# Drivers of cost in healthcare

Objectives

- The task is to analyze the clinical and financial data of patients hospitalized for a certain condition
- Some variable names and patient_id's have been anonymized in this dataset.
- You are required to join the data given in different tables, and find insights about the drivers of cost of care.

## WORKFLOW

**General sequence of events**

1) Merging datasets

2) Cleaning merged datasets

3) Feature Engineering 1

4) Data Visualisation

5) Fitting Random Forest

6) Feature Engineering 2

7) Feature Engineering for Elastic Net Regression

8) Elastic Net Regression

**Details of each segment**

Pre-processing (Merging and cleaning dataset)

- Set datatype for each column
- Remove duplicates for each dataset (Could interfere with merging)
- Merge datasets (bill_amount, bill_ids, clinical_data, demographics)
- Check if each column has consistent data format, and levels are consistent (for categorical variables). If wrong, rectify (e.g. 4 genders because of different spelling)
- Null values for datasets before and after merging
    - Fill in missing values with appropriate values
- Check for conflicting variables e.g. **date of discharge** *before* **date of admission**
- Identifying any patients not on ops medication (Not a problem)
- Generate brief description of categorical and numerical variables

Feature Engineering 1

- Creation of variables which can be derived from dataset such as: **age at admission**, **days admitted**, **BMI**, **number of drugs** etc
- Sum bills of each patient per admission date due to 4 bills per patient per admission
- Count number of admissions per patient

Data visualisation

- Split into categorical and continuous variables
- Univariate analysis: Distribution plots (Continuous variables) + Count plots (Categorical variables)
- Bivariate plots: Violin plots (For categorical against continuous) & scatter plots (For continuous against continuous)
    - Roughly gauge if there are any
- Correlation plots for continuous variables

Random Forest (RF)

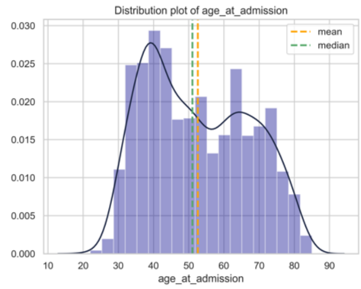
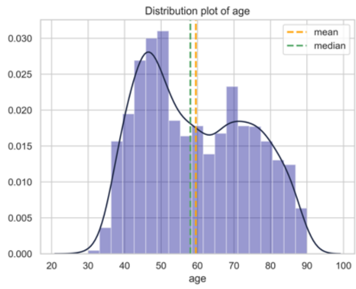
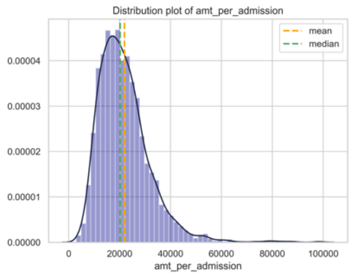
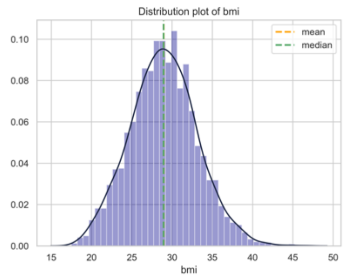
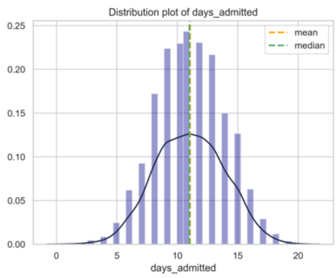
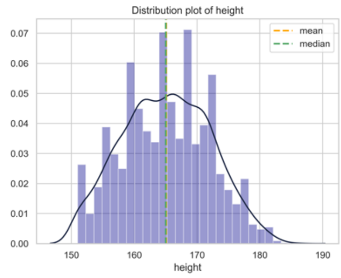
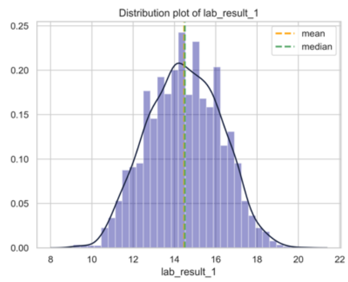
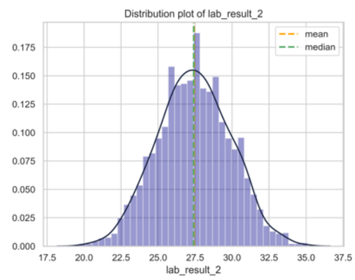
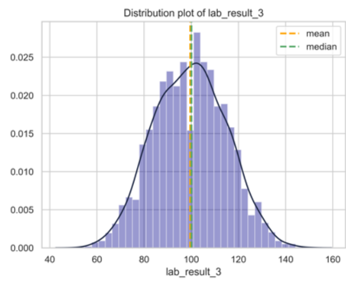
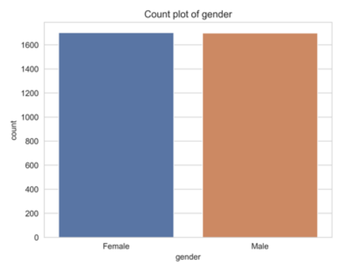
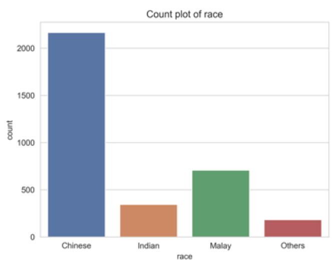
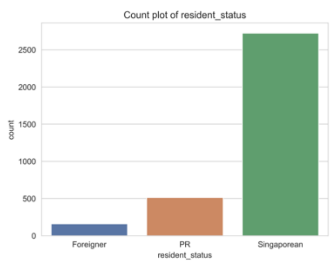
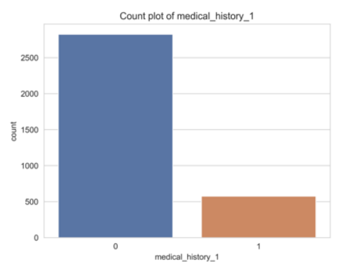
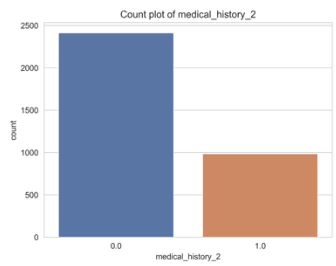
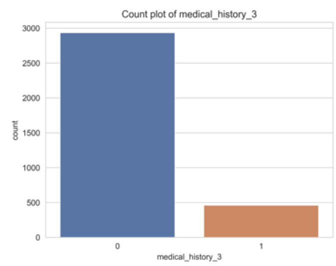
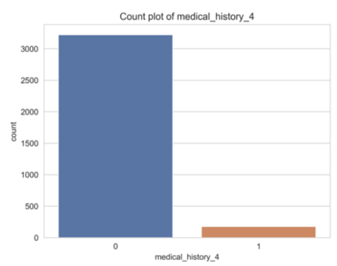
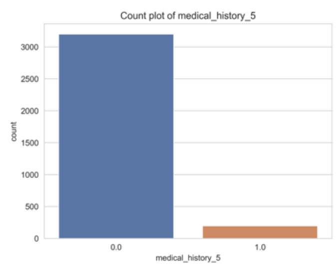
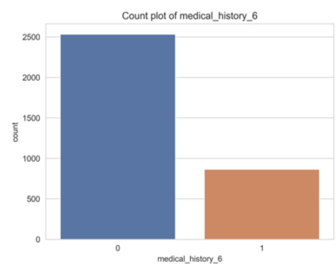
- Creation of dummy variables for multi-level categorical variables. Will be used for all models hereafter.
- Split into training set + validation set (80%) + test set (20%)
    - Set seed 2604. This seed will be used hereafter.
    - Training set (n = 2176) + Validation set (n = 544) [5 K-Folds to be used, with shuffling]
    - Test set (n=680)
- Random Forest (RF)
    - Make Root Mean Squared Log Error (RMSLE) scorer for future algorithms
        - Because want to penalise underpredicting cost of care heavier than over predicting.
    - Use GridSearch CV + parameter grid to identify good parameters for full set of data
        - {'max_features': ['auto', 'log2', 'sqrt'], 'n_estimators': [100,200,300,400, 500,700,900], 'max_depth': [1,5,10,15,20], 'min_samples_leaf': [1,5,10,15] }
        - Best parameters: {'max_features': ['auto'], 'n_estimators': [700], 'max_depth': [20], 'min_samples_leaf': [1]}
    - Metrics for model success: Average cross-validation RMSLE and out-of-bag (OOB) score ($R^2$).
    - Pseudo-backwards elimination to remove variables which do not contribute to the model to create a parsimonious model (also helps with interpretation of model). Even better if it helps to improve metrics, which did happen.
    - Feature Engineering 2 occurs here as well. Features such as number of diseases and pre-operation medicine is added here.
    - Add to best set of variables, and filter out redundant features again.
    - Identify predictive power of best model (using $R^2$) by fitting best trained model to final test set.

