3.05 Feature Engineering

Feature Engineering

- Process of extracting useful features from raw data using:
  - Mathematics
  - Statistics
  - Domain Knowledge

Feature Engineering

- Creating new features or transforming your existing features to get the most out of your data.

- Transformation of raw data into features that best represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

Feature Engineering

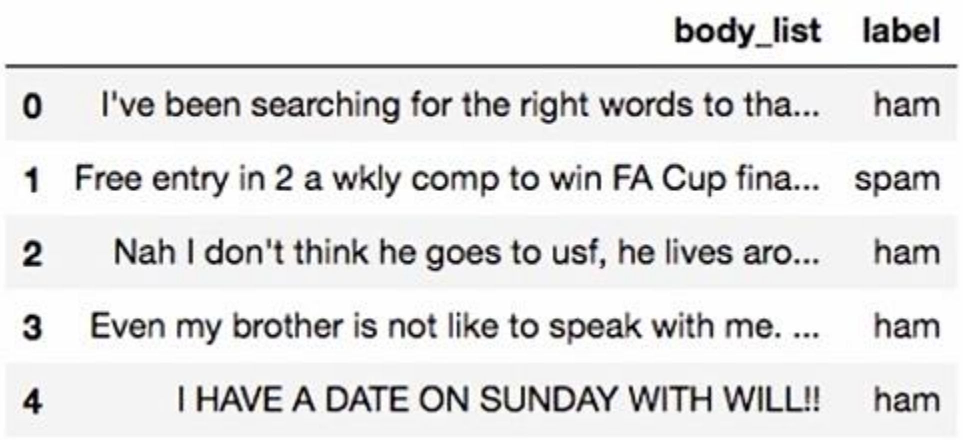
- “What is the best representation of the sample data to learn a solution to your problem?”

- Better features means:
  - Flexibility → Get good results with sub-optimal models
  - Simpler Models → Get good results with wrong parameters (mean, prop)
  - Better Results

Feature Engineering - Example

- Perform Feature Engineering on data set containing email messages

- Goal is to increase prediction accuracy of spam email detection



	body_list	label
0	I've been searching for the right words to tha...	ham
1	Free entry in 2 a wkly comp to win FA Cup fina...	spam
2	Nah I don't think he goes to usf, he lives aro...	ham
3	Even my brother is not like to speak with me. ...	ham
4	I HAVE A DATE ON SUNDAY WITH WILL!!	ham

Feature Engineering - Example

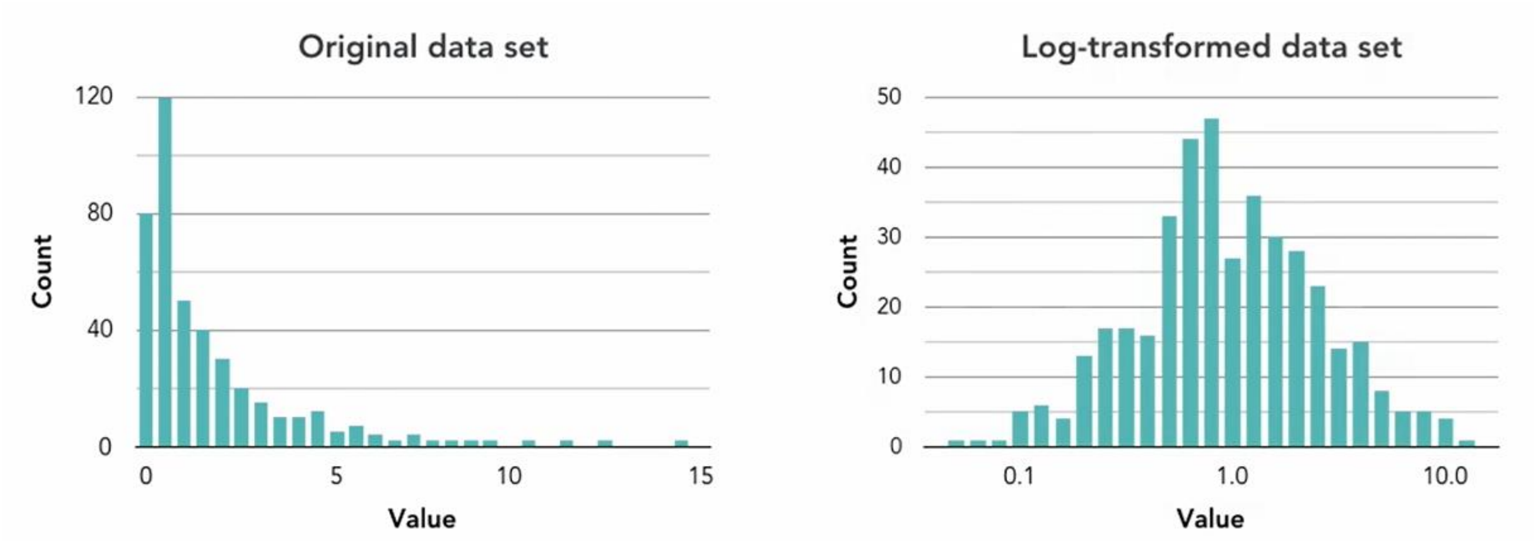
- Create features (for use case in previous slide)

  1. Length of text field
  2. Percentage of characters that are punctuation in the text
  3. Percentage of characters that are capitalized
  4. Anything else?

