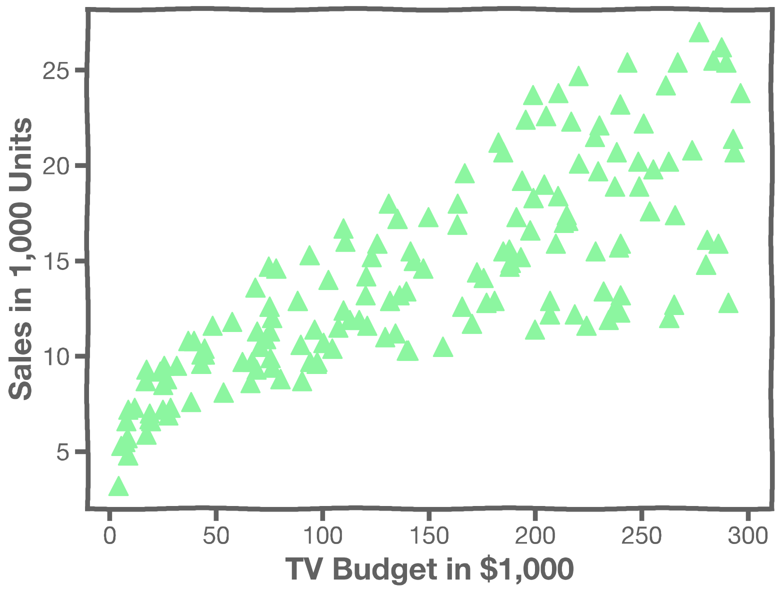
Examples of kNN

Bookmark this page

Example: Predicting Sales

Going back to our example data, what we show here is a plot of the sales in thousands of units versus the TV budget in thousands of dollars. The motivation here is to predict the sales of the response variable, y, given the predictor, x, the TV budget. So we want to build a model to **predict** sales based on TV budget.

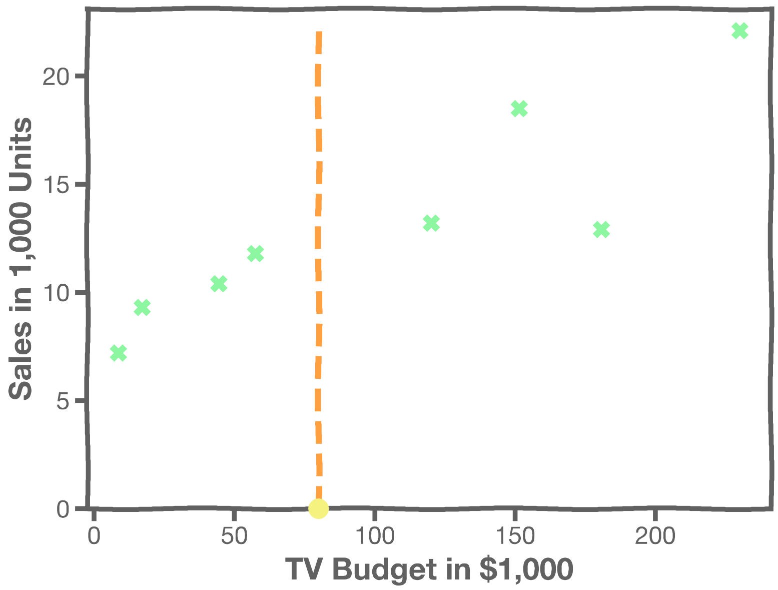
Here is a plot of sales (in thousands of units) for different TV budgets (in thousands of dollars):



Statistical Model

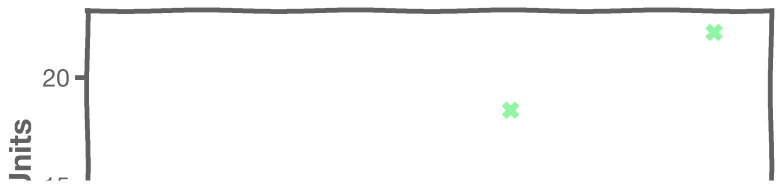
How do we predict, y, for some x? The goal here is to predict the sales given some TV budget; we want to find the value of the sales for a TV budget of about 75,000, or the sales for a TV budget of about 160,000.

To simplify, we will start by looking at eight points on the plot.

