

Prediction of HOLS Report

Introduction

- The main aim of our experiment is to find the hospital length of stay given the data of the patient.
- Initially two data frames were combined Hospital-length-of-stay and Data-at-admission.
- All the data column from dataset Data-at-admission were taken as the independent variables and the column 'Hospital-length-of-stay' from the dataset Hospital-length-of-stay was taken as the dependent variable
- **Initially** we had Total columns: 759, Total rows: 508
- They were matched on the basis on the column 'id' in Data-at-admission.csv, 'parent_id' in Hospital-length-of-stay.csv

Handling null values in the dataset:

1. Column 'height' and 'weight' null values were handled by setting the null values as their mean.
 2. Column 'smoking_history's null value was handled by filling it with 0 because it was observed that there were 1's for rows that had a smoking history.
 3. The columns ['year_they_quit','Methylprednisolone Acetate Injectable Suspension Usp','Prozac Capsules','Rybelsus'] were dropped as 'year_they_quit' wasn't seen as a significant column that would provide insight on the model and the rest of the columns had 508 null values.
- After handling missing values, **Total rows:** 506 rows x 747 columns

Types of columns in the dataset

Type	Counts
Medications	694
Reasons for admit	27
Comorbidities	18
Personal	7

Dependent and Independent Variables

- Y is the target variable - 'Hospital-length-of-stay', X is the independent variable apart from

Feature Study

Feature: Hospital_length_of_stay (Target variable)

count	506.000000
mean	12.454545
std	11.920318
25%	5.000000
50%	8.000000
75%	15.000000
max	81.000000
min	1.000000

- Then we were able to filter out the columns out of the 750 columns that did not have a 0 value in any of the rows. We divided the dataset into 6 parts, df_20_30 means that data points for patients having 20-30 days of HLOS and so on. We did not segregate the data for lower than 20 as the model was initially predicting the values accurately for values for HLOS under 20. With the following result:

Dataset	Row vs Column
df_20_30	36 rows vs 172 columns
df_30_40	27 rows vs 140 columns
df_40_50	12 rows vs 102 columns
df_50_60	5 rows vs 45 columns
df_60_70	2 rows vs 35 columns
df_70_80	1 row vs 17 columns

Findings:

- **Df_20_30**
 - Mean age: 64.80555555555556
 - Sex 0-20,1-16
 - Smoking history 0-33,1-3
 - Most rows in the dataset had the following comorbidities:
 - **admission_disposition**
 - **Hypertension**

- **Diabetes**
 - **Other**
- **Df_30_40**
 - Mean age: 69.4074074074074
 - Sex 0-10,1-17
 - Smoking history 0-23,1-4
 - Most rows in the dataset had the following comorbidities:
 - **admission_disposition**
 - **Hypertension**
 - **Other**
 - **reason_for_admission_COVID-19 [U07.1]**
- **Df_40_50**
 - Mean age: 71.83333333333333
 - Sex 0-3,1-9
 - Smoking history 0-11,1-1
 - Most rows in the dataset had the following comorbidities:
 - **admission_disposition**
 - **Hypertension**
 - **Other**
 - **Diabetes**
 - **reason_for_admission_COVID-19 [U07.1]**
- **Df_50_60**
 - Mean age: 68.4
 - Sex 0-3,1-2
 - Smoking history 0-4,1-1
 - Most rows in the dataset had the following comorbidities:
 - **admission_disposition**
 - **Hypertension**
 - **All had other Other**
 - **Diabetes**
 - **reason_for_admission_COVID-19 [U07.1]**
- **Df_60_70**
 - Mean age: 61.0
 - Sex 0-1,1-1
 - Smoking history 0-1,1-1
 - Most rows in the dataset had the following comorbidities:
 - **admission_disposition**
 - **Hypertension**
 - **All had other Other**
 - **Diabetes**
 - **reason_for_admission_COVID-19 [U07.1]**
 - **reason_for_admission_Pneumonia [J18.9]**
- **Df_70_80**
 - Mean age: 62.0
 - Sex 0-0,1-1

- Most rows in the dataset had the following comorbidities:
 - **admission_disposition**
 - **History of cancer [now in remission]**
 - **other Other**
 - **reason_for_admission_COVID-19 [U07.1]**
 - **Duloxetine**
 - **Tamsulosin**

Experiment 1- Applying different ML algorithms and techniques in the original dataset

