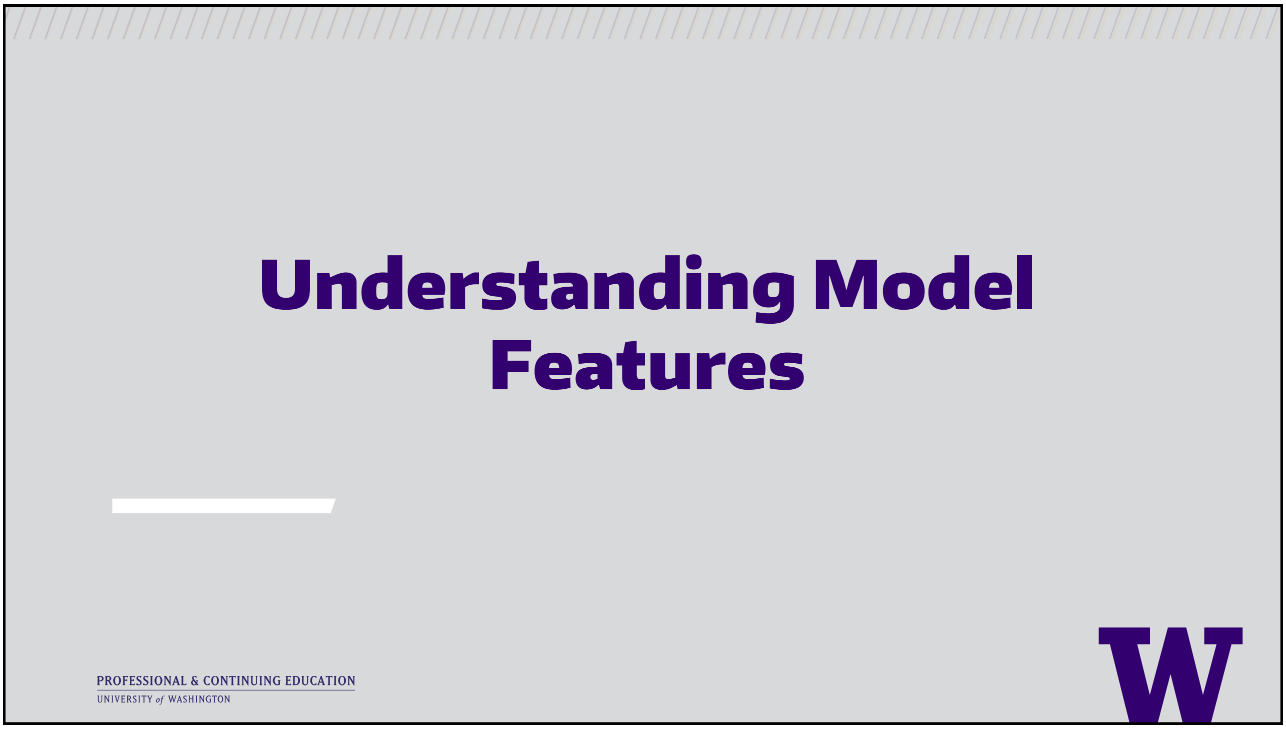
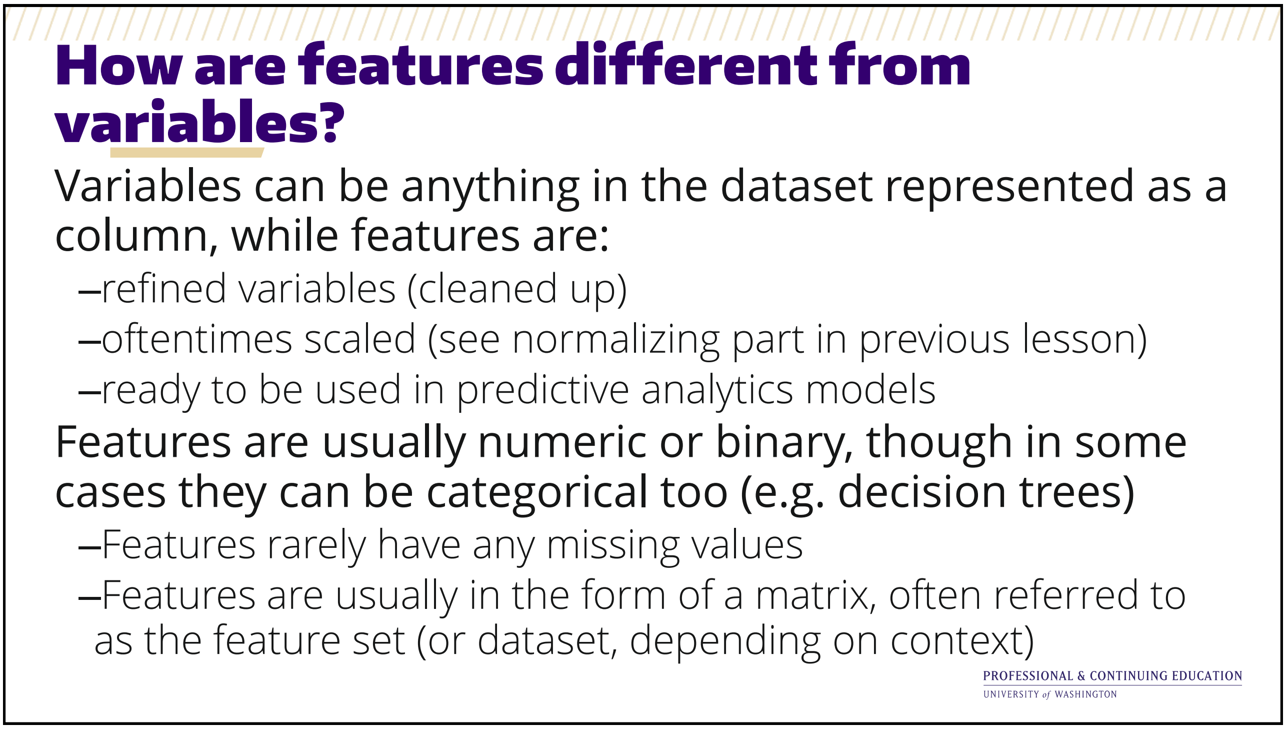


