Hospital Data Analysis and Modeling

Saana Kuusela

About me

Saana Kuusela

Education
- Aalto University, Master of Science (Technology), Information Networks
- University of Helsinki, Master of Arts (Medicine) in Speech-Language Pathology
- University of Helsinki, Bachelor of Arts (Medicine) in Speech-Language Pathology
- Lyseonpuisto high school Rovaniemi, graduated in top 5% nationwide

Recent work experience
- Unity Technologies, Data Scientist (internship)
- Yourspace Oy/Kelvin Analytics, Software Developer
- Healthcare, research and clinical work as a speech-language pathologist

Interests
- Data Science and healthcare data analytics driven decision making process

Work flow

Joining the datasets

Cleaning the data

Null value analysis

Feature engineering

Explanatory data analysis

Feature correlations

Predictive models

- Predicting frequent visits with Logistic Regression model
- Predicting hospital admission costs with Random Forest Regression model
- Predicting hospital admission costs with Neural Network model

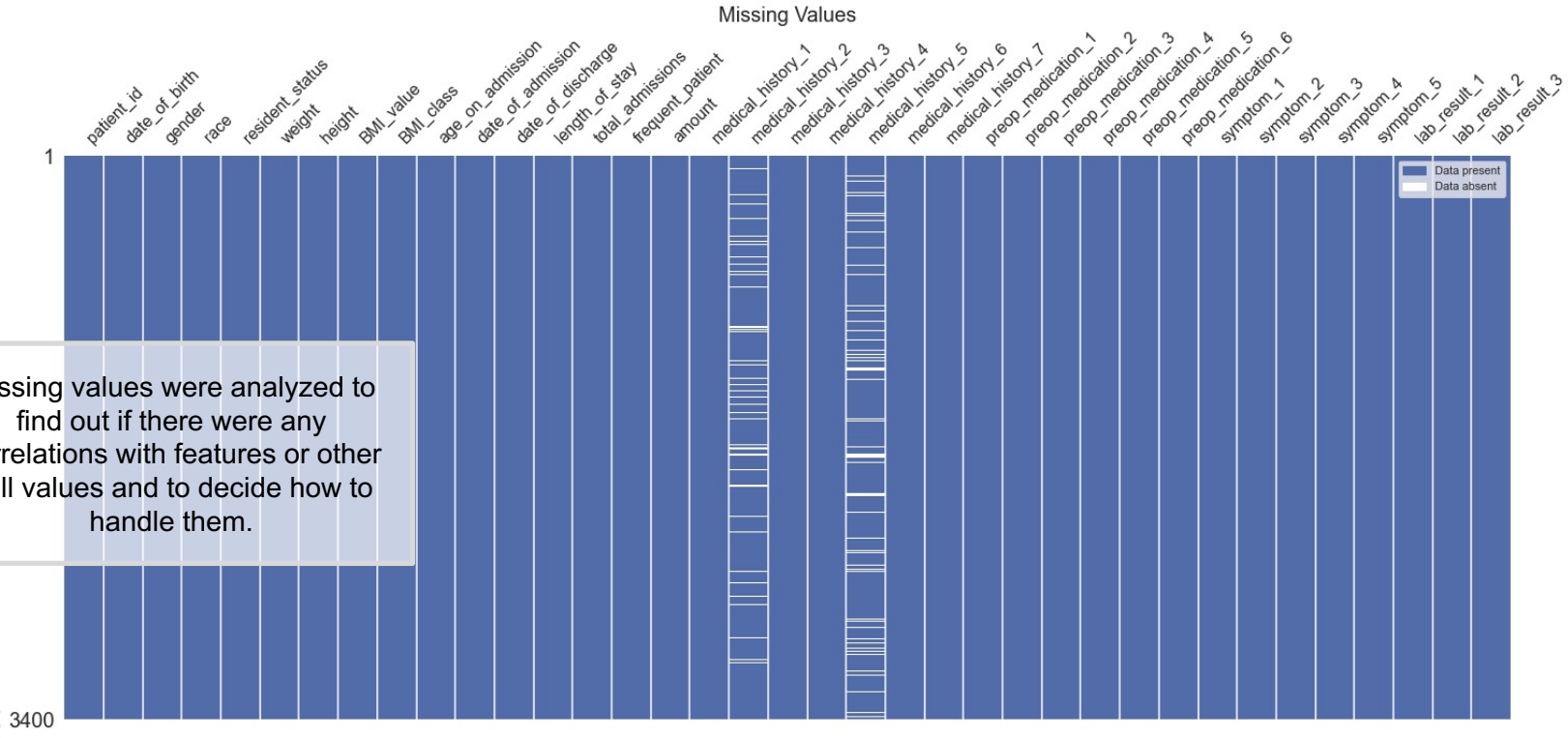
Tuning the models and evaluating their performance

Action points

1.

