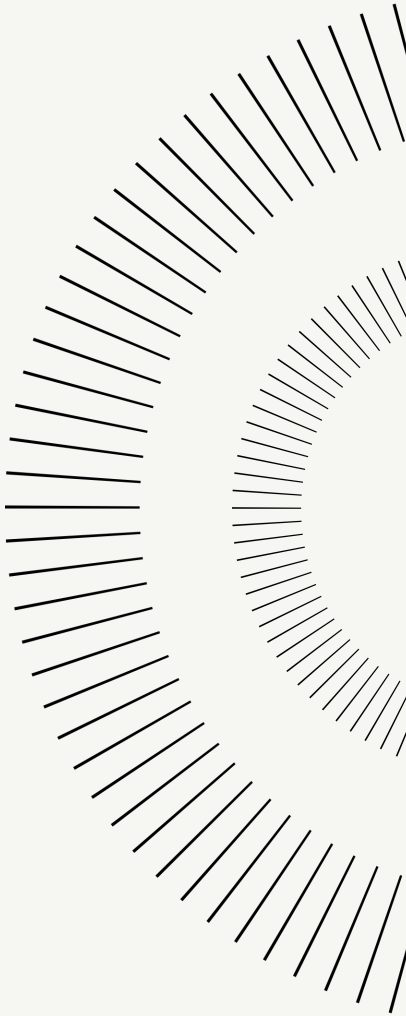
AUC is Worthless

Lessons in transitioning from academic to business data science

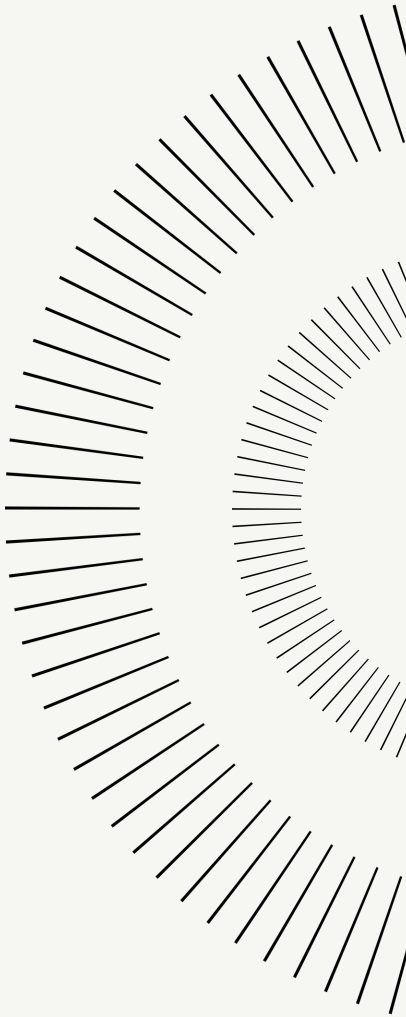


Dillon R. Gardner - PyData 2022

My background

- PhD in experimental physics from MIT
- Data scientist for 7+ years
    - Electricity markets
    - Programmatic advertising
    - Agtech
    - Fintech
- Worked in companies from 4 people to ~800 people
- Held roles as an individual contributor and as a manager
- Been laid off

Preamble:
Job Descriptions



OPENINGS

Senior Data Scientist

Technical

We are looking for someone who:

- Finds innovative solutions to problems in statistics, experiment design and distributed machine learning (we use Python libraries and Spark extensively)
- Sets high personal standards for research design that increase iteration speed and ensure reproducibility of results
- Approaches data as an artefact of the real-world and develop repeatable strategies for validating results against ground truth
- Develops techniques to detect and explore anomalies in our data, uncover the source and address in a way that improves our product
- Thrives in high-performing, collaborative teams with short, high-impact research cycles

APOLLO
AGRICULTURE

Senior Data Scientist

AMSTERDAM / DATA SCIENCE / FULL TIME

You:

- Have strong technical skills regarding data analysis, statistics, machine learning and programming.
- Experience with building scalable machine learning models in Python/R (or equivalent).
- Able to confidently collect and curate data, ingest it, explore it, and clean it.
- Experience in using SQL/BigQuery to query and manipulate large data sets for analysis.
- Are an expert in Experimental Design, Satellite Imagery, Credit Modeling, or Quantitative Agronomy.
- Are driven to find new insights in data and pose new questions that help shape our business.
- Are comfortable explaining complex solutions to non-data science peers across multiple continents and cultures.
- Are analytically rigorous, but recognize when complete is better than perfectAre excited about the mission of transforming the livelihoods of smallholder farmers.
- Are excited to spend time with our customers and learn from them.
- Know how to build end-to-end machine learning solutions to real-world problems, from building the model, to training and evaluating it, to iterating and deploying it for both batch and online predictions.
- Have product management skills to maximize the impact of Data Science work.
- Have experience with cloud platforms (GCP, AWS, Azure, etc.)

Google Careers

Data Scientist, Engineering

Google In-office: Mountain View, CA, USA Sunnyvale, CA, USA

Preferred qualifications:

- PhD degree in a quantitative discipline.
- 4 years of relevant work experience, including expertise with statistical data analysis such as linear models, multivariate analysis, stochastic models, sampling methods.
- Applied experience with machine learning on large datasets.
- Experience articulating and translating business questions and using statistical techniques to arrive at an answer using available data.
- Demonstrated leadership and self-direction. Willingness to both teach others and learn new techniques.
- Demonstrated skills in selecting the right statistical tools given a data analysis problem. Effective written and verbal communication skills.

MIT OpenLearningLibrary
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Course

Introduction to Machine Learning

GENERAL ASSEMBLY

Data Science Immersive

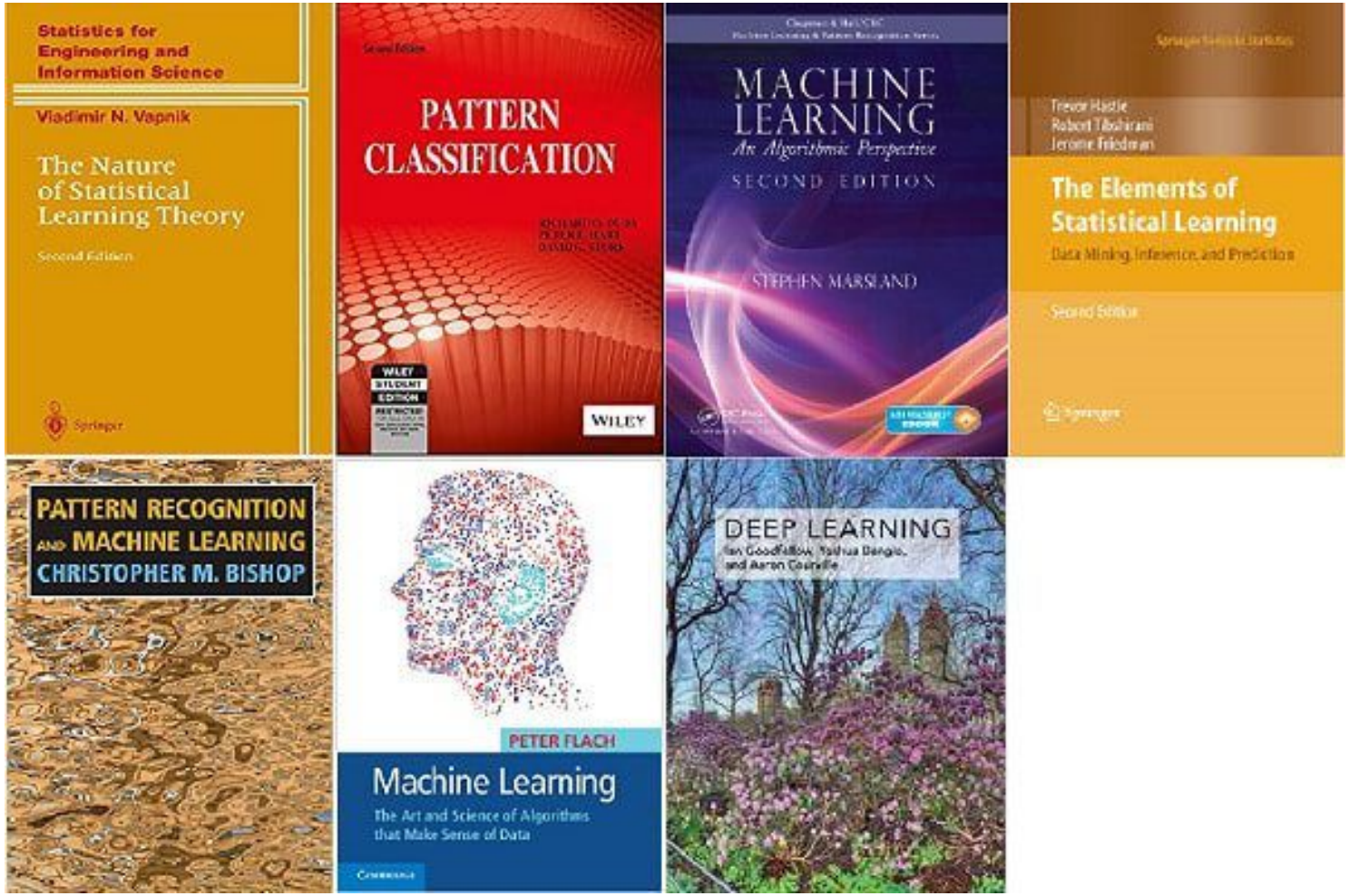
Unit 1: Data Science Fundamentals

Unit 2: Exploratory Data Analysis

Unit 3: Classical Statistical Modeling

Unit 4: Machine Learning Models

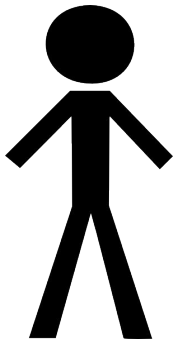
Unit 5: Advanced Topics and Trends


Statistics for Engineering and Information Science
Vladimir N. Vapnik
The Nature of Statistical Learning Theory
Second Edition
Springer
PATTERN CLASSIFICATION
WILEY STUDENT EDITION
WILEY
MACHINE LEARNING
An Algorithmic Perspective
SECOND EDITION
STEPHEN MARSLAND
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer
PATTERN RECOGNITION AND MACHINE LEARNING
CHRISTOPHER M. BISHOP
PETER FLACH
Machine Learning
The Art and Science of Algorithms that Make Sense of Data
DEEP LEARNING

What do companies want?

