Validating, Feature Engineering, and Other Improvements

David Quigley
CSCI 5622
2021 Fall

COMPUTING AND SOFTWARE

Career & Internship Fair

Tuesday, Oct. 5

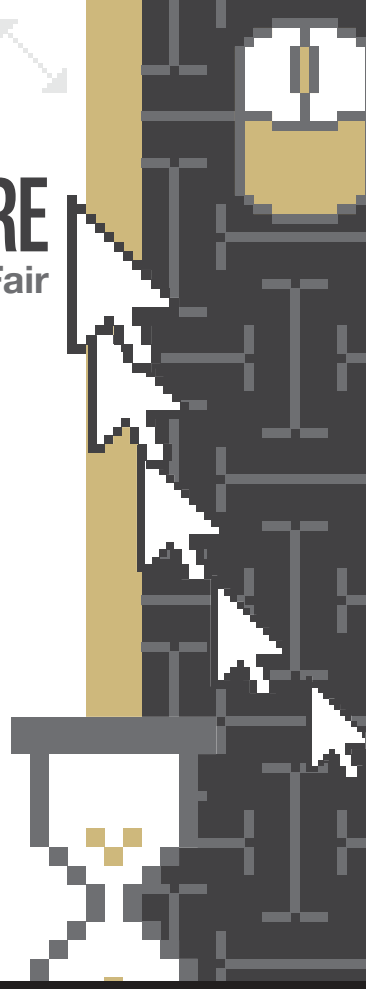
11 am - 4 pm

VIRTUAL: On Handshake

COMPLETE YOUR PROFILE THEN SIGN UP!

Employers from a variety of companies will be recruiting for computer science, software development or engineering and all levels. Speak with engineers one-on-one and learn about a broad spectrum of work opportunities.

colorado.edu/career/fairs-events



Career Services
UNIVERSITY OF COLORADO BOULDER

ProReady

colorado.edu/career

- Accenture
- Alarm.com
- Altia
- Amazon
- AMERGINT Technologies
- Big Compass
- CACI
- Caterpillar Inc.
- Cigna
- Cloud Campaign
- Comcast - Central Division
- CommScope
- Congruex
- CoreLogic
- Credit One Bank
- Danaher Corporation
- DISH
- Encompass Technologies
- Epic
- Expedia Group
- Faegre Drinker Biddle & Reath LLP
- FIS
- Garmin
- Hill AFB Civilian Engineering
- ICR, Inc.
- KPMG LLP
- L3Harris Technologies
- Lennox International
- Lucid Software
- Manifold
- MiTek Inc.
- National Security Agency (NSA)
- Nelnet
- NetApp
- Omitron, Inc.
- Palski & Associates, Inc.
- Progressive Insurance
- Qualcomm
- Raytheon Technologies
- Real-Time Innovations (RTI)
- realtor.com
- Ricoh USA, Inc
- RSM US LLP
- Sandia National Laboratories
- Seagate Technology
- Silicon Labs
- Spectrum
- Splunk
- TASC
- Trimble, Inc.
- Visa, Inc.
- Workiva
- Xilinx

Selected Employers Seeking Computing Talent:

Blue Horseshoe Solutions, Inc.
Capitalize
Colorado Dept of Transportation
Comcast - Central Division
CONMED
Credit One Bank
Deloitte
Esri
Fast Enterprises, LLC
Hill AFB Civilian Engineering
Hitachi ABB Power Grids
Holland & Hart LLP
Idaho National Laboratory
Jacobs
Johns Hopkins University Applied Physics
Keck Graduate Institute
Keysight Technologies
L3Harris Technologies
Lockheed Martin
Lumen
MIT Lincoln Laboratory
National Security Agency (NSA)
NetApp
Parsons
Procter & Gamble (P&G)
Raytheon Technologies
RSM US LLP
Sandia National Laboratories
Seagate Technology
Silicon Labs
SK hynix NewCo -- New Storage Solutions Company
Space Dynamics Laboratory
Spectrum
VMware, Inc.
Xcel Energy

Currently 79 total employers, see the full list on Handshake

Course Logistics

- Project Phase 2: Due 9/30 (Thursday!)
- Project Phase 2.1: Due 2(ish) weeks after Phase 2
  - i.e. Approximately 10/14 (we will distribute them manually and it will be due two weeks later).
- Problem Sets: ~~Problem~~ Set 1 Feedback expected Thursday 9/23
  - ~~Currently anticipating a minor delay.~~ Today 11:59pm
- Problem Set 2: Due 10/7

