Session 3
Metrics, Feature Engineering, Feature Selection

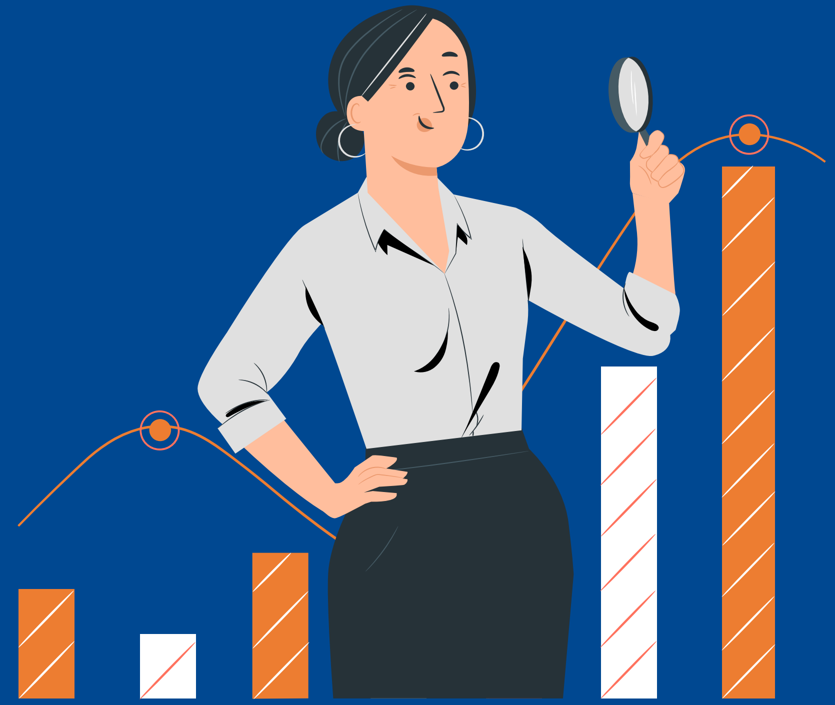
AI and Machine Learning

Hult International Business School

Michael de la Maza

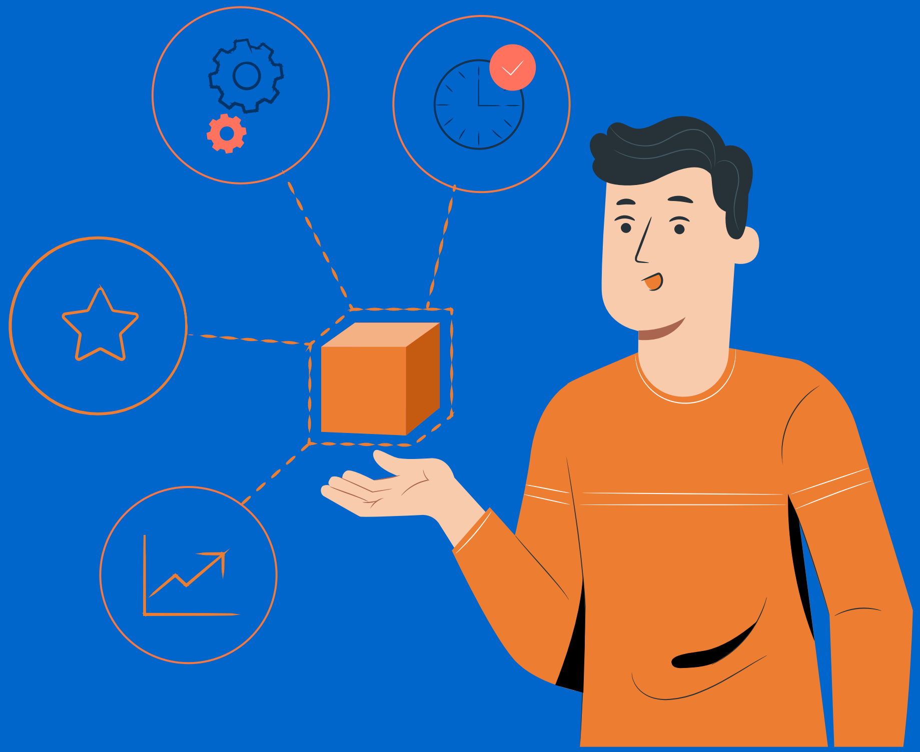
Version 1.0

Metrics



Metrics for classification problems

- Accuracy
- Precision
- Recall
- F1
- AUC (Area under curve)



Which metric to use for classification problem

ACCURACY: Good to use when classes are balanced and care equally about getting both classes right and wrong

- Good: Three classes, each with 30% – 40% of the instances
- Bad: Two classes, one with 90% of the instances (e.g., medical condition)

