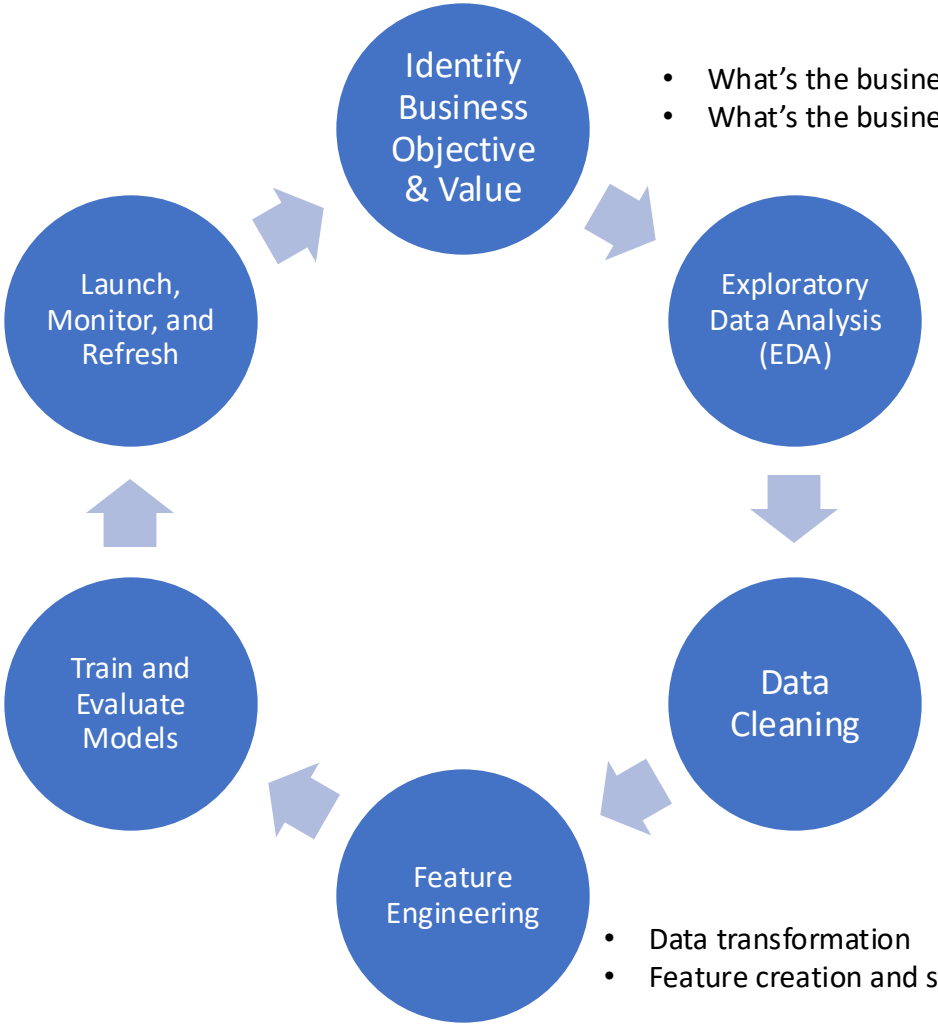
Machine Learning Workflow Overview

Harry Jiannan Wang, Ph.D.

Professor of Management Information Systems

University of Delaware

Machine Learning Workflow



- What's the business goal of this analysis?
- What's the business value if done right

- Data types
- Basic stats
- Visualizations

- Missing data
- Outliers
- Data errors

- Data transformation
- Feature creation and selection

- Cross Validation
- Model Selection
- Parameter Tuning

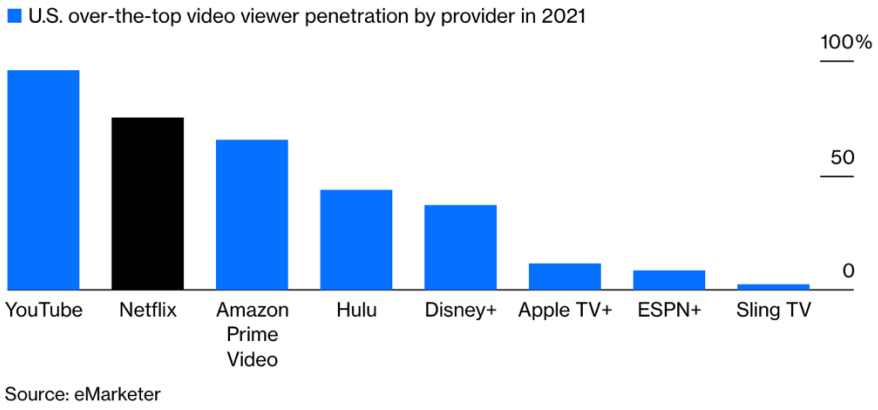
- Launch, Monitor and Analyze Performance
- Collect Additional Data
- Refresh Model