- The number of columns were reduced in dimension with the help of PCA. We saw an elbow point at 6-8 however it explained around 42 percent of the variance of the data. Therefore, to achieve 80,85 and 90 percent of the data variance we had to reduce to 80, 105 and 140 dimensions, respectively.
 - a. Findings at 80 percent variance:
 - i. Using Lasso Regression, alpha=0.1:
 - 1. Mean Squared Error for training: 118.23190612205381
 - 2. Mean Squared Error for testing: 152.71776717173663
 - b. Findings at 85 percent variance:
 - ii. Using Lasso Regression, alpha=0.1:
 - 3. Mean Squared Error for training: 113.90381297745256
 - 4. Mean Squared Error for testing: 156.22622050093702
 - c. Findings at 90 percent variance:
 - iii. Using Lasso Regression, alpha=0.1:
 - 5. Mean Squared Error for training: 108.6513559222108
 - 6. Mean Squared Error for testing: 156.76659453490643
- Findings for Lasso regression for the initial dataset:
 - d. 31 important columns were calculated with the help of the coefficients
 - e. The model was created around those 31 features where:
 - iv. Alpha = 0.1
 - v. Mean Squared Error train: 113.32840599373615
 - vi. Mean Squared Error test: 151.27098598631906
 - f. It was also found that an increase in the alpha value decreased the error in the training set however increased the testing set error
- Findings for Ridge regression for the initial dataset:
 - g. Findings for alpha value = 1
 - vii. Mean Squared Error train: 37.39640727980008
 - viii. Mean Squared Error test: 197.2250962042391
 - h. Findings for alpha value = 1000
 - ix. Mean Squared Error train: 138.06498367265505
 - x. Mean Squared Error test: 149.3518668165006

With the results we concluded that there were not enough data points. Therefore, we decided to generate datapoints with the help of SMOGN.

Parameters used for SMOGN to generate data:

- 5 datasets were generated.
- Parameters used: `thr.rel_grid <- expand.grid(dist = "Euclidean", thr.rel = seq(0.2, 0.8, by = 0.01), C.perc = "extreme", p = 1, k = 2, repl = TRUE, pert = 0.01)`

Findings:

- Results for original dataset

	Train MSE (mean \pm std)	Test MSE (mean \pm std)
Lasso alpha 0.1	104.39 \pm 4.52	128.8 \pm 26.56
Lasso alpha 10	140.32 \pm 6.64	119.46 \pm 26.15
Ridge alpha 0.1	10.97 \pm 1.85	260.86 \pm 25.58
Ridge alpha 10	76.86 \pm 4.8	135.97 \pm 24.62
RandomForest	18.64 \pm 0.86	114.17 \pm 19.08
MLP	108.59 \pm 8.75	128.06 \pm 25.16
Logistic Regression	205.64 \pm 14.02	273.61 \pm 40.32

- Results for original HOLS data and the SMOGN data total with 1024 rows

	Train MSE (mean \pm std)	Test MSE (mean \pm std)
Lasso alpha 0.1	142.84 \pm 5.14	168.95 \pm 15.63
Lasso alpha 10	274.27 \pm 4.3	259.66 \pm 9.64
Ridge alpha 0.1	26.12 \pm 1.38	166.47 \pm 12.37
Ridge alpha 10	90.82 \pm 3.21	124.57 \pm 14.01
RandomForest	9.01 \pm 0.67	74.23 \pm 14.13
MLP	16.98 \pm 1.09	116.07 \pm 5.76
Logistic Regression	29.66 \pm 1.81	81.46 \pm 17.05

Models and their hyper parameters:

Lasso alpha 0.1 hyperparameters:

`{'alpha': 0.1, 'copy_X': True, 'fit_intercept': True, 'max_iter': 1000, 'positive': False, 'precompute': False, 'random_state': None, 'selection': 'cyclic', 'tol': 0.0001, 'warm_start': False}`

Lasso alpha 10 hyperparameters:

`{'alpha': 10, 'copy_X': True, 'fit_intercept': True, 'max_iter': 1000, 'positive': False, 'precompute': False, 'random_state': None, 'selection': 'cyclic', 'tol': 0.0001, 'warm_start': False}`

Ridge alpha 0.1 hyperparameters:

```
{'alpha': 0.1, 'copy_X': True, 'fit_intercept': True, 'max_iter': None, 'positive': False, 'random_state': None, 'solver': 'auto', 'tol': 0.0001}
```

