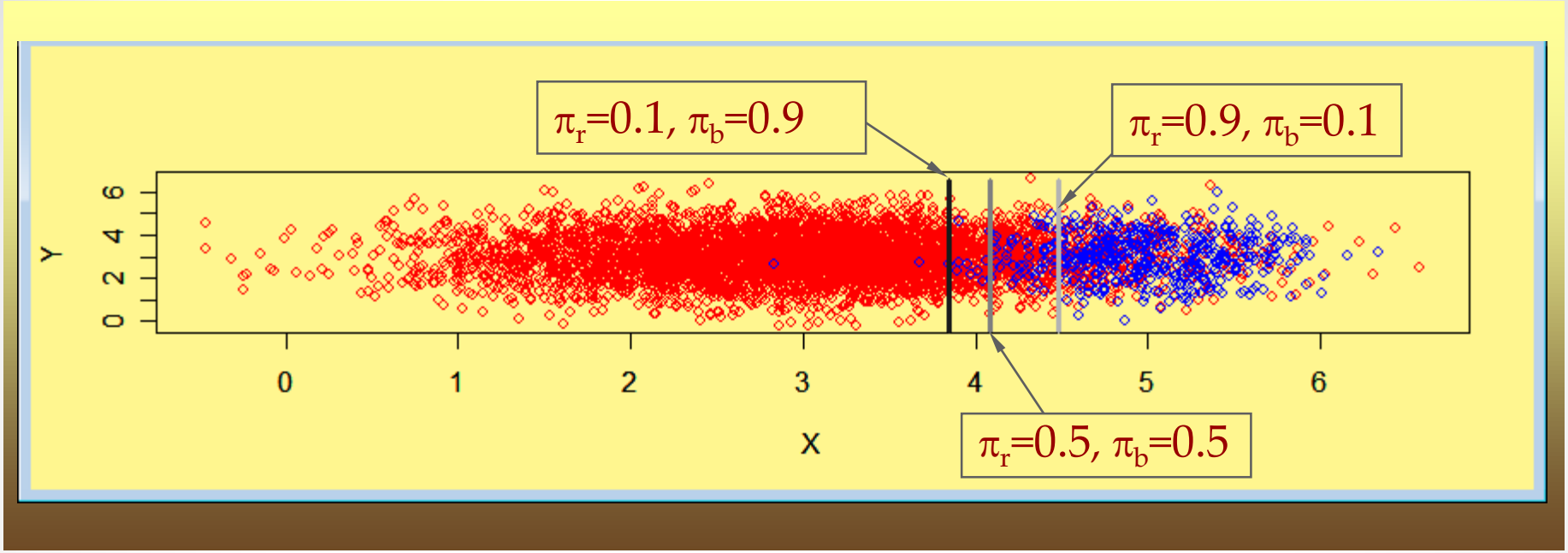
  - Priors (also called within-node probabilities) work in conjunction with costs

Fundamental Controls

- PRIORS and COSTS controls will make profound implications for all stages of tree development and evaluation
- These controls lie at the core of successful model building techniques



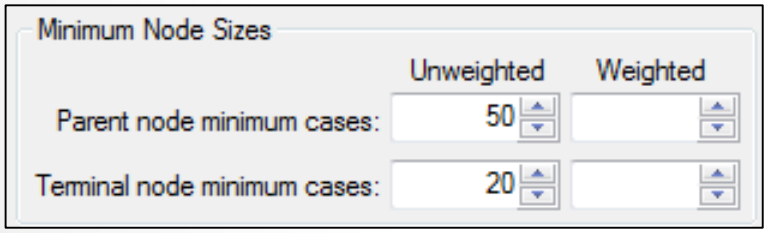
How Priors and Costs Work?



