Feature Engineering

Week 11 Day 02

DS 3000 – Foundations of Data Science

1

Reminders

HW 8

Released today

Tuesday, November 26

FP4

Study the description

2

Outline

Exhaustive Parameter Tuning

Feature Engineering

Feature Selection

Feature Extraction from Text

3

Feature Engineering

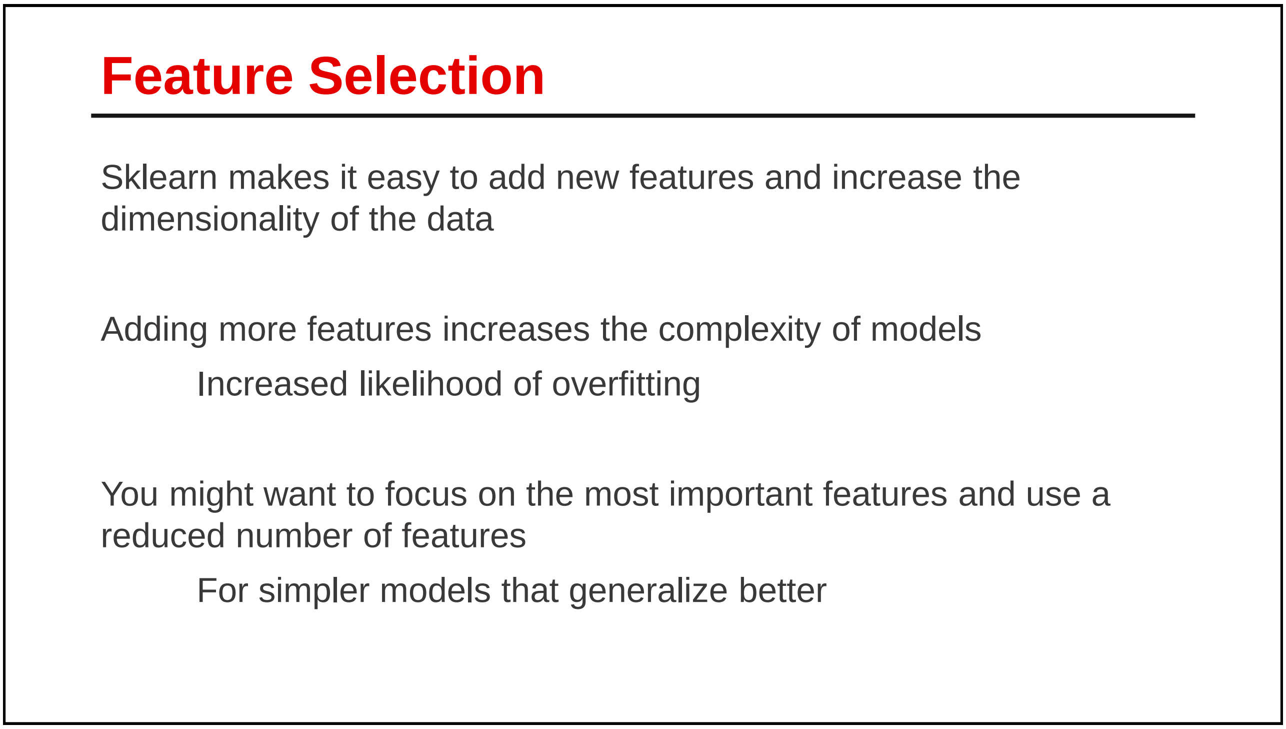
The process of representing raw data in a meaningful way for ML tasks

Need to quantify the properties of the data

These are the variables based on which you will make predictions

Known as **features**, predictors, or attributes (sometimes IVs too)

4



Feature Selection

Sklearn makes it easy to add new features and increase the dimensionality of the data

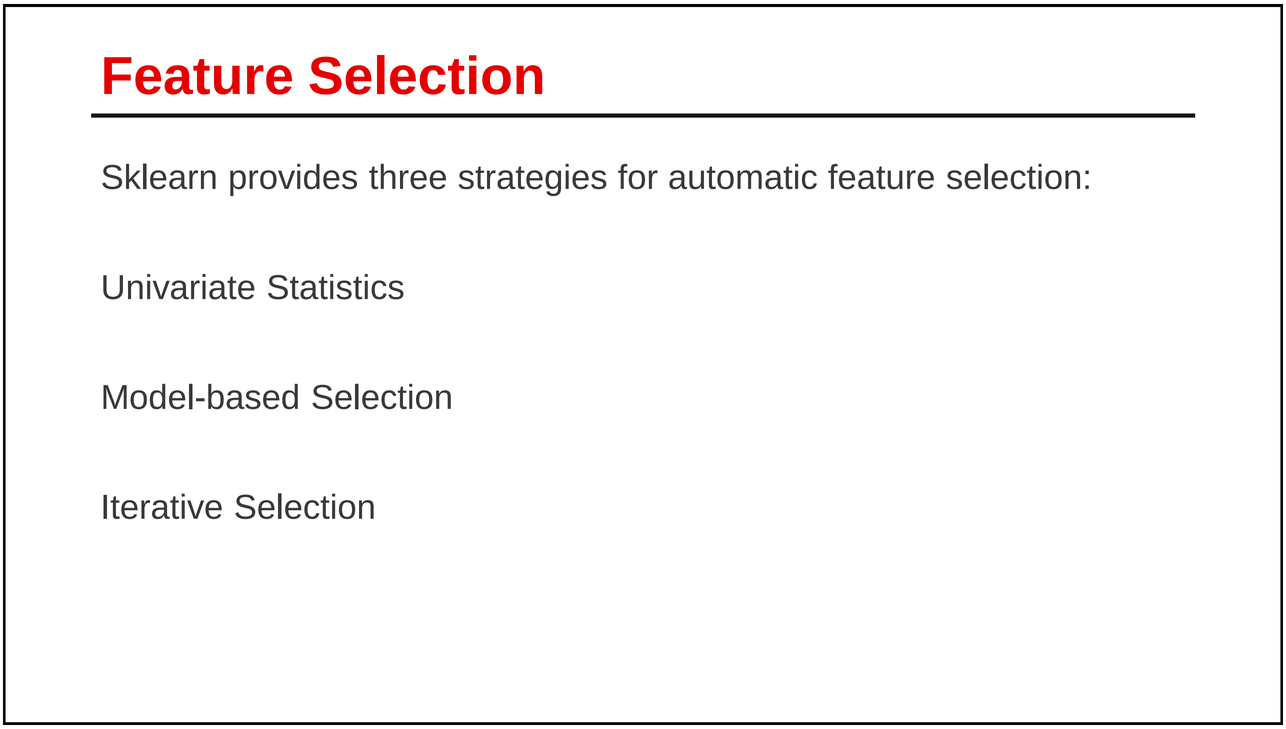
Adding more features increases the complexity of models

Increased likelihood of overfitting

You might want to focus on the most important features and use a reduced number of features

For simpler models that generalize better

5



Feature Selection

Sklearn provides three strategies for automatic feature selection:

Univariate Statistics

Model-based Selection

Iterative Selection

6

7

Feature Extraction from Text

Bag-of-Words Representation

Discards most of the structure of the input text

e.g., paragraphs, sentences, etc.

Counts how often each word appears in each text in the corpus

Bag-of-words representation involves three steps:

Tokenization, Vocabulary Building, and Encoding

8

Bag-of-Words Processing

Tokenization

Split each document or string into words (token)

Vocabulary building

Collect a vocabulary of all words that appear in any of the documents or strings and number them (typically in alphabetical order)

Encoding

For each document or string, count the occurrence of each word in this document or string