virtusa
Accelerating Business Outcomes

Insurance claim prediction - Using Machine Learning

VLIFE

By

Gagan K Shetty

Under the guidance of

Ankit Khatri and Avinash Reddy

A Look At The Problem

- **Majority of Americans do not know details about their health plans**
  - UnitedHealthcare's 2016 "Consumer Sentiment Survey" tested consumer knowledge of how insurance works by asking participants to define the four basic insurance components:
    - premium
    - deductible
    - co-insurance
    - out-of-pocket maximum
  - A mere **7%** of those surveyed had a complete understanding of the four concepts. While most understood the terms premium and deductible, only **about a third** understood the other two terms.

A Look At The Problem

- A Carnegie Mellon University economist, George Loewenstein, conducted a similar study that was later published in the Journal of Health Economics. He surveyed 202 employees' understanding of their employer-sponsored health insurance by testing their knowledge of those same four insurance terms. According to the survey, only **11%** could figure out what their **insurance would cover** given a hypothetical four-day stay in a hospital for a procedure.
- Given the paucity of the knowledge that the people possess about how insurance works, most individuals end up overestimating or underestimating the amount covered by their insurance.
- Having a system that does the calculation can aid people by giving them a better understanding of their expenditure on health care.

Who will be benefitted by this?

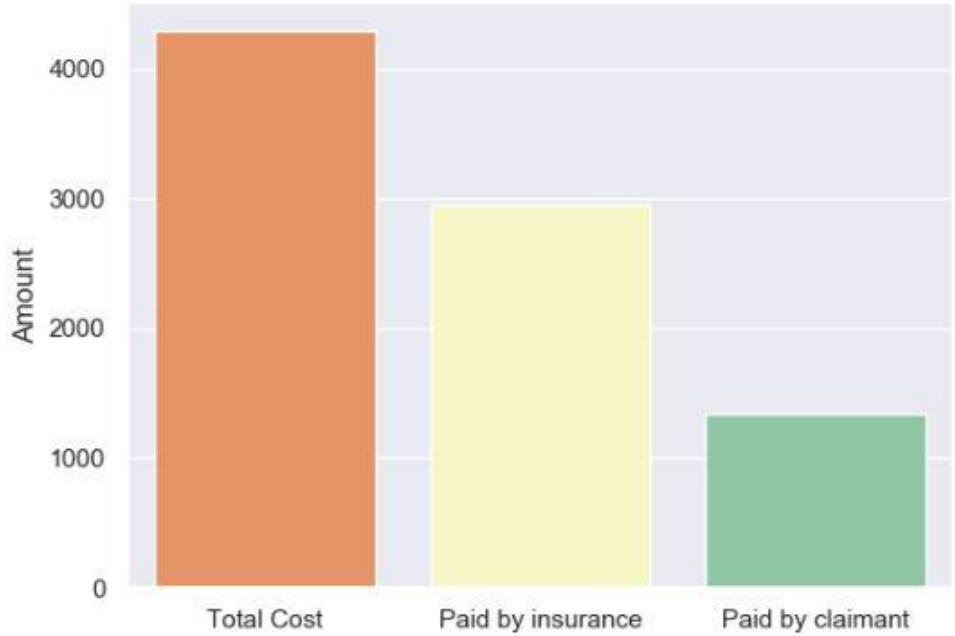
- There are two main perspectives:-
  - The patient's perspective : A patient who just made a visit to the hospital and finds out that he might have to undergo treatment for a particular disease can get a better understanding of the amount he is expected to pay, instead of just getting worried about the medical bills, or overestimating the amount covered by his insurance.
  - The insurance provider perspective : As an insurance provider, to maximise profit, the company must effectively price their insurance premium and assess the risk involved for a particular individual. Using predictive modelling over the data procured for previous insurance claims, an insurance company can effectively price the insurance premium. If they find that the individual has a high risk of claiming severe amounts, they may increase the premium to reflect the increased risk.
- The predictive model and the user interface built primarily focuses on the patient's perspective.

