MSBA 5510
Dominican University of California

Week 1: Data Analysis and Model Building Processes (draft)

Daqing Zhao
Summer 2023

About the Course

- Capstone Projects, in One of Three Groups
- Elect a project leader, a presentation leader and a technical leader
- Each member keeps a Google doc diary, shared with team and prof
- Choose a Data Set
- Conduct Exploratory Data Analysis
    - SQL queries, Data Visualizations
- Design Analysis and Modeling Objectives
    - Use techniques we learned, may learn some new tools
- Execute Analysis
- Analysis and Model Interpretation
- Presentations
    - PowerPoint, each member present a few slides

Learning Outcomes and Requirements

- Be able to build machine learning models using some opensource platform, and handle large data sets
- Be able to optimize and select good models
- Be able to prepare data sets from a database
- Be able to explain database concepts
- Be able to explain modeling concepts
- Be able to explain how models make prediction
- Be able to identify common problems in models
- Be able to QA own work
- Know own roles and those of others
- Be able to construct and explain models in business solutions
- Interact with teachers/students, ask questions

Learning Objectives This Week

- Knowing the business process of modeling building
- Learn about relations between modeling and solving business problems
- Knowing what data to request
- Learning to watch for and to avoid/correct modeling problems
- Knowing the iterative nature of modeling
- Learning to communicate with business partners
- Being able to prepare data from database tables
- SQL queries
- Use SQL to prepare data for machine learning

Typical Data Analysis Process

- Data collection
- Data selection
- Exploratory data analysis
- Data transformation and feature engineering
- Definition of target variables
- Modeling training, model assessment and model optimization
- Model selection
- Model interpretation
- Model scoring deployment
- Model validation
- Model usage and results analysis

Business Objectives of Analysis

- Targeted Marketing Messages
  - Emails, SMS, APP personalization
  - Customer acquisition, lead conversion, marketing, relationship management, sales, customer service
  - Media buys
- Product Recommendation in eCommerce
  - Personalization in real time
- Media Effective Analysis
  - Is the campaign working?
  - How can we better spend the media budget?
- Product Performance
  - Is the New Homepage or APP Better Than the Old One?

Business Objectives of Analysis

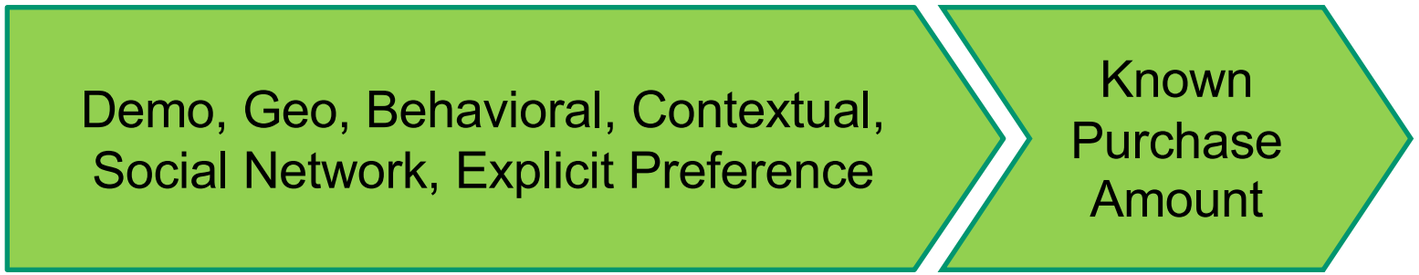
- Customer identification
- Customer segmentation, customer preference profiles, customer lifetime value
- Customer Lifetime Value (CLTV)
  - Who are our Best Customers?
  - At each customer level, how the CLTV changes in time
  - Their Profiles and Where to Acquire Them?
- Financial Forecast, sales, customer numbers, marketing spend, purchases, by segment, product category
  - For Marketing
  - For Finance

Business Objectives of Analysis

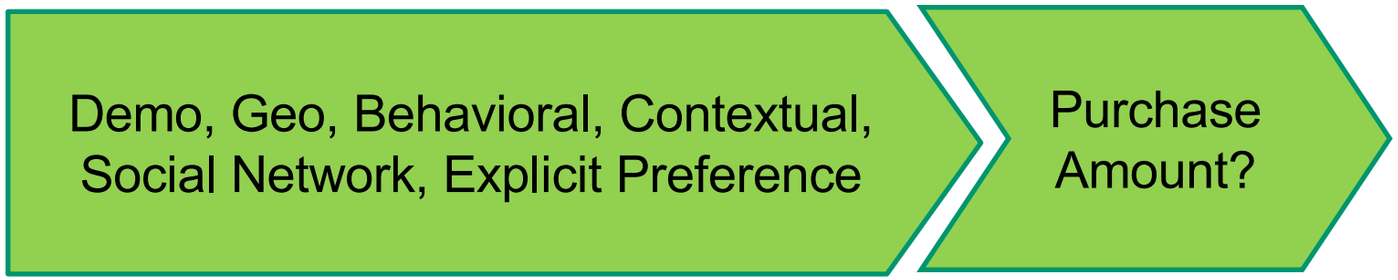
- Others
  - Supply Chain
  - Inventory Management, which products stocked in store?
  - Store Location, opening new and close low performing stores
  - Delivery Optimization, stores are fulfillment centers, which store should fulfill the online order?

Predictive Models

Collect Data, Build Model



Predict future cases



Feature Engineering

- Binary variables
- Clustering of variables
- Principal component analysis
- Recency, frequency, monetary, and variety
- Target encoding
- Embedding
- NLP processing, topic generation, sentiment analysis, word frequency, bag of words, word2vec, word sequence
- Neural network autoencoding
- Social network analysis
- Domain knowledge
- Variable selection

Feature Engineering

- Target encoding
- Discretizing continuous variables
- Embedding categorical variables
- Cluster selected variables
- Principal Component Analysis
- Transformation of target, such as log
- Generating features from unformatted data

Feature Selection

- Plot target against each predictor
- Calculate some metrics such as mutual information or Pearson corr with the target
- Use machine learning algorithms to identify important variables
- Decide modeling is for predictive accuracy or finding important predictor or driver of target
- Refine selection iteratively

Seasonality Features

- Weekly, monthly, yearly seasonality
- Cosine and Sine functions
- Lagging variables
- Current trend for this time period
- Decay functions
- Moving averages

Supervised versus Unsupervised

- Unsupervised
  - Summary
  - Compression, lossy and lossless
  - Characteristic of input data
  - General
  - Cluster analysis
- Supervised
  - Variable importance and selection biased toward supervision
  - Train for relation between predictors and target
  - Specific for one tasks, such as recognizing a picture

Supervised Machine Learning

- Classification
  - Binary, binomial
  - Multinomial
- Regression
  - Root mean squared errors, absolute errors
  - Heteroscedasticity

Data cleaning and organization

- Outlier identification and removal
- User generated data, typos
- Data normalization on range and mean
- Hierarchical data categories
- Unformatted data
- High dimensional data
- Correlated data
- Case coverage and data completeness
- Causality, association and correlation

Three Independent Data Sets

Training Data Set, Validation Data Set and Testing Data Set

Training data set is used to determine regression coefficients

Validation Data Set is used to decide variable selections, transformations, these are called model hyperparameters

Testing Data Set is used only to evaluate the model after using the training and validation data sets

Testing Data Set should NEVER be used to train the model or to select the model hyperparameters

When we don’t too little enough data, we may use cross validation to help

Model accuracy versus data size

- In general, the more data the better
- More rows, more columns/features
- Some models have limited capacity, resulting in diminishing improvement of accuracy as data size increases
- For example, simple linear models cannot capture nonlinear dependence of target on predictors
- Simple decision trees also hard for linear relations
- Good models can continue to improve accuracy as data size increases

Assessment of Models

- Training accuracy
- Validation accuracy
- Test accuracy
- Test on time shifted data set
- Continuous test for accuracy over time
- Data quality may change over time
- Uncaptured predictors may change over time
- People’s behavior may change over time
- Probability score versus ranking

Best models

- Best models depend on the problem to be modeled
- Best models solve the problem and are the simplest
- Linear, nonlinear models
- Nonparametric models
- Decision tree models
- Ensemble models
- Robust models
- Training time versus model accuracy
- Data freshness versus model accuracy
- Model accuracy versus model granularity
- Time series, freshness, seasonality, streaming, GAN

Agile and Iterate

- Go to Market Quickly
- Communicate with team members often
- Work with Partners to Identify Issues
- Find an Acceptable Solution
- Test and Optimize
- Iterate
- Limit investment
- Earn time and credibility
- Flush our problems early

Appendix

Passive and Active Data Collection

- Web log data
- Transaction data
- Credit card data
- Tracking pixel embedding
- Preference profiles
- Surveys and Reviews
- Third party data append
- Product data
- Aggregate sales data
- IoT Data

Typical Web Data

- Cookie ID, a little file web engines write on your computer
- IP address
- Referral URL
- Browser Type
- OS and version
- ISP
- Device
- Session ID
- View on Page or Product ID
- Time stamp
- User ID for registered customers
- Click on Page or Product ID or Content ID
- Search Keyword
- Conversion with Order ID

- IP address can be mapped to some geo information
- Pages are tagged with tracking pixels

Typical Transaction Data

- Customer ID
- Purchase Date
- Promote Type
- Order ID
- Store ID
- Payment Type
- Amount

- Order Detail table :
- Product ID
- Order ID
- Number of Items
- Price
- Discount

Additional Tables:

Customer table
Product table
Campaign table
Marketing Response table

Typical Registration data

- Name
- Demo
- Address
- Email
- Permission for Email Marketing
- Acquisition channel
- Credit Card information
- Registration date time
- Preference profile
- DMA
-

Marketing Channels

- Search Engine Marketing
- Organic Site Traffic
- Email Marketing
- APP
- SMS
- Display
- Social Network
- Direct Mail
- Affiliate Marketing
- TV

Marketing Response Data

- SEM, landing page CTR, engine CTR, CPC, QS, Avg Rank, Revenue, query category
- Display ads, publisher, publiusher content category, impression, session depth, page views, clicks, conversions
- Emails, sends, opens, clicks, conversions, product, content
- Affiliate, lead price, lead demo, lead conversion rate

Product Data

- Product ID
- Product name
- Product Category Hierarchy
- Brand
- Vendor
- Regular price
- Release date
- Product attributes
- Inventory level

Additional Tables:

Brand table
Vendor table

Third Party Append Data

- Personal data relating to credit cards
  - Name, address, gender, age, marital status, prior address, household data
- Customer identification algorithm
- Credit Card Data (aggregate anonymized)
- Census data
- Auto registration data
- Property tax data
- Product registration data
- Survey data
- Preference model prediction

Acxiom

Acxiom Data Aggregator
Customer Segmentation
Customer Identification
Data Management Platform (DMP), resells media
Personicx
Try:

https://isapps.acxiom.com/personicx/personicx.aspx

https://media.cmgdigital.com/shared/news/documents/2014/02/03/see_p_240_PersonicX_Binder.pdf