Understanding Model Features

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



How are features different from variables?

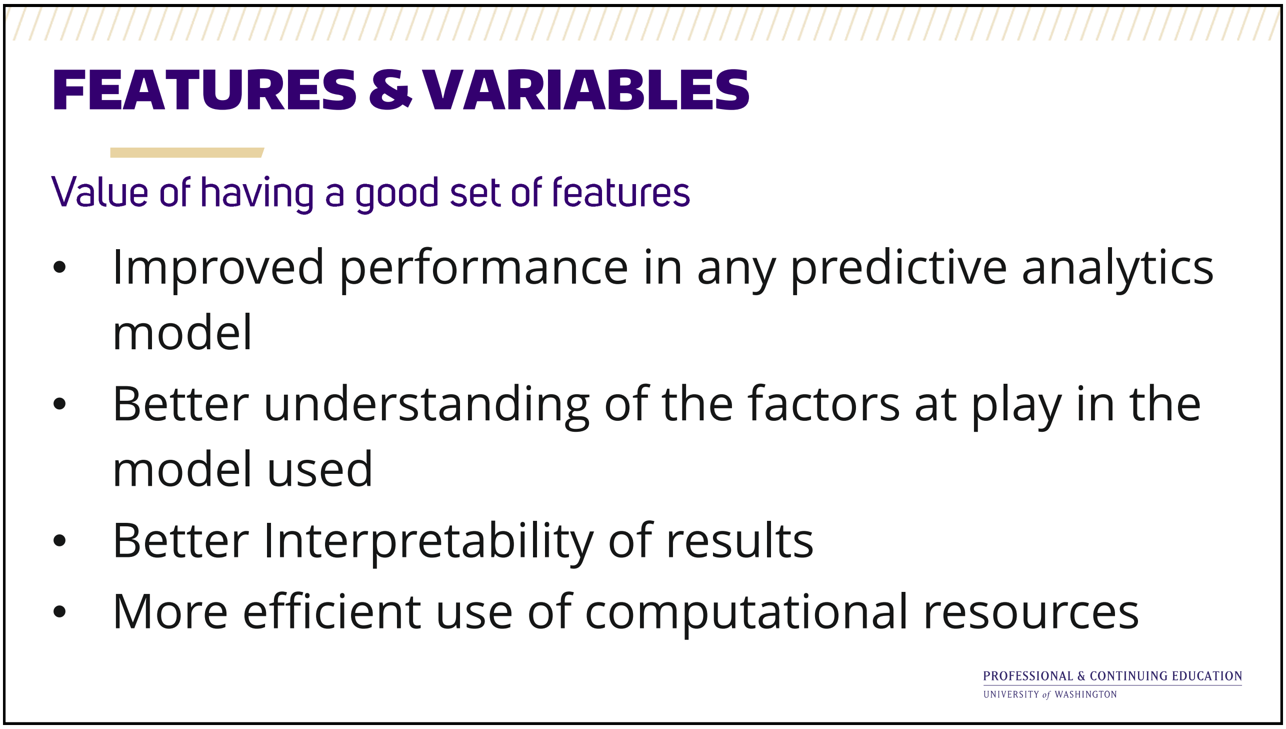
Variables can be anything in the dataset represented as a column, while features are:

