Hospitalisation Cost Drivers

Rifqi Alkhatib

Holmusk Healthcare Data Challenge

Agenda

1. Problem Statement
2. Data Exploration & Cleaning
3. Feature Engineering
4. Exploratory Data Analysis (EDA)
5. Modelling
6. Conclusion

Problem Statement

Problem Statement

In order to combat the issue of high hospitalisation bills, the Ministry of Health (MOH) wants to understand the drivers of cost of care for patients hospitalised for a certain condition

Data Provided:

Clinical & financial data of patients hospitalised for a certain condition (Jan 2011 – Jan 2016)

Data Exploration & Cleaning

Datasets

- 4 separate datasets
  - Bill Amount
  - Bill ID
  - Clinical Data
  - Demographics
- Clean & merge into 1 dataframe

Datasets – Merging into 1 Dataframe

1. Merge Bill Amount & Bill ID (on bill_id)
    - ‘bill’
2. Left join Demographics onto Clinical Data (on patient_id)
    - ‘patient’
3. Left join ‘patient’ onto ‘bill’ (on patient_id)
    - 13600 rows, 32 variables

Datasets – Merging Bills from Same Admission

- Multiple bills for same patient for each hospitalization
  - Bills from different departments

	patient_id	date_of_admission	bill_amount	date_of_discharge
366	b2d15cda8c4e1f86ba43356434df6718	2011-02-26	2444.80	2011-03-08
367	b2d15cda8c4e1f86ba43356434df6718	2011-02-26	1455.54	2011-03-08
368	b2d15cda8c4e1f86ba43356434df6718	2011-02-26	19943.02	2011-03-08
371	b2d15cda8c4e1f86ba43356434df6718	2011-02-26	1447.26	2011-03-08
1124	b2d15cda8c4e1f86ba43356434df6718	2011-06-02	1045.39	2011-06-08
1126	b2d15cda8c4e1f86ba43356434df6718	2011-06-02	1460.12	2011-06-08
1127	b2d15cda8c4e1f86ba43356434df6718	2011-06-02	1426.59	2011-06-08
1128	b2d15cda8c4e1f86ba43356434df6718	2011-06-02	9087.35	2011-06-08
3972	b2d15cda8c4e1f86ba43356434df6718	2012-06-21	1516.63	2012-06-29
3975	b2d15cda8c4e1f86ba43356434df6718	2012-06-21	1188.14	2012-06-29
3977	b2d15cda8c4e1f86ba43356434df6718	2012-06-21	6502.11	2012-06-29
3979	b2d15cda8c4e1f86ba43356434df6718	2012-06-21	8472.64	2012-06-29

	patient_id	date_of_discharge	bill_amount	date_of_admission
90	b2d15cda8c4e1f86ba43356434df6718	2011-03-08	25290.62	2011-02-26
273	b2d15cda8c4e1f86ba43356434df6718	2011-06-08	13019.45	2011-06-02
986	b2d15cda8c4e1f86ba43356434df6718	2012-06-29	17679.52	2012-06-21

Datasets – Inconsistent Clinical Data

- Clinical Data - Multiple entries for same patient
  - Clinical data inconsistent across admissions

	id	date_of_admission	date_of_discharge	medical_history_1	medical_history_2	medical_history_3
88	b2d15cda8c4e1f86ba43356434df6718	2011-02-26	2011-03-08	0	1	0
273	b2d15cda8c4e1f86ba43356434df6718	2011-06-02	2011-06-08	0	0	1
986	b2d15cda8c4e1f86ba43356434df6718	2012-06-21	2012-06-29	1	0	0

Additional Data Preprocessing

- Dropping duplicate bill amounts
  - Same patient, different admissions / bill_id, identical amount
- Adjusting bill amount for inflation
  - To Jan 2016 Consumer Price Index
- Dropping bill ID
  - No relation to patient clinical & demographic data