PRECISION: Good to use when minimizing false positives

- Good: Filtering out spam (i.e., don’t want to say something is spam when it isn’t)
- Good: Detective (i.e., don’t want to accuse the wrong person of committing a crime)

RECALL: Good to use when want to get all positives

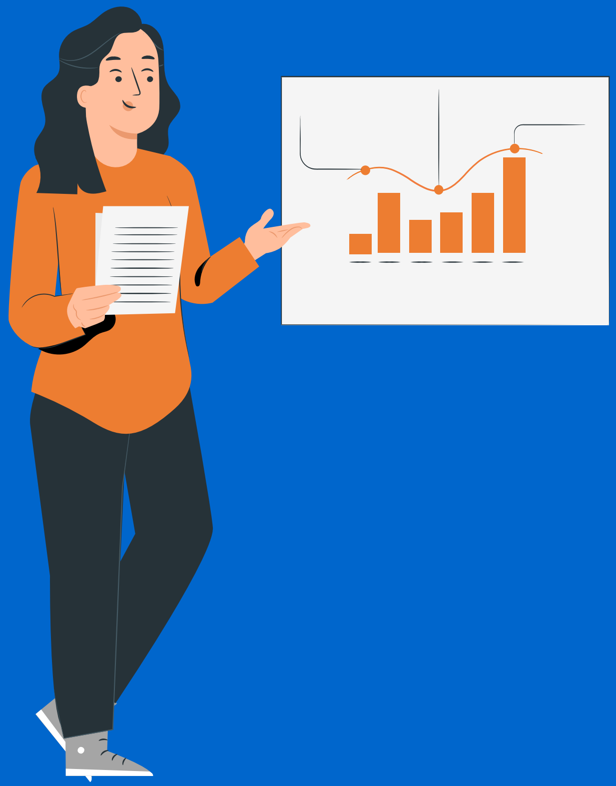
- Good: Detecting a medical condition

F1: Good to use when balancing precision and recall

- Good: Evaluating search engine. Want to get all relevant results and show only relevant results.
- Bad: Classifying emails into categories. Certainly care more about errors in some categories (e.g., work) than in other categories (e.g., fun vs. entertainment).

AUC: Good to use when want measure across all prediction thresholds

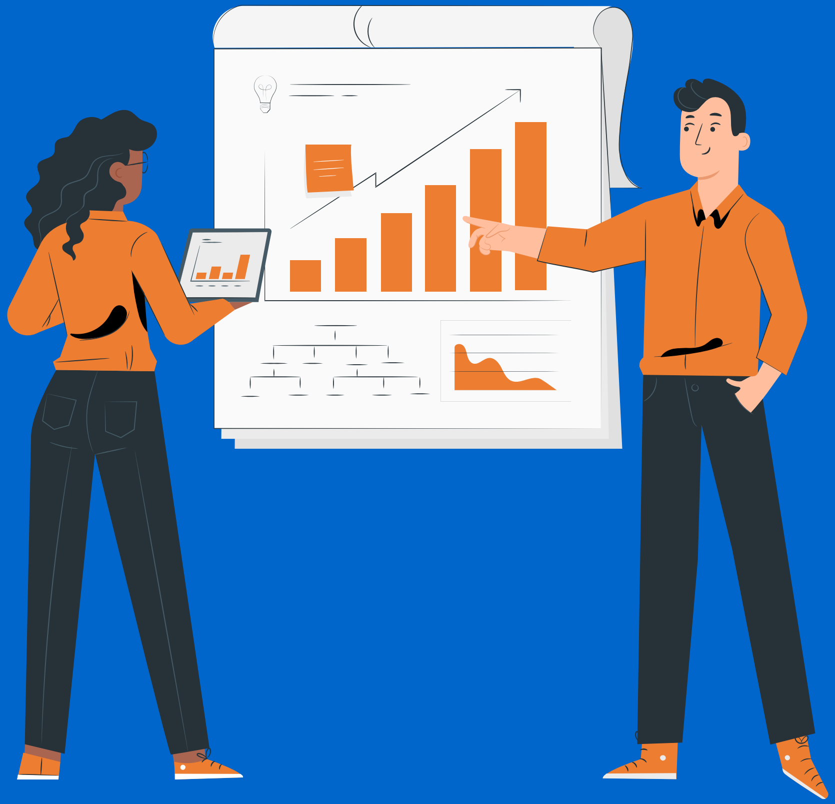
- AUC = 1. Perfect. | AUC = 0.5. Random. | AUC < 0.5. Something is wrong. Would improve by just flipping classes!



