Introduction

Motor insurance is at the core of Direct Line Group's business and we are constantly exploring new ways of improving. We are currently in the process of launching a Data Science Innovation squad to look at various machine learning pricing techniques.

The Challenge

The goal of this exercise is to explore the attached data to better understand what drives claims and who is more likely to claim, and by how much. There is no "right" answer (we haven't even consciously stated a question!) and we know that there are many different ways to approach this, so pick something you think is appropriate and tell us why. We expect the entire piece to take a few hours of your time at most, as we know everyone is busy!

The technical part of the challenge is to provide a Jupyter notebook to do this. In the notebook, you should also provide brief rationale on why you have chosen a particular methodology. This should outline the pros and cons of the solution, any considerations and give reference to any potential alternative solutions that were considered. Be prepared to discuss this work when you visit us.

Feel free to modify, supplement, enhance or transform the data as you see fit.

Mandatory Goals:

- Build a model to identify better/worse risks when it comes to providing car insurance
- Clearly communicate the methodology with consideration of the pros and cons of the provided solution

Desirable Goals:

- Accuracy of the solution, based on whatever metrics you deem fit
- Any non-technical considerations hat can be drawn out

Data Set

A single file has been provided, with data from the US, containing information about drivers and whether they claimed or not in the past year. Where they have claimed, the dataset outlines how much they were paid out. The fields included in the dataset include:

REFERENCE - ID Variable • BIRTHDAY - Birthday • AGE - Age of the driver. • CAR_COST - Value of vehicle. • CAR_AGE - Vehicle age. • CAR_TYPE - Type of car. • CAR_USE - Vehicle use. • CLM_FREQ – Number of claims in the past 5 years. • EDUCATION - Education level. • HOME_CHILDREN – Number of children at home. • HOME_VAL - Home value. • INCOME - Income. • OCCUPATION - Job type. • CHILD_DRIVE – Number of driving children • MARITAL_STATUS - Marital status. • MVR_PTS - Motor vehicle record points. • OLDCLAIM - Total claim value in the past 5 years • REVOKED – Has license been revoked in the past 7 years. • Gender – M/F • PTIF – The length of time the policy has been in force. • TRAVTIME - Distance to work • CITY_RURAL - City vs. rural home area. • JOB_TENURE - Years on job. • CLAIM_FLAG: Was there a claim this year? • CLM_AMT: If there was a claim, for how much was it?

Your submission must include:

The code in a Jupyter notebook along with an HTML version of it.

A brief rationale on your chosen methodology, considering pros and cons, findings, considerations and potential next steps. Just an outline of your thoughts will suffice, this does not need to look super slick!

Please provide all of the above submission items in a single zip file which should be returned via email. If during the recruitment process you require any reasonable adjustments, please just let us know.

By accepting this test you also agree to delete all test related documents and data no later than two weeks after receiving.