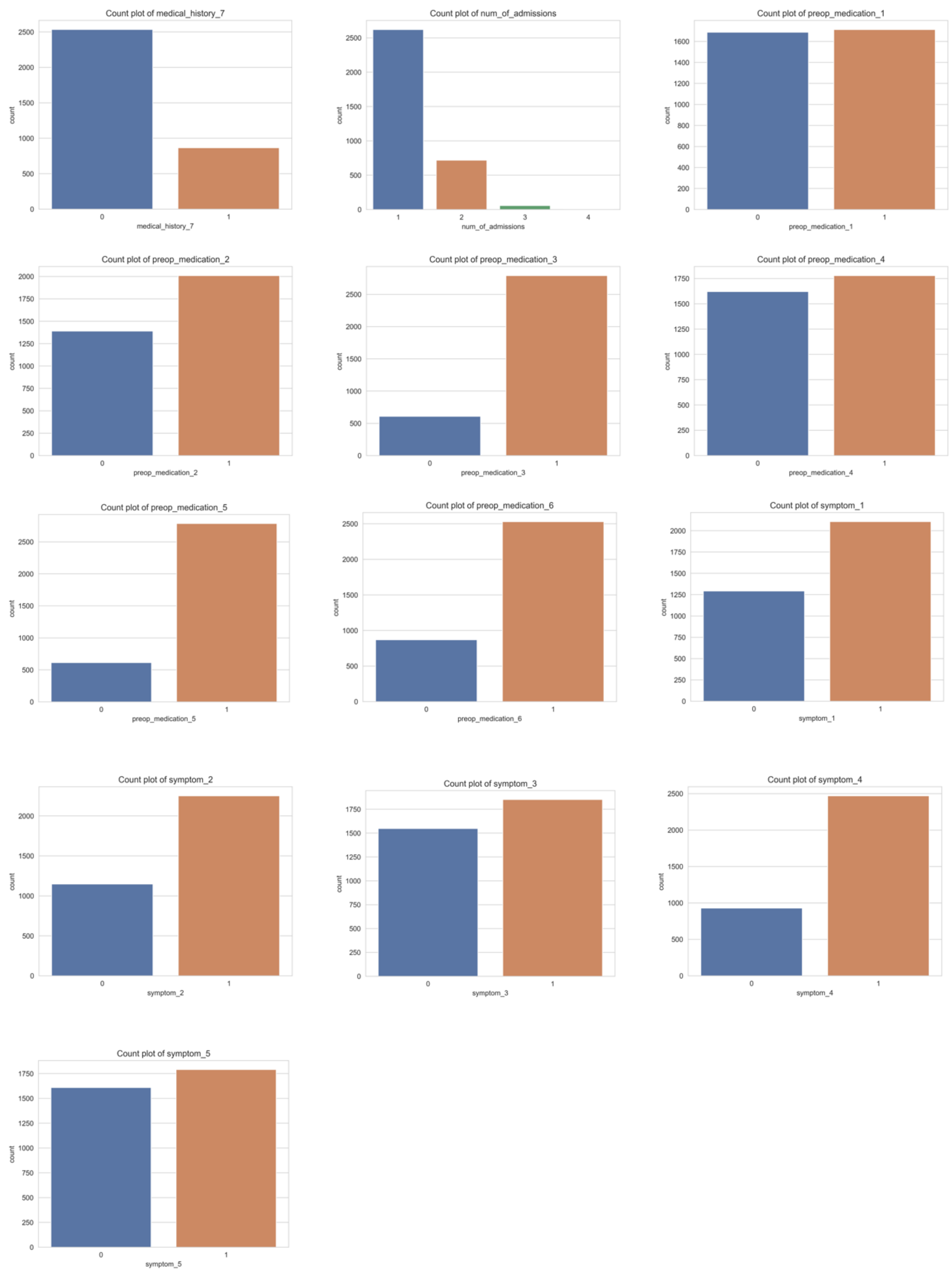
    - Best variables to train RF: **Malay**, **Chinese**, **Foreigner**, **PR**, **Age at admission**, **Weight**, **Lab result 2**, **Symptom 5**, **Medical History 1 & 6**, **Number of Symptoms** and **Number of Diseases**
- Elastic Net
    - Feature engineering 3: Generated several transformed variables such as log, square-root and standardisation of continuous variables to fit linear regression assumptions
        - Log-transform total **cost**
        - Square-root transform **weight**
        - Standardisation was performed for training and test set each. Transformed **age at admission** and **lab result 2**
    - Use GridSearch CV + parameter grid to identify good parameters for full set of data
        - {'alpha': [0.0005, 0.1, 0.5, 0.9, 0.95, 0.99, 1], 'l1_ratio':[0.0, 0.1, 0.2, 0.3...,1]}
        - Best parameters: Alpha = 0.0005, L1 ratio = 0.0
    - Identified cross-validation RMSLE, train set RMSLE and test set RMSLE
    - First performed on best set of variables selected by RF (Vanilla). After which, replace the original variables with the transformed variables.
    - Compare model performance to Random Forest (Control)
    - Further reduction of model size as much as possible (Removed **Lab-result 2** because removal did not change test score)