- –refined variables (cleaned up)
- –oftentimes scaled (see normalizing part in previous lesson)
- –ready to be used in predictive analytics models

Features are usually numeric or binary, though in some cases they can be categorical too (e.g. decision trees)

- –Features rarely have any missing values
- –Features are usually in the form of a matrix, often referred to as the feature set (or dataset, depending on context)

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

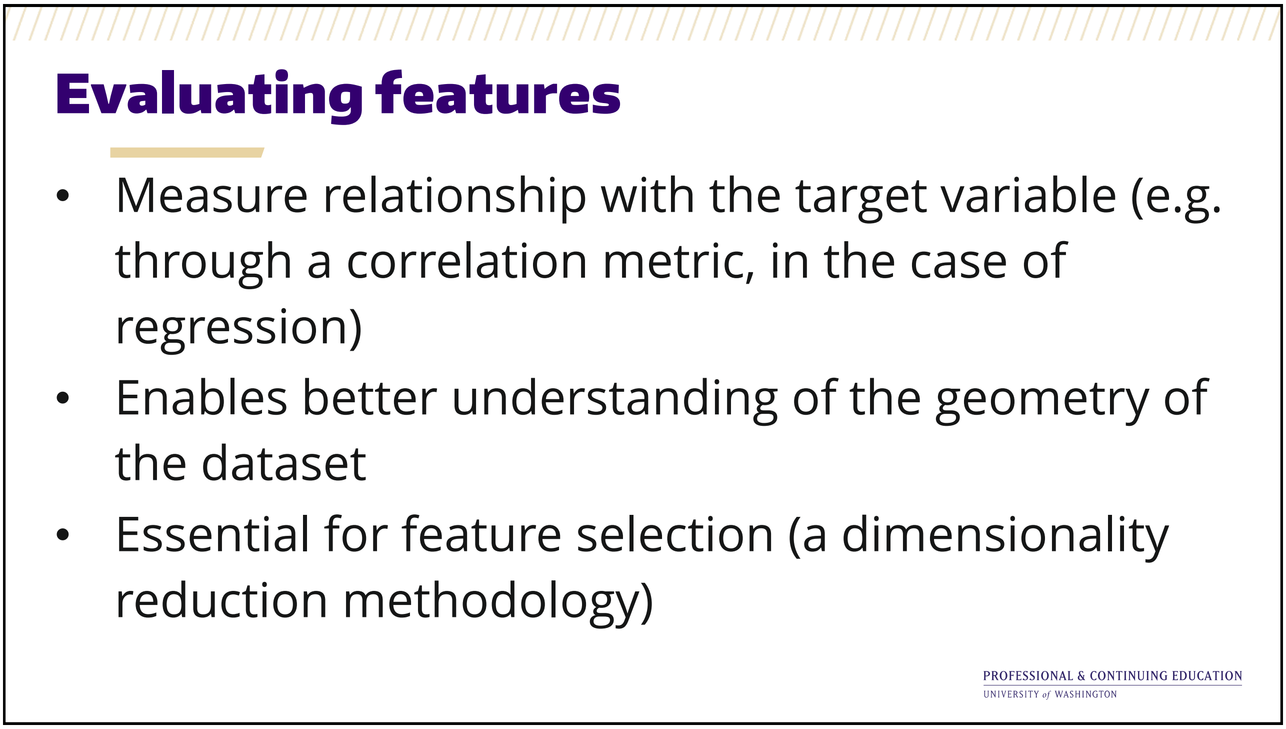


FEATURES & VARIABLES

Value of having a good set of features

- Improved performance in any predictive analytics model
- Better understanding of the factors at play in the model used
- Better Interpretability of results
- More efficient use of computational resources