Feature Engineering

Feature Engineering

- Length of Hospitalisation (days)
  - Longer hospitalization → higher costs from daily charges
- Patient Age (years)
  - Older patients → more complications
- Body Mass Index, BMI
  - Estimate risk for obesity-related diseases

Feature Engineering

- Extract year & month of hospitalization
  - Study trends year-to-year & month-to-month

- Number of times hospitalised (on admission date)
  - Study effect of repeated hospitalisations on the bill amount
  - Incremental (1 → 2 → 3)

	patient_id	date_of_admission	hosp_no
90	b2d15cda8c4e1f86ba43356434df6718	2011-02-26	1
273	b2d15cda8c4e1f86ba43356434df6718	2011-06-02	2
986	b2d15cda8c4e1f86ba43356434df6718	2012-06-21	3

Feature Engineering

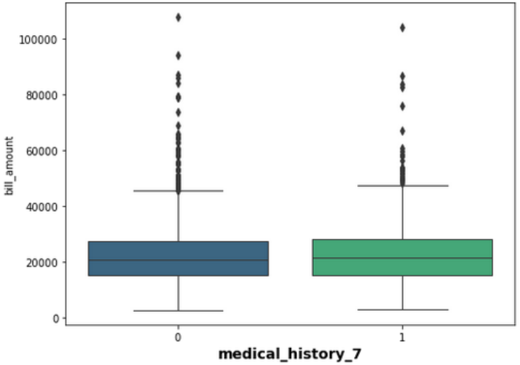
- Summed clinical features
  - Medical History, Preop Medication, Symptoms
  - Study relationship between total occurrences and bill amount
    - More history / meds / symptoms → higher costs

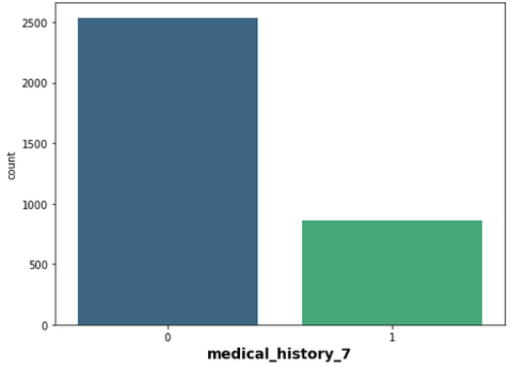
- Initial Feature Elimination
  - Patient ID
  - Date of admission & discharge
  - DOB

EDA

EDA (Categorical) – Medical History

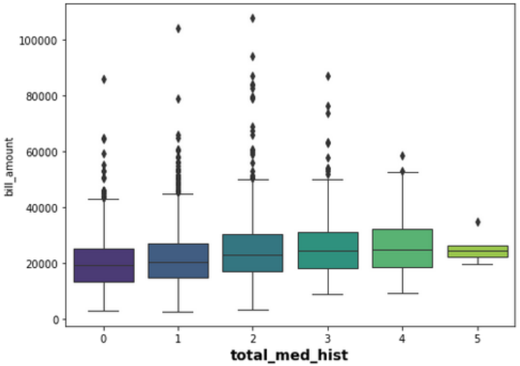
Generally higher bill for patients with each medical history

