POC – Insurance Claims Prediction

vLife Use Case
Insurance Claims Prediction Using Machine Learning

Background
In UnitedHealthcare’s 2017 Consumer Sentiment Survey, United Healthcare discovered that merely 9% of the Americans understand the four basic insurance components — premium, deductible, co-insurance, out-of-pocket maximum.

Given the lack of the knowledge that people possess about how insurance works, most individuals end up overestimating or underestimating the amount covered by their insurance.

We need a system that reduces unexpected financial burdens by giving people a better understanding of their expenditure on health care is highly desirable.


Approach
In this proof of concept, we used Multilayer Perceptron Regressor to predict the insurance claim amount and the responsibility amount(amount paid by the individual) for the patient with Ischemic heart disease or Coronary heart disease.

For this purpose, we built two models, one to predict the insurance claim amount and the other to predict the responsibility amount.
Insurance Claims Prediction using Machine Learning

Datasets

• **MIMIC III**, a vast medical database of EMR/EHR records, contains *multiple co-dependent tables* of patient data. This contains *biological* and *demographic* information of patients in the USA anonymized to avoid privacy concerns.

• **CMS SynPUF** US government research data contains the insurance claims of many individuals. It contains *demographic* information of patients in the USA and their *insurance details*, such as claim amount, responsibility amount, etc.

Challenges:

• One of the challenges in having a *single model* to predict the cost for *multiple diseases* was *not robust* enough, so we had to restrict the scope to just one disease.

• We chose *ischemic heart disease or coronary heart disease*, as the number of cases was sufficient.
Solution Architecture

Solution Approach

• The first step is to analyze the insurance claims (CLAIMS) data and filter out individuals with heart disease.
• Next, analyze the MIMIC (EMR) data.
• Merged the two databases based on five factors; disease, age, length of stay, gender, race.
• Performed univariate and bivariate analysis.
• Feature engineering was embedded into a pipeline which automatically performs standard scaling, non-linear transformation, feature selection, and training of the model.
• Built a predictive model using MLP classifier and 4 hidden layers.
• The model predicts the insurance claim amount and the responsibility amount (amount paid by the individual).
Output and Opportunities

Output
• The model can predict the total cost and out-of-pocket expenses for the medical procedure with **85%** accuracy.

Potential Opportunities
• Currently we have done just for one disease, coronary heart disease. The model can be scaled to cover other diseases.
• The model can be further developed to provide more insights and can be leveraged to minimize/mitigate these out-of-pocket expenses.
• Current solution is using ICD9 code systems, in order to use ICD10 code systems we can incorporate conversion systems.
THANK YOU