Fundamentos de Machine Learning para Geometalurgia

Data Preparation

24 de abril al 4 de mayo 2023

Agenda

Machine Learning basis

python

NumPy

pandas

matplotlib

seaborn

Case study

Univariate
Exploratory Data Analysis (EDA)

Data Preparation

Regression model (proxy) for geometallurgical parameter Ai

Scikit-learn

Simple and efficient tools for predictive data analysis. Built on NumPy, SciPy, and matplotlib.

Scikit-learn is an open source machine learning library. It provides various tools for model fitting, data preprocessing, model selection, model evaluation, and many other utilities.

https://scikit-learn.org/

Scikit-learn

https://scikit-learn.org/

```python
"""
    Author: Dario Radečić
    Article: Let's Make a KNN Classifier from Scratch
    Publication: Towards Data Science
"""

class KNearestNeighbors(object):
    def __init__(self, k):
        self.k = k

    @staticmethod
    def _euclidean_distance(v1, v2):
        v1, v2 = np.array(v1), np.array(v2)
        distance = 0
        for i in range(len(v1) - 1):
            distance += (v1[i] - v2[i]) ** 2
        return np.sqrt(distance)

    def predict(self, train_set, test_instance):
        distances = []
        for i in range(len(train_set)):
            dist = self._euclidean_distance(train_set[i][:-1], test_instance)
            distances.append((train_set[i], dist))
        distances.sort(key=lambda x: x[1])

        neighbors = []
        for i in range(self.k):
            neighbors.append(distances[i][0])

        classes = {}
        for i in range(len(neighbors)):
            response = neighbors[i][-1]
            if response in classes:
                classes[response] += 1
            else:
                classes[response] = 1

        sorted_classes = sorted(classes.items(), key=lambda x: x[1], reverse=True)
        return sorted_classes[0][0]

    @staticmethod
    def evaluate(y_true, y_pred):
        n_correct = 0
        for act, pred in zip(y_true, y_pred):
            if act == pred:
                n_correct += 1
        return n_correct / len(y_true)
```

https://towardsdatascience.com/lets-make-a-knn-classifier-from-scratch-e73c43da346d

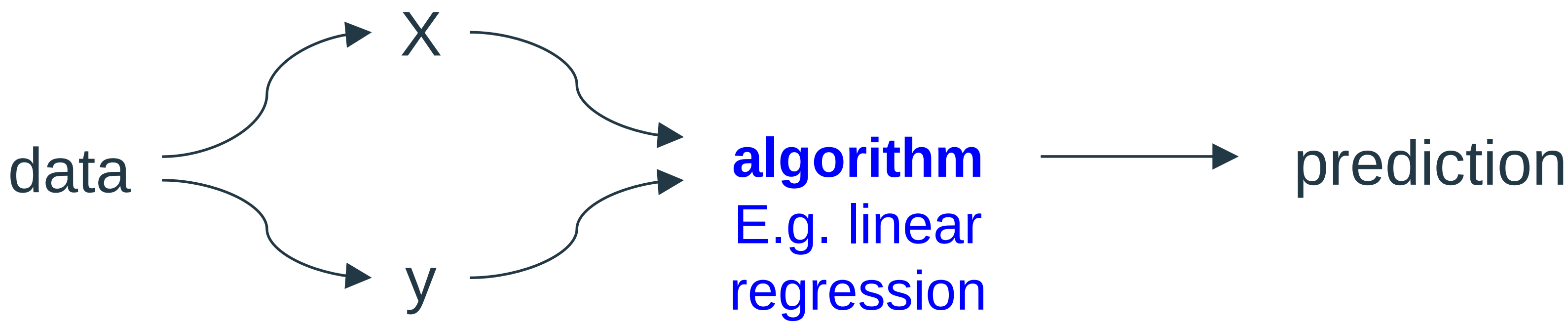
Feature-engine

Feature-engine is an open source Python library with the most exhaustive battery of transformers to engineer features for use in machine learning models.

