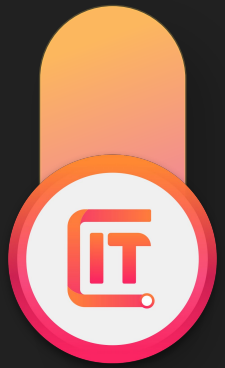
Data Analytics

Lecture 4

Modeling and Data Manipulation



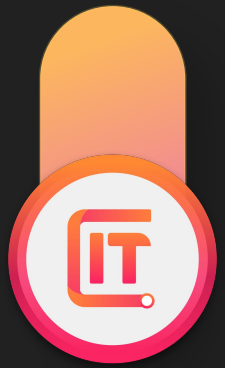
Progress

- Introduction, Methodology, Quiz to understand your SQL/ basic data handling proficiency
- What we are dealing with:
  - Evolving production data
    - Meta data
    - Big data
  - Ad-hoc, external data
    - lifecycle management
    - reference data
  - Tactics: (Agile, Github)
  - Classifications
    - Type of data (categorical, nominal, ordinal)
    - Data Universe quadrant
    - Capability quadrant
    - Killer tools Future tools (D3, GraphQL)
- Methodology
  - CRoss-Industry Process for Data Mining (CRISP-DM)
    - Other alternatives
  - Case: Aircraft incident
    - Objective:what we want to do?
    - Data Understanding:
    - Data Cleansing:
    - Data Enhancing:
    - Modeling/Analysis
    - Evaluation
    - Deployment

Lecture 3

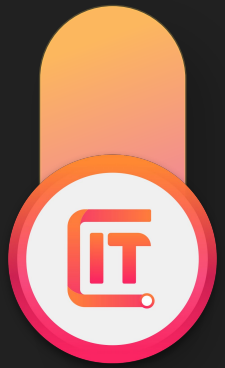
- Deliver Value to Business - communication
  - Goal
    - What's the right goal?
    - Project governance applies
  - Story telling
    - Case: multinational manufacturer
    - Case: Hanse Rosling
  - Visualisation
    - Principle
    - Examples
    - Pitfalls (pie chart, 3D, redundant elements)
  - Common pitfall
    - correlation not causation
    - Too good to be true predictor
  - What users are thinking
    - Trade off of modeling accuracy/complexity(can the rules be interpreted?)
    - Manage cost of deployment

Lecture 4 today



- Data modeling and manipulation
  - Machine Learning Model Quadrant
    - generative model/discriminative model
  - Feature:
  - Dataset Split: Training/Test/Validation
  - Hyperparameters
    - 80:20 rule(spend more time in feature engineering)
    - How to find them - # of clusters
    - Learning rate - NN
    - Pruning- single decision tree - Level of depth - random forest
  - Measuring Model accuracy and effectiveness
    - Regression: R2, MSE
    - Classification - Measuring ROC,Gain, precision/recall
    - Confusion matrix,Type 1 /2 Error
    - Ensemble Learning

