Data Analysis Task

Payal Gupta

Problem Statement:

Analyse clinical and financial data of patients and find insights about the drivers of cost of care

Given Dataset:

- Billing Data
  - Bill Id
  - Patient Id
  - Billing amount
- Demographic Data
  - Gender
  - Race
  - Resident Status
  - Date of Birth

- Clinical Data
  - Patient Id
  - Date of Admission
  - Date of Discharge
  - Medical History
  - Preop Medication
  - Symptoms
  - Lab Result
  - Weight
  - Height

Analysis Pipeline:

1. **Data Preprocessing and Feature Engineering:**
    a. Join the multiple datasets to a single meaningful dataset.
    b. Handle the missing data values.
        i. NaN and empty values are replaced by 0 in medical history and pre_op medication columns.
    c. Calculate Age column using DOB data.
    d. Maintain Data consistency.
        i. 'F', 'm' are replaced by Female and Male respectively. Similarly, Singaporean and Singapore Citizen are considered as same. 'Yes' and 'No' are replaced by '1' and '0'. Respectively.
        ii. Replace lower case characters with Upper in 'indian' and 'chinese'.
    e. Convert Date string columns into Datetime format.
    f. Create categorical features for age, race, gender, and resident status.

Analysis Pipeline:

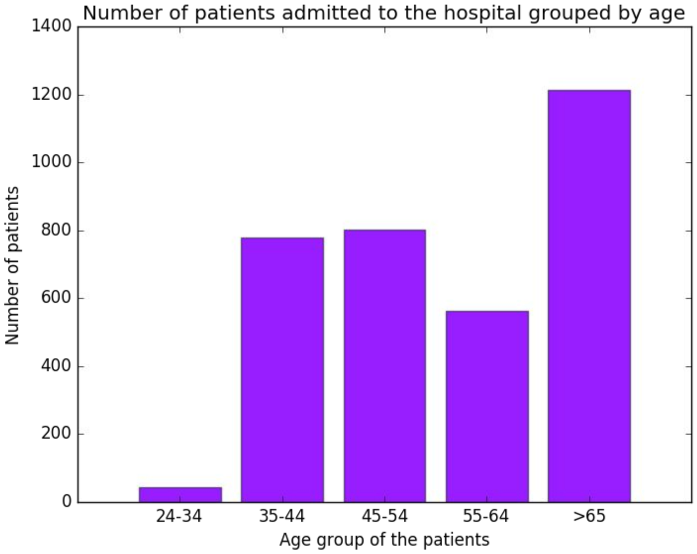
2. **Machine learning model to rank the feature importance**
    a. Random Forest regression model fitted on the Dataset.
    b. Target variable chosen as Bill amount.
    c. R2 evaluation metric for model accuracy.
    d. Train model
    e. Evaluate the most important features for cost regression model.

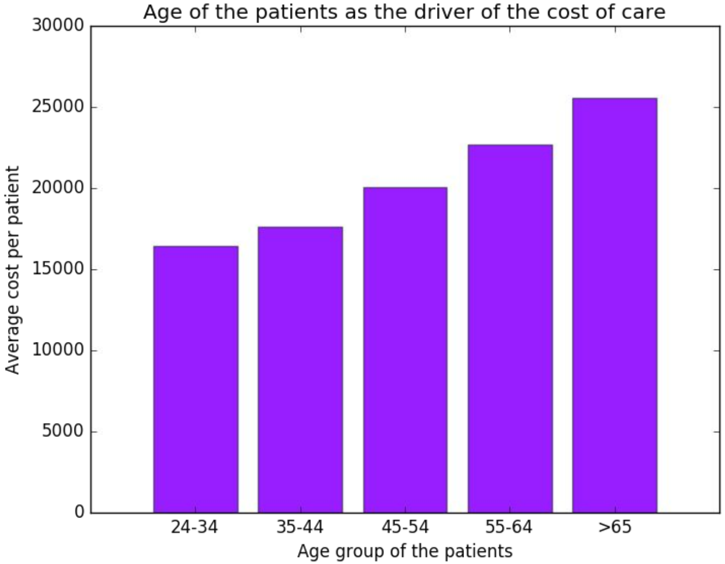
3. **Analyse the important features in the dataset for cost driven features.**

Age as a factor of Cost

- Based on age group data, total number of patients and total cost in each category are grouped and calculated from the dataset. Average cost per patient for each age group is plotted.

- The bar chart shows that the patients in the age group >65 years require maximum health care services and drive the cost of the health care.



