Intro to Machine Learning with Python

Based on Kaggle Learn
V. Alavi

“Machine intelligence is the last invention that humanity will ever need to make.”
~Nick Bostrom

How Models Work

Introduction

We'll start with an overview of how machine learning models work and how they are used.

This may feel basic if you've done statistical modeling or machine learning before. Don't worry, we will progress to building powerful models soon.

Scenario

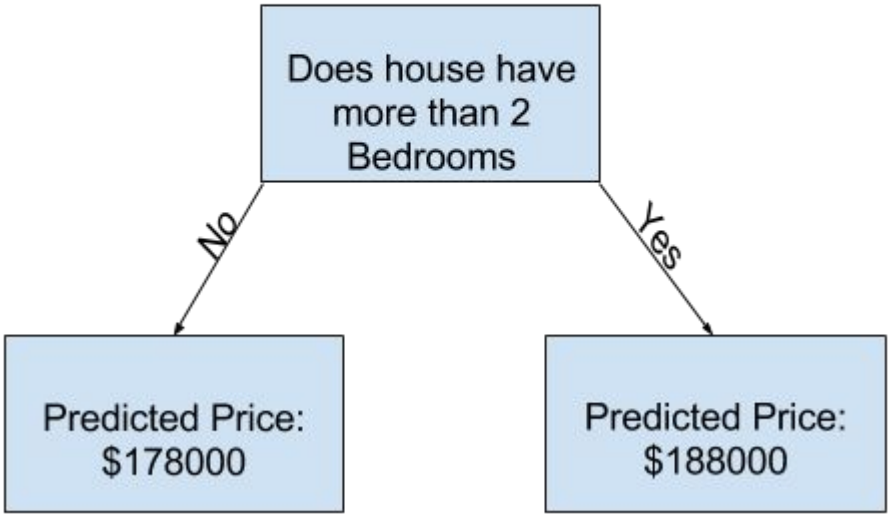
Your cousin has made millions of dollars speculating on real estate. He's offered to become business partners with you because of your interest in data science. He'll supply the money, and you'll supply models that predict how much various houses are worth.

Decision Tree

We'll start with a model called the Decision Tree. There are fancier models that give more accurate predictions. But decision trees are easy to understand, and they are the basic building block for some of the best models in data science.

For simplicity, we'll start with the simplest possible decision tree.

Sample Decision Tree



It divides houses into only two categories.

Sample Decision Tree

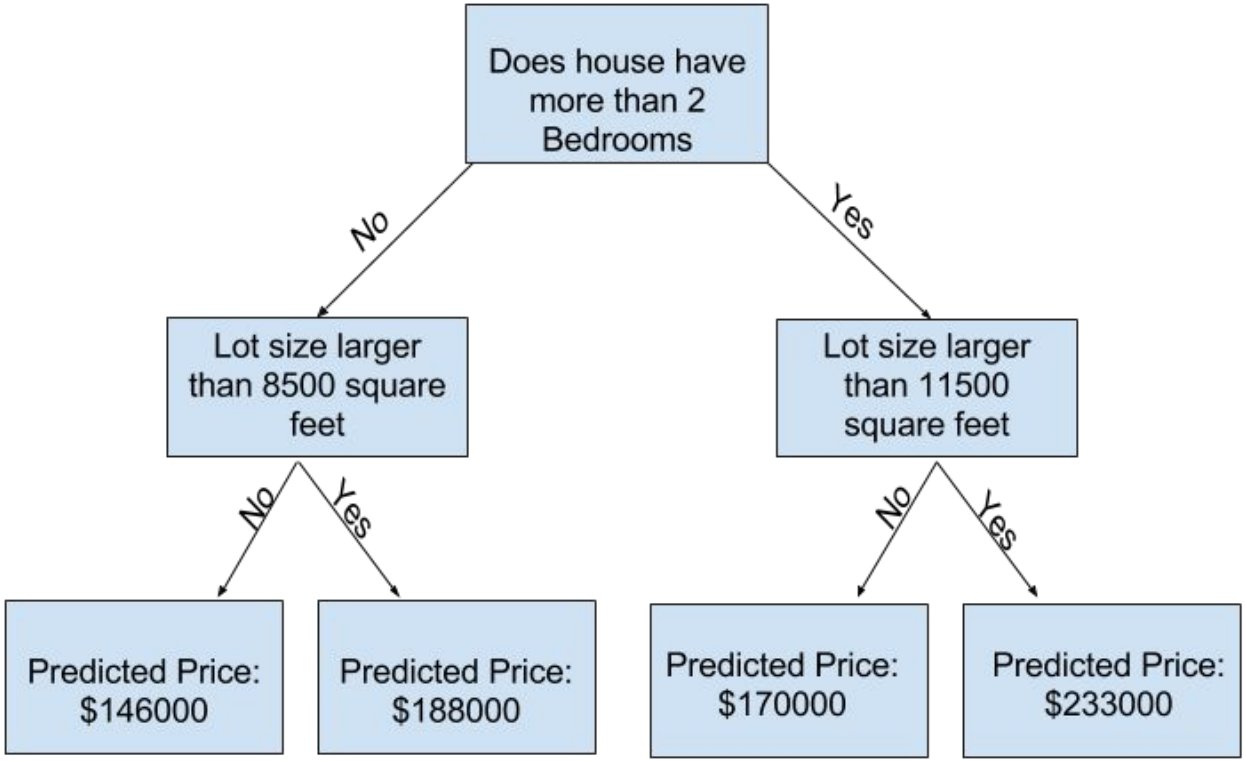
We use data to decide how to break the houses into two groups, and then again to determine the predicted price in each group. This step of capturing patterns from data is called **fitting** or **training** the model. The data used to **fit** the model is called the **training data**.

After the model has been fit, you can apply it to new data to predict prices of additional homes.

Improving the Decision Tree

The biggest shortcoming of this model is that it doesn't capture most factors affecting home price, like number of bathrooms, lot size, location, etc.

You can capture more factors using a tree that has more "splits." These are called "deeper" trees.



A decision tree that also considers the total size of each house's lot

Conclusion

You predict the price of any house by tracing through the decision tree, always picking the path corresponding to that house's characteristics. The predicted price for the house is at the bottom of the tree (**leaf**).

The splits and values at the leaves will be determined by the data, so it's time for you to check out the data you will be working with.

Basic Data Exploration

Using Pandas to Get Familiar With Your Data

The first step in any machine learning project is familiarize yourself with the data. You'll use the Pandas library for this. Pandas is the primary tool data scientists use for exploring and manipulating data.

Pandas

The most important part of the Pandas library is the DataFrame. A DataFrame holds the type of data you might think of as a table. This is similar to a sheet in Excel, or a table in a SQL database.

Pandas has powerful methods for most things you'll want to do with this type of data.

Your First Machine Learning Model

Selecting Data for Modeling

Your dataset had too many variables to wrap your head around, or even to print out nicely. How can you pare down this overwhelming amount of data to something you can understand?