EMMA ding

emmading.com

info@datainterviewpro.com

Feature Selection

Lesson Structure

📌 Interview Questions

- Why use feature selection?
- How do you select features in general?
- How to do feature selection if you have 10000 features?
- How to calculate feature importance?

▼ Feature Selection

Select a **subset** of the original features for model training.

Is usually used as a pre-processing step before doing the actual learning.

🐥 There is no best feature selection method.

▼ Advantages

- Avoid the curse of dimensionality
- Improves predictive performance and interpretability of models
    - Shorten training times → improve computational efficiency
    - reduce generalization error of the model by removing irrelevant features or noise
    - Improves the predictive power of the model if a model suffers from overfitting

▼ Domain knowledge is important!

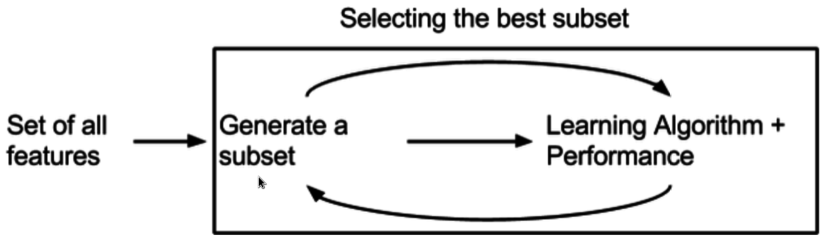
- Understand the business problem: know which features matter and which ones don't
- Consult with domain experts

- Exploratory data analysis (EDA)

Feature Selection Methods

▼ Intrinsic methods

- Embedded methods or implicit methods
- Have feature selection naturally **embedded** with the training process



▼ Tree-based models

- Search for the best feature to split node so that the outcomes are more homogeneous with each new partition.
- If a feature is not used in any split, it's independent of the target variable

▼ Regularization models

- L1-regularization penalizes many of estimated coefficients to zero → only keep features with non-zero coefficients
- Models use regularization, e.g. linear regression, logistic regression, SVMs.

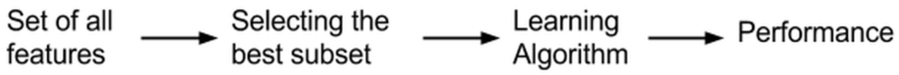
▼ Pros and Cons

- ✔️ Fast because feature selection is embedded within model fitting process
- ✔️ No external feature selection tool is needed.
- ✔️ Provides a direct connection between feature selection and the object function (e.g. maximize information gain in decision trees, maximize likelihood function in logistic regression) which makes it easier to make informed choice.
- ❌ Model-dependent and the choice of models is limited.

▼ Filter methods

- Select features that correlate well with target variable.
- Evaluation is independent of the algorithm.

- The search is performed only once.