## Interpretation by segment of ML pipeline

Data visualisation (All results shown below)

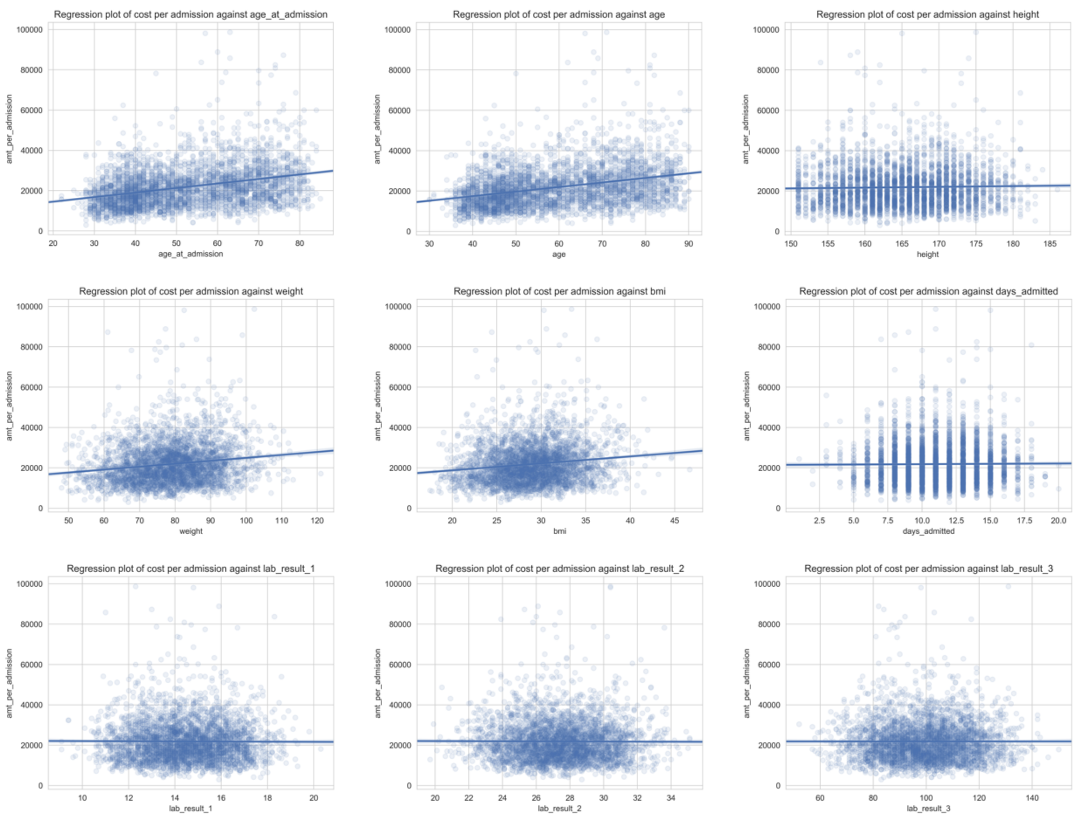
- Univariate analysis
    - Continuous variables
        - Most continuous variables are approximately normally distributed, with the exception of cost, which is quite right skewed. **BMI** was more normally distributed than either height or weight individually. **Age** is not normally distributed.
    - Categorical variables
        - **Gender** appears to be completely even. This is a **Singaporean** and **Chinese**-dominated sample.
        - **Pre-op medication 1** appears to be completely even. Most patients take **pre-op medication 3,5 and 6**. Most patients show **symptom 4**.