Case 1: Customer Churn Prediction

Customer churn (or customer attrition) : customers stop being a paying client (e.g., cancel the service) during a period of time.

REPORT: APPLE TV+ HAS 20% 'CHURN RATE' COMPARED TO ABOUT 3% FOR NETFLIX

Archived Post Dennis Sellers December 9, 2020



https://www.appleworld.today/2020/12/09/report-apple-tv-has-20-churn-rate-compared-to-about-3-for-netflix/
https://www.bloomberg.com/opinion/articles/2021-01-19/netflix-hbo-max-disney-face-reckoning-with-churn
https://hbr.org/2014/10/the-value-of-keeping-the-right-customers

The Value of Keeping the Right Customers

by Amy Gallo

October 29, 2014

Depending on which study you believe, and what industry you're in, acquiring a new customer is anywhere from five to 25 times more expensive than retaining an existing one. It makes sense: you don't have to spend time and resources going out and finding a new client — you just have to keep the one you have happy. If you're not convinced that retaining customers is so valuable, consider research done by Frederick Reichheld of Bain & Company (the inventor of the net promoter score) that shows increasing customer retention rates by 5% increases profits by 25% to 95%.

Case 2: Click-Through Rate (CTR) Prediction

- Predict whether an online ad will be clicked or not
- Cost-Per-Click (CPC) is one of the most common pricing models for online advertising (CPC, CPM, CPA, etc.)
- Different ad design can lead to different CTR

Top Templates Ranked by CTR

排行	模板	平均点击率
①		1.59 %
②		1.1 %
③		1.09 %

Top Designs Ranked by CTR

https://rainlocal.com/2019/05/09/most-common-pricing-models-for-online-advertising/
https://www.wordstream.com/blog/ws/2017/07/05/online-advertising-costs
https://onlymega.com/blog/increase-your-display-ads-click-through-rate-with-these-hacks

Image sources: Taobao.com

Identify Business Objective and Values

Customer Churn Case

Identify potential churn customers and work on retention

CTR Case

Find insights to help create designs with potential better CTR

Image credits: forbes.com, and gleadz.com

Exploratory Data Analysis (EDA)

"Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods."
- IBM

- data types (number, text, object, etc.)
- continuous/discrete
- basic stats (min, max, std, median, etc.) using boxplot
- Plots: outlier via boxplot, frequency via histogram
- scales and distributions of different features

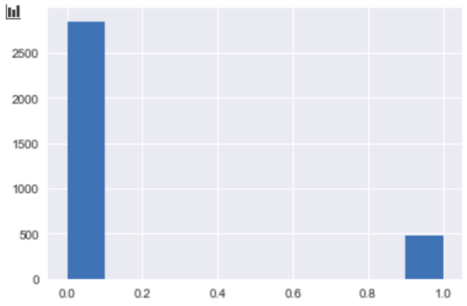
Feature (Attribute)

Target (Label)

Gender	Age	Payment Method	Churn	LastTransaction
male	64	credit card	loyal	98
male	35	cheque	churn	118
female	25	credit card	loyal	107
female	39	credit card		177
male	39	credit card	loyal	90
female	28	cheque	churn	189
female	21	credit card	loyal	102
male	48	credit card	loyal	141
female	70	credit card	churn	153
male	36	credit card	loyal	46