Lecture 4/5



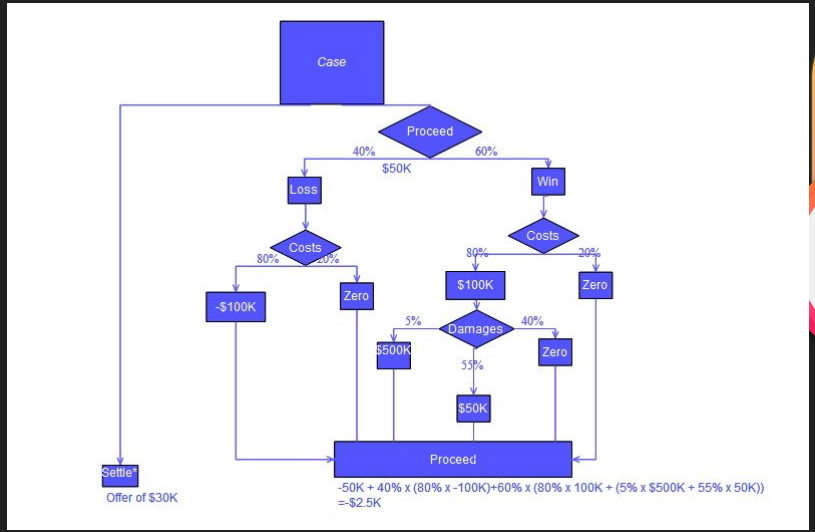
- Case 1: predict customer response
  - Feature selection
  - Supervised learning- which model to choose?
  - Model evaluation and setup A/B test
- Case 2: segmentation
  - Clustering
  - Interpret result to users
- Regression
- Time Series
  - Outlier/purification
  - Tool: FB
- Association Rules
- NN
- Text Mining
  - Tools
- Case3: Link Analysis on blockchain

- Stand Up

学远示例：
杭州房地产数据分析：
方法论很清晰，即使外行人也可以很好的理解它的条理， 层次递进。可能的问题是为什么要选用收入，生命周期 这些变量，ppt没写，只能推 测讲解的时候会口头提到

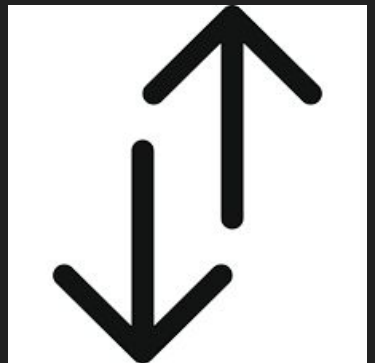
Advantage of GUI

- Data flows in different branches for different purpose;
- Intuitive
- Easy to combine

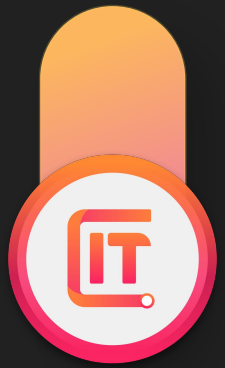


Python :

- Even notebook has only 1 flow: Up, Down



Keep AGILE



1 lecture = 1 sprint

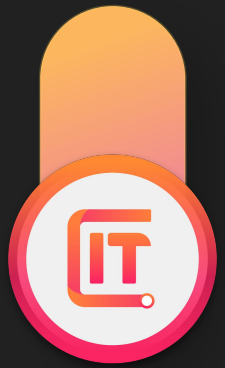
Sprint 2, homework requirement not clear

=> delay in sprint review

=> start/end of sprint 3 20 minutes late

(we are not agile YET, when sprint time ends, sprint ends)

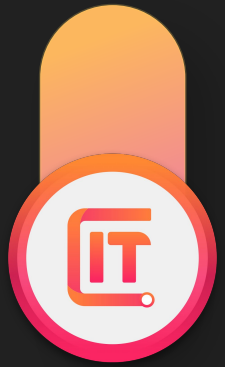
Big data - about correlation, can be misleading



Stat: those have 10 min breaks between work, after 5 years 40% more likely to have cancer