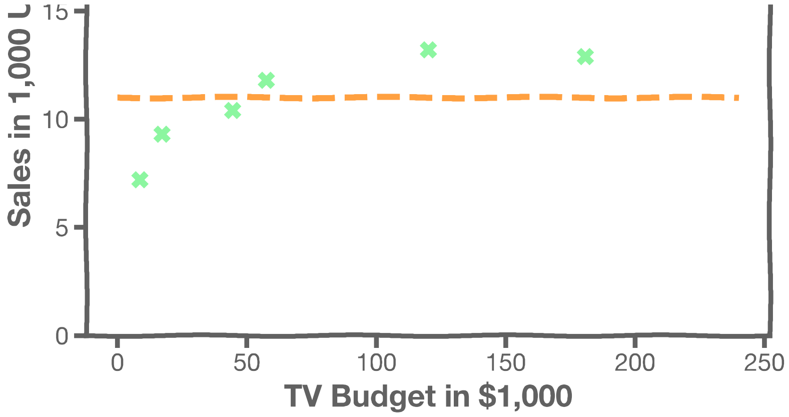


We can first come up with a very simple model, called a naïve model, by taking the mean of all the sales y values for all our observations: $\frac{1}{n}\sum_{i=1}^{n} y_i$

For all TV budgets, the naïve model would predict the average sales and can be used as a baseline later.



Calculator



Simple Prediction Model

We can do better than the naïve model. Let's think of a different type of simple prediction model. We motivate ourselves with an example from everyday life.

If you go to a doctor with some symptoms, for example your tummy is hurting. The doctor will think about the other patients they have seen with similar symptoms and give the same treatment to you. Thus, one type of simple prediction model is where we find the most similar predictor data and predict y.

How do we find $\hat{y}_q$ at some x_q?

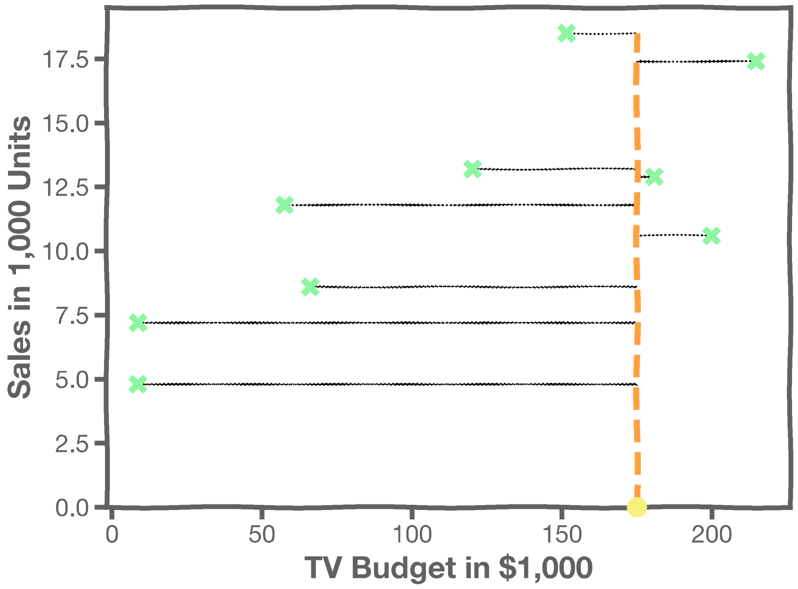
- Find distances to all other points $D(x_q, x_i)$
- Find the nearest neighbor, (x_p, y_p)
- Predict $\hat{y}_q = y_p$. In other words, what we predict for y is the same as the y for the nearest neighbors.

Let's apply this type of model to our data. If we want to know the sales given, say, a $175,000 budget for TV advertising, we look at similar examples - those who are the nearest neighbors to our TV budget. We find the nearest neighbor by finding the smallest distance. In this way we find the most similar example we have in our data. Once we have that, our prediction will be identical to the nearest neighbor's y sales.

We can then do the same for all points x to get a simple prediction model using nearest neighbor.

Extend the Prediction Model

We can extend the model to more than one neighbor to any number "k" of nearest neighbors. So in the example, we can take the 2 nearest neighbors (`k=2`) which are circled in red, and average their y values, and that is our sales prediction.



Generalizing, we take the k nearest neighbors, find the y values and average them. We measure the distance to all other points and find the k nearest neighbors and take the average of the y values of the k nearest neighbors.

