Data Analytics Task

Purpose

The purpose of this task is to allow you to demonstrate your technical skills as well as your ability to communicate results effectively.

Background

Imagine that you are currently employed as a consultant. Your client is a small life insurer named DataIns. DataIns offers a product that covers the risk of being hit by a car. If a policyholder is hit by a car, then the policyholder is paid a lump sum amount equal to the medical costs incurred by the policyholder and the policy contract is terminated. The cost of purchasing this policy is $0.20 per day.

Objective

The CFO of DataIns has approached you with a question: he would like to know whether there are certain policyholder segments that are more profitable than others. To help you answer the CFO's question, some claims and exposure data have been provided as at 2018-11-30.

In a brief email addressed to , please discuss the following points:

- Which types of customers are more profitable.
- Which types of customers are especially risky (i.e. claim more).
- Which types of customers should DataIns focus on marketing to.

Please attach any code/working to your email and ensure that any insights suggested in your email are reproducible.

Tip: The extent to which you use R (preferably the data.table and ggplot2 packages) will reflect favourably on your application.

Data dictionary

Attached are two data files.

The first is an exposure file that records each policy ever sold by DataIns. The columns in this are:

- ID – a unique identifier for the policyholder.
- SMOKER – a TRUE/FALSE field that records whether the policyholder is a smoker or not.
- OCCUPATION – a field that records the policyholder's occupation.
- POLICY_START_DATE – the date at which time the policyholder bought their policy.
- POLICY_END_DATE – the date at which the policyholder ended their policy.
    - A POLICY_END_DATE represents either a claim or lapsation.
    - If the POLICY_END_DATE is missing, assume that the policy has not ended yet.
- TOTAL_PREMIUM_PAID_TO_DATE – the amount of premiums that DataIns has received from the policyholder to date.

The second file is a claims file that records whether a policy has received a claim payment from DataIns. You can assume that the claims file is complete (i.e. there are no outstanding claims that are not recorded in the claims file). The claims file has the following columns:

- ID – the policyholder identifier.
- CLAIM_SIZE – the amount paid out to the policyholder.
- CLAIM_DATE – the date of the claim payment.