End-to-End Machine Learning Project

Look at the Big Picture

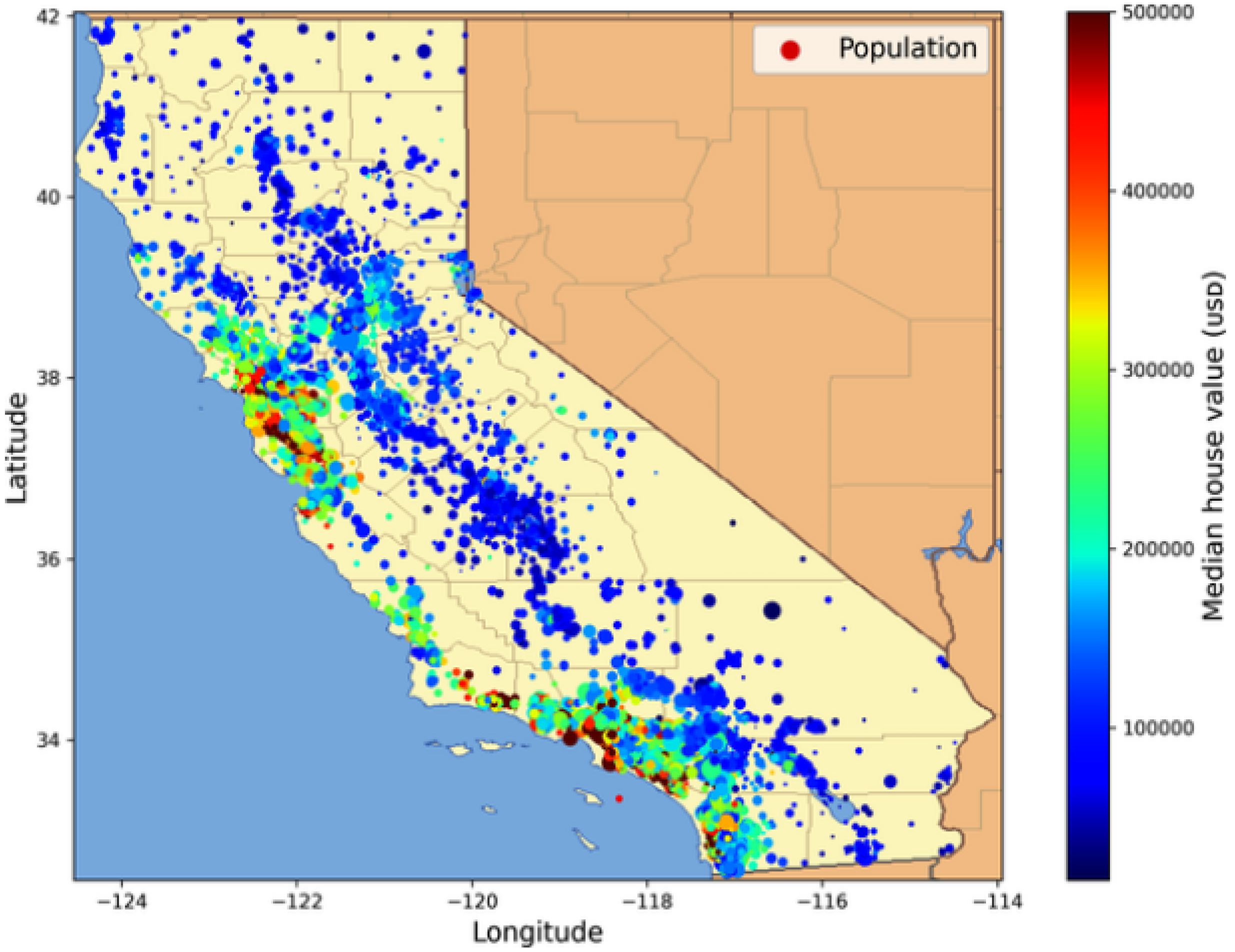


Figure 2-1. California housing prices

-
- The model should learn from this data and be able to predict the median housing price in any district.

Frame the problem

- The first question to ask your boss is what exactly the business objective is.
- How does the company expect to use and benefit from this model?
  - Ex: Your boss answers that your model’s output will be fed to another machine learning system. This downstream system will determine whether it is worth investing in a given area. Getting this right is critical, as it directly affects revenue.
- The next question to ask your boss is what the current solution looks like (if any).
  - Ex: Our boss answers that the district housing prices are currently estimated manually by experts (but only 30% predictions are right).