Personicx Clusters

70 PersonicX Classic Clusters Organized by Cluster Code

Cluster #	Group #	Group Name	Cluster Name	Age	Marital Status	Home Ownership	Kids	Income	Rank	Urbanicity	Rank	Net Worth	Rank
01	11B	Fortunes & Families	Summit Estates	36-55	Married	Owner	School-age Kids	Wealthy	1	City & Surrounds	35	$2MM+	1
02	15M	Mature Wealth	Established Elite	46-65	Married/Single	Owner	No Kids	Wealthy	2	City & Surrounds	12	$2MM+	2
03	15M	Mature Wealth	Corporate Clout	46-65	Married	Owner	No Kids	Wealthy	3	City & Surrounds	33	$1MM-$2MM	3
04	11B	Fortunes & Families	Skyboxes & Suburbans	36-65	Married	Owner	School-age Kids	Wealthy	4	Suburbs & Towns	51	$1MM-$2MM	4
05	19M	Golden Years	Sitting Pretty	46-55	Married	Owner	No Kids	Wealthy	7	Suburbs & Towns	56	$100K-$999K	7
06	07X	Cash & Careers	Shooting Stars	30-45	Married	Owner	No Kids	Wealthy	6	Suburbs & Towns	43	$100K-$499K	9
07	11B	Fortunes & Families	Lavish Lifestyles	36-55	Married	Owner	School-age Kids	Wealthy	5	Suburbs & Towns	57	$100K-$499K	10
08	19M	Golden Years	Full Steaming	56-65	Married/Single	Owner	No Kids	Affluent	13	Suburbs & Towns	47	$500K-$999K	6
09	19M	Golden Years	Platinum Oldies	66+	Married/Single	Owner	No Kids	Upper Middle	24	City & Surrounds	11	$500K-$999K	5
10	07X	Cash & Careers	Hard Chargers	30-45	Single	Owner	No Kids	Affluent	8	Suburbs & Towns	41	$50K-$499K	11
11	08X	Jumbo Families	Kids & Clout	36-45	Married	Owner	School-age Kids	Affluent	9	Suburbs & Towns	44	$50K-$499K	12
12	08X	Jumbo Families	Tots & Toys	30-45	Married	Owner	Toddlers/Preschool	Affluent	10	City & Surrounds	28	$5K-$499K	29
13	12B	Flush Families	Solid Single Parents	36-55	Single	Owner/Renter	Kids; Age Mix	Affluent	14	City & Surrounds	21	$25K-$499K	22
14	16M	Aging Upscale	Career Centered Singles	46-65	Single	Owner	No Kids	Affluent	12	City & Surrounds	37	$25K-$499K	20
15	16M	Aging Upscale	Country Ways	46-65	Married	Owner	No Kids	Affluent	11	Rural	70	$25K-$999K	16
16	14B	Our Turn	Country Single	36-65	Single	Owner	No Kids	Upper Middle	18	Rural	65	$25K-$499K	21
17	12B	Flush Families	Apple Pie Families	46-65	Married	Owner	School-age Kids	Upper Middle	15	City & Surrounds	32	$25K-$999K	15
18	02Y	Taking Hold	Married Sophisticates	30-35	Married	Owner	No Kids	Upper Middle	19	Suburbs & Towns	49	$25K-$499K	18
19	08X	Jumbo Families	Country Comfort	36-55	Married	Owner	Kids; Age Mix	Upper Middle	16	Rural	59	$25K-$499K	33
20	07X	Cash & Careers	Dynamic Duos	36-45	Married	Owner	No Kids	Upper Middle	20	Suburbs & Towns	48	$25K-$499K	19
21	02Y	Taking Hold	Children First	18-29	Married/Single	Owner/Renter	Kids; Age Mix	Upper Middle	27	Suburbs & Towns	52	<$500K	49
22	14B	Our Turn	Fun & Games	46-55	Married	Owner	No Kids	Upper Middle	17	Suburbs & Towns	53	$25K-$499K	23
23	16M	Aging Upscale	Acred Couples	56-65	Married	Owner	No Kids	Upper Middle	21	Suburbs & Towns	55	$25K-$499K	32
24	02Y	Taking Hold	Career Building	24-29	Single	Renter/Owner	No Kids	Upper Middle	23	City & Surrounds	38	<$100K	56
25	20S	Active Elders	Clubs & Causes	66-75	Married/Single	Owner	No Kids	Upper Middle	22	Suburbs & Towns	54	$25K-$499K	25
26	07X	Cash & Careers	Savvy Singles	30-45	Single	Renter/Owner	No Kids	Upper Middle	26	City & Surrounds	23	<$500K	50
27	08X	Jumbo Families	Soccer & SUVs	30-45	Married	Owner	School-age Kids	Upper Middle	29	Suburbs & Towns	39	$5K-$499K	37
28	20S	Active Elders	Suburban Seniors	76+	Married/Single	Owner	No Kids	Upper Middle	25	City & Surrounds	34	$25K-$499K	24
29	09B	Middling Singles	City Mixers	36-45	Single	Owner/Renter	No Kids	Upper Middle	28	Downtown Metro	1	<$999K	17
30	02Y	Taking Hold	Spouses & Houses	24-29	Married	Owner	No Kids	Middle	31	Suburbs & Towns	42	$25K-$499K	35
31	14B	Our Turn	Mid Americana	46-65	Married	Owner	No Kids	Middle	34	Suburbs & Towns	46	$25K-$999K	13
32	14B	Our Turn	Metro Mix	46-65	Married/Single	Owner	No Kids	Middle	32	Downtown Metro	2	$100K-$999K	8
33	14B	Our Turn	Urban Tenants	46-65	Single/Married	Renter	No Kids	Middle	30	Downtown Metro	4	<$100K	57
34	03X	Transition Time	Outward Bound	30-45	Married	Owner	No Kids	Middle	33	Rural	67	<$250K	44
35	09B	Middling Singles	Solo and Stable	36-45	Single	Owner	No Kids	Middle	37	City & Surrounds	19	<$500K	40
36	20S	Active Elders	Raisin' Grandkids	66+	Married/Single	Owner	Kids; Age Mix	Middle	35	City & Surrounds	24	$10K-$999K	14
37	05X	Family Focused	Cartoons & Carpools	30-45	Married	Owner	Kids; Age Mix	Middle	43	City & Surrounds	20	<$500K	47
38	13B	True Blues	Midtown Minivanners	46-65	Married	Owner	School-age Kids	Low Middle	49	City & Surrounds	13	<$500K	38
39	01Y	Beginnings	Early Parents	18-29	Single/Married	Renter/Owner	Kids; Age Mix	Low	60	City & Surrounds	30	<$50K	59
40	18M	Mature Rustics	The Great Outdoors	46-65	Married	Owner	No Kids	Low Middle	45	Rural	68	$10K-$999K	31
41	03X	Transition Time	Truckin' & Stylin'	30-45	Single/Married	Owner/Renter	No Kids	Middle	39	Rural	60	<$100K	58
42	04X	Flying Solo	First Mortgage	24-35	Single	Owner	No Kids	Low Middle	41	City & Surrounds	16	<$500K	46

BEGINNINGS—EARLY PARENTS

Cluster 39 (Group 1Y)

At a mean age of 25, Early Parents represents one of the youngest of the segments. It contains single and married parents in their mid-20s whose spending habits and leisure time are heavily influenced by their young children.

BEGINNINGS

LOW

CITY & SURROUNDS

ABOUT BEGINNINGS—EARLY PARENTS

Early Parents ranks among the nation's lowest clusters for income and net worth (ranked 60th for income and 59th for net worth). The majority of the cluster (71%) is high school educated. There is a fairly even distribution of families in this ethnically diverse cluster who own and rent their homes, but quite consistently, the length of residence is short. With limited financial resources, leisure time is focused on less expensive entertainment, often spent with their children. Their spending habits are also centered on their new family status — they are over four times more likely to be out buying strollers and high chairs. Even their choices in reading material — *Parenting*, *Baby Talk* and *American Baby* — reflect the toddler-centric nature of this cluster.

WHEN THEY GREW UP...

- *The Notebook* is released
- Facebook is introduced for the first time
- *Lost* is a TV hit
- Tom Brady and the New England Patriots dominate the NFL

CLUSTER SIZE

Households: 1,951,400
% U.S. Households: 1.58%

FINANCIAL/INSURANCE:

Money Orders
Check Cashing Services
H&R Block (on-site)
Medicaid
One insured auto

SHOPPING:

Impulse buyers
Wal-Mart/Wal-Mart Supercenter
Convenience Stores
Auto Zone
Foot Locker
Hot Dogs
Ice Cream
Juice Drinks

RADIO/TV:

Nickelodeon
Disney
Urban radio
Family Guy
Sports
Go Diego Go
Telemundo
Soap Operas/Novelas

MAGAZINES/NEWSPAPERS:

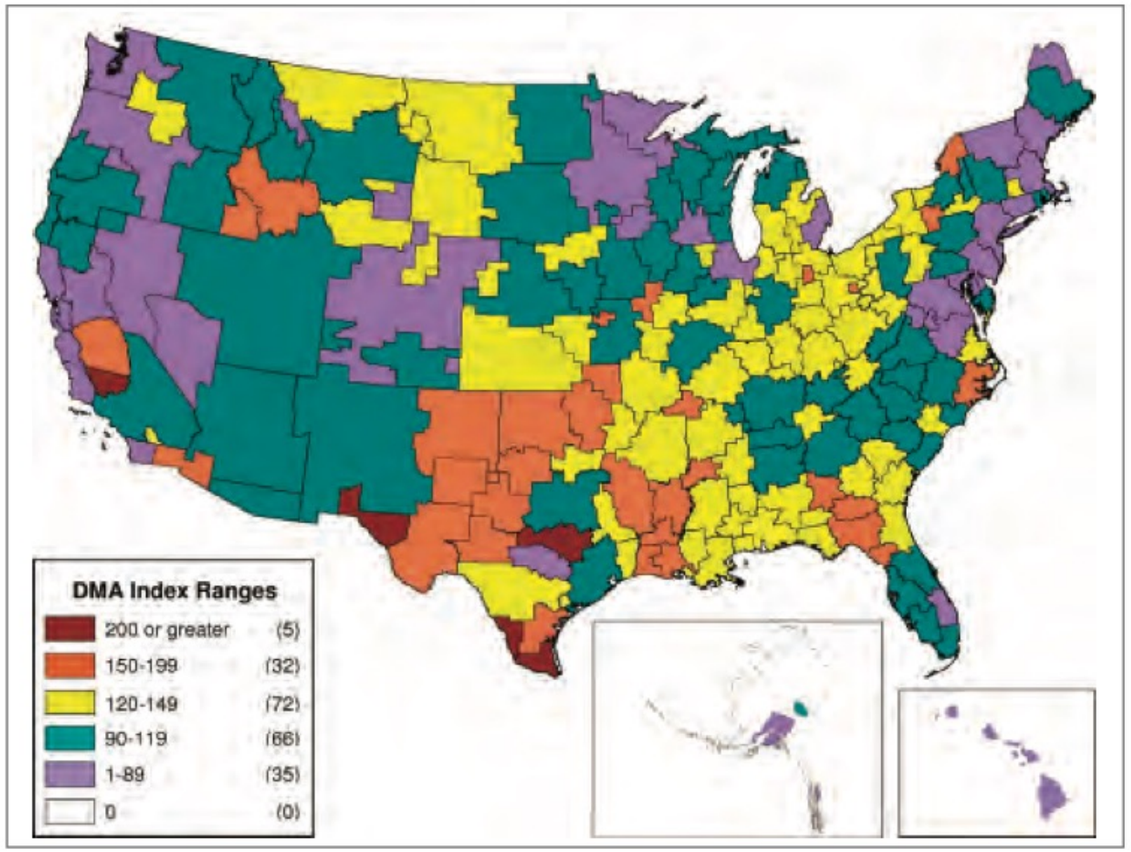
Baby Talk
Parent and Child
American Baby
Parenting
Official X-Box Magazine
Cosmopolitan

ACTIVITIES:

Computer Entertainment/Games
Go to Movies
Roller Skating
Hunting

Young, Parents, Less Educated

DISTRIBUTION OF BEGINNINGS—EARLY PARENTS



Computer Entertainment/Games
Go to Movies
Roller Skating
Hunting
Football

COMPUTERS/ONLINE:
MySpace
Disney
Job Searching
Yahoo
Play Video Games Online
Chat Rooms
Instant Messaging

AC-0753-10 10/10

INFOBASE-X® DEMOGRAPHIC CHARACTERISTICS BEGINNINGS—EARLY PARENTS

Cluster 39 (Group 1Y)

	Group %	National %	Index
Age—Head of Household			
18-23 Years	27.6	2.3	1208
24-29 Years	70.1	7.6	919
30-35 Years	2.3	10.5	22
36-45 Years	0.0	21.5	0
46-55 Years	0.0	21.5	0
56-65 Years	0.0	15.4	0
66-75 Years	0.0	10.7	0
76+ Years	0.0	10.5	0
MEAN AGE			**24.8**
Estimated Income			
<$15,000	27.0	10.1	267
$15,000-$19,999	12.9	4.7	272
$20,000-$29,999	26.6	8.8	301
$30,000-$39,999	33.5	10.8	309
$40,000-$49,999	0.0	11.2	0
$50,000-$74,999	0.0	24.2	0
$75,000-$99,999	0.0	13.9	0
$100,000-$124,999	0.0	6.3	0
$125,000-$149,999	0.0	4.7	0
$150,000+	0.0	5.1	0
Presence and Age of Children			
No Children Present	0.0	65.9	0
Ages 0-2	35.1	6.5	543
Ages 3-5	28.4	7.0	406
Ages 6-10	27.1	11.1	245
Ages 11-15	13.2	10.6	124
Ages 16-17	14.4	7.5	192
Marital Status			
Single	61.9	44.0	141
Married	38.1	56.0	68

	Group %	National %	Index
Length of Residence			
<2 Years	28.2	12.7	222
2-5 Years	46.7	30.2	154
6-14 Years	16.5	32.0	53
15+ Years	8.6	25.0	34
Market Value of Home			
<$50,000	17.7	10.3	172
$50,000-$99,000	30.9	19.1	162
$100,000-$124,999	11.3	8.8	128
$125,000-$149,999	8.5	8.0	106
$150,000-$199,999	11.7	12.9	90
$200,000-$299,999	10.6	15.9	67
$300,000-$500,000	6.6	14.7	45
$500,000+	2.8	10.5	27
Dwelling Unit Size			
Single Family Dwelling	75.5	86.0	88
Multiple Family Dwelling	24.5	14.0	174
Occupation			
Professional/Technical	13.4	30.4	44
Administrative/Managerial	3.8	6.8	57
Sales/Service	3.1	1.7	175
Clerical/White Collar	32.2	16.7	193
Craftsman/Blue Collar	20.8	18.7	111
Student	5.2	0.8	655
Housewife	12.9	6.0	216
Retired	0.5	12.1	4
Other	6.1	3.8	150
Self Employed	2.1	3.1	67
Education			
Completed High School	71.0	53.1	134
Completed College	26.5	33.1	80

