Features

Rayid Ghani

Slides liberally borrowed and customized from lots of excellent online sources

Tabular vs non-tabular data

- Structured data has some pre-existing features
- Video, Audio, Text, etc. need to have features extracted first.
- We'll talk about structured data today.

What we'll cover today

- Feature Creation/Engineering
- Feature Selection [not going to cover]

Why do we care?

- Features are hints you give your model
- Feature generation has the most impact on your model's performance
- Complexity in features allows us to use less complex models that are faster to run, easier to understand and easier to maintain.

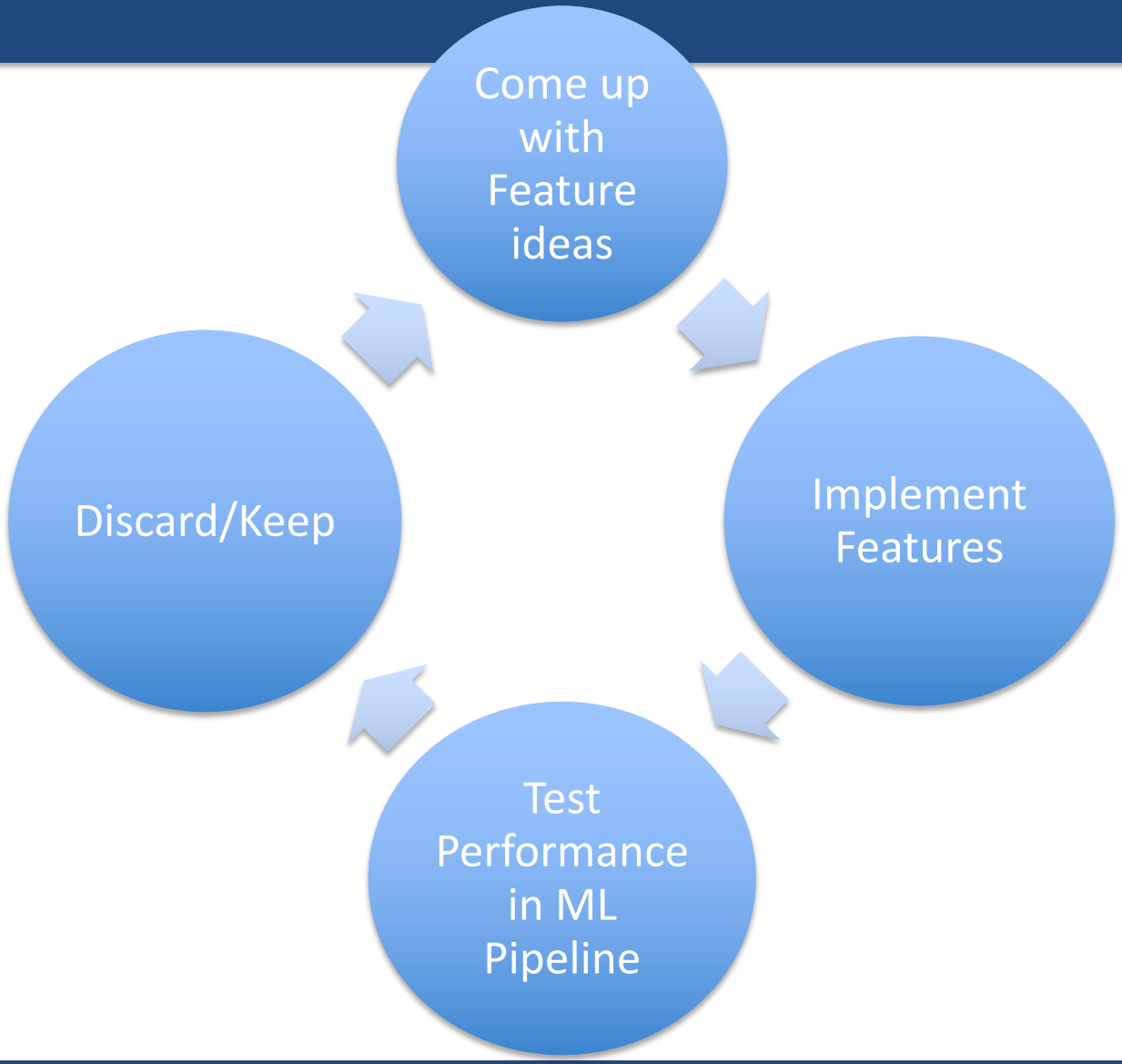
Types of features

- Simple/raw
- Dummy
- Discretization
- Aggregations
  - Spatial
  - Temporal
  - Spatiotemporal
- Disaggregations
- Expert rules/heuristics

Feature Generation

- Raw
- Categorical to Binary (Dummies)
- Features for missing values
- Discretization
- Date/Time Features
- Scaling/Normalizing
- Transformations
- Aggregations (space, time, space and time)
- Disaggregations
- Relative (compared to the average…)
- Interactions
- Expert rules

Feature Generation Process



Raw Features

- Demographic for example
  - Gender
  - Race
  - Location
- Other common attributes

Categorical to Binary

- One vs All (Dummy Variables)
  - What if there are 1000s of values?
- Groups
- Presence Vs Absence
- Other

Missing Values

- Impute (Fill in) missing values based on why you think they may be missing and what you want the model to do with those missing values
  - Mean/median/mode

- Typically, also add binary feature (dummy) for missing vs not missing in case “missingness” is predictive of the outcome

Discretization

- Equal width bins
- Equal size bins
- Entropy-based bins
- Domain-Specific bins (infant, KG, Elementary school age, middle school age, etc.)