Metrics for regression problems

- Mean absolute error (MAE)
- Mean squared error (MSE)
- Root mean squared error (RMSE)
- R^2

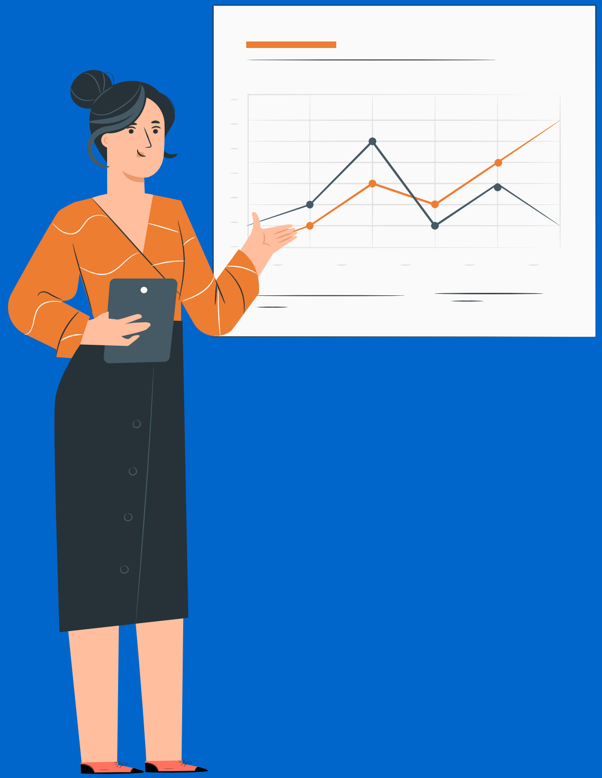
Will often use RMSE and R^2



Feature Engineering

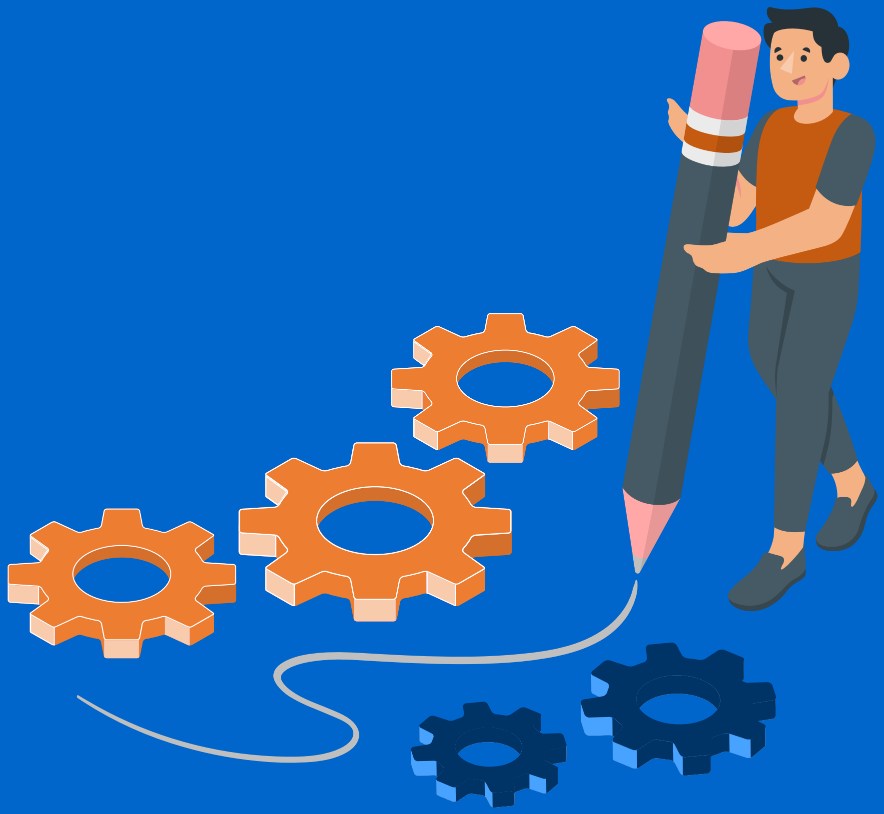
Intuitive Explanation

- In this course, we define feature engineering to mean creating derived (or calculated) features from the 'raw' features.
- Example: Housing database
  - Raw features
    - Square feet
    - Price
  - Derived feature
    - Cost per square foot = Price / Square feet
- The goal of feature engineering is to make things 'easier' for classification/regression algorithms
- min_impurity_decrease:A node will not be split if the purity decrease is less than this value.



