4th McMaster/Co-operators Problem-Solving Workshop

Qualification round information

February 2023

Introduction

Thank you for your interest in participating in this 4th workshop! We hope you will have a good experience in answering this year's problematic.

Qualification round problematic

The problematic for the qualification round is about predicting the probability of cross-selling another product to an existing client.

In a general insurance company like Co-operators who sells multiple products, the analytical professionals use insights from data to determine which profile of existing customer is mostly like to purchase another product from the company. This will help the company to target its marketing communication to the right group of audience. In this problematic, we ask you to predict the probability of an existing customer who has a home insurance policy with the company to purchase an automobile insurance policy in the next one-year period.

You are provided with a dataset which contains information about existing clients who has a home insurance policy with a general insurance company, and an indicator on whether this client purchased an automobile policy from the company in the next one-year period.

The data dictionary in the next section will go into more details about the information available. The data are anonymized for information security purposes.

Data dictionary

This dataset contains information on client level. For each existing client, you have information about the client, the home insurance he/she has with the company and some information on his/her auto insurance need.

We assume each client is insured under his/her own home insurance policy and has completely separated auto insurance need from other clients (i.e. there is no situation such that two clients are from the same home or is driving the same car).

Here are the variable definitions of the dataset.

policyid: an integer to uniquely identify the policy the client is insured under. ids are thus all different.

responseVariable: 0 if the client did not purchase an auto insurance policy in the next one-year period; 1 if the client purchased an auto insurance policy in the next one-year period

Gender: Gender of the customer

policyHolderAge: Age of the policy holder

hasCanadianDrivingLicense: 1 if the client has a Canadian driving license; 0 if the client does not have a Canadian driving license

territory: Anonymized code of geographical territory of where the client's address is

hasAutoInsurance: 1 if the client currently has auto insurance; 0 if the client currently does not have auto insurance

hadVehicleClaiminPast: Whether the client has any auto insurance claim in the past.

homeInsurancePremium: The premium of the client's current home insurance.

saleChannel: Anonymized code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.

isOwner: 1 if the client is the owner of the insured home; 0 if the client is not the owner of the insured home

rentedVehicle: 1 if the vehicle is rented; 0 if the vehicle is not rented

hasMortgage: 1 if the insured home has mortgage on it; 0 is the insured home does not have mortgage on it; N/A if the insured is not the owner of the home

nbWeeksInsured: Number of weeks the insured has been with the company. All clients have only been with the company for less than a year (52 weeks) by the valuation date of the data.

vehicleStatus: Categorical variable based on the age of the vehicle the client drives. Old > Recent > New

Datasets for the qualification round

Two datasets are provided for the qualification round.

The first one **<TrainingDataset_2023Qualification.csv>** contains 149,780 lines and all the 15 columns defined above. This is meant to be the training dataset on which you will build your model. A good practice could be to slice this training dataset to use a train-test modulization approach, or alternatively one could prefer to use a cross validation modulization approach, as to reduce the risk of overfitting.

The second one **<ScoringDataset_2023Qualification.csv>** contains 40,118 lines and all the columns defined above, except for the responseVariable. This is meant to be the dataset on which you will submit prediction result of your model and that will be used for ranking the teams in the qualification round. We call it the scoring dataset.

What you have to deliver

By the end of the qualification round, you will have to predict whether the client in the scoring dataset will purchase an automobile insurance policy in the next one-year period. Your prediction should be either 1 (the client will purchase an automobile insurance policy) or 0 (the client will NOT purchase an automobile insurance policy). You should add this prediction to the scoring dataset in a column named predictedResponseVariable and send us back a csv file with only two columns: policyid and predictedResponseVariable. A submission example **<SubmissionExample_2023Qualification.csv>** has been provided for your reference. Please respect the csv file type and DON'T FORGET TO INCLUDE both policyid and predictedResponseVariable!

Your csv file has to be sent by email to mcmaster_workshop@cooperators.ca by the end of the qualification round on **March 1st, 2023**.

If you have any questions, you can send them at this email address as well.

Method for ranking the teams

Your goal is to predict for the profile of client who is more likely to purchase an automobile insurance policy in the next one-year period. We will evaluate your model by its ability in selecting profiles of clients that did purchase an automobile insurance policy.

We have the information concerning the observed indicator of whether the client purchased an automobile policy for the 40,118 lines of the scoring dataset, i.e. the missing column responseVariable. We will merge this information to your scored dataset (using policyid as a key for merging, this is why this column has to be there!), then calculate AUC on the scored

dataset (definition of AUC: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc). The teams which produced the highest AUC will be invited to the final in-person workshop day at McMaster University with team members from Co-operators on site!

Resources

Here is an example to model a 0/1 response variable using binomial GLM in R:

https://bookdown.org/ndphillips/YaRrr/logistic-regression-with-glmfamily-binomial.html

Many more resources are available on the internet using machine learning algorithm, look for them!

There will be some office hours provided by Actuarial or Business Intelligence team at Co-operators to answer any question you have. Stay Tuned!

Final remarks

We have given here an example of a GLM built in R, but you are free to use whatever model you want to answer the problematic and whatever software to build your model, as long as you do build a model on your own and return a csv file with minimally your predictions and the identifiers.

Have fun!

4th McMaster/Co-operators Problem-Solving Workshop organizing committee

mcmaster_workshop@cooperators.ca