UNIVERSIDAD ICESI
DATA ANALYTICS
Aníbal Sosa and Andrés Aristizábal

Class Project

Description

This data set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe. The third factor 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/speciality, etc...), and represents the average loss per car per year.

Attribute Information:

Attribute: Attribute Range

1. symboling: -3, -2, -1, 0, 1, 2, 3.
2. normalized-losses: continuous from 65 to 256.
3. make:
alfa-romero, audi, bmw, chevrolet, dodge, honda,
isuzu, jaguar, mazda, mercedes-benz, mercury,
mitsubishi, nissan, peugot, plymouth, porsche,
renault, saab, subaru, toyota, volkswagen, volvo

4. fuel-type: diesel, gas.
5. aspiration: std, turbo.
6. num-of-doors: four, two.
7. body-style: hardtop, wagon, sedan, hatchback, convertible.
8. drive-wheels: 4wd, fwd, rwd.
9. engine-location: front, rear.
10. wheel-base: continuous from 86.6 120.9.
11. length: continuous from 141.1 to 208.1.
12. width: continuous from 60.3 to 72.3.
13. height: continuous from 47.8 to 59.8.
14. curb-weight: continuous from 1488 to 4066.
15. engine-type: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. num-of-cylinders: eight, five, four, six, three, twelve, two.
17. engine-size: continuous from 61 to 326.
18. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. bore: continuous from 2.54 to 3.94.
20. stroke: continuous from 2.07 to 4.17.
21. compression-ratio: continuous from 7 to 23.
22. horsepower: continuous from 48 to 288.

23. peak-rpm: continuous from 4150 to 6600.
24. city-mpg: continuous from 13 to 49.
25. highway-mpg: continuous from 16 to 54.
26. price: continuous from 5118 to 45400.

The idea of this project is to predict the insurance risk rating of a car as well as characterizing the different segments of the population.

Steps to follow

1. **Cleaning and EDA**: Check for data quality issues.
   You must evaluate the quality of the data, as well as understanding the relationship between features and the target variable.
2. **Predictive models**: Train predictive models (at least 3) that will allow you to estimate the insurance risk rating of a car from the values of the other variables. Choose the best model, looking for its optimal parameters.
   You must include a section in which you establish the evaluation protocols and the models' training and evaluation processes.
3. **Dimensionality reduction:** Considering all the variables, perform a principal component analysis (PCA), choosing the number of components necessary to preserve at least 80% of the original representation.
4. **Characterizing the cars:** With the data in its new representation, perform a segmentation, establishing the best number of clusters between 3 and 5. Characterize the clusters with respect to the original variables.

Rubric

Data Quality	Data Visualization	Extracting insights from data	Uderstanding and cleaning data	Training and evaluation protocols	Training the 3 models	Dimensioanlity reduction with PCA	Characterizing the groups of cars	TOTAL
0.3	0.7	0.7	0.3	0.3	1.0	0.7	1.0	5.0