Feature-engine

https://trainindata.medium.com/feature-engine-a-new-open-source-python-package-for-feature-engineering-29a0ab88ea7c

Machine Learning Process



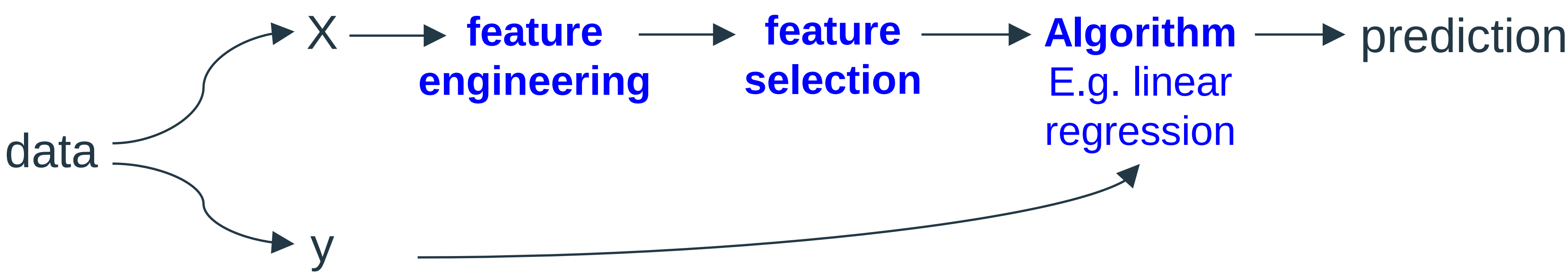
X
(predictor matrix)

information such as mineralogy, minzone, etc.

y
(target variable)

Ai values

Machine Learning Process



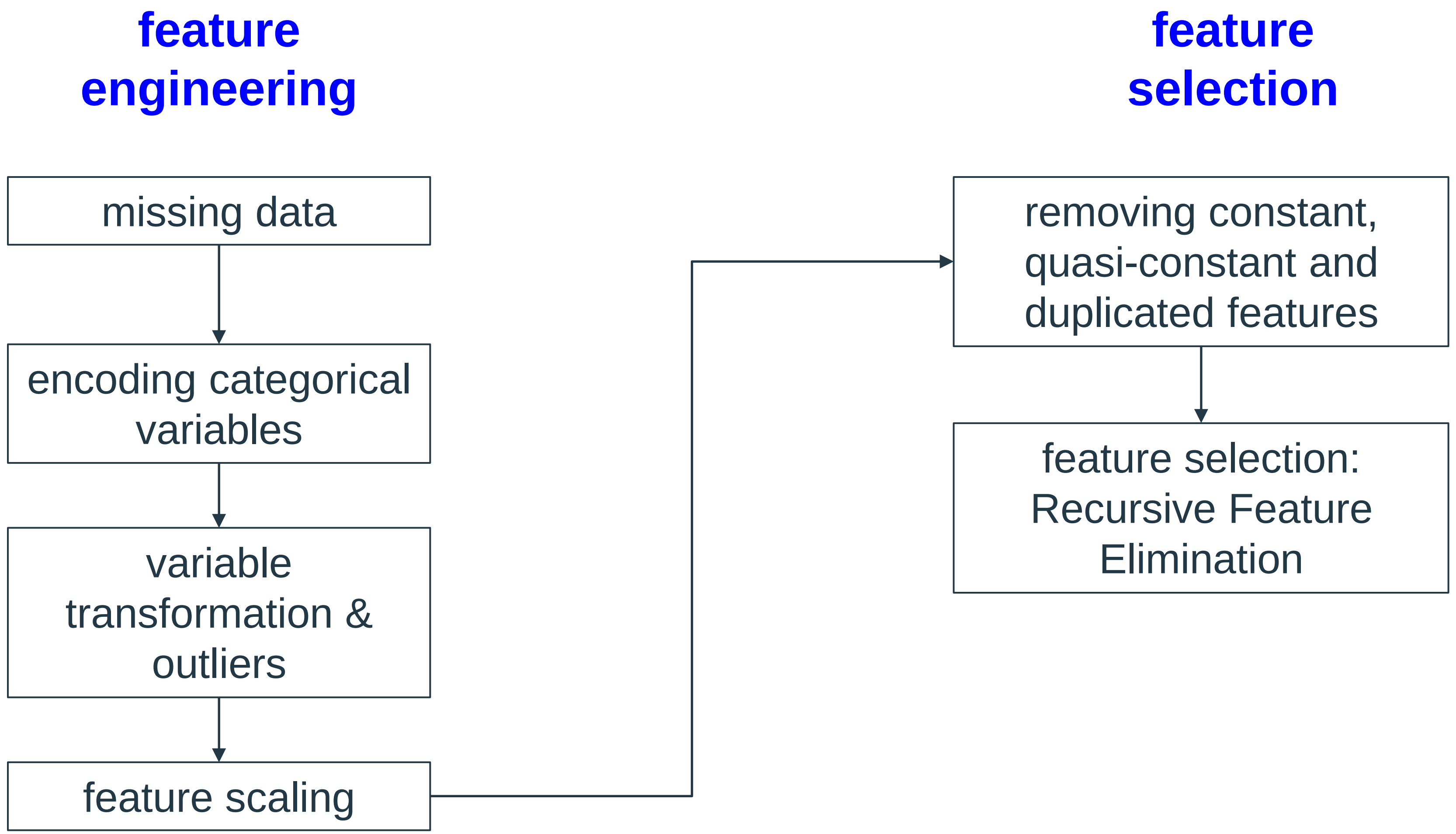
X
(predictor matrix)