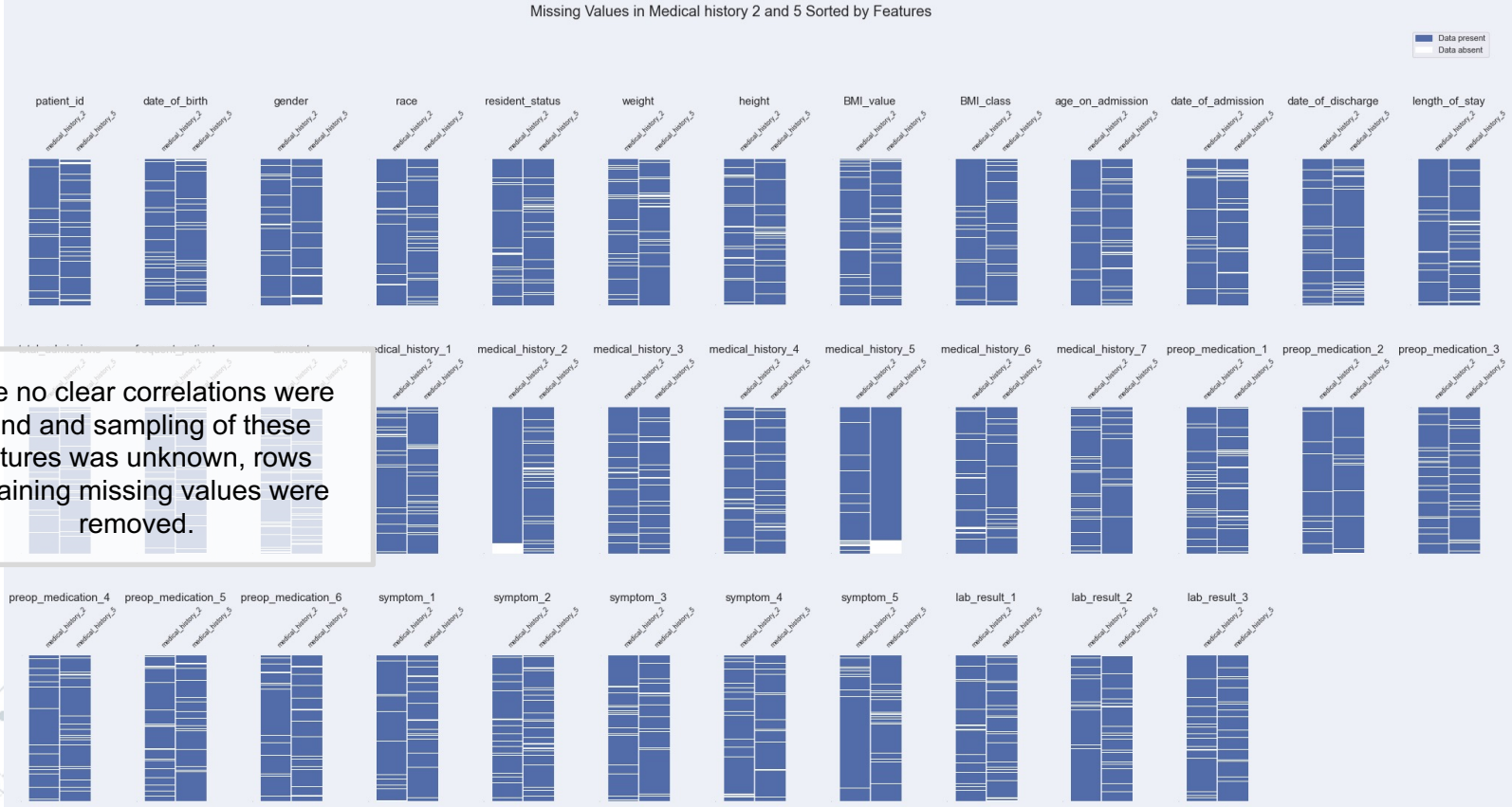
Data Preprocessing

Missing Value Analysis



Missing values were analyzed to find out if there were any correlations with features or other null values and to decide how to handle them.

Missing Value Analysis



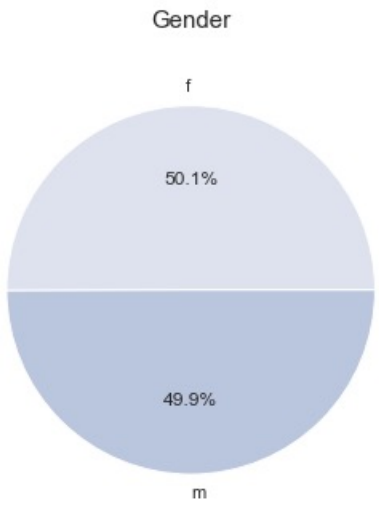
Since no clear correlations were found and sampling of these features was unknown, rows containing missing values were removed.

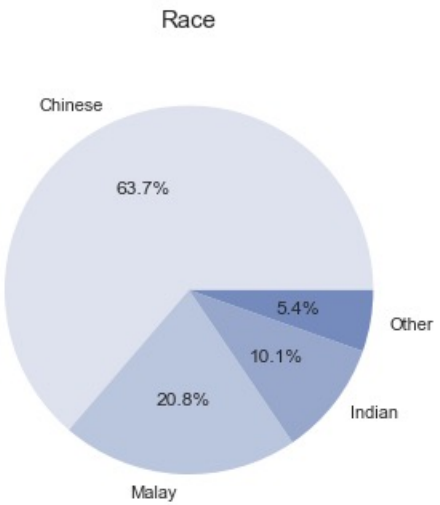
1.