Data Point

Data Type



Loyal vs. Churn

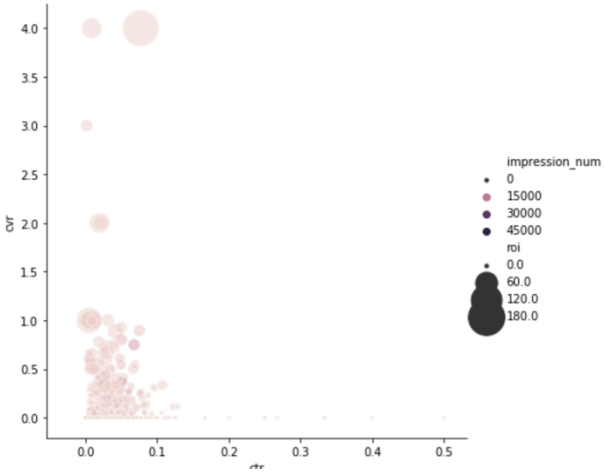
	impression_num	ctr	cvr	roi
count	1000.000000	1000.000000	1000.000000	1000.000000
mean	1051.163000	0.028546	0.088156	2.131125
std	2520.817927	0.036047	0.268881	7.623011
min	1.000000	0.000000	0.000000	0.000000
25%	61.750000	0.007915	0.000000	0.000000
50%	248.500000	0.022727	0.000000	0.000000
75%	936.000000	0.039281	0.066667	1.750409
max	35961.000000	0.500000	4.000000	168.251874

Basic CTR Stats

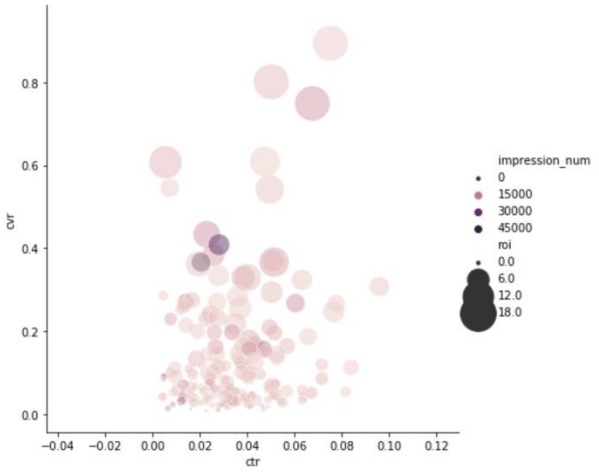
https://www.ibm.com/cloud/learn/exploratory-data-analysis

Data Cleaning

- Avoid GIGO (Garbage In Garbage Out)
- Remove irrelevant data (e.g., deleting ads, sponsorships from podcast transcript for better summarization algorithm training)
- Handle missing data
- Detect outliers
- Correct data errors



Impression > 1000
0<CTR<1



	Missing Data	Outliner	Data Error
Example	31 out of 58 columns have missing data	CTR > 0.7	CVR > 1
Solution	• Remove • Fill (Avg, Next Value, Predicted Value)	• Remove • Analyze and Fix	• Remove • Analyze and Fix

Feature Engineering

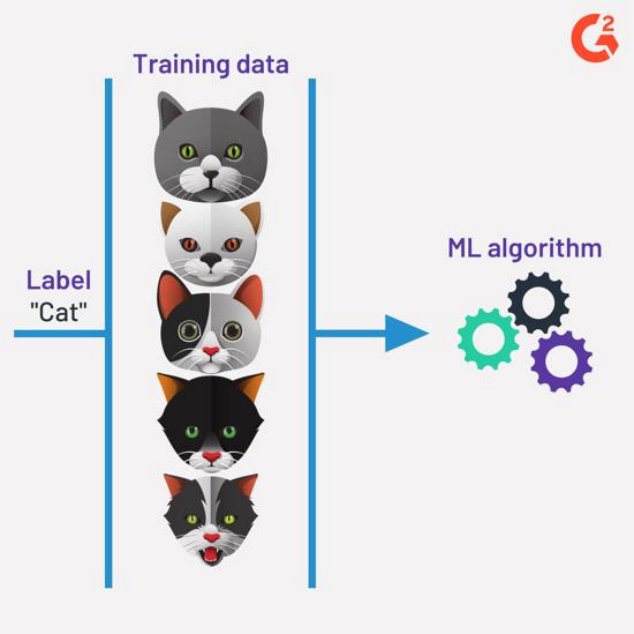
- Data Transformation
  - Numerical features
    - Binning
    - Log and power transformation
    - Scaling/Normalization
  - Categorical features
    - Encoding (ordinal, one-hot, etc.)
- Feature Extraction
  - Image feature extraction
  - Image embedding
  - Text embedding
- Feature Selection
  - Dimensionality reduction
  - Feature correlation and importance

