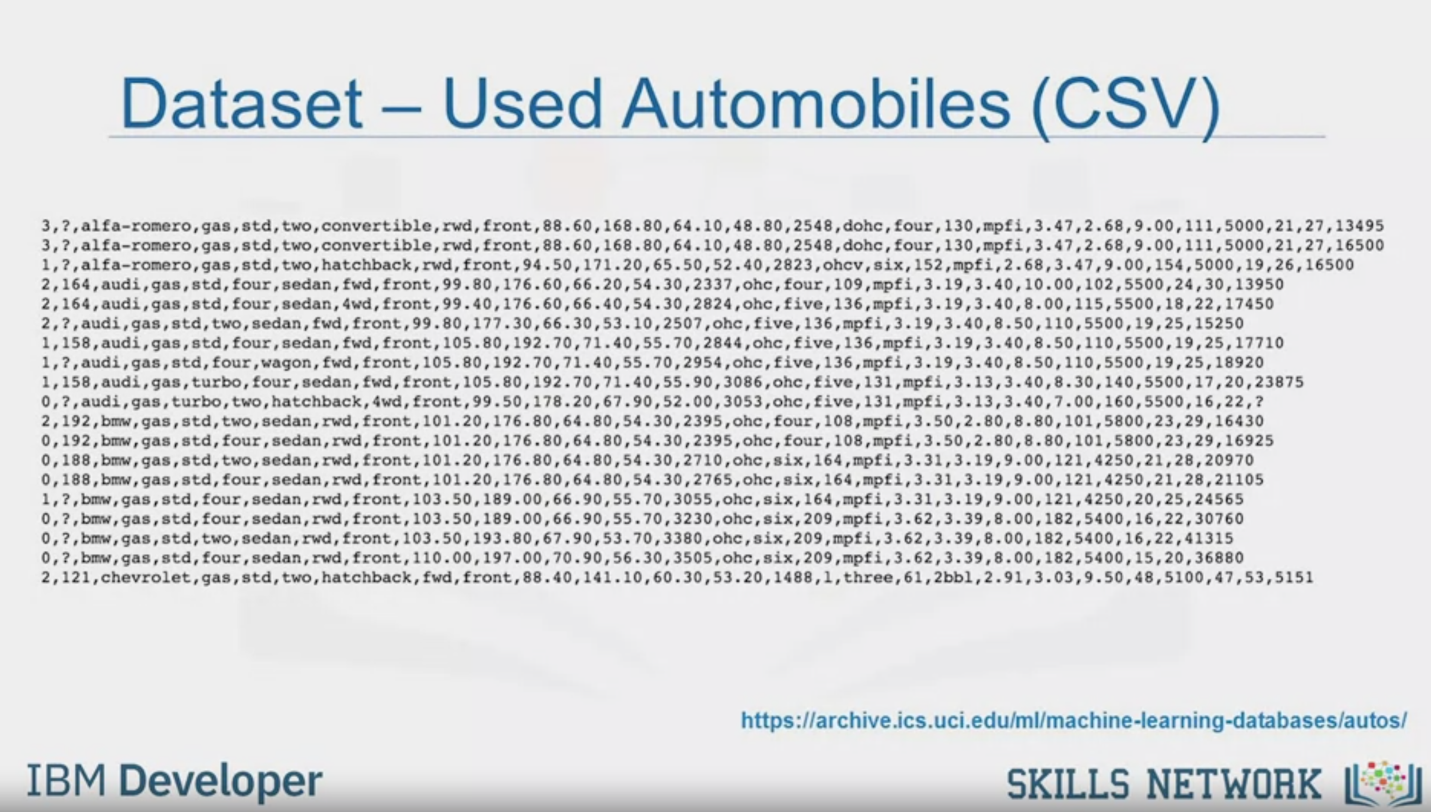


Understanding the Data

IBM Developer