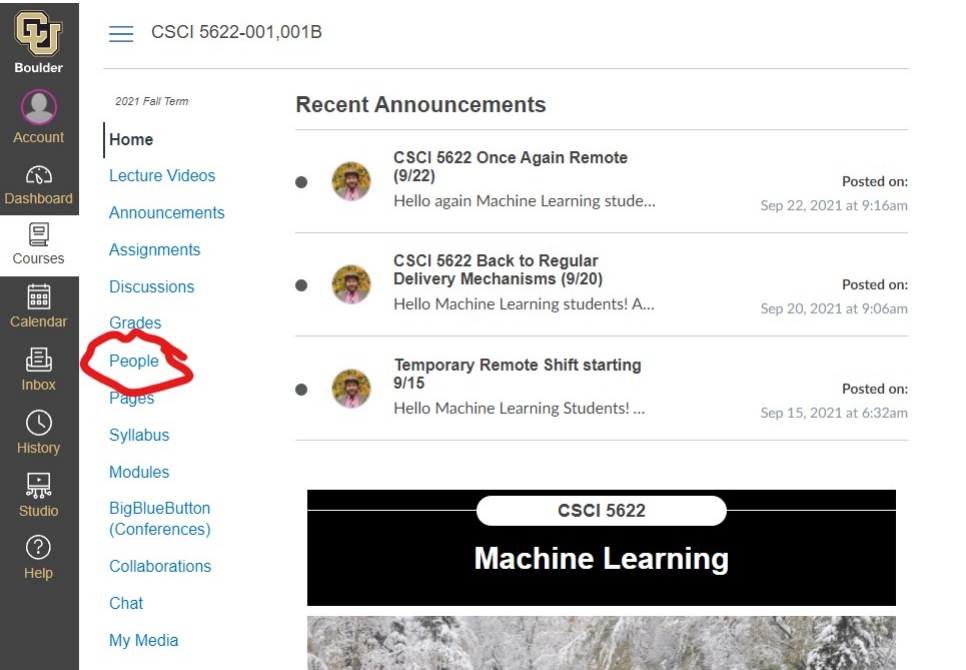
Who Will I Review For Phase 2.1?

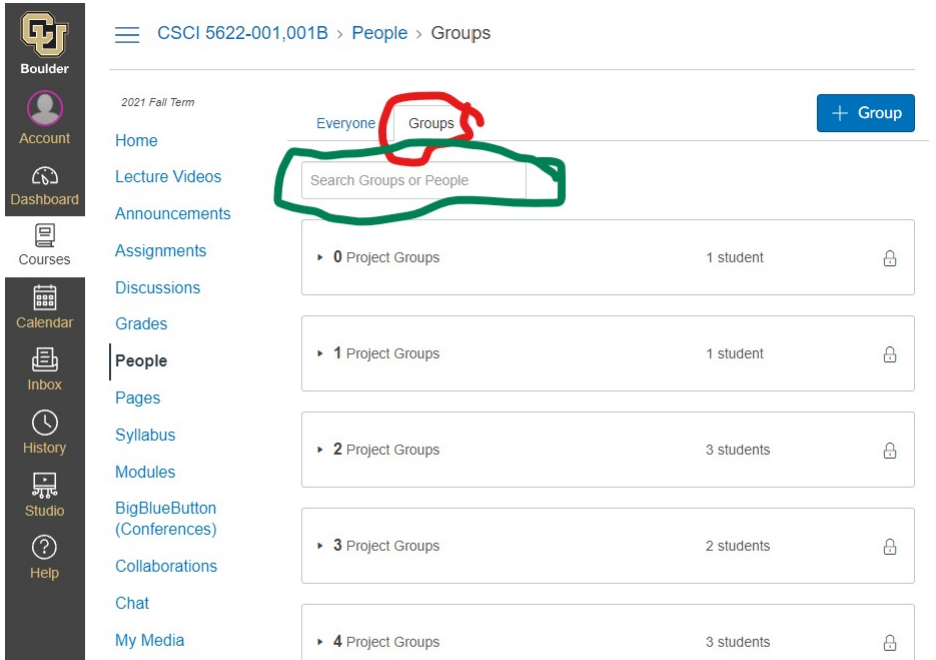
- <Your Project Group> + 1
- Except Group 26 will be reviewing Group 0 and Group 27 will be reviewing Group 26 (it's complicated, I'll email you)

What's <Your Project Group> For Me?

Viewing Project Groups

- Canvas → People (Blue Link on Left) → Groups (Tab on Top) → Search (Yourself)





But Also, Don't Worry!

- In Class:
  - We watch everyone's pitch videos
  - We ask questions about them
  - Hopefully you're there to discover what kinds of questions get asked about your projects – it will help improve your project
  - Hopefully you're there to get a head start on 2.1
- After Class:
  - We (the Instructional Team) will distribute videos and papers from authors to reviewers.
  - You will have two weeks (upon receiving digital copies of the materials, even if it isn't until Friday or the weekend) to review these materials and provide feedback.

Resources for Your Projects

- Some Groups have been asking about computing power above and beyond your laptops – we have that available!
  - Campus-wide Systems Engineering – OIT
    - Should be self-serve for basic needs
  - CS-Specific OpenStack Deployment
    - Email Me, I just put in a request for an instance for this course we should be able to share

Where We Left Off: Feature Engineering

Sequential Data

Sequential Data

I have a series of inputs that I want to combine to do classification

- We have a series of locations, can we gauge trajectory?
- We have 100 readings from an accelerometer, can we get gait?
- We have recent weather readings, can we better predict it?

Sequential Data

I have a series of inputs that I want to combine to do classification

- We have a series of locations, can we gauge trajectory?
- We have 100 readings from an accelerometer, can we get gait?
- We have recent weather readings, can we better predict it?

For those who need advanced techniques quickly for their projects, consider one of the textbooks for the class, Bayesian Reasoning & Machine Learning Ch. 23 - 26

Sequential Data

Timestamp	**User**	***Action***
100	David	*Copy*
150	David	*Paste*
200	David	*Copy*
250	David	*Paste*

Sequential Data – Past Actions as Features

Timestamp	**User**	**PrevAction**	***Action***
100	David	Login	*Copy*
150	David	Copy	*Paste*
200	David	Paste	*Copy*
250	David	Copy	*Paste*

Sequential Data – Past Actions as Features

Timestamp	**User**	**PrevAction**	***Action***
100	David	Login	*Copy*
150	David	Copy	*Paste*
200	David	Paste	*Copy*
250	David	Copy	*Paste*

Can this scale? Does this give us enough info?