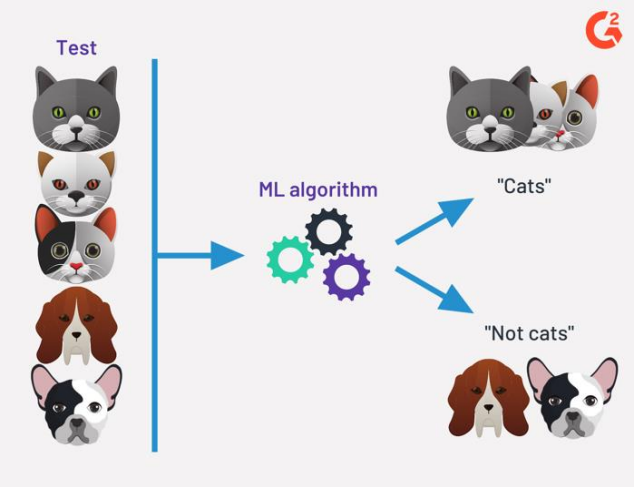
Feature	Left Design	Right Design
# of People	1	1
# of Products	1	4
Logo	Yes	Yes
3B	Beauty	Baby
# of Words in Main Tile	7	2
Main Color	blue	white

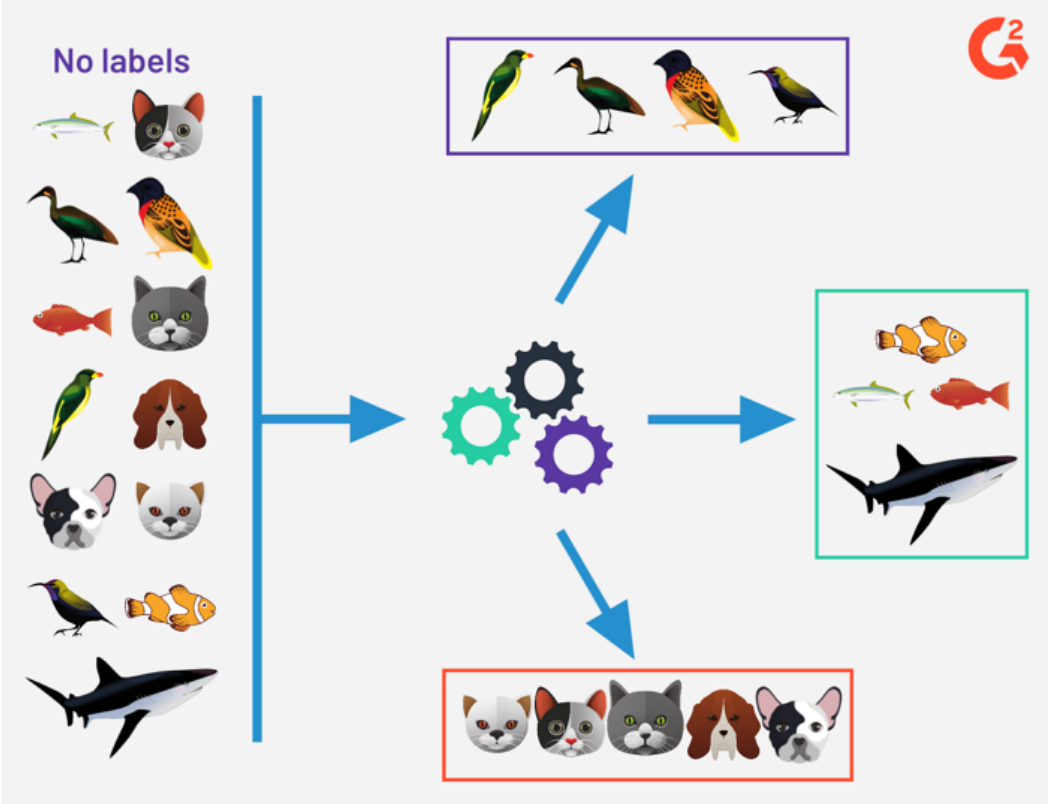
Modeling

- A model is a pattern, is a *summarization* of relationships in the data.
  - Younger customers are more likely to churn
  - Low GPA students are likely to switch majors
  - If Balance >= 50K and Age > 45, then loan default = 'no'
  - Default amount = 0.001*Income+2*Age
- Machine learning algorithm categories
  - Supervised learning, e.g., regression, classification
  - Unsupervised learning, e.g., clustering, association mining
  - Reinforcement learning, e.g., multi-armed bandit, Q-learning