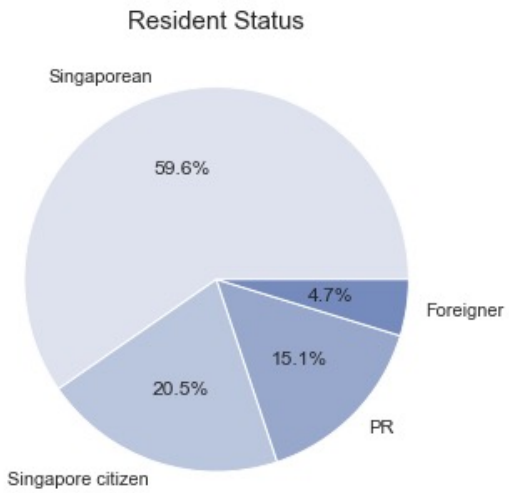
Patients

Patients

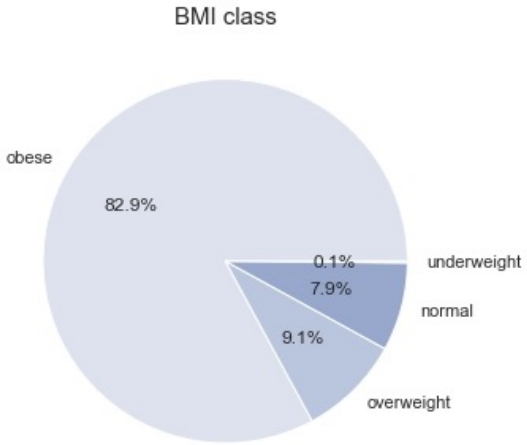
n = 3000

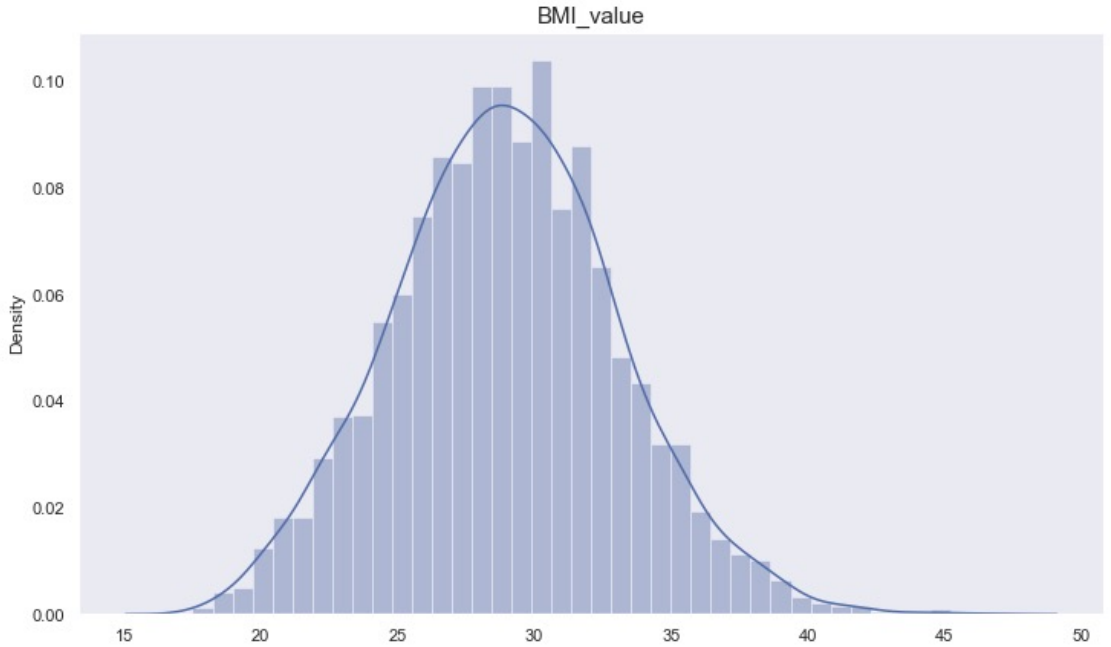




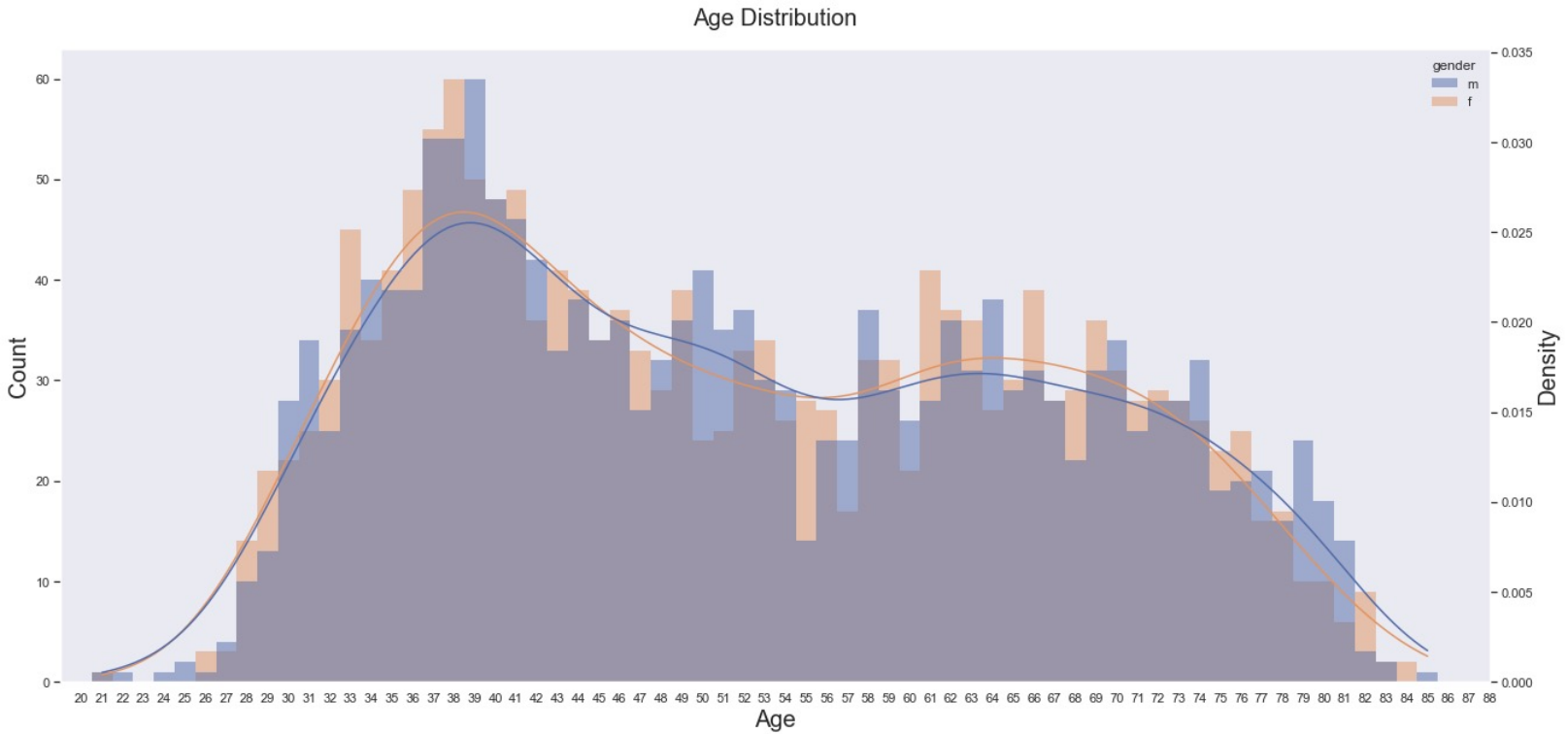


BMI (Asia-Pacific)