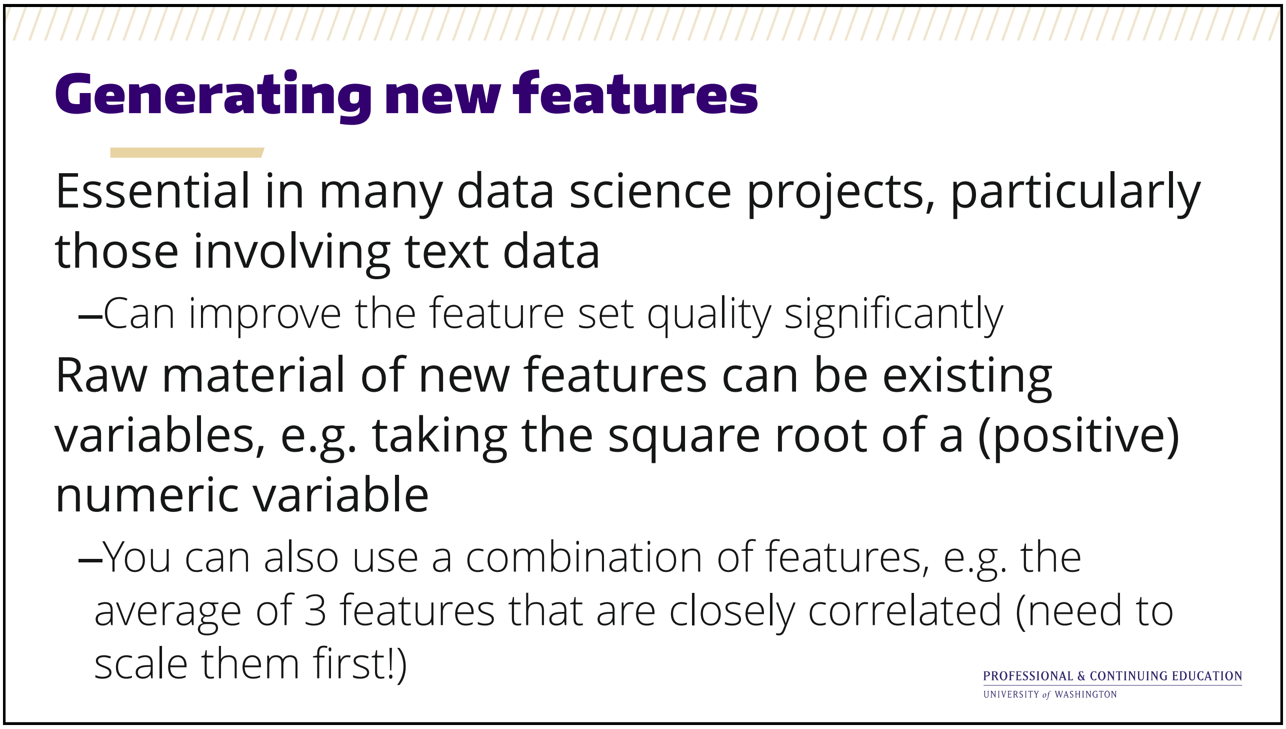
PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



Evaluating features

- Measure relationship with the target variable (e.g. through a correlation metric, in the case of regression)
- Enables better understanding of the geometry of the dataset
- Essential for feature selection (a dimensionality reduction methodology)