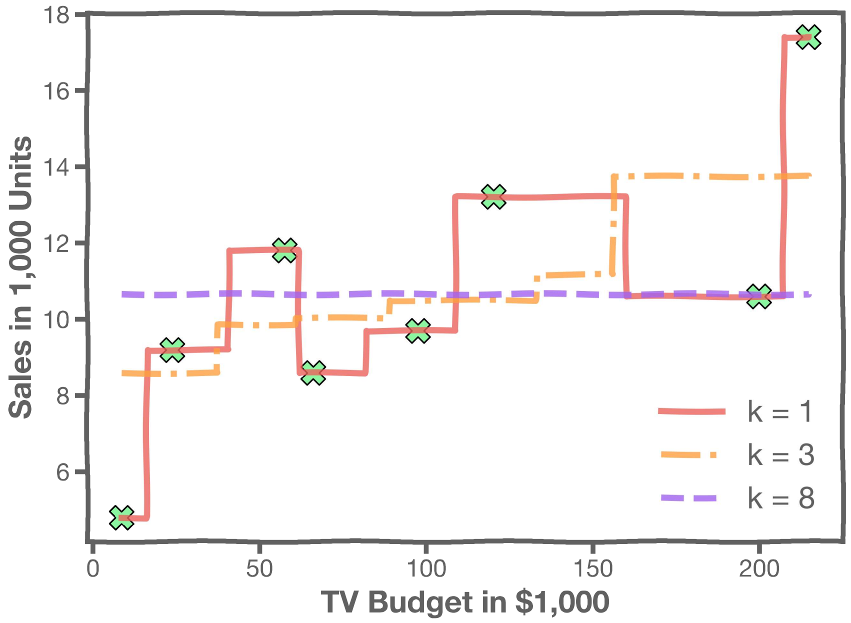
What is $\hat{y}_q$ at some x_q?

Calculator

- Find distances to all other points $D(x_q, x_i)$
- Find the k-nearest neighbor, $(x_{q1}, \dots, y_{qk})$
- Predict $\hat{y_q} = \frac{1}{k}\sum_{i}^{k} y_{qi}$

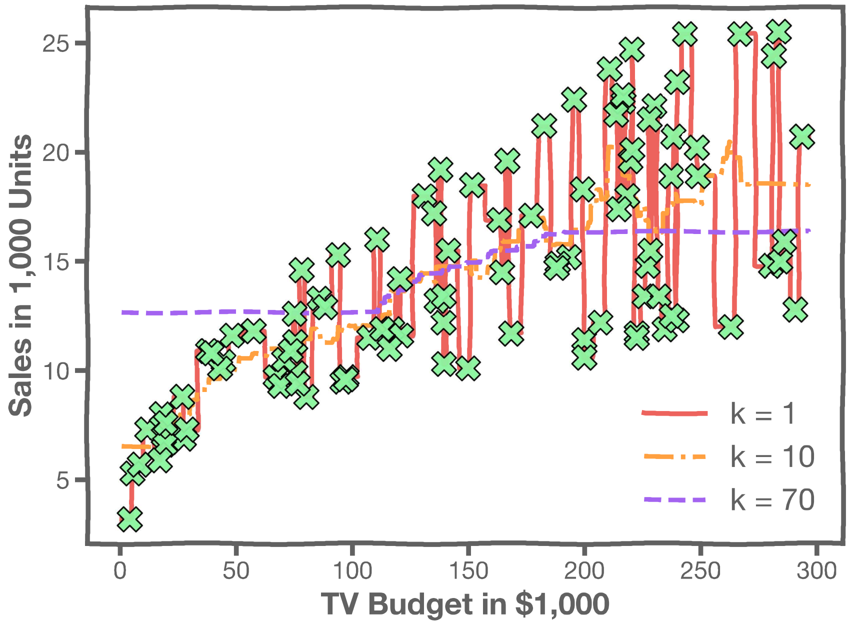
Simple Prediction Models

We can do the modeling for any number of nearest neighbors. k equal to 1, 3 and 8 are shown in the plot below. We see these horizontal lines in the plot that is the region where the center point in the horizontal line is the nearest neighbor to everything else. When k is equal to 1 it goes through every point. When k is equal to 8, since we only have 8 points, are returning back to the naïve model where all the points are averaged.



Simple Prediction Models with All Data

Using more data, we can try different k-models. The plot below shows when k=1, 10 and 70. When k is equal to 1 it goes through every point as expected. When k is equal to 10 the line becomes a little bit more descriptive and for k is equal to 70 it goes closer and closer to the average or naïve model.



📕 k-Nearest Neighbors - kNN

kNN is a non-parametric learning algorithm, meaning that there is no assumptions about the underlying data distribution. The k-Nearest Neighbor Algorithm can be described more formally. Given a dataset $D = (X_1, y_1), \dots, (X_N, y_N)$, for every new X:

1. Find the k-number of observations in D most similar to X:

$$\left(X^{(n_1)}, y^{(n_1)}\right), \ldots, \left(X^{(n_k)}, y^{(n_k)}\right)$$

   These are called the k-nearest neighbors of x

2. Average the output of the k-nearest neighbors of x

$$\hat{y} = \frac{1}{K} \sum_{k=1}^{K} y^{(n_k)}$$

Discussion Board (External resource)

Click OK to have your username and e-mail address sent to a 3rd party application.

< Previous Next >

Calculator