Strive

SCHOOL

/ Feature Selection

Is the process where you (automatically or manually) select those features which contribute most to your prediction variable.

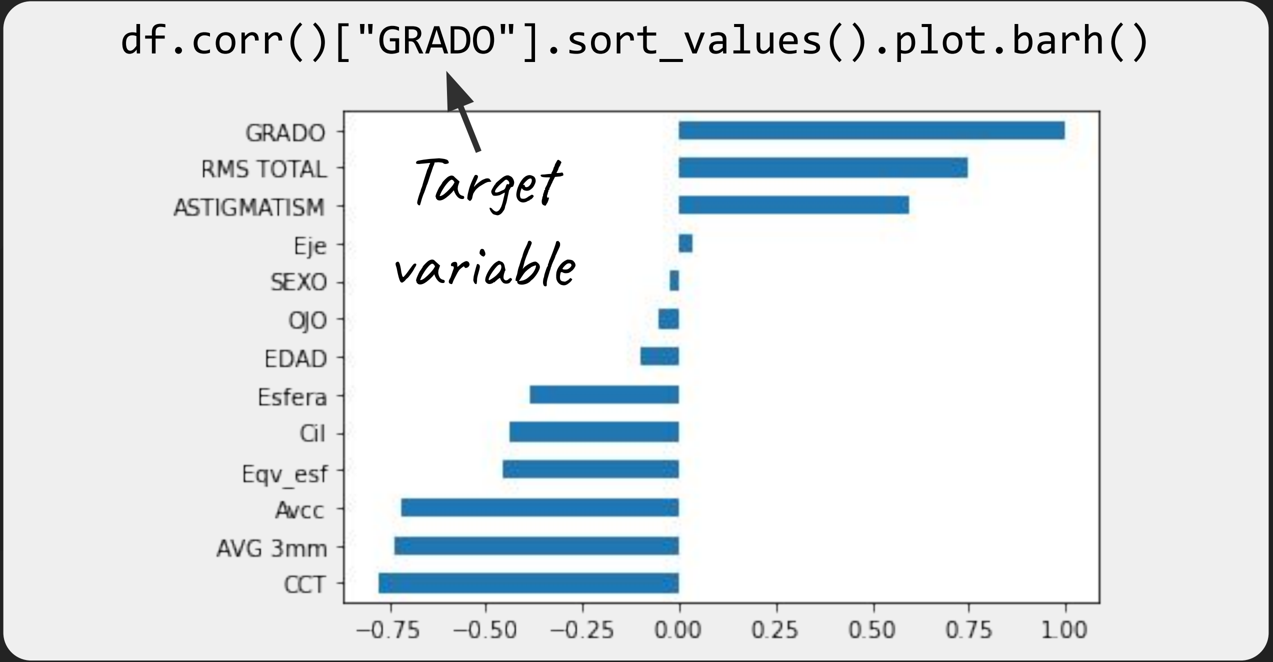
Feature Selection

/ You need to understand the problem and develop an intuition about those features you will need to solve your problem and remove unnecessary features.

What features will I need to predict Y ??

Correlation with target variable

/ Correlation with the target variable is a great tool. And a great place to start. You can plot correlation in one line:



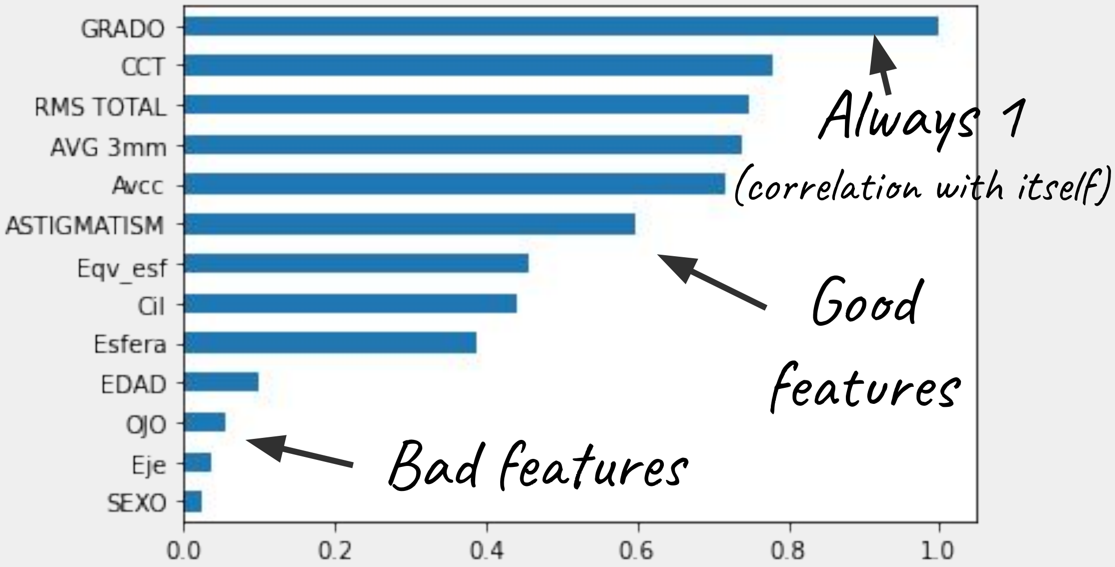
```
df.corr()["GRADO"].sort_values().plot.barh()
```