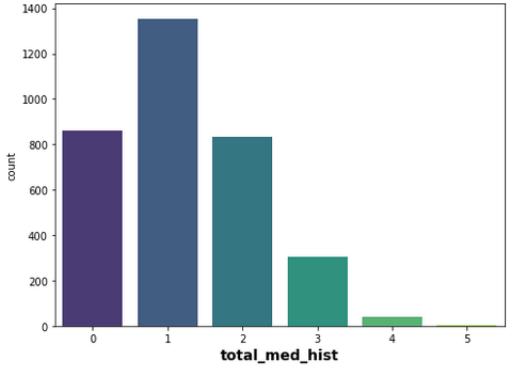




More patients without each medical history

More medical history, higher median bill amount

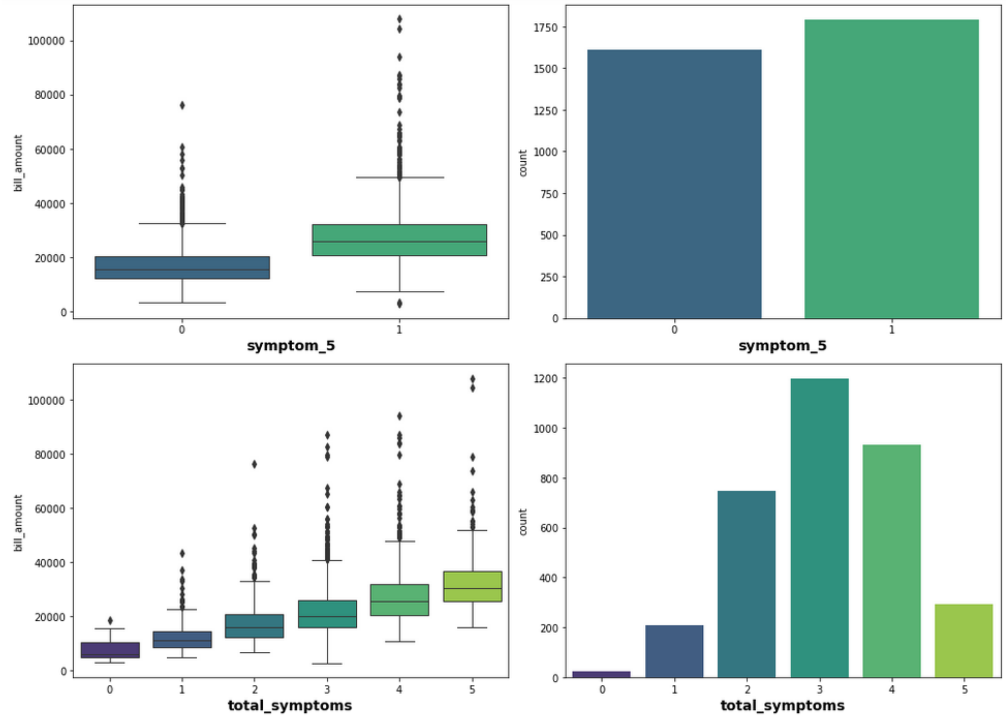




Most patients only have a few medical histories

EDA (Categorical) – Symptoms

Higher median bill for patients with each symptom



More patients with each individual symptom

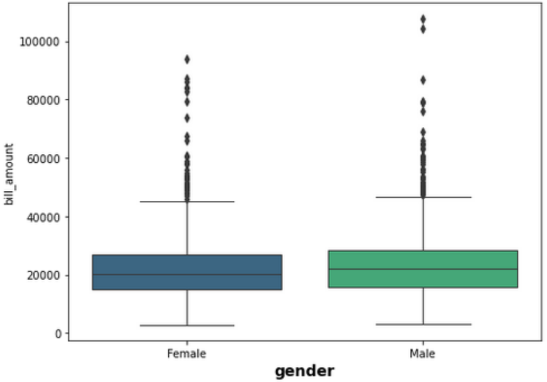
More symptoms, higher median bill amount

