Senior Data Scientist Task

Why this task?

The task below is designed to help us understand how you approach data analysis & data modelling. It should take you no more than 3-4 hours to complete. Don't worry if you have not managed to do everything you would have liked to do within that timeframe - just email back what you have and add notes on what you would have done if you were given more time.

The scenario

The company is growing fast, and our Salary Sacrifice product is flying off the shelves. Website traffic is increasing, the Customer Success team is doing an excellent job at generating interest in existing customers.

Currently, the sales team is asked to get in touch with every single lead (i.e. potential sale) that is created on our CRM. They're doing an excellent job at converting leads into sales, but with the volume of leads growing rapidly, we'll soon have too much workload for our EV sellers.

After chatting to the Head of EV Sales, you've had an idea: what if we were able to focus EV sellers' attention on "good" leads exclusively, by calculating people's propensity to buy?

The Task

You want to use customer data to develop a model that calculates the probability of a lead converting into a sale.

Use the dataset attached to:

- Perform some exploratory data analysis (what does the data tell us about who buys with us?)

- Build a model to predict propensity to buy

- Propose next steps to the Head of EV Sales

The Data

The data was collated by querying our data platform. It contains 1 row per lead. The dataset contains 50,000 rows.

A lead is generated in our CRM as soon as a person logs into our website. Some of the information recorded is manually inserted by the person logging in.

The dataset is a snapshot of the lead's attribute after their first website session, together with information about whether they eventually converted to a sale or not.

The schema:

- `lead_id` – the unique identifier of the lead
- `lead_entry_time` – the time the person logged into the website for the first time (i.e. the time the lead was generated)
- `has_placed_order` – a boolean expressing whether the lead converted into a sale
- `time_order_placed` – a timestamp showing when the lead converted to a sale (if ever). It will be NULL if the lead never converted
- `salary` – salary of the lead
- `date_of_birth` – date of birth of the lead
- `n_engaged_minutes` – how many minutes they spent actively engaging with content on the website
- `n_of_cars_viewed` – how many cars they viewed on our website
- `company_industry` – industry of the company they work for (broad category)
- `day_of_week` – day of the week the lead was generated
- `price_of_last_car_viewed` – price (per month) of the last car selected (which you can assume, for the purpose of this exercise, to be the car they are most interested in)
- `step_reached_in_website` – how far they got in the online journey. The options are:
    - `homepage` (they got to the homepage but no further)
    - `car_clicked` (they clicked on one or multiple cars to view them)

    - `car_configured` (they clicked on one or multiple cars, and configured their quote)
    - `quote_generated` (they generated a final quote)
- `how_did_you_hear_about_us` – how they heard about our service

The Deliverables

You can submit your work however you like (scripts, notebook, etc.), but your model code should use **Python**.

Requirements:

- Please ensure you include all of your exploratory data analysis, as well as the conclusions you are drawing from it
- Ensure your code is reproducible and well commented
- Include model evaluation and what conclusions you draw from it
- Please include text explaining your proposed next steps (this can be rough bullet points)

As you complete the exercise, you should **keep the following questions in mind:**

- Who seems more likely to buy a car with us?
- How do you propose the Head of EV Sales make use of your work?
- Are there risks associated with using this model? How would you mitigate them?
- What does a “productionised” version of this work look like?