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



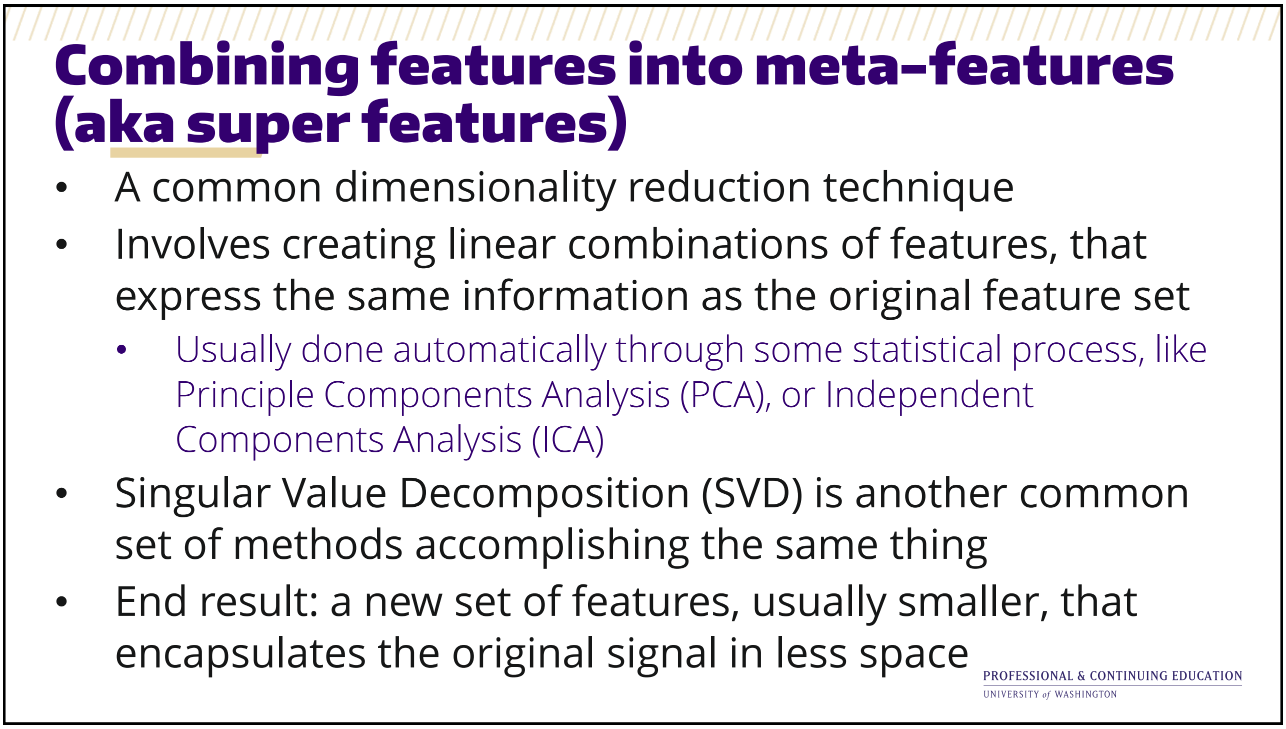
Generating new features

Essential in many data science projects, particularly those involving text data

- –Can improve the feature set quality significantly

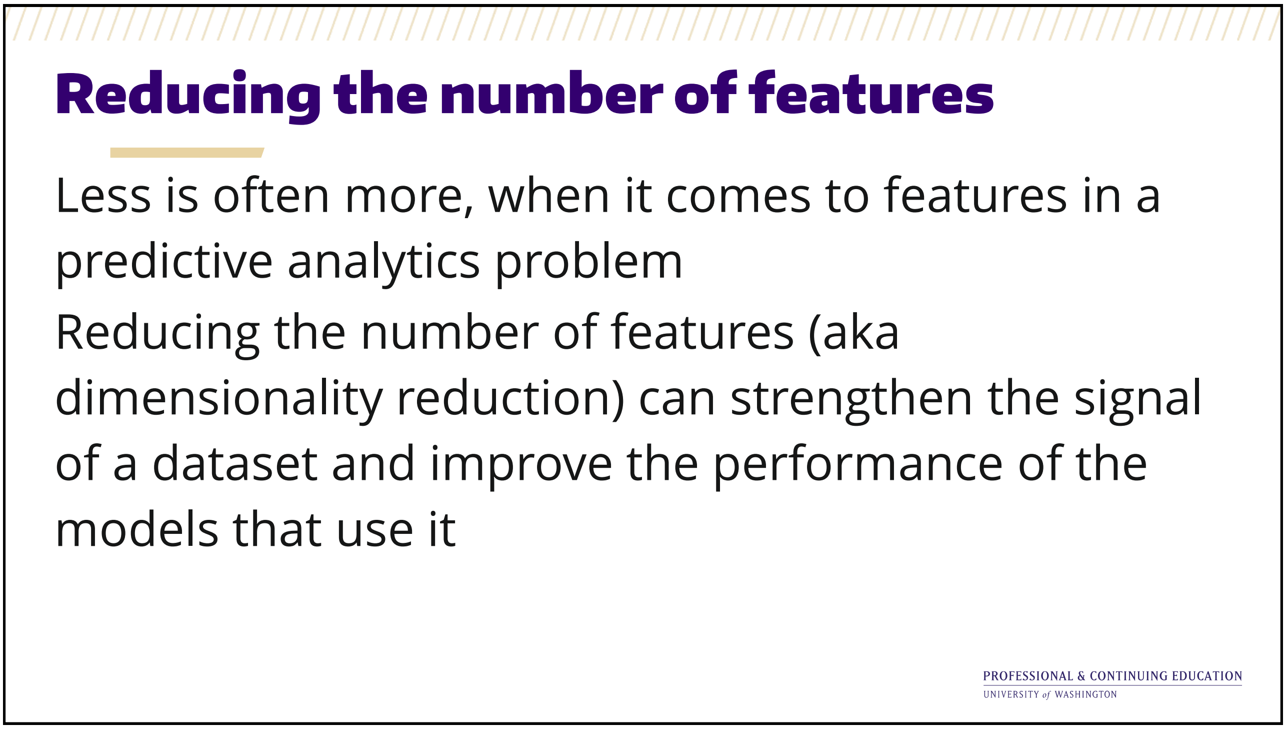
Raw material of new features can be existing variables, e.g. taking the square root of a (positive) numeric variable

- –You can also use a combination of features, e.g. the average of 3 features that are closely correlated (need to scale them first!)