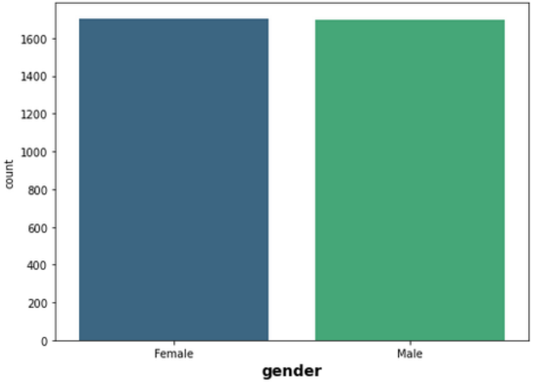
Most patients have at least 2 symptoms

EDA (Categorical) – Demographic Data

Gender

- Even balance of males & females
- Slightly higher median bill amount for males

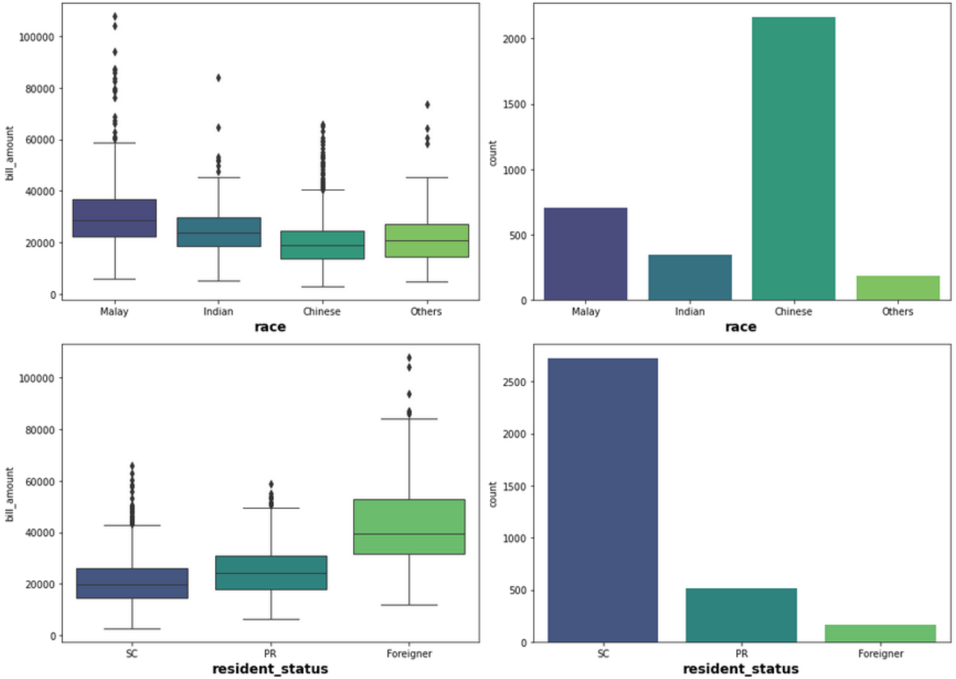




EDA (Categorical) – Demographic Data

Race & Resident Status

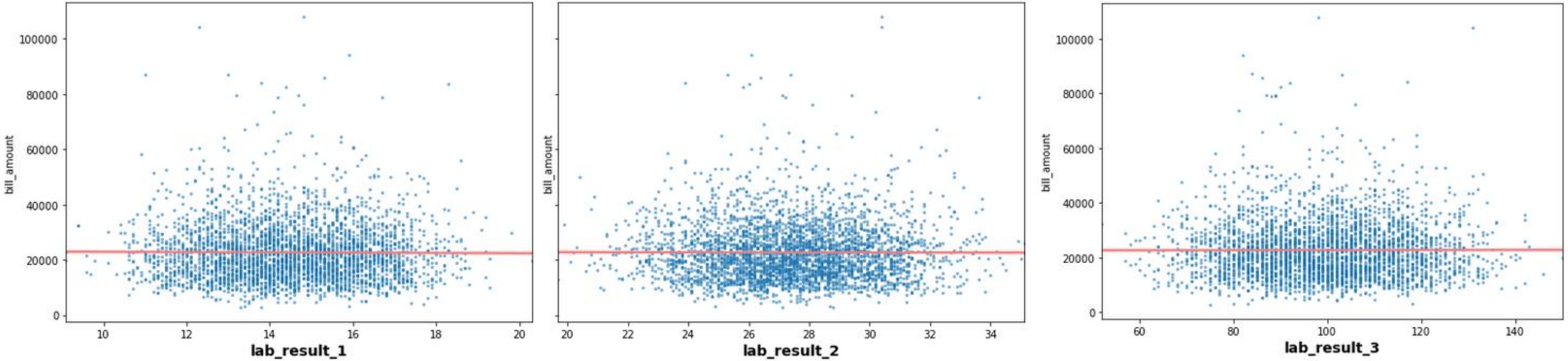
- Minority races overrepresented
- Foreigners underrepresented
- Noticeable differences in median bill amounts