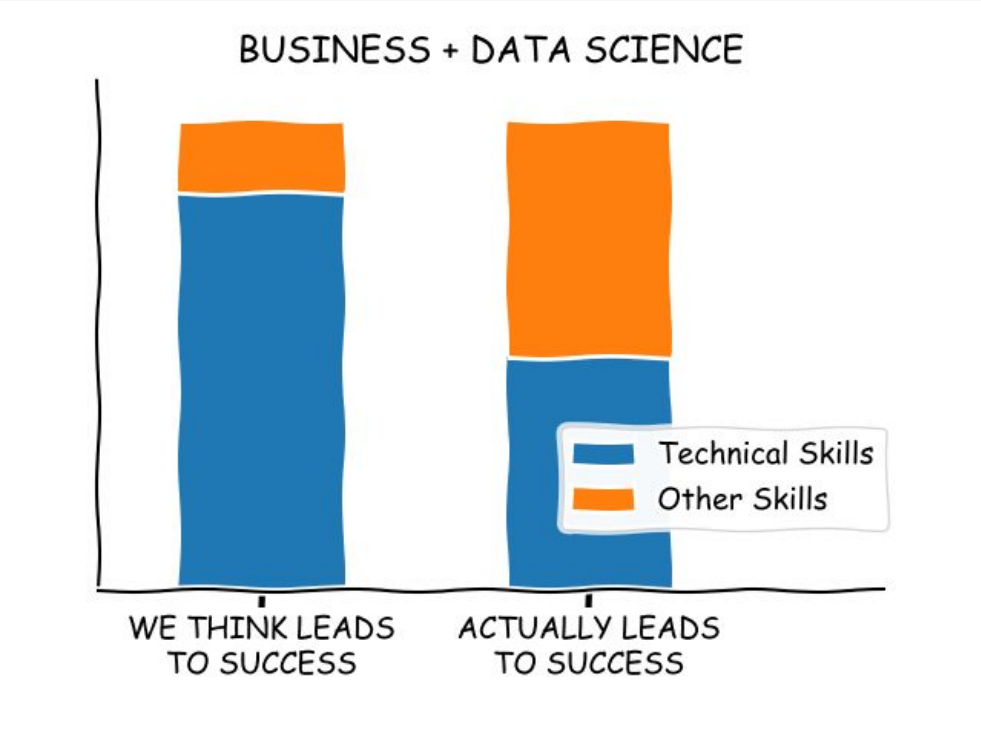
Many people, many requirements

- Machine learning + Programming
- Statistics + Math
- Domain expertise



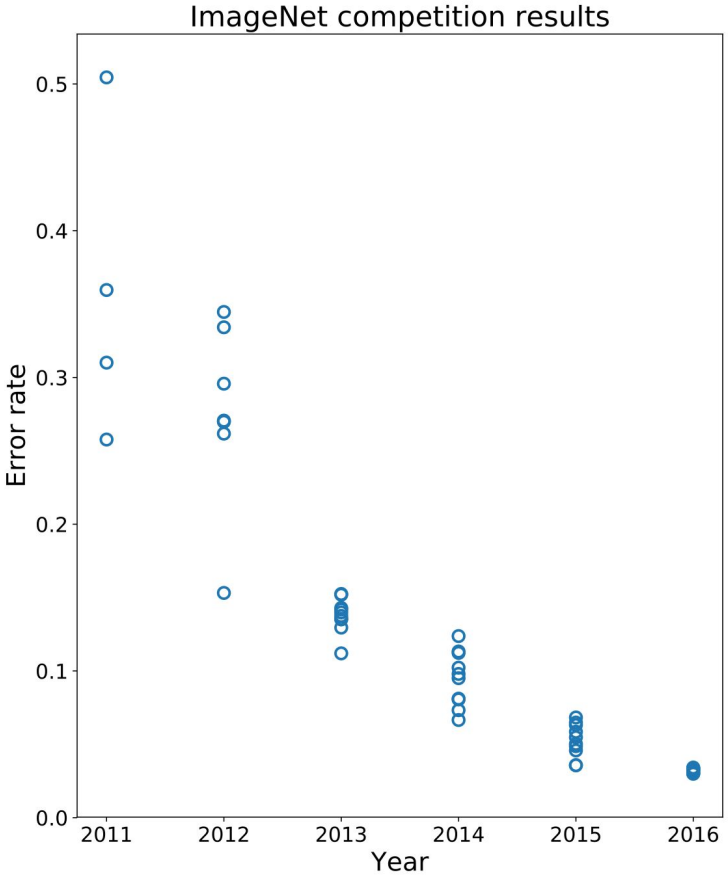
- Real world data
  - Not toy problems
- Communication
  - How do we make sense of the work?
- Product Management
  - How do we get value out of the work?

Mismatch between training and skills for success

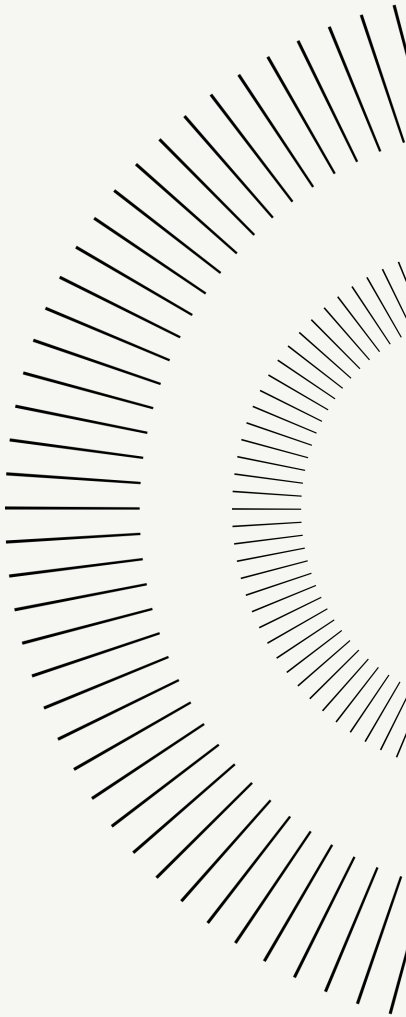


What causes the mismatch?

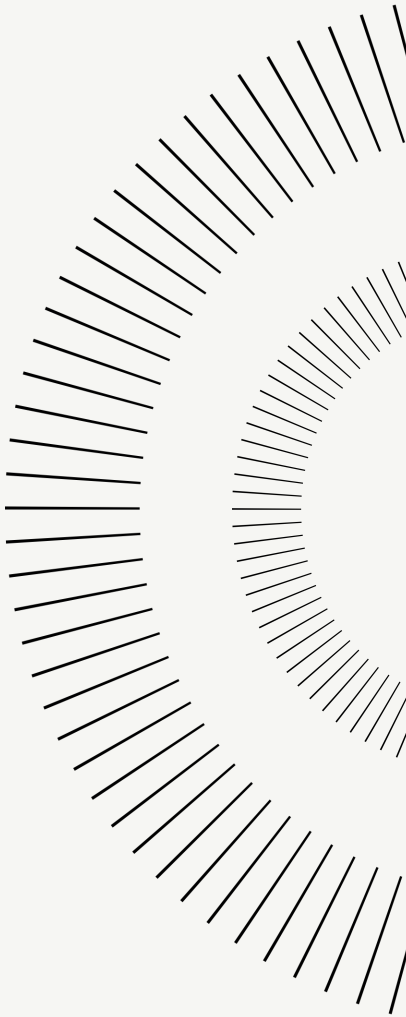
- Classes taught by academics whose other job is researching how to expand what is possible
- We are more comfortable *assessing* technical skills
- We are more comfortable *teaching* technical skills
- We might disdain business skills?



