Pricing game data dictionary

December 2019

Everything except the last two columns in the data can be used for predictions. The test data includes all columns except the last two.

1. **id_policy**: A unique ID for contracts
2. **pol_bonus**: The bonus/malus system is compulsory in France. The coefficient is attached to the driver. It starts at 1 for young drivers (i.e. first year of insurance). Then, every year without claim, the bonus decreases by 5% until it reaches its minimum of 0.5. Without any claim, the bonus evolution would then be: 1, 0.95, 0.9, 0.85, 0.8, 0.76, 0.72, 0.68, 0.64, 0.6, 0.57, 0.54, 0.51, 0.5. Every time the driver causes a claim (only certain types of claims are taken into account), the coefficient increases by 25%, with a maximum of 3.5. Thus, the range of pol bonus extends from 0.5 to 3.5 in the dataset.
3. **pol_coverage**: The coverage are of 4 types : Mini, Median1, Median2 and Maxi, in this order. Mini policies cover only Third Party Liability claims, whereas Maxi policies covers all claims, including Damage, Theft, Windshield Breaking, Assistance, etc.
4. **pol_duration**: Policy duration represents how old the policy is. It is expressed in year, accounted from the beginning of the current year i.
5. **pol_sit_duration**: Represent how old the current policy characteristics are. This is different from pol_duration, because the same insurance policy could have evolved in the past (e.g. coverage, or vehicle, or drivers, ...).
6. **pol_pay_freq**: The price of the insurance coverage can be paid annually, bi-annually, quarterly or monthly. Be aware that you must provide a yearly cotation in your answer to the pricing game.
7. **pol_payd**: The pol_payd is a boolean (i.e. a string with Yes or No), which indicates whether the client has subscribed a mileage-based policy or not.
8. **pol_usage**: The policy use describes what usage the driver makes from the vehicle, most of time. There are 4 possible values : WorkPrivate which is the most common, Retired which is presumed to be aimed at retired people (who also are presumed to drive less). Professional which denotes a professional usage of the vehicle, and AllTrips which is quite similar to Professional (including professional tours).
9. **pol_insee_code**. INSEE code is 5-digits alphanumeric code used by the French National Institute for Statistics and Economic Studies (hence INSEE) to identify communes and departments in France. There are about 36,000 ‘communes’ in France, but not every one of them is present in the dataset . The first 2 digits of insee code identifies the ‘department’ (they are 96, not including overseas departments).
10. **drv_drv2**: This boolean (Yes/No) identifies the presence of a secondary driver on the contract. There is always a first driver, for whom characteristics (age, sex, licence) are provided, but a secondary driver is optional, and is present in roughly a third of contracts.
11. **drv_age1**: This is the age of the first driver in years.
12. **drv_age2**: When drv_drv2 is Yes, then the secondary driver’s age is present. When not, this is 0.
13. **drv_sex1**: European rules force insurers to charge the same price for women and men. But gender can still be used in academic studies, and that’s why drv_sex1 is still available in the datasets. For our purposes, let’s assume that this can be used as discriminatory variable in this pricing game.
14. **drv_sex2**: The gender of the second driver if present in the data.
15. **drv_age_lic1**: The age of the first driver’s driving license in years.
16. **drv_age_lic2**: The age of the second driver’s license. Be wary of outliers in the data.
17. **vh_age**: This variable is the vehicle’s age, the difference between the year of production and the current year. One can consider values of 1 or 2 to correspond to new vehicles.

18. **vh_cyl**: The engine cylinder displacement is expressed in ml in a continuous scale.

19. **vh_din**: A representation of the engine power. Don't be surprised to find correlations between cylinder displacement, power, speed or even the value of the vehicle.

20. **vh_fuel**: Diesel, Gasoline or Hybrid. This is the fuel type of the vehicle.

21. **vh_make**: The make (brand) of the vehicle. As the database is built from a French insurance set, the three major brands are Renault, Peugeot and Citroën.

22. **vh_model**: As a subdivision of the make, vehicle is identified by its model name. The are about 100 different make names in the datasets, and about 1,000 different models. If you used the, consider maybe concatenation of the two variables.

23. **vh_sale_begin**: The difference between the current year and the year in which the vehicle first went into production.

24. **vh_sale_end**: The difference between the current year and the year in which the vehicle officially went out of production.

25. **vh_speed**: This is the maximum speed of the vehicle, as stated by the manufacturer.

26. **vh_type**: Tourism or Commercial. You'll find more Commercial types for Professional policy usage than for WorkPrivate.

27. **vh_value**: The vehicle's value (replacement value) is expressed in euros.

28. **vh_weight**: The weight (in kg) of the vehicle.

29. **town_mean_altitude**: The altitude of the town centre.

30. **town_surface_area**: Approximate surface area of the town.

31. **population**: The population of the town

32. **commune_code**: Commune, canton, city district, and regional department codes can act as lookup code to figure out exactly which area the policy belongs too.

33. **canton_code**

34. **city_district_code**

35. **regional_department_code**

36. (TARGET VARIABLE) **claim_amount**: The amount of a claim expressed in euros for the year (a value of zero indicates that no claim was made).

37. (TARGET VARIABLE) **made_claim**: A binary variable with a value of 1 indicating that a claim has been made.