- We have a mixture of two overlapping classes
- The vertical lines show root node splits for different sets of priors (the left child is classified as red, the right child is classified as blue)
- Varying priors provides effective control over the tradeoff between class purity and class accuracy
- Cost is another mechanism to control over the tradeoff by weighting false positive and false negative rate

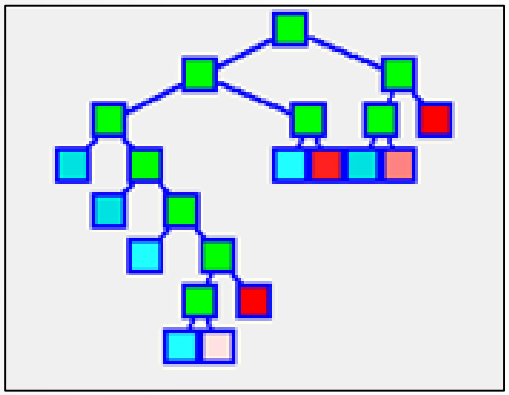
Tip 3: Decide on Segment Size

- In our first CART model, some of the terminal nodes (segments) had as few as 11 records
- You, as the analyst, have control over the size of these segments
- Imposing limits on both parent and terminal nodes will control the size of your tree



- A node cannot be split any further unless it contains at least 50 records
- A terminal node cannot be created from a split unless it holds at least 20 records

Customer Segments



Terminal Node 1		
Class	Cases	%
1	143	96.0
2	6	4.0
N = 149		

Terminal Node 8		
Class	Cases	%
1	22	55.0
2	18	45.0
N = 40		

Terminal Node 11		
Class	Cases	%
1	236	51.6
2	221	48.4
N = 457		

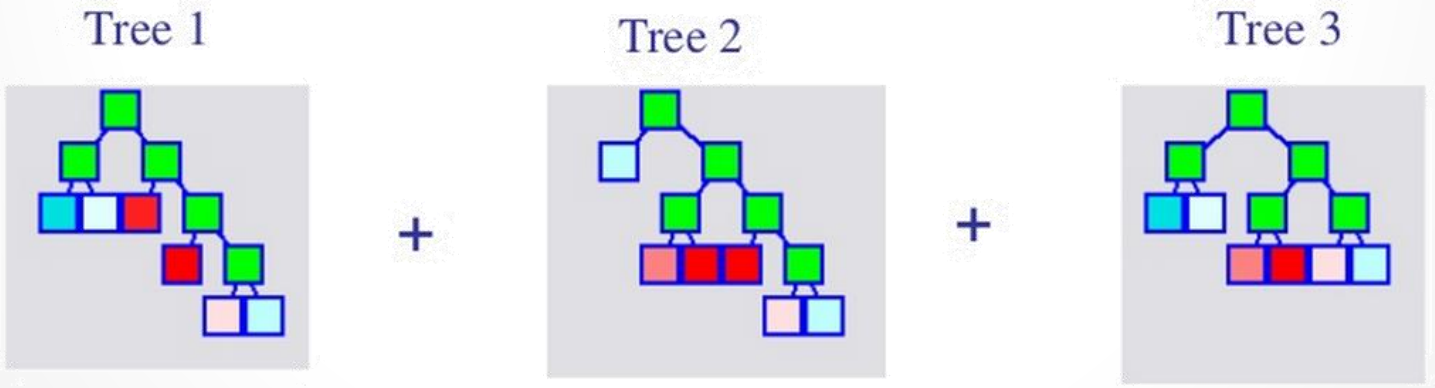
Actual Class	Total Class	Percent Correct	1 N = 411	2 N = 589
1	700.00	52.43%	367.00	333.00
2	300.00	85.33%	44.00	256.00
Total:	1,000.00			
Average:		68.88%		
Overall % Correct:		62.30%		
Specificity		52.43%		
Sensitivity/Recall		85.33%		
Precision		43.46%		
F1 statistic		57.59%		

- CART has identified an 11-node tree with desired low false "good" rate
  - Segment 1 (96% good): customers with credit history, no other installments, and either no checking account or a checking account with at least 200 DM
  - Segment 8 (45% bad): customers who have either no checking or a checking account with at least 200 DM, BUT have other installments and are using the credit line for business, education, or a car
  - Segment 11 (48% bad): customers who have less than 200 DM in their checking account and have had credit history for at least a year