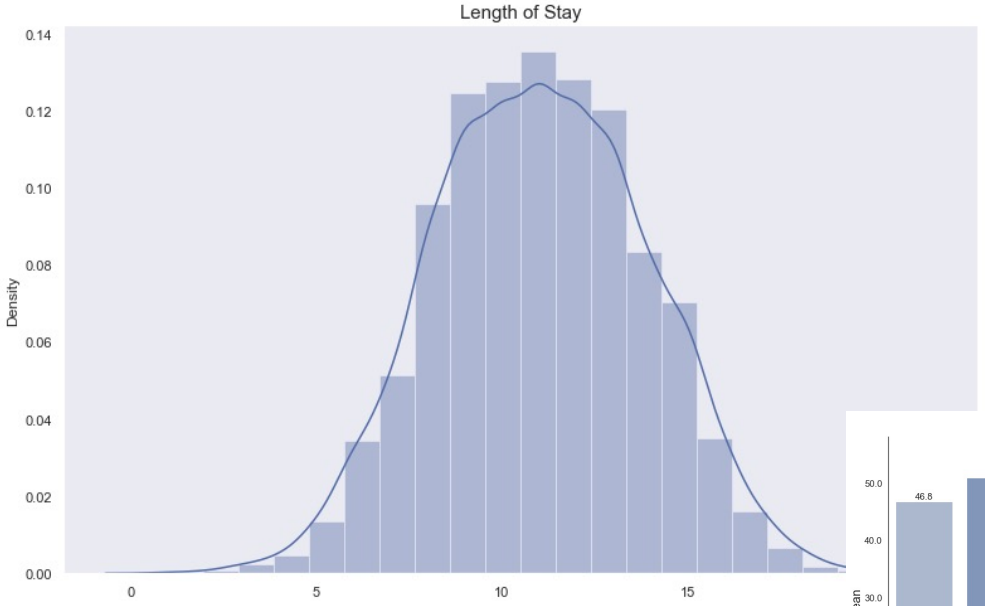




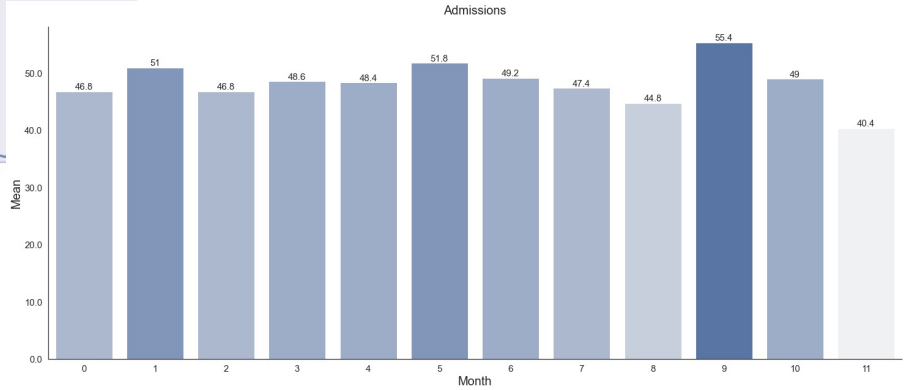
Age


Age Distribution
gender
m
f
Count
Density
Age
0
10
20
30
40
50
60
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88

Admissions



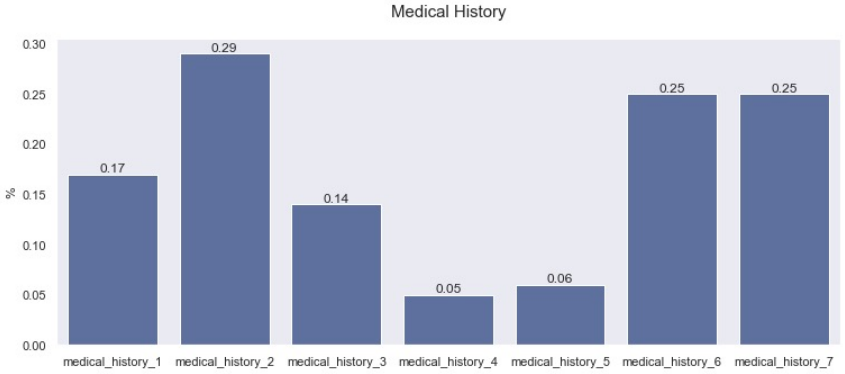
	Mean (STD)	Min-Max	Count
Length of Stay	11.0 (2.9)	1-20	3400