Marital Status			
Single	75.1	44.0	171
Married	24.9	56.0	44

Estimated Net Worth			
< $1	62.2	9.4	665
$1-$4,999	16.6	8.8	189
$5,000-$9,999	5.1	5.5	93
$10,000-$24,999	2.5	5.2	47
$25,000-$49,999	3.1	8.6	37
$50,000-$99,999	3.7	10.0	38
$100,000-$249,999	4.3	20.6	21
$250,000-$499,999	1.7	15.0	11
$500,000-$999,999	0.7	9.4	8
$1,000,000-$1,999,999	0.0	3.6	1
$2,000,000+	0.0	4.1	0

Home Ownership Status			
Renter	78.9	23.2	340
Home Owner	21.1	76.8	27

Population Density – HH per Sq. Mile			
0-24	7.9	8.8	89
25-83	8.6	9.3	92
84-1,015	28.6	33.4	86
1,016-3,015	37.6	33.2	113
3,016-5,440	13.6	8.3	164
5,441-9,948	3.3	3.6	93
9,949+	0.3	3.4	10

Education			
Completed High School	76.2	53.1	143
Completed College	21.5	33.1	65
Completed Graduate School	1.5	13.2	12
Attended Vocational/Technical	0.7	0.6	117

Ethnicity			
Caucasian	61.2	74.3	82
African American	18.2	10.3	178
Hispanic	16.4	10.7	154
Asian	3.0	3.6	84
Other	1.1	1.1	97

Household Size			
One Person Household	48.5	24.7	196
Two Person Household	27.5	28.8	95
Three Person Household	14.4	21.3	68
Four Person Household	5.6	13.0	43
Five+ Person Household	3.9	12.1	33

Mail Responsive			
Mail Order Responsive	34.2	76.8	44
Mail Order Buyer	34.0	76.6	44
Mail Order Donor	0.1	2.8	3

Buying Channel Preference – Decile			
Top Internet Decile	8.9	8.9	100
Top Mail Decile	0.2	11.9	1
Top Phone Decile	0.0	12.0	0

GOLDEN YEARS—PLATINUM OLDIES

Cluster 09 (Group 19M)

These well-heeled retirees and soon-to-be retirees living in the outer edges of the city are enjoying the fruits of their lifetime labor. They are active pillars of their communities, dedicated grandparents, and interested in maintaining their health and fitness.

GOLDEN YEARS

UPPER-MIDDLE

CITY & SURROUNDS

ABOUT GOLDEN YEARS—PLATINUM OLDIES

At the age of 66+, Platinum Oldies households have established themselves financially and socially. These elderly couples (62%) and singles enjoy high net worth (fifth), are extremely well educated and rank first for home home values of $300,000 to $500,000. They are financially secure and very well vested in a mix of

WHEN THEY GREW UP...

- President Kennedy killed In Dallas, TX
- Valium is developed
- Julia Child makes her

CLUSTER SIZE

Households: 3,062,500
% U.S. Households: 2.47%

FINANCIAL/INSURANCE:

Securities: $150,000+
Medicare
Tax Exempt Funds
Use Full-Service Brokerage Firm
Annuities

SHOPPING:

Catalogs
L.L. Bean
Trader Joe's
Macy's
Pathmark
Costco
Bed Bath & Beyond

RADIO/TV:

Tournament of Roses Parade
Golf Channel
Live from Lincoln Center
The O'Reilly Factor
60 Minutes
News Radio

MAGAZINES/NEWSPAPERS:

Appendix

Feature Engineering

- Once we have a data set with predictors and target, we may do some transformations based on what we have
- The goal is to improve model performance
- Use domain knowledge and insight we find in EDA
- We can do experiments and improve accuracy
- The process is iterative and time consuming
- A simple example is creating indicator variables
- Remove irrelevant predictors
- Smoothing to reduce noise when sample is small
- Feature engineering is key to good model building

Feature Engineering

- Some variables are redundant
- Duplicate, or transformation of some variable
- Some feature may come from different sources, say age from date of birth and from self reporting
- Some variables may not be a simple transformation of another variable, but can be a complex linear combination of many variables
- Such variables do not provide additional information for prediction

Feature Engineering

- Auto encoding
- High dimensional categorical variables, use others
- Use numbers as discrete variables
- Caution leakage when doing feature engineering
- Missing value imputation
- https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02

Feature Engineering

- Normalization
- Subtract mean and divide by standard deviation
- Turn numbers into ranges or categories
- Log transform, sqrt, plus 1
- Revenue
- Link function, logit, Poisson, ordinal, multinomial
- Image data, labeling, categorization, feature identification
- Standardize, normalize

Feature Engineering

- Labeled encoding
  - Use some sequence number as values for a category
  - For color yellow is 1, red is 2, blue is 3, black is 4, ...
- One hot encoding
  - Create a new variable for each categorical value use 1 and 0
- Frequency encoding
  - Use frequency of value as value
- Target mean, median encoding
  - Use mean or median of target value as value
  - CA converts at 0.025, NY converts at 0.035

Feature Engineering

- Binning is to put values in some bins or buckets
  - Ways to handle sparse data
- Binning so that each bin has the same population and use population mean or median as value
- Binning using histogram so that value interval are the same
- Clustering of a set of variables
- Cluster using cluster id or distance to cluster center as new variables
- Find interaction variables, use products, powers or tree rules

Feature Engineering

- Weight of Evidence (WoE):

  Ln(event%/non-event%)

  WoE = 0 if event% = non-event%

  |WoE| → large when event% and non-event% are very different

- Information Value (IV):

  ∑ (event% - non-event%)*Ln(event%/non-event%)

  sum is over all values of a group variable (e,g.,age)

Network Analysis

- https://www.jessesadler.com/post/network-analysis-with-r/
- Certain information is not from any specific predictor but the relationships among predictors or values of predictors
- Network information is a key type of such information
- Nodes and edges, directed, undirected, node and edge attributes
- Degree, cluster, closeness, betweenness centrality, hub
- For more, Stanford has a course:
- https://www.youtube.com/watch?v=3IS7UhNMQ3U&list=RDCMUCBa5G_ESCn8Yd4vw5U-gIcg&index=2

Network Analysis Use Cases

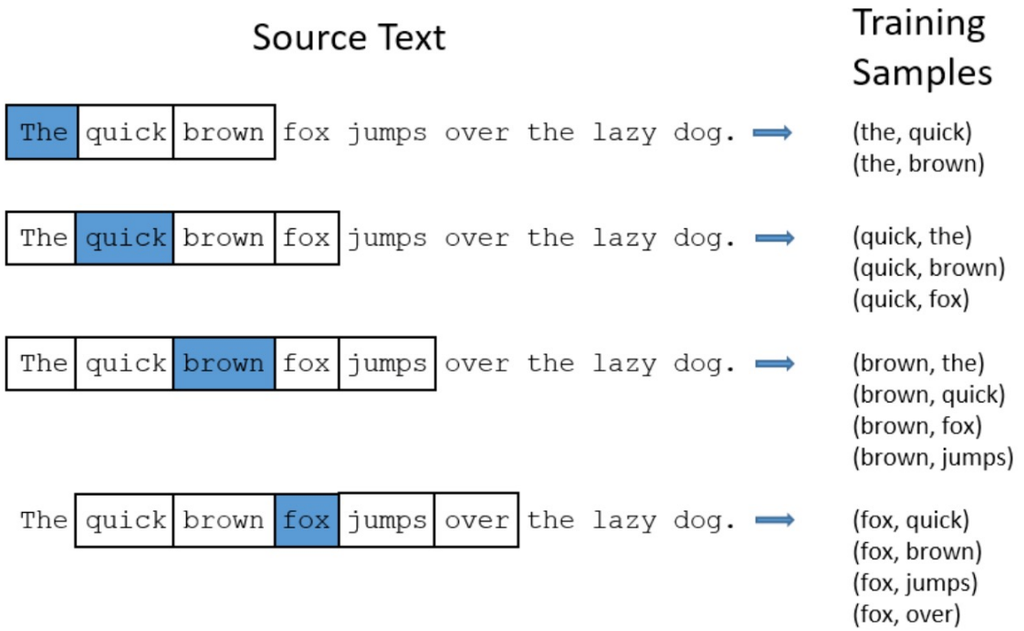
- Fraud detection input data graph, presence of edges between input data, such as share same street, zip, dob, last name
- Influencer marketing
- Products with radiated sales

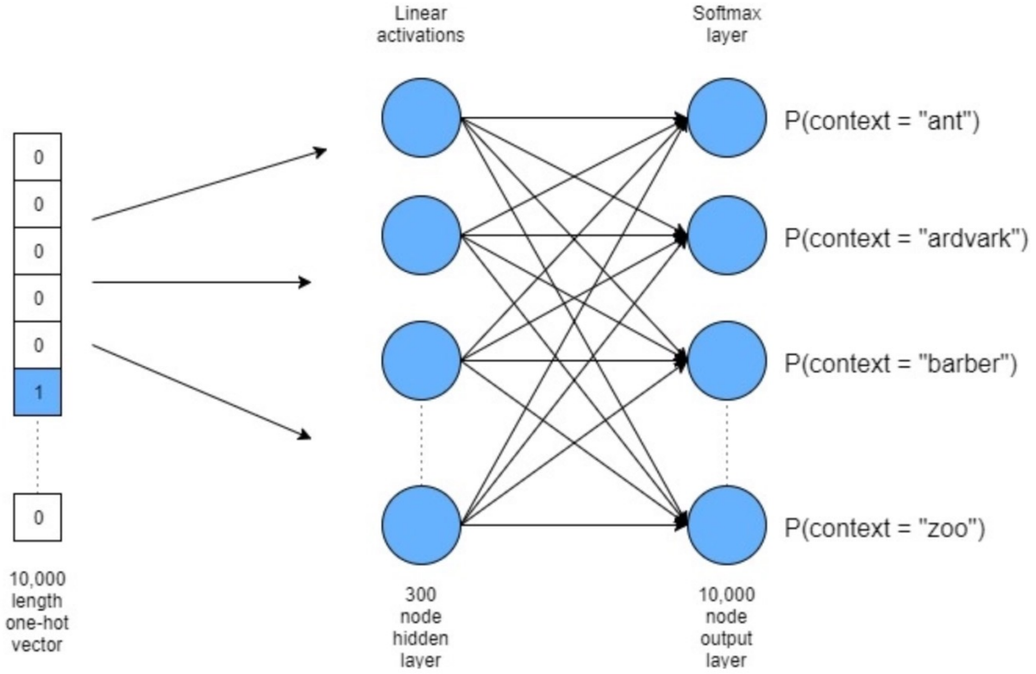
Tf-idf, LDA, Word2Vec

- Tf-idf is term frequency inverse document frequency
- LDA is local Dirichlet allocation
- Word2Vec
- Transfer learning embedding

word2vec

- Google methodology to represent words
- Each word is represented by an array of numbers or a vector
- We can set the number of numbers, dimension of the vector





https://towardsdatascience.com/word2vec-skip-gram-model-part-1-intuition-78614e4d6e0b

word2vec

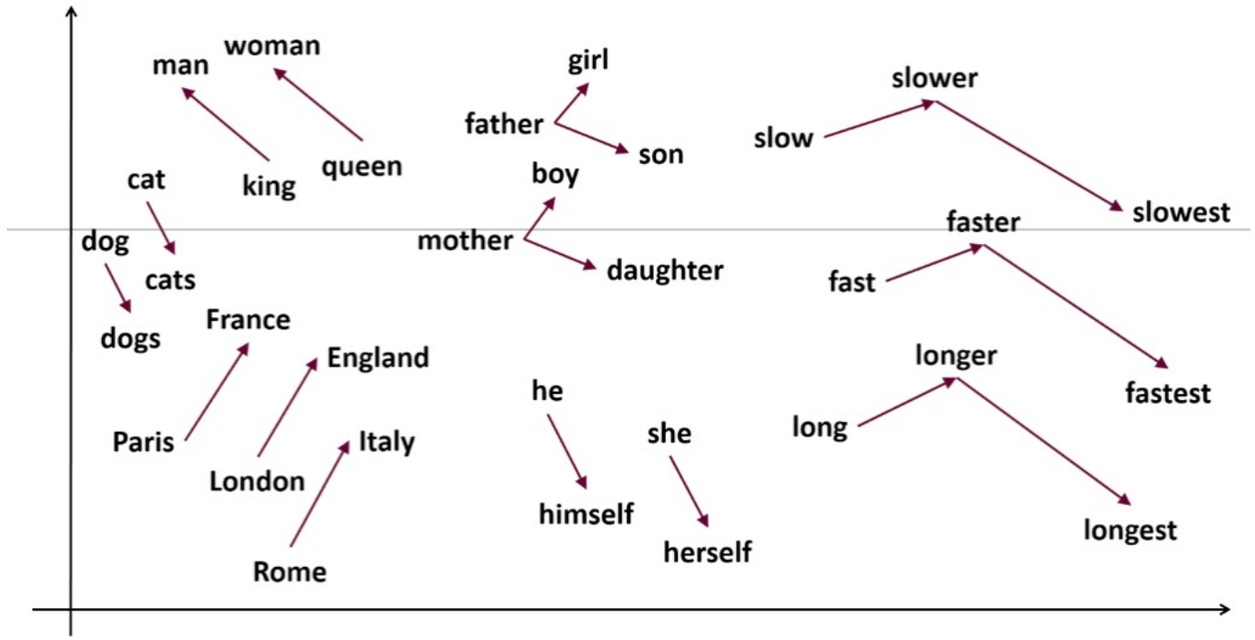
Words that appear in similar contexts are similar
In contexts of sentences where “president” is mentioned,

- POTUS,
- White House,
- Joe Biden,
- 46

are also more likely to appear
Vectors for these words are similar

word2vec

- Google methodology to represent words
- Each word is represented by an array of numbers or a vector
- We can set the number of numbers, dimension of the vector



Word2Vec

- Use Word2Vec as feature engineering technique for texts
- Input is a single column with rows of words delimited by NA
- http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-munging/tokenize.html#tokenize
- Output is a vector of numbers
- https://github.com/h2oai/h2o-3/blob/master/h2o-r/demos/rdemo.word2vec.craigslistjobtitles.R
- See R Markdown for Craigslist in Moodle, later today

Two Purposes of Modeling

- First purpose, to predict the event
- For example, customer A is 40% likely to buy item X in next week
- Typical conversion probability is < 1%
- Second purpose, to analyze why the event happened
- Promotion event, TV advertising, but there is winter storm, holiday or not
- How much promotion or TV advertising influenced the sale event?