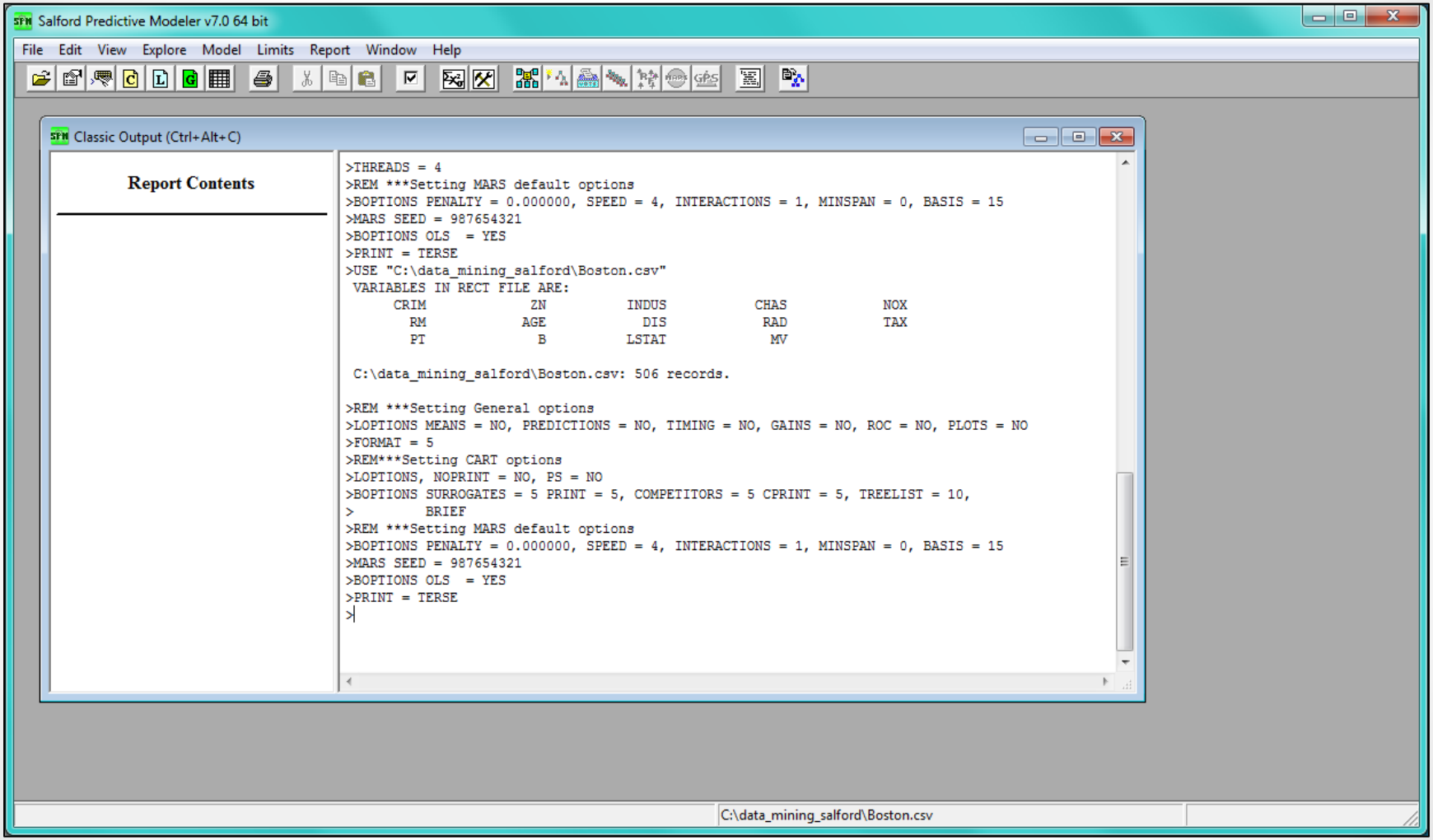
Tip 4: Use TreeNet to Optimize

- TreeNet aka Stochastic Gradient Boosting
- Small CART trees built in an error-correcting sequence