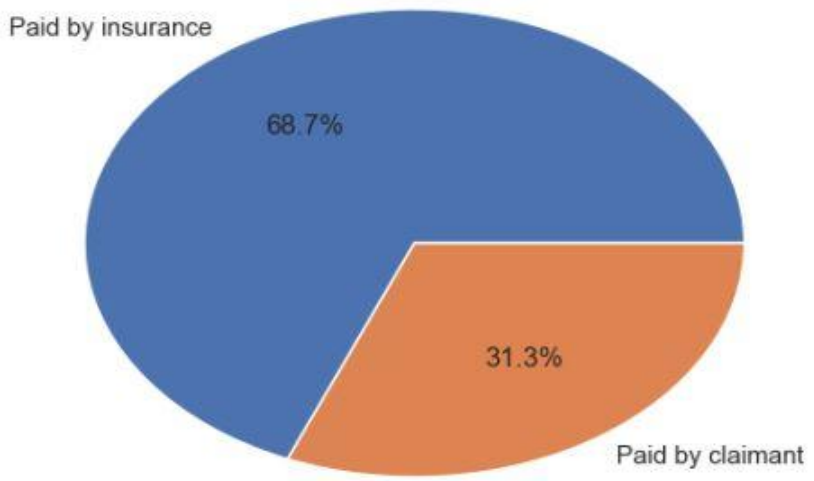
Approach and Outcome

- **CMS SynPUF US gov research data** contains the insurance claims of many individuals.
  - This contains **demographic** information of patients in USA and the **insurance details** such as claim amount, responsibility amount, etc.
- **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 due to privacy reasons.
- This use case aims at predicting the insurance claim amount and the responsibility amount(amount paid by the individual).
- For this purpose, we build two models, one to predict the insurance claim amount and the other to predict the responsibility amount.

The Desired Output

- The use case aims to have an interface where a person can enter his details and the model predicts the amount that insurance will cover and the amount he would have to pay on his own.
- An example for a particular test case is given below.





Technical Requirements

- The technical resources used are:
  - IPython Notebook
  - Python libraries:
    - Numpy
    - Pandas
    - Sci-Kit Learn
    - MatplotLib
    - Seaborn
    - Django
    - Scipy Stats
  - API to access the data.

Exploring the Data

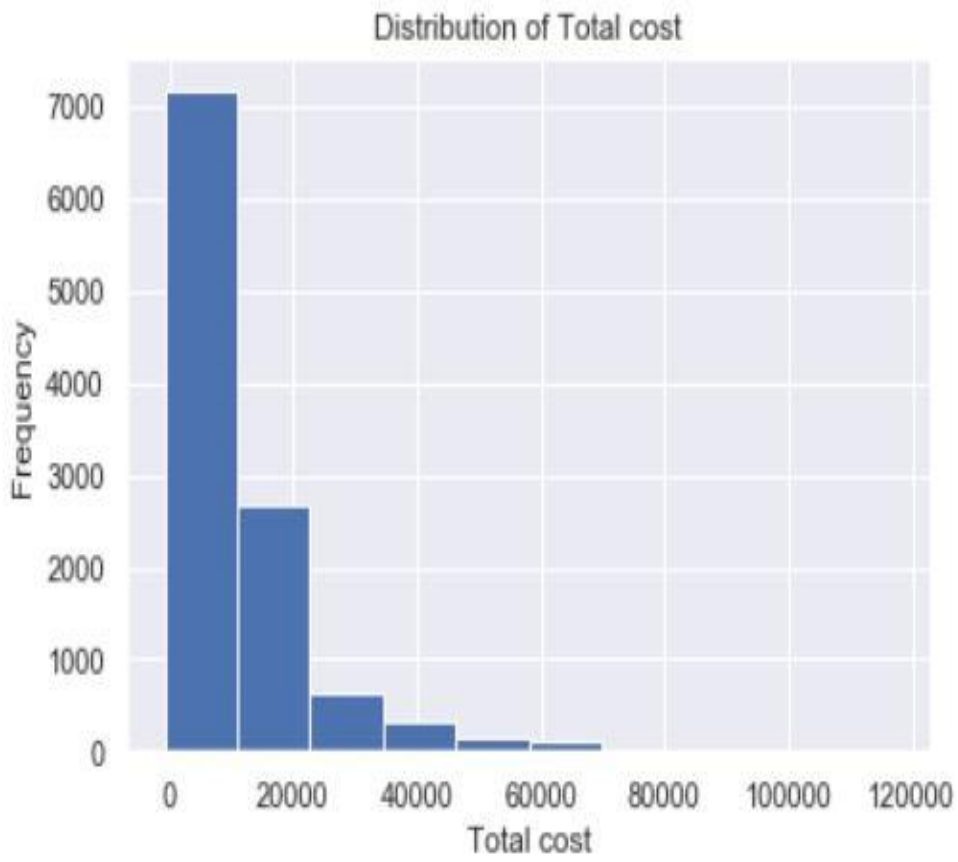
- The first step is to analyse the insurance claims database. This database consisted of around 11,000 individuals who had claimed for **ischemic heart disease**.
- This database contains information about the patients such as the demographics, insurance claim amount etc.
- The next step is to analyse the data in the mimic database. This database consists of the patient details such as demographics, procedures, icu stays, etc.
- After analysing both the databases, we imputed the insurance claim amount into the patient database by merging based on 5 factors:-
  - Disease
  - Age
  - Length of stay
  - Gender
  - Race

Data Structure and Parameters

- This section aims at **understanding the nature of the data**, forming conclusions on the **correlation of variables** and cleaning up erroneous values.
- The data is spread across **multiple tables** through unique links.
- A deeper look at the tables is required to **extract features** that *may* be relevant to the model.
- These can be later validated through **statistical testing.**
- First the data from the claims database is filtered based on the **disease**.
- Next the insurance claims cost are imputed into the Electronic health records based on the five features discussed before.