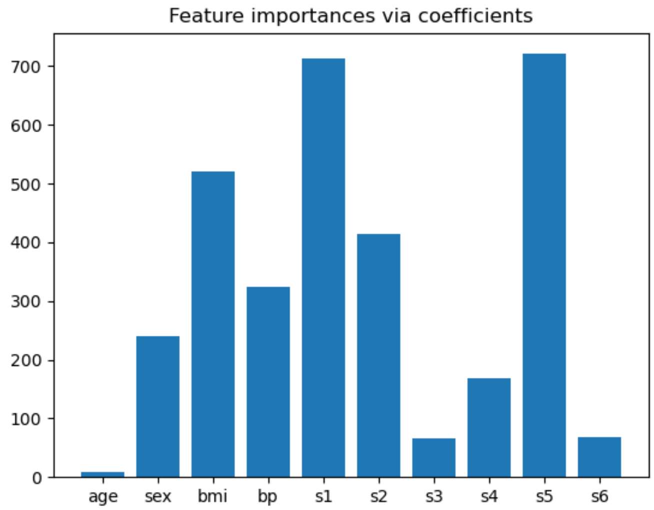


Univariate statistical analysis

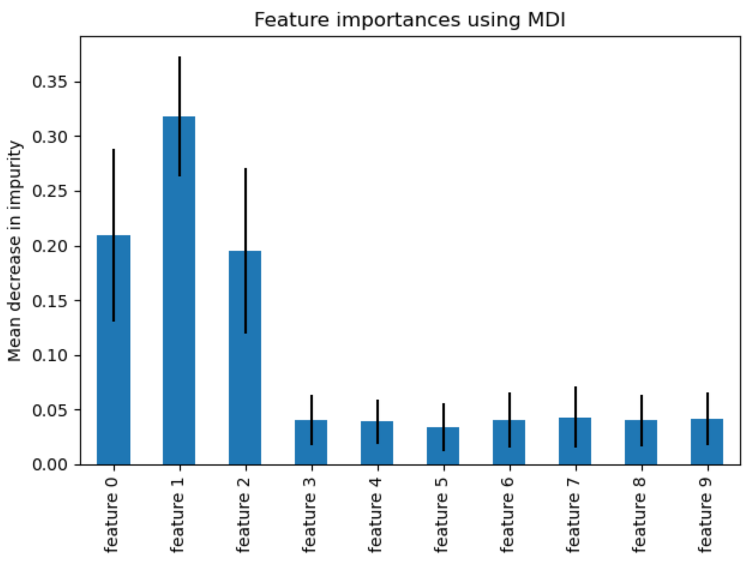
- Analyze how each feature correlates with the target variable and select the ones with higher correlations.

Feature Importance-based

- Use feature importance scores to select features to keep (highest scores) or delete (lowest scores).
    - Coefficients as feature importance, e.g. linear regression, logistic regression.



    - Impurity-based feature importances, e.g. tree-based models.



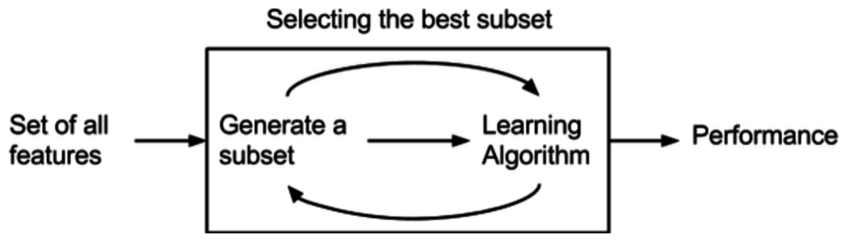
Impurity-based feature importance

▼ Pros and Cons

- ✔️ Simple and fast.
- ✔️ Can be effective at capturing the large trends in the data.
- ❌ Tend to select redundant features.
- ❌ Ignore relationships among features.

▼ Wrapper methods

- Iterative process that repeatedly add subsets of feature to the model and then use the resulting model performance to guide the selection of the next subset.



▼ Sequential feature selection (SFS)

- A family of **greedy search algorithms** that are used to automatically select a subset of features that are most relevant to the problem.

https://scikit-learn.org/stable/modules/feature_selection.html#sequential-feature-selection

- ▼ **Forward** SFS
  - Iteratively finds the best new feature to ***add*** to the set of selected features.
  - Start with ***zero*** feature and find the one feature that maximizes a cross-validated score when a model is trained on this single feature.
  - Once that first feature is selected, we repeat the procedure by adding a new feature to the set of selected features.
  - The procedure stops when the desired number of selected features is reached.
- ▼ **Backward** SFS
  - Start with ***all*** the features and sequentially ***remove*** features from the set until the desired number of features is reached.

▼ **Pros and Cons**

- ✔️ Search for a wider variety of feature subsets than other methods.
  - Consider features that are already selected when choosing a new feature.
- ❌ Have the most potential to overfit the features to the training data.
- ❌ Significant computation time when the number of features is large.