information such as mineralogy, minzone, etc.

y
(target variable)

Ai values

Machine Learning Process



Encoding

Some algorithms, such as decision tree can work directly with categorical data. However, most require inputs or outputs variables to be numeric value. This means that any categorical data must be mapped to integers.

https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd

Label Encoding

Converting each category in a column to a number.

Lithology	Code
Gravel	1
Andesite	2
Tuff	3
Porphyry	4

Any problem with this approach?

The algorithm might misunderstand that data has some kind of hierarchy/order 1 < 2 < 3 < 4 and might give 4X more weight to ‘Porphyry’ in calculation then than ‘Gravel’.

https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd

One Hot Encoding

OHE converts each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns.

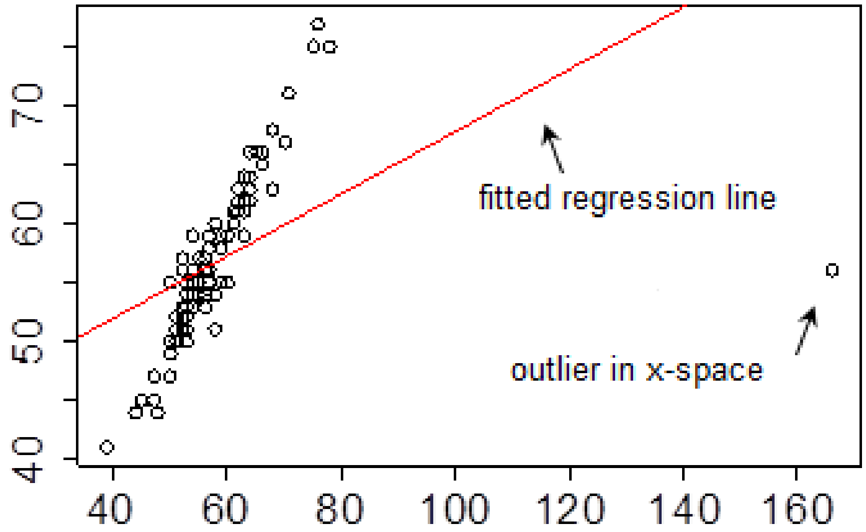
Sample	Lithology
A	Gravel
B	Gravel
C	Andesite
D	Andesite
E	Tuff
F	Porphyry
G	Tuff
H	Tuff