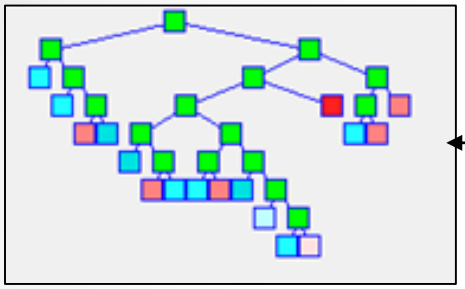
1. Begin with small tree as initial model
2. Compute residuals from this model for all records
3. Grow a second small tree to predict these residuals
4. And so on…



Building a TreeNet Model



Model Results Summary

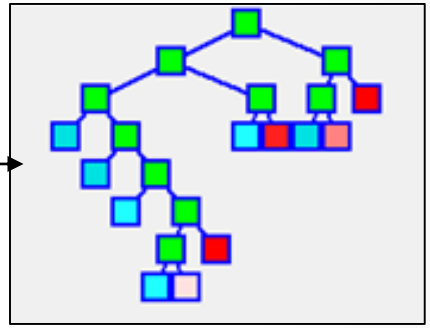


Actual Class	Total Class	Percent Correct	1 N = 571	2 N = 429
1	700.00	69.71%	488.00	212.00
2	300.00	72.33%	83.00	217.00
Total:	1,000.00			
Average:		71.02%		
Overall % Correct:		70.50%		

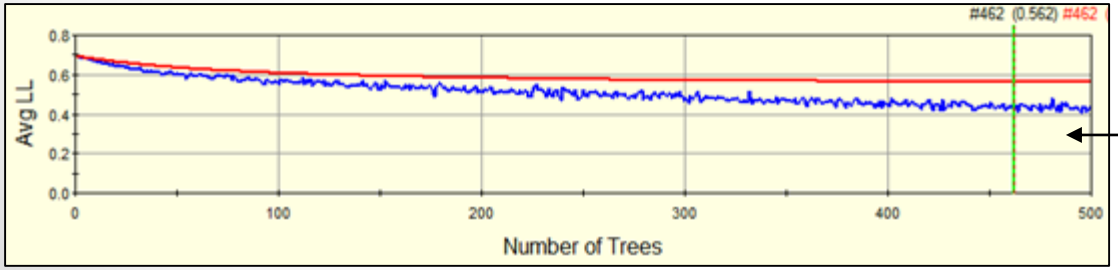
1st CART model - no parameter adjustments

2nd CART model - priors, costs, node limits

Actual Class	Total Class	Percent Correct	1 N = 411	2 N = 589
1	700.00	52.43%	367.00	333.00
2	300.00	85.33%	44.00	256.00
Total:	1,000.00			
Average:		68.88%		
Overall % Correct:		62.30%		



TreeNet model

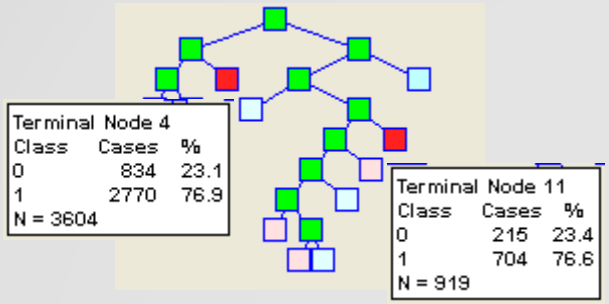


Actual Class	Total Class	Percent Correct	1 N = 354	2 N = 646
1	700.00	45.14%	316.00	384.00
2	300.00	87.33%	38.00	262.00
Total:	1,000.00			
Average:		66.24%		
Overall % Correct:		57.80%		