Claims database

- This is the univariate distribution of the total cost. Since the distribution doesn't follow a **normal distribution**, it is not possible to synthesize a total cost into the electronic health records database just based on mean and standard deviation of the data.
- Instead, we directly merged the databases by sampling the distribution formed after grouping the data based on the five parameters in the claims database.



Method for merging the databases

- First form multiple groups based on the 5 different merging parameters in the claims database. Each of this groups will form a distribution.
- For every record in the EHR database, find which group(group in the claims database) the record belongs to, and take a random sample from the group. Append the total cost and payable amount to the EHR record.
- Once we have imputed the total cost and payable amount into the ehr database, we can proceed with statistical analysis and predictive modelling.

Data Categorization

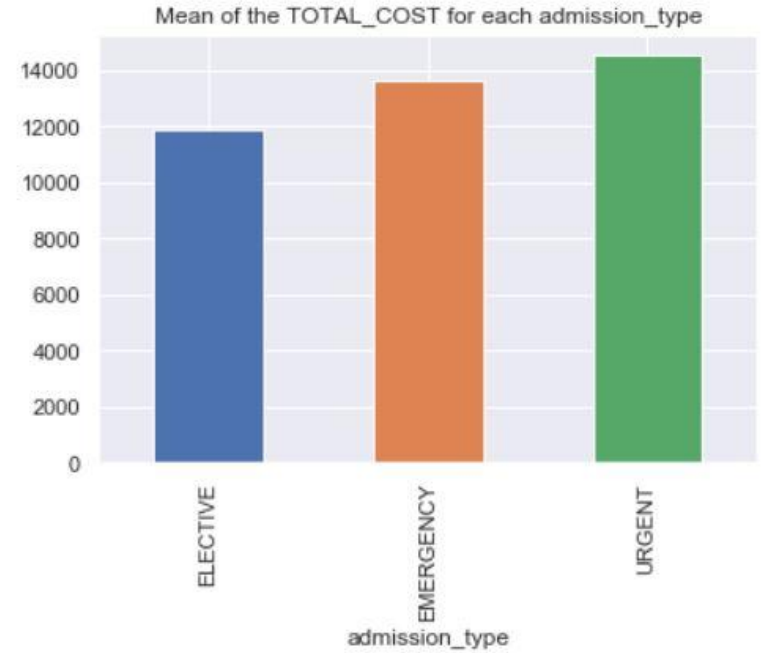
- A lot of categorical variables are **too diverse** for their own good and won't help us in making inferences.
- Some of the simplifications we shall implement are:
  - Reducing all races to White , Black, Hispanic or others.
  - Age, although continuous, is **binned into groups** of 10
  - Length of stay is also binned into 10 separate bins.

Bivariate Visualization

- We start analyzing the categorical and continuous variables versus our target variable, **Total cost and the Gap**
- We compared every independent variable with the Total cost and Gap variable to find out if there exists a correlation between them.
- The main variables considered are:-
  - Admission Type
  - Admission Location
  - Insurance
  - Gender
  - Ethnicity
  - Age
  - Length of stay