Logistic Regression

- Coefficients
- Logit P = 0.01 + 0.5*A – 0.28*B
- Variables with larger coefficients are more important
- Variables ranges also important
- Use L1 and L2 regularizations, Elastic Net lambda and alpha, to select only useful predictors
- Interpretation is not unique

Simple Decision Trees

- If A > 2 then (Node 2) P = 0.5, else (Node 3) P = 0.02
- Node 2 if B > 0.2 then (Node 4) P = 0.4 else (Node 5) P = 0.6
- Predictors split more cases, reduce more loss, appear more often in the tree are more important

More Accurate Models

- More accurate models have larger capacity for more complex relations, but it is harder to describe in simple terms
- Two purposes of interpreting a model
- a) aggregate effects
- b) reason code for each case, at the individual level
- Rank list of important variables
- Quantify contribution of a variable in prediction

Interpretable Models

- Build accurate Machine Learning Model with model possibly being complex
- Build simpler models to approximately describe how the accurate model works
- Use same predictors but predicted value as targets
- Score each case using simple model to get interpretations

K Cluster and Linear Models

- Local Surrogate, or LIME (for Local Interpretable Model-agnostic Explanations)
- Segment customers in some way and build linear models for each segment
- Locally interpret the models near the customer in question
- Simple reason code can be provided but not consistent
- Two different segmentation methods can lead to very different reason codes

K Cluster and Linear Models

- Describe Nearest Neighbor type of interpretation
- The more similar the given case is relative to the neighbors, the more accurate the interpretation
- Depend on sample size and stability of distributions

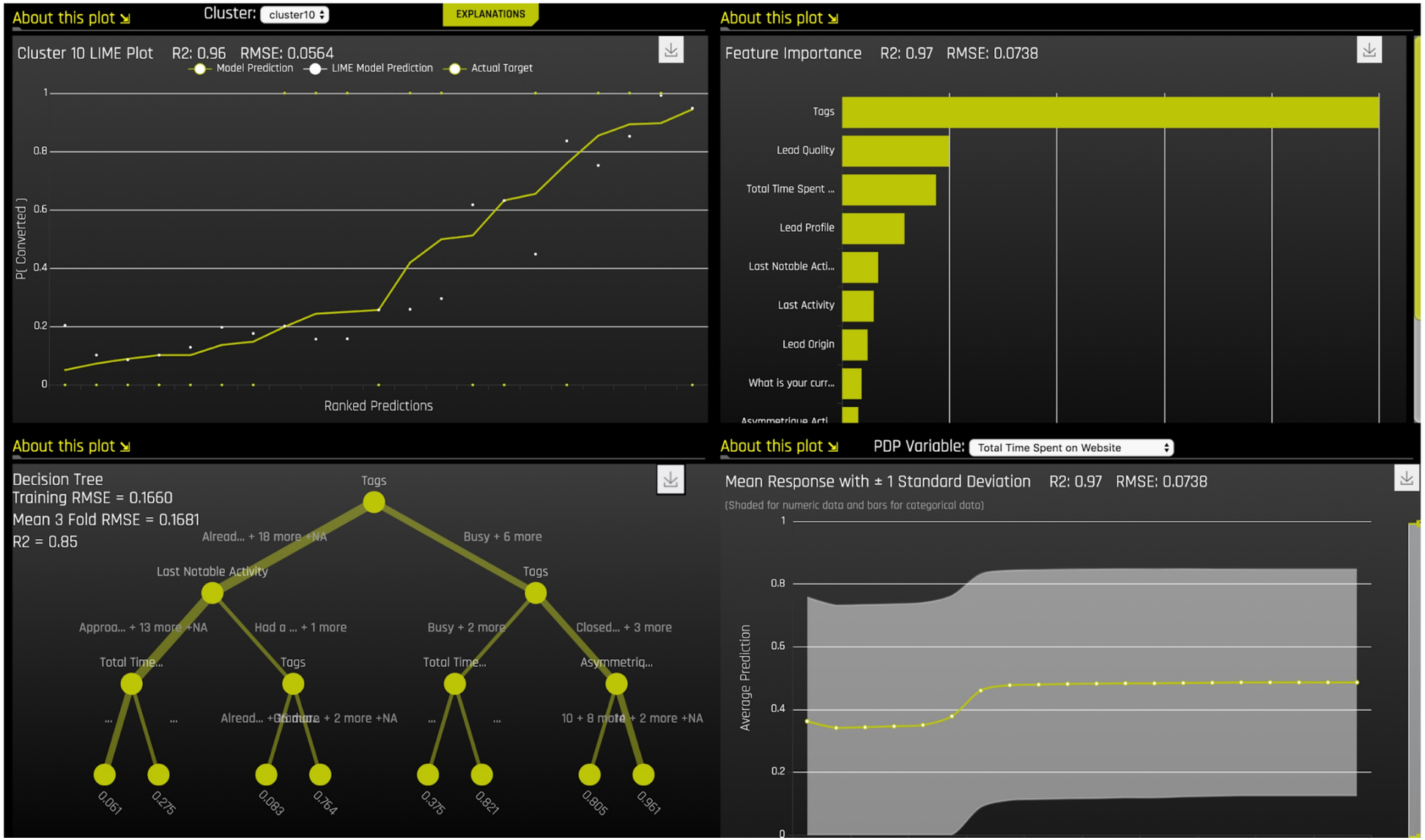
Decision Trees

- Simple tree tells a story by itself
- Rank import variable by how much each contribute to differentiation of cases
- Aggregate entropy difference for each splits to give value of importance of a variable
- Decision trees are unstable
- Split may happen differently, some variables may correlate with others
- Ranking is not always consistent
- Depend on presence of other variables

H2O.ai: Driverless AI

- Commercial product
- https://aquarium.h2o.ai/login
- Free two-hour training sessions
- Upload the Leads.csv data
- Start Experiment
- Choose 4 for Accuracy, 2 for Time, 4 for Interpretability

Surrogate Models by Driverless AI



Variable Importance

- In XGBoost
- Weight or frequency, the number of times a predictor is used in splitting
- Gain is average gain in loss function for all splits by this predictor
- Cover is the number of splits weighted by the number of cases affected by each split
- The 3 metrics may result in very different ordering

What Qualify as Best Models

- Best models depend on the problem to be modeled
- Best models solve the problem and are the simplest
- Linear, nonlinear models
- Non parametric models
- Decision tree models
- Ensemble models
- Robust models
- Training time versus model accuracy
- Data freshness versus model accuracy
- Model accuracy versus model granularity
- Time series, freshness, seasonality, streaming, GAN

What Does a Model Predict?

- A model does not predict what a person will or will not do
- It predicts in a significantly large group of people sharing given characteristics will have a probability value of having some behavior, with a score of say 0.65
- “Blue collar rust belt white older men” are 65% likely to vote for Trump, but this only says if we sample 1000 such people, 65% +/- 3% would vote for Trump
- Probability may change also but rank is more stable
- John Smith is such a person, the model will either be right and wrong about his voting behavior
- The model does not say that he would vote for Trump 65% of the time

Targets and Predictors

Unsupervised Learning

Supervised Learning

Binary Target

Numeric Target

Multiclass Classification

One versus all

One versus one

Data Formats for Models

- Flat files, target and a tuple of predictors
- A mixed list of numerical, categorical, ordinal vars, and a target
- Format usually are in:
- Text files
- Database tables
- R data frames
- H2O data frames
- JSON
- on HDFS
- on Spark RDD

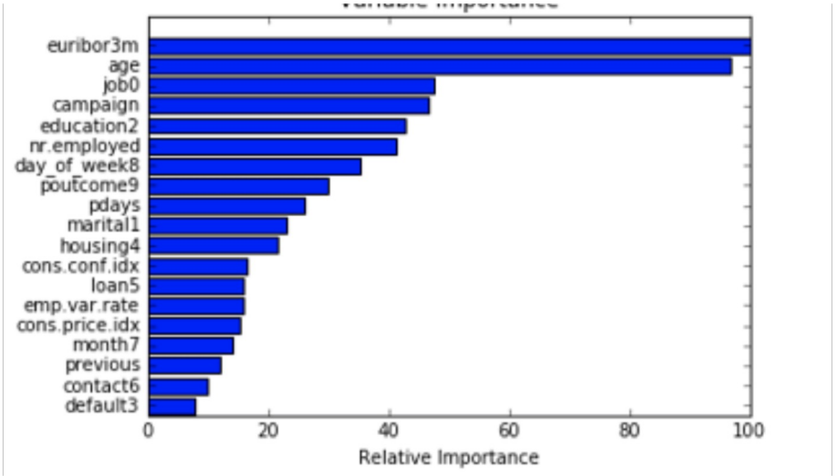
- Pros and cons considerations:
- Flexibility
- Enforce data type
- Compressed
- Fast to access and write back
- Easy to manipulate
- Easy for ML algorithms to access

Modeling Technical Process

- Data Collection
- Data Cleaning
- Feature Engineering or Data Transformation
- Target Definition
- Model Optimization
- Model Validation
- Model Deployment
- Model Monitoring
- Feedback Data Collection
- Iterative Refinement

Variable Selection

- Mutual Information
- Correlation
- Principal Component Analysis
- Model Based Selection
- Leave One Variable Out, or Leave One Covariate Out



Linear Models

- Linear regressions, least squares deviation, absolute deviation

$$\Sigma\,(\,Y - \hat{Y}\,)^2 \quad \text{vs} \quad \Sigma\,|\,Y - \hat{Y}\,|$$

- Assumed functional form, called **Cost Functions or Loss Functions or Objective Functions**
- A systematic optimization process to fit the model parameters
- We can get into a local minimum
- Convergence may be slow or difficult to get

Most Important Part of Modeling Work

It is not modeling, but collecting, cleaning and organizing the data
Find the smoking gun, you don't need circumstantial evidence
Make sure you understand the context of the data collection
Data always have issues, and they are as clean as the efforts you put in to analyze it
Some say 80-90% of the modeling work is data munging
When you make mistakes in your data preparations, you may waste a lot of your modeling efforts
Garbage in, garbage out.
But in this lecture, our focus is on model building

Multiple Linear Models

Cars mpg depend on more than one factors
The simplest multivariate model s are multiple linear models

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$, where β_j's are
β_0 is the intercept and β_j is the slope along x_j direction
$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$,is the model predicted
ϵ_i 's are the errors
x_k may be numeric, binary, nominal predictors
For binary, nominal predictors, indicator variables are used.