A Trip To Business School



Scoring Functions



Loss Function - what an ML algorithm tries to minimize

Scoring Function - what a business wants to minimize (other terms used are **cost** and **utility**)

Loss Function - need to have nice mathematical properties

Scoring Function - can be very arbitrary

Loss + Scoring

Loss

- RMSE
- Cross-Entropy
- Regularization
- …

Model optimization

Academic Scoring

- Accuracy
- F1-score
- AUC
- …

Model selection

Loss + Academic Scoring + Business Scoring

Loss

- RMSE
- Cross-Entropy
- Regularization
- …

Model optimization

Academic Scoring

- Accuracy
- F1-score
- AUC
- …

Model selection

Business Scoring

- Profit
- Loss Ratio
- Model Complexity
- Explainability

Business optimization

Loss + Academic Scoring + Business Scoring

Loss

- RMSE
- Cross-Entropy
- Regularization
- …

Academic Scoring

- Accuracy
- F1-score
- AUC
- …

Business Scoring

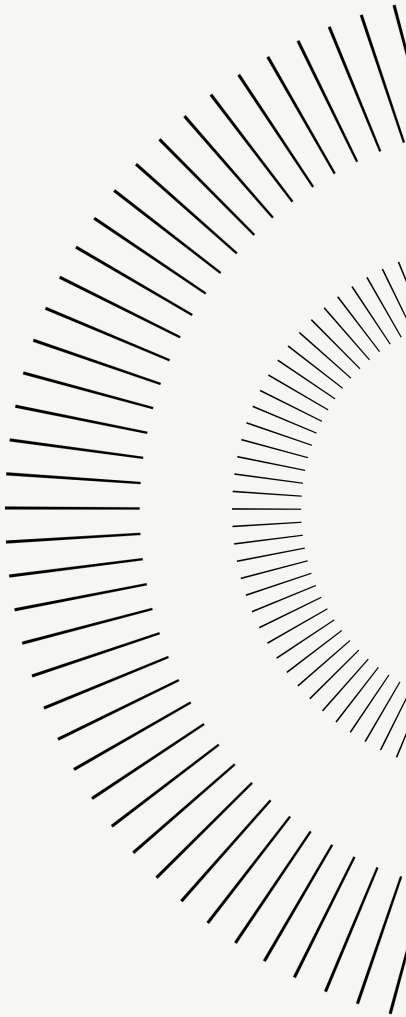
- Profit
- Loss Ratio
- Model Complexity
- Explainability