Sequential Data – Past Actions as Features

Timestamp	**User**	**N-2Action**	**PrevAction**	***Action***
100	David	N/A	Login	*Copy*
150	David	Login	Copy	*Paste*
200	David	Copy	Paste	*Copy*
250	David	Paste	Copy	*Paste*

You could do this forever...

Curse of Dimensionality

+samples

Features have a *value* and a *cost*

- Value: benefit to your training / accuracy
- Cost: Computation time, memory
  Time & Effort to collect

If you're going to include a feature, it better be worth it!

Lots of ways to evaluate / measure the *usefulness* of a feature...

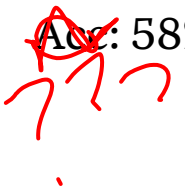
https://www.kdnuggets.com/2017/04/must-know-curse-dimensionality.html

Curse of Dimensionality – Leave-one-out

Train & evaluate your classifier while leaving out one feature at a time.

~~Log Sq Ft~~	# Bedrooms	House Lat.	House Long.	*Cost*
~~3.2~~	4	121.33	47.34	*500K*
~~3.09~~	3	121.33	55.23	*450K*
~~2.87~~	2	121.33	55.34	*200K*
~~3.06~~	2	130.99	47.34	*1500K*

Acc: 58%



Curse of Dimensionality – Leave-one-out

Train & evaluate your classifier while leaving out one feature at a time.

Log Sq Ft	**~~# Bedrooms~~**	**House Lat.**	**House Long.**	***Cost***
3.2	~~4~~	121.33	47.34	*500K*
3.09	~~3~~	121.33	55.23	*450K*
2.87	~~2~~	121.33	55.34	*200K*
3.06	~~2~~	130.99	47.34	*1500K*

Acc: 68%

Curse of Dimensionality – Leave-one-out

Train & evaluate your classifier while leaving out one feature at a time.

Log Sq Ft	**# Bedrooms**	~~**House Lat.**~~	**House Long.**	***Cost***
3.2	4	~~121.33~~	47.34	*500K*
3.09	3	~~121.33~~	55.23	*450K*
2.87	2	~~121.33~~	55.34	*200K*
3.06	2	~~130.99~~	47.34	*1500K*

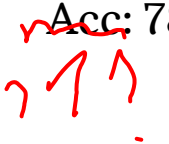
Acc: 65%

Curse of Dimensionality – Leave-one-out

Train & evaluate your classifier while leaving out one feature at a time.

Log Sq Ft	**# Bedrooms**	**House Lat.**	~~**House Long.**~~	***Cost***
3.2	4	121.33	~~47.34~~	*500K*
3.09	3	121.33	~~55.23~~	*450K*
2.87	2	121.33	~~55.34~~	*200K*
3.06	2	130.99	~~47.34~~	*1500K*

Acc: 78%



Curse of Dimensionality – Feature Optimization

Select-K-Best – Leave out all but K features, with the K features being the highest scoring on your test

Selecting to a threshold – you can define a threshold of some sort to determine what features are “good enough”

https://scikit-learn.org/stable/modules/feature_selection.html

Curse of Dimensionality - Sparse Data

Rating-GoT	Rating-BB	Rating-F	Rating-AD
0	0	0	0
4	2	5	4
5	0	0	0
0	5	0	0

Curse of Dimensionality – One-hot Encoding

How many “colors” are there?

- in your crayon box?
- in computational space (256 * 256 * 256)
- for a dog?

What benefits do you get from granularity of color?

What problems do you get from granularity of color?

Categorical Data – Feature Hashing

1) Choose however many features you're willing to accept (n)
    - based on size constraints, etc.
2) Create a unique hash function to encode your categories
    - Each one should be unique

Now you have a set of n-dimensional vectors that represent the variable*

*concerns include collisions, interpretation from classifiers

Categorical Data – Feature Binning

Even if they don't have *order*, features may have *clusters*.

- Copy / Paste / Cut are "edit" actions
- Typing letters are "generate" actions

Put these items into clusters before you encode these features!

This should be motivated by some theory about your data / problem!

Curse of Dimensionality - Sequential Data

How far back do we want to keep our window?

Do we want to keep our expanding feature set?

Sequential Data - Categorization

1) Build a separate category for each possible action sequence (or action cluster, see binning)

    1) Edit, edit edit; edit, edit, generate; edit, generate, edit; ...

2) Build a separate feature for each kind of sequence.

Evaluating Models

Back to it!