Case Study 1 - Customer Retention

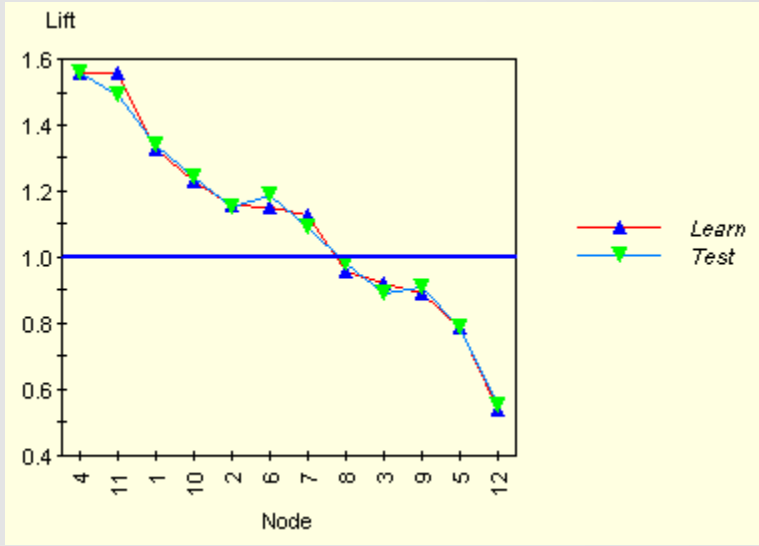
- Have a dataset containing information about last year renewals of insurance policies for the existing customers
  - About 100,000 observations randomly split into 50% learn and 50% test samples
  - Overall renewal rate set at about 50%
- Want to build a segmentation model to identify segments likely to renew
- Key predictors available
  - GENDER – customer's gender
  - AREA – customer's area
  - POLICY_TYPE – type of policy
  - POLICY_AGE – years with the customer
  - RESTRICTION – policy restrictions
  - AGE – customer age
  - POLICYCHANGE – recent change in premium
  - MARKETINTENSITY – how competitive the current market is

CART Model



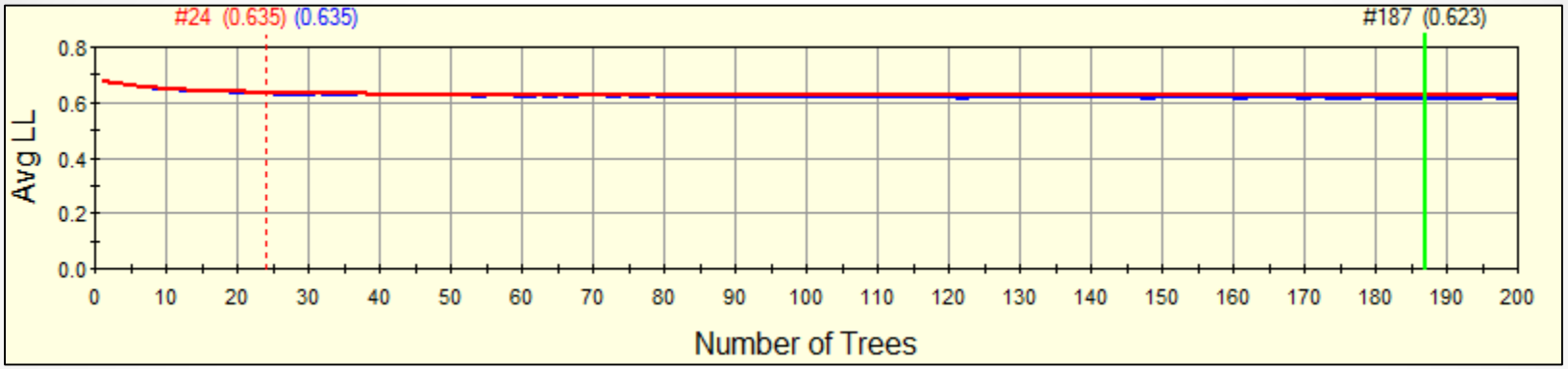
Test Sample Prediction Success Table

Actual Class	Total Cases	Percent Correct	Predicted Class 0 N=25220	Predicted Class 1 N=23397
0	24,640	65.64	16,173	8,467
1	23,977	62.27	9,047	14,930
Total:	**48,617.00**			
	Average:	63.95		
	Overall % Correct:	63.98		



