CS 307

Fall 2023

Dalpiaz

Week 12

Some Practical Considerations

DATA ETHICS

JUST BECAUSE YOU CAN,
DOES NOT MEAN YOU SHOULD!

"IT'S TOUGH TO MAKE PREDICTIONS, ESPECIALLY ABOUT THE FUTURE"

- YOGI BERRA

PREDICT TIME VERSUS TRAIN TIME

- DO YOU EXPECT 'NEW" DATA TO COME FROM THE SAME DISTRIBUTION AS THE TRAIN DATA?

FORECASTING

- WHAT FEATURES WILL BE AVAILABLE IN 'NEW' DATA WHEN YOU NEED TO MAKE A PREDICTION?

Missing Data

- Why is it missing?
  - Survey non-response
  - Formatting
  - Measurement malfunction
  - Informative

Missing Data

- Remove rows with missingness
- Remove columns with high missingness
- Replace with a constant
- Impute with mean/median/mode of column
- Impute with consideration for other columns
- Using missingness as information

"LOOK AT THE DATA"

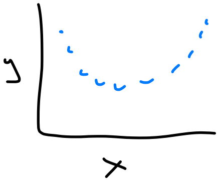
- WHAT IS THE SOURCE? HOW WAS COLLECTED / GENERATED?
  ↳ WHAT FEATURES ARE DOMAIN RELEVANT.
- n? p? SIZE?

- PRINT DATA FRAME
  .head()   "glimpse"
  .tail()

PAIRS PLOT

- VARIABLE DATA TYPES

- SUMMARY STATISTICS / VISUALIZATIONS
- HOW MANY MISSING? WHERE MISSING?



CONDITIONAL

No free Lunch

Just use RF!

- linear model
- linear model w/ regularization
- Kaggle → Boosting

One-hot vs Dummy vs Ordinal

		One-hot			Dummy	
X	ORD	X_A	X_B	X_C	X_B	X_C
A	0	1	0	0	0	0
B	1	0	1	0	1	0
C	2	0	0	1	0	1
A	0	1	0	0	0	0
B	1	0	1	0	1	0
B	1	0	1	0	1	0
C	2	0	0	1	0	1
C	2	0	0	1	0	1
A	0	1	0	0	0	0

Feature Engineering

- One-Hot
- Scaling
- Count Vectorization
- Manual