SKILLS NETWORK

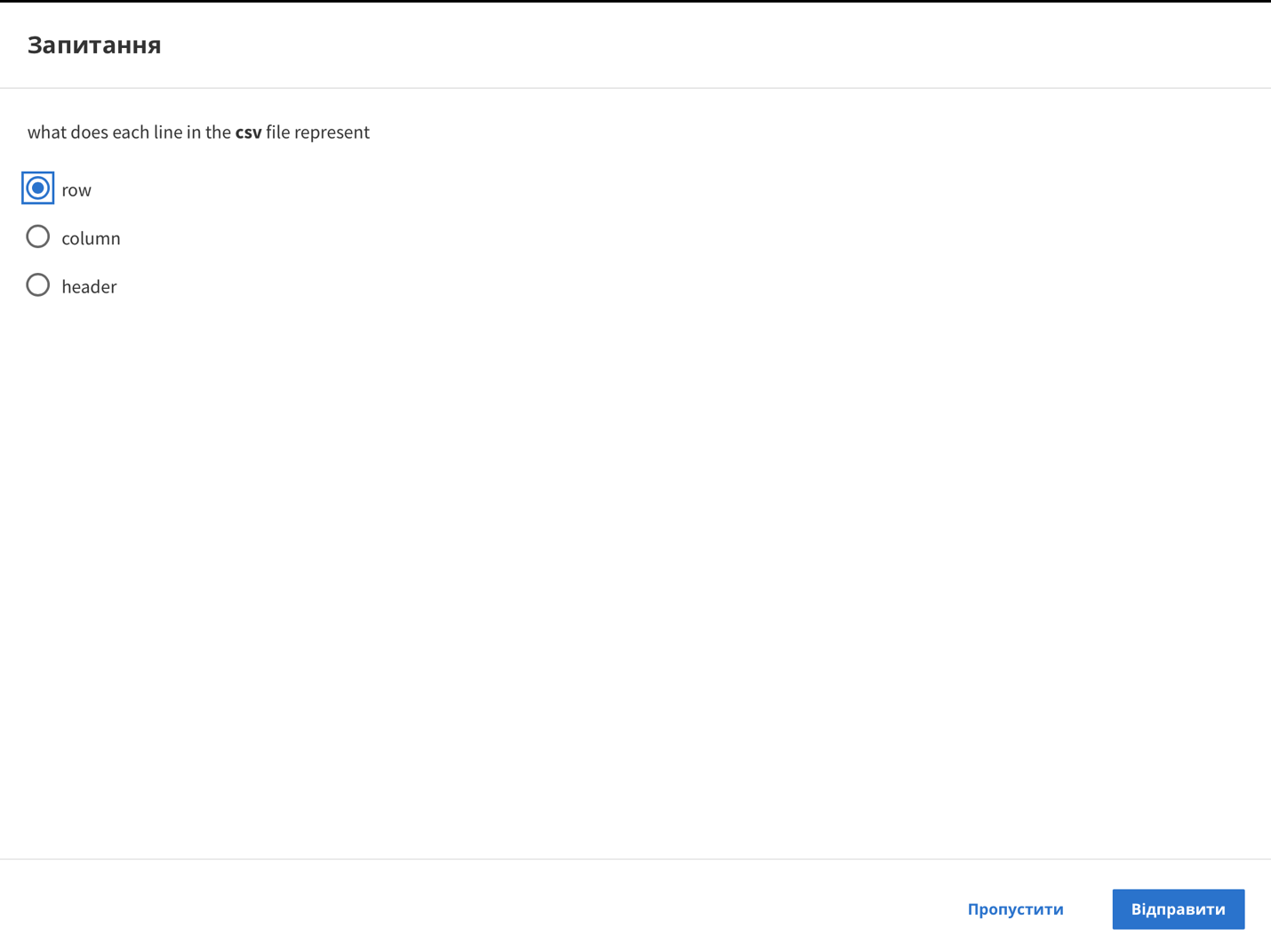


Dataset - Used Automobiles (CSV)

