7.04 PCA

Principal Component Analysis (PCA)

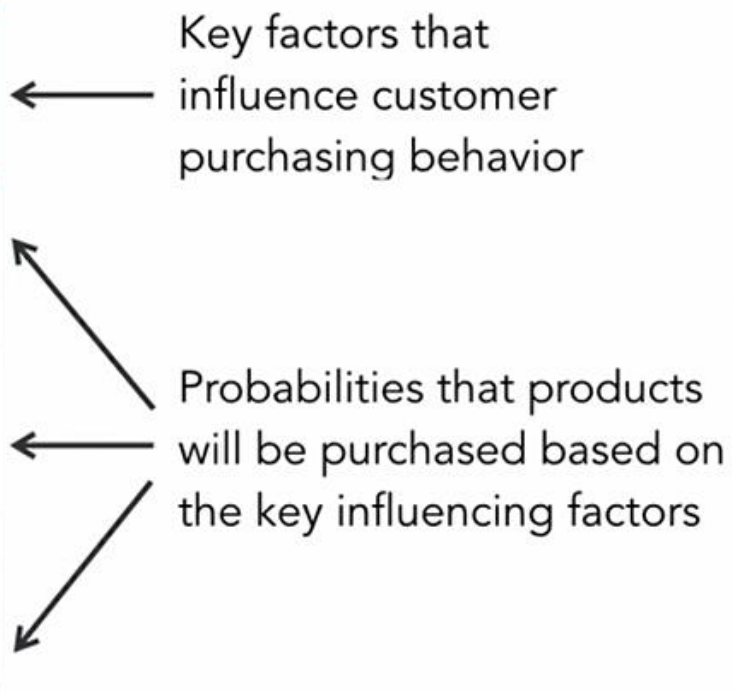
- It is an unsupervised ML algorithm that discovers the relationships between variables and reduces variables down to uncorrelated and synthetic representations called principal components (PCs)

- PCs are uncorrelated features that embody a data set's important info (e.g. variance) with the redundancy, noise, and outliers stripped out

- Use PCs as input variables for ML algorithms to generate predictions from these compressed representations of your data

Principal Component Analysis (PCA)

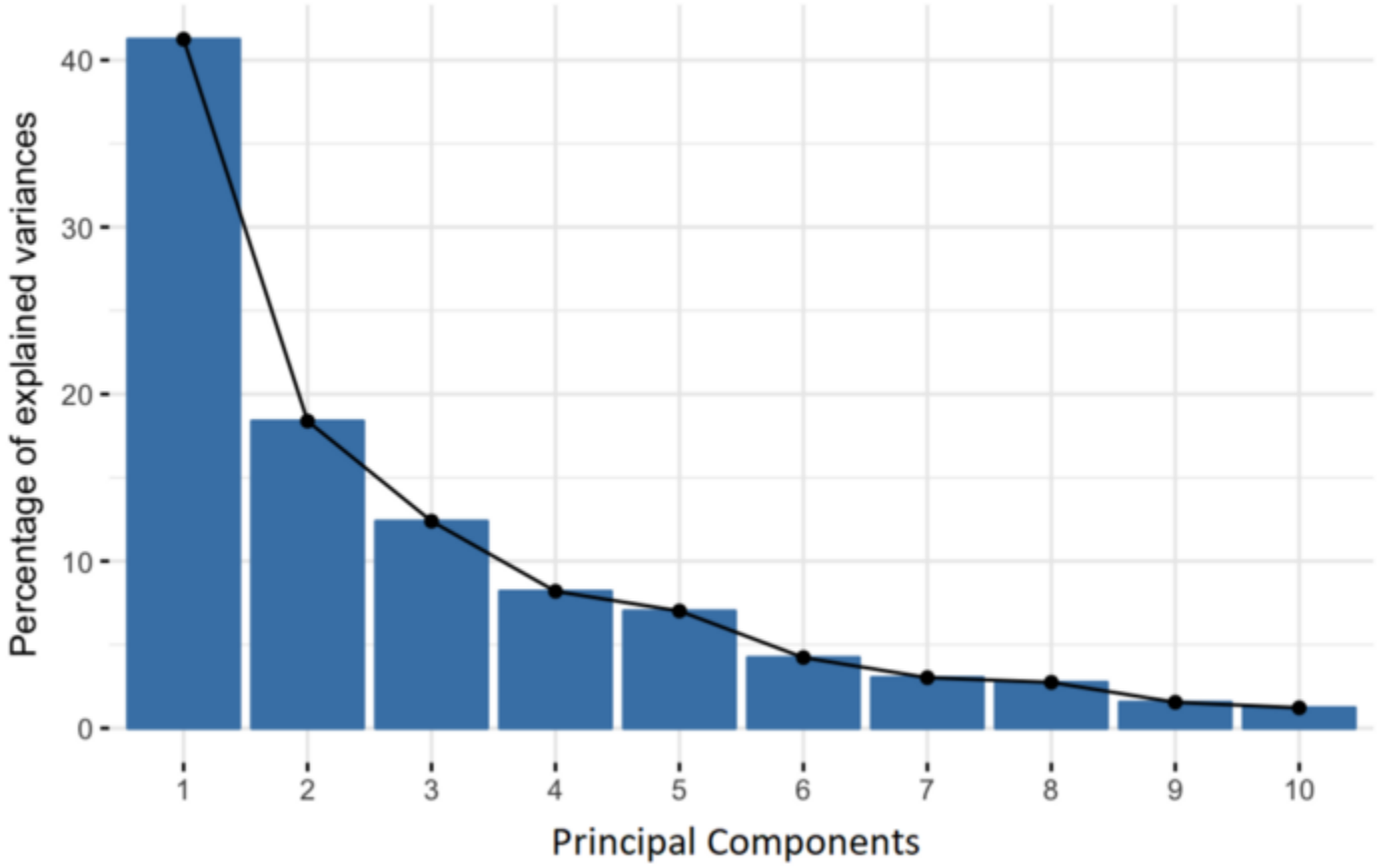
Customer Name	Customer Age	Customer Income	Product Name	Product Price	Sale Status
Ron	55	$45,000	Toothpaste	4.29	1
Tiffany	72	$15,000	Shampoo	5.99	0
Jennifer	47	$65,000	Hair Color	8.99	1



Principal Components

- Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables.
- These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components.
- Example: 10-dimensional data gives you 10 principal components, but PCA tries to put maximum possible info in the first component, then maximal remaining info in the second and so on, until having something like shown in the next slide.

Principal Components



Percentage of Variance (Information) for each by PC

Principal Components

- Organizing information in principal components this way, will allow you to reduce dimensionality without losing much information, and this by discarding the components with low information and considering the remaining components as your new variables.

- An important thing to realize here is that, the principal components are less interpretable and don't have any real meaning since they are constructed as linear combinations of the initial variables.

Principal Components

- Uncorrelated features that embody a data set's important info (e.g. variance) with the redundancy, noise, and outliers stripped out

- Use them as input variables for ML algorithms to generate predictions from these compressed representations of your data

Deciding what components to keep

- The explained variance ratio tells us how much info is compressed into the first few components
- When deciding how many components to keep, look at the percent of cumulative variance
- Ensure at least 70% of the data set’s original info is retained

PCA Use Cases

- Fraud Detection
- Spam Detection
- Speech Recognition
- Image Recognition