- CART has identified a 12-node tree with good agreement on renewal rates between the learn and the test partitions
  - Segment 4 (77% renewal): customers with at least 8-year history are likely to renew when market intensity is low
  - Segment 11 (77% renewal): when market intensity is high only the most loyal customers (11 or more years history) are likely to renew provided that there were no significant premium increase
  - Segment 12 (27% renewal): when there is a significant premium increase we are likely to lose our customers in the intense market conditions

TreeNet Model



- TreeNet increased 1.19% in accuracy comparing to CART
- Learning rate and number of trees can be adjusted in setting to boost model performance
- Class Weights can also be played around with to boost model performance
- Partial dependency plots examine underlying relationships between predictors and target variable

Actual Class	Total Class	Percent Correct	0 N = 23499	1 N = 25111
0	24,681.00	63.24%	15,609.00	9,072.00
1	23,929.00	67.03%	7,890.00	16,039.00
Total:	48,610.00			
Average:		65.14%		
Overall % Correct:		65.11%		

Case Study 2 - Gym Example

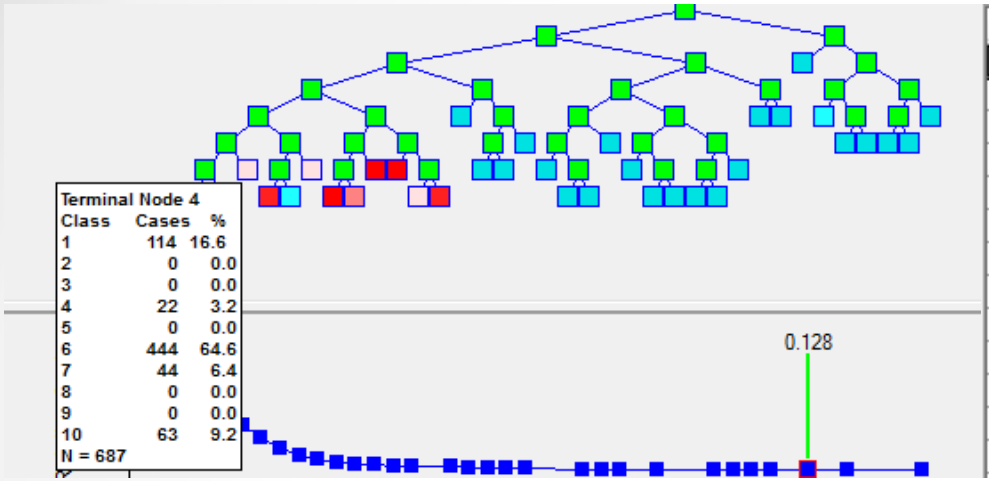
- The original data comes from a financial institution but was disguised as a health club
- Problem: need to understand a market research clustering scheme
- Clusters were created externally using 18 variables and conventional clustering software
- Need to find simple rules to describe cluster membership
- CART tree provides a nice way to arrive at an intuitive story

Variable Dictionary

ID	Identification # of member
CLUSTER	Cluster assigned from clustering scheme (10 level categorical coded 1-10)
ANYRAQT	Racquet ball usage (binary indicator coded 0, 1)
ONRCT	Number of on-peak racquet ball uses
ANYPOOL	Pool usage (binary indicator coded 0, 1)
ONPOOL	Number of on-peak pool uses
PLRQTPCT	Percent of pool and racquet ball usage
TPLRCT	Total number of pool and racquet ball uses
ONAER	Number of on-peak aerobics classes attended
OFFAER	Number of off-peak aerobics classes attended
SAERDIF	Difference between number of on- and off-peak aerobics visits
TANNING	Number of visits to tanning salon
PERSTRN	Personal trainer (binary indicator coded 0, 1)
CLASSES	Number of classes taken
NSUPPS	Number of supplements/vitamins/frozen dinners purchased
SMALLBUS	Small business discount (binary indicator coded 0, 1)
OFFER	Terms of offer
IPAKPRIC	Index variable for package price
MONFEE	Monthly fee paid
FIT	Fitness score
NFAMMEN	Number of family members
HOME	Home ownership (binary indicator coded 0, 1)