Types of Errors (Week 2)

Classified As *Ground Truth*	**C**	**~C**
C	True Positive (Hit)	False Negative (Miss)
~C	False Positive (False Alarm)	True Negative (Correct Rejection)

Types of Errors

Predicted Positive Rate (Precision) = Hits / (Hits + False Alarm)

Classified As *Ground Truth*	**C**	**~C**
C	True Positive (Hit)	False Negative (Miss)
~C	False Positive (False Alarm)	True Negative (Correct Rejection)

Types of Errors

True Positive Rate (Sensitivity, Recall) = Hits / (Hits + Miss)

Classified As *Ground Truth*	**C**	**~C**
C	True Positive (Hit)	False Negative (Miss)
~C	False Positive (False Alarm)	True Negative (Correct Rejection)

Types of Errors

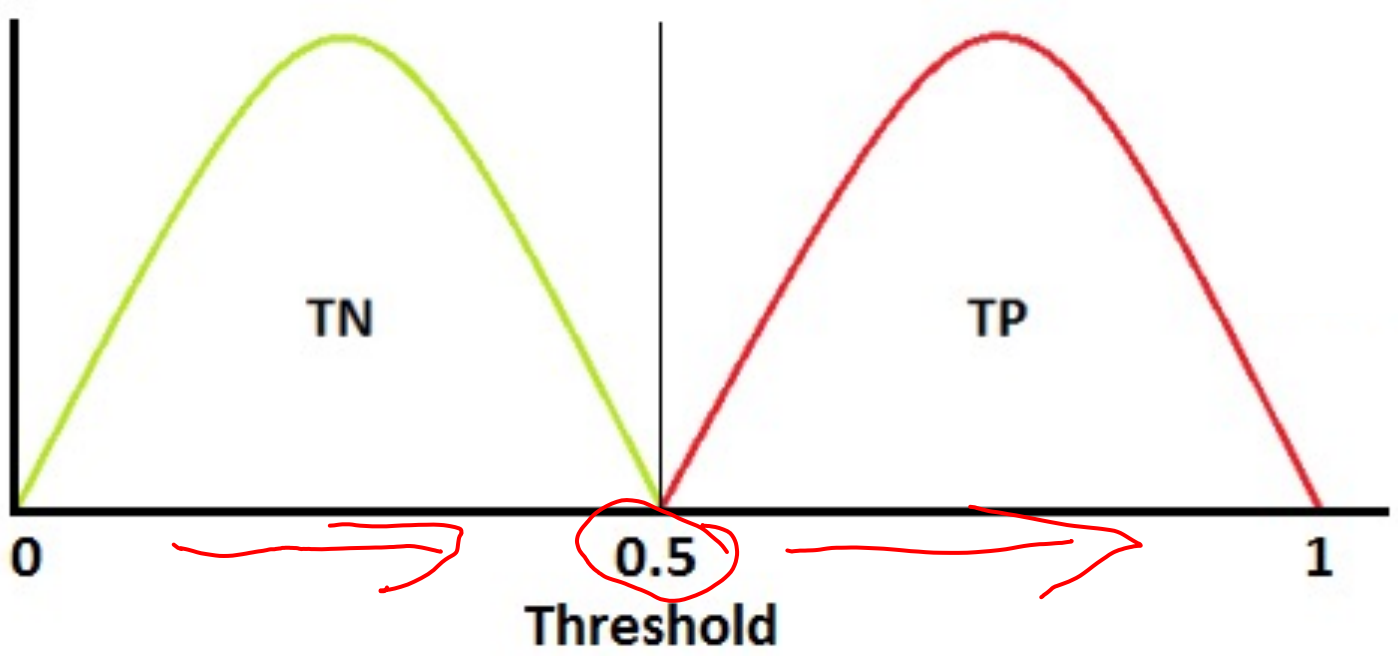
False Positive Rate (Specificity) = Corr. Rej. / (False Alarm + Corr. Rej.)

Classified As *Ground Truth*	**C**	**~C**
C	True Positive (Hit)	False Negative (Miss)
~C	False Positive (False Alarm)	True Negative (Correct Rejection)