OHE →

Gravel	Andesite	Tuff	Porphyry
1	0	0	0
1	0	0	0
0	1	0	0
0	1	0	0
0	0	1	0
0	0	0	1
0	0	1	0
0	0	1	0

Outliers

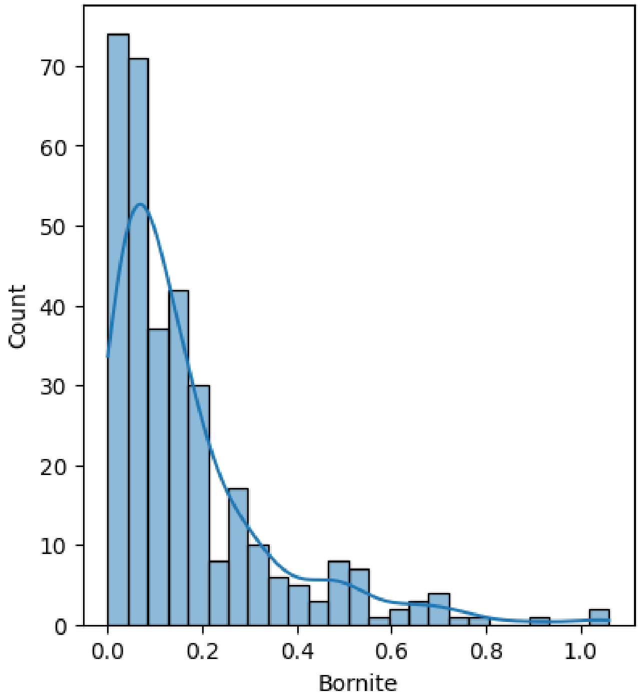
Outliers can effect regression, producing a less accurate prediction.

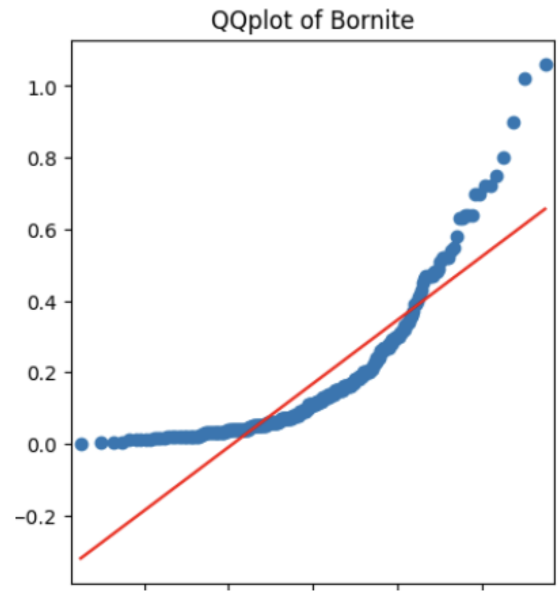


https://towardsdatascience.com/linear-regression-assumptions-why-is-it-important-af28438a44a1#:~:text=The%20linear%20regression%20algorithm%20assumes,important%20to%20validate%20this%20assumption.

Variable transformation

For some algorithms, such as linear regression, it is assumed that there is a linear relationship between continuous predictors and target. Otherwise, the accuracy of the regression may be reduced.





https://towardsdatascience.com/linear-regression-assumptions-why-is-it-important-af28438a44a1#:~:text=The%20linear%20regression%20algorithm%20assumes,important%20to%20validate%20this%20assumption.

Variable transformation

The aim is transform non normal data to data that fits a normal distribution. Transformation is done by the use of functions such as logarithm, squared root, power, etc. In this course is used Yeo-Johnson transformation (data includes zeros), which is similar to Box-Cox group of transformations. The parameter **λ** produces the best fitting transformation.