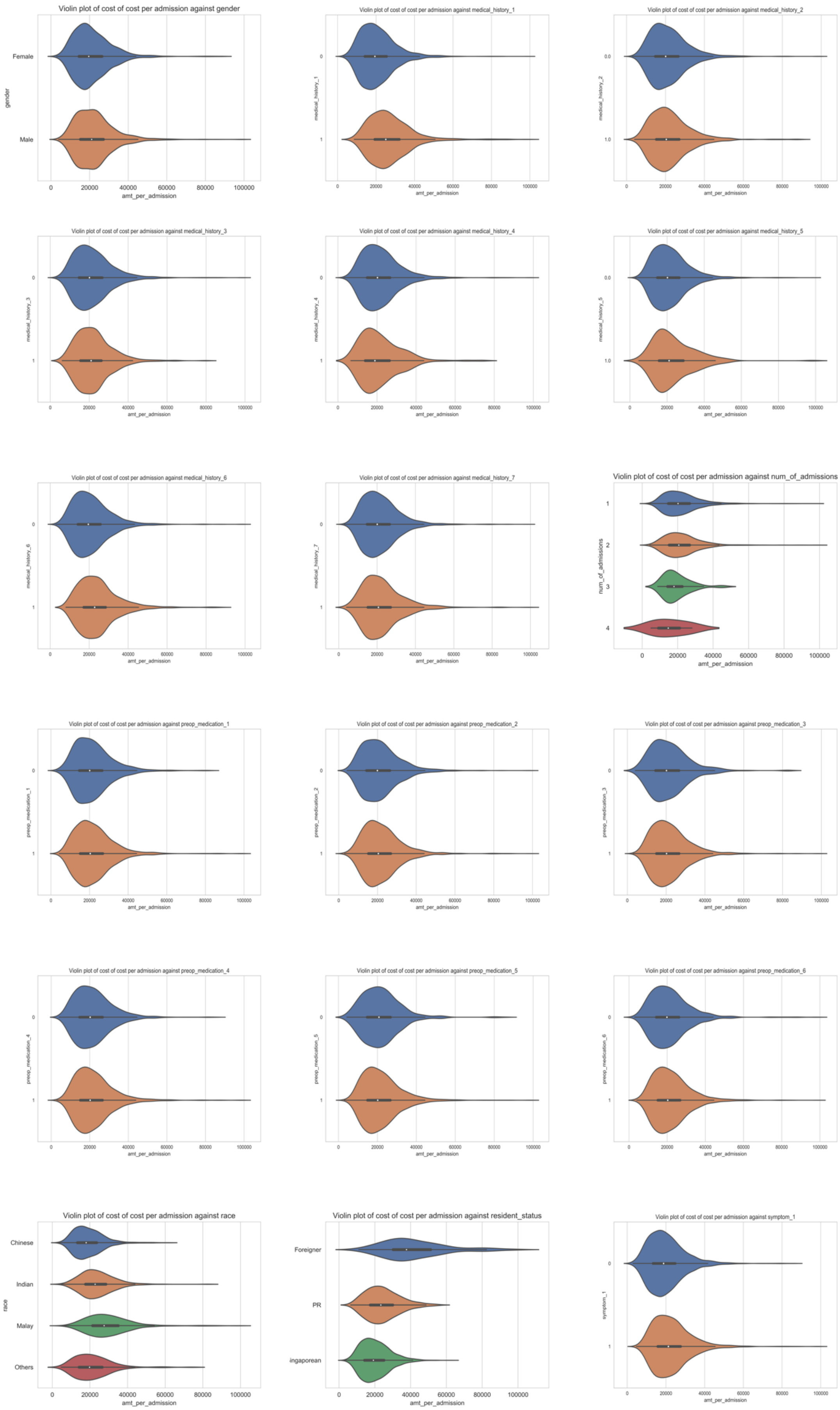

Distribution plot of age_at_admission
mean
median
age_at_admission

Distribution plot of age
mean
median
age

Distribution plot of amt_per_admission
mean
median
amt_per_admission

Distribution plot of bmi
mean
median
bmi

Distribution plot of days_admitted
mean
median
days_admitted

Distribution plot of height
mean
median
height

Distribution plot of lab_result_1
mean
median
lab_result_1

Distribution plot of lab_result_2
mean
median
lab_result_2

Distribution plot of lab_result_3
mean
median
lab_result_3

Count plot of gender
count
Female
Male
gender

Count plot of race
count
Chinese
Indian
Malay
Others
race

Count plot of resident_status
count
Foreigner
PR
Singaporean
resident_status

Count plot of medical_history_1
count
medical_history_1

Count plot of medical_history_2
count
medical_history_2

Count plot of medical_history_3
count
medical_history_3

Count plot of medical_history_4
count
medical_history_4

Count plot of medical_history_5
count
medical_history_5

Count plot of medical_history_6
count
medical_history_6

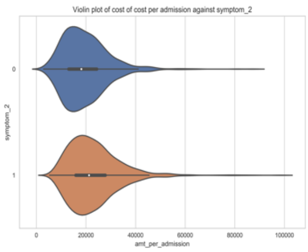


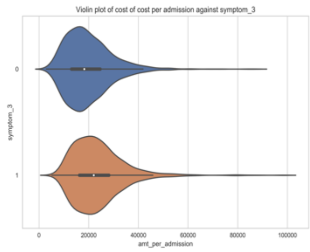

- Bivariate analysis
  - Categorical vs Total Cost
    - Males seem have higher costs than females. For race, **Chinese < Others < Malay < Indians** in terms of costs. **Foreigners** pay

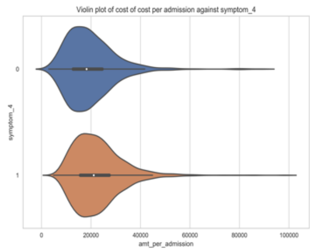
    the most, while **PR** pays less, but still pays more than Singaporeans, which makes sense.
  - **Medical history 1 and 6** seems to have higher costs due to the higher median are having a larger distribution having higher costs. **Number of admissions** doesn't say anything useful. **Pre-op medication** doesn't indicate much. Having symptoms generally indicate slightly higher cost except **symptom 5** which shows substantially increased cost.
- Continuous vs Total Cost
  - **Age, age at admission and BMI** shows clearest trend of linear increase. **Weight** shows steeper trendline than height, so using **height** might dilute the effect of weight in **BMI**
  - Interestingly, **number of days** admitted does not contribute at all. All of the lab test results don't show anything.
- Correlation heatmap
  - **Age at admission** correlates to **age**, as expected. Similar to BMI, height and weight. No correlation between other variables. Seems like linear models should do fine as long as correlated variables are removed.