Model optimization

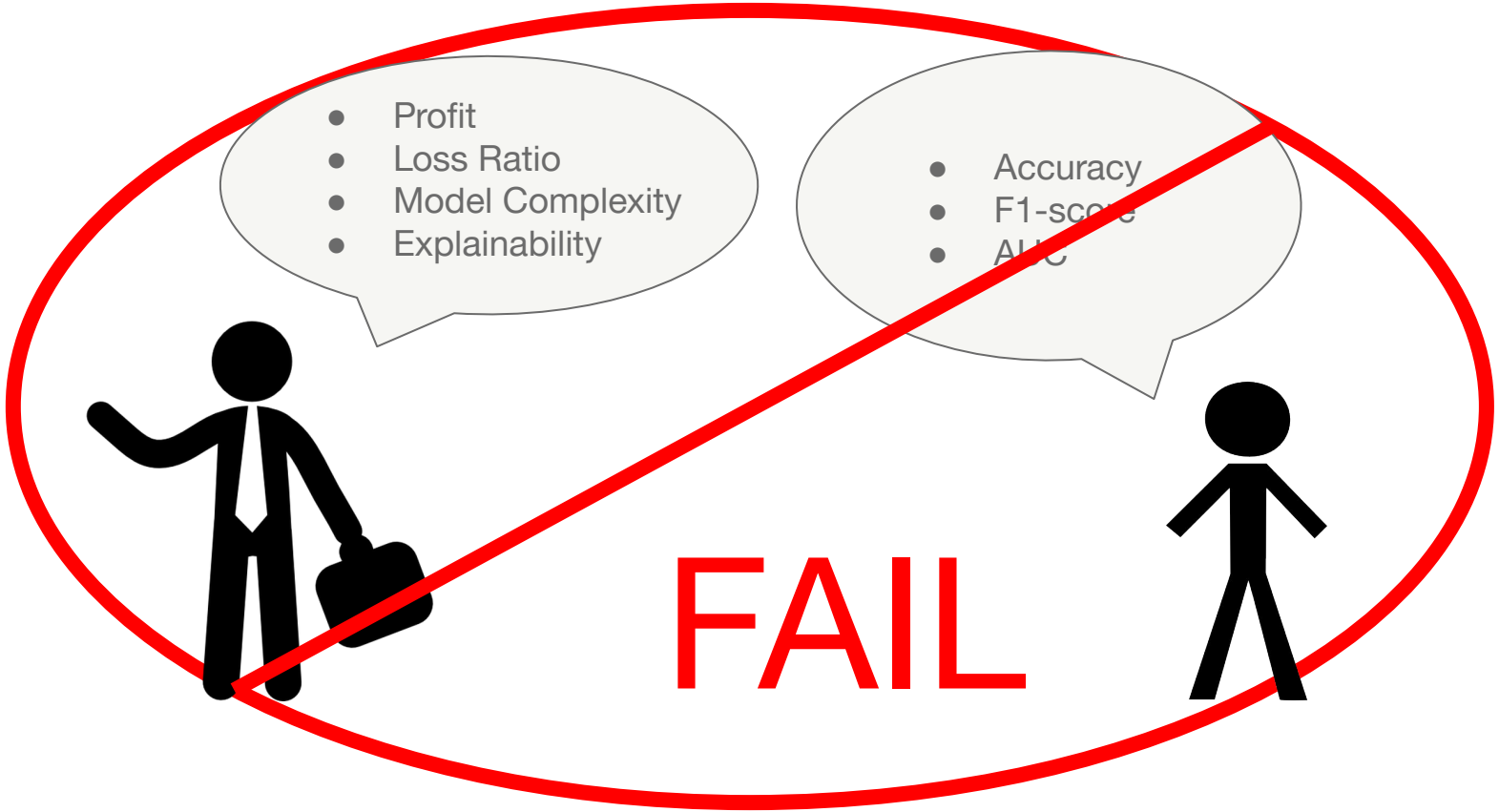
Model selection

Business optimization

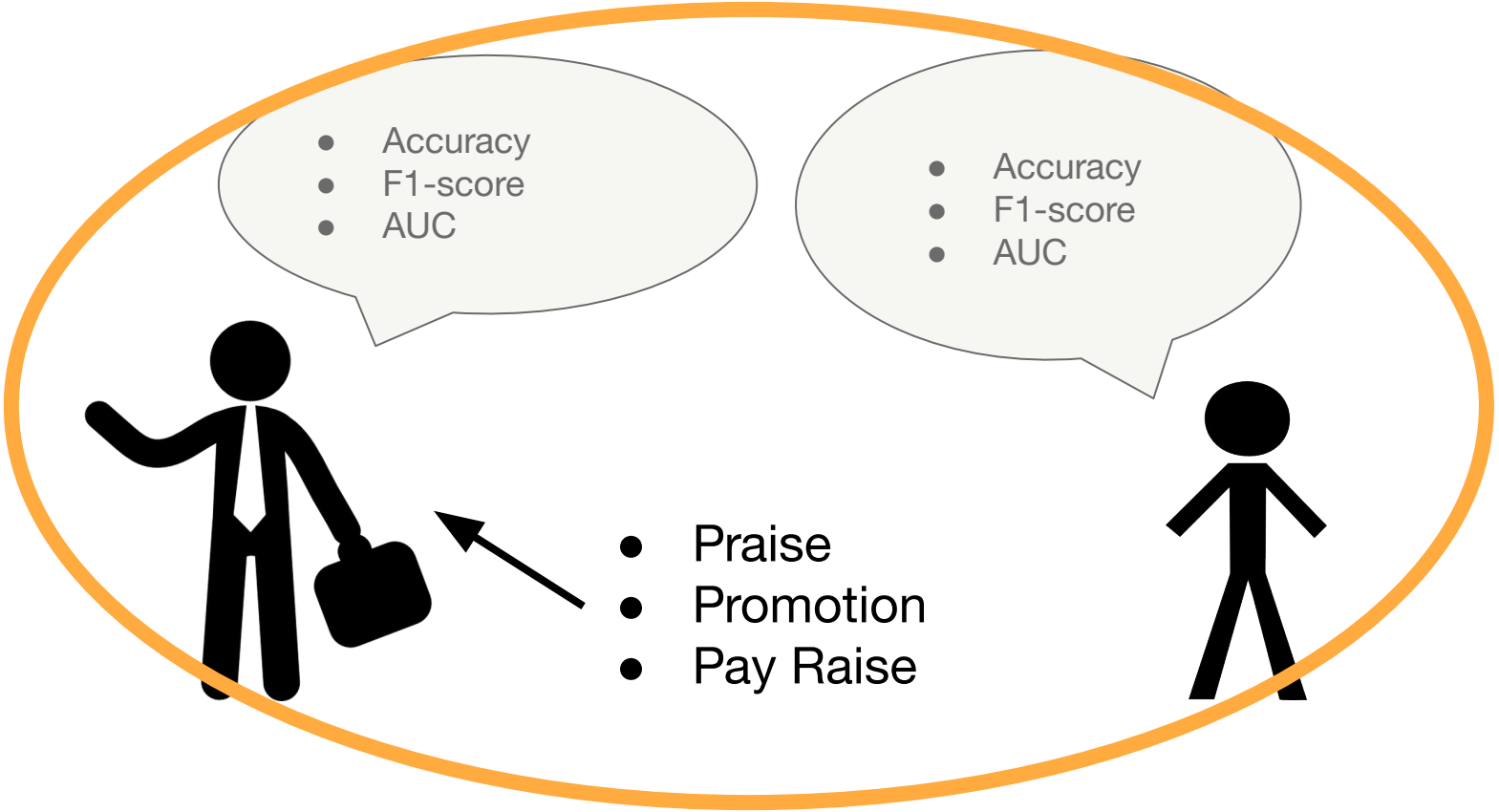
How products are discussed



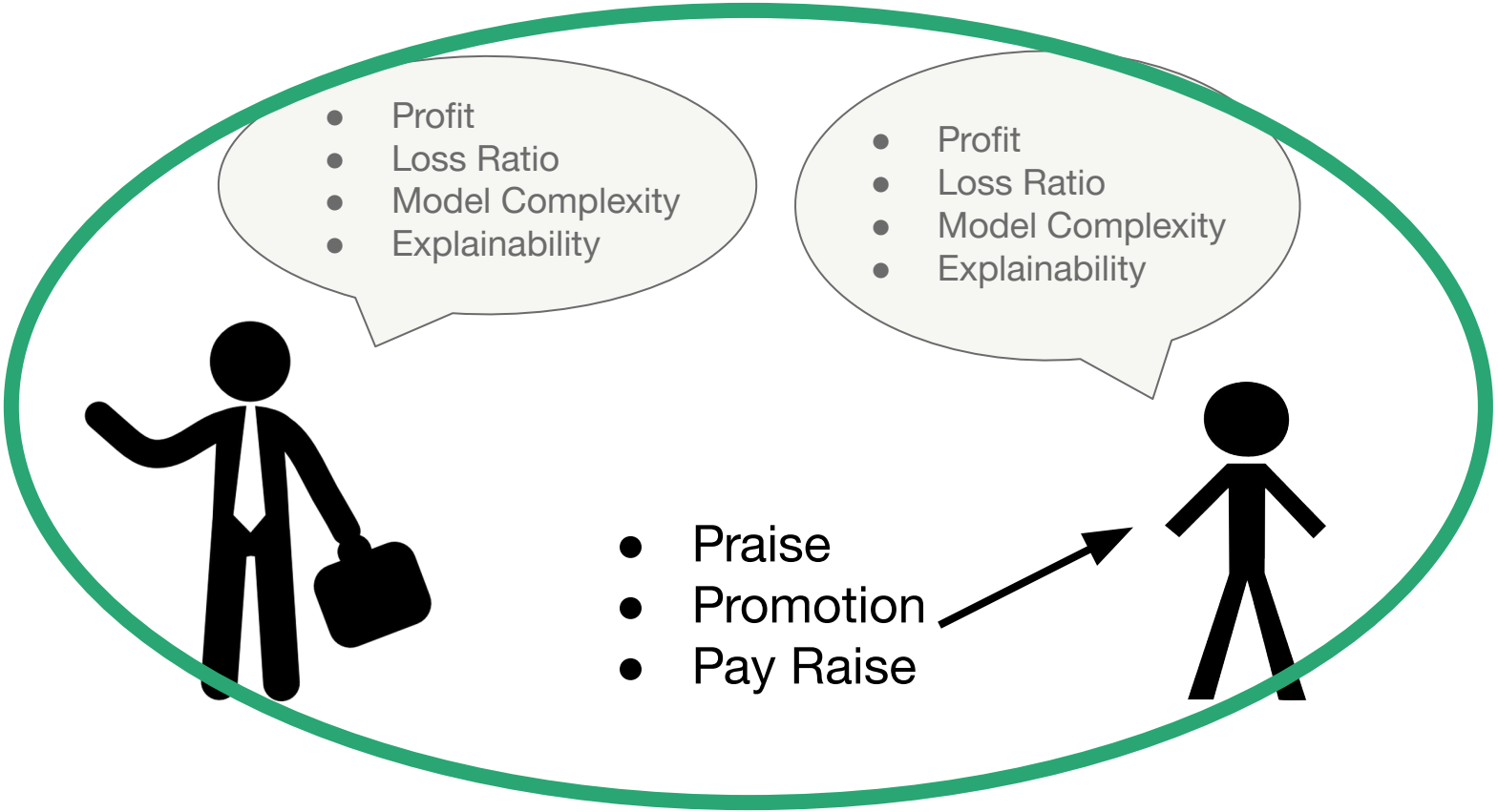
Business - DS Communication



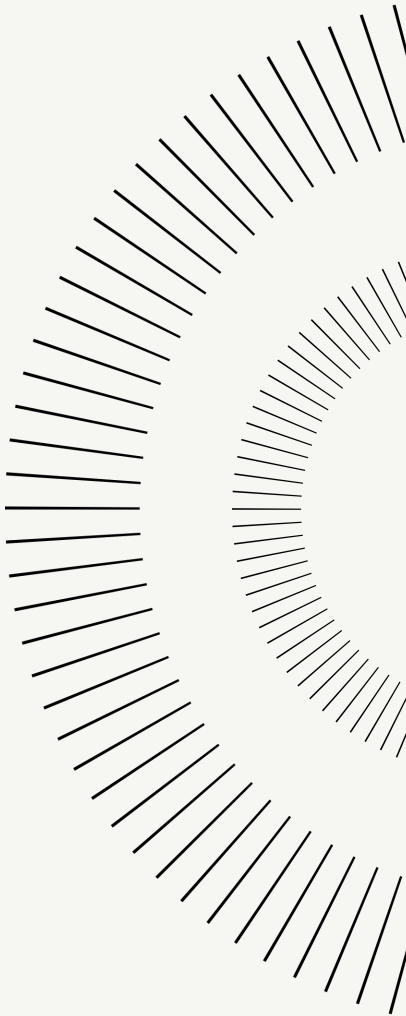
Business - DS Communication



Business - DS Communication

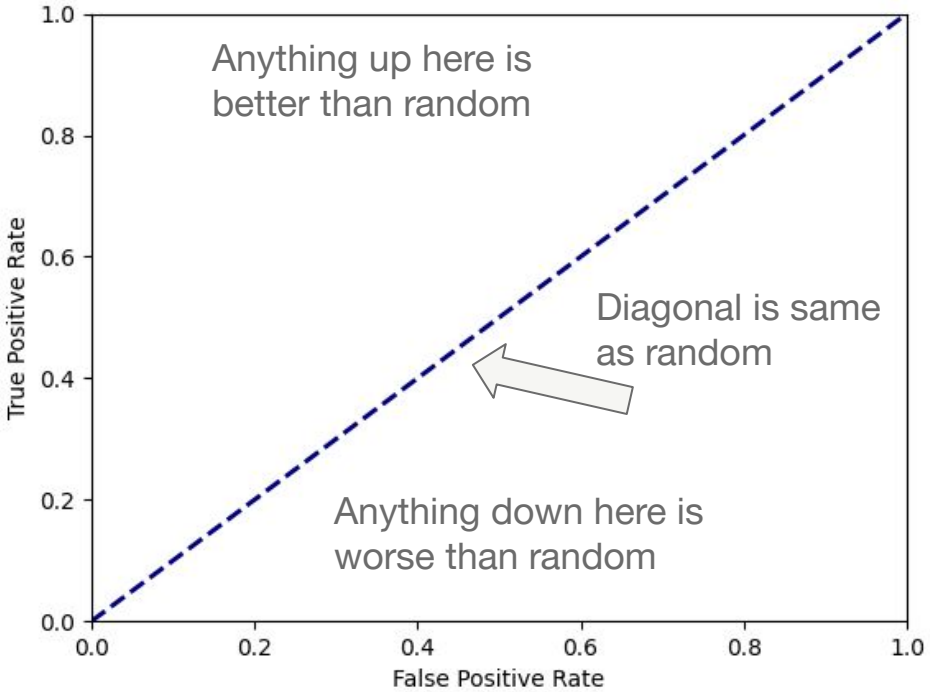


How to get there?