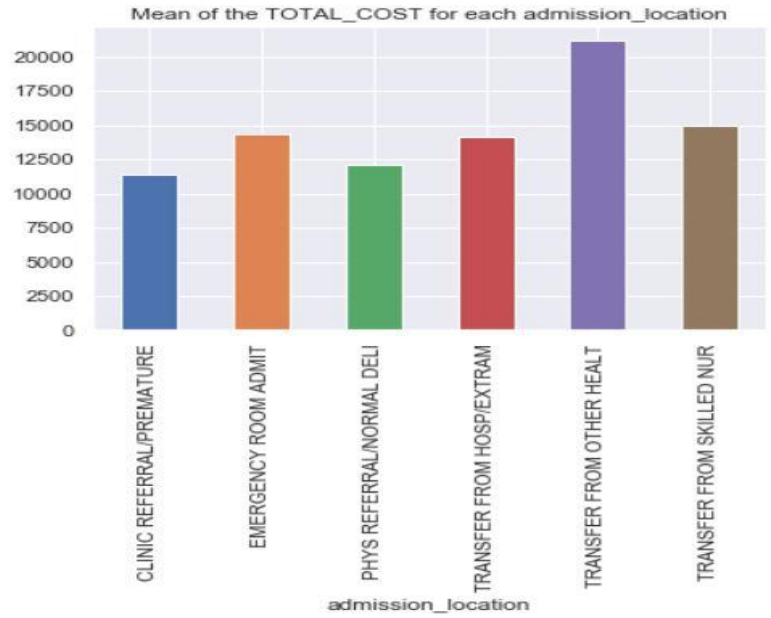
Admission Type

- **Urgent admissions** have a higher proportion of Total cost.
- This could possibly be due to symptoms of an injury that start appearing much later and end up requiring more remedial treatment.



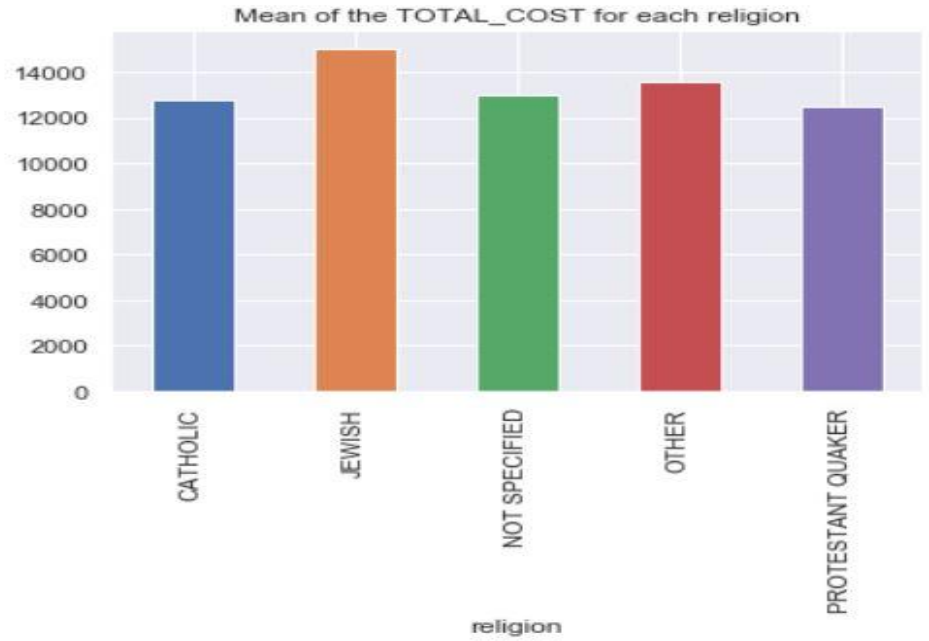
Admission Location

- **Transfers from other health facilities** again have higher correlations with the total cost.
- **Clinical referrals and normal deliveries** have lower costs.



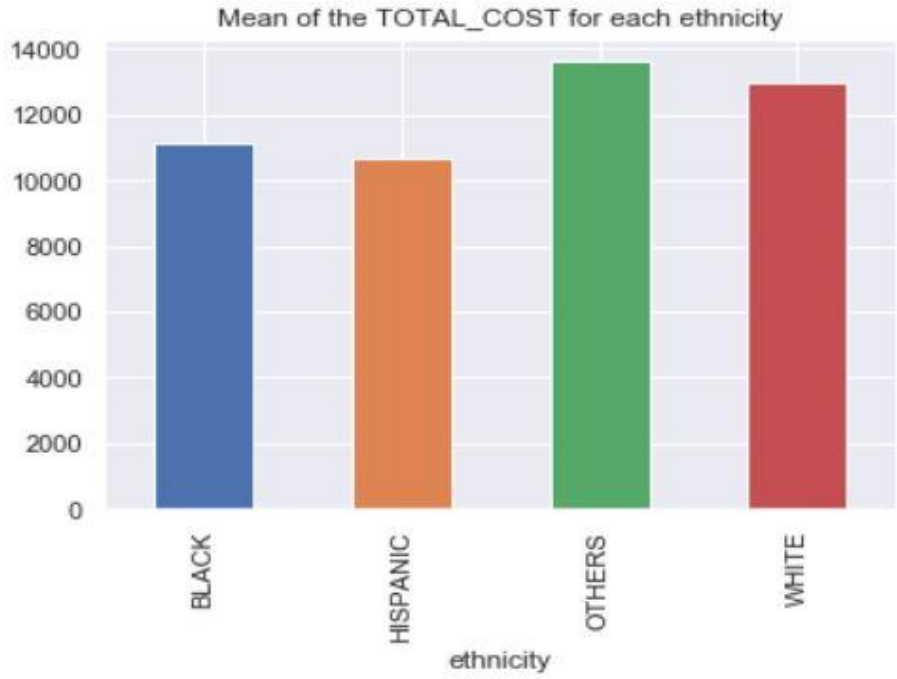
Religion

- There is not much to infer here other than the surprisingly high amount paid by **Jews**. Religion could represent **different lifestyles or beliefs** which could affect whether someone takes certain medications or not and thus may have an effect.



