Lorica Health Data Scientist Test

Submission:

You have three days from the time this is sent to you to submit your results. Please use either R or python code and submit your final code along with a report with your responses to the questions below (notebooks or markdown files are fine). The aim of this test is to see how you approach the problem, not necessarily to train the most accurate model possible.

Background

This test involves analysing a synthetically generated clinical data set from the Centres for Medicare and Medicaid Services (CMS). The claims data contained in the file *claim_outpatient.csv* relates to outpatient claims, where the doctor has not written an order to admit the patient to hospital. Typically, these claims are for emergency services, observation, outpatient surgery (single day), lab tests, x-rays, etc. Whilst in hospital, the patient is assigned diagnosis codes in the ICD9 scheme and may also have a diagnosis code for when they were admitted. Any procedures performed are also coded in the ICD9 scheme. The data also contains HCPCS codes (Healthcare Common Procedure Coding System), which are used to identify products, services, supplies, and pharmaceutical drug codes. All the claims contained in this data set have a 'J' HCPCS code which is used to identify drugs.

Clinical claims data may contain coding errors resulting in waste in the healthcare system, or over-charging of some items resulting in abuse of the system. The coding system may also be used fraudulently, resulting in large amounts of money being paid to providers and physicians for services not delivered. The overarching goal of this analysis is to identify these behaviours in the data, and the most common approach is to use outlier detection methods to flag unusual behaviour.

Data Files:

- *claim_outpatient.csv* – claim level data
- *beneficiary.csv* – patient information
- *data_dictionary.xlsx*

Tasks:

1. Load both files and explore the data.
2. Merge the two data sets together based on DESYNPUF_ID (patient / beneficiary id).
3. Create an 'Age' column on the merged data set at the date of the claim (if the claim date is missing, use a default date of 2010-12-31).
4. Train a linear regression model on claim benefit (CLM_PMT_AMT) using the available data as features, where appropriate. Evaluate the model and comment on the performance.
5. Improve on the linear regression model
    - You can create additional features, apply additional feature engineering, or use another type of model, but your technique must still predict benefit.
    - Evaluate your model and comment on the performance, comparing results to the first linear regression model.
6. Use your model to identify outliers based on benefit in the data. Discuss some characteristics of the outliers found.
7. How else would you approach the task of identifying waste, abuse and fraud in this sort of data? Describe an approach, no need to implement it.