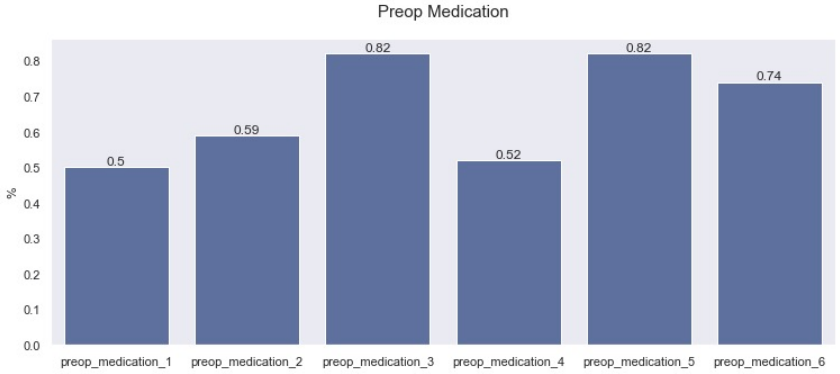


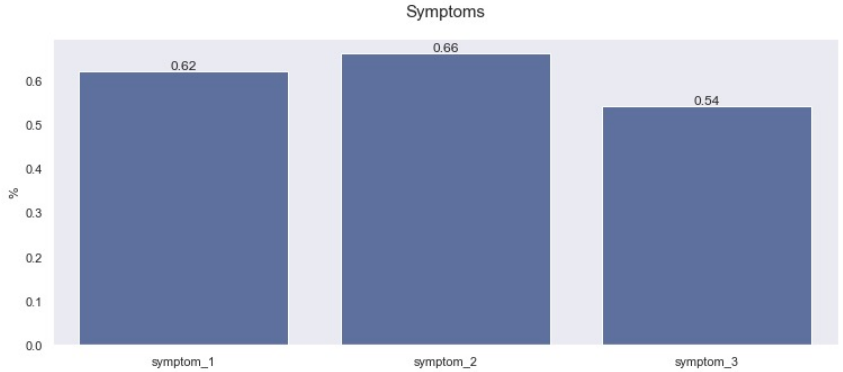
3.

Categorical Medical Data

Symptoms, Preop Medication and Medical History



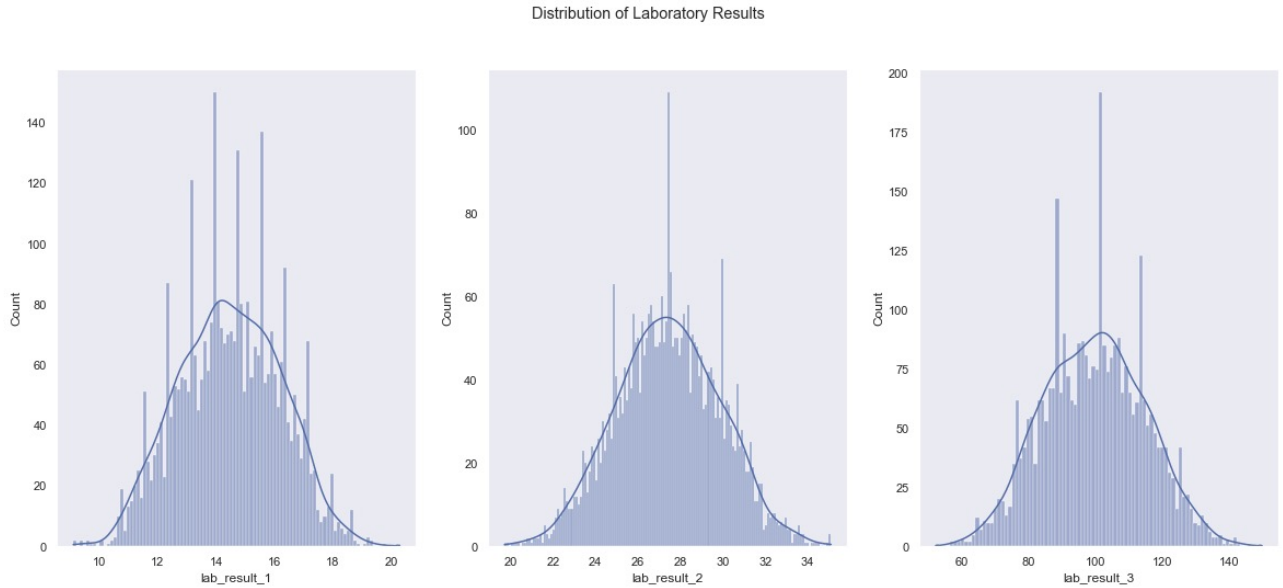




4.

Laboratory Results

Laboratory Results

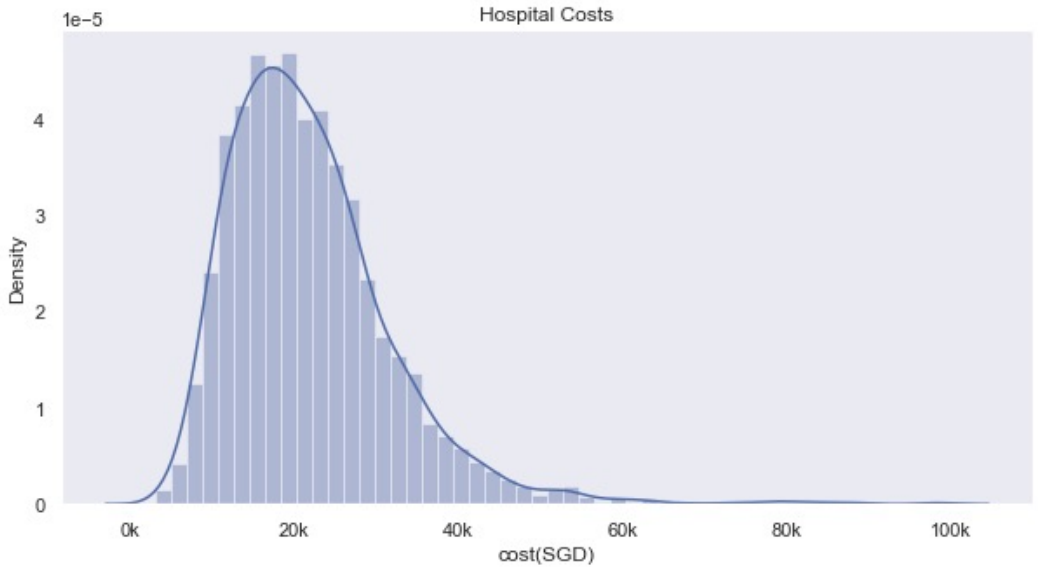


	Mean (STD)	Min-Max	Count
Lab result 1	14.5 (1.7)	9.1-20.3	3400
Lab result 2	27.4 (2.5)	19.7-35.1	3400
Lab result 3	99,5 (15.3)	52.0-150.0	3400

2.