<https://archive.ics.uci.edu/ml/machine-learning-databases/autos/>

IBM Developer

SKILLS NETWORK



Question

what does each line in the csv file represent

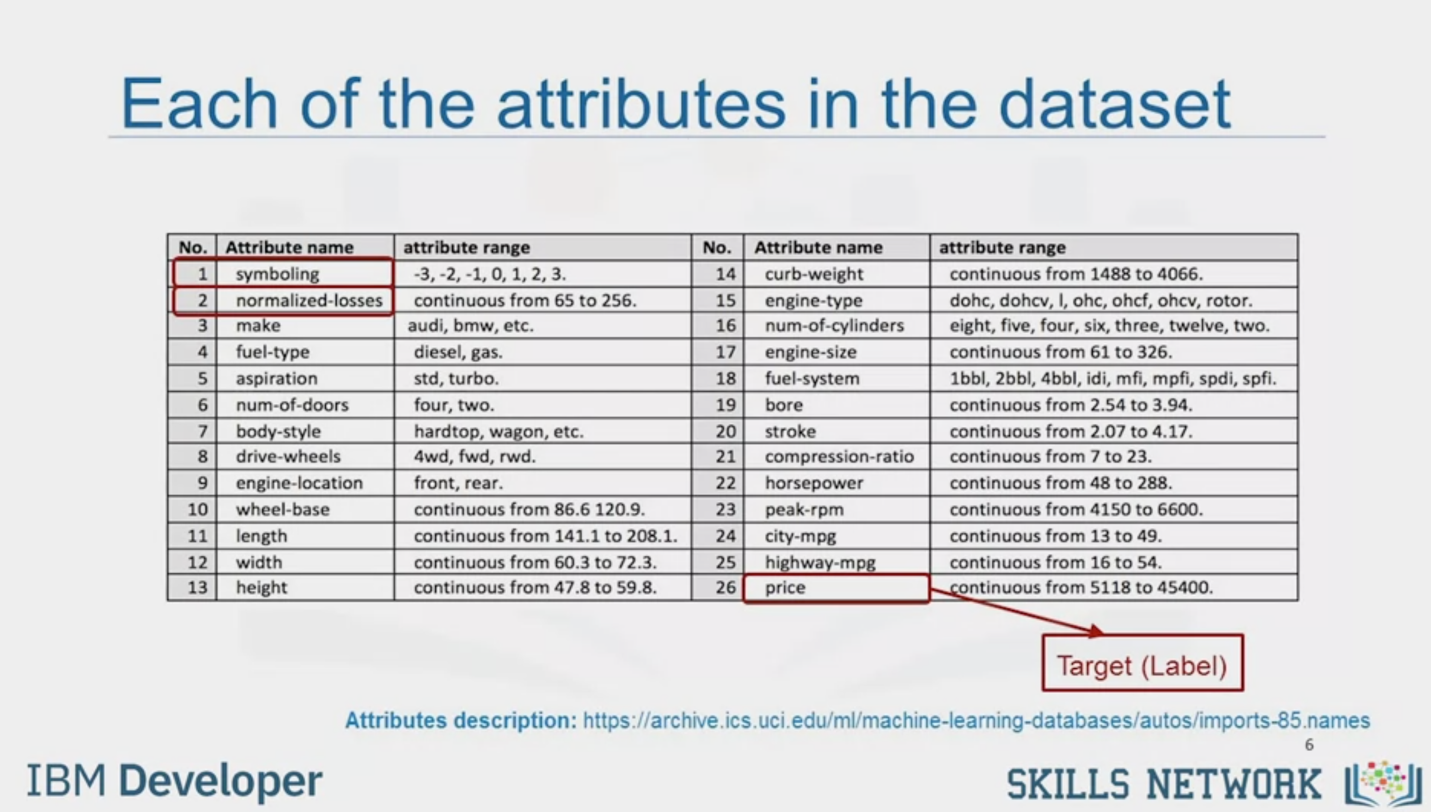
row

column

header

Skip

Send



Each of the attributes in the dataset

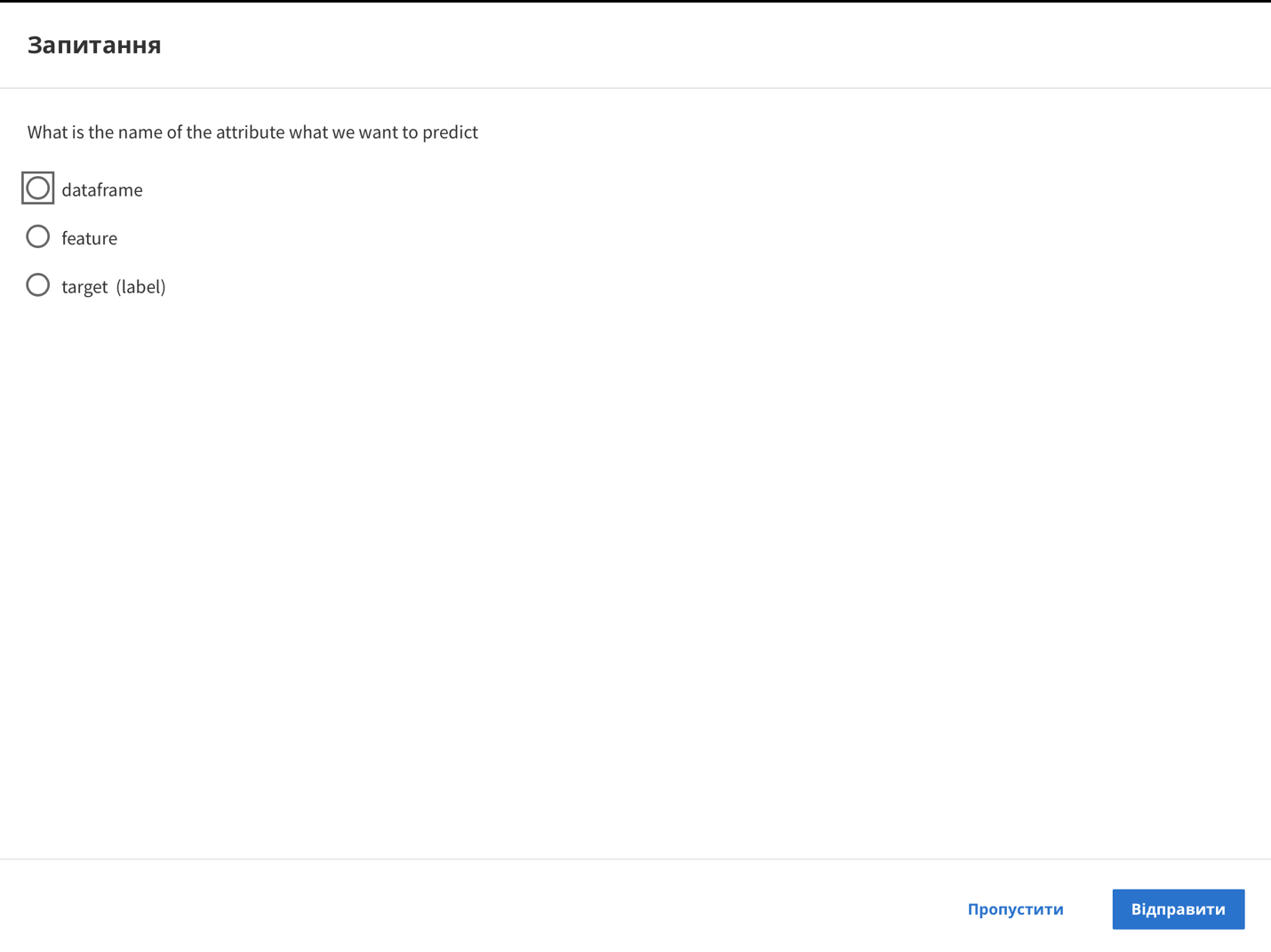
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No. | Attribute name | attribute range | No. | Attribute name | attribute range |
| 1 | symboling | -3, -2, -1, 0, 1, 2, 3. | 14 | curb-weight | continuous from 1488 to 4066. |
| 2 | normalized-losses | continuous from 65 to 256. | 15 | engine-type | dohc, dohcv, I, ohc, ohcf, ohcv, rotor. |
| 3 | make | audi, bmw, etc. | 16 | num-of-cylinders | eight, five, four, six, three, twelve, two. |
| 4 | fuel-type | diesel, gas. | 17 | engine-size | continuous from 61 to 326. |
| 5 | aspiration | std, turb. | 18 | fuel-system | 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi. |
| 6 | num-of-doors | four, two. | 19 | bore | continuous from 2.54 to 3.94. |
| 7 | body-style | hardtop, wagon, etc. | 20 | stroke | continuous from 2.07 to 4.17. |
| 8 | drive-wheels | 4wd, fwd, rwd. | 21 | compression-ratio | continuous from 7 to 23. |
| 9 | engine-location | front, rear. | 22 | horsepower | continuous from 48 to 288. |
| 10 | wheel-base | continuous from 86.6 120.9. | 23 | peak-rpm | continuous from 4150 to 6600. |
| 11 | length | continuous from 141.1 to 208.1. | 24 | city-mpg | continuous from 13 to 49. |
| 12 | width | continuous from 60.3 to 72.3. | 25 | highway-mpg | continuous from 16 to 54. |
| 13 | height | continuous from 47.8 to 59.8. | 26 | price | continuous from 5118 to 45400. |