Feature Engineering - Example

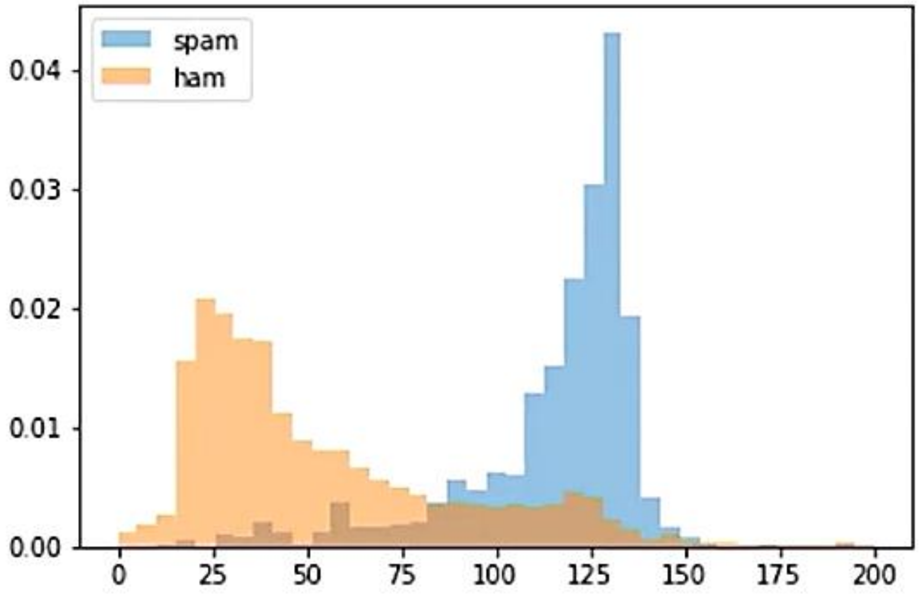
- Transformations

    1. Power Transformations (square, square root, etc.)
    2. Standardizing data

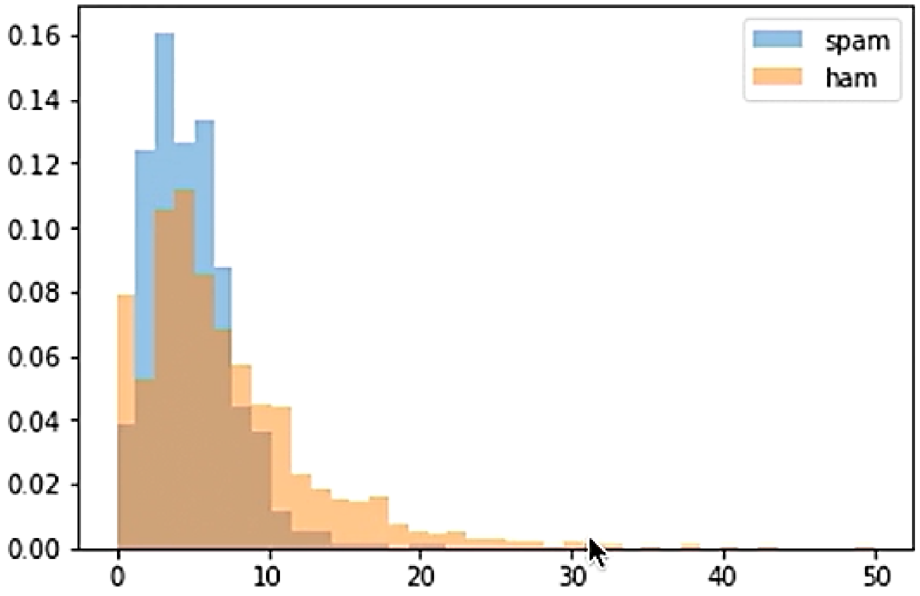


Feature Engineering - Example

- How do we **evaluate the new feature** prior to modelling?
  - Assess degree of separation between normalized plots of feature values by target group



X = Message Length of Text Message
y = Frequency



X = % of Punctuations in Text Message
y = Frequency