Readme

1. All the scripts including the HTML file can be (once downloaded) - accessed via Jupyter Lab

2. Brief testing of Github Codespace for instant reproducible runs
    - Very low memory on a basic account
    - New – not sure if a long term project / plan
    - Handy tool to avoid downloads and dependencies etc
    - Not sure how long data sits in memory yet – have not been through the documentation
        i. i.e seems sometimes packages have to be reinstalled with each run after a certain period of time
    - Linux Based

3. **Engine**
    - Contains the scripts used to power the GBM model and general code

4. **Project**
    - Contains the script to build
        i. EDA
        ii. Feature engineering & Data Cleaning
        iii. Data Modelling
5. EDA Profile needs to be downloaded to view the HTML file for viewing – should open on any browser – or as stated above in Jupyter Lab

Data Used

1. Completely generated by samples – so will differ on each iteration
    a. Locations are completley random so on any given run their may be no way to get a post code – the latitudes & longitudes are random so it could be a desert, lake, ocean, etc which don't possess a postcode – just a demo
    b. Could alternatively opt for the package in Python - **pip install random-address**
    c. Could alternatively create random addresses from a given shapefile – plenty of approaches – this is simply a demo

Process & Findings

1. Code written in Jupyter Lab
    a. Script provides documentation for the process step by step
    b. A regression model was used

    c. Currently working on a classification model
2. All the data are Numeric or Date objects

3. No Missing data to either exclude or impute

4. Data Cleaning
    a. The file Profile contains the initial EDA stages to view the data
    b. Cleaned Age to a more sensible bracket between 18 & 80 for this exercise
    c. Date formatting to make consistent

5. Modelling

6. Futher investigation
    a. The extreme tails on factors such as
        i. **Claims Amount** – Where are these occurring & why.
            1. Is there a specific time, date, region, etc. associated with these?
    b. Correlation between **Claims Amount** & **Purchase Price**?

Future Upgrades to the current modelling process

1. Better understanding of the variables which are unclear to me
2. Feature enrichment
    a. More factors
        i. Geographic data
        ii. Credit scores
        iii. Occupation
        iv. Policy data (for example is it a New Line of Business or a Renewal? MTA?)
3. More data for a more accurate model
4. AvE charts
5. SHAP vs PDP
6. Maps – how granular by Area, Region, Postcode(may be too granular – unless the first segment of the postcode I.e NW1 rather than NW1 E193 – so split after the space 0
7. Parameter Tuning for the model

8. Compare results with current models
9. Best strategy
    a. Which type of model to build
    b. Which platform
    c. How to present the end results

Notes

- RMSE Metric is arbitrary – can easily be changed
- In past experiences I used Classification modeling for Elasticity
    - Attempt here was to try a regression model
    - Can easily adapt and fit instead a classification model
        - 0 & 1 scale
- At the moment the model show's likely hood of conversion
    - Dummy data so this is a just an exercise