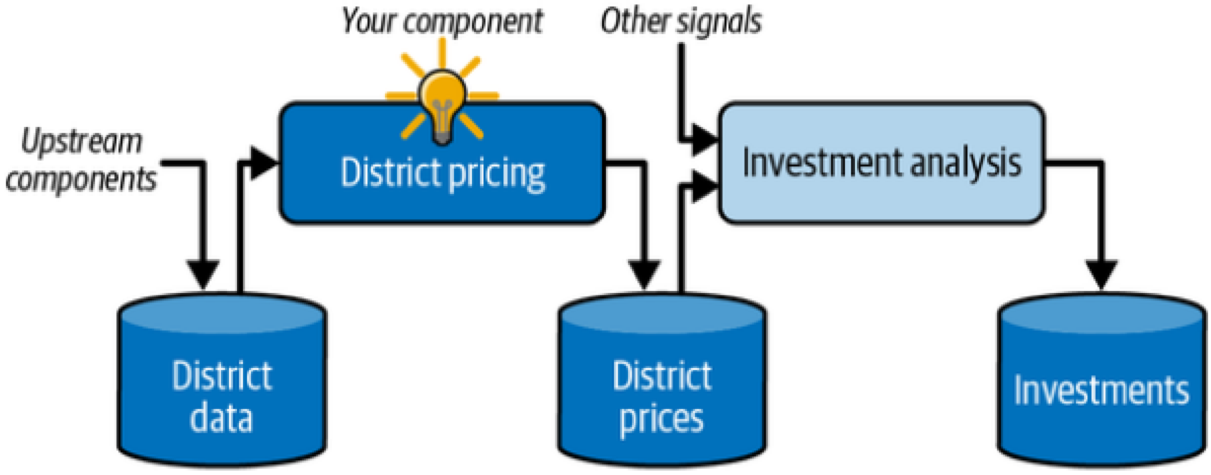


Figure 2-2. A machine learning pipeline for real estate investments

-
  - **Pipeline**: Is a sequence of data processing components. In Machine Learning, they are essential because raw data almost always requires many transformations before it is useful.
- With all this information, you are now ready to start designing your system:
  - First, determine what kind of training supervision the model will need: is it a supervised, unsupervised, semi-supervised, self-supervised, or reinforcement learning task?
  - Is it a classification task, a regression task, or something else?
  - Should you use batch learning or online learning techniques?
  - For this example:
    - Supervised Learning: Because the model is trained using data that already contains the correct answer (labels).
    - Regression: Because the ultimate goal is to predict a specific numerical value.
    - Multiple Regression: Because several features are used to calculate the prediction.
    - Univariate Regression: Because, despite using many inputs, only a single output value is predicted for each case.
    - Batch Learning: Because there is no continuous, real-time data flow; the system can be trained only once with the entire available dataset.