Feature Scaling

- Usually a good idea to scale features to have similar range: [-1,1] or [0,1] for example
  - Be careful with outliers

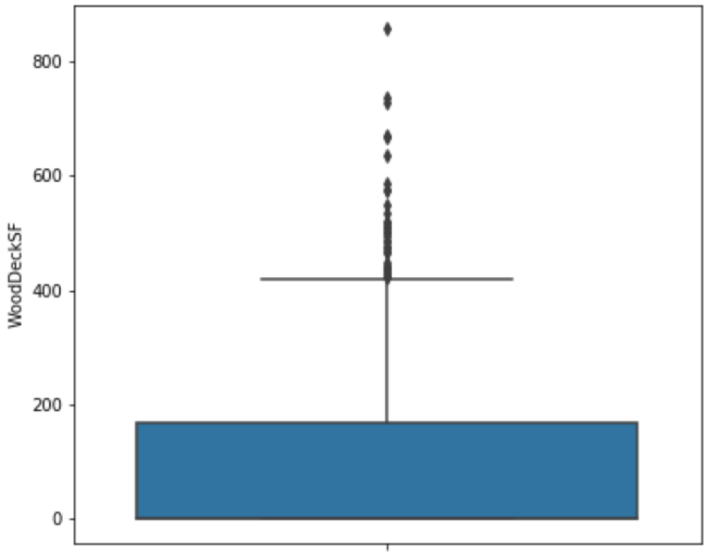
- Standardize/Normalize
  - Zero mean and unit variance

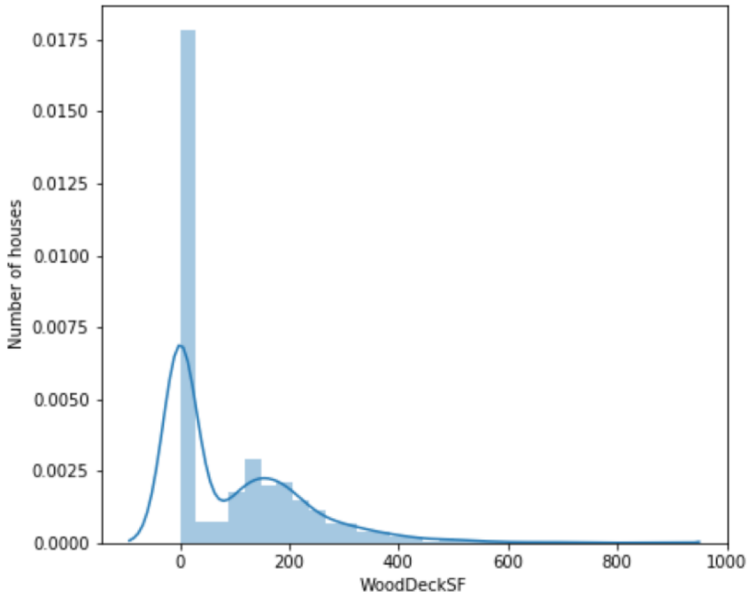
$$x_{new} = \frac{x - \mu}{\sigma}$$

  - Sklearn.preprocessing.normalize

Dealing with Outliers

- Use Boxplots to find them

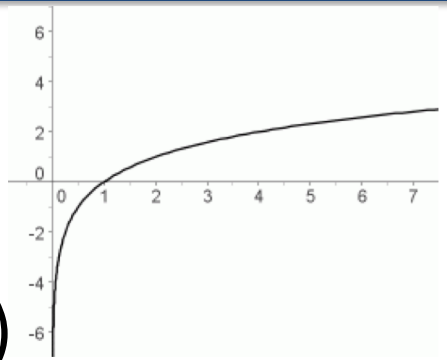




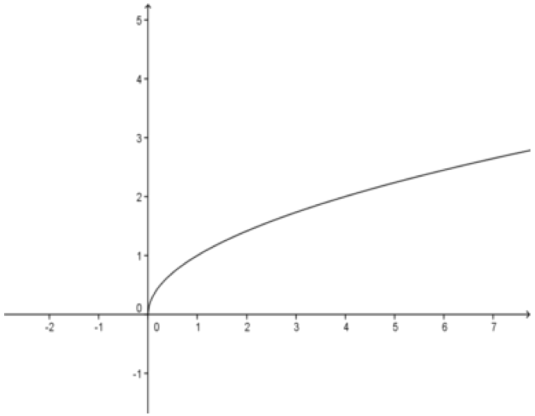
- Top-coding, bottom-coding
- Do not remove them (unless it's a data entry error)

Feature Transformations

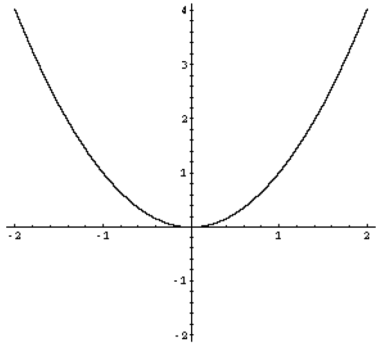
- Non-linear
  - Log (decreasing marginal utility)



  - (Square) Root



  - Squared



$y = x^2$

Aggregations

- Date differences (# of days since…)
- Aggregates over different time periods
  - Min, max, avg, stdev
  - Avg spend in the past 3 months
- Relative aggregates
  - 1.5x avg spend
- Distances
- Aggregates over different distances
- Seasonality
- slope

Feature Interactions

- Generate features for combination of features
  - Age x gender
- Allows you to use linear models but still model non linear relationships
- Random Forests are one way of discovering useful interactions

Features are also model-dependent

- Decision trees may need differences between values (dates, amounts, etc.)
- Linear models may need …

Exercise to do before next week

- Create a spreadsheet with a list of features for your project