Combining features into meta-features (aka super features)

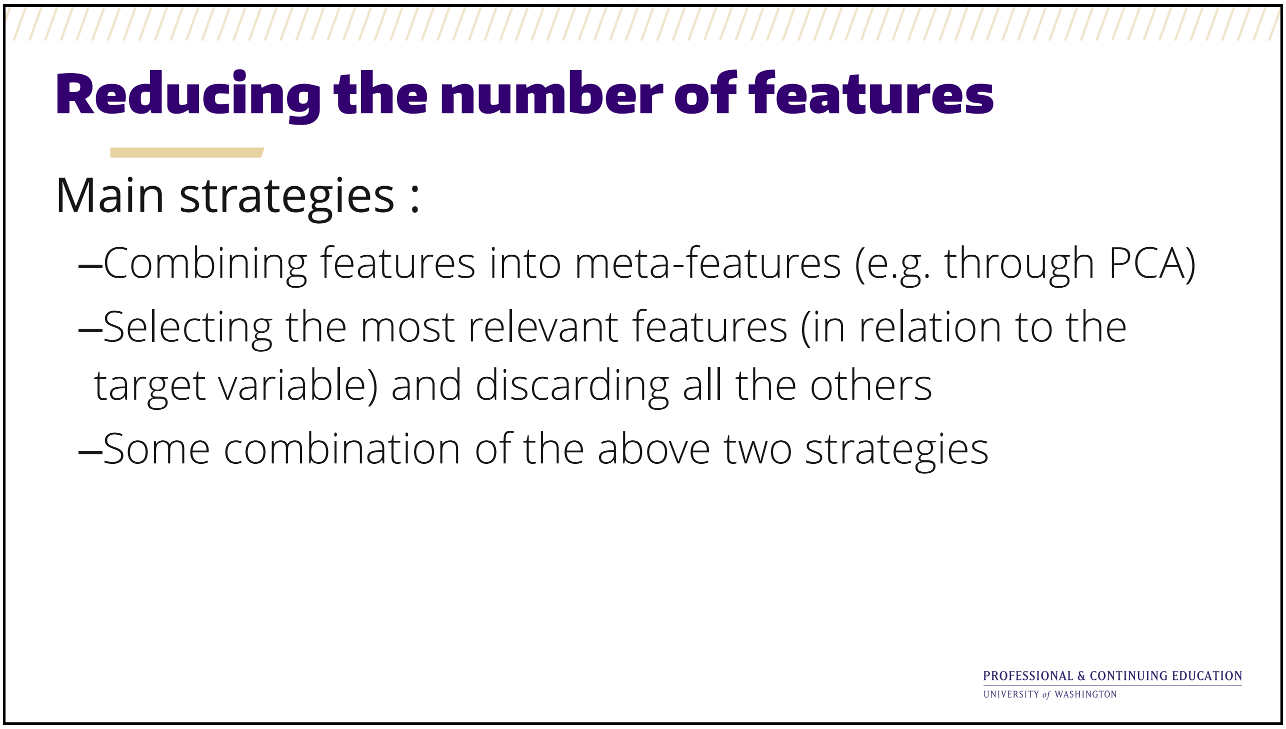
- A common dimensionality reduction technique
- Involves creating linear combinations of features, that express the same information as the original feature set
  - Usually done automatically through some statistical process, like Principle Components Analysis (PCA), or Independent Components Analysis (ICA)
- Singular Value Decomposition (SVD) is another common set of methods accomplishing the same thing
- End result: a new set of features, usually smaller, that encapsulates the original signal in less space