https://www.kaggle.com/kanncaa1/why-gun-violance-increase-in-texas

论点-论据 - 还是从现象中找结论？



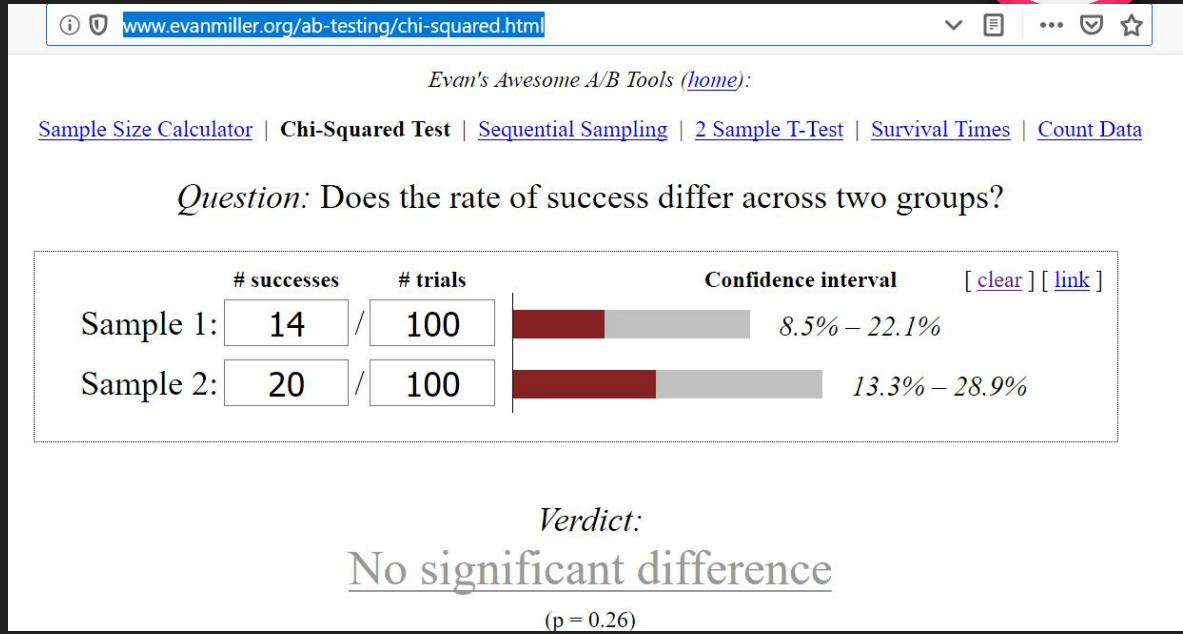
数据科学中的“科学”

达尔文，假说
毛泽东-农村
林彪-应用大数据，Hannibal - campaign

数据只能证伪，以及提出假说，不能证实

So you need

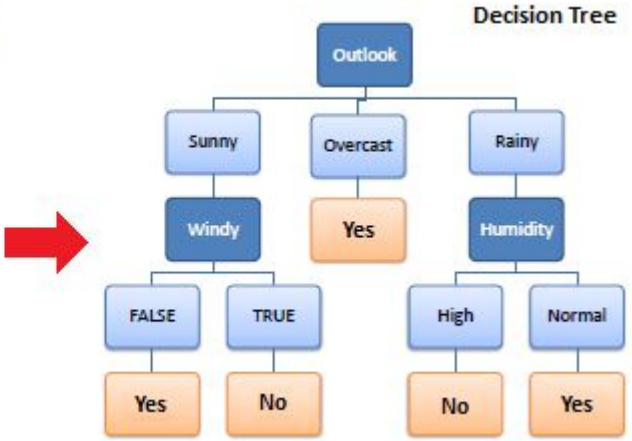
可检验性（testable）
1. Can be tested
2. producible



Decision tree

Predictors | Target

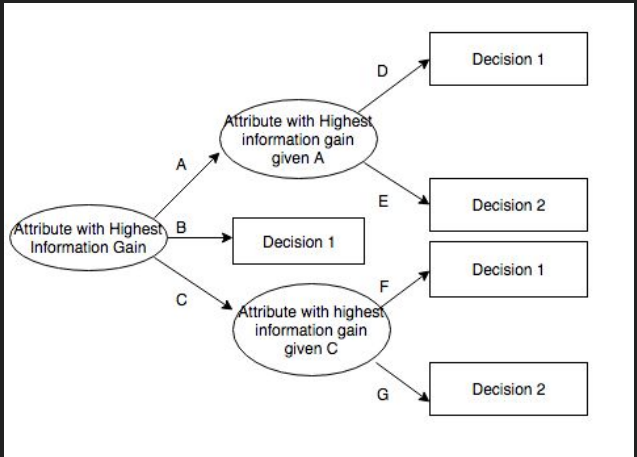
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