Target variable

Correlation with target variable

/ Even better is to show the absolute values of the correlation.

```
df.corr()["GRADO"].abs().sort_values().plot.barh()
```



Always 1

(correlation with itself)

Good features

Bad features

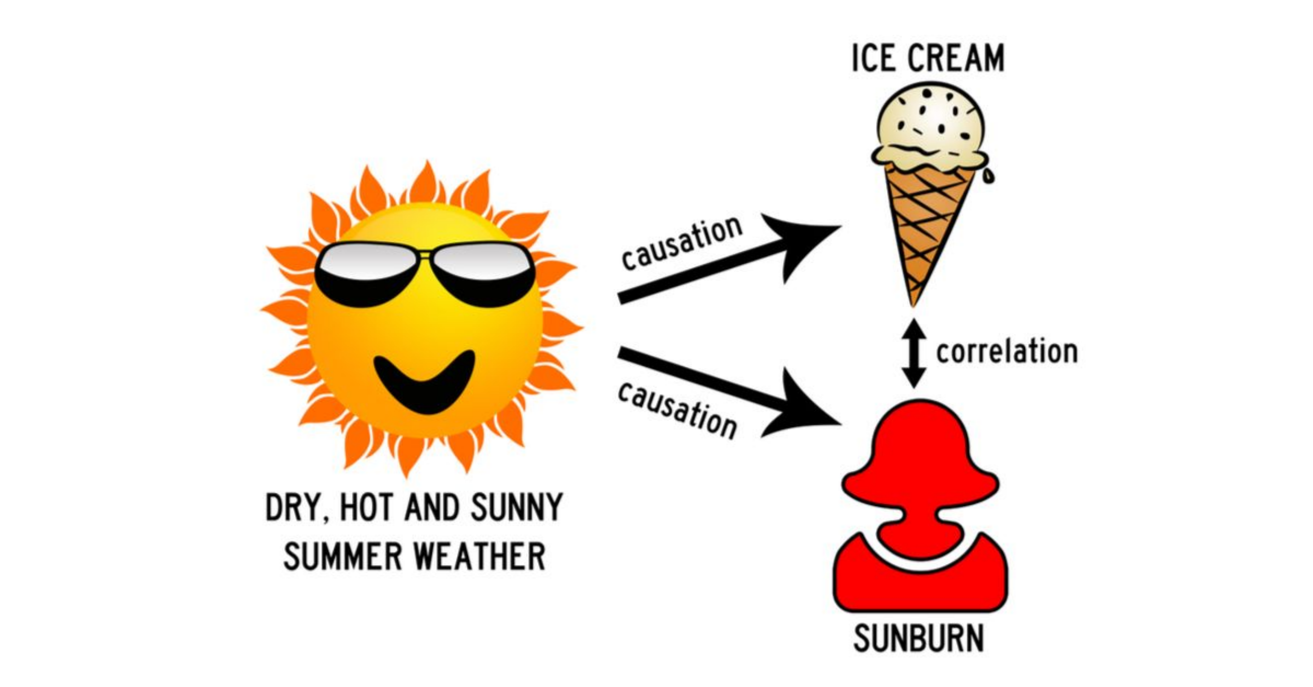
Correlation with target variable

/ You need to be careful with correlation because:

1. This is Pearson Correlation, and is only for numerical variables.
2. Correlation only shows that the feature is important alone. But it could be that somes features are bad on their own, but good together.
3. Correlation is not causation.

Correlation VS Causation

Always look for Causation. Correlation is misleading.