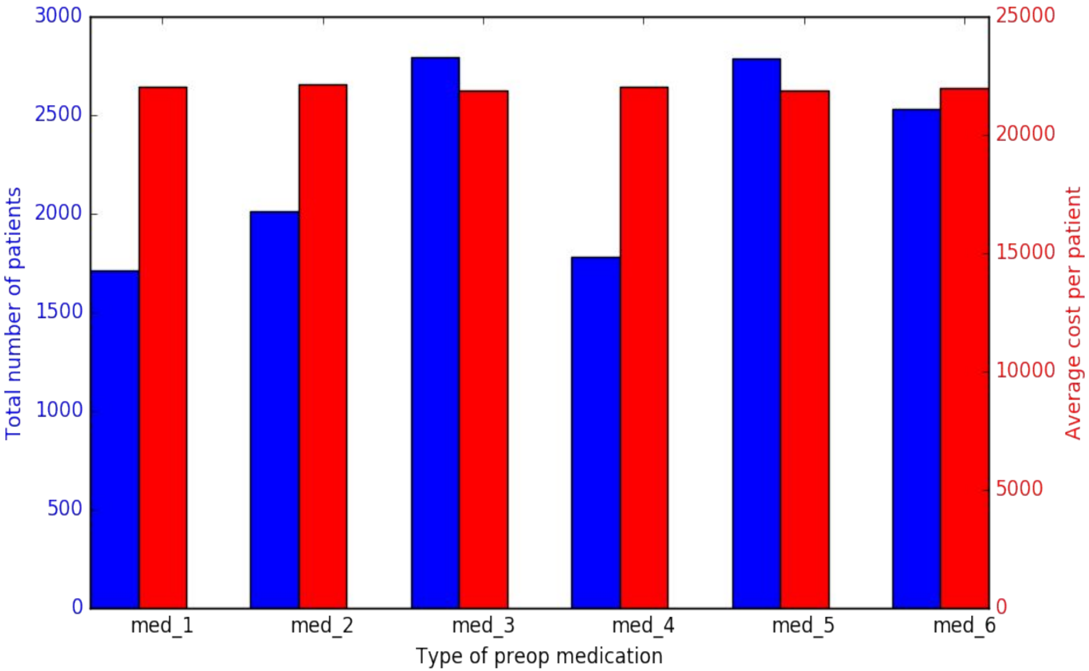


Preop Medication as a factor of Cost

- Based on Preop Medication data, total number of patients for each preop medication category are grouped are calculated from the dataset.

- Average cost per patient for each medication is plotted.

- The bar chart shows that a large number of patients require preop medication 3 and 5 and the average cost of medication per patient is also high. Thus, the above preop medication are one of the factors of cost to health care.

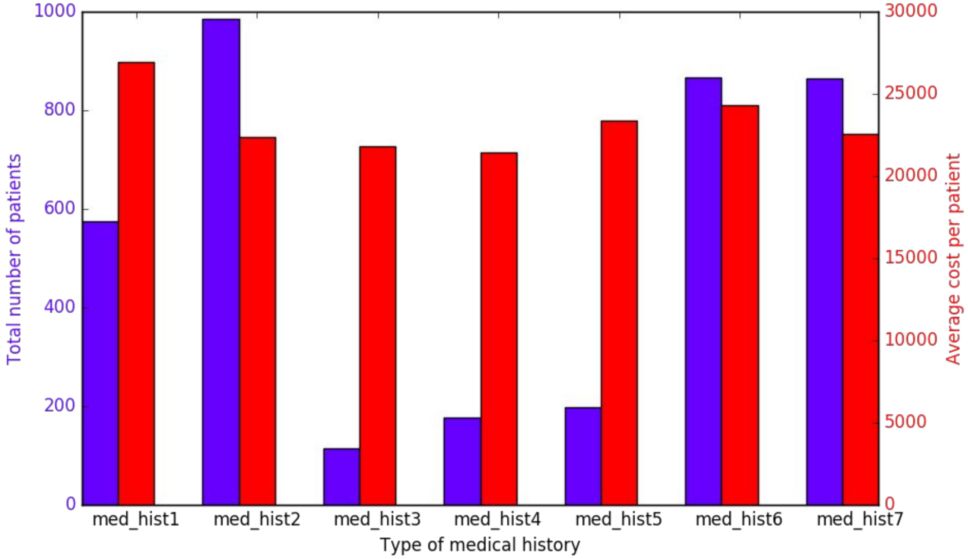


Medical History as a factor of Cost

- Based on medical history data, total number of patients for each medical history are calculated from the dataset.