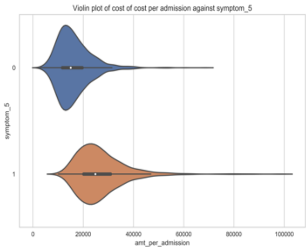


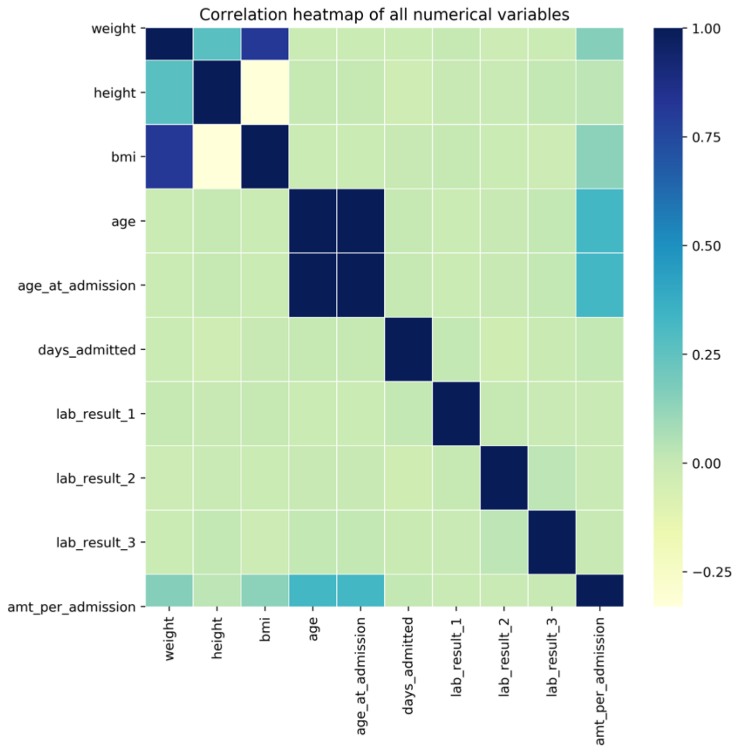

Violin plot of cost of cost per admission against gender
Female
Male
gender
amt_per_admission
Violin plot of cost of cost per admission against medical_history_1
Violin plot of cost of cost per admission against medical_history_2
Violin plot of cost of cost per admission against medical_history_3
Violin plot of cost of cost per admission against medical_history_4
Violin plot of cost of cost per admission against medical_history_5
Violin plot of cost of cost per admission against medical_history_6
Violin plot of cost of cost per admission against medical_history_7
Violin plot of cost of cost per admission against num_of_admissions
num_of_admissions
Violin plot of cost of cost per admission against preop_medication_1
Violin plot of cost of cost per admission against preop_medication_2
Violin plot of cost of cost per admission against preop_medication_3
Violin plot of cost of cost per admission against preop_medication_4
Violin plot of cost of cost per admission against preop_medication_5
Violin plot of cost of cost per admission against preop_medication_6
Violin plot of cost of cost per admission against race
Chinese
Indian
Malay
Others
race
Violin plot of cost of cost per admission against resident_status
Foreigner
PR
ingaporean
Violin plot of cost of cost per admission against symptom_1
symptom_1