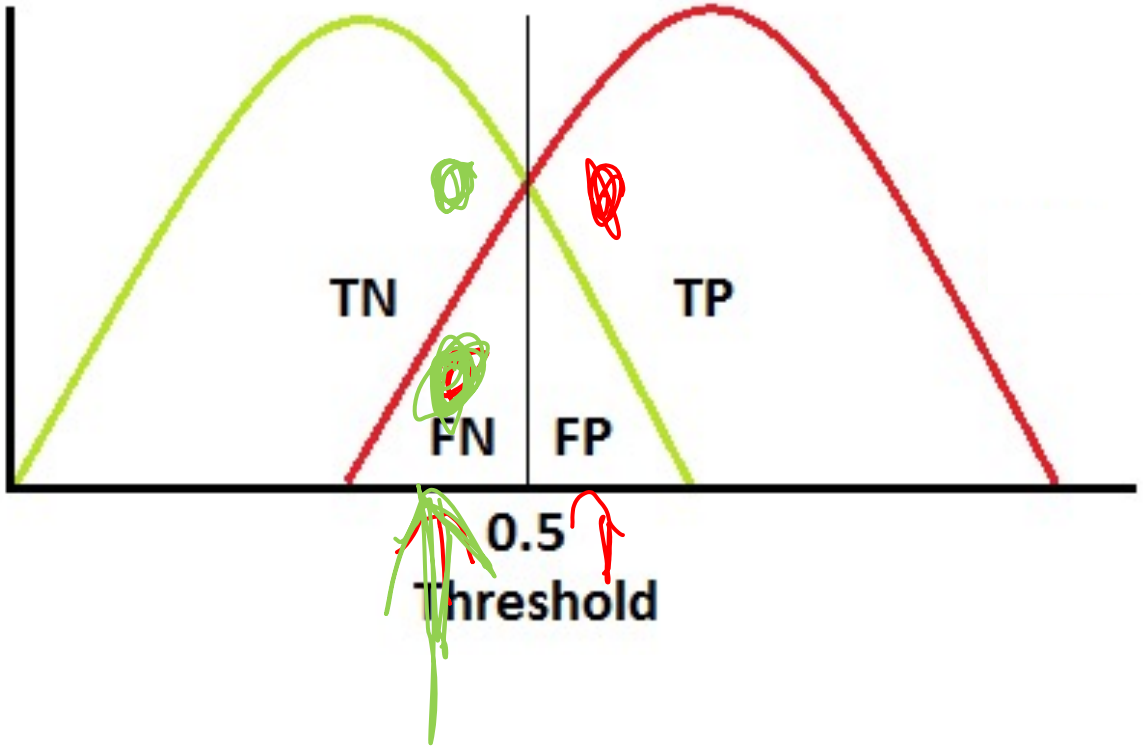
Confusion Matrix

Classified As *Ground Truth*	**C**	**~C**
C	Count of True Positives (Hit)	Count of False Negatives (Miss)
~C	Count of False Positives (False Alarm)	Count of True Negatives (Correct Rej.)

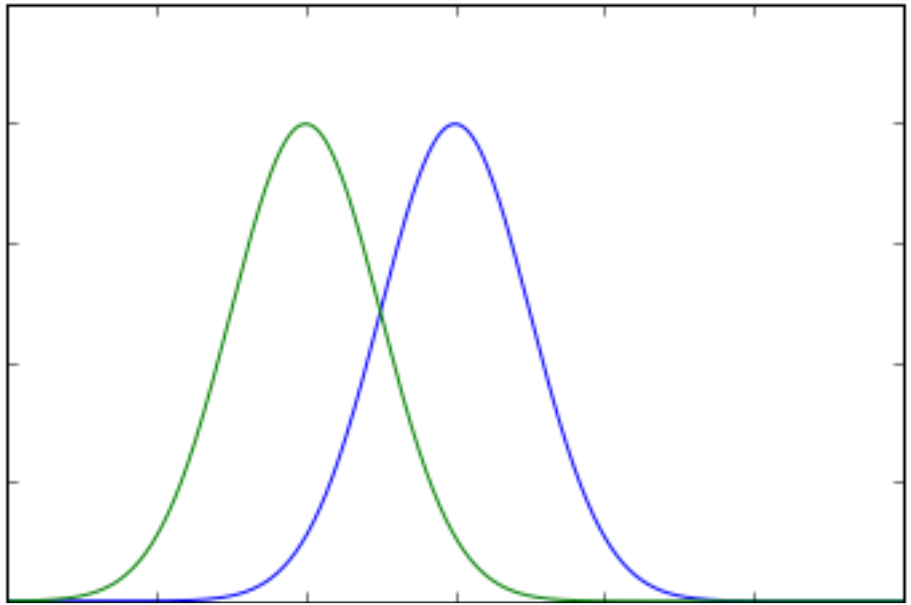
Classification Errors – Regression Model



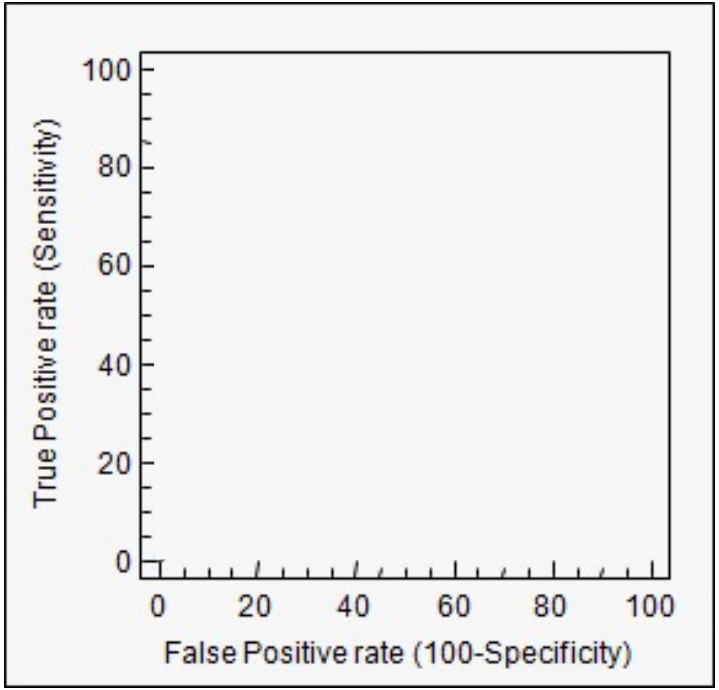
Classification Errors – Regression Model