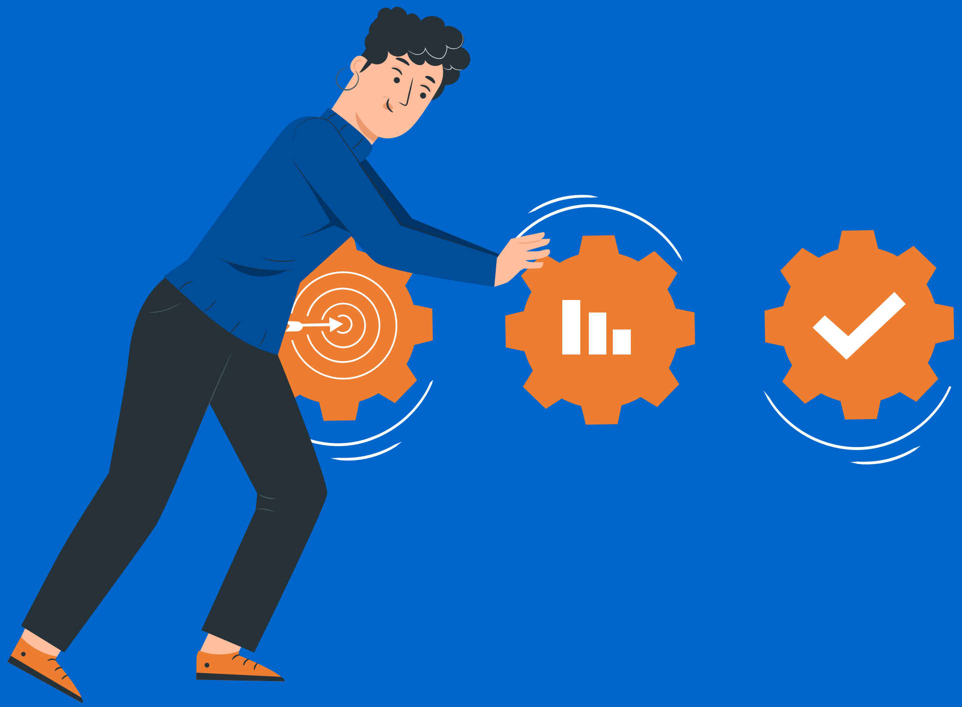
Overall process

- Create many derived values
- Select a subset of those derived values (feature selection)
- Run machine learning algorithm

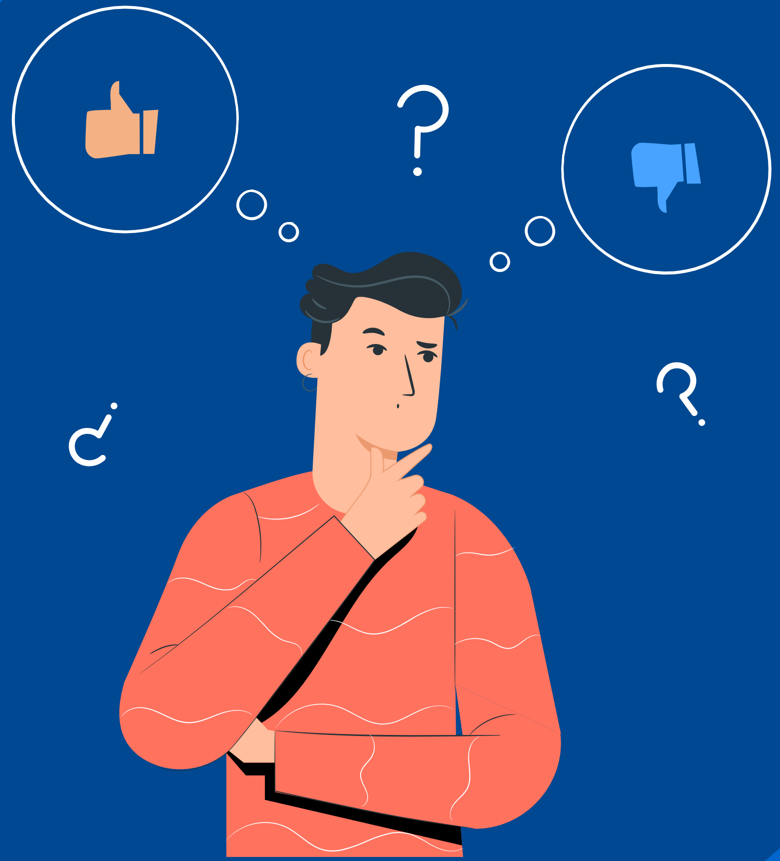


Types of feature engineering

- Polynomial features
- Binning
- Logarithm
- Normalization / scaling



Feature Selection



Feature selection methods

- Eliminate highly correlated features
- Recursive Feature Elimination (RFE)

  Def RFE(features, model, num_features_to_select):
  - While number of features > num_features_to_select:
    - Train the model with all features
    - Rank features based on importance
    - Remove the least important feature
  - Return the selected features
- SelectFromModel