- Average cost per patient for each type is plotted.

- The bar chart shows that patients with medical history 3, 4, 5 adds the high cost to health care per patient. However, the number of patients with the above medical history are fewer than patients with medical history 2. Hence, medical history 2 add more cost to health care.

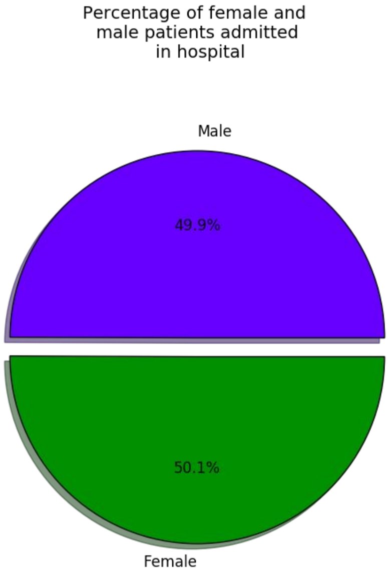


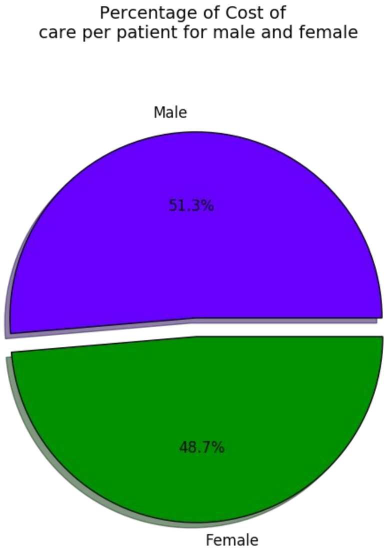
Gender as a factor of Cost

- Based on gender data, total number of patients in each category are grouped and calculated from the dataset.

- Average cost per patient for each gender is plotted.

- The bar chart shows that although the number of male patients are lower than female patients but the cost per male patient is higher than cost of care per female patient.

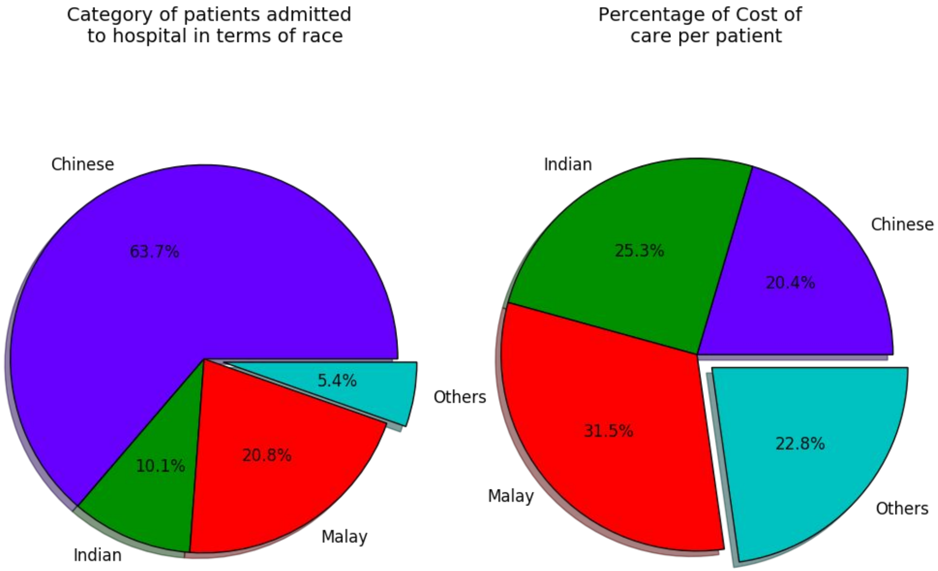

Percentage of female and male patients admitted in hospital


Percentage of Cost of care per patient for male and female

Race as a factor of Cost

- Based on race data, total number of patients and total cost in each category are grouped and calculated from the dataset. Average cost per patient for each category is plotted.

- The bar chart shows that the patients in the ‘others’ category drive a substantial amount to the health care cost even though the number of patients admitted are less compared to other races.



Conclusion:

- Analysis of the features determining the cost of care was performed.
- The important features among all the variables were selected using the regression model built to estimate the bill amount.
- To determine the feature importance, random forest algorithm is chosen with cross-validation split of 10.
- The mean R2 metric value on the given dataset is obtained as 89%.
- The important features in the order of their ranking are:
  - Age group
  - Race
  - Medical History
  - Gender
  - Preop Medication