Multiple Linear Models

- Multiple linear regressions may use
- least squares deviation or absolute deviation:
  $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ vs. $\sum_{i=1}^{N}|y_i - \hat{y}_i|$
- Assumed functional form, called **Cost Functions or Loss Functions or Objective Functions**
- A systematic optimization process to fit the model parameters

We can get into a local minimum

Convergence may be slow or difficult to get

Model Building

The model building process is to minimize SSE by varying β_0 and β_j's,

$$SSE = \epsilon_i{}^2 = \sum_{i=1}^{N} (yi - \hat{y}_i)^2 = \sum_{i=1}^{N} (yi - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki})^2$$

$$\frac{\partial}{\partial \beta_0}(SSE) = \frac{\partial}{\partial \beta_0}\left(\sum_{i=1}^{N} (yi - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki})^2\right)$$

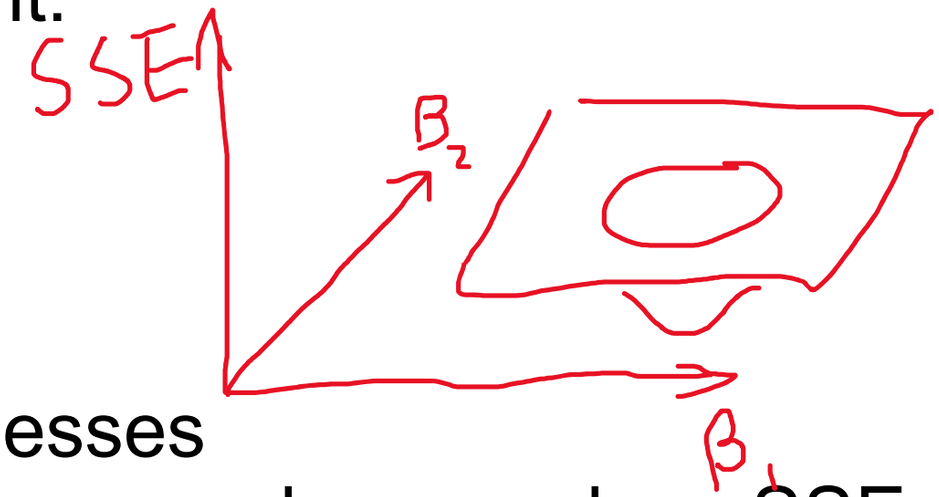
$$\frac{\partial}{\partial \beta_J}(SSE) = \frac{\partial}{\partial \beta_j}\left(\sum_{i=1}^{N} (yi - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki})^2\right)$$

$$\nabla(SSE) = \frac{\partial SSE}{\partial \beta_0} + \frac{\partial SSE}{\partial \beta_1} + \frac{\partial SSE}{\partial \beta_2} + \cdots$$

Derivatives of SSE on Parameters

Sensitivity of SSE to a small change in one parameter while keeping all other parameters constant:

$$\frac{\partial SSE}{\partial \beta_i} \approx \left(\frac{\Delta SSE}{\Delta \beta_i}\right)_{\beta_{j \neq i}}$$



Initially, all parameters are random guesses
We find change of which variable by how much can reduce SSE
We adjust the parameters to minimize SSE
We do this iteratively

Updating Model Parameters

Start with an initial set of parameters
Find the direction and slope of change in parameter that will reduce SSE
Move the parameters by a step in that direction
Loop
Use the current parameters
Find the direction and slope of change in parameter that will reduce SSE
Move the parameters by a step in that direction
Until the minimum is reached
The optimization is called to have converged
At minimum, any movement of any parameter will increase SSE

Multiple Linear Regression Model in R

> lmresmtcars=lm(mpg~am+qsec+wt, data=mtcars)
> summary(lmresmtcarint)

```
> lmresmtcars=lm(mpg~am+qsec+wt, data=mtcars)
> summary(lmresmtcars)

Call:
lm(formula = mpg ~ am + qsec + wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4811 -1.5555 -0.7257  1.4110  4.6610

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6178     6.9596   1.382 0.177915
am            2.9358     1.4109   2.081 0.046716 *
qsec          1.2259     0.2887   4.247 0.000216 ***
wt           -3.9165     0.7112  -5.507 6.95e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared:  0.8497,    Adjusted R-squared:  0.8336
F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Linear Models with Categorical Predictors

```
> lmiris<-lm(Sepal.Length~., data=iris)
> summary(lmiris)
Sepal.Length depends on
  species, but is also proportio
  to Sepal Width, Petal Length
  and Width
> class(iris$Species)
[1] "factor"
```

```
> names(iris)
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
> lmiris<-lm(Sepal.Length~., data=iris)
> summary(lmiris)

Call:
lm(formula = Sepal.Length ~ ., data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.79424 -0.21874  0.00899  0.20255  0.73103

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.17127    0.27979   7.760 1.43e-12 ***
Sepal.Width         0.49589    0.08607   5.761 4.87e-08 ***
Petal.Length        0.82924    0.06853  12.101  < 2e-16 ***
Petal.Width        -0.31516    0.15120  -2.084  0.03889 *
Speciesversicolor -0.72356    0.24017  -3.013  0.00306 **
Speciesvirginica  -1.02350    0.33373  -3.067  0.00258 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3068 on 144 degrees of freedom
Multiple R-squared:  0.8673,    Adjusted R-squared:  0.8627
F-statistic: 188.3 on 5 and 144 DF,  p-value: < 2.2e-16
```

Prediction Using the Model

```
newiris<-data.frame(Sepal.Width=c(3.0, 3.0),Petal.Length=c(5.5,
4.0),Petal.Width=c(2.0, 1.0),Species=c("virginica", "versicolor"))

newiris
  Sepal.Width Petal.Length Petal.Width    Species

1           3          5.5           2  virginica
2           3          4.0           1 versicolor

predict(lmiris,newiris, interval = "confidence")

       fit      lwr      upr
1 6.565966 6.479474 6.652459
2 5.937192 5.794768 6.079615
```

Indicator Variables

Retained: yes|no can be transformed into retained: 1 or 0
Online_shopper: yes|no can be transformed into Online_shopper: 1 or 0
State: “CA”,”AZ”,”NV”,”WA”,”OR”,“NY”…
For 50 values in state, 49 new binary variables are created, leaving one as reference level, represented by 0 value in all 49 binary state

customer	State	State_ca	State_az	State_nv	State_wa	State_or	State_ny
1	CA	1	0	0	0	0	0
2	NV	0	0	1	0	0	0
3	OR	0	0	0	0	1	0
4	NY	0	0	0	0	0	1
5	WA	0	0	0	1	0	0

Indicator Variables

Age_group: “<18”(reference),”18-25”,”26-35”,”36-45”,”46-55”,“>55”

Creating and modifying variables are called data transformation or feature engineering which helps build better models

customer	age	Age_lt18	Age_18t25	Age_26t35	Age_36t45	Age_46t55	Age_gt55
1	18	0	1	0	0	0	0
2	37	0	0	0	1	0	0
3	70	0	0	0	0	0	1
4	15	1	0	0	0	0	0
5	28	0	0	1	0	0	0

Linear Model Parameter Estimation

Sum of squares of estimate

$SSE = \sum(\hat{y} - y)^2$, sum is over all data points in our training data

Where $\hat{y} = \beta_0 + \beta_1 x_1$

We optimization algorithms to find the set of model parameters for which SSE is minimized

Trade-off Between Complex and Simpler Models

When we add more parameters, we increase the degrees of freedom of the model

We often can improve R^2 by including more predictors

But our model may “overfit”, memorizing the training data and don’t generate well when we apply it to new data

We should include a predictor variable in the model only when it can help reduce the sum of squares significantly

AIC is one such metric that include such balanced considerations

https://www.youtube.com/watch?v=4al2LfJz6Q8

Issues of Linear Modeling

Leakage
Such as using same day SP 500 Index to predict stock price
Confounding Variables
Such as “reported age” versus age from birthdates
Biased sampling
For example, we conduct an online survey for online and offline customers
Multicollinearity
An example of correlating new_var = 10*hp + 2*wt + 3*am

Issues of Linear Modeling

Multicollinearity
This can be from dummy variable encoding for all values, one value is redundant
Omit the last dummy variable and called it the reference level
Say gender as a binary variable

person	gender	ind_male	ind_female
1	female	0	1
2	male	1	0

Variable Inflation Factor for Collinearity

Variable Inflation Factor is defined: $VIF = \frac{1}{1 - R_{X_J|X_{-J}}{}^2}$

$R_{X_J|X_{-J}}{}^2$ is the R^2 when we regress X_J using all X_{-J}'s

$$X_J = \sum_{X_i \neq X_j} C_i X_i$$

For example, do a regression for

cyl~disp+hp+drat+wt+qsec+vs+am+gear+carb

get R^2

If R^2 for this regression is close to 1, then there is significant collinearity.

Variable Inflation Factor for Collinearity

Usually VIF should not be larger than 5 to 10
Remove those variable with VIF above 10
library(car)
vif(lm_result)

```
> vif(lmres)
      cyl      disp        hp      drat        wt      qsec        vs        am      gear
15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873  4.648487  5.357452
     carb
 7.908747
> vif(lmresfinal)
      wt     qsec       am       hp
3.964515 3.216021 2.541527 4.922129
```

Variable Selection

```
lmres<-lm(mpg~cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb,data=mtcars)
summary(lmres)
```

```
> lmres<-lm(mpg~cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb,data=mtcars)
> summary(lmres)

Call:
lm(formula = mpg ~ cyl + disp + hp + drat + wt + qsec + vs +
    am + gear + carb, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Variable Selection

Stepwise regression

```
install.packages(“MASS”)
library(MASS)

lmres<-lm(mpg~cyl+disp+hp+drat+wt+qsec+vs+am
                    +gear+carb,data=mtcars)

step<-stepAIC(lmres, direction="both")
step$anova

lmresfinal<-lm(mpg~wt+qsec+am,data=mtcars)
summary(lmresfinal)
```

Variable Selection

step<-stepAIC(lmres, direction="both")
step$anova

```
> step$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

Final Model:
mpg ~ wt + qsec + am


    Step Df   Deviance Resid. Df Resid. Dev      AIC
1                               21   147.4944 70.89774
2  - cyl  1 0.07987121        22   147.5743 68.91507
3   - vs  1 0.26852280        23   147.8428 66.97324
4 - carb  1 0.68546077        24   148.5283 65.12126
5 - gear  1 1.56497053        25   150.0933 63.45667
6 - drat  1 3.34455117        26   153.4378 62.16190
7 - disp  1 6.62865369        27   160.0665 61.51530
8   - hp  1 9.21946935        28   169.2859 61.30730
>
```

Variable Selection

Stepwise regression

```
> summary(lmresfinal)

Call:
lm(formula = mpg ~ wt + qsec + am, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4811 -1.5555 -0.7257  1.4110  4.6610

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6178     6.9596   1.382 0.177915
wt           -3.9165     0.7112  -5.507 6.95e-06 ***
qsec          1.2259     0.2887   4.247 0.000216 ***
am            2.9358     1.4109   2.081 0.046716 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared:  0.8497,    Adjusted R-squared:  0.8336
F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Interaction Terms

```
> lmresmtcarint=lm(mpg~am+qsec+wt+qsec:wt+am:wt,
  data=mtcars)
> summary(lmresmtcarint)
```

```
> lmresmtcarint=lm(mpg~am+qsec+wt+qsec:wt+am:wt, data=mtcars)
> summary(lmresmtcarint)

Call:
lm(formula = mpg ~ am + qsec + wt + qsec:wt + am:wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6264 -1.4660 -0.3559  1.1520  3.9559

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.1094    23.5809  -0.853 0.401568
am           14.0026     3.3918   4.128 0.000334 ***
qsec          2.6831     1.3002   2.064 0.049171 *
wt            6.6931     7.4051   0.904 0.374379
qsec:wt      -0.5401     0.4137  -1.306 0.203141
am:wt        -4.1411     1.1815  -3.505 0.001675 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.057 on 26 degrees of freedom
Multiple R-squared:  0.9023,    Adjusted R-squared:  0.8835
F-statistic:    48 on 5 and 26 DF,  p-value: 2.606e-12
```

Interaction Terms

> lmresmtcarint=lm(mpg~am+qsec+wt+I(am*wt), data=mtcars)
> summary(lmresmtcarint)
Better model!