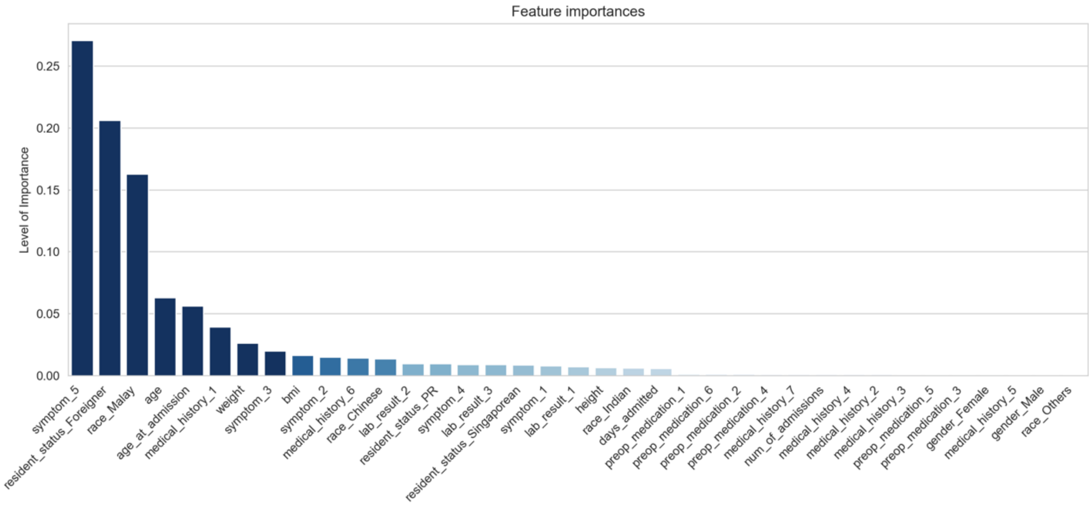




Random Forest

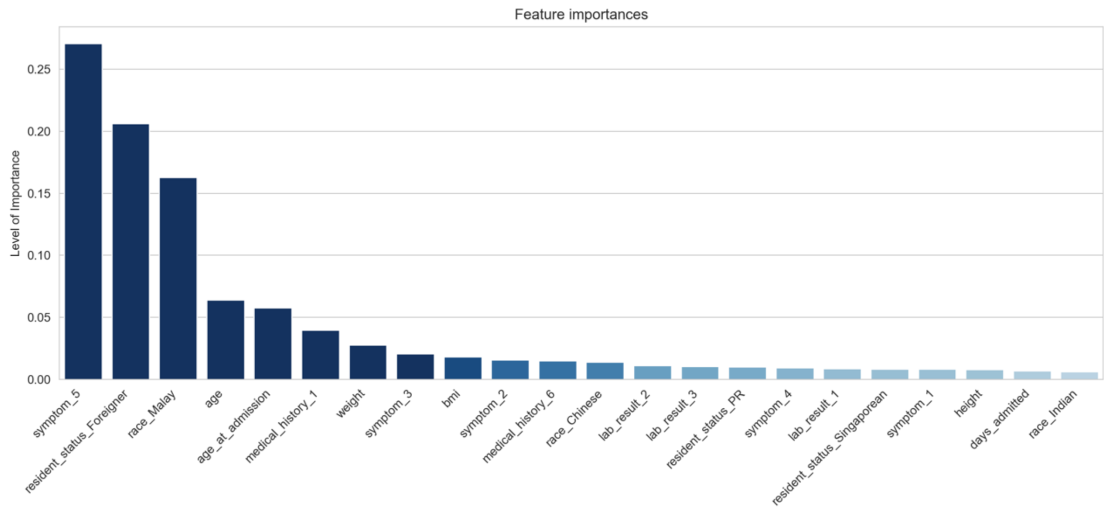
- Random forest quite suitable for predicting cost of care based on the various demographics and clinical data
- Only 12 variables were selected and useful in predicting cost of care. **Symptom 5** has the most importance, followed by being a **Foreigner** and **Malay**. Being of a **Chinese race** is also important in driving cost, but not as

much as being **Malay**. **Indians** and **Others** were removed from the model. Perhaps it could be a disease which tends to appear in **Malays**. Being a **PR** is also relatively informative. **Age at admission** was found to be more important than **age in 2020**.

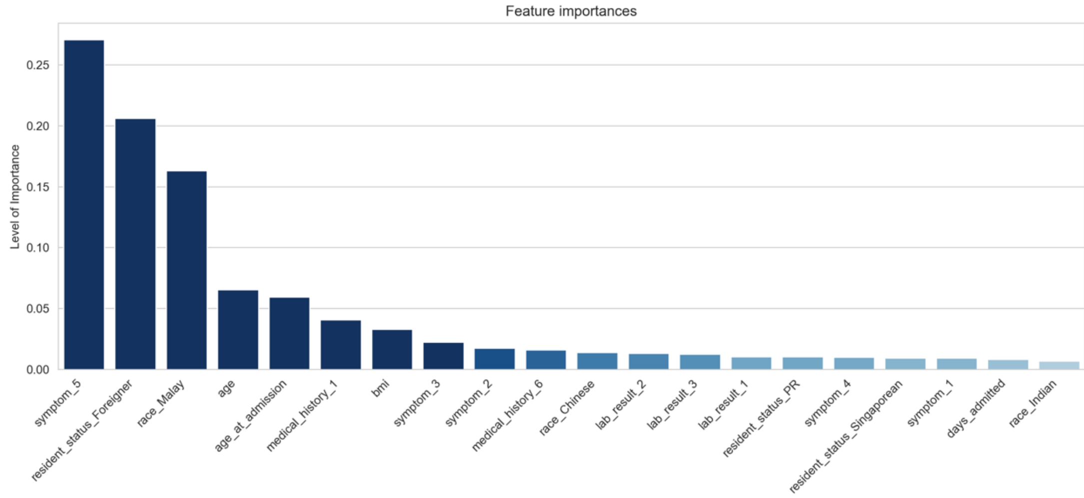
- **Weight** was found to be more important than **height and BMI**. **Medical history 1 and 6** are important. **Lab results 2** is substantial. Lastly, the only generated variables which were important was found to be number of diseases and symptoms, although they are at the end of the list.
- The below are feature importance plots using sklearn. The higher the level of importance, the more it helps with the prediction of total cost of care.