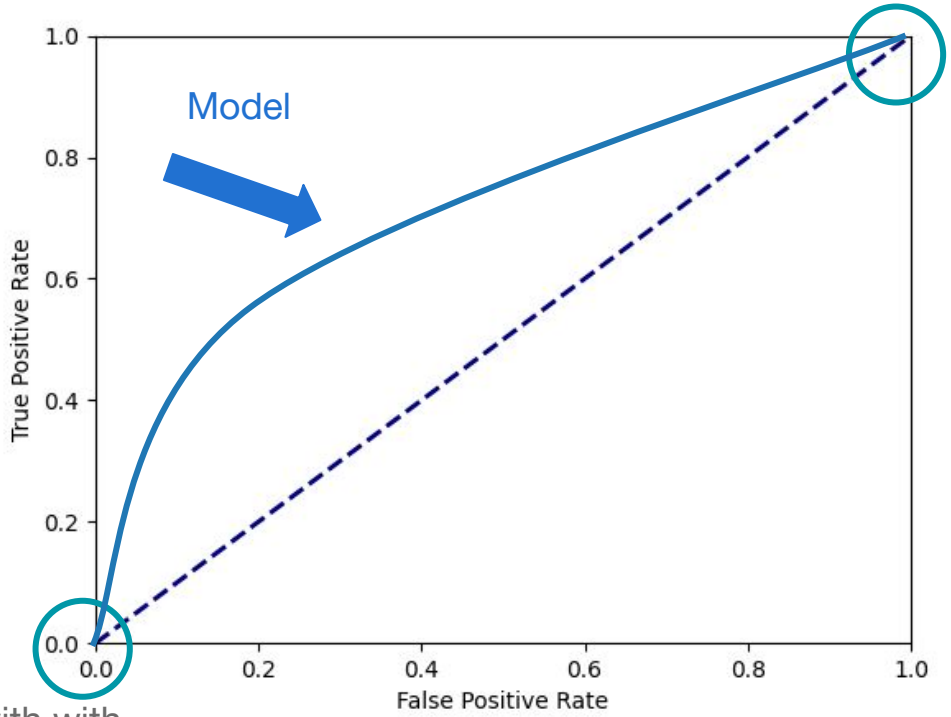


Quick Aside - ROC curves and AUC

Trade-off between True Positive Rate and False Positive Rate



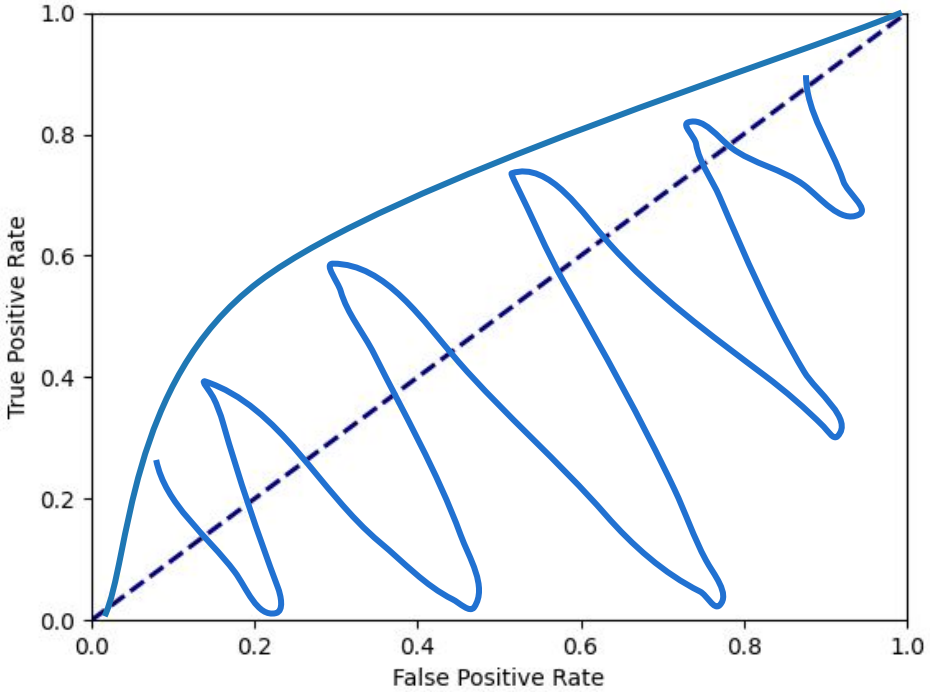
Quick Aside - ROC curves and AUC



End here with with 100% acceptance

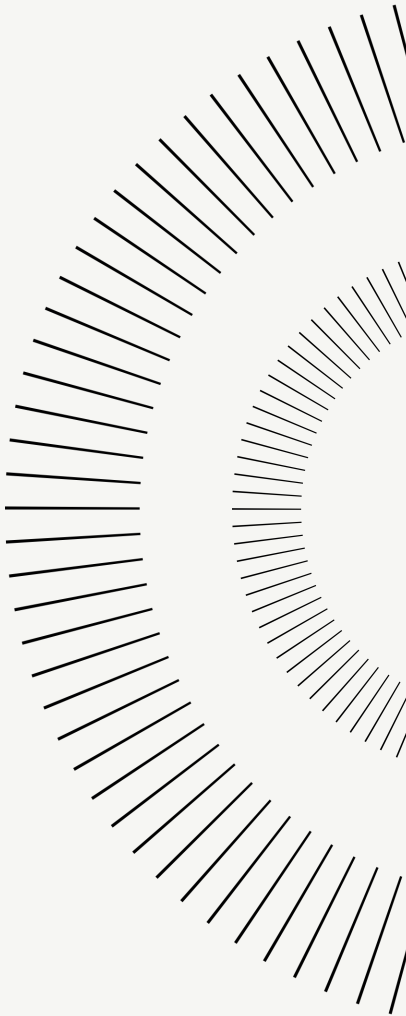
Start here with with 100% rejection

Quick Aside - ROC curves and AUC

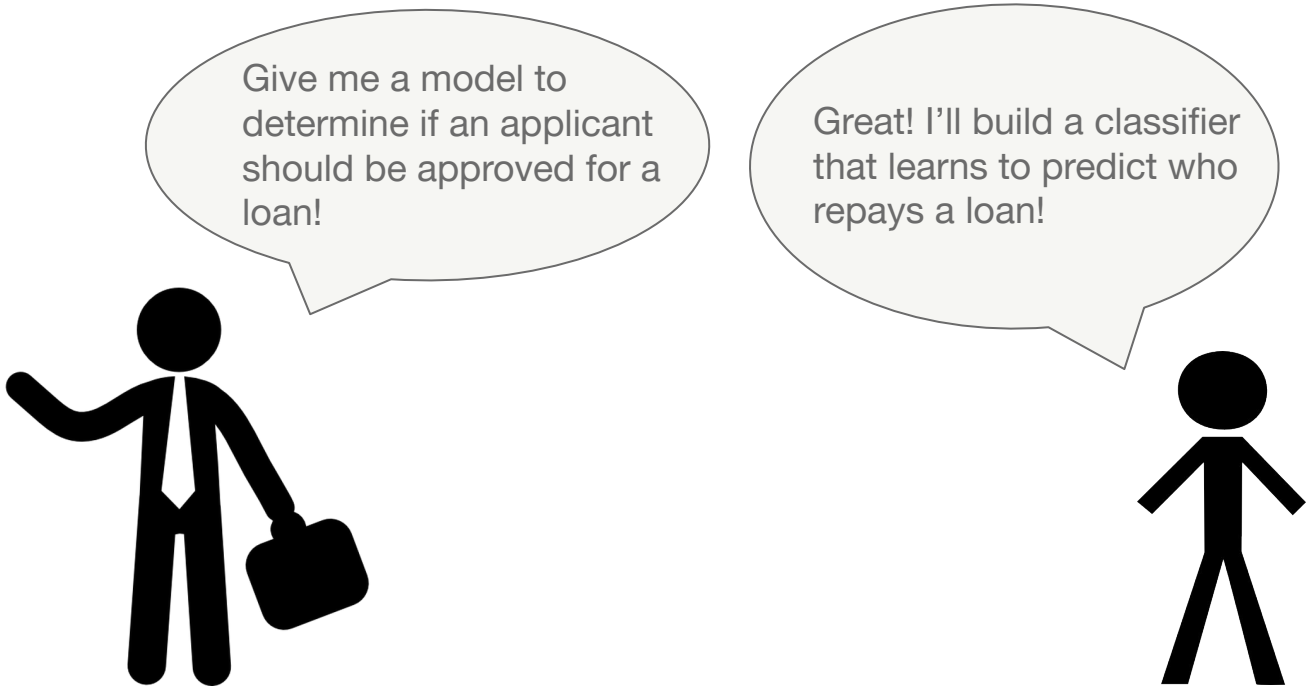


The Area Under the Curve (AUC)

Example - Loan Applications

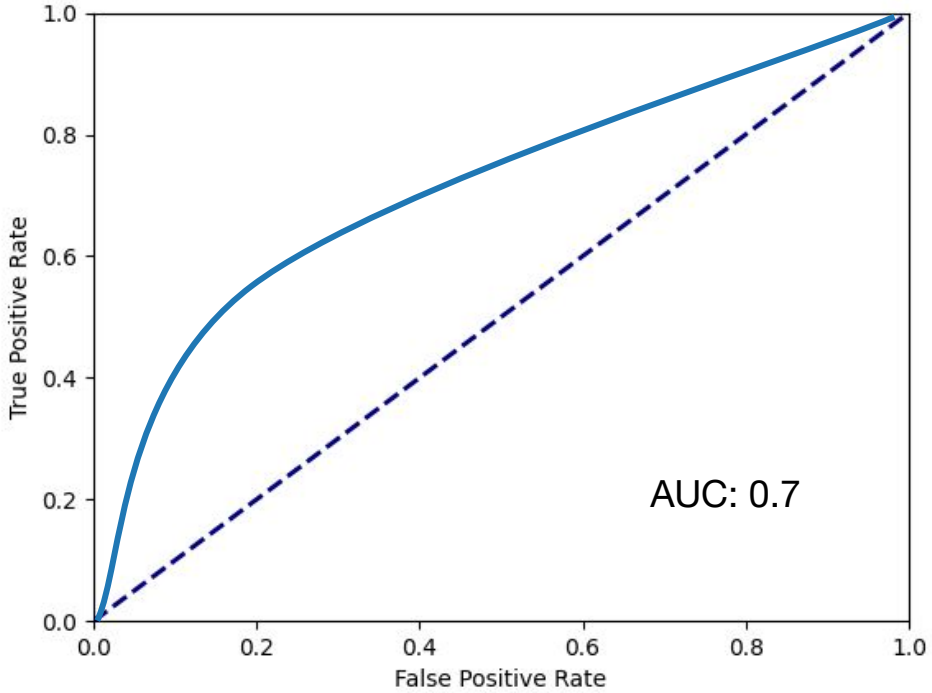


Example - Loan Applications



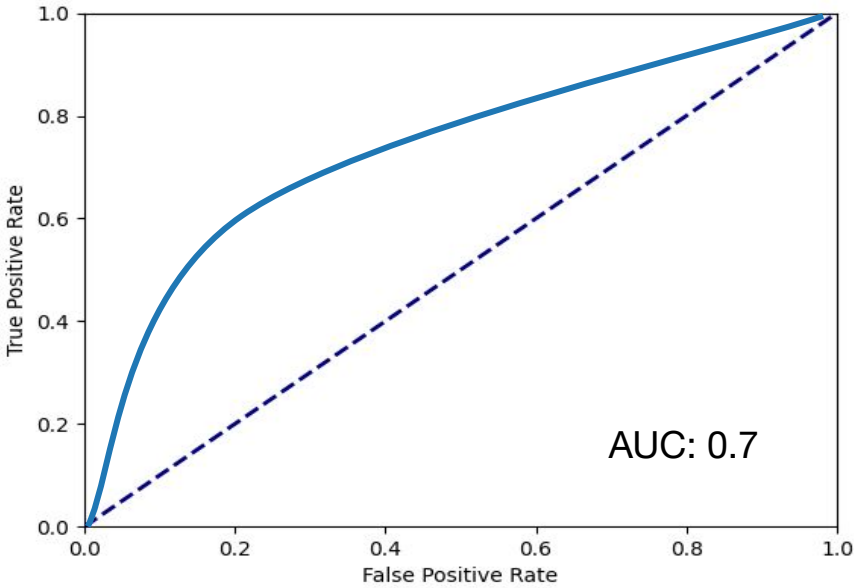
* Everyone either repays in a lump-sum payment or defaults (no-partial payment)