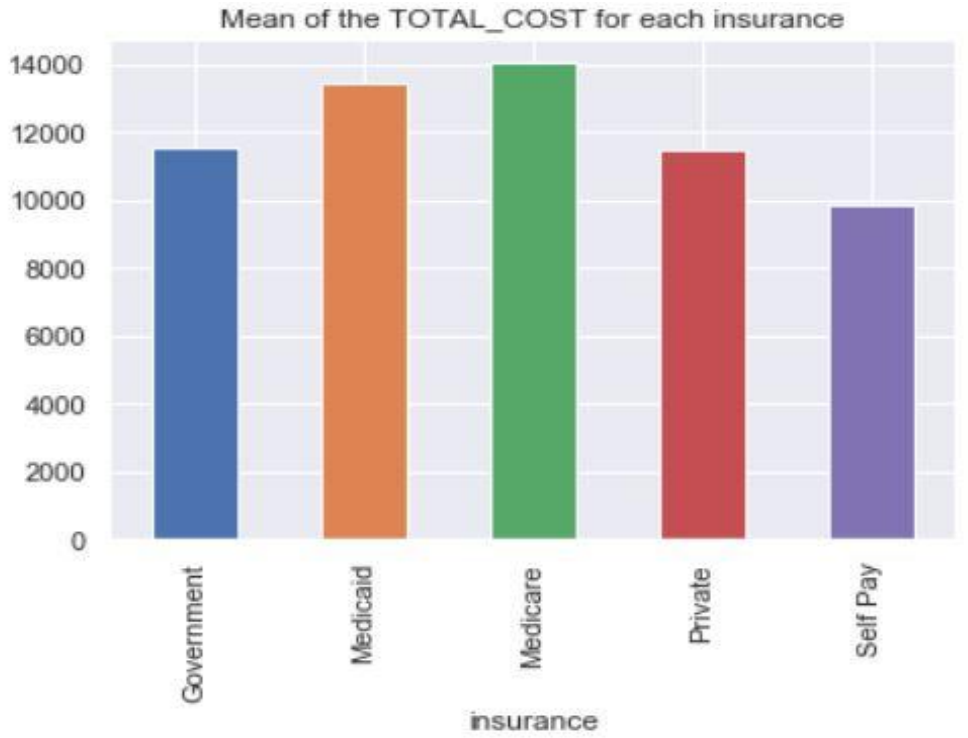
Ethnicity

- **Blacks, Hispanics** have lower expenditure for the disease. Again, this maybe due to cultural lifestyles which is not necessarily captured explicitly.



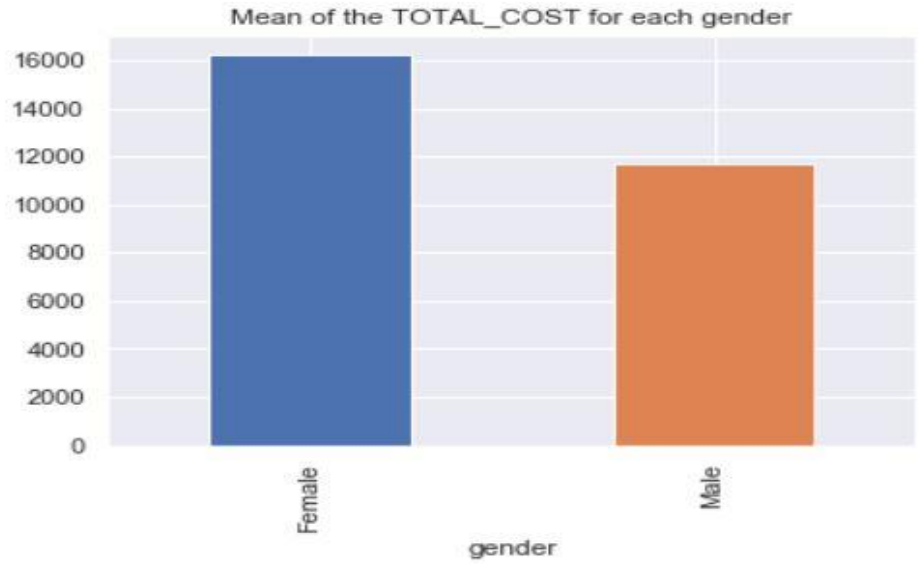
Insurance

- Insurance shows much higher cost for **Medicare** and **Medicaid** which cater to elderly and poor respectively. This could reflect elders having **inherently higher** risk of developing other symptoms and poor people having higher risks due to a worse off **lifestyle**.
- Self pay is lowest which could be because patients would not be willing to pay the high hospital admission fees out of their pocket