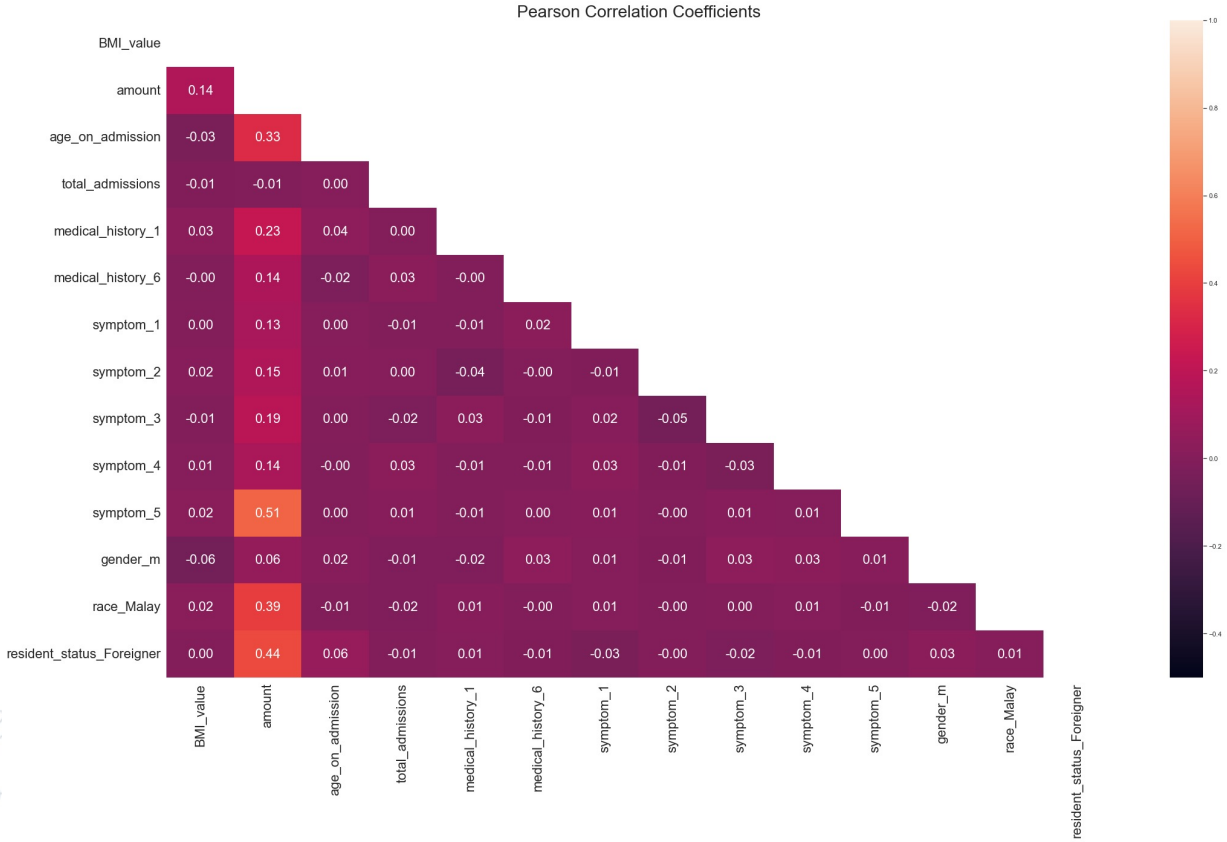
Hospital Admission Costs

Hospital Admission Costs



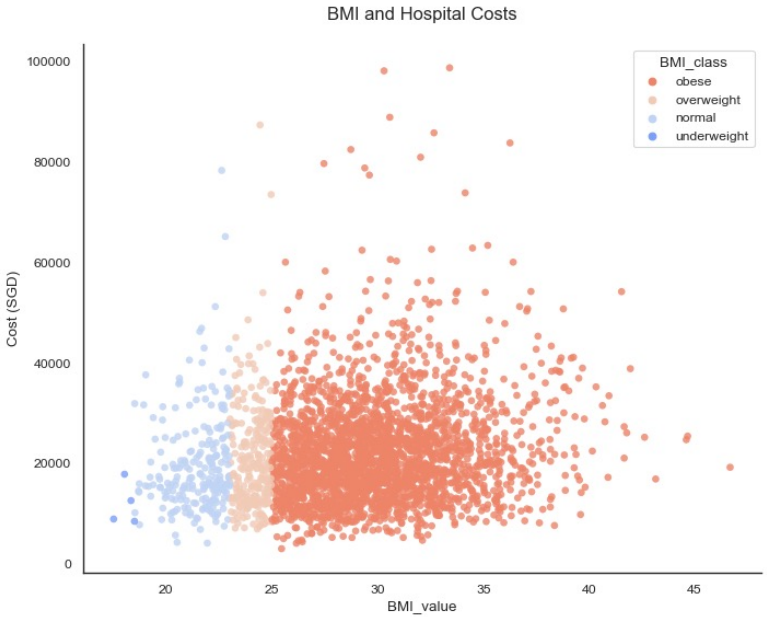
	Mean (STD)	Min-Max	Count
All	21,859 (10,155)	2,946-98,724	3400
Women	21,273 (9,982)	2,946-88,874	1702
Men	22,446 (10,295)	4,027-98,724	1698