Ridge alpha 10 hyperparameters:

```
{'alpha': 10, 'copy_X': True, 'fit_intercept': True, 'max_iter': None, 'positive': False, 'random_state': None, 'solver': 'auto', 'tol': 0.0001}
```

RandomForest hyperparameters:

```
{'bootstrap': True, 'ccp_alpha': 0.0, 'criterion': 'squared_error', 'max_depth': None, 'max_features': 1.0, 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'monotonic_cst': None, 'n_estimators': 100, 'n_jobs': None, 'oob_score': False, 'random_state': None, 'verbose': 0, 'warm_start': False}
```

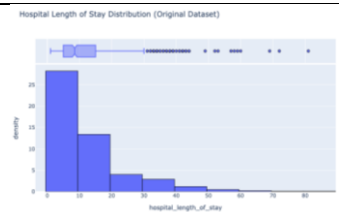
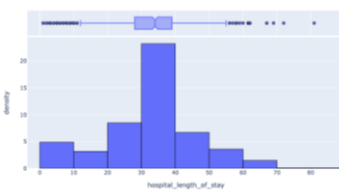
MLP hyperparameters:

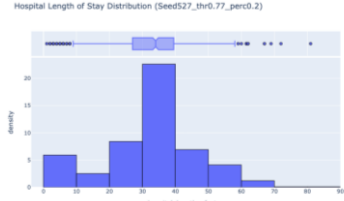
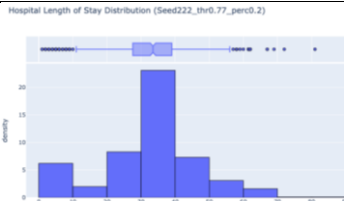
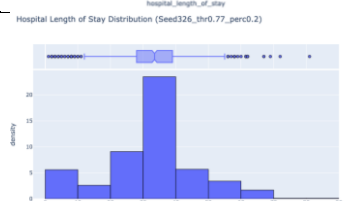
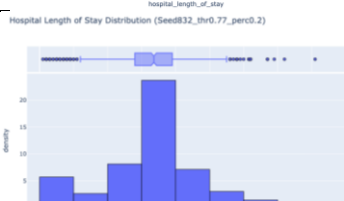
```
{'activation': 'relu', 'alpha': 0.0001, 'batch_size': 'auto', 'beta_1': 0.9, 'beta_2': 0.999, 'early_stopping': False, 'epsilon': 1e-08, 'hidden_layer_sizes': (100,), 'learning_rate': 'constant', 'learning_rate_init': 0.001, 'max_fun': 15000, 'max_iter': 500, 'momentum': 0.9, 'n_iter_no_change': 10, 'nesterovs_momentum': True, 'power_t': 0.5, 'random_state': None, 'shuffle': True, 'solver': 'adam', 'tol': 0.0001, 'validation_fraction': 0.1, 'verbose': False, 'warm_start': False}
```

Logistic Regression hyperparameters:

```
{'C': 1.0, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'l1_ratio': None, 'max_iter': 100, 'multi_class': 'deprecated', 'n_jobs': None, 'penalty': 'l2', 'random_state': None, 'solver': 'lbfgs', 'tol': 0.0001, 'verbose': 0, 'warm_start': False}
```

Analyzing the original dataset and the newly generated datasets

Description	Distribution
Original Dataset	
Synthetic dataset - Seed410_thr0.77_perc0.2	

Synthetic dataset - Seed527_thr0.77_perc0.2	
Synthetic dataset - Seed222_thr0.77_perc0.2	
Synthetic dataset - Seed326_thr0.77_perc0.2	
Synthetic dataset - Seed832_thr0.77_perc0.2	

To check the authenticity of the synthetic data generated, we produced a way to validate the data. We will be listing out the columns for each bin of data i.e., 10-20, 20-30, 30-40, 40-50, 50-60, 60-70 and 70-80 where there is our presence of values in a single row. This means that columns without values in any of the datapoints (rows) are removed. With this we can compare the original dataset with the newly generated dataset.