EDA (Categorical) – Dropped Features

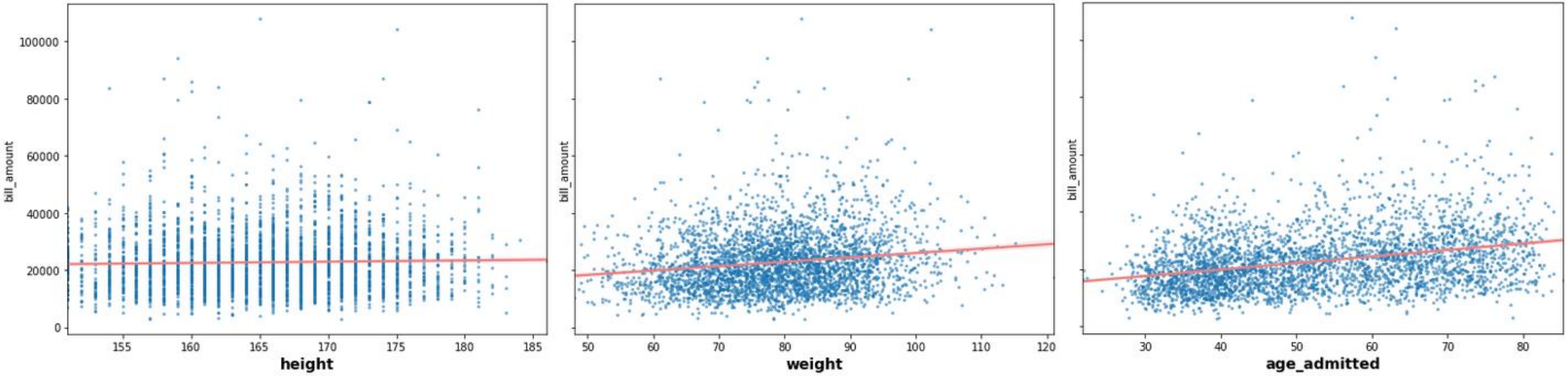
- No clear trend
  - Preop Medications (Individual & Total)
  - Days in Hospital
  - Hospitalisation Number
  - Month Admitted
- Not useful
  - Year Admitted

EDA (Continuous) – Lab Results



- No clear relationship
- Patients charged for lab test regardless of result
- Dropped

EDA (Continuous) – Physical Features

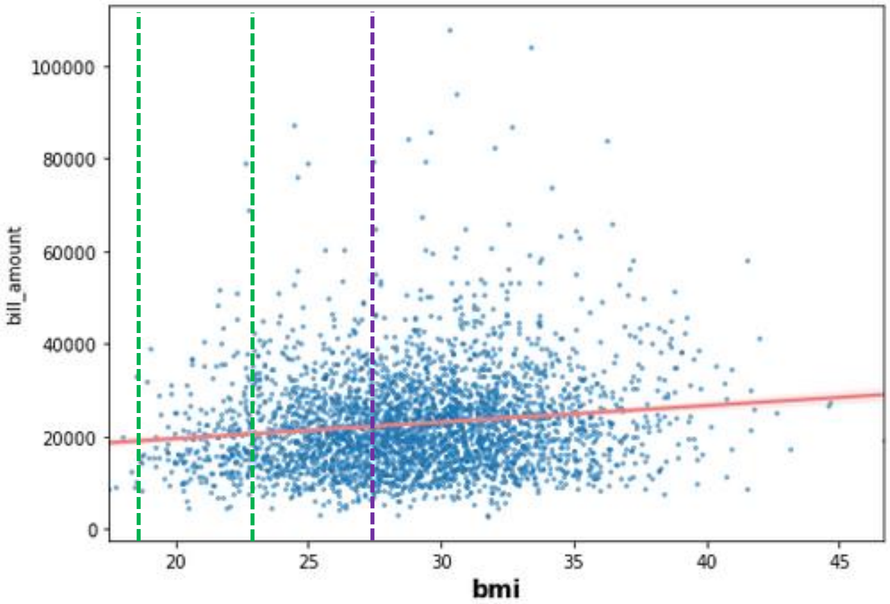


- No clear relationship for height – dropped
- Slight positive correlation for weight & age