Type of input variable:
classification (离散discrete) - information gain (ID3), or convert to discrete (binary - C4.5
regression(连续continous)

Decision tree

ID3算法（Iterative Dichotomiser 3 ）
Smaller Tree Better（Occam's raizor）



$$E(T,X) = \sum_{c \in X} P(c)E(c)$$

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

E(PlayGolf, Outlook) = **P**(Sunny)***E**(3,2) + **P**(Overcast)***E**(4,0) + **P**(Rainy)***E**(2,3)

= (5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971

= 0.693

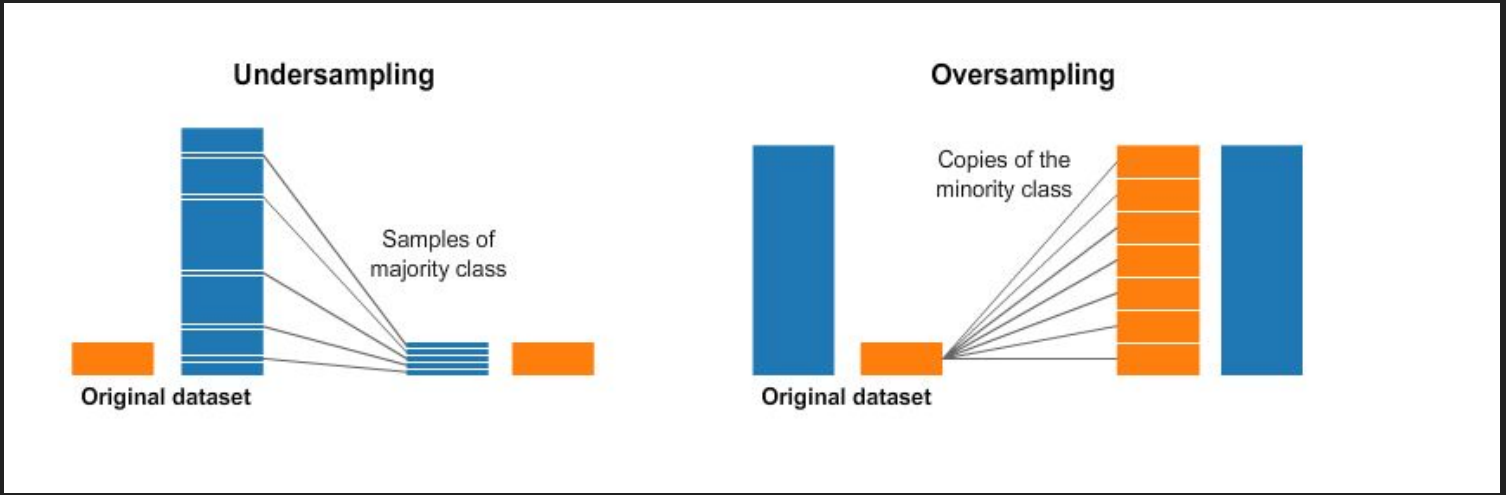
Entropy(PlayGolf) = Entropy (5,9)

= Entropy (0.36, 0.64)

= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)