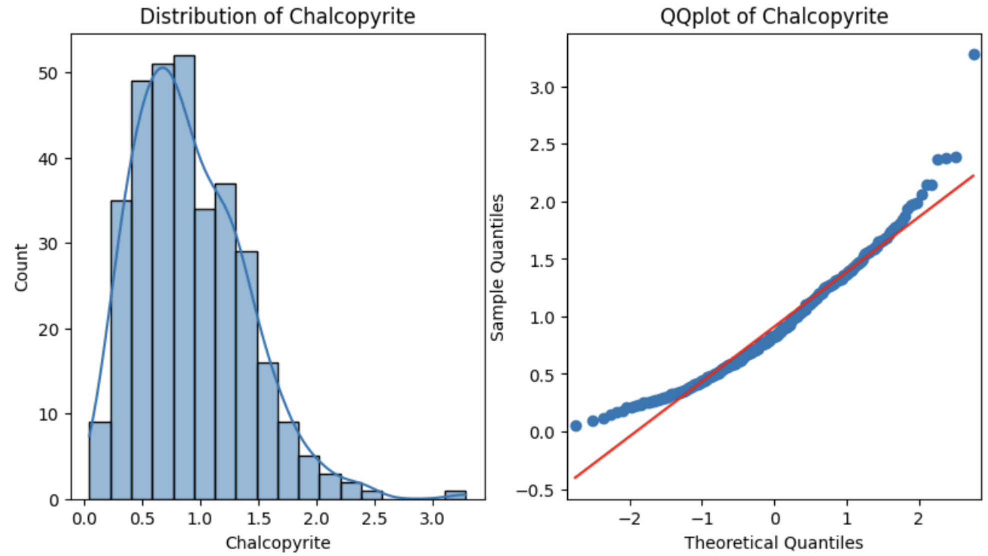
$$y = x^{\lambda} \; for \; \lambda \neq 0 \; and \; y = \ln(x) \; for \; \lambda = 0$$

$$\boxed{\begin{array}{lr}
\lambda = -1.0\,, & x_i(\lambda) = \dfrac{1}{x_i} \\
\lambda = -0.5\,, & x_i(\lambda) = \dfrac{1}{\sqrt{x_i}} \\
\lambda = \;\;\;0.0\,, & x_i(\lambda) = \ln(x_i) \\
\lambda = \;\;\;0.5\,, & x_i(\lambda) = \sqrt{x_i} \\
\lambda = \;\;\;2.0\,, & x_i(\lambda) = x_i^2
\end{array}}$$

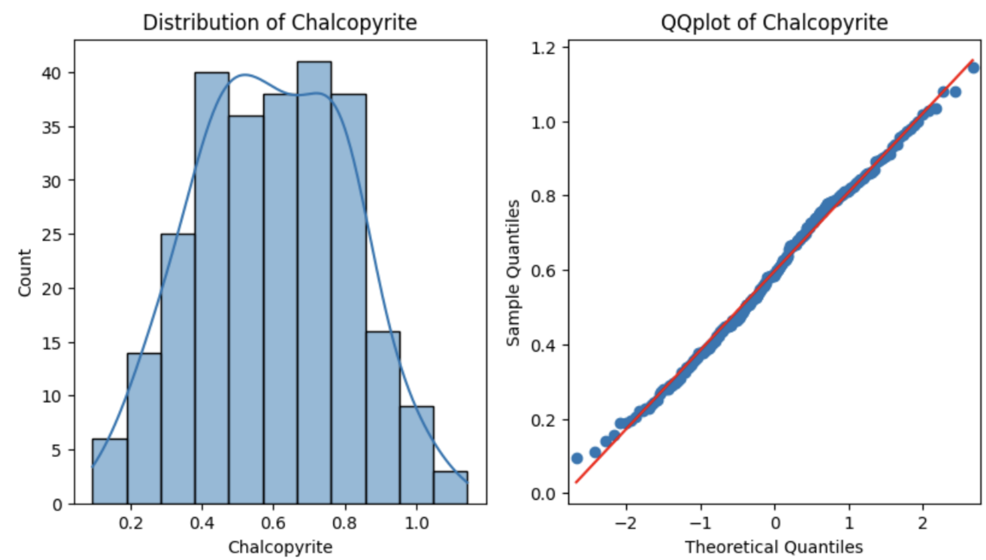
https://sigmamagic.com/blogs/how-do-i-transform-data-to-normal-distribution/

