Data Science Take Home Exercise

These are the instructions for the take home exercise for candidates interviewing. Please do not share these instructions, the accompanying data, or your analysis with anyone. You may spend as much time on this exercise as you like between now and the mutually agreed deadline and use whatever resources you wish to look things up if you get stuck. However, we estimate that most candidates should be able to complete the tasks in approximately 3 hours and will be able to do so using standard tools. You may use Python programming language and associated libraries you wish to answer the exercise.

Instructions

Your client is a car insurance company. They want to price their car insurance competitively, which means having a good model for customers at risk of getting into accidents. They have shared with you a sample of data in CSV format (attached to the email that included these instructions) that they would like you to analyse. Each row corresponds to a customer, the **outcome** column records whether the customer made a claim in the previous year or not. The client has informed you that the other columns should be self-explanatory.

Note: The data for this exercise has been generated randomly, so may display some regularity that would not be expected of real-world data

Exploratory Questions

The client has some specific questions that they would like you to answer that they have not been able to answer themselves.

Perform any pre-processing / cleaning of the data necessary to answer these questions.

1. What proportion of customers with a credit score below 0.2 made a claim in the last year?
2. What is the average number of speeding violations among customers with driving experience between 20 and 29 years (inclusive)?
3. What if you consider only the people in this group who drive a sports car?
4. What is the standard deviation in annual mileage?

As well as these specific questions, you suspect that they just want to understand their data better.

1. Are there particular customer types?
2. What do claimants have in common?
3. How does the number of claims vary between postcodes?

Explore the data and present some of your findings to help the client understand their data better. These could be summary statistics, or visualisations.

In addition to understanding the data they have; the client is interested to know how they should collect data in the future in order to better support data science work.

1. Are there any problems with the data you have been given that should be kept in mind when modelling?
2. Has the client collected the right data for their business needs?
3. What recommendations would you make to the client for future data collection?

Modelling

The client is interested to know if the customer data can be used to predict the likelihood that a claim is made in the next year. Your task is to investigate this and make a recommendation. You should complete the following tasks:

1. Briefly discuss any assumptions being made about the data
2. Build a proof-of-concept model to predict the **outcome** column from the customer data, including any necessary data processing
3. Test your model using appropriate metrics and state how you would expect it to perform on unseen data.
4. The client is keen to be able to interpret the model you build and would be particularly interested in understanding which features are most important to the model's decisions.
5. They are also new to data science and interested in how this exciting new model you've built them works. Write a brief (no longer than a paragraph) description of how your model works that can be understood by someone without a technical background.

Submitting

Please send us all the code you wrote, plus documentation and answers to questions by email in reply to the email originally sent to you. We recommend a **Colab** in **iPYNB** and **HTML** as an ideal format for showing code and results, documentation and presentation of the results into a separate file.