= 0.94

Over sampling, Prunning, Validation



https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

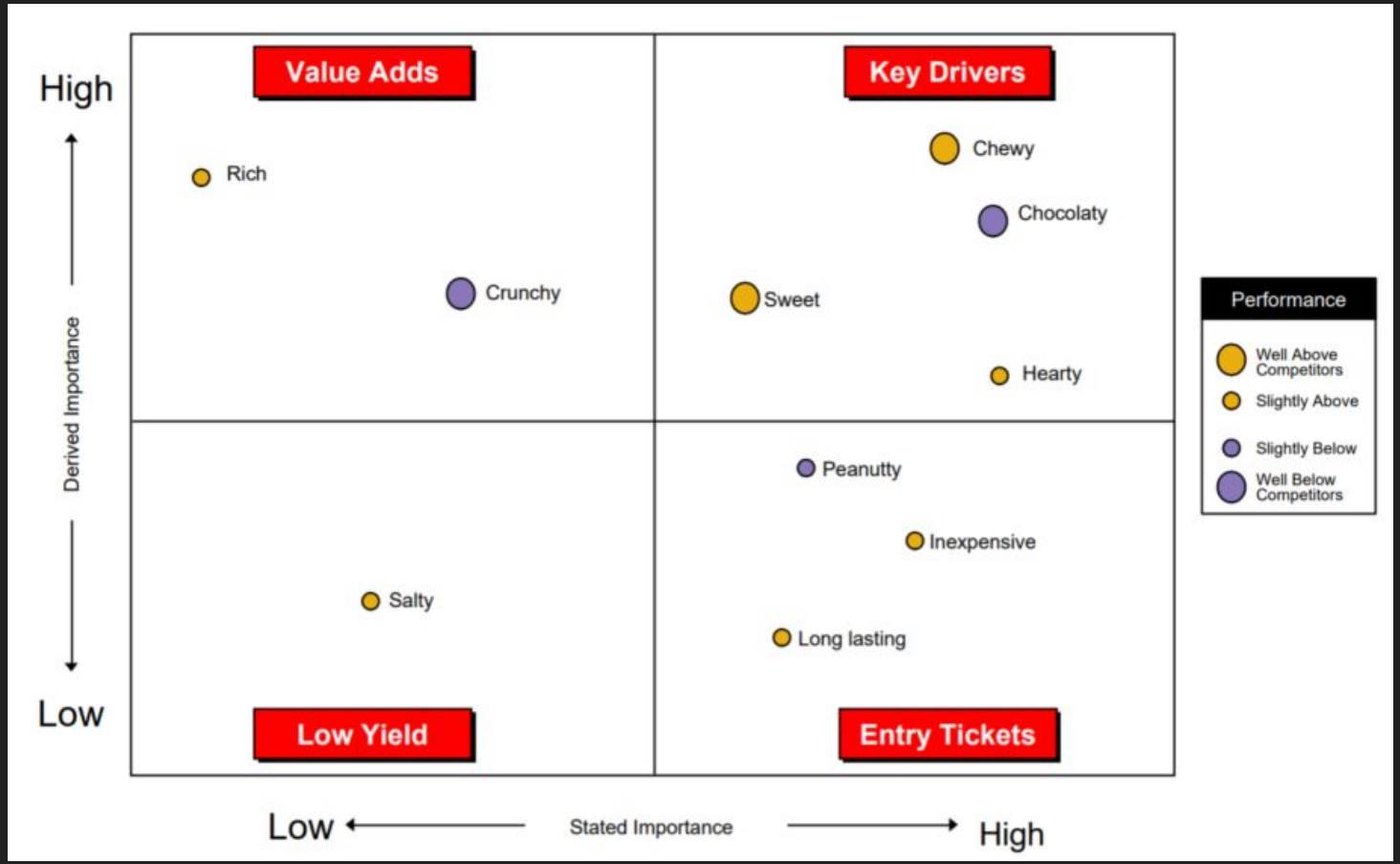
Clustering

Segmentation

The dividing of a market's customers into subgroups in a way that optimizes the firm's ability to profit from the fact that customers have different needs, priorities, and economic levers.

Keep in mind the end goal of enhancing profitability, as this can help increase the actionability of the segmentation. At each step ask "how can these results help improve profits?"