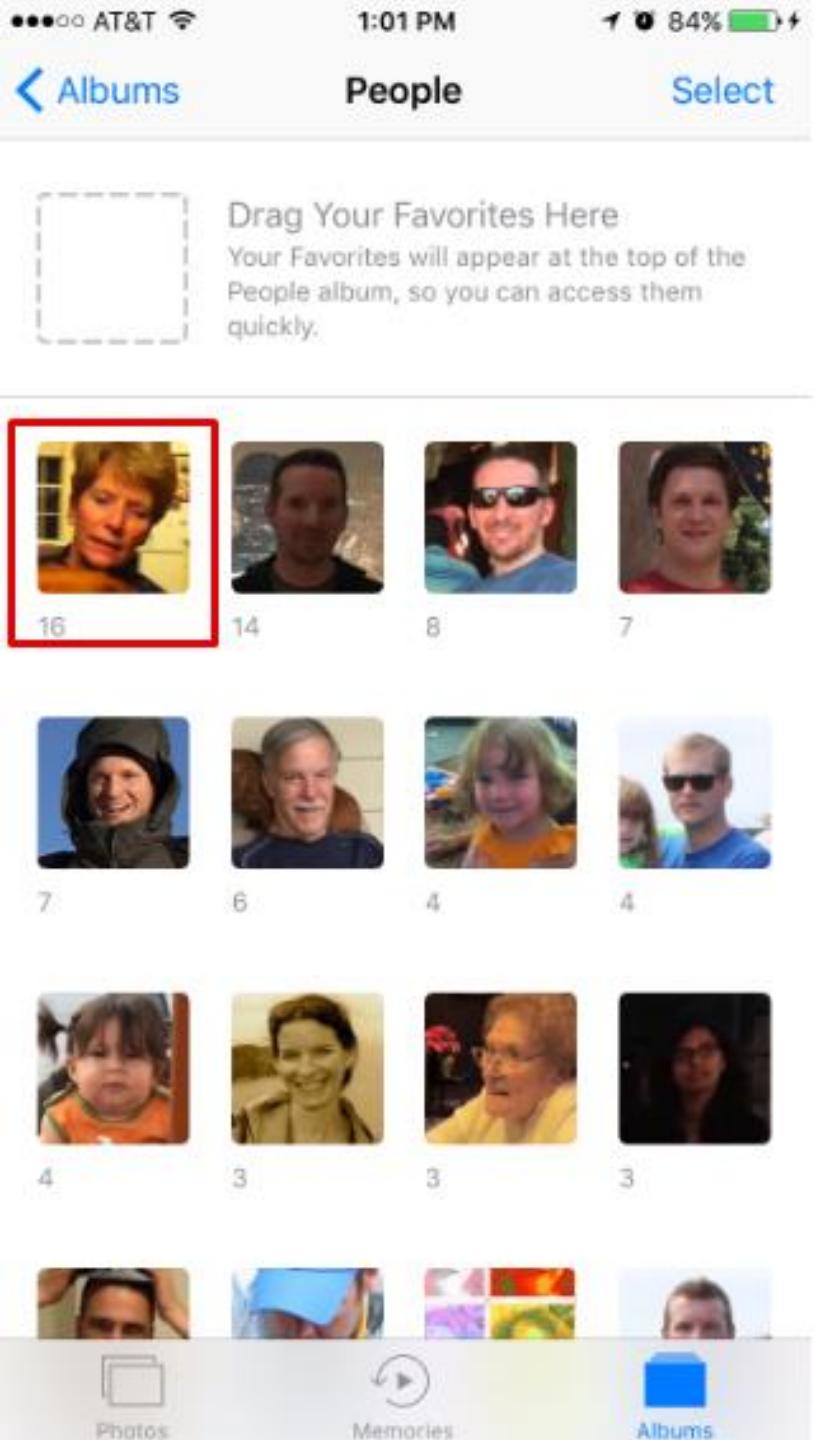
Supervised vs. Unsupervised Learning





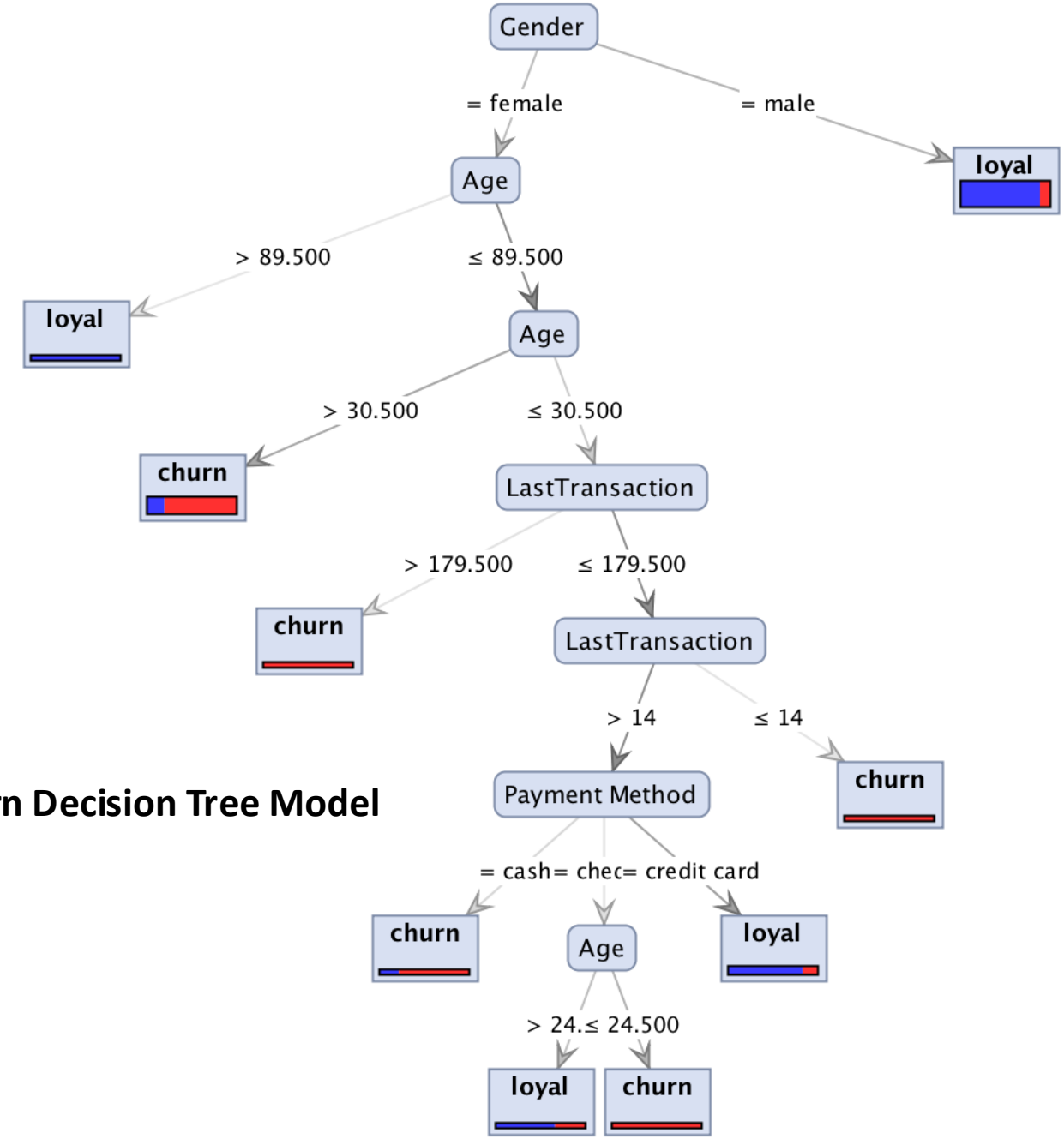


https://learn.g2.com/supervised-vs-unsupervised-learning



Supervised?
Unsupervised?

Customer Churn Decision Tree Model



Model Evaluation, Tuning, and Selection

- Dataset is randomly divided to three subsets to evaluate the model
  - Training set used to build predictive models
  - Validation set used to assess the performance of the models, fine tune models' parameters, and select the best-performing model
  - Test set used to assess the likely future performance of a model. If a model fit to the training set much better than it fits the test set, overfitting is probably the cause, e.g., a very complex decision tree
- Choose different algorithms to build the model
- Performance of different combinations of model parameters should be evaluated to select the best performing model