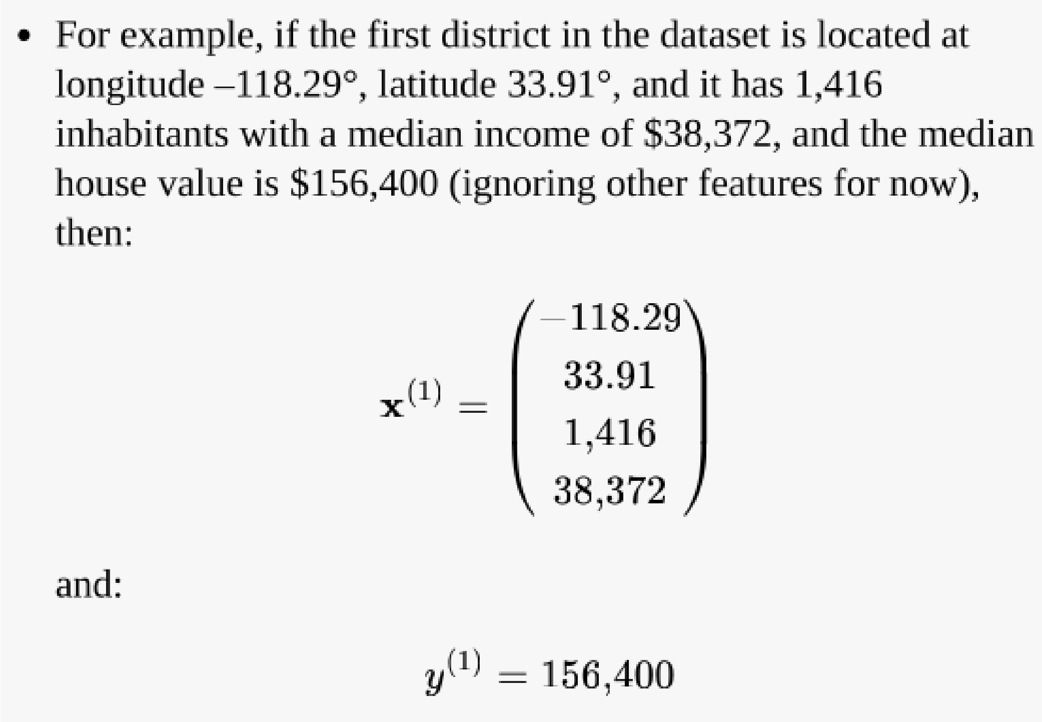
Select a performance Measure

- A typical performance measure for regression problems is the root mean square error (RMSE).

Equation 2-1. Root mean square error (RMSE)

$$\mathrm{RMSE}\left(\mathbf{X}, h\right) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(h\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right)^2}$$

  -
  - Notations:
    - m is the number of instances in the dataset
    - x(i) is a vector of all the feature values (excluding the label) of the ith instance in the dataset, and y(i) is its labe.

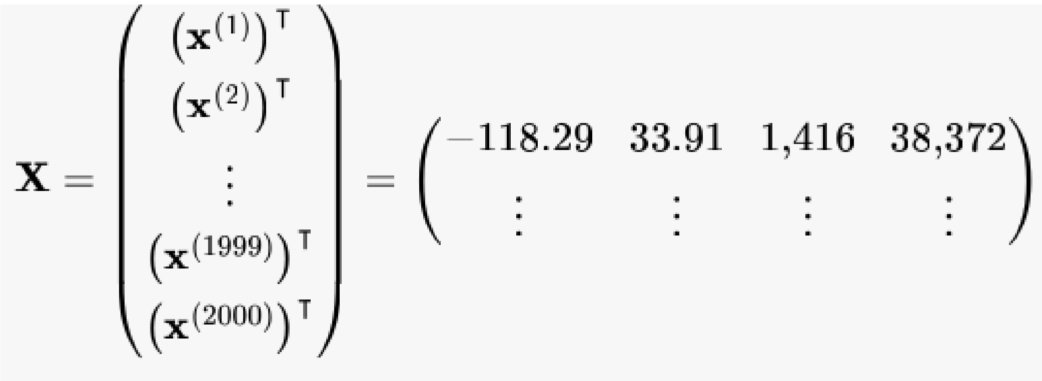
    - 

      - For example, if the first district in the dataset is located at longitude –118.29°, latitude 33.91°, and it has 1,416 inhabitants with a median income of $38,372, and the median house value is $156,400 (ignoring other features for now), then:

      $$\mathbf{x}^{(1)} = \begin{pmatrix} -118.29 \\ 33.91 \\ 1{,}416 \\ 38{,}372 \end{pmatrix}$$

      and:

      $$y^{(1)} = 156{,}400$$

    - X is a matrix containing all the feature values (excluding labels) of all instances in the dataset.
    - 

      $$\mathbf{X} = \begin{pmatrix} \left(\mathbf{x}^{(1)}\right)^{\mathsf{T}} \\ \left(\mathbf{x}^{(2)}\right)^{\mathsf{T}} \\ \vdots \\ \left(\mathbf{x}^{(1999)}\right)^{\mathsf{T}} \\ \left(\mathbf{x}^{(2000)}\right)^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} -118.29 & 33.91 & 1{,}416 & 38{,}372 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

    - h is your system’s prediction function, also called a hypothesis.
    - RMSE(X,h) is the cost function measured on the set of examples using your hypothesis h.

- In some context, you may prefer to use another function:

  - If there are many outlier districts. In that case, you may consider using the mean absolute error (MAE)

    - $$\mathrm{MAE}\left(\mathbf{X}, h\right) = \frac{1}{m} \sum_{i=1}^{m} \left| h\left(\mathbf{x}^{(i)}\right) - y^{(i)} \right|$$

- Both the RMSE and the MAE are ways to measure the distance between two vectors: the vector of predictions and the vector of target values.

Check the assumptions

- The Risk: You could spend months perfecting a regression model (predicting exact prices), only to discover that the next system converts those prices into simple categories ("cheap," "expensive"). If that were the case, your problem should have been one of classification from the start, not regression.

- The Lesson: Confirming exactly what the downstream system needs ensures you're addressing the right problem before investing time in development.