EDA (Continuous) – Physical Features

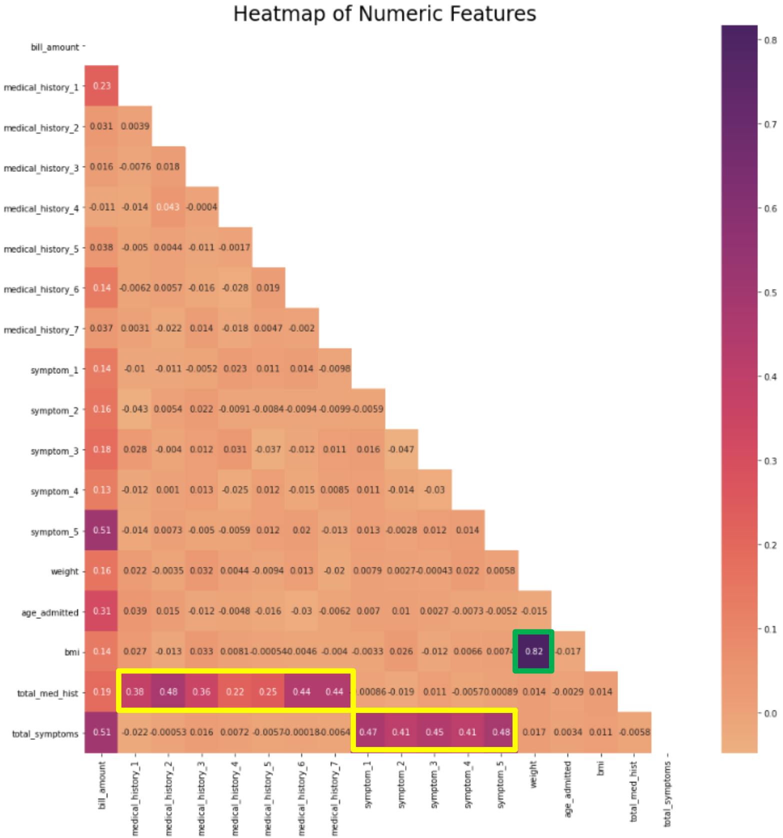
BMI

- Singapore healthy range: 18.5 – 22.9 kg / m^2
- Most patients overweight
  - Many at high risk
- Very few underweight
- Slight positive correlation



EDA – Heatmap of Numeric Features

- Look out for multicollinearity
- High correlation between weight and BMI
  - Dropped BMI
- Moderate correlation between total_med_hist, total_symptoms and their components
  - Keep