Other methods

/ Besides Correlation, it exits other methods of Feature Importance:

- Feature importance of some models (XGBoost, LightGBM).
- Permutation Feature importance.

Advanced Methods of Feature Selection

Reference:

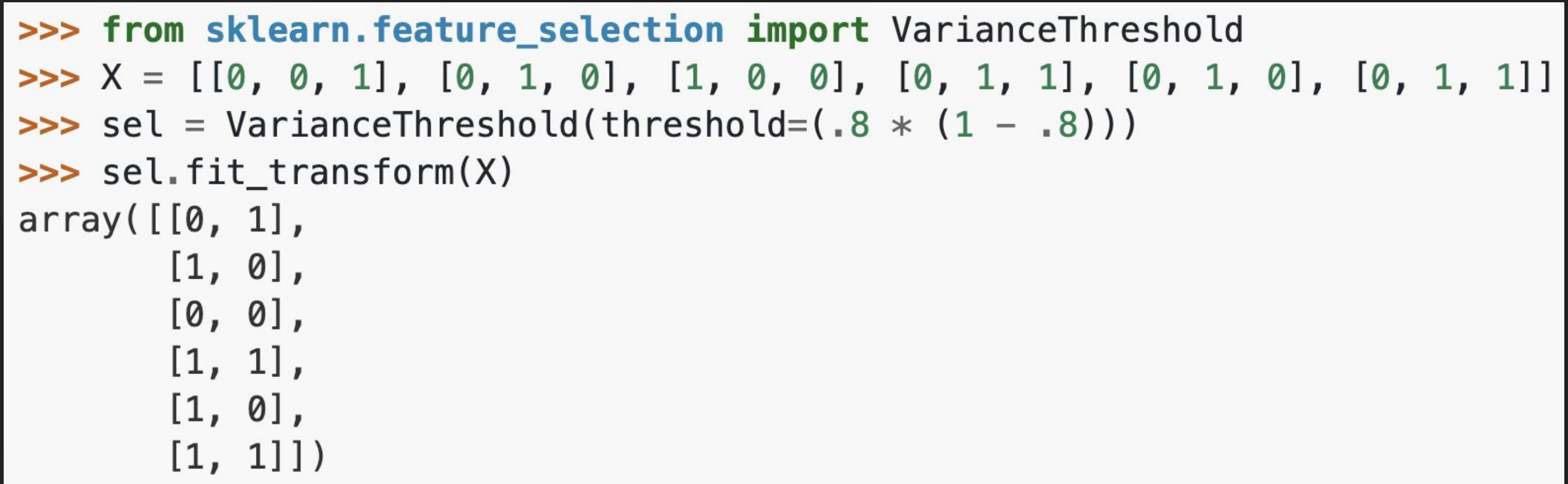
- Sklearn Feat. Sel. Documentation
- machinelearningmastery.com
- Boruta-py

- Variance Threshold
- Univariate feature selection
  - Mutual information
- LASSO
- Wrapper: Su usa un classificador
  - MultiObjectiveEvolutionarySearch: Mejor para muchas generaciones. 10000 Evals
  - PSO: Particule Search optimization
  - RFE: Recursive Feature Elimination
  - SelectKBest
- Filters:
  - InfoGAIN: Cantidad de informacion
  - Correlation Feature Selection

Variance Threshold

/ It removes features with low variance (below some threshold).

zero-variance features == features that have the same value in all rows



```
>>> from sklearn.feature_selection import VarianceThreshold
>>> X = [[0, 0, 1], [0, 1, 0], [1, 0, 0], [0, 1, 1], [0, 1, 0], [0, 1, 1]]
>>> sel = VarianceThreshold(threshold=(.8 * (1 - .8)))
>>> sel.fit_transform(X)
array([[0, 1],
       [1, 0],
       [0, 0],
       [1, 1],
       [1, 0],
       [1, 1]])
```

Types of correlation according variables

/ There are many types of correlation:

- Numerical + Numerical
    - Pearson’s correlation coefficient (linear). f_regression()
    - Spearman’s rank coefficient (nonlinear).
- Numerical + Categorical
    - ANOVA correlation coefficient (linear). f_classif()
    - Kendall’s rank coefficient (nonlinear).
- Categorical + Categorical
    - Chi-Squared test (contingency tables). chi2()
    - Mutual Information. mutual_info_classif() and mutual_info_regression()

/ Q&A

What are your doubts?

Strive

SCHOOL