Matrix

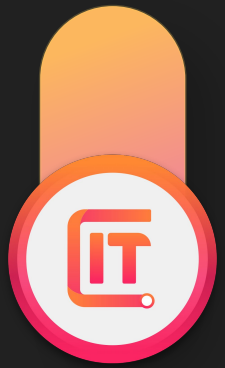

High
Low
Derived Importance
Value Adds
Key Drivers
Low Yield
Entry Tickets
Rich
Crunchy
Chewy
Chocolaty
Sweet
Hearty
Salty
Peanutty
Inexpensive
Long lasting
Low
Stated Importance
High
Performance
Well Above Competitors
Slightly Above
Slightly Below
Well Below Competitors

Decision tree Lab

High accuracy model-

- Too good to be true?(self-predicting)
- Useless (99% false = all false

https://www.kaggle.com/diegosch/classifier-evaluation-using-confusion-matrix



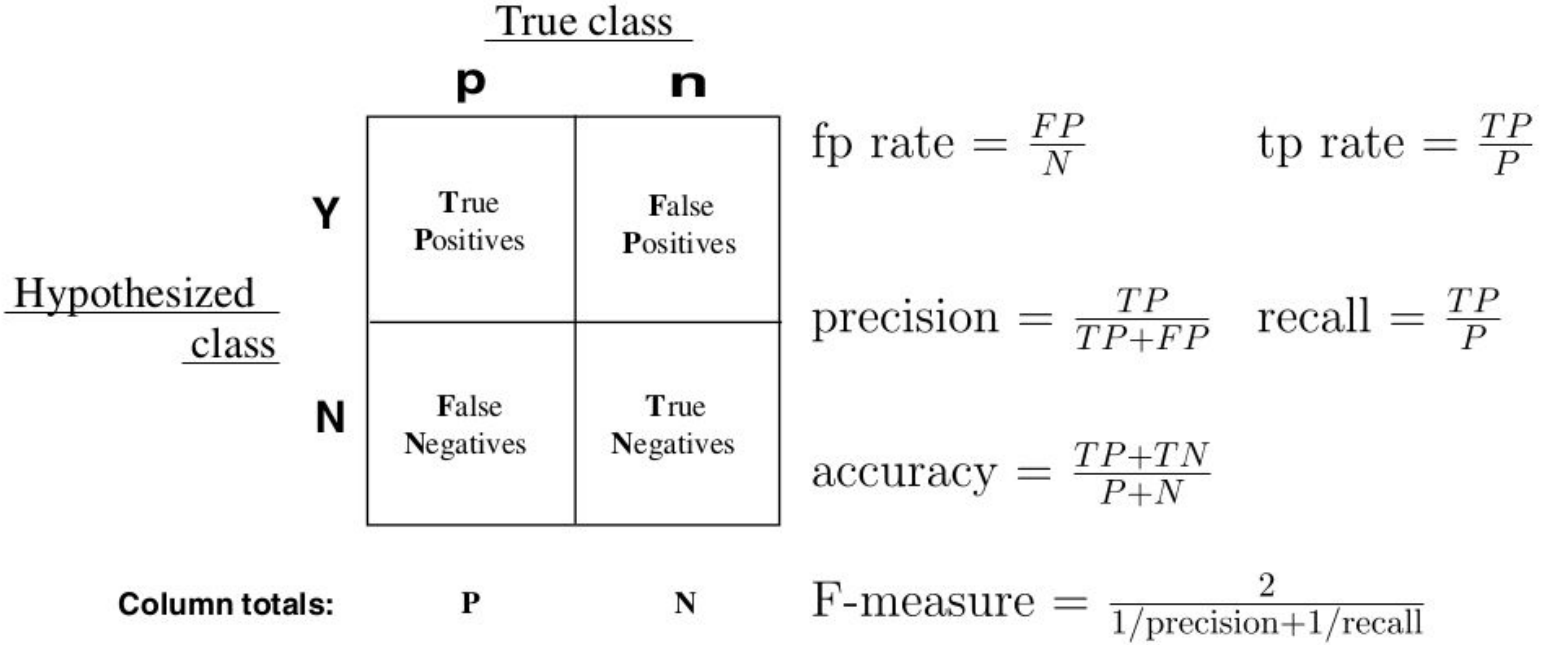


Fig. 1. Confusion matrix and common performance metrics calculated from it.

