Chapter 3

Models 1.0 Categorical Models

We often use models to predict. The most basic predictive models rely on categories. These models are even simpler than the geometric supertanker model that I described above. A category model creates a subset and then makes some prediction about that subset. Often that prediction relates to how that subset differs from its complement – the subset of everyone else. If the set is people, the subset might be people from Wisconsin. The prediction might be that people from Wisconsin eat more cheese.

Now, you might say that this hardly seems like a model. It seems more like a claim, a claim that may or may not be true. That's true but simple claims like this can be thought of as models in two ways. They can be *theoretical models* if we have some rationale for making the claim. Formally, the statement: *Wisconsin has many dairy farms and farmers promote cheese consumption, therefore people in Wisconsin should eat more cheese* would be called a *hypothesis*. Underlying that hypothesis is a *model* of how the world works but that model is not formal. In this book, I'm describing formal models – models with explicit assumptions.

Here's why I want to call this a model. Suppose that you and I are given an enormous pile of home loans that a bank has made over the past decade. You read each file carefully and use your judgment to make predictions as to whether or not each loan defaulted. I on the other hand use an extremely simple model. I look at the ratio of average income over the previous three years to the amount of the monthly payment. I then divide the files into those that have a high income to payment ratio and those that have a low income to payment ratio. Those in the first pile, I predict won't default. Those in the second pile, I predict will default. This model isn't entirely objective, I'll have to use my judgement for where to make the cutoff. Who will predict more accurately? Evidence from similarly constructed experiments says that I will. Stark categorical models outperform experts in predicting loan defaults.

Variation

I'm going to use categorical models to introduce some basic concepts and measures from statistics. Suppose that I have data on the values of four homes.

Home	Value
A	200K
B	300K
C	500K
D	600K

Notice that these values differ. The first measure that I want to introduce is the *mean.* That's just the average value of the four homes, which is 400K. The second measure, called the *variance*, captures how much the values differ from the mean. Variance equals the average squared difference from each value to the mean. In this example, variance can be computed as follows:

$$\text{Variance} = (200 - 400)^2 + (300 - 400)^2 + (500 - 400)^2 + (600 - 400)^2$$

$$\text{Variance} = 40,000 + 10,000 + 10,000 + 40,000 = 100,000$$

Often categorical distinctions provide powerful insights. Again, these categorical distinctions are not models per se, but the building blocks of models. They're the nouns. In the categorical models, we're not using models to gain deep insights, we're just using them to organize how we think of the world. To measure how well categorical models organize information, statisticians use a measure called R^2, the percentage of the variation explained when the data gets binned into categories. For example, suppose that I place the four houses into two categories. The first contains houses A and B are newer homes and the second contains C and D which are older homes. I can write this more formally as *New* $= \{A, B\}$ and *Old* $= \{C, D\}$. The mean value of the houses in *New* equals 250K and the mean value of houses in *Old* equals 550K. I can now compute the variance within each of the two categories.

$$\text{Variance}(\textit{New}) \ = (200 - 250)^2 + (300 - 350)^2 = 2500 + 2500 = 5000$$

$$\text{Variance}(\textit{Old}) = +(500 - 550)^2 + (600 - 550)^2 = 2500 + 2500 = 5000$$

Summing the variation across the two categories gives 10,000. Recall that original data had a variation of 100,000. This means that we've explained 90% of the variation in the data by binning it into two categories. The statistics R^2 in this case equals 0.9 or 90%.

$$R^2 = \frac{\text{Variance Explained}}{\text{Total Variance}}$$

The higher the R^2, the more variance that's explained and the better the categorization. For example, suppose that houses A and D are brick and houses B and C are wooden. Categorizing the houses into brick and wooden will explain none of the variation, so $R^2 = 0$.