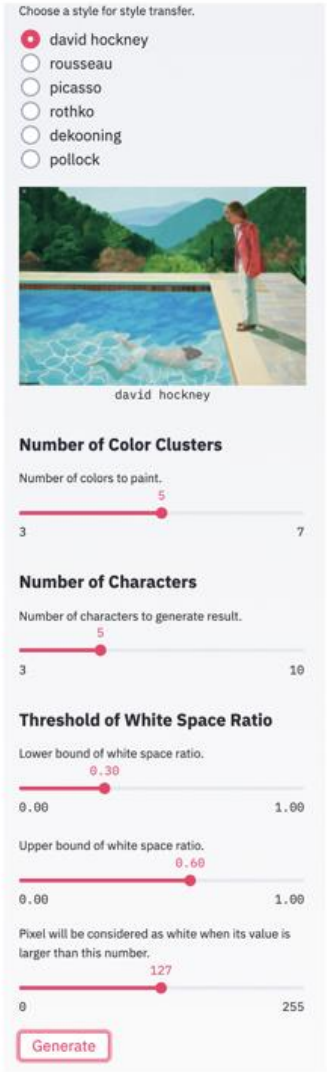
http://harrywang.me/2020/05/09/calligraphy.html

Data Apps

- Tools like Streamlit and Dash allows data scientists to quickly develop interactive data apps to showcase the models

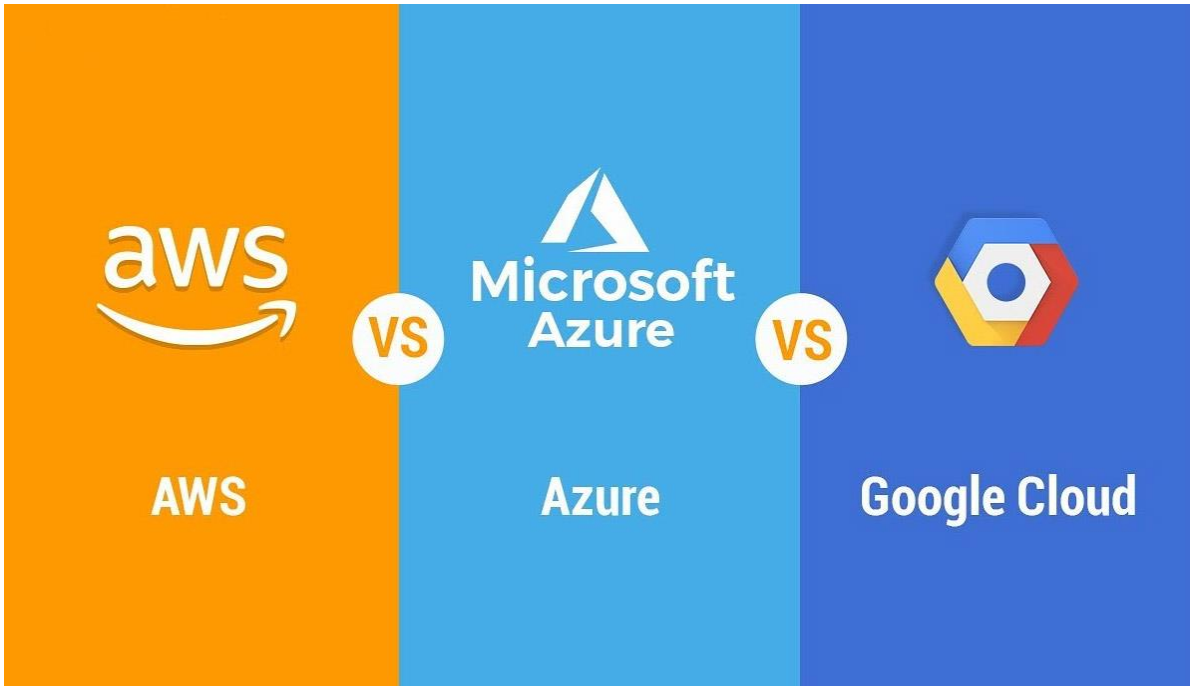
Streamlit

gradio



Deployment

- Deployment turns the model to the “live” mode for prediction
- Model packaging and optimization
- Model serving (self hosting vs. cloud vendors)
- Periodic model updates and redeployment



A Simple E2E ML Workflow Demo and Exercise

- Make sure to run the following commands in **base virtual environment** to install required packages (numpy, pandas, matplotlib, scikit-learn)
- Type up and Run churn-mini.ipynb (don't worry about the details for now – just want to show how you can build a machine learning model to do prediction with less than 20 lines of code)