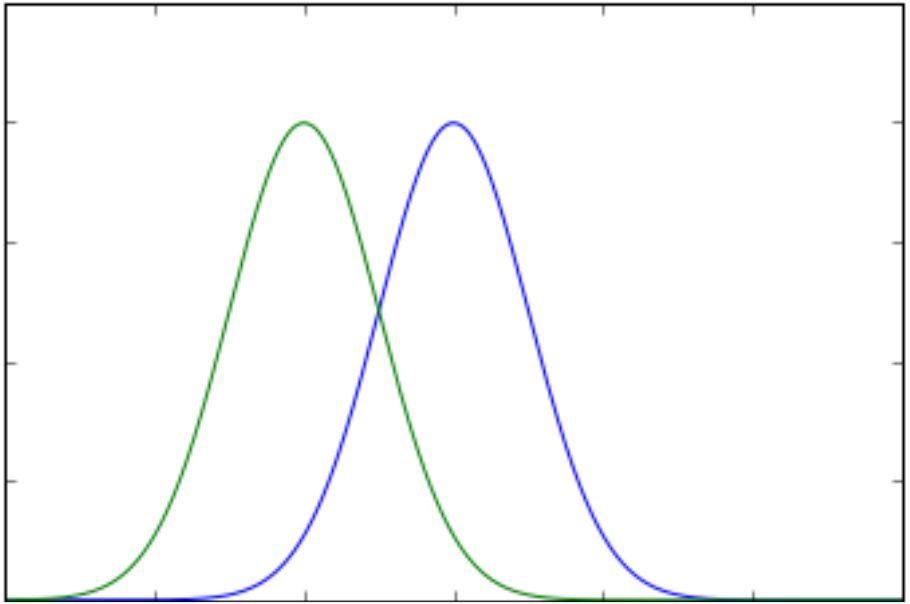


Gaussian Distributions

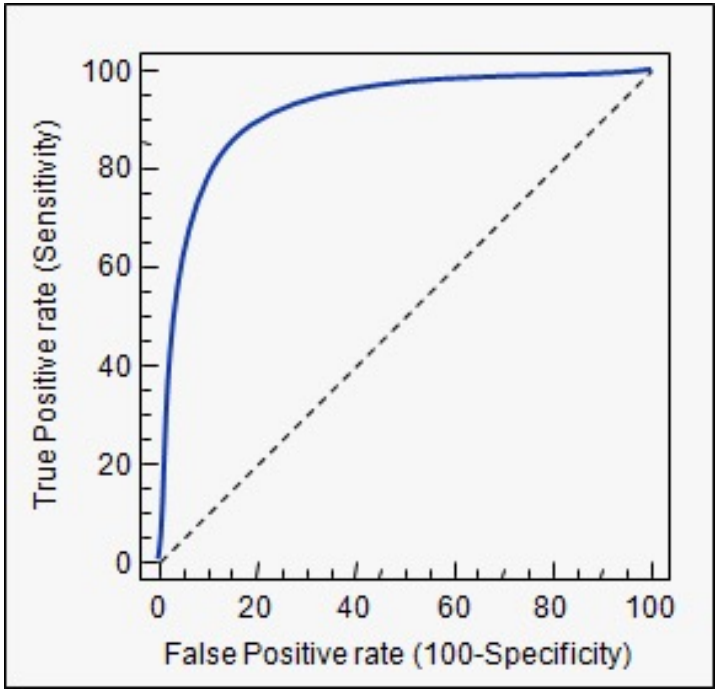


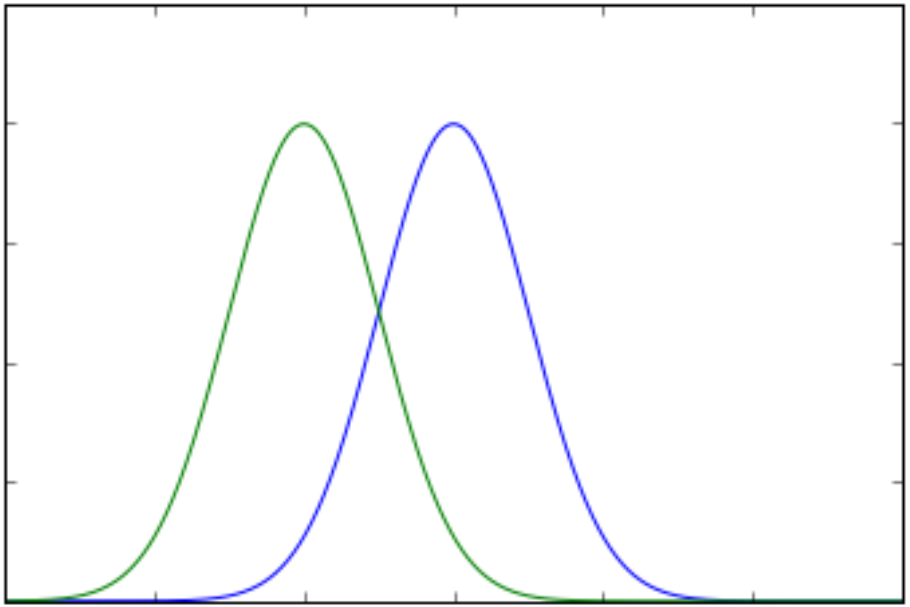
Classification Errors – Regression Model



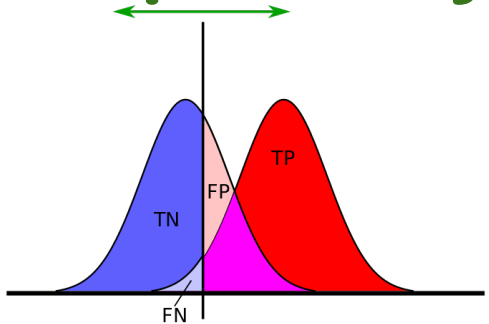


Classification Errors – ROC Curve (receiver operating characteristic)