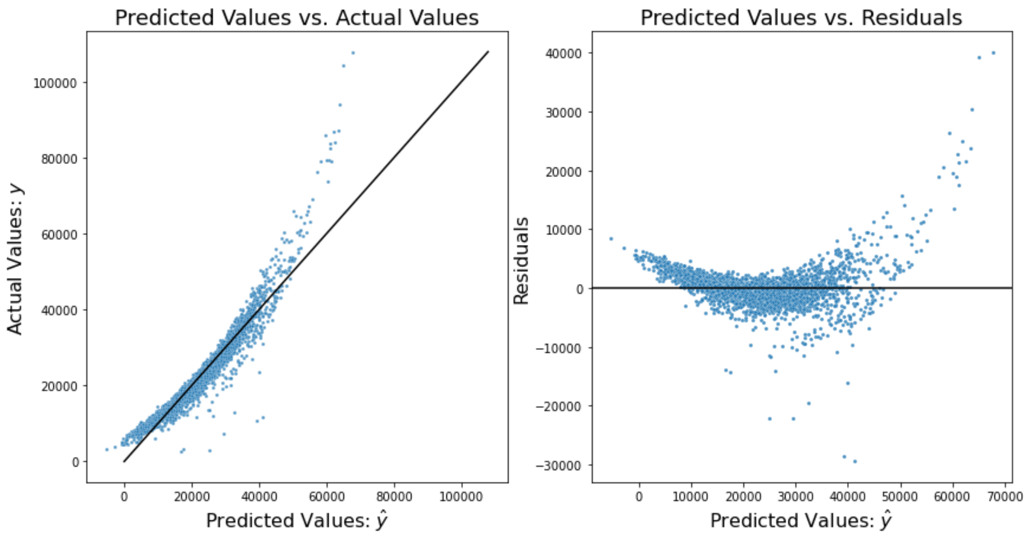
Final Feature Set

- 19 independent variables
  - Medical History (1 – 7 & total)
  - Symptom (1 – 5 & total)
  - Weight
  - Gender
  - Race
  - Resident Status
  - Age

Modelling

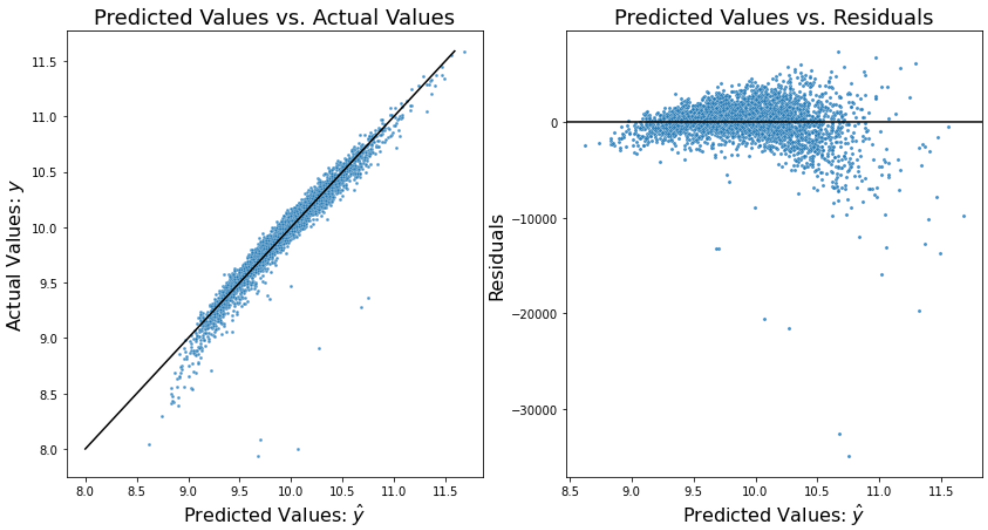
Initial Model

- Fit in all features with bill amount as y
- Clear non-linearity of predictions
- Heteroscedasticity of residuals
- RMSE: 3180.9



Model with Log Transformed Target Variable

- Fit in all features with log(bill amount) as y
- Improved linearity of predictions
- Reduced heteroscedasticity of residuals
- RMSE: 2236.6



Model Analysis

- R-squared & Adj. R-squared = 0.941
  - Model able to explain 94.1% of changes in target variable
  - Almost all variables are contributing properly
- Prob (F-statistic) = 0.0
  - At least one independent variable has significant effect

- Equation for MLR model:

$$\log(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

  - 1 unit increase in X_1 → β_1 increase in log(y)