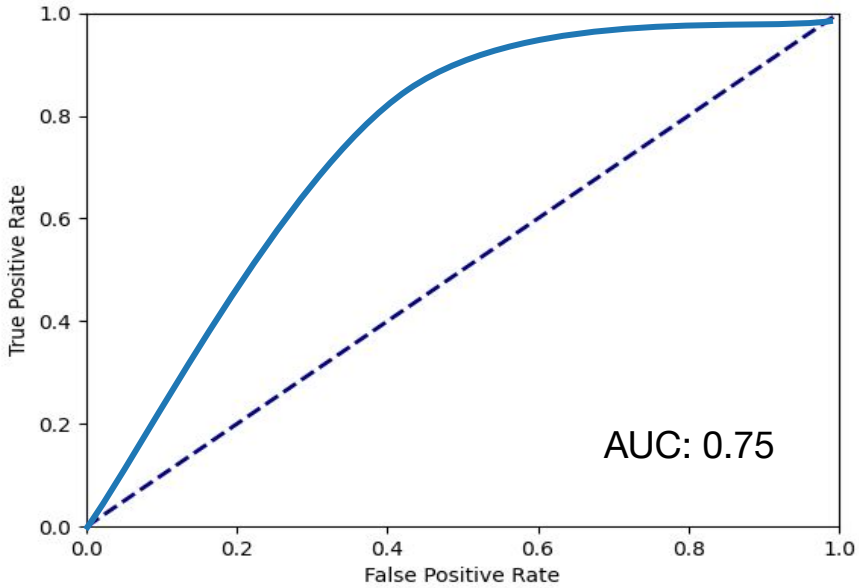
Success! I made a model!



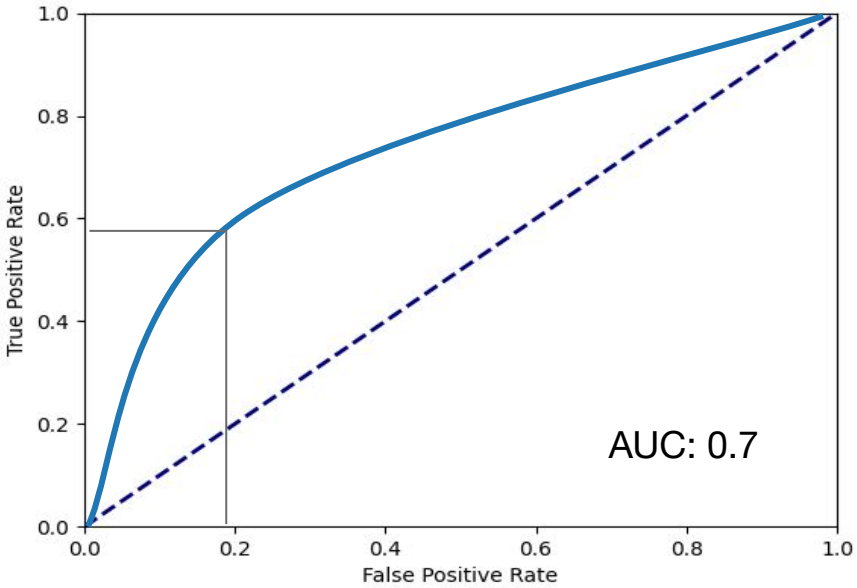
- Is this any good?
- Is it good enough to be worth the cost of using it?
- Is it better than no model at all?
- How do you know?

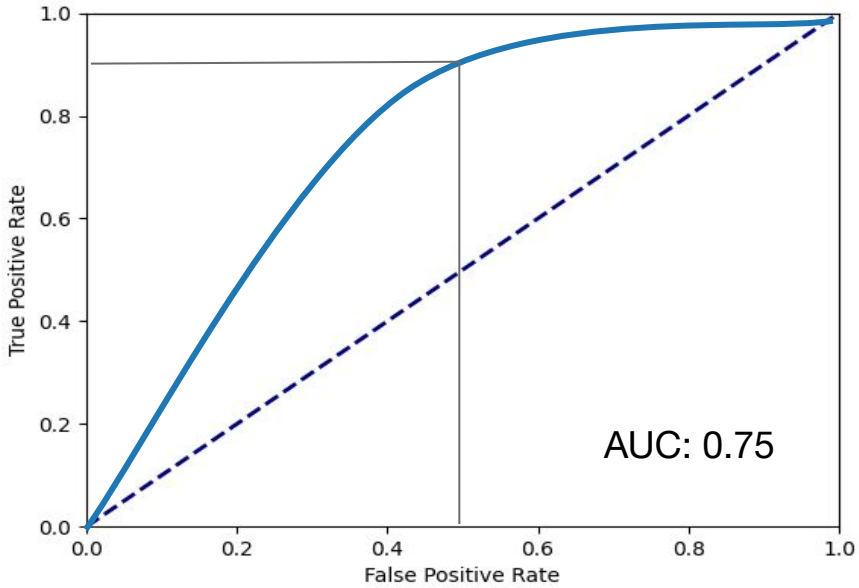
Which model is better?





Which TPR-FPR trade-off is better?





Confusion Matrix

		Predicted condition	
	Total population = P + N	**Positive (PP)**	**Negative (PN)**
Actual condition	**Positive (P)**	**True positive (TP)**	**False negative (FN)**
	Negative (N)	**False positive (FP)**	**True negative (TN)**

Confusion Matrix - Utility

What does the business gain/lose for each outcome

		Predicted condition	
	Total population = P + N	**Positive (PP)**	**Negative (PN)**
Actual condition	**Positive (P)**	U_{TP}	U_{FN}
	Negative (N)	U_{FP}	U_{TN}

What does the business get for each outcome?

True Positive = ???

False Positive = ???

True Negative = ???

False Negative = ???

Talk to the CFO!

Or CEO, Project Manager, etc.

True Positive = Price (P) +
Cost of Goods and Services (COGS) +
Interest (I) +
Customer Acquisition Cost (CAC)

False Positive = COGS + Interest + CAC

True Negative = CAC

False Negative = CAC

Is this right? - Maybe!

True Positive = Price (P) +

Cost of Goods and Services (COGS) +

Interest (I) +

Customer Acquisition Cost (CAC) +

Value of Learning (VOL)

False Positive = COGS + I + CAC + VOL

True Negative = CAC

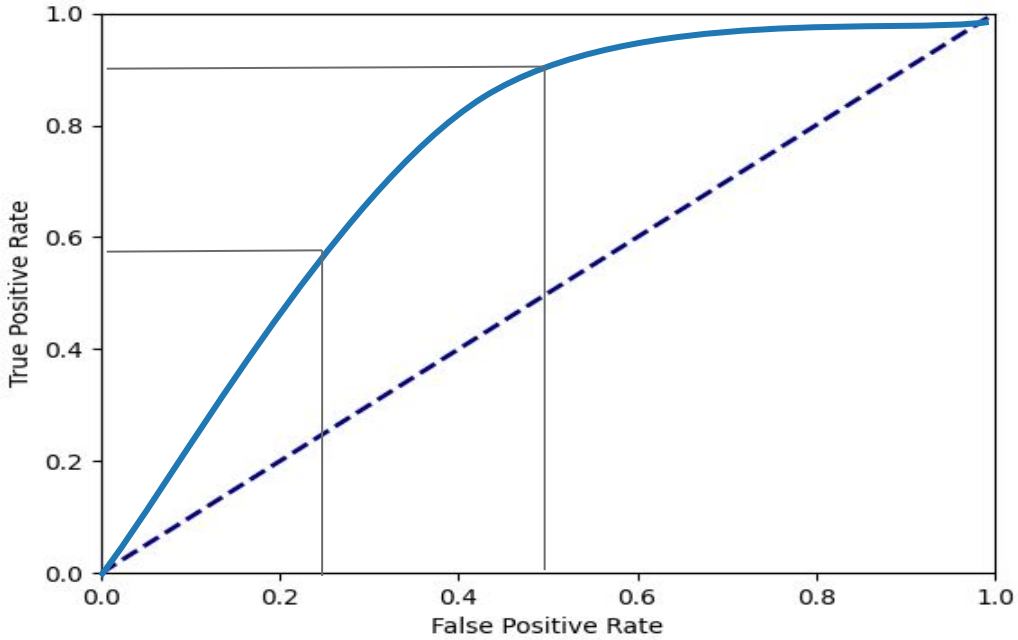
False Negative = CAC

Confusion Matrix - Utility

Cherry pick numbers to make examples work well

		Predicted condition	
	Total population = P + N	**Positive (PP)**	**Negative (PN)**
Actual condition	**Positive (P)**	**110**	**-10**
	Negative (N)	**-162**	**-10**