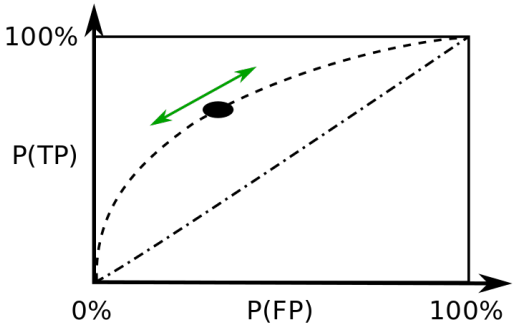




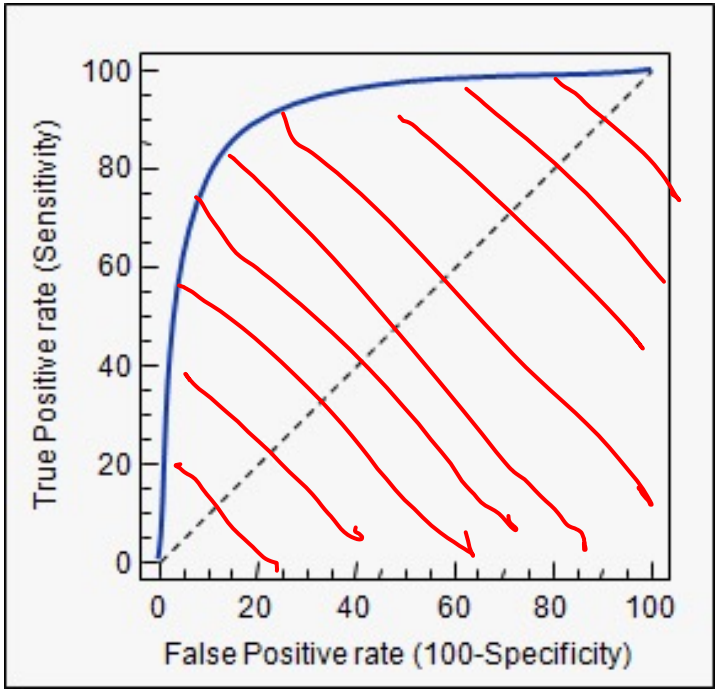
Adapting True Pos rate vs. False Pos rate (sensitivity vs. specificity)