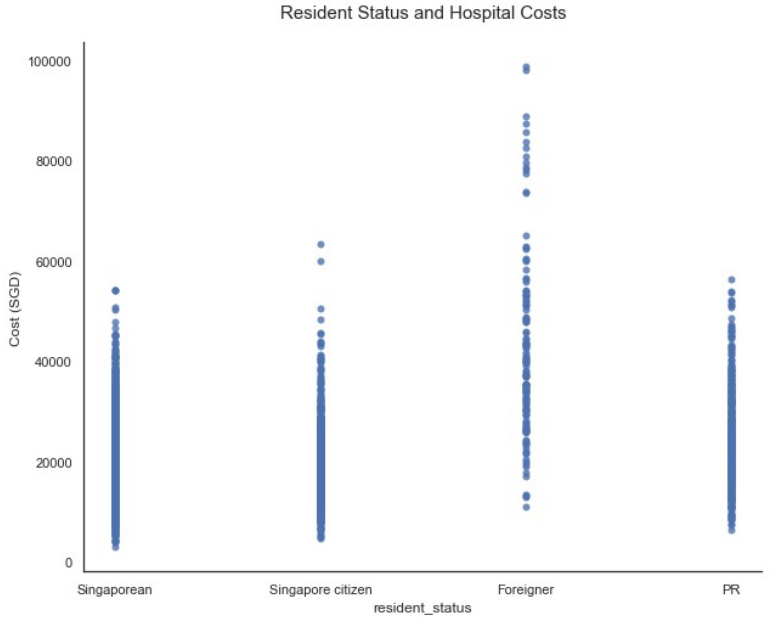
Features Contributing to Hospital Admission Costs



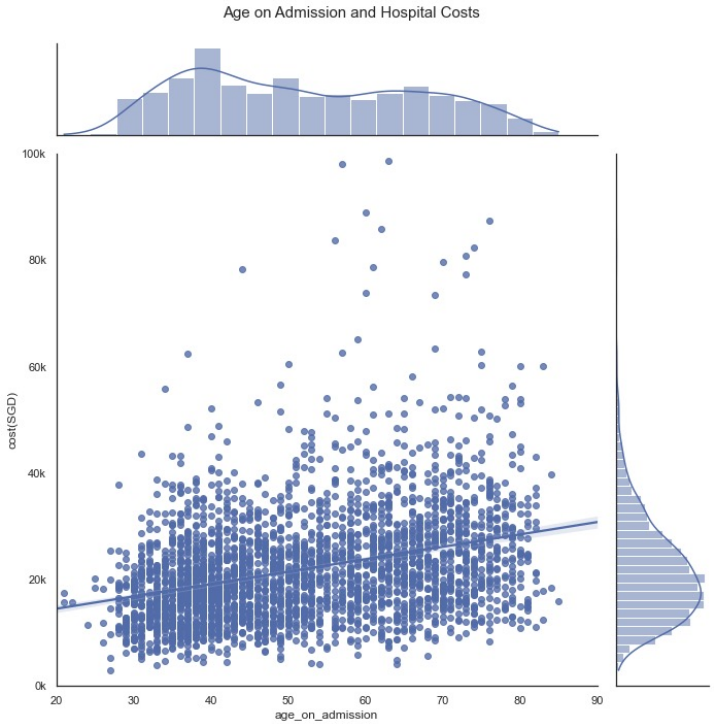
- Moderate positive correlation with costs: Foreigner, symptoms 5 and 6
- Weak positive correlation with costs: being Malay, age on admission, and medical history 1
- Additionally, there was some very weak positive correlations e.g., between BMI value and amount.

Features Contributing to Hospital Admission Costs





Features Contributing to Hospital Admission Costs



2.

Frequent Hospital Admissions

Frequent Hospital Admissions

Frequent patient was defined as a patient that will have subsequent admission during the year after the first admission.

Since the data was gathered before 28.12.2015, patients that were admitted after 28.12.2014 were excluded from the analysis.

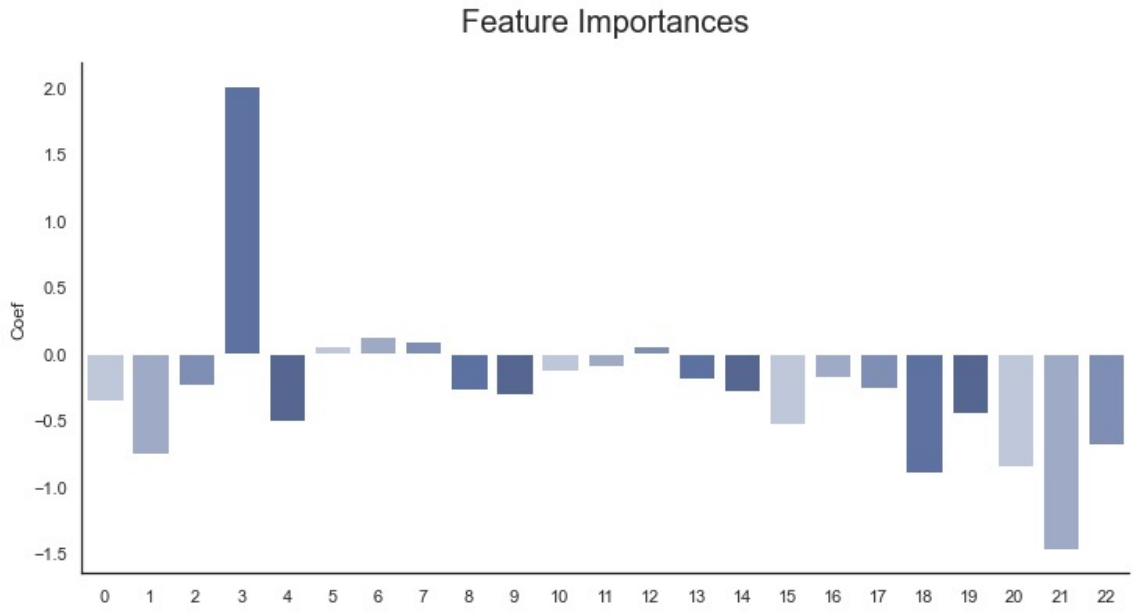
Foreigners were excluded since they are likely to not have subsequent visits because of their primary healthcare provider might be located elsewhere.

There were no statistically significant differences between these groups.