Reducing the number of features

Less is often more, when it comes to features in a predictive analytics problem

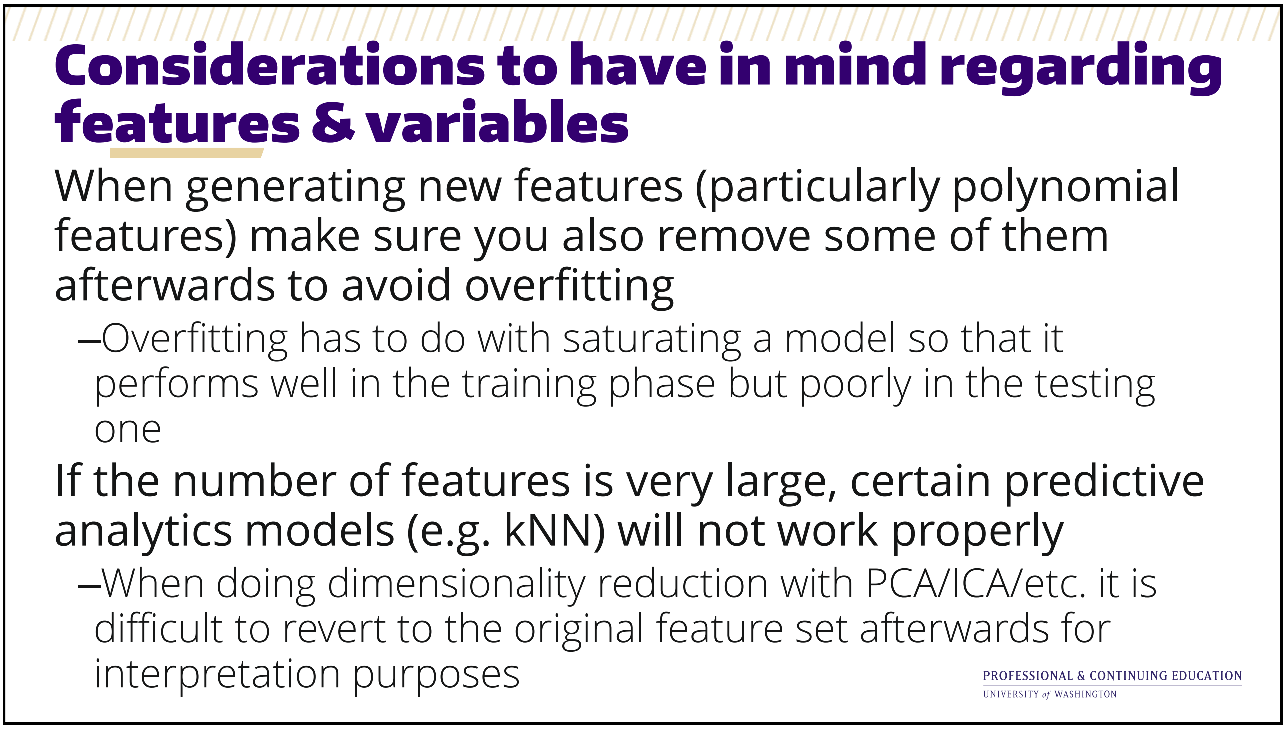
Reducing the number of features (aka dimensionality reduction) can strengthen the signal of a dataset and improve the performance of the models that use it



Reducing the number of features

Main strategies :

- Combining features into meta-features (e.g. through PCA)
- Selecting the most relevant features (in relation to the target variable) and discarding all the others
- Some combination of the above two strategies



Considerations to have in mind regarding features & variables

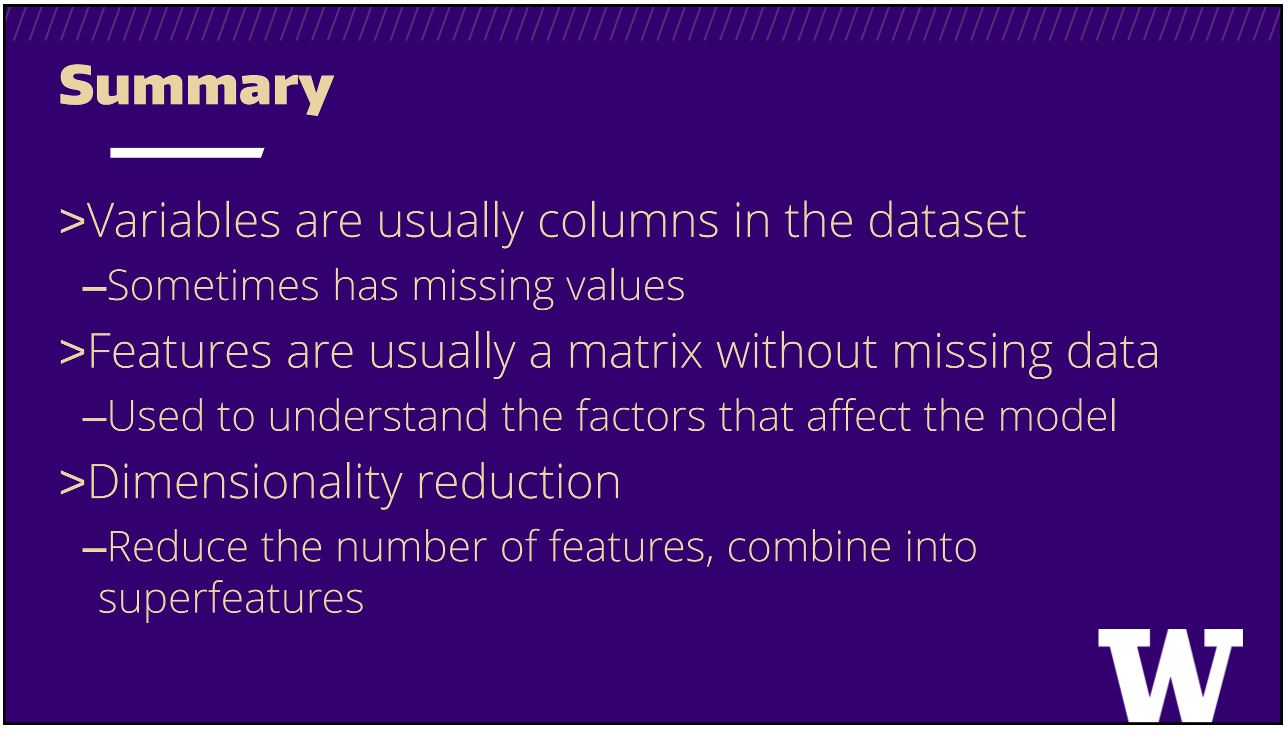
When generating new features (particularly polynomial features) make sure you also remove some of them afterwards to avoid overfitting