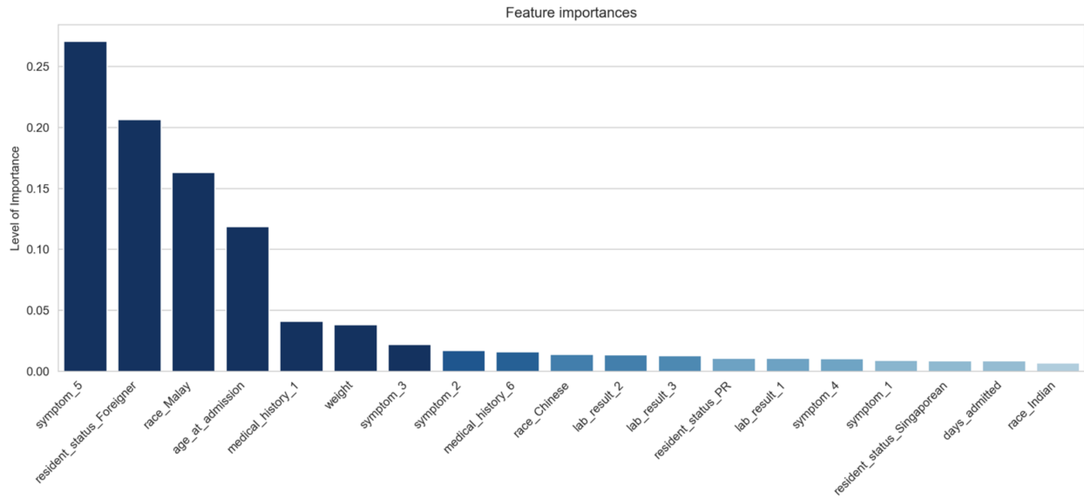

*Included all variables. Cross-validation RMSLE = 0.139, OOB Score = 0.898



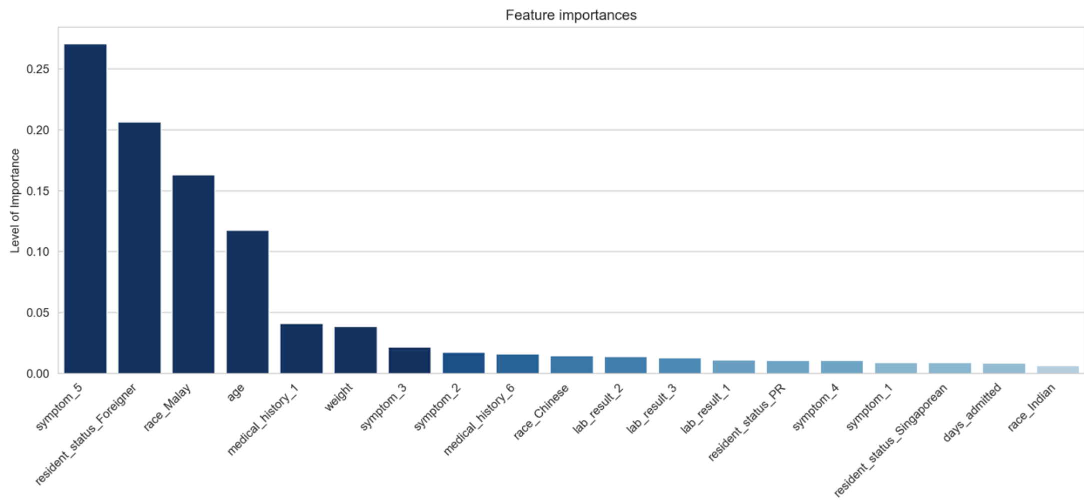

*After removing variables close to 0. Cross-validation RMSLE = 0.137, OOB Score = 0.902



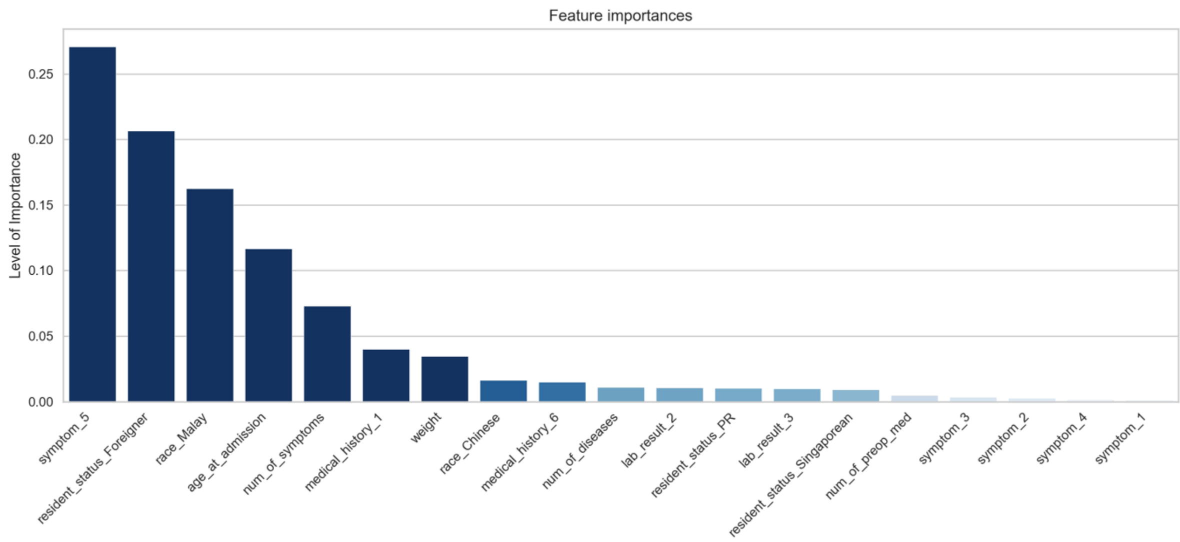

*Removed height, and weight to see if bmi can represent both. Cross-validation RMSLE = 0.135, OOB Score = 0.904



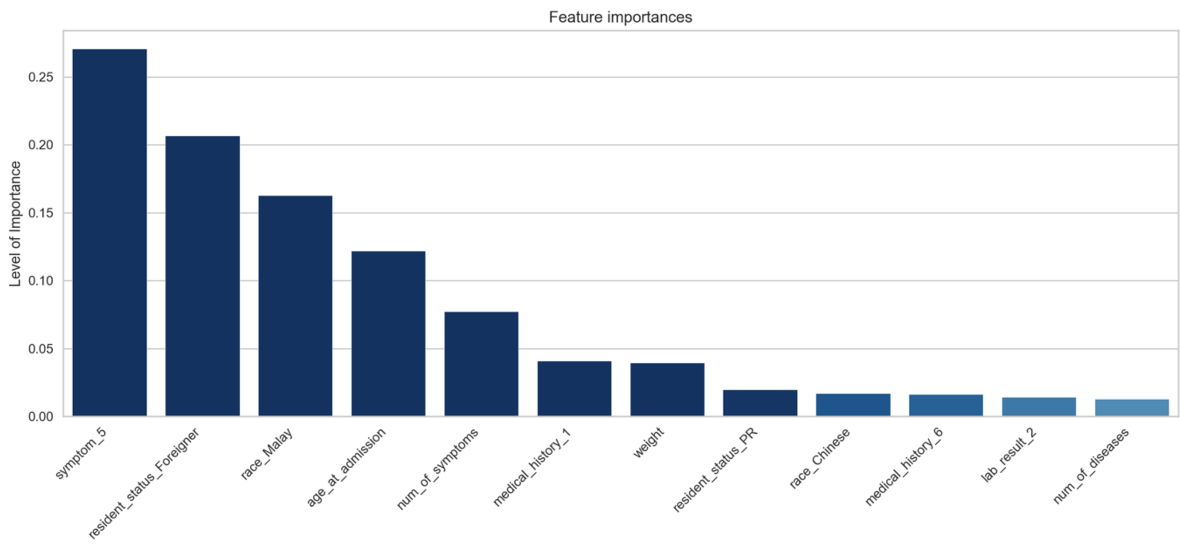

*Used weight instead of bmi and height to see if it's better than bmi. Cross-validation RMSLE = 0.133, OOB Score = 0.906



