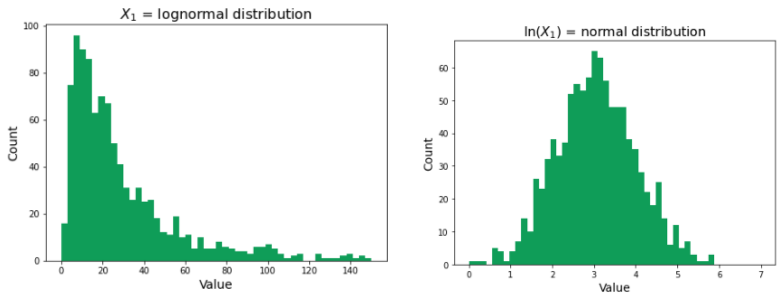
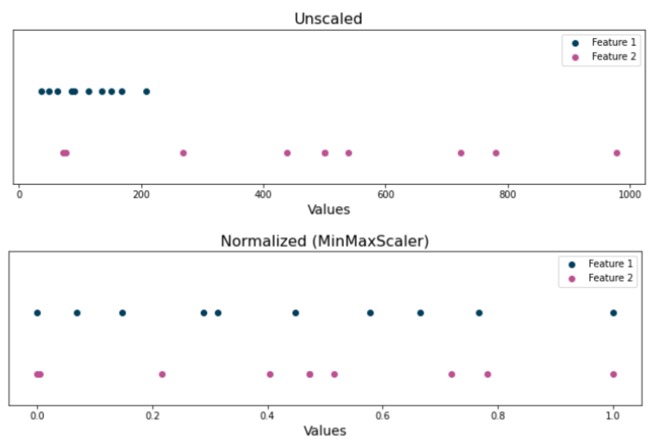
MODULE 2

Feature engineering
Process of using practical, statistical an data science knowledge to select, transform or extract characteristics, properties and attributes from raw data

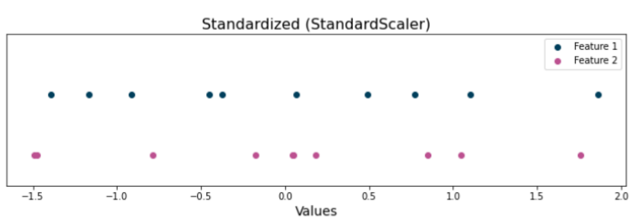
- **Feature selection**
  Select the features in the data that contribute the most to predicting the response variable
- **Feature transformation**
  Modifying existing features in a way that improves the accuracy
  - **Log normalization** – take the log of skewed feature to reduce the skew



  - **Scaling** – adjust the range of feature's value by applying normalization function
    - **Normalization (MinMaxScaler)** – transform data to reassign each value to fall within range [0,1]



    - **Standardization (StandardScaler)** – transform each value within a feature so they collectively have a mean = 0 and std = 1.



  - **Encoding** – Convert categorical data to numerical data
- **Feature extraction**
  Taking multiple features to create a new one that would improve the accuracy of the algorithm

Generally, there are **three types of features**:

1. **Predictive**: Features that by themselves contain information useful to predict the target
2. **Interactive:** Features that are not useful by themselves to predict the target variable, but become predictive in conjunction with other features
3. **Irrelevant:** Features that don't contain any useful information to predict the target

Class – for categorical variables, it means different possible values that each can take

Class imbalance – When a dataset has a predictor variable that contains more instances of one outcome than another

Class balancing
The process of changing the data by altering the number of samples to make the ratios of classes more balanced.

- **Downsampling** - Remove observations from the majority class
    - Mostly use with large datasets (>10,000)
    - Usually random removal works well
- **Upsampling** - increase the number of observations in minority class
    - Usually use with small dataset
    - Methods to add observations:
        - Duplicate samples of the minority class
        - Create synthetic, unique observation of the minority class
- **When** to do:
    - Extreme imbalance (<1%)
    - The model doesn't fit well due to few samples in minority class
    - Not when you need to use your model's output class probabilities in downstream model

Naïve Bayes

A supervised classification technique that is based on Baye's Theorem with an assumption of independence among predictors

- One of the simplest classification algorithm but still able to produce valuable results
- Really low training
- The drawback is few datasets have truly conditionally independent features
- ***BernoulliNB***: Used for binary/Boolean features
- ***CategoricalNB***: Used for categorical features
- ***ComplementNB***: Used for imbalanced datasets, often for text classification tasks
- ***GaussianNB***: Used for continuous features, normally distributed features
- ***MultinomialNB***: Used for multinomial (discrete) features

Accuracy – The number of correct predictions divided by the total number of predictions

$$Accuracy = \frac{No.of\ correct\ predictions}{No.of\ total\ predictions}$$

Precision – Proportion of positive predictions that were correct to all positive predictions

$$Precision = \frac{True\ Positives}{True\ Positives\ +\ False\ Positives}$$

Recall – Proportion of actual positives that were identified correctly to all actual positives

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

F1 score – Harmonic mean of precision and recall

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

- The score range from 0 to 1
- Penalizes low values of either metric prevents one very strong factor (precision or recall) from carrying the other when it is weaker

$\boldsymbol{F_\beta}$ **score** - Represents how many times more important recall is compared to precision

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

Accuracy, precision, recall and F1 score are useful for measuring unbalanced classes