	Frequent	Non-Frequent
Mean BMI value	29.08	28.99
Mean age	50.46	51.37
Mean length of Stay	11.08	11.05
Mean of total admissions	2.15	1.17
Cost (SGD)	20,8k	20,9k
Count	2016	96

Predicting Frequent Hospital Admissions

Frequent hospital admissions were predicted with *logistic regression model*



0 BMI_value
1 age_on_admission
2 length_of_stay
3 amount
4 medical_history_1
5 medical_history_2
6 medical_history_3
7 medical_history_4
8 medical_history_5
9 medical_history_6
10 medical_history_7
11 preop_medication_4
12 preop_medication_5
13 preop_medication_6
14 symptom_1
15 symptom_2
16 symptom_3
17 symptom_4
18 symptom_5
19 gender_m
20 race_Indian
21 race_Malay
22 race_Other

2.

Predictive Modeling: Hospital Admission Costs

Random Forest Regression Model

Random Forest Regression model was trained and validated with k-fold cross validation. Pipeline was constructed to normalize the input data, select the best features and train the model.

Parameter selection was done inside grid search using SelectKBest. 20 parameters were chosen.

Parameters of Random Forest Regression were tuned inside grid search.

feature	importance
symptom_5	0.214388
race_Malay	0.076463
age_on_admission	0.065916
resident_status_Singaporean	0.061729
medical_history_1	0.029184
symptom_3	0.029121
symptom_1	0.015855
medical_history_6	0.013738
preop_medication_5	0.011512
symptom_4	0.010426

	MAE	RMSE	Explained Variance
Train	2 177	3 162	0.90

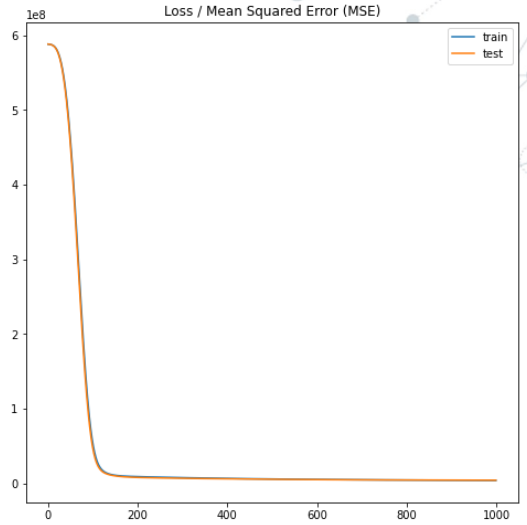
Neural Network Model

Neural Network model was defined as

```
keras.models.Sequential([
    keras.layers.Dense(200, input_dim = x_train.shape[1], activation='relu'),
    keras.layers.Dense(100, input_dim = x_train.shape[1], activation='relu'),
    keras.layers.Dense(1, activation='linear')
])
```

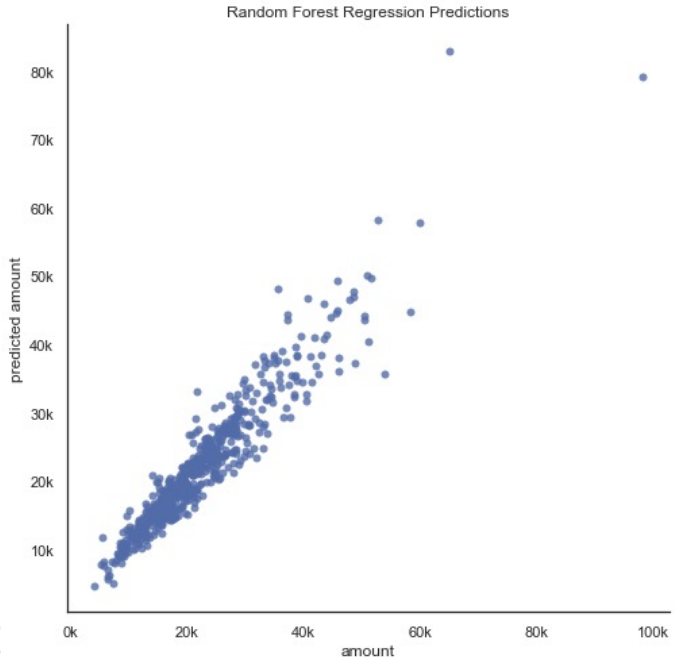
Adam was used as an optimization solver. Mean squared error (MSE) was used as loss function and root mean squared error (RMSE) and mean absolute error (MAE) as additional metrics.

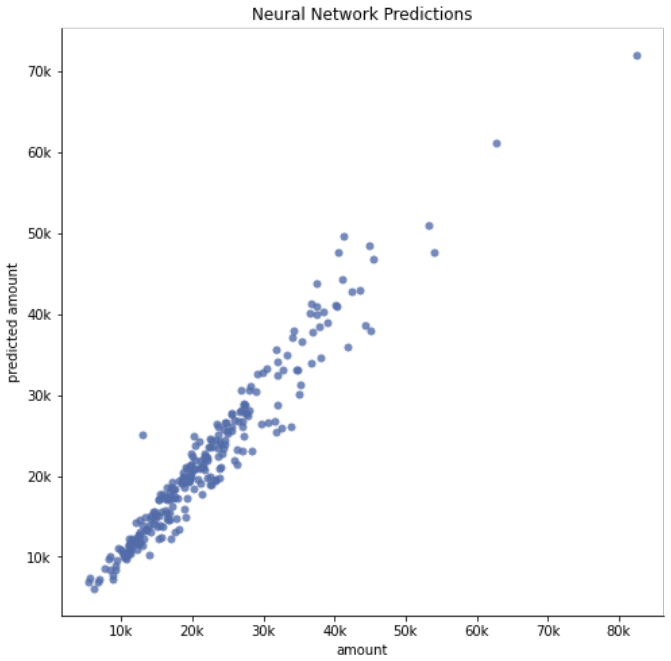
Epochs = 1000 and batch size = 500 were used when fitting the model. Model was tuned manually.



	MAE	RMSE
Train	1 331	1 931
Test	1 516	2 019

Model Comparison





Actions

- Obesity Prevention
- healthcare services for elderly people
- Patients 30-40 years of age
- Patients around 50-60 years of age
- Foreigner worker’s healthcare services

- Data enrichments: socio-economic status
- Data amount

Thank you!