Feature Pair Plots – Check comparison between pairs of features

* Check feature correlations to each other and target
  + Numerical to each other and target
  + Categorical to each other, or to numerical and target

With KNN check accuracy with different values of k

There is also KNNRegressor

* Regression evaluation types: R2

Good idea to use KNN as baseline first

Do Normalization

* MinMax Scaling
  + Scale train and test data separately, but create the scalar using only the train data
    - Otherwise data leakage

Regularization less important as amount of training data increases

* Makes sense, with more data, less likely to overfit (or create complex model)

Use make\_pipeline to create the full normalization / scaling, training, testing process

Use validation\_curve to see changes of results with different parameter values

* But not for model tuning. Use GridSearch for that

Companies can use multiple evaluation metrics on a trained model to decide which models to use

* Choose evaluation methods that match goal of application

Use DummyClassifier or DummyRegressor to check that your test scores are not just the same as random classification, most frequent class, etc. (for examples)

* Provides null metric (baseline)
* If classifier score close to dummy score:
  + Ineffective, erroneous, missing features
  + Poor choice of kernel or hyperparameter
  + Large class imbalance

For data with imbalanced classes, usually not a good idea to use Accuracy (AUC is better)

Use classification\_report to get print out of accuracy, precision, recall and F1 together

Regression evaluation uses r2\_score, mean\_squared\_error, etc. Not necessarily same as classification

Best to use K-fold cross validation and GridSearchCV for optimizing classifiers

* Default uses accuracy for evaluation
* GridSearchCV has class\_weight parameter to set how much weight is given to each class while training

Some algorithms can learn the best imputation values for data in a dataset

* Can use a predictive algorithm like kNN to get imputed values

CatBoostClassifier can work with categorical data