Train In Data

Correlation

Correlation

- Correlation is any statistical relationship or association between 2 random variables.
- Correlations are useful because they can indicate a predictive relationship.
- If a feature is highly correlated with the target variable, we can predict the target values from the feature values.

Pearson's correlation coefficient

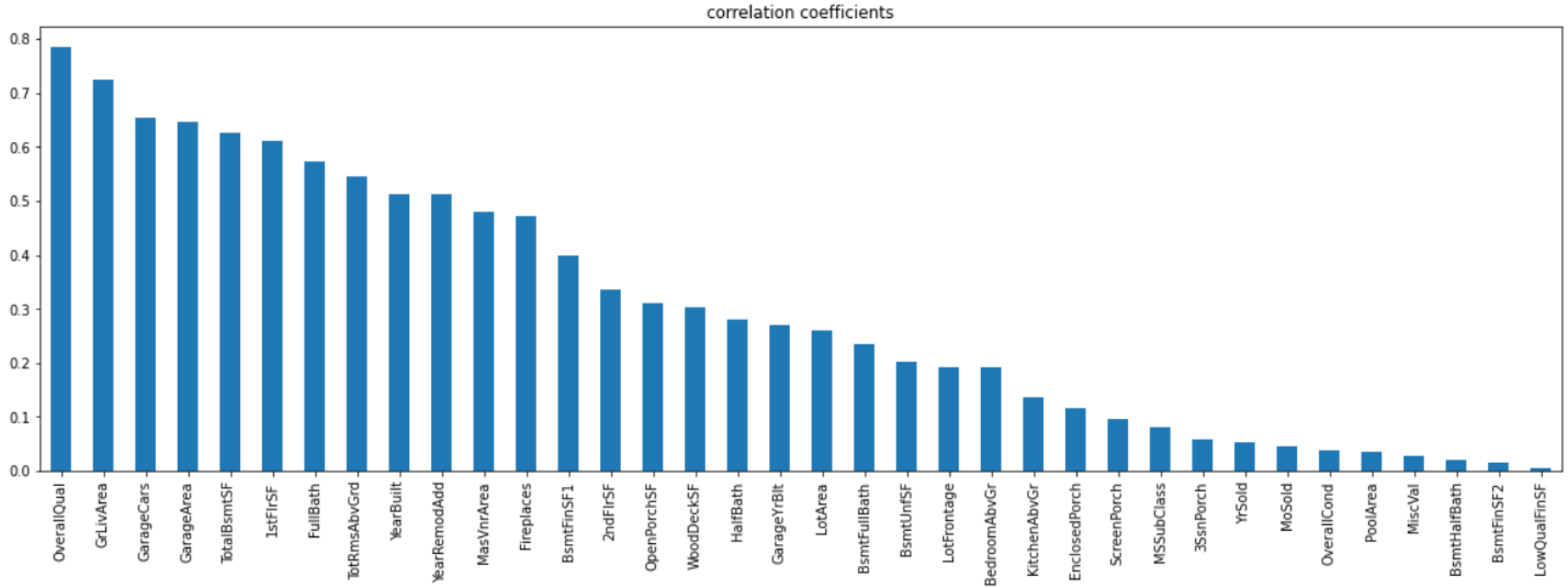
$$R = \sum \frac{(X - \overline{X})(Y - \overline{Y})}{\sigma_x \sigma_y}$$

R is large when the more X deviates from its mean, the more Y deviates from its mean.

R varies between -1 and 1, and we care about the absolute values (for feature selection).

Ranking the correlation coefficients

We can rank the features based on their correlation coefficient with the target.



Asses correlation statistically

$$t = \frac{r\sqrt{N-2}}{1-r^2}$$

R is Pearson's correlation coefficient.

N is the sample size.

t follows the *t*-student's distribution.

Knowing t ➔ p-value ➔ the probability of getting that R if 2 samples are NOT statistically associated.

Correlation assumptions

Pearson's correlation evaluates linear associations.

If we use the t-statistic ➔ it assumes that the variables are normally distributed.

Correlation: Scikit-learn

- **`f_regression`**: returns t and p-values.
  - Rank features
  - Larger t or smaller p-values ➔ important features
- **`SelectKBest:`** select best k features
- **`SelectPercentile:`** select features in top percentile

Train In Data

THANK YOU

www.trainindata.com