Model Analysis – Coefficients

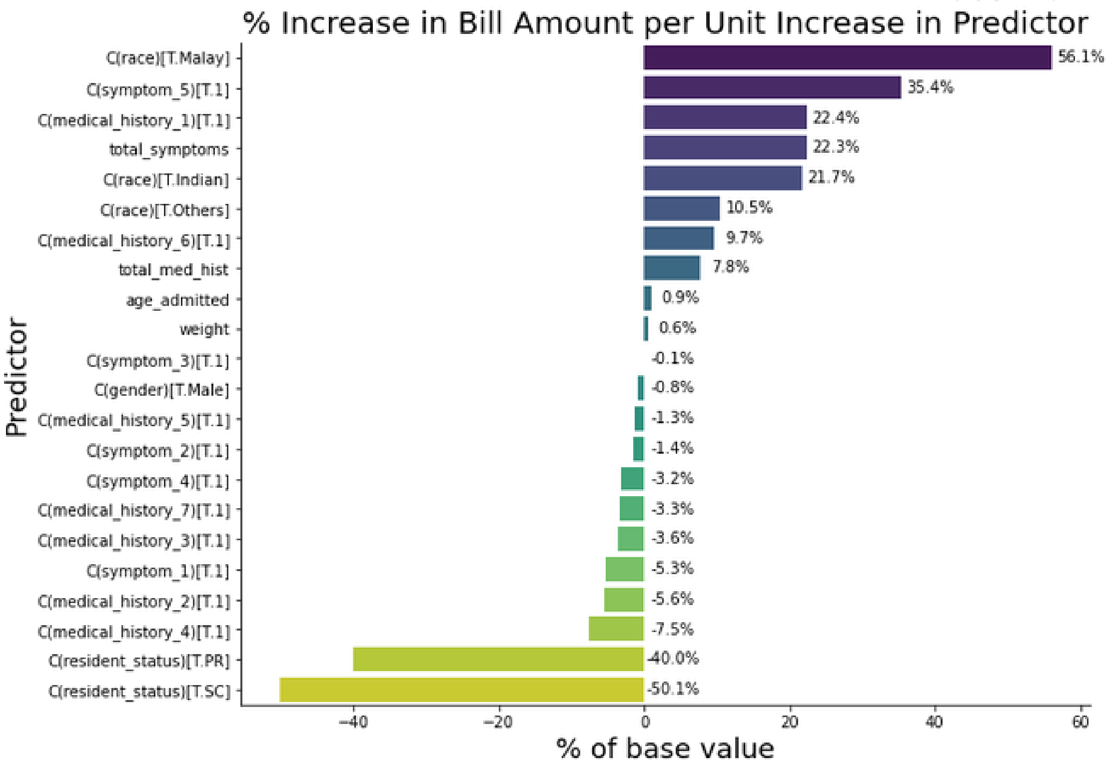
- log(bill amount) does not make sense
  - Exponentiate

$$y = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}$$

  - 1 unit increase in X_1 → e^{β_1} times increase in y compared to base value
- Base value
  - Bill amount when all other coefficients set to 0 → $y = e^{\beta_0}$
    - Female, Chinese, Foreigner
    - No medical history & symptoms
    - Hypothetical weight & age = 0
  - Base Bill Amount = $5607.82

Model Analysis – Coefficients

- Race an important feature
- Certain symptoms & medical histories have greater impact
- Resident status also important
- Gender, age & weight not very important
- Total medical histories & symptoms have greater impact



Conclusion

Recommendations

- Conduct further studies into race-specific differences
  - Results indicate race plays a huge role in patient’s cost of care
  - Studies to identify underlying causes
  - Develop targeted measures to equalise cost of care
- Target symptom_5, medical_history_1 & medical_history_6 for early intervention
  - Studies show that early intervention and prevention highly effective at saving costs
  - Mass media campaigns targeting these 3 features
  - Too late once hospitalised

Limitations

- Ambiguity of bills
  - Multiple bills per hospitalisation
  - Nett or gross amounts
    - Subsidies, insurance, etc
- Lack of context
  - Clinical features difficult to understand without knowing how data is collected
  - Inconsistencies in data
- Addressing anonymity
  - Inevitable in healthcare
  - More domain knowledge
    - Enables formulation of more reasonable assumptions

Thank You