Calculate utility for each point in TPR - FPR space



Time for the Algebra Crank

Goal: Calculate expected utility in TPR / FPR space

$$r_p = \frac{P}{N+P}$$

prevalence - probability an example is positive

$$r_n = \frac{N}{N+P}$$

“negavence” - probability an example is negative

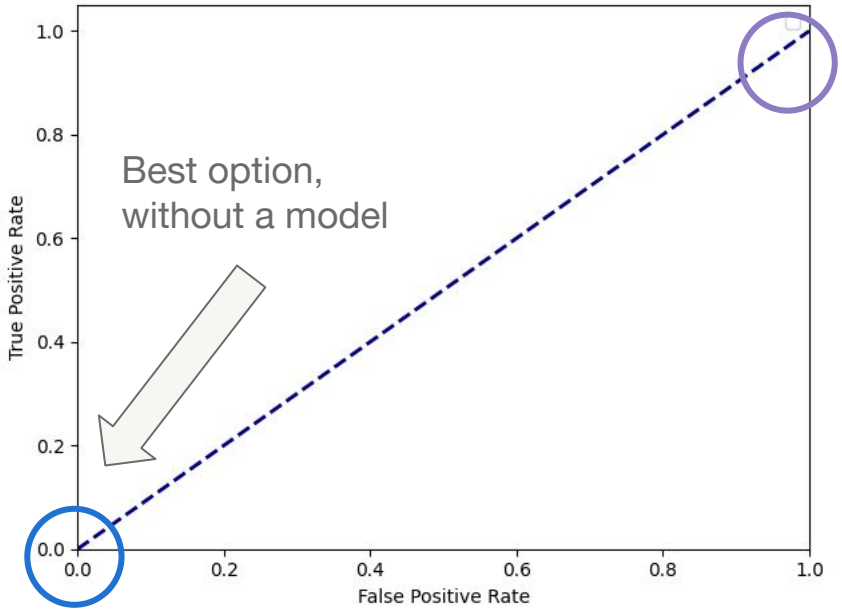
$$E[U] = u_{TP} * TPR * r_p \ + \ u_{FN} * (1 - TPR) * r_p + \\ u_{FP} * FPR * r_n \ + \ u_{TN} * (1 - FPR) * r_n$$

Value of 100% rejection vs acceptance

		Predicted condition	
	Total population = P + N	**Positive (PP)**	**Negative (PN)**
Actual condition	**Positive (P)**	**110**	**-10**
	Negative (N)	**-162**	**-10**

prevalence = $r_p = 0.5$

Accepting everyone is worth -26



Rejecting everyone is worth -10

Time for the Algebra Crank

Curves of constant utility in TPR / FPR space

$$TPR = m * FPR + b$$

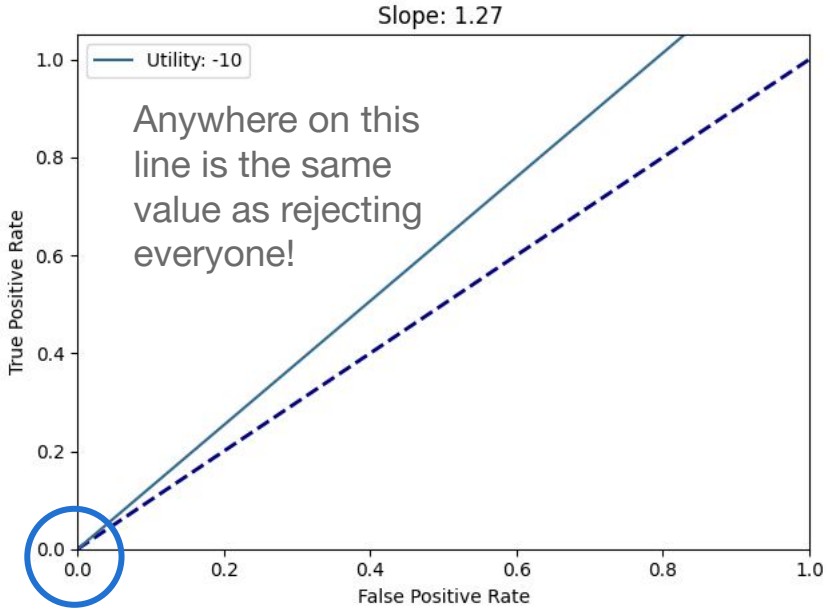
$$m = \frac{u_{TN} - u_{FP}}{u_{TP} - u_{FN}} * \frac{r_n}{r_p}$$

$$b = \frac{E[U] - u_{FN} * r_p - u_{TN} * r_n}{(u_{TP} - u_{FN}) * r_p}$$

Lines of Constant Utility

		Predicted condition	
	Total population = P + N	**Positive (PP)**	**Negative (PN)**
Actual condition	**Positive (P)**	**110**	**-10**
	Negative (N)	**-162**	**-10**

prevalence = r_p = 0.5



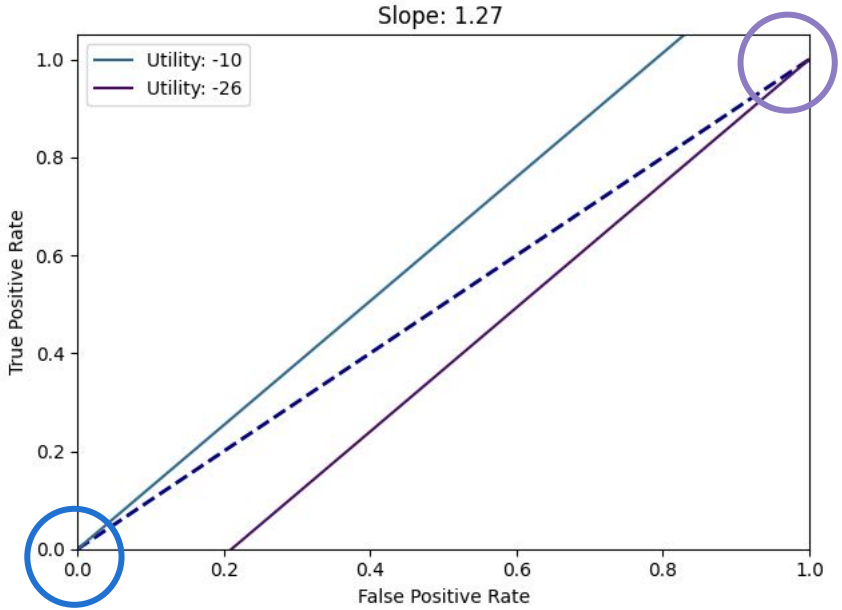
Rejecting everyone is worth -10

Lines of Constant Utility

		Predicted condition	
	Total population = P + N	**Positive (PP)**	**Negative (PN)**
Actual condition	**Positive (P)**	**110**	**-10**
	Negative (N)	**-162**	**-10**

prevalence = r_p = 0.5

Accepting everyone is worth -26



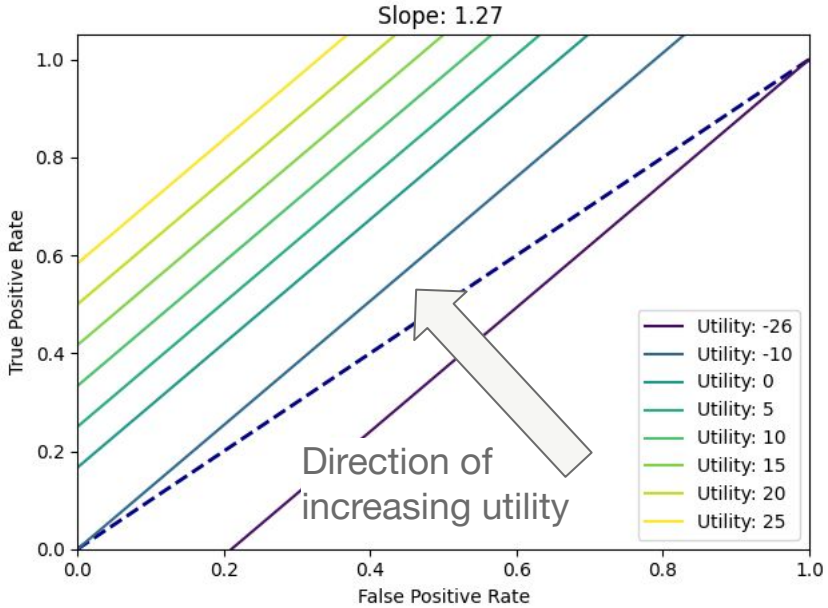
Rejecting everyone is worth -10

Lines of Constant Utility

Slope > 1 => Reject everyone if you don't have a model

		Predicted condition	
	Total population = P + N	**Positive (PP)**	**Negative (PN)**
Actual condition	**Positive (P)**	**110**	**-10**
	Negative (N)	**-162**	**-10**

prevalence = r_p = 0.5



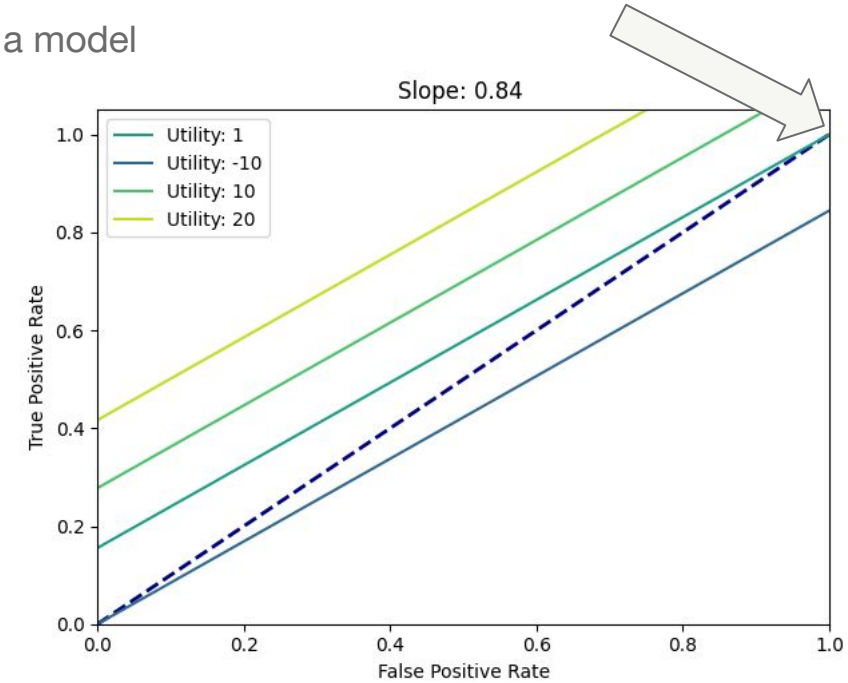
Increase in prevalence

Slope < 1 => Accept everyone if you don't have a model

		Predicted condition	
	Total population = P + N	**Positive (PP)**	**Negative (PN)**
Actual condition	**Positive (P)**	**110**	**-10**
	Negative (N)	**-162**	**-10**

