PREDICTING HOUSE PRICES USING MACHINE LEARNING

MODEL TRAINING

METHODS

In order to answer the question about what machine learning method is better to use for the house price problem the algorithms k-NN and Random Forest, as motivated in section 2.1, have been compared in terms of their prediction accuracy. Instead of implementing the algorithms from scratch for this study, algorithms from the scikit-learn library have been used. It is a state-of-the-art library part of the scikit suite of scientific toolkits for Python. We have also used the "Our Python" data analysis library Pandas

CLEANING DATA

Machine learning algorithms are largely implemented to only take data that is in a numeric format as input. More than half of the columns in the Ames Housing data set are non-numerical and need to be encoded, in this case using one-hot encoding and labeling. Additionally, various columns contain some empty values that have been dealt with in different ways as described in section.

ENCODING CATEGORICAL DATA

Many of the variables of the data set are categorical, and take on a limited set of values. One example is the nominal variable ”Street” which represents the type of road access to the property and takes on the values ”Grvl” for gravel and ”Pave” for paved.

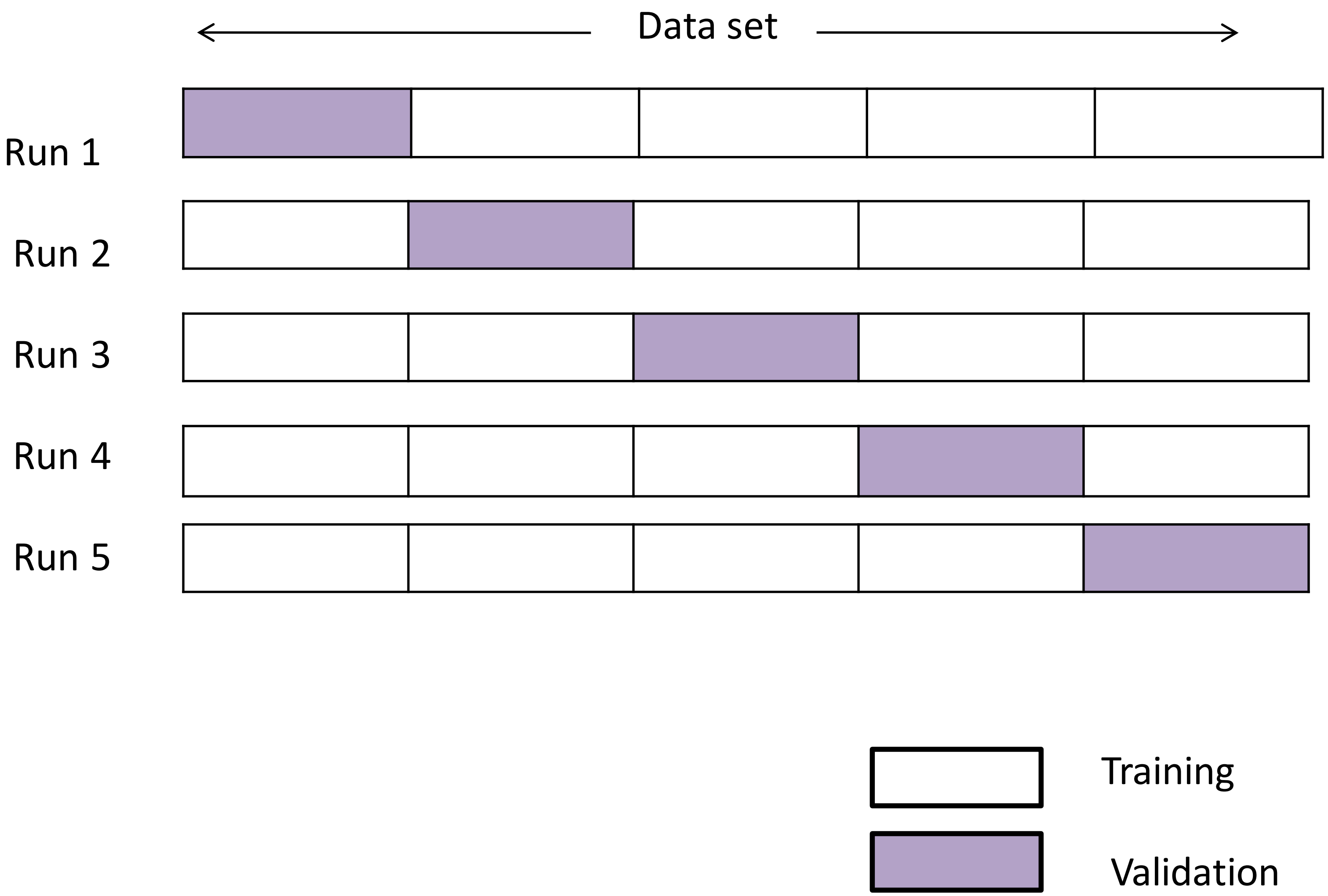
LAND SLIDE
GTL
MOD
SEV
GTL

LAND SLOPE
1
2
3
1

LABEL OF ORIGINAL VARIABLES

Splitting the data

The data set is used in two ways. First to train the algorithm, and then to test it, and for these intents we have split the set in two. The ratio between the number of rows in the training data and the test data needs to be carefully selected. If the test data is too small the result is less convincing since it is not tested on a large variety of rows. Increasing the test data size improves reliability but reduces the number of rows in the training data which causes the model to predict worse. A way of mitigating the effects of a larger test set is to use cross-validation, which is used for this experiment.



Data set

Run 1

Run 2

Run 3

Run 4

Run 5

Training

Validation

THANK YOU!