2023 Travelers Analytics Case Competition

InsNova Auto Insurance Company Modeling Problem

1) Business problem

You work for InsNova Auto Insurance Company, an Australian company. Your business partner, who is not familiar with statistics at all, would like you to create a rating plan based on the historical auto claim data. Your business partner is concerned about segmentation as well as competitiveness, as there are several other competitors in the market.

For this case competition, your group's task is to provide a method for predicting the claim cost for each policy and to convince your business partner that your predictions will work well.

2) Data Description

The modeling data is attached with this email (InsNova_train.csv). The InsNova data set is based on one-year vehicle insurance policies from 2004 to 2005. There are 45,239 policies, of which around 6.8% had at least one claim. The data is split to two parts: training data and validation data. In the validation data, claim_cost, claim_ind and claim_counts are omitted. You can build your model on the training data. In the end, use your best model to score the validation data. We will evaluate your model based on your validation data prediction.

Variable information in the data:

- ID: policy key
- Veh_value: market value of the vehicle in $10,000's
- Veh_body: Type of vehicles
- Veh_age: Age of vehicles (1=youngest, 4=oldest)
- Veh_color: Color of vehicles
- Engine_type: Engine type of vehicles
- Max_power: Max horsepower of vehicles
- Driving_history_score: Driving score based past driving history (higher the better)
- Gender: Gender of driver
- Area: Driving area of residence
- Dr_age: Driver's age category from young (1) to old (6)
- Marital_status: Marital Status of driver (M = married, S = single)
- E_bill: Indicator for paperless billing (0 = no, 1 = yes)
- Time_of_week_driven: Most frequent driving date of the week (weekdays vs weekend)
- Time_driven: Most frequent driving time of the day
- Trm_len: term length (6-month vs 12-month policies)
- Credit_score: Credit score
- High_education_ind: indicator for higher education
- Exposure: The basic unit of risk underlying an insurance premium
- Claim_ind: Indicator of claim (0=no, 1=yes)
- Claim_counts: The number of claims
- Claim_cost: Claim amount

3) Modeling

Each group may have at most 5 people and will:

a. Work together within group but not between groups (They can't provide extra info or help – please contact us directly if anything comes up).
b. Build a model to predict the claim cost and submit the predicted cost for the Validation data as a csv file. The format requested is provided in the attachment.
c. Prepare a presentation for your business partner to summarize your analysis results. You do not need to explain the problem, just summarize what you did and what you found (see questions to answer in section 8 below).
d. Each group can make at most 3 submissions per day

4) Benchmark Model

The benchmark will be LightGBM model. We will provide it before the first optional submission.

5) Model Evaluation & Competition Logistics

The model will be evaluated using the Gini index. We will calculate your score once you submit your result.

The teams scoring better than the benchmark will move on to the second stage, the virtual live presentations. Each qualifying team will give a 5-7 minute presentation on the above questions followed by a 3 minute Q&A session. The top teams will be eligible for consideration as the overall campus winner.

The winning campus team will join other winning teams for a virtual job shadow day at the Travelers Hartford campus and make final presentations to a panel who will determine the ultimate winner! The ultimate winner each year going forward will be engraved on a trophy which will be showcased at Travelers for posterity!

6) Contacts:

a. Business problem: John Scheele JTSCHEEL@travelers.com; Gazi Inkiyad GINKIYAD@travelers.com; Wenye Qiu WQIU@travelers.com
b. Kaggle tech: Ziye Tao ZTAO@travelers.com; Susie Li: SLI6@travelers.com; Swapnanil Banerjee: SBANER19@travelers.com
c. Campus representative:
   - UCONN: Lubing Wang LWANG5@travelers.com,
   - UMass Amherst: Ji Ah Lee JLEE26@travelers.com
   - Duke: Kim Cheng KCHENG@travelers.com; Yi Yi YYI3@travelers.com
   - University of Iowa: Yong Qiao YQIAO@travelers.com
   - Rutgers University: Zexi Song ZSONG@travelers.com

7) Presentation instructions

Submit your code with documentation along with your presentation answering the following questions:

Recall that the audience of the presentation is your non-statistician business partner

a. What methods did you consider (you don't have to have actually tried all of these methods)?
b. What method did you choose in the end and why?
c. How did you do the variable selection?
d. What variables help explain pure premium?

e. What other variables not in the data set do you think might be useful?

Some bonus items to think about and address if possible:

f. How did you test the assumptions of this method?
g. How did you evaluate your model (e.g. fit statistics, over-fitting, etc.)?
h. Any concerns about the resulting model?
i. What questions to you have about the data?

Disclaimer

InsNova Insurance Company and the data is a fictitious example used for the purpose of this competition only.