Variable transformation

before transformation

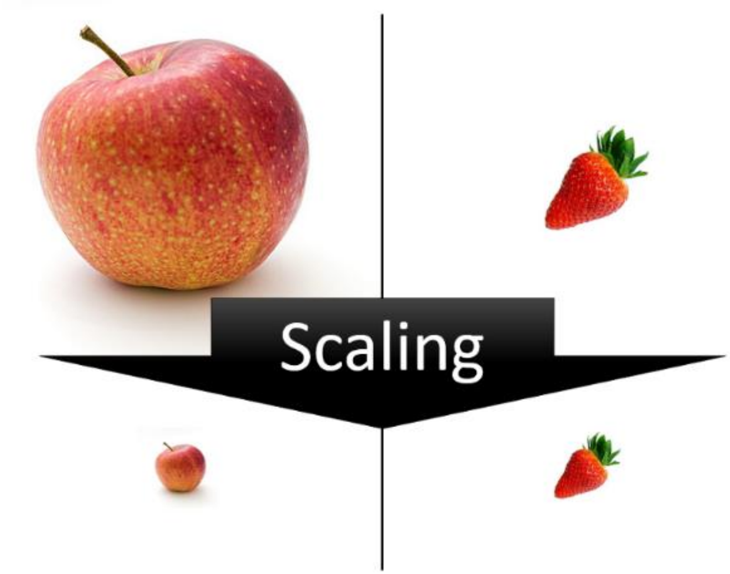


after transformation



Feature scaling

Essential for machine learning algorithms that calculate distances between data, such as K-nearest neighbors (KNN) and K-Means. If not scale, the feature with a higher value range starts dominating when calculating distances. Scaling can significantly improve model performance.