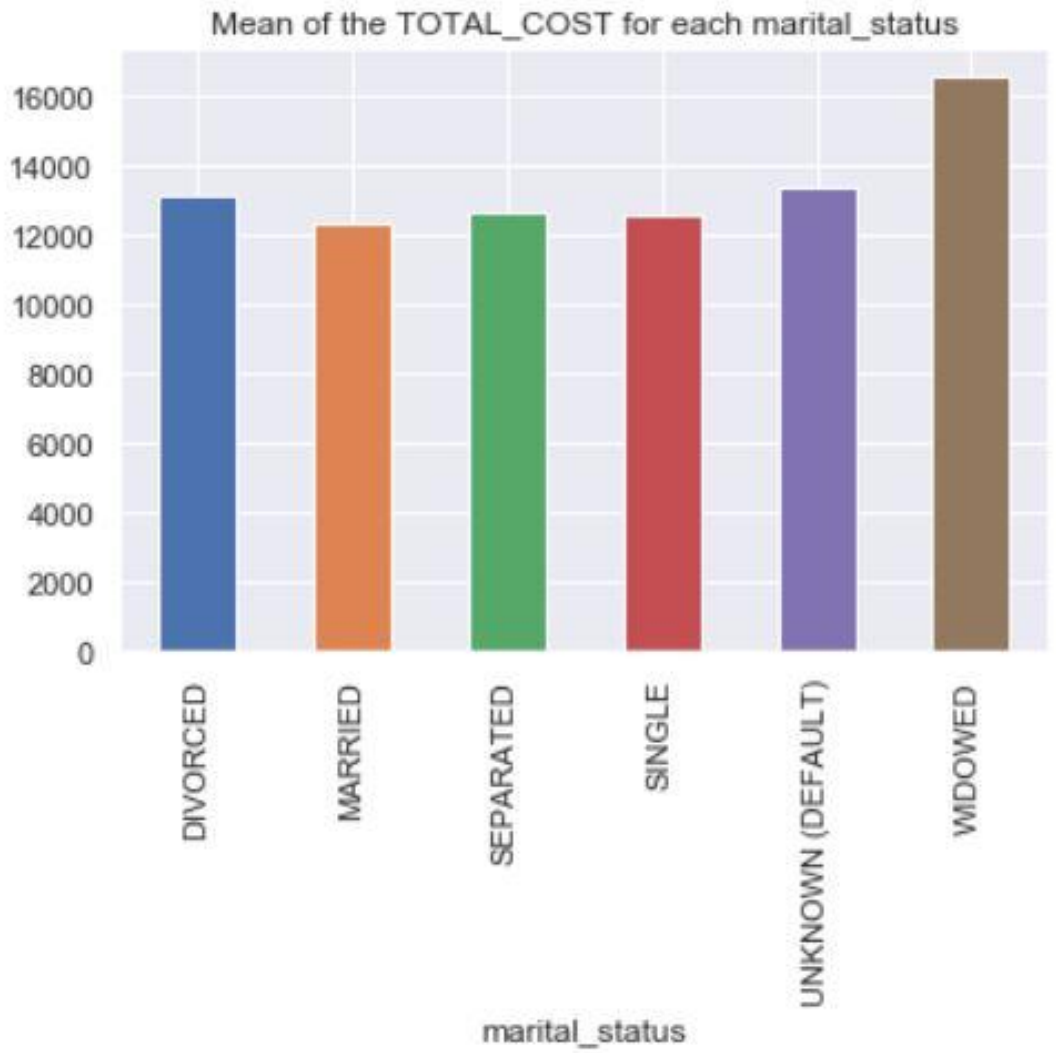
Gender

- Gender does show a significance in the total cost incurred for the treatment of the disease.
- This could be due to a difference in the effectiveness of the treatment among the different genders.



Marital Status

- Marital status has a biased trend to those individuals who were **Widowed.**



Feature Engineering

- Apart from Univariate and bivariate analysis, we plotted multivariate graphs and used statistical analysis methods.
- Since the data we had was not normally distributed, we had to use the Kruskal Wallis H test which is a non parametric test used to compare a categorical independent variable with the continuous dependent variable.
- Another technique used was the transformation of the Total cost into a **normally distributed** variable using the box cox transformation so we can use parametric tests such as anova.
- Once statistical tests are performed, we **one hot encode** all the categorical variables.
  - One hot encoding simply makes **every categorical value a feature** unto itself which is represented in a **binary format**.