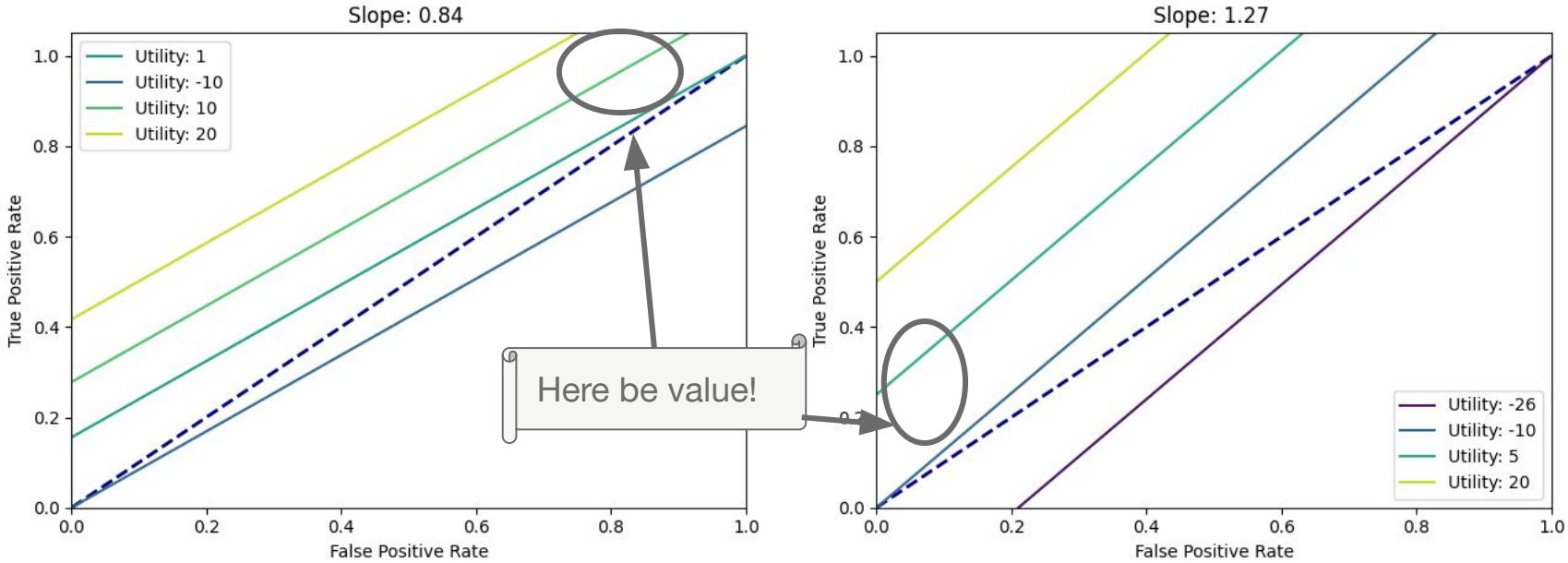
prevalence = r_p = 0.6

Best option with no model!



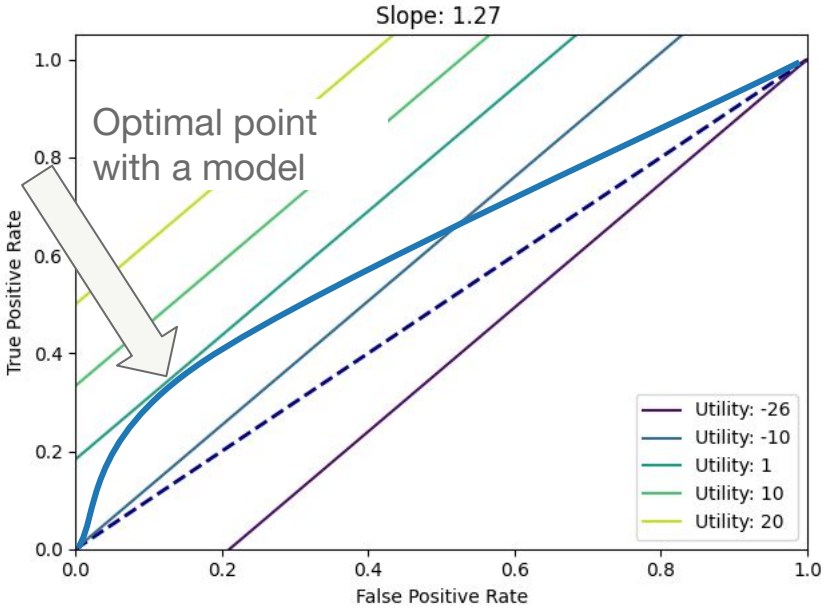
Slope determines region to focus!

Switches relative to 1



Optimal Choice with a Model

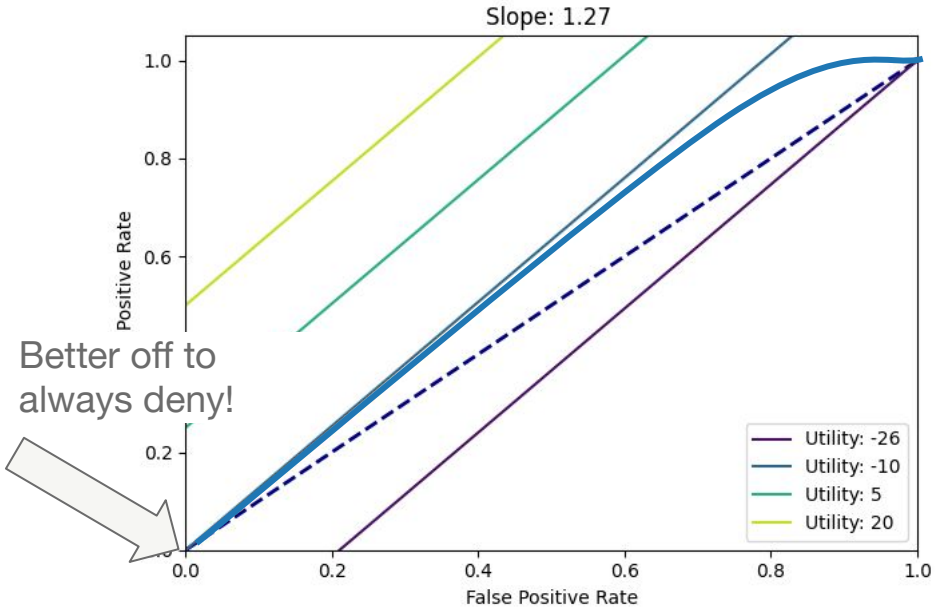
Constrained optimization problem - Lagrange Multiplier



- Optimal choice will be either
  - Tangent line (where slope of ROC curve is equal to slope)
  - 100% reject or deny
- Knowing nothing about the trade-offs, regions of high curvature are very likely to be optimal points

Optimal Choice with a Model (2)

Change the utility functions and model



- Optimal choice will be either
  - Tangent line (where slope of ROC curve is equal to slope)
  - 100% reject or deny
- Knowing nothing about the trade-offs, regions of high curvature are very likely to be optimal points
- No value in model even with same AUC because performant in the wrong region!

Slope is what matters

Intercept determines absolute value, but not tradeoffs

- When slope is really big, requires very high precision model
  - Hard to beat just reject all
- When slope is really small, requires very high recall model
  - Hard to beat just accept all
- Any thing in all terms cannot change decision
  - Sunk cost fallacy!

$$m = \frac{u_{TN} - u_{FP}}{u_{TP} - u_{FN}} * \frac{r_n}{r_p}$$

Slope is what matters

Intercept determines absolute value, but not tradeoffs

- When slope is really big, requires very high precision model
  - Hard to beat just reject all
- When slope is really small, requires very high recall model
  - Hard to beat just accept all
- Any thing in all terms cannot change decision
  - Sunk cost fallacy!

True Positive = Price (P) +

Cost of Goods and Services (COGS) +

Interest (I) +

Customer Acquisition Cost (CAC)

False Positive = COGS + Interest + CAC

True Negative = CAC

False Negative = CAC

Utility + ROC

- Curves of constant utility are (typically) straight lines
  - Slope determines behavior in absence of model
  - Slope determines what region of ROC space to try to live
- Optimal points are tangents
  - Almost always at areas of large curvature
- If the model results are bad… maybe change the business
  - Reduce customer acquisition cost
  - Change price

When does this model break?

- No feedback loops
  - Changing acceptance rates or unit economics can impact who applies
  - Change data that are needed to apply
- Can have multiple thresholds - reject, accept, and human review
- Can have variable loan terms depending on risk
- Model as an unending stream of applicants with no limit to capital
- Not always so simple to define what the business is trying to optimize
  - Maybe a bigger problem with the business!

Conclusion/Recommendations

- Business Logic = Algebra ML = Calculus
  - If we can learn backprop, SVM, etc. we can learn some business
- Understand business outcomes *early* leads to better outcomes
- Talk to the business side of the company
  - We’re don’t need to do it all by ourselves
  - Ensure that everyone agrees with the model!

Questions, Comments, Disagreements?

dillon@gardner.fyi
www.linkedin.com/in/dillon-r-gardner