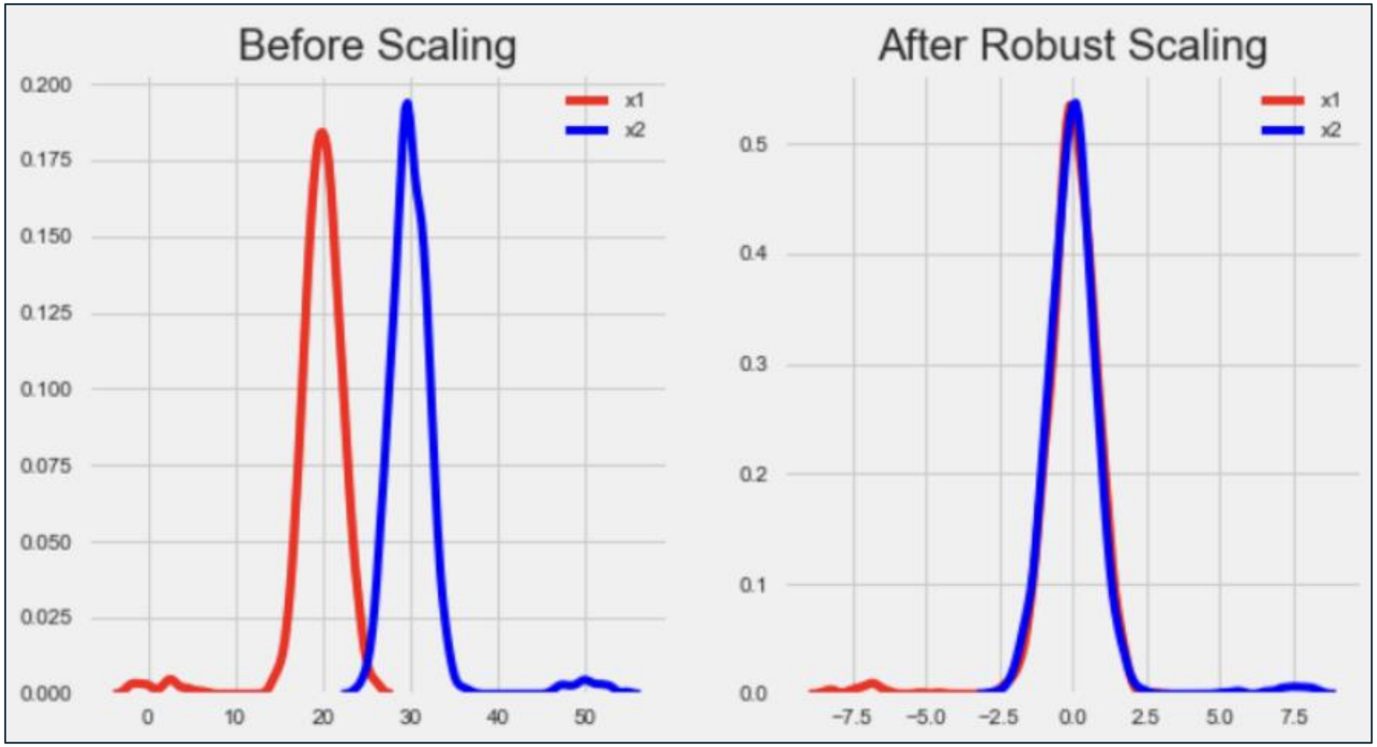
https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35

Robust Scaler

Scales features using statistics that are robust to outliers. This method removes the median and scales the data in the range between 1st quartile and 3rd quartile:

$$\frac{x_i - Q_1(x)}{Q_3(x) - Q_1(x)}$$

Robust Scaler

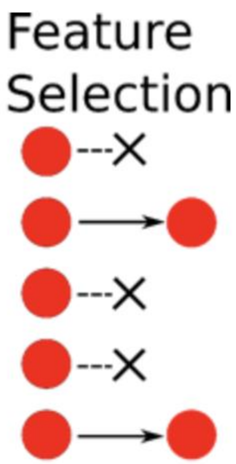


https://www.geeksforgeeks.org/standardscaler-minmaxscaler-and-robustscaler-techniques-ml/

Feature selection

Process of reducing the number of input variables when developing a predictive model. Advantages:

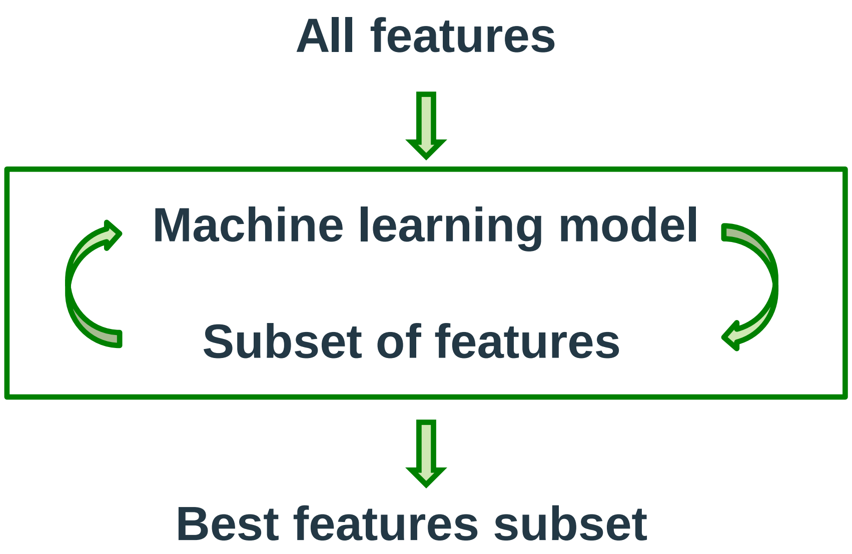
- Simple models are easier to interpret
- Shorter training times
- Reducing overfitting
- Easy to implement
- Reducing data error