Target (Label)

Attributes description: https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.names

IBM Developer

SKILLS NETWORK



Question

What is the name of the attribute what we want to predict

dataframe

feature

target (label)

Skip

Send



SKILLS NETWORK

IBM Developer

In this video, we'll be looking at the dataset on used car prices.

The dataset used in this course is an open dataset by Jeffrey C. Schlemmer.

This dataset is in CSV format,

which separates each of the values with commas,

making it very easy to import in most tools or applications.

Each line represents a row in the dataset.

In the hands-on lab for this module,

you'll be able to download and use the CSV file.

Do you notice anything different about the first row?

Sometimes the first row is a header,

which contains a column name for each of the 26 columns.

But in this example, it's just another row of data.

So, here's the documentation on what each of the 26 columns represent.

There are a lot of columns and I'll just go through a few of the column names,

but you can also check out the link at the bottom of

the slide to go through the descriptions yourself.

The first attribute, symboling,

corresponds to the insurance risk level of a car.

Cars are initially assigned a risk factor symbol associated with their price.

Then, if an automobile is more risky,

this symbol is adjusted by moving it up the scale.

A value of plus three indicates that the auto is risky.

Minus three, that is probably pretty safe.

The second attribute, normalized-losses,

is the relative average loss payment per insured vehicle year.

This value is normalized for all autos within

a particular size classification, two door small,

station wagons, sports specialty,

etc., and represents the average loss per car per year.

The values range from 65 to 256.

The other attributes are easy to understand.

If you would like to check out more details,

refer to the link at the bottom of the slide.

Okay, after we understand the meaning of each feature,

we'll notice that the 26 attribute is price.

This is our target value or label in other words.

This means price is the value that we want to predict from the dataset and

the predictors should be all the other variables listed like symboling,

normalized-losses, make, and so on.

Thus, the goal of this project is to predict price in

terms of other car features. Just a quick note.

This dataset is actually from 1985.

So, the car prices for the models may seem a little low.

But just bear in mind that the goal of this exercise is to learn how to analyze the data.