predictors

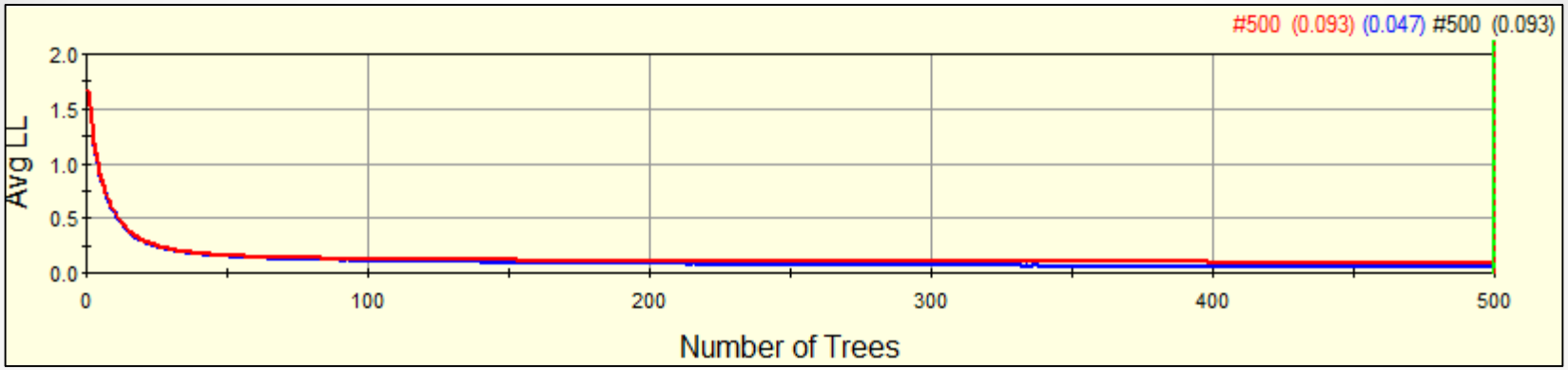
CART Model



Variable	Score
FIT	100.0000
ANYRAQT	94.9987
PLRQTPCT	93.0132
TPLRCT	92.7796
CLASSES	92.0764
ONRCT	89.8164
NSUPPS	73.8529
NFAMMEM	69.9809
ONAER	65.5616
SAERDIF	62.1419
OFFAER	52.5561
PERSTRN	52.2003
HOME	50.9585

Actual Class	Total Class	Percent Correct	1 N = 2583	2 N = 1293	3 N = 4078	4 N = 1436	5 N = 1597	6 N = 1868	7 N = 1573	8 N = 3156	9 N = 2191	10 N = 720
1	2,826.00	84.93%	2,400.00	0.00	4.00	21.00	22.00	245.00	36.00	15.00	5.00	78.00
2	1,430.00	88.53%	0.00	1,266.00	63.00	0.00	68.00	25.00	8.00	0.00	0.00	0.00
3	4,034.00	95.93%	0.00	13.00	3,870.00	0.00	0.00	0.00	46.00	105.00	0.00	0.00
4	1,818.00	69.86%	23.00	2.00	26.00	1,270.00	3.00	11.00	298.00	0.00	185.00	0.00
5	1,503.00	95.08%	0.00	9.00	0.00	0.00	1,429.00	41.00	0.00	24.00	0.00	0.00
6	1,786.00	83.82%	137.00	0.00	0.00	1.00	75.00	1,497.00	18.00	29.00	0.00	29.00
7	1,342.00	84.35%	7.00	1.00	11.00	74.00	0.00	19.00	1,132.00	0.00	92.00	6.00
8	3,082.00	96.72%	0.00	0.00	101.00	0.00	0.00	0.00	0.00	2,981.00	0.00	0.00
9	2,001.00	95.40%	1.00	2.00	3.00	70.00	0.00	0.00	16.00	0.00	1,909.00	0.00
10	673.00	90.19%	15.00	0.00	0.00	0.00	0.00	30.00	19.00	2.00	0.00	607.00
Total:	20,495.00											
Average:		88.48%										
Overall % Correct:		89.59%										