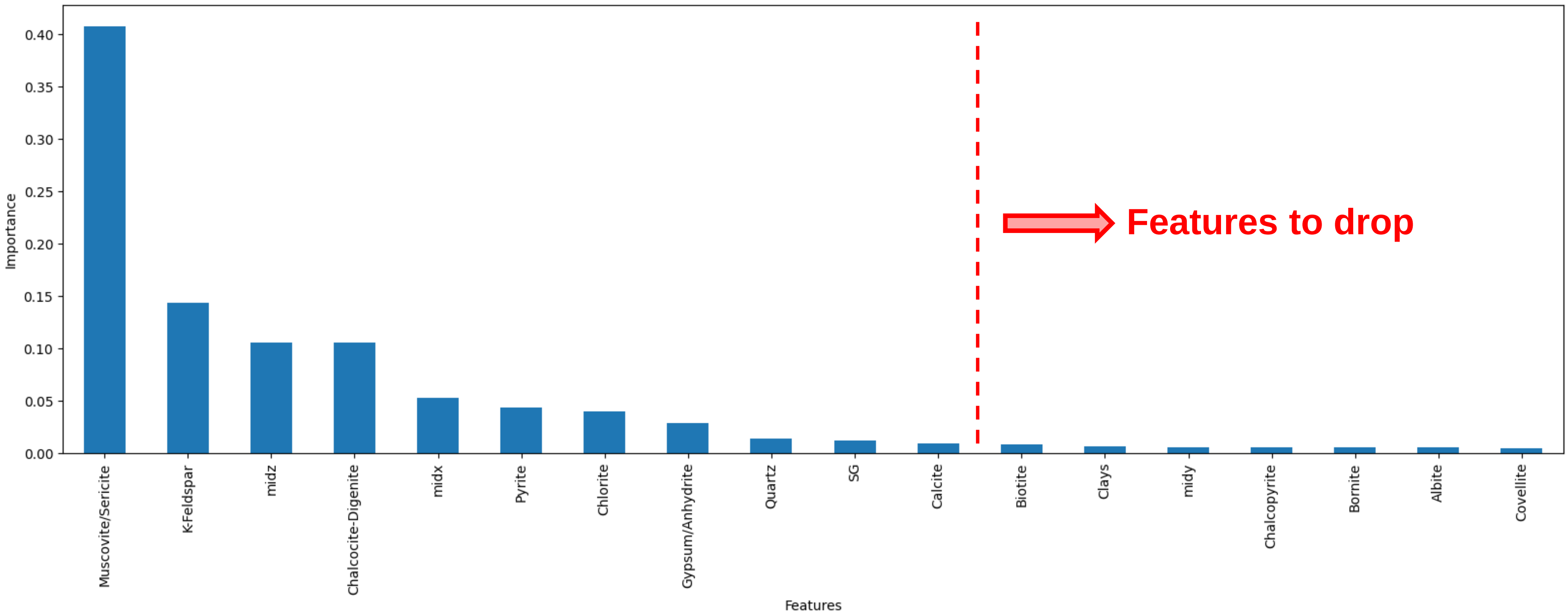
https://www.udemy.com/course/feature-selection-for-machine-learning/learn/lecture/9341700#overview
https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Recursive Feature Elimination (RFE)

RFE starts by building a model on the whole set of predictors and computing an importance score for each predictor. The least important predictor(s) are then sequentially removed, the model is re-built, and importance scores are computed again.



Recursive Feature Elimination (RFE)




Q
&
A