```
> lmresmtcarint=lm(mpg~am+qsec+wt+I(am*wt), data=mtcars)
> AIC(lmresmtcarint)
[1] 144.3736
> lmresmtcarsfinal=lm(mpg~am+qsec+wt, data=mtcars)
> AIC(lmresmtcarsfinal)
[1] 154.1194
```

```
> lmresmtcarint=lm(mpg~am+qsec+wt+I(am*wt), data=mtcars)
> summary(lmresmtcarint)

Call:
lm(formula = mpg ~ am + qsec + wt + I(am * wt), data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5076 -1.3801 -0.5588  1.0630  4.3684

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.723      5.899   1.648 0.110893
am            14.079      3.435   4.099 0.000341 ***
qsec           1.017      0.252   4.035 0.000403 ***
wt            -2.937      0.666  -4.409 0.000149 ***
I(am * wt)    -4.141      1.197  -3.460 0.001809 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.084 on 27 degrees of freedom
Multiple R-squared:  0.8959,    Adjusted R-squared:  0.8804
F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

Logistic Regression

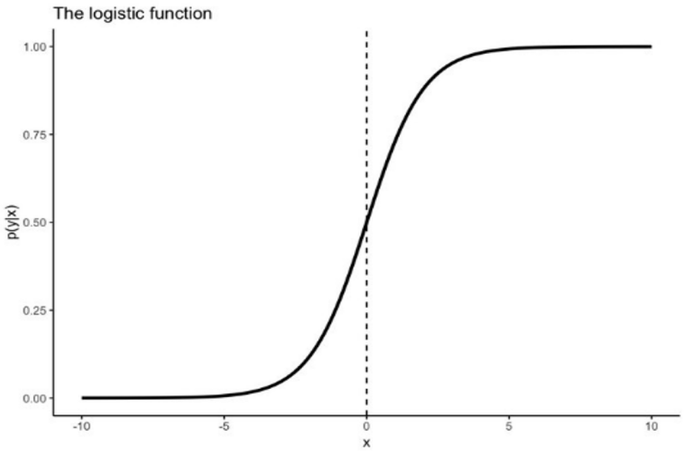
Logit link function

One of Generalized Linear Models

$$\log(OR_i) = \log\frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \ + \cdots + \beta_k x_{ki} + \epsilon_i = \boldsymbol{\beta} \cdot \boldsymbol{x} + \ \epsilon$$

$$OR(\boldsymbol{x}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{x}, \boldsymbol{\beta})}{1 - p(\boldsymbol{x}, \boldsymbol{\beta})} = e^{\boldsymbol{\beta} \cdot \boldsymbol{x}}$$

$$p(\boldsymbol{x}, \boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta} \cdot \boldsymbol{x}}}{1 + \ e^{\boldsymbol{\beta} \cdot \boldsymbol{x}}}$$



Likelihood Function

Given data,

$$\mathrm{L}(y, x, \beta) = \prod_{i=1}^{N} p^{y_i} (1-p)^{1-y_i}$$

If y is 0, then (1 – p) is maximized
If y is 1, then p is maximized
P is the logit link function

$$p(x, \beta) = \frac{e^{\beta \cdot x}}{1 + e^{\beta \cdot x}}$$

https://data.princeton.edu/wws509/notes/c3.pdf
https://data-flair.training/blogs/r-nonlinear-regression/

Likelihood Function

Likelihood estimation, given data minimize negative likelihood of model parameters

$$\mathrm{L}(y, x, \beta) = \prod_{i=1}^{N} \left(\frac{e^{\beta \cdot x}}{1+ e^{\beta \cdot x}}\right)^{y_i} \left(1 - \frac{e^{\beta \cdot x}}{1+ e^{\beta \cdot x}}\right)^{1-y_i}$$

$$= \prod_{i=1}^{N} \left(\frac{e^{\beta \cdot x}}{1+ e^{\beta \cdot x}}\right)^{y_i} \left(1 - \frac{e^{\beta \cdot x}}{1+ e^{\beta \cdot x}}\right)^{1-y_i}$$

$$logL = \sum_{i=1}^{N} \left(log\left(\frac{e^{\beta \cdot x}}{1+ e^{\beta \cdot x}}\right)^{y_i} + log\left(1 - \frac{e^{\beta \cdot x}}{1+ e^{\beta \cdot x}}\right)^{1-y_i}\right)$$

Cost Function is: - Log L or -LL

$$-\frac{\partial}{\partial \beta_j} LogL = -\frac{\partial}{\partial \beta_j} \left(\sum_{i=1}^{N} y_i \beta \cdot x - \sum_{i=1}^{N} y_i log\left(1 + e^{\beta \cdot x}\right) - \sum_{i=1}^{N} (1 - y_i) \log\left(1 + e^{\beta \cdot x}\right)\right)$$

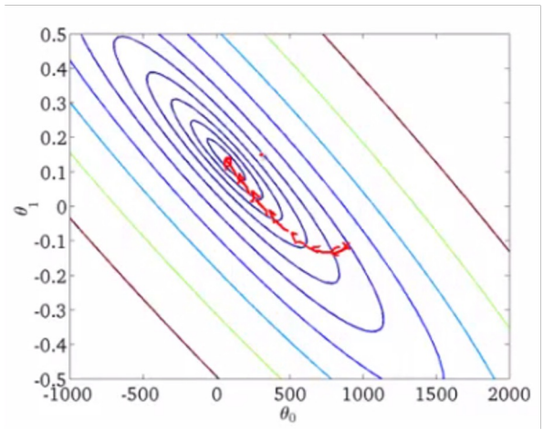
$$= -\frac{\partial}{\partial \beta_j} \left(\sum_{i=1}^{N} y_i \beta \cdot x - \sum_{i=1}^{N} log\left(1 + e^{\beta \cdot x}\right)\right) = -\sum_{i=1}^{N} y_i x_{ij} + \sum_{i=1}^{N} e^{\beta \cdot x} x_{ij} / \left(1 + e^{\beta \cdot x}\right)$$

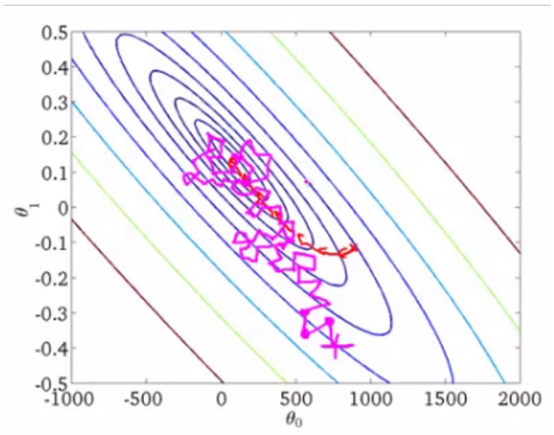
Optimization Algorithms

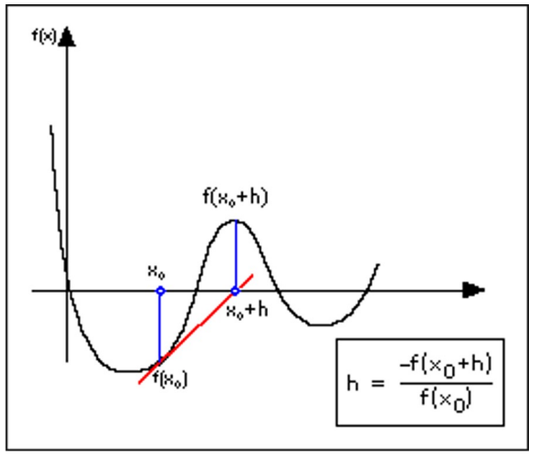
- When we change the parameters, cost function will go up or down
- Just like a rolling hill, and in **high dimension**
- How to get the minimum the quickest?
- Optimization can be very slow but now there is GPU and distributed computing clusters
- Adjust step size
- Go fast in right direction
- Go slow near the minimum
- Avoid going back and forth very fast in one direction but not going downhill globally quickly

Gradient Descent

- On a landscape, the gradient has a magnitude and a direction at each point
- Gradient is maximum when it is steepest going down comparing to all other directions
- Gradient magnitude is the slope of going downhill
- Water flow down the mountain following the path of steepest descent



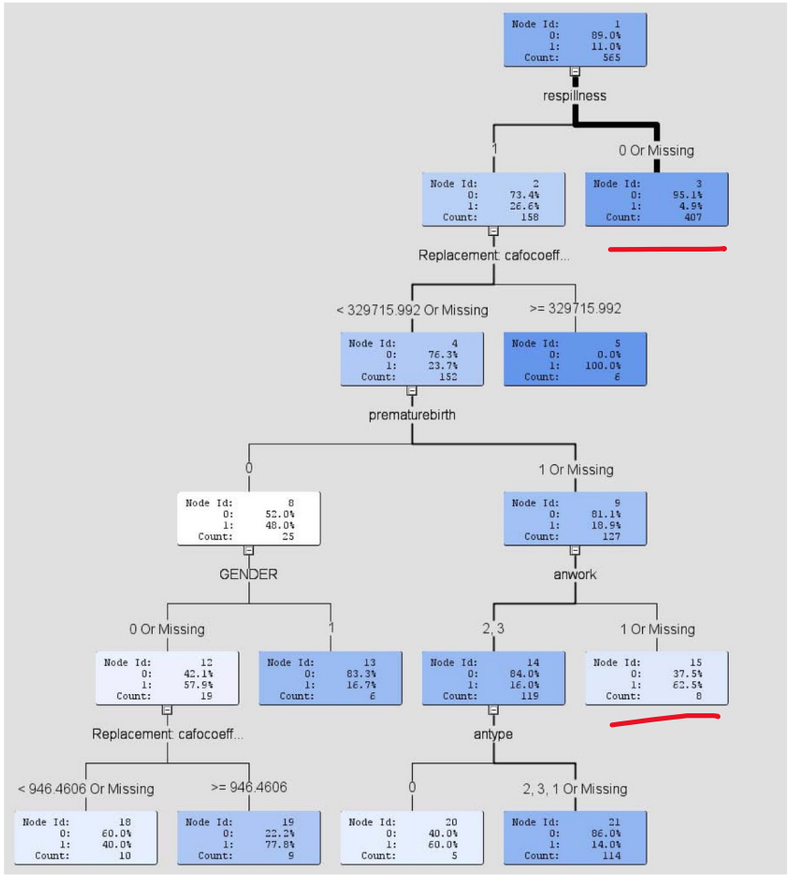


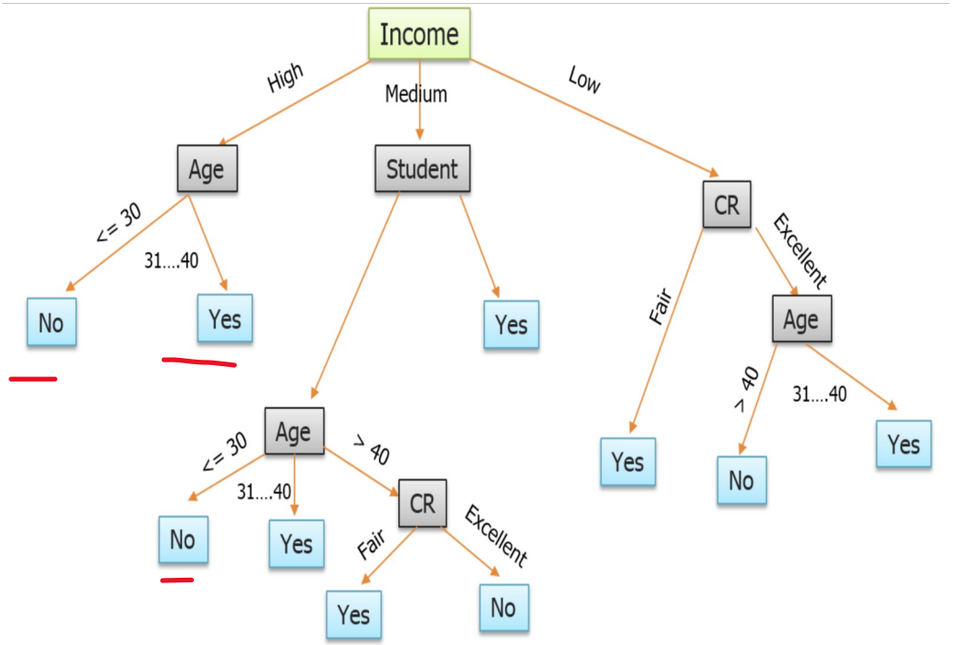


Decision Trees

- **CART**, CHAID, C4.5/C5
- CART, Classification and Regression Trees uses GINI, binary trees
  - Regression tree using Least Squared Deviation or Least Absolute Deviation
- CHAID, Chi-squared Automatic Interaction Detector uses Chi-squared tests
  - Non-binary splits (into multiple branches)
  - Regression using F-tests
- C4.5/C5 entropy difference to split into multiple branches
- Grow trees and prune trees

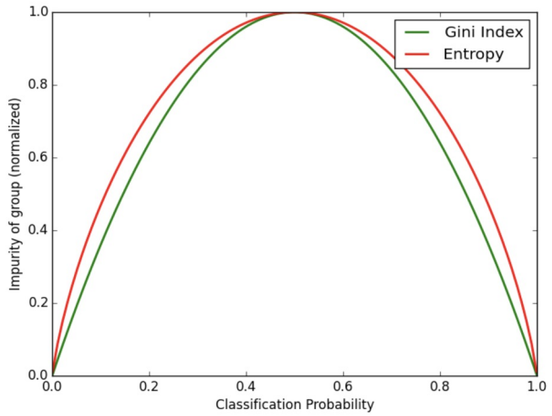
CART and CHAID Decision Tree





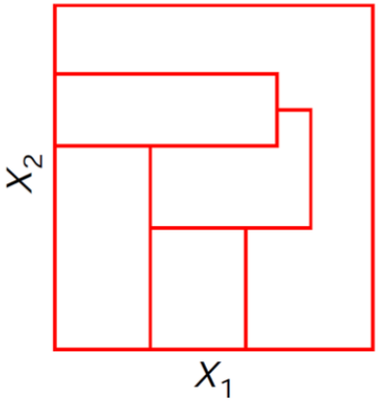
Entropy: $- \sum P(i) \log P(i)$

Gini: $1 - \sum P(i)^2$

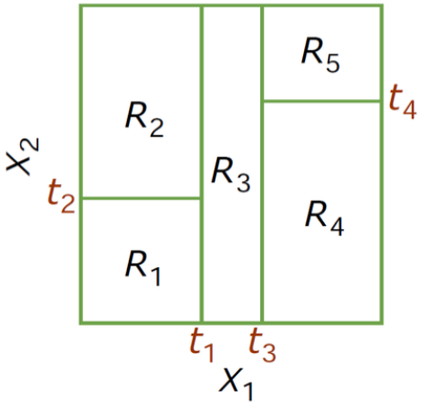


https://support.sas.com/resources/papers/proceedings13/089-2013.pdf

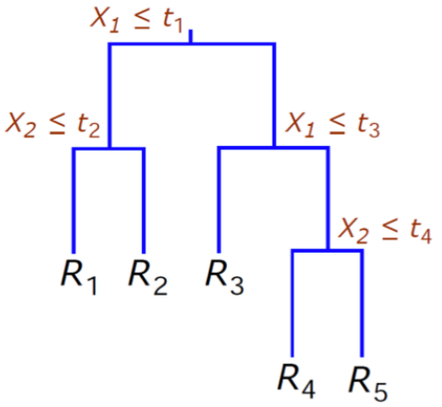
Partitions and CART



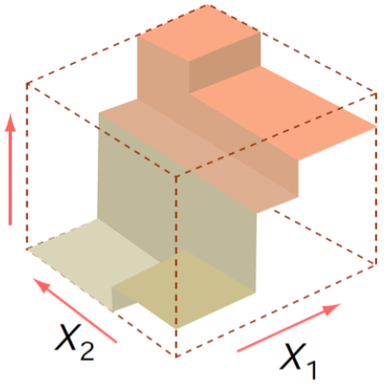
(a) General partition that cannot be obtained from recursive binary splitting.