Feature Engineering

- Other techniques used for feature engineering are:
  - Recursive feature elimination, where a subset of features are considered and recursively eliminated based on the feature importances.
  - Model based selection, where the best set of features are selected for a given model.
- This feature engineering was embedded into a pipeline which automatically performs the standard scaling, non linear transformation, feature selection and training of the model.

Building a Predictive Model

- The models we apply (in order) are:
  - **Linear Regression**
  - **Ridge Regression**
  - **Lasso**
  - **Elastic nets**
  - **Extra trees regressor**
  - **Multi layer perceptrons**
- Each of these are tuned and tested on training and testing sets to evaluate the Explained variance score.

Evaluation Metrics

- For evaluating algorithms we use
  - Explained variance score
  - Mean absolute error
  - Mean squared error

- The main criterion is the **Explained variance score.** This is an estimate of how much of the variance is explained by our model. Our priority is to improve the explained variance score.

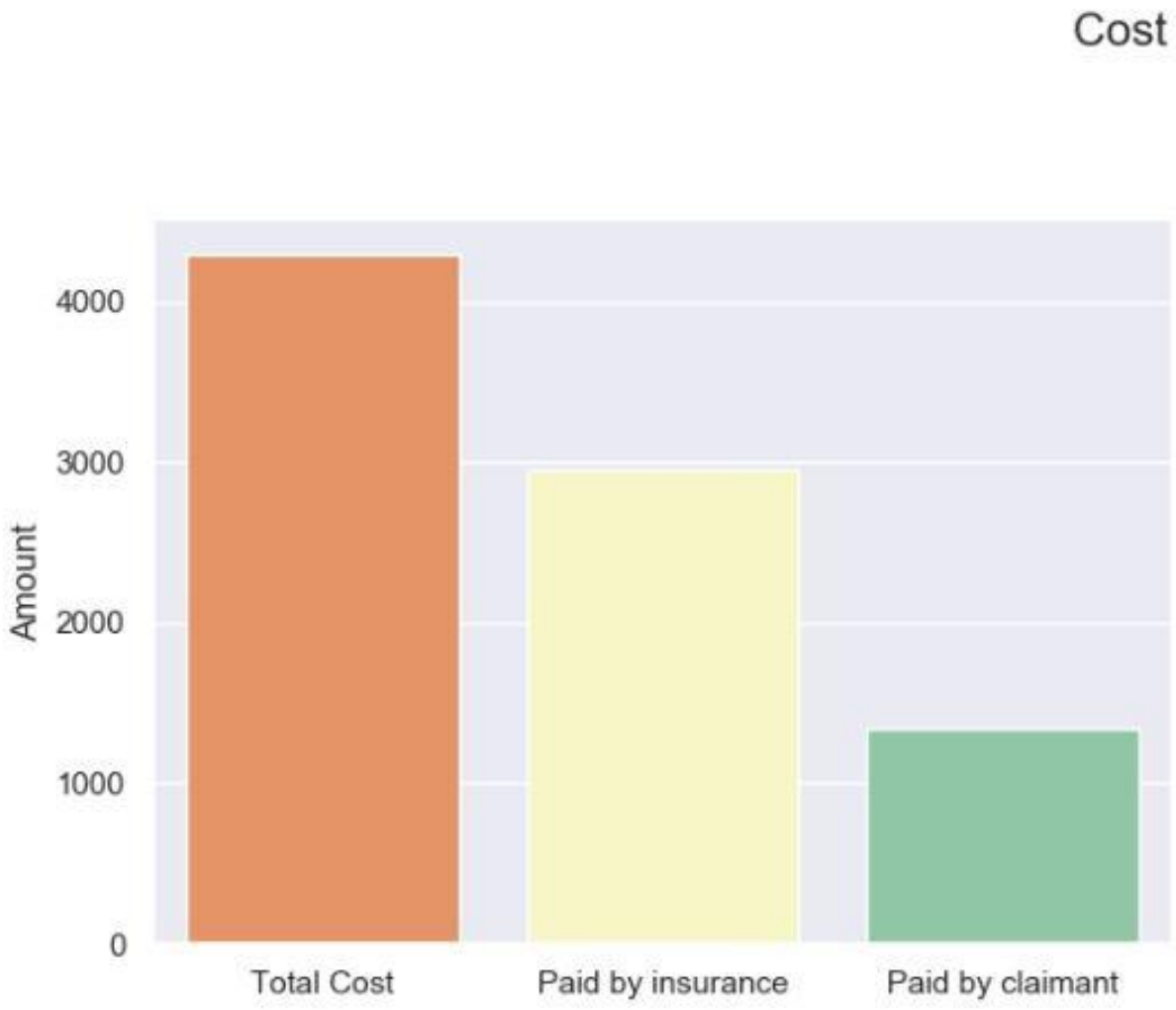
Model statistics for Total cost prediction