TreeNet Model



Actual Class	Total Class	Percent Correct	1 N = 2843	2 N = 1424	3 N = 4049	4 N = 1857	5 N = 1504	6 N = 1755	7 N = 1330	8 N = 3100	9 N = 1964	10 N = 669
1	2,826.00	96.36%	2,723.00	0.00	2.00	21.00	2.00	62.00	5.00	3.00	1.00	7.00
2	1,430.00	97.27%	0.00	1,391.00	30.00	1.00	8.00	0.00	0.00	0.00	0.00	0.00
3	4,034.00	97.92%	4.00	14.00	3,950.00	4.00	0.00	3.00	1.00	58.00	0.00	0.00
4	1,818.00	94.06%	24.00	2.00	8.00	1,710.00	0.00	0.00	42.00	0.00	32.00	0.00
5	1,503.00	97.54%	0.00	9.00	2.00	0.00	1,466.00	10.00	0.00	14.00	0.00	2.00
6	1,786.00	93.45%	70.00	6.00	0.00	6.00	20.00	1,669.00	2.00	9.00	0.00	4.00
7	1,342.00	94.19%	3.00	1.00	5.00	49.00	0.00	5.00	1,264.00	1.00	12.00	2.00
8	3,082.00	97.83%	7.00	0.00	50.00	0.00	8.00	2.00	0.00	3,015.00	0.00	0.00
9	2,001.00	95.90%	1.00	1.00	2.00	66.00	0.00	0.00	12.00	0.00	1,919.00	0.00
10	673.00	97.18%	11.00	0.00	0.00	0.00	0.00	4.00	4.00	0.00	0.00	654.00
Total:	20,495.00											
Average:		96.17%										
Overall % Correct:		96.42%										

TreeNet gives better model in terms of accuracy in most cases while CART gives a tree with segments where you can visualize groups/nodes

What We Learned

- CART is a powerful tool for customer segmentation
  - Supervised analysis (unsupervised is also supported)
  - Priors and Costs settings gives us the flexibility to adjust the sensitivity, specificity, false positive and false negative rate
  - Interaction detection
  - Automatically handles missing values
- TreeNet boosts accuracy
  - Produce models with better accuracy in most cases
  - Class Weights setting allows us to set corresponding class rate for better accuracy (similar to Priors)
  - Dependency plots reveal underlying relationships between target variable and predictors
- Advanced features
  - Hotspot detection for identifying richest nodes (after varying priors)
  - Predictor selection (shave variables one at a time)
  - RuleLearner for extracting important rule sets in an ensemble

Other Applications

- **Business**: Marketing strategies, Targeted Sales, Fraud Detection
- **Drug Discovery**: Better Profiling including Adverse Events and Healthcare Outcomes
- **Insurance Premium Optimization**: finding variables, often non-intuitive, providing clues into a prospect or customer's purchasing behavior and risk level
- **Environmental**: Decision making for Environmental Management, Population Dynamics, Habitat Suitability
- **Epidemiology**: Risk Analysis, Population Dynamics

References

- StatLog: Comparison of Classification Algorithms on Large Real-World Problems
  - R.D. King , C. Fengy , and A. Sutherlandz
  - http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.25.8985&rep=rep1&type=pdf
- StatLog German Credit Database
  - https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)