(b) Partition of a two-dimensional feature space by recursive binary splitting, as used in CART, applied to some fake data.



(c) Tree corresponding to the partition in the top right panel.



(d) A perspective plot of the prediction surface.

http://www.stats.ox.ac.uk/~flaxman/HT17_lecture13.pdf

Decision Trees

- Recursively split data to improve purity
- Find the best predictor and best place to cut
- To get child nodes of higher purity (entropy)
- Successively pick the best variable to split the current node (recursive)
- One variable can be used many times
- It does not assume any distribution
- Reproducibility is ensured by using validation and test data sets.

Decision Trees

- Splits by nature are nonlinear
- Splits to reduce entropy or Gini for purity
- For regressions to reduce least squares deviations
- Tree building processes are greedy
- Simple trees may not be accurate
- Tree building calculations are efficient
- For each split, we scan and search over all predictors
- At most 2^d splits, at maximum depth d

Decision Trees: Pros/Cons

- Simple trees are easy to interpret
- No need to do data transformations that are monotonic functions
- No need to assume any distribution
- Missing data do not break the prediction, may make prediction less accurate
- Tree structure unstable for sample variations
- Bad for linear patterns

Decision Tree Hyper Parameters

- Max tree depth
- Split criteria, Gini, entropy, twoing

  Similar shapes
- Number of child nodes, binary is sufficient
- Min support of leaf nodes
- Cross validation
- Validation data file

Problems of Modeling

- Lack of data, data quality issues
- Inaccuracy, not converging
- Overfitting
- Collinearity
- Data sparsity
- High dimensionality
- Model degrade quickly
- Leakage
- Heteroscedasticity

Leakage examples

- Leakage occurs when in training data information of predictors is present in target variables
- Include market index to predict stock price
- Include data fields available only for customers with behavior to be predicted
- Display ads appear on certain pages, then one can predict that clicks happen on pages of ads
- Customer intent prediction using data of part of the purchasing process
- Data transformation using information of entire data such as normalization or PCA and then use cross validation

Titanic Data Set Example

VARIABLE DESCRIPTIONS

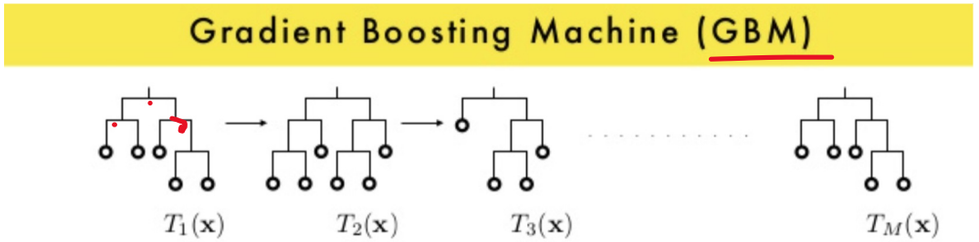
Pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
survival	Survival (0 = No; 1 = Yes)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare (British pound)
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
boat	Lifeboat
body	Body Identification Number
home.dest	Home/Destination

Good ML Algorithms

- Accurate, have model capacity, limited training time, robust
- Easy to tweak model hyper parameters
- Do not require a lot of programming
- Generalize better
- Can scale, training and scoring large data sets
- Missing data, data issues, errors, large numbers of categories
- Outliers
- Prevent overfitting
- Automated feature generation
- Can score new cases quickly

Improving Simple Decision Trees

- Bagging and Ensemble models
  - Take bootstrap samples and build a set of trees, models
- Random Forests
  - Select a sample, and a subset of predictors
- Gradient Boosting Machines
  - Build a sequence of additive shallow trees
- Rulefit
  - Use a set of rules from a set of trees as predictors to fit a linear model



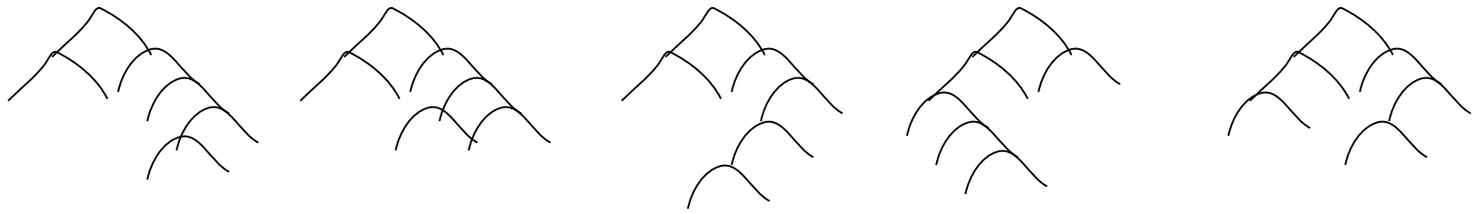
$$f_i(\mathbf{x}) = f_{i-1}(\mathbf{x}) + \lambda T_i(\mathbf{x}; \hat{\Theta}_i)$$

Gradient Boosting Machines

- Build a series of shallow CART trees added together
- Each tree stays once it is built
- Next tree will be built more on where the previous trees don't do well
- Each tree focuses somewhat different parts of the data set
- Weight at each step calculated to minimize loss function
- Each tree is multiplied by a learn rate
- Smaller learn rate limit each tree's contribution to allow slow learning
- Each tree is shallow, to reduce greedy learning

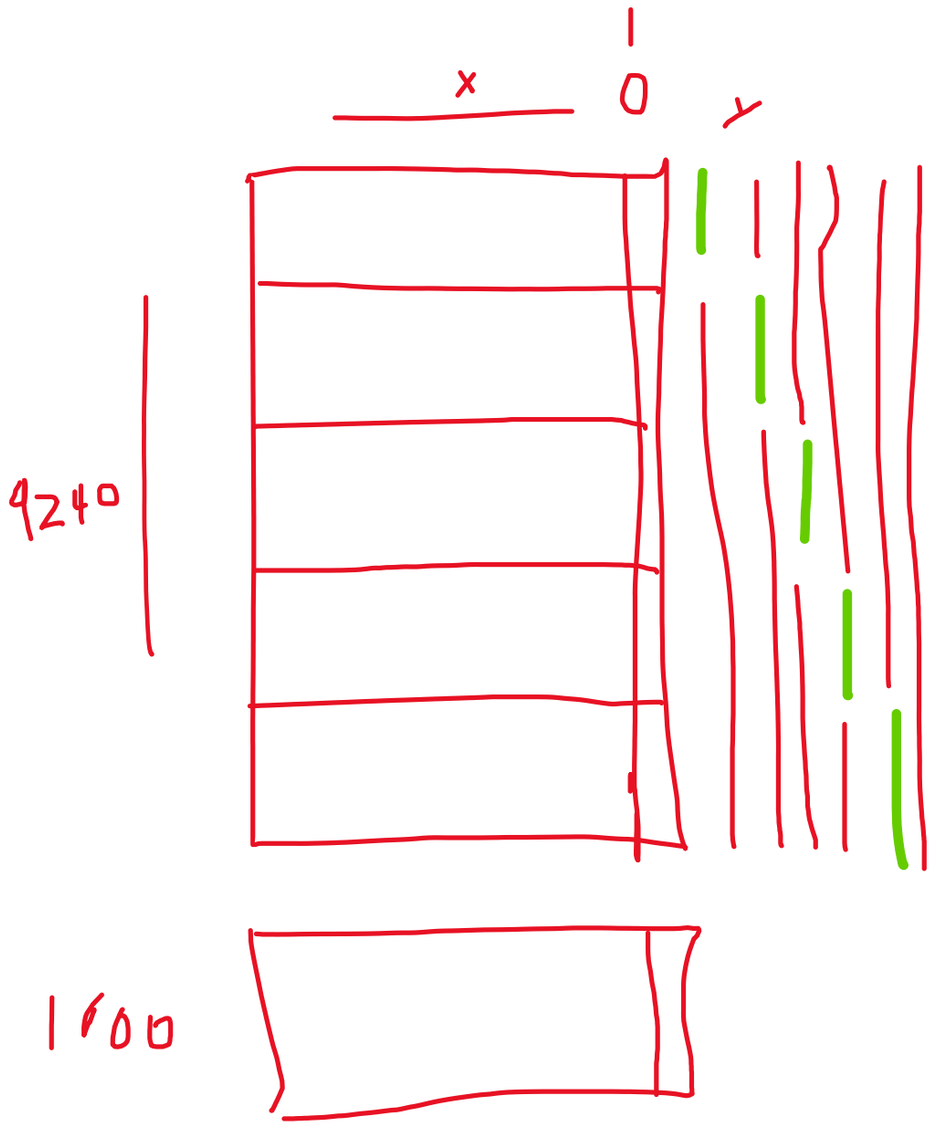
Random Forest

- Random Forest builds an ensemble of independent CART trees
- Each tree is built on a bootstrap sample from the training data
- Each tree only use a fraction of the p predictors, usually sqrt(p)
- Results are averaged or obtained by voting, it is based on a committee of independent predictions
- Usually the number of trees are very large, 100s to 1000s
- Random Forests make estimation based on a principle of Crowdsourcing
- Random Forests are robust, stable and widely used



GBM Hyper Parameters

- Number of trees
- Learn rate
- Max tree depths
- Min support for leaf nodes
- Binomial (Bernoulli) or Gaussian loss function
- Cross validation
- Validation file
- Huber M outlier threshold
- Platt Scaling normalization of distribution



XGBoost

- Another GBM implementation
- Having more regularization to control overfitting
- Weighted Quantile Sketch for rank approximation
- Sparsity aware split for speed
- Utilization of GPU, much faster than CPU
- Work together with max tree size

LightGBM

Work well on large data sets

Leaf wise tree growth

- Grow branches with high impurity reduction first

Gradient-based One-Side Sampling

- Sample biased for less well-trained side, Alpha hyper parameter

Exclusive Feature Bundling (EFB)

- Find similar predictors and bundle them together

Uses GPU

Fast and accurate

Best First Tree

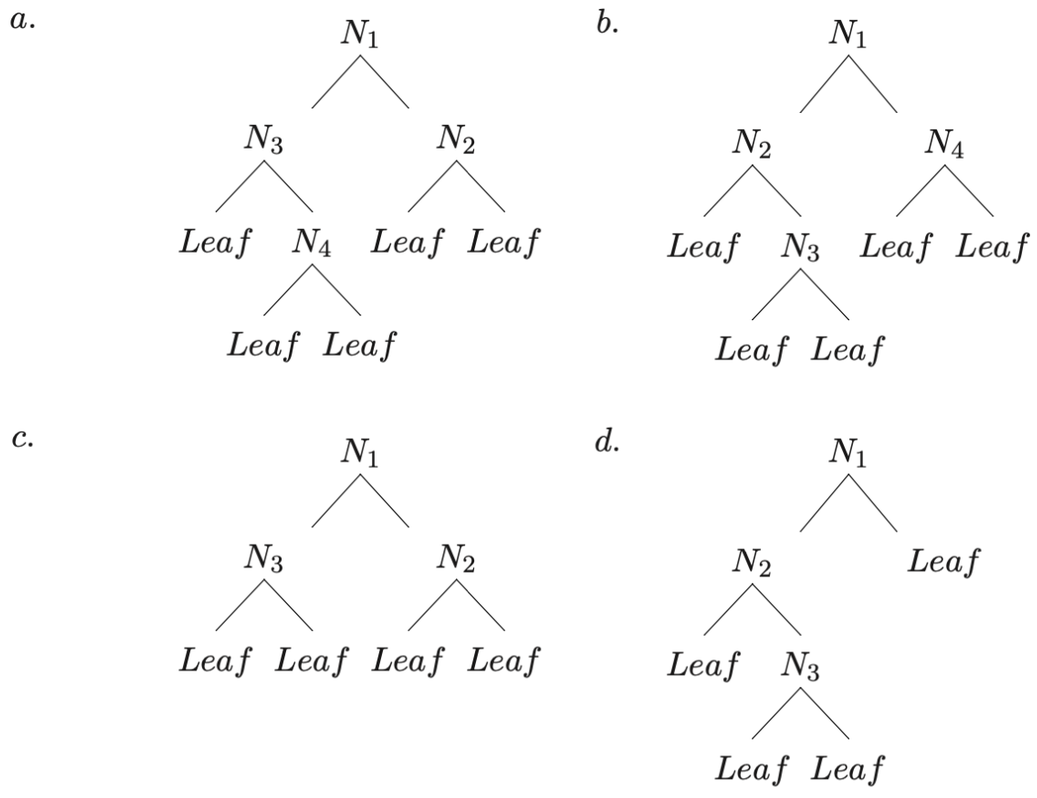


Figure 3.2: (a) The fully-expanded best-first decision tree; (b) the fully-expanded standard decision tree; (c) the best-first decision tree with three expansions from (a); (d) the standard decision tree with three expansions from (b).