Black/whitebox

Measuring modeling quality - AUC, Lift, Precision

- 客户信用风险评分（SVM，决策树，神经网络）
-
- 市场风险评分建模（逻辑回归和决策树）
-
- 运营风险评分建模（SVM）
-
- 欺诈检测（决策树，聚类，社交网络）

应用模型	特征值属性个数	AUC值范围
信用风险评分	10-15	70%－85%
行为风险评分	10-15	80%－90%
流失预警	6-10	70%－90%
欺诈侦测（保险）	10-15	70%－90%

五 数据模型评价的方法

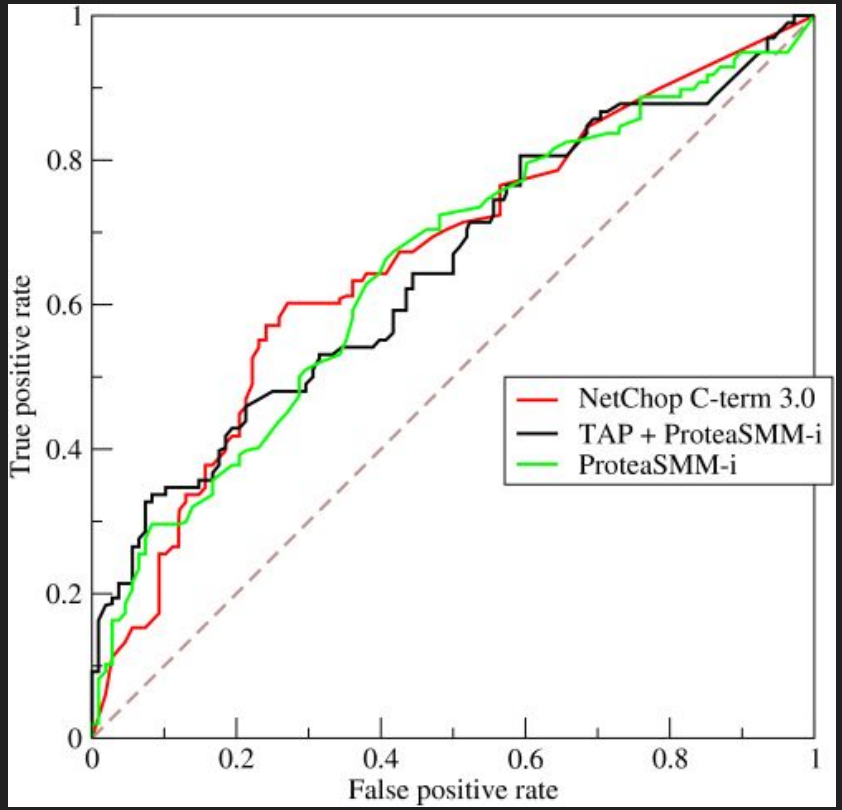
1 AUC值判别法

AUC小于0.7识别能力很弱
AUC在0.7-0.8之间识别能力可接受
AUC在0.8-0.9 之间识别能力卓越
AUC大于0.9 模型出现意外

讨论的ROC和AUC。下图是一个ROC曲线的示例[2]。

正如我们在这个ROC曲线的示例图中看到的那样，ROC曲线的横坐标为false positive rate（FPR），纵坐标为true positive rate（TPR）。下图中详细说明了FPR和TPR是如何定义的。