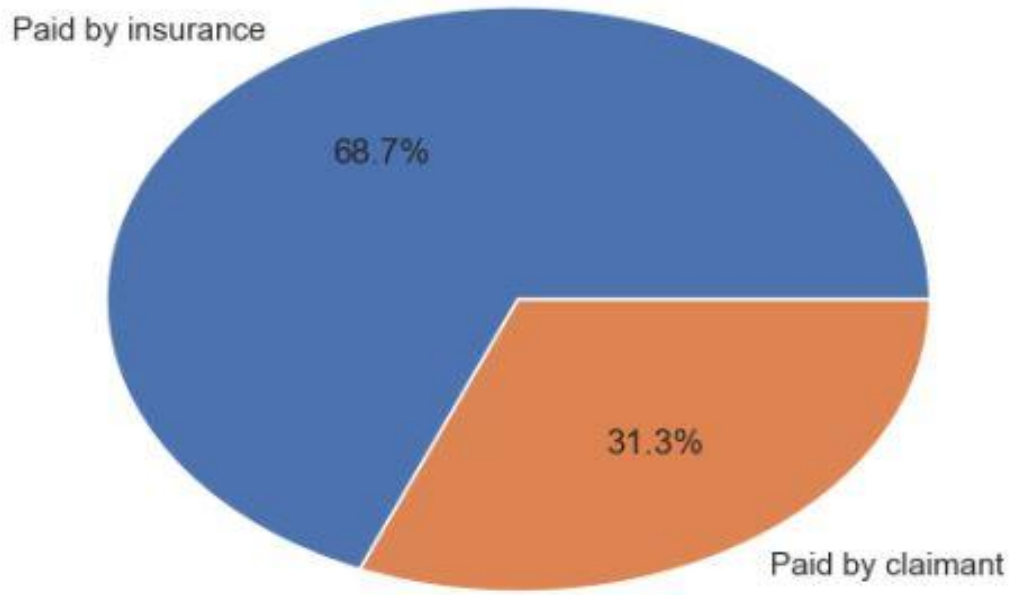
Model	Explained variance Score	Comments
Multilayer Perceptron	0.85	This model does not over fit the data
Extra trees regressor	0.83	This model has a close enough score to MLP, but overfits the data
Ridge regression	0.34	Not reliable.
Lasso	0.37	Not reliable.
Linear regression	0.32	Not reliable

Model statistics for Gap prediction

Model	Explained variance Score	Comments
Multilayer Perceptron	0.89	This model does not over fit the data
Extra trees regressor	0.84	This model has a close enough score to MLP, but overfits the data
Ridge regression	0.37	Not reliable.
Lasso	0.40	Not reliable.
Linear regression	0.32	Not reliable

Sample Output





Challenges

- The data accessed through the APIs was **initially not sufficient** and the model can be **expected to improve** through more training data as is the case for a variety of models.
- 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.
- A more diverse **knowledge of statistical methods** from my side would have aided in selecting the correct features manually.
- Having someone with **medical domain knowledge** would have boosted the validity and confidence of the chosen predictors **based on their inputs,** and insights on otherwise hard to find relations could have been acquired.
- More time to **read research journals** based on the fairly extensive work already done.
- Exposure to **other medical databases** to find common trends.

The Way Forward

- The major challenge now is the seamless integration of this model into patient records.
  - For this , one could use NLP to derive similar feature names from the foreign database.
  - ICD9 Code systems have been updated to ICD10 nowadays, hence a conversion system must be put in place
  - If patient data is segregated into multiple tables (like MIMIC III), then a localized program has to be written to combine them into a singular record.
- A better UI should be made that is directly linked to the patient records where the doctor can simply see the model output aligned with the record itself.
- A more robust model to predict the total cost for multiple diseases.

References

- MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available at: http://www.nature.com/articles/sdata201635
- Pollard, T. J. & Johnson, A. E. W. The MIMIC-III Clinical Database http://dx.doi.org/10.13026/C2XW26 (2016).
- Rao, R., Landi, W., & Rucker, D. (2005). U.S. Patent Application No. 10/812,589.
- Robinson, J. C., Luft, H. S., Gardner, L. B., & Morrison, E. M. (1991). A method for risk-adjusting employer contributions to competing health insurance plans. Inquiry, 107-116.

THANK YOU