- Overfitting has to do with saturating a model so that it performs well in the training phase but poorly in the testing one

If the number of features is very large, certain predictive analytics models (e.g. kNN) will not work properly

- When doing dimensionality reduction with PCA/ICA/etc. it is difficult to revert to the original feature set afterwards for interpretation purposes

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



Summary

>Variables are usually columns in the dataset

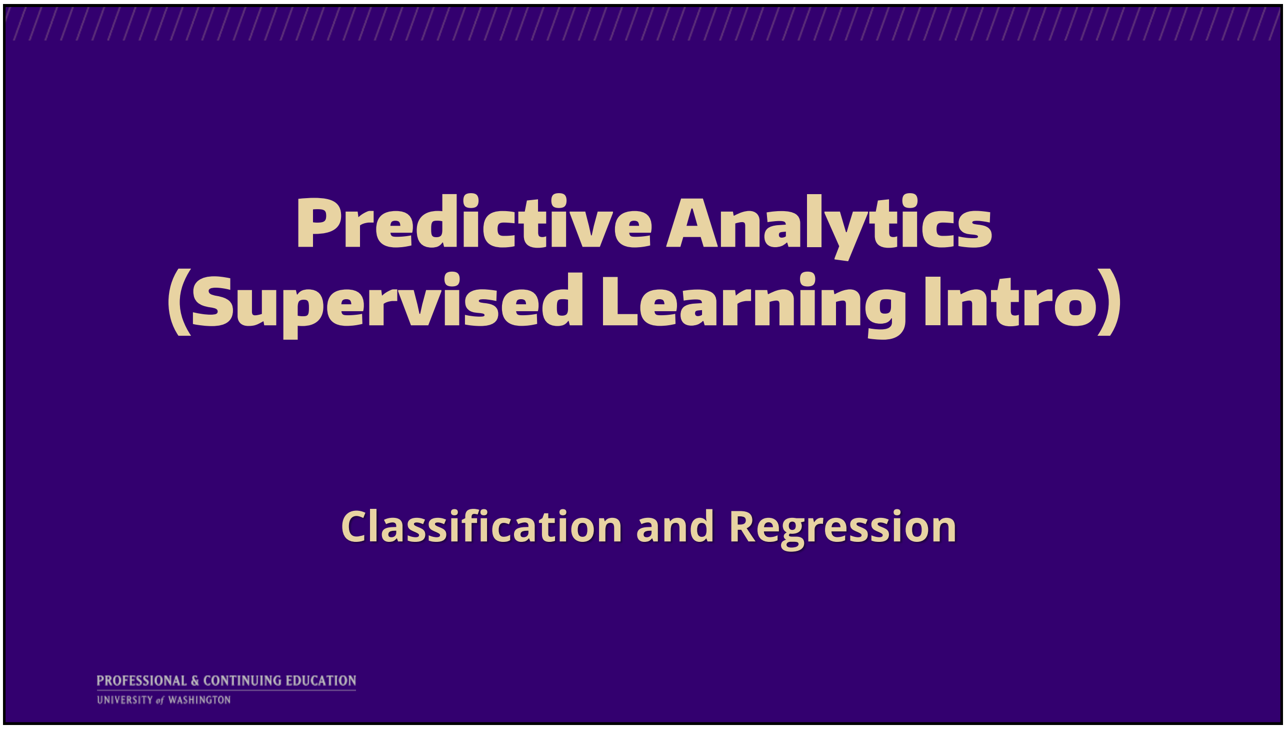
- Sometimes has missing values

>Features are usually a matrix without missing data

- Used to understand the factors that affect the model

>Dimensionality reduction

- Reduce the number of features, combine into superfeatures



Predictive Analytics (Supervised Learning Intro)

Classification and Regression

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON