Data Science in Action

Machine Learning for Self-Driving Cars

CELEHS, Harvard University

Jesse Gronsbell

Outline

- ⋆ Train-test split
- ⋆ Measures of predictive performance
- ⋆ Evaluating fairness

Today: Evaluating predictive performance

Question

How do we evaluate the performance of a classifier?

Airbnb Data: What predicts high occupancy rate?

host_response_rate	100	88	100	90	100	100	100	100	100	100
host_listings_count	1	1	2	5	2	1	5	2	4	4
host_total_listings_count	1	1	2	5	2	1	5	2	4	4
latitude	38.98	39.00	38.91	38.91	38.91	38.91	38.85	38.83	38.84	38.84
longitude	-77.02	-77.04	-77.03	-77.02	-77.02	-77.03	-77.00	-77.01	-76.98	-76.98
accommodates	4	1	4	2	5	2	3	4	4	5
bathrooms	1.5	1	1	1	2.5	1	1	2.5	1.5	1.5
bedrooms	2	1	1	1	2	1	1	2	1	1
beds	2	1	2	1	2	1	2	2	1	1
price	97	55	150	138	283	89	55	130	94	64
weekly_price	580	299	1100	1000	1650	524	295	800	559	379
monthly_price	2100	999	3700	2494	4400	1848	650	2800	1869	1129
security_deposit	250	100	100	250	500	200	150	300	95	95
guests_included	4	1	2	1	4	1	1	1	1	1
minimum_nights	4	3	2	1	3	1	1	2	1	1
maximum_nights	1125	1125	365	365	1125	1125	1125	1125	31	90
number_of_reviews	5	1	84	47	15	3	5	1	18	115
review_scores_rating	88	100	99	92	100	93	84	100	98	94
review_scores_accuracy	9	10	10	9	10	9	10	10	10	10
review_scores_cleanliness	9	6	10	8	10	9	8	10	10	10
review_scores_checkin	10	10	10	10	10	10	10	10	10	10
review_scores_communication	10	10	10	10	10	10	9	10	10	10
review_scores_location	9	10	10	8	10	10	6	6	9	9
review_scores_value	9	10	9	9	10	10	8	10	10	9
calculated_host_listings_count	1	1	2	4	1	1	5	2	3	3
reviews_per_month	0.22	1	2.91	0.87	1.06	0.64	2.73	0.45	0.99	6.1

$p = 26$ and $n = 580$

Airbnb data

Step 1: Train-test split

- To start, we need to the split the data into two sets
  - **Training set**: The data used to train the model
  - **Test set**: The data used to evaluate the model

Article by Cassie Kozyrkov

Step 1: Train-test split

Question

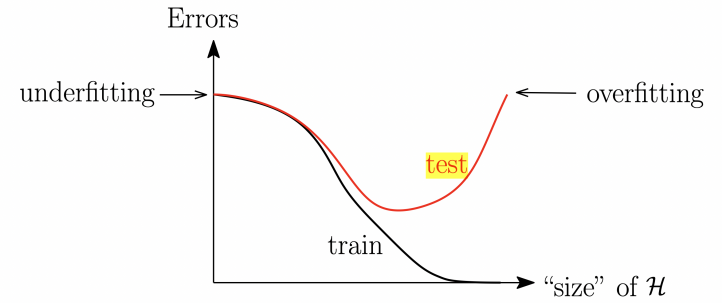
Why do we need to do this?

Step 1: Train-test split

Question

Why do we need to do this?

- ⋆ To accurately evaluate a model



Airbnb data: Model accuracy

Using 80% of the data for training we obtain the following results.

Logistic regression model

Train accuracy: 0.72
Test accuracy: 0.65

Airbnb data: Model accuracy

Using 80% of the data for training we obtain the following results.

Logistic regression model

Train accuracy: 0.72
Test accuracy: 0.65

Question

Is this good performance?

Always keep in mind...

- The **prevalence**, or $P(Y = 1)$, of the outcome
- In the Airbnb data, the prevalence of high occupancy is 34%
  - Predicting 0 for all observations would yield 66% accuracy

Always keep in mind...

- The **prevalence**, or $P(Y = 1)$, of the outcome
- In the Airbnb data, the prevalence of high occupancy is 34%
  - Predicting 0 for all observations would yield 66% accuracy

Question

Is the logistic regression model still useful?

Other criteria to evaluate the model

Question

Can you think of any other performance metrics besides accuracy?

Other criteria to evaluate the model

Question

Can you think of any other performance metrics besides accuracy?

- ⋆ False Positives, True Positives
- ⋆ False Negatives, True Negatives
- ⋆ Sensitivity, Specificity
- ⋆ Positive and Negative Predictive Values
- ⋆ ...

Confusion matrix

	Observed Positive	Observed Negative
Predicted Positive	True Positives (TP)	False Positives (FP)
Predicted Negative	False Negatives (FN)	True Negatives (TN)

Some scenarios

Question

When do we care about false positives more?

Some scenarios

Question

When do we care about false positives more?

- ⋆ Predicting an invasive medical procedure

Some scenarios

Question

When do we care about false negatives more?

Some scenarios

Question

When do we care about false negatives more?

- ⋆ Disease screening

Defining prediction metrics

	Observed Positive	Observed Negative
Predicted Positive	True Positives (TP)	False Positives (FP)
Predicted Negative	False Negatives (FN)	True Negatives (TN)

Let P be all observed positives and N all the negatives.

⋆ **Accuracy**: $(\mathrm{TP} + \mathrm{TN})/(\mathrm{P} + \mathrm{N})$

Defining prediction metrics

	Observed Positive	Observed Negative
Predicted Positive	True Positives (TP)	False Positives (FP)
Predicted Negative	False Negatives (FN)	True Negatives (TN)

- ⋆ **Sensitivity**: $\mathrm{TPR} = \mathrm{TP}/\mathrm{P} = P(\hat{Y} = 1 | Y = 1)$
  - ○ Also called **Recall**
- ⋆ **Specificity**: $\mathrm{TNR} = \mathrm{TN}/\mathrm{N} = P(\hat{Y} = 0 | Y = 0)$
  - ○ $\mathrm{TNR} = 1 - \mathrm{FPR} = 1 - \mathrm{FP}/\mathrm{N}$

Defining prediction metrics

	Observed Positive	Observed Negative
Predicted Positive	True Positives (TP)	False Positives (FP)
Predicted Negative	False Negatives (FN)	True Negatives (TN)

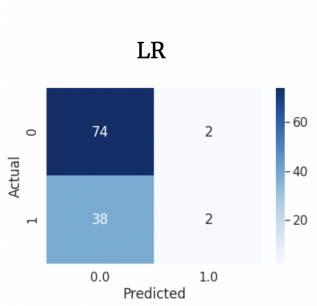
- ⋆ **Positive predictive value** or **Precision**
  $PPV = \frac{TP}{TP+FP} = P(Y = 1 | \hat{Y} = 1)$
- ⋆ **Negative predictive value**
  $NPV = \frac{TN}{TN+FN} = P(Y = 0 | \hat{Y} = 0)$

Defining prediction metrics

	Observed Positive	Observed Negative
Predicted Positive	True Positives (TP)	False Positives (FP)
Predicted Negative	False Negatives (FN)	True Negatives (TN)

- **F1**: 2 (PPV*TPR)/(PPV+TPR) = 2TP/(2TP+FP+FN)

Airbnb data: Evaluating predictive performance



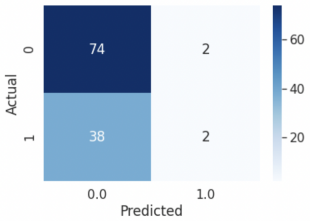
	LR
Accuracy	.65
Sensitivity/Recall	.05
Specificity/TNR	.97
PPV/Precision	.5
NPV	.66
F1	.09

Question

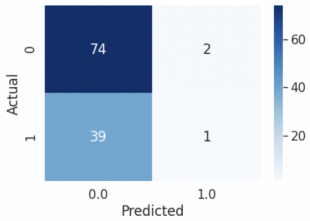
What are the take-aways here?

Always look at the confusion matrix

LR

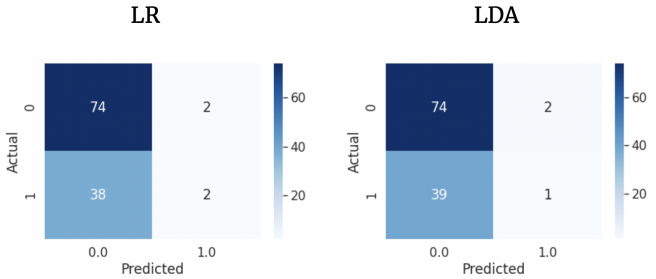


LDA



	LR	LDA
Accuracy	.65	.65
Sensitivity/Recall	.05	.025
Specificity/TNR	.97	.97
PPV/Precision	.5	.33
NPV	.66	.65
F1	.09	.05

Always look at the confusion matrix



	LR	LDA
Accuracy	.65	.65
Sensitivity/Recall	.05	.025
Specificity/TNR	.97	.97
PPV/Precision	.5	.33
NPV	.66	.65
F1	.09	.05

There is a difference of only 1 classification, but the metrics are wildly different!

Can we do a better job at evaluation?

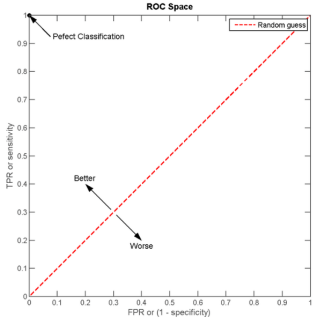
- ⋆ Remember that the logistic regression model gives us a probability of $Y = 1$ for each value of X
- ⋆ We don't have to evaluate performance at a single threshold!

Receiver Operating Characteristic (ROC) Curve

- First used during World War II to analyze radar signals following the attack on Pearl Harbor 1941
- Plots the TPR (sensitivity/recall) vs. FPR (1 - specificity)
  - Visual of the trade off between TPR and FPR across different thresholds for classification
- Now it's the most commonly used classifier evaluation criterion!

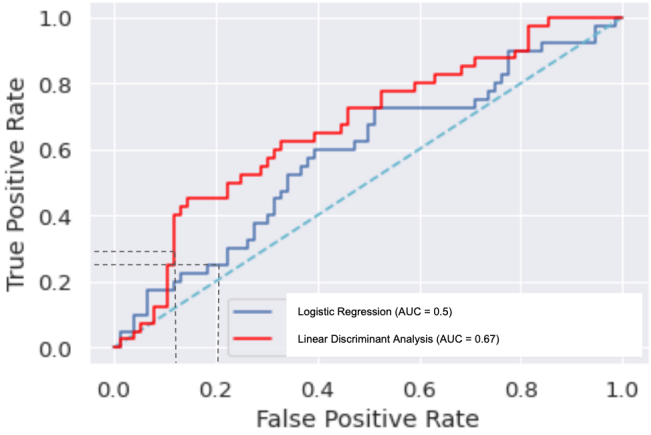
Computing the ROC curve

- ⋆ Order probabilities from highest to lowest
- ⋆ Start from the highest probability and draw a threshold at each point, each time checking TPR and FPR

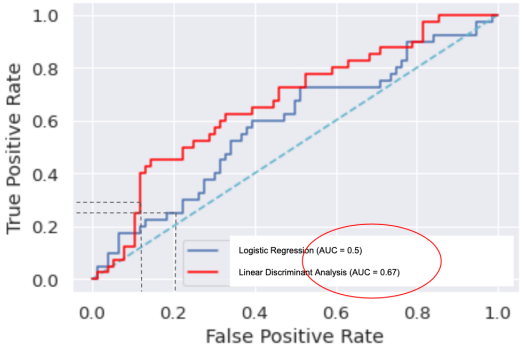


Animation

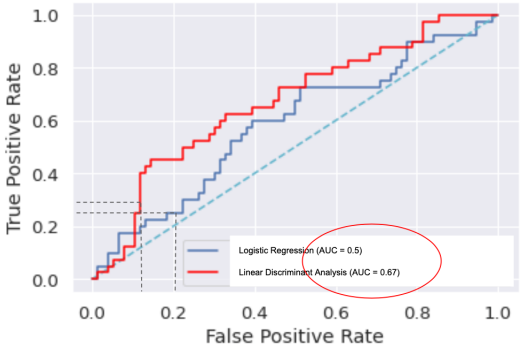
Airbnb data: ROC curve



Airbnb data: ROC curve



Airbnb data: ROC curve



⋆ **AUC**: The probability that a positive will have a higher predicted probability than a negative

ROC Curve Summary

Advantages

- ⋆ Captures a lot of information
- ⋆ Visualizes the behavior of the model at different thresholds
- ⋆ Can be used when we have imbalanced data

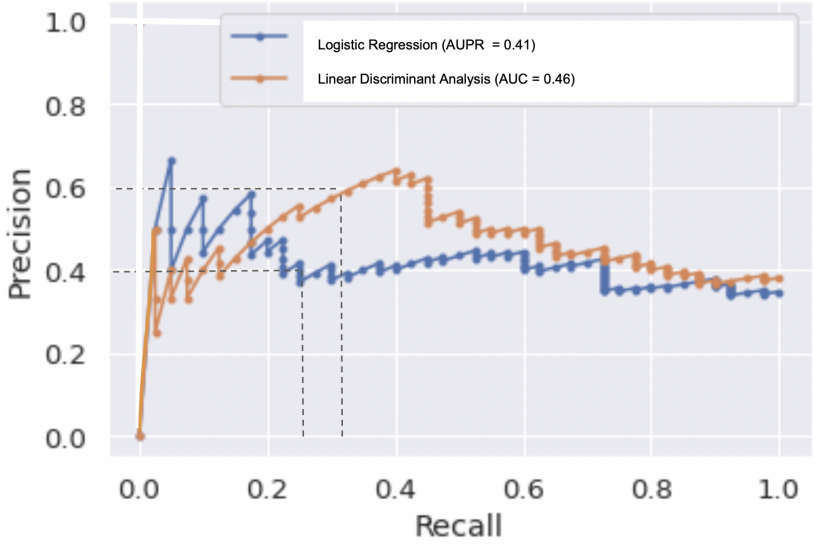
Disadvantages

- ⋆ If the data is severely imbalanced, each data point will make a big difference
- ⋆ Doesn't summarize information about PPV or NPV

Precision-Recall (PR) Curve

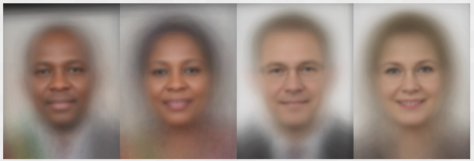
- Precision = TP/(TP+FP)
  - Proportion of correctly made positive predictions out of all positive predictions made
- Recall = TPR = TP/P
  - Proportion of correctly made positive predictions out of all that positive examples
- **PR Curve**: Compute the precision vs. recall for each threshold!
  - Focus is on the positive class

PR Curve


Logistic Regression (AUPR = 0.41)
Linear Discriminant Analysis (AUC = 0.46)
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Bias in classification

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



http://gendershades.org/overview.html

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Slide from Irene Chen

Evaluating fairness

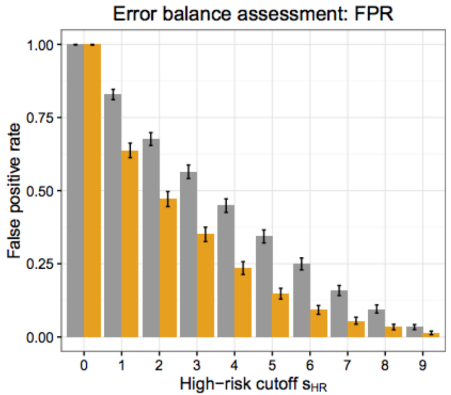
Arvind Narayanan ✓
@random_walker

I wrote up a 2-pager titled "21 fairness definitions and their politics" based on the tweetstorm below and it was accepted at a tutorial for the Conference on Fairness, Accountability, and Transparency!
Here it is (with minor edits):
docs.google.com/document/d/1bn...
See you on Feb 23/24.

> **Arvind Narayanan** ✓ @random_walker · Nov 6, 2017
> When I tell my computer science colleagues that there are so many fairness definitions, they are often surprised and/or confused. [Thread]
> twitter.com/random_walker/...
> Show this thread

Evaluating fairness

- ⋆ One idea is to check for no differences in performance across protected groups
  - eg. Equal FPR across two groups



Additional references

- ⋆ Elements of Statistical Learning Textbook
- ⋆ An introduction to ROC analysis by Tom Fawcett
- ⋆ ROC vs. PR Curve