For the original dataset:

Bin	Names	Counts_Personal	Counts_Comorbidities	Counts_Reason_Admit	Counts_Medications
0 10-20	[id, hospital_length_of_stay, age, sex, height...	9	18	15	345
1 20-30	[id, hospital_length_of_stay, age, sex, height...	9	16	8	162
2 30-40	[id, hospital_length_of_stay, age, sex, height...	9	16	8	142
3 40-50	[id, hospital_length_of_stay, age, sex, height...	9	13	3	59
4 50-60	[id, hospital_length_of_stay, age, sex, height...	8	12	4	31
5 60-70	[id, hospital_length_of_stay, age, sex, height...	9	5	2	8
6 70-80	[id, hospital_length_of_stay, age, sex, height...	8	4	1	4

For synthetic datasets

	Bin	Names	Counts_Personal	Counts_Comorbidities	Counts_Reason_Admit	Counts_Medications
0	10-20	[id, hospital_length_of_stay, age, sex, height...	9	15	6	143
1	20-30	[id, hospital_length_of_stay, age, sex, height...	9	16	10	130
2	30-40	[id, hospital_length_of_stay, age, sex, height...	9	18	10	214
3	40-50	[id, hospital_length_of_stay, age, sex, height...	9	17	9	116
4	50-60	[id, hospital_length_of_stay, age, sex, height...	9	14	8	55
5	60-70	[id, hospital_length_of_stay, age, sex, height...	9	10	4	23
6	70-80	[id, hospital_length_of_stay, age, sex, height...	8	4	1	4

	Bin	Names	Counts_Personal	Counts_Comorbidities	Counts_Reason_Admit	Counts_Medications
0	10-20	[id, hospital_length_of_stay, age, sex, height...	9	16	5	126
1	20-30	[id, hospital_length_of_stay, age, sex, height...	9	16	9	119
2	30-40	[id, hospital_length_of_stay, age, sex, height...	9	18	10	214
3	40-50	[id, hospital_length_of_stay, age, sex, height...	9	17	9	116
4	50-60	[id, hospital_length_of_stay, age, sex, height...	9	14	8	55
5	60-70	[id, hospital_length_of_stay, age, sex, height...	9	10	4	23
6	70-80	[id, hospital_length_of_stay, age, sex, height...	8	4	1	4

	Bin	Names	Counts_Personal	Counts_Comorbidities	Counts_Reason_Admit	Counts_Medications
0	10-20	[id, hospital_length_of_stay, age, sex, height...	9	16	4	99
1	20-30	[id, hospital_length_of_stay, age, sex, height...	9	17	9	125
2	30-40	[id, hospital_length_of_stay, age, sex, height...	9	18	10	214
3	40-50	[id, hospital_length_of_stay, age, sex, height...	9	17	9	116
4	50-60	[id, hospital_length_of_stay, age, sex, height...	9	14	8	55
5	60-70	[id, hospital_length_of_stay, age, sex, height...	9	10	4	23
6	70-80	[id, hospital_length_of_stay, age, sex, height...	8	4	1	4

	Bin	Names	Counts_Personal	Counts_Comorbidities	Counts_Reason_Admit	Counts_Medications
0	10-20	[id, hospital_length_of_stay, age, sex, height...	9	17	5	132
1	20-30	[id, hospital_length_of_stay, age, sex, height...	9	16	10	127
2	30-40	[id, hospital_length_of_stay, age, sex, height...	9	18	10	214
3	40-50	[id, hospital_length_of_stay, age, sex, height...	9	17	8	114
4	50-60	[id, hospital_length_of_stay, age, sex, height...	9	14	8	55
5	60-70	[id, hospital_length_of_stay, age, sex, height...	9	10	4	23
6	70-80	[id, hospital_length_of_stay, age, sex, height...	8	4	1	4
	Bin	Names	Counts_Personal	Counts_Comorbidities	Counts_Reason_Admit	Counts_Medications
0	10-20	[id, hospital_length_of_stay, age, sex, height...	9	13	8	100
1	20-30	[id, hospital_length_of_stay, age, sex, height...	9	16	10	123
2	30-40	[id, hospital_length_of_stay, age, sex, height...	9	18	10	214
3	40-50	[id, hospital_length_of_stay, age, sex, height...	9	17	9	116
4	50-60	[id, hospital_length_of_stay, age, sex, height...	9	14	8	55
5	60-70	[id, hospital_length_of_stay, age, sex, height...	9	10	4	23
6	70-80	[id, hospital_length_of_stay, age, sex, height...	8	4	1	4