Assessment Task

Please use R or Python to answer the below questions concisely using statements, graphs and tables where applicable, making use of any packages of your choice. Jupyter Notebook or R script(s), R Markdown are all acceptable. You have 1 week to complete this test so please do not feel rushed and it is “open book”, so feel free to use any online or offline resources during this time.

After completion, please email back an archived folder with all your answers (data, scripts, graphs etc.) and your initials in the file name.

If anything is unclear or you are unsure how to answer a question, please note this down in the email and do not worry.

This is an exchange of ideas exercise; we will be paying more attention to how you put your thoughts forward and how clear they are other than the performance of the model.

**Objective:**

Domestic and General is interested in using Data Science to lead their pricing strategy on their new business book. In this exercise, you will be tasked with producing a set of pricing analysis and a demand model.

**Data:**

One file is provided, **Train \_Test.csv**, please split the file into 2 sets as more convenient to you. A description of each variable is provided on the second page. Data at Domestic and General is generally collected at the contact centre and as such the data may have mistakes or missing values. You may choose to impute or omit these errors in your analysis.

**Tasks:**

1. Explore the DS\_train.csv dataset conduct any EDA that you feel will help you understand the data.
2. Compare the price elasticity of the two strategies in the dataset, flat and random price.
   1. Price elasticity
3. Build how many demand models you want and report on the metric(s) of your choice. The model should predict the likelihood of the customer accepting an offer.
   1. Hyper parameter optimisation may take a lot of time to complete, this is not a requirement, but I’d like your comments on how you would approach this.
4. Engineer at least 1 new feature to improve the demand model. Assess whether the new feature(s) should be included.
5. For the top 5 features, please provide a partial independence plot. What’s the most important and why (qualitatively speaking)?
6. Last question, now we’d like to see the impact of a 10% flat price increase in the demand. Use your model to recalculate the probability demand with the new 10% increase and compare the results after the change.

**DS\_Train.csv variables are as follows:**

1. **Claims\_Amount (float):** sum of total value of claims in the last period
2. **Claims\_Count (int):** number of claims in the last period
3. **Purchase\_Date (date):** date when the appliance was purchased
4. **Purchase\_Price (float):** value paid for the appliance
5. **Sale\_Flag (binary):** sale flag (this is the target)
6. **Period\_of\_Cover (int):** how long the manufacturer’s guarantee was whether it was annual or extended
7. **Premium (float):** base premium
8. **Plan\_Flag (binary):** flag whether the customer has or has had a plan with D+G
9. **Age (int):** Age of appliance in days
10. **Price\_Diff (float):** price difference between base premium and sold premium
11. **Cover\_Start\_Date:** date when the cover will start
12. **Plan\_Count (int):** number of plans the customer has
13. **Price test (binary):** indicates the price test group (not a modelling feature)
    * ‘1’: random price
    * ‘0’: no increase
14. **Account (cat):** type of account (this has been encoded already)
15. **Category (cat):** category of appliance (this has been encoded already)