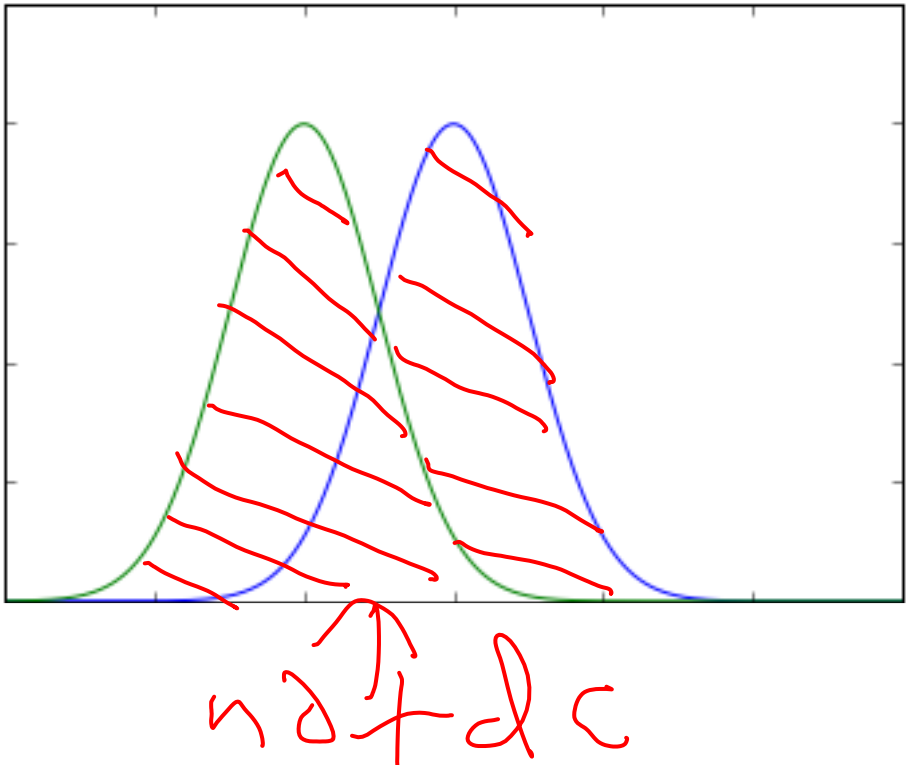


TP	FP
FN	TN

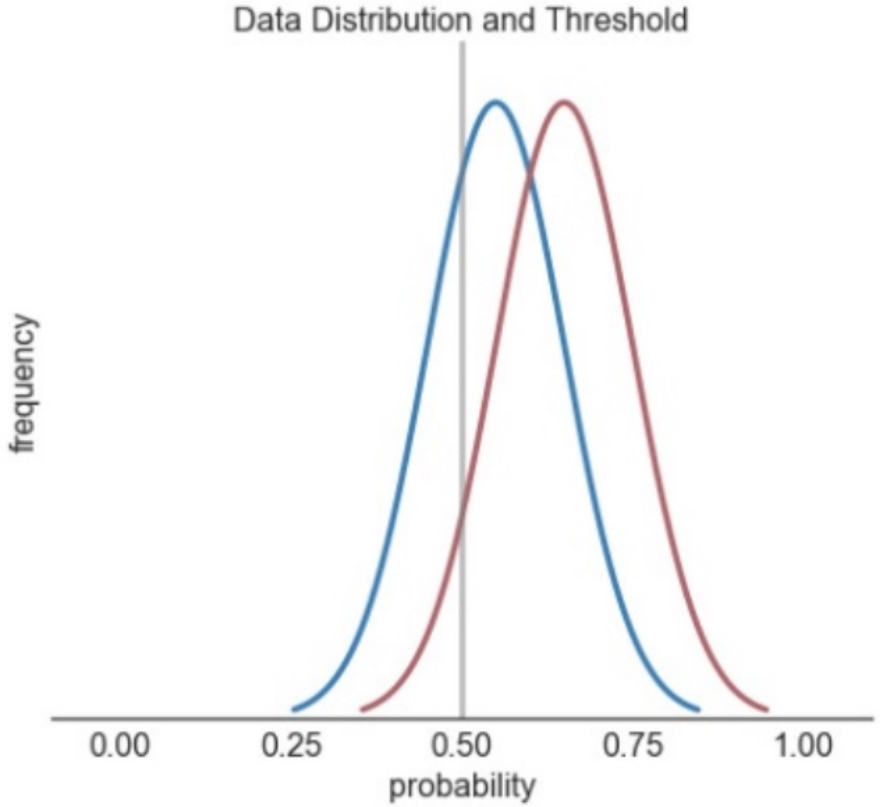


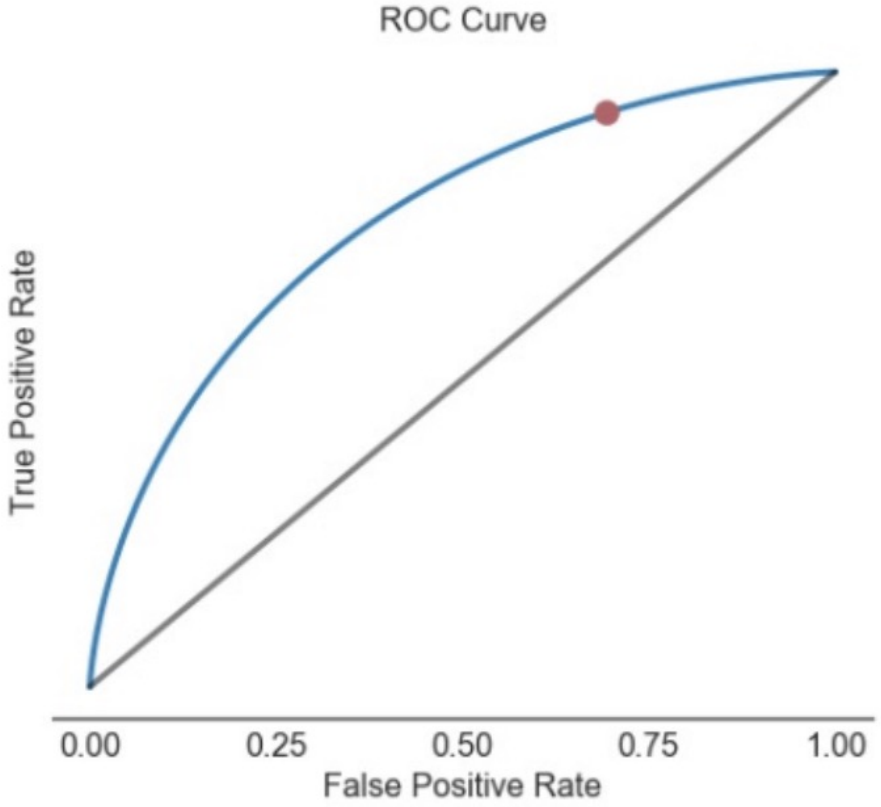
ROC Curve vs AUC



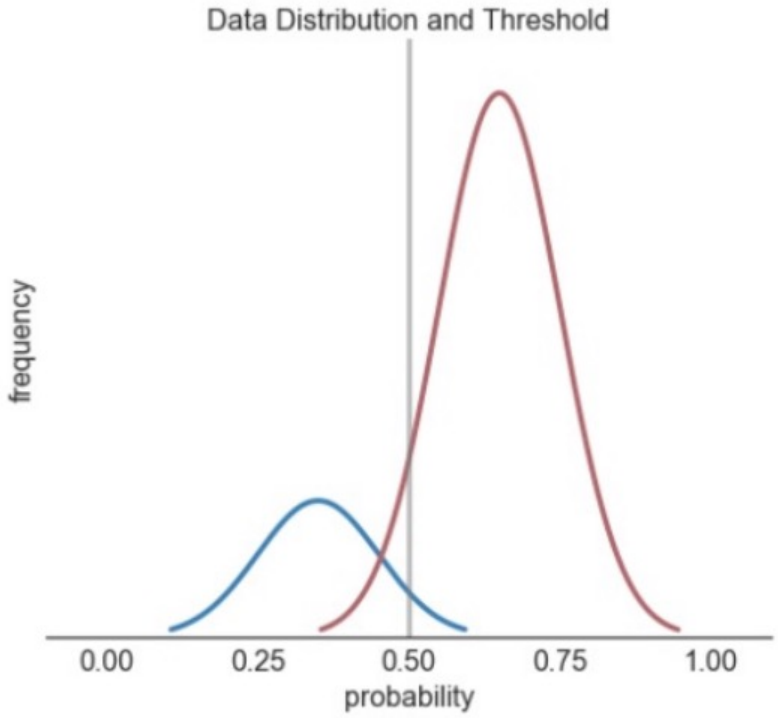


Sharpness of ROC Curve





Non-Identical Distributions - Skew



Non-Identical Distributions - Skew

END