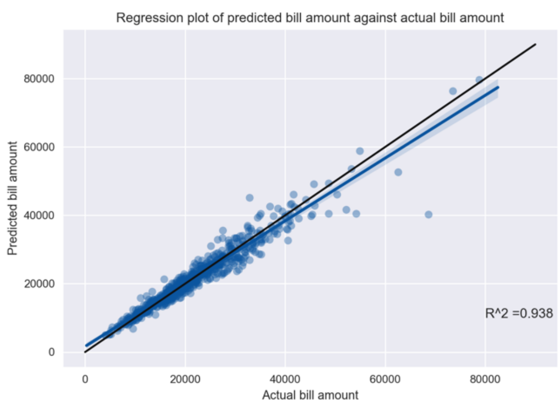

*Used age instead of age at admission to see which is better. Cross-validation RMSLE = 0.132, OOB Score = 0.906




*Included engineered features. Cross-validation RMSLE = 0.126, OOB Score = 0.914



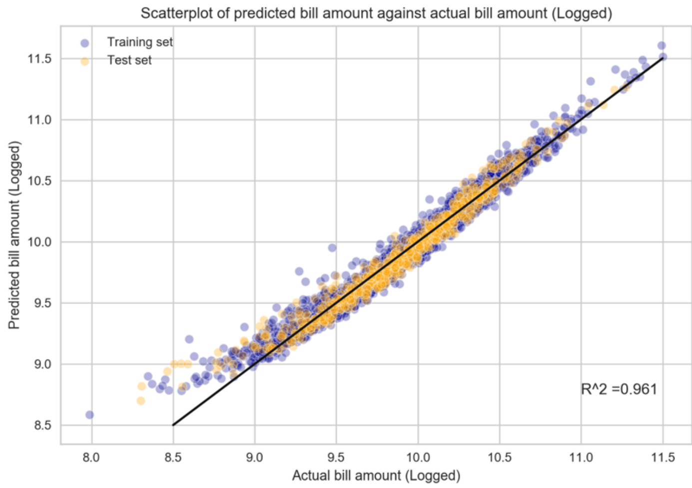

*Iterated removal of variables which improves score or parsimony of model to reach best set of variables. Cross-validation RMSLE = 0.101, OOB Score = 0.937



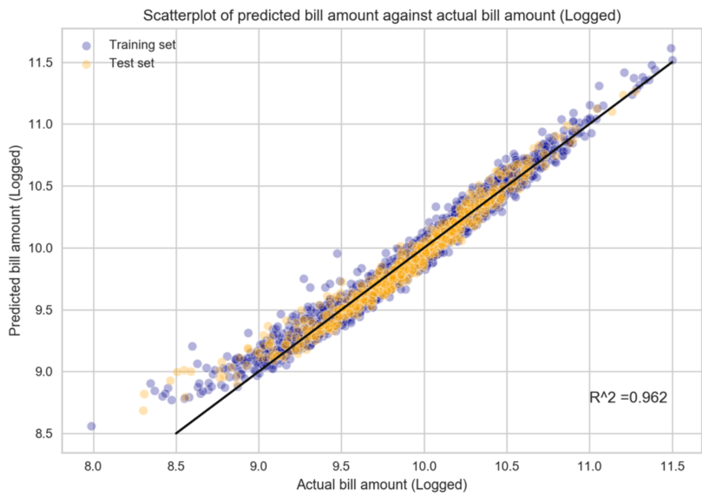

*Regression plot of trained RF model using best set of variables to predict on test set. $R^2$ score was calculated.

Elastic Net

- No overfitting
- The Elastic Net model did not perform much worse when fitting test set to model. The decrease in RMSLE score was marginal.
- Here, it can be seen that the RMSLE function helped with over-predicting cost of care, which RF was unable to do. Over-prediction was also not exaggerated and predictions was quite close to the diagonal line (which represents points where $R^2 = 1.00$).
- In defence of the random forest model, it was not trained with total cost on the log scale, which could explain was why it was prone to outliers
- When using transformed variables (standardised weight and lab test results 2, and square-rooted age at admission), the scores ($R^2$, train and test RMSLE) only improved slightly.




*Scatterplot of train and test set predicted vs actual cost of care, using the original variables




*Scatterplot of train and test set predicted vs actual cost of care, using the transformed variables

Insights

- Because BMI was high, and weight was an effective predictor of cost of care from the ML models, condition could be a diet-related disease
- Disease possibly prevalent in Malays.
- Older they are, higher the cost according to the scatterplot. Also significant from ML models
- Possible Coronary heart disease/Amputation given hospital stay (scatterplot)
- Not going to talk much about the foreigners, because it is understood that cost for foreigners would be more
- Polypharmacy possible, which contributes to total cost. Also indicates underlying illness/prone to disease. This could lead to further complications after surgery (assuming that they are going for surgery, which they should)
- Pre-ops medication, Lab test results and days admitted did not matter at all
- Admittedly, length of stay (LOS) in hospital was surprising, because from what I understand, LOS was found to be very highly correlated with healthcare costs.

Concluding statements

- Age, Race, Weight, Citizenship, Medical history and Symptoms contribute the most to cost of care for the particular condition
- Random Forests and Elastic Net regression are able to accurately predict cost of care just using these 6 classes of variables ($R^2 > 0.90$)
- Possible to target interventions or craft policies based on these 6 classes of variables