接下来我们考虑ROC曲线图中的四个点和一条线。第一个点，(0,1)，即FPR=0, TPR=1，这意味着FN（false negative）=0，并且FP（false positive）=0。Wow，这是一个完美的分类器，它将所有的样本都正确分类。第二个点，(1,0)，即FPR=1，TPR=0，类似地分析可以发现这是一个最糟糕的分类器，因为它成功避开了所有的正确答案。第三个点，(0,0)，即FPR=TPR=0，即FP（false positive）=TP（true positive）=0，可以发现该分类器预测所有的样本都为负样本（negative）。类似的，第四个点（1,1），分类器实际上预测所有的样本都为正样本。经过以上的分析，我们可以断言，ROC曲线越接近左上角，该分类器的性能越好

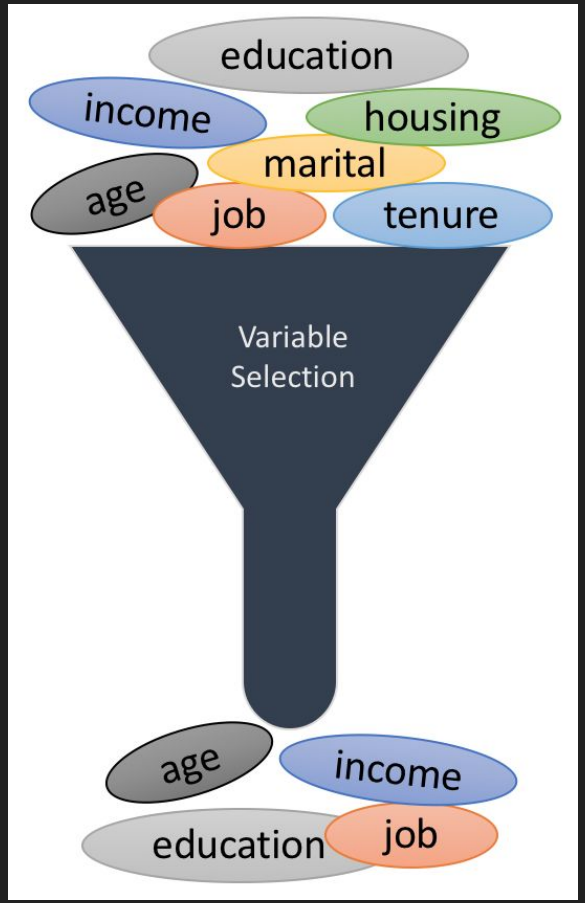


RFM Model-Alteryx

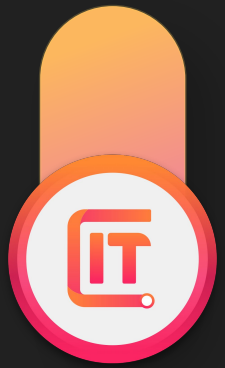
哪个国家/机型容易injure？continent. How many years old, airline history

Group model to derive attribute

用IV WOE来量化变量选择

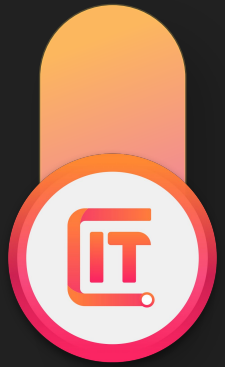


Correlation not Causation



- Which month are most of our customers born?
- Highest value woolies EDR member?

Kaggle建模冠军之犀利风格



“建模前我根本不看数据，5分钟跑一遍模型出来，再看feature”

简单树模型的组合，秒杀全部-Occam Raizor

“Forgot all the tools” -儿戏的随机数选的用Matlab来开发

数据侦查犹如汉尼拔打佩鲁贾 要日常做 不靠取巧

没有一个分类器解决不了的分类*问题*. 如果有, 就多用几个 www.k2data.com.cn/?p=4992

Incomplete data in, disaster out

Read and Exercise

Forecasting: Principles and Practice - OTexts

https://mp.weixin.qq.com/s/NSM98pmbq1ThQDfP0tYwbg

https://www.kaggle.com/diegosch/classifier-evaluation-using-confusion-matrix

https://www.kaggle.com/gargmanish/how-to-handle-imbalance-data-study-in-detail