Best split has maximum reduction of Gini Index

- https://researchcommons.waikato.ac.nz/handle/10289/2317

CatBoost

- CatBoost implements symmetric trees
- Ordered boosting
  - Use time and if not available, assign data points random time
  - Random Permutations on data sets for time
  - Train on early data points and test on later one
- Train only log number of data, leave one out for testing on error
- Response coding, represent each categorical feature using the mean of target values over all data points
- CatBoost combines multiple categorical features

- https://medium.com/@hanishsidhu/whats-so-special-about-catboost-335d64d754ae

Random Forest Hyper Parameter

- Number of trees
- Max tree depths
- Min support for leaf nodes
- Binomial (Bernoulli) or Gaussian loss function
- Sampling rate
- Predictor sampling rate
- Cross validation
- Validation file

Customer Behaviors

- Recency, Frequency, Monetary, Variety
- Regression to the mean
- Transaction, browsing, marketing responses
- Intent, purpose of shopping
  - Needs, value and quality, research/compare, deal
  - Explore, shopping experience, entertainment

Logistic Regression

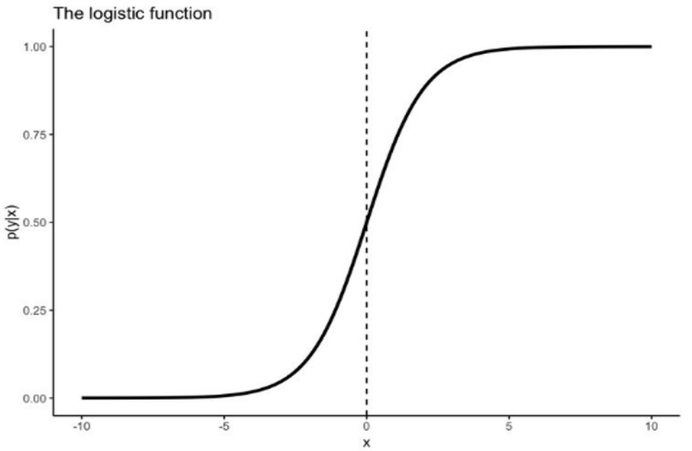
Logit link function

One of Generalized Linear Models

$$\log(OR_i) = \log\frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \ + \cdots + \beta_k x_{ki} + \epsilon_i = \boldsymbol{\beta} \cdot \boldsymbol{x} + \ \epsilon$$

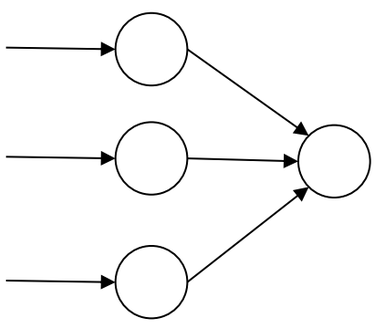
$$OR(\boldsymbol{x}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{x}, \boldsymbol{\beta})}{1 - p(\boldsymbol{x}, \boldsymbol{\beta})} = e^{\boldsymbol{\beta} \cdot \boldsymbol{x}}$$

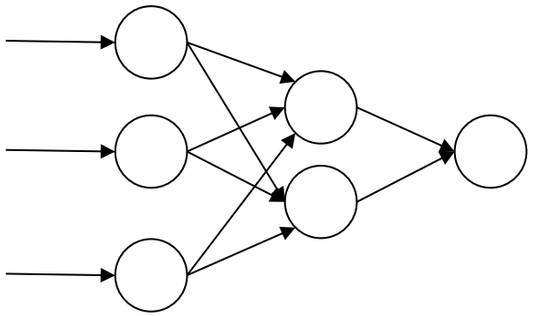
$$p(\boldsymbol{x}, \boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta} \cdot \boldsymbol{x}}}{1 + \ e^{\boldsymbol{\beta} \cdot \boldsymbol{x}}}$$



Improving Logistic Regressions

- Neural networks are a network of logistic regressions, each stage is fed to the next stage
- Each stage can be a layer of many logistic regressions
- Each logistic regression is a neuron
- A sequence of arrays of neurons, one layer feeding to another





Parameters for Logistic Regression

- L1 and L2 regularization, elastic net
- Indicator variables for categorical variables
- Putting continuous variables into some buckets
- Automatically finding interactions
- Neural network architecture
- Layers and number of neurons

L1 Regularization

$$\text{Cost} = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2 + \lambda \sum_{j=0}^{M} |W_j|$$

L2 Regularization

$$\text{Cost} = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2 + \lambda \sum_{j=0}^{M} W_j^2$$

Loss function

Regularization Term

L2, L1 Regularizations

- L2 regularization adds to loss function a fraction of the sum of square of all parameters
- This tend to reduce the size of the parameters, and penalize unnecessarily complex models
- L1 regularization adds the sum of absolute values of coefficients
- A combination of L1 and L2 forms Elastic Net which works well
- Lambda strength, alpha weight of L1 vs L2

Grid Search of Hyper Parameters

Cartesian grid search

Random search

Bayesian search (Python Hyperopt)

Biased Sampling

- In a typical marketing campaign, conversion rate is small
- Most customers do not convert
- How do you predict who is going to convert?
- For example, 1% customers converts, 10K out of 1 Million
- Typical model has accuracy of 80%, if we identify everyone as none converters, the error will be only 1%
- We generally estimate error rates for both converters and none converters
- For each percent of errors, it is 100 converters, but 10K non converters
- We often over sample converters so that model can predict converters better

Biased Sampling

- Customers form distributions in dimensions of age, income, house value
- If we sample from entire population, we make better predictions for typical customers
- Most valuable customers may not be typical, they are under represented in the sample
- We may build separate models for various segments based on demographics or lifetime value

H2O Modeling Platform

- H2O in-memory map reduce and custom compressions
- H2O Demo: 100 million rows and 50 columns, logistic regression on a 16 node cluster, finishes in 11 seconds
- A cluster of 20 machines
- Customer sample of 7 million, 4000+ columns
- Train 100 models, R interface, in two days
- Model accuracy by ROC AUC distributions, robustness
- Score all customers in about 1.5 hours
- Demo of H2O

H2O.ai

H2O.ai's Architecture

- In Memory MapReduce
- Custom compression, columnar
- Use all CPUs and cores on a server and multiple servers in a cluster
- Some algorithms use GPU, if available
- Open source H2O 3
- Commercial product Driverless AI

AutoML in H2O

- The Automatic Machine Learning (AutoML)
- Automates the supervised machine learning model training process.
- The current version of AutoML trains and cross-validates
- a Random Forest (DRF),
- an Extremely-Randomized Forest (DRF/XRT),
- a random grid of Generalized Linear Models (GLM)
- a random grid of XGBoost (XGBoost),
- a random grid of Gradient Boosting Machines (GBM),
- a random grid of Deep Neural Nets (DeepLearning), and
- 2 Stacked Ensembles
- one of all the models, and one of only the best models of each kind.

H2O.ai Installation

- unzip downloaded
- java -jar h2o.jar
- install.packages("/path/h2o-version/R/h2o_version.tar.gz", repos = NULL, type = "source")
- demo(h2o.gbm)

Kaggle: Lead Scoring

- Kaggle Competition Data Set
- https://www.kaggle.com/ashydv/lead-scoring-logistic-regression

Homework

- Read Cases Studies in Big Data Analytics in Online Marketing Cases
- Install h2o.ai open source platform h2o 3
  - https://s3.amazonaws.com/artifacts.h2o.ai/releases/ai/h2o/dai/rel-1.8.9-17/index.html

- This requires installation of java 8
- Mac OS X
  - https://java.com/en/download/help/mac_install.xml

- Windows 10 63-bit
  - https://java.com/en/download/faq/java_win64bit.xml

Assignment

Use H2O Flow Upload Leads.csv, Split the Data into a Training Dataset 60%, a Validation Dataset 20% and a Test Dataset 20%

Train GLM, GBM, Xboost, Random Forest, Auto ML models

Change Hyperparameters for Each Model to Optimize ROC AUC

Compare Models to Identify the Best Model and Hyperparameters

Document the Optimization Efforts and select the best model

RuleFit

Stability of Variable Importance

- Many variables are correlated, so if we remove other predictors, the importance of a variable will increase
- If we use different modeling methods, the variable importance will be different
- When use different metrics, the rank can be different
- LOCO or LOVO, leave one covariate (variable) out
- Often variable importance are different using different models or model hyper parameters
- They are not consistent or stable

Partial Dependence Plots

- Contribution of a predictor averaged over other variables
- Over simplify the relationship
- It does not describe local behavior
- It also depends on distribution and sample size near the point of estimation
- Say there is a big jump in college graduation rate at age 70, which may come from a few data points

Monotonic Constraints

- Model loss function may be flat
- Apparent variable dependence may be unstable
- We may add constraints and still get good models
- If we know one variable should have a positive (or negative) effect consistently
- We can put a monotonic constraint on it
- For example, higher discount should lead to higher sales, or else equal
- Use together with Partial Dependence Plots

Shapley Values

- From Game Theory allocating contribution of each player in a group of players
- Average contribution of a variable A
- We can build two sets of models
- Set 1: Average contribution with a combination of a subset of variables including the variable A in question
- Set 2: Average contribution without the variable A in all possible combination of variables
- The difference in average model accuracy between set 1 and set 2
- Shapley values are consistent given model algorithm

Shapley Values Definition

- A set of M predictor variables has M! combinations
- For each variable A, S is a subset (unordered) of M without A
- Construct one, two, three, …, until M coalitions according to order in a combination, for each coalition S with order, calculate model performance with and without variable A
- Repeat for all the rest of M! combinations
- Shapley value is the average of all differences
- S has |S|! (M-|S|-1)! possible combinations
- For S = {X,Y,Z} has 6, XYZ, XZY, YXZ, YZX, ZXY, ZYX
- Shapley Value for A = $(1/M!)\ \Sigma\ |S|!\ (M-|S|-1)!\ (V(S \cup A) - V(S))$
- sum is over all coalition S not including A

Shapley Ratio Properties

- Efficiency, contributions from variables should add up to the difference
- Symmetry, if two variables have the same contributions in all coalitions, the two variables should be the same
- Dummy variables do not add value
- Additive using average contributions from different models
- https://christophm.github.io/interpretable-ml-book/shapley.html

Monte Carlo Estimation of Shapley

- M! is very large, ~(M/e)^M
- If M=100, M! > 10^100
- Exact calculation of Shapley value is very expensive
- Need to build 2M! models
- Use Monte Carlo algorithm to estimate Shapley values
- Take a random sample of size K << M! combinations, calculate the average of sample
- This gives an approximation of Shapley value, with the error being proportional to $1/\sqrt{(K)}$

Asymptotic Results

- As we include more terms, or as N increases
- The errors will decrease as $\frac{1}{\sqrt{N}}$

Driverless AI by H2O.ai

- Commercial product
- Free academic license
- Feature engineering
- Model interpretation
- Model deployment
- Runs in https://aquarium.h2o.ai

Data and Analytics Related Roles

- Product Engineers
- Service Engineers
- Platform Engineers
- Data Engineers
- Data Warehousing Engineers
- **Business Analysts**
- Reporting Developers
- **Test Analysts**
- **Modeling Analysts**
- Marketing or CRM Managers
- Project owners
- Product owners
- Business owners