

Die Another Day (DAD) Hospital Case

Instructions:

- You have to use Python/R for the complete analysis. Use Jupyter notebook/RStudio for documentation and python/R code.
-

We will be using DAD hospital data in this exercise. Refer the **Exhibit 1** to understand the feature list. Use the DAD Hospital data and answer the below questions.

1. Load the dataset in Jupyter Notebook/RStudio
2. Build a correlation matrix between all the numeric features in the dataset. Report the features, which are correlated at a cut-off of 0.70. What actions will you take on the features, which are highly correlated?
3. Build a new feature named BMI using body height and body weight. Include this as a part of the data frame created in step 1.
4. Past medical history code has 175 instances of missing value (NaN). Impute 'None' as a label wherever the value is NaN for this feature.
5. Create a new data frame with the numeric features and categorical features as dummy variable coded features. Which features will you include for model building and why?
6. Split the data into training set and test set. Use 80% of data for model training and 20% for model testing.
7. Build a model using age as independent variable and cost of treatment as dependent variable.
 - Is age a significant feature in this model?
 - What inferences can be drawn from this model?
8. Build a model with statsmodel.api/lmStepAIC to estimate the total cost to hospital. How do you interpret the model outcome? Report the model performance on the test set.
9. This is python specific step. Build a model with statsmodel.formula.api to estimate the total cost to hospital and report the model performance on the test set. What difference do you observe in the model built here and the one built in step 8.
10. Build a model using sklearn package (Python) or glmnet (caret package) to estimate the total cost to hospital. What difference do you observe in this model compared to model built in step 8.

Exhibit 1

Variable	Description
Age	Age of the patient in years
Body Weight	Weight of the patient in Kilograms
Body Height	Height of the patient in cm
HR Pulse	Pulse of patient at the time of admission
BP-High	High BP of patient (Systolic)
BP-Low	Low BP of patient (Diastolic)
RR	Respiratory rate of patient
HB	Hemoglobin count of patient
Urea	Urea levels of patient
Creatinine	Creatinine levels of patient
Marital Status	Marital status of the patient
Gender	Gender code for patient
Past Medical History Code	Code given to the past medical history of the Patient
Mode of Arrival	Way in which the patient arrived the hospital
State at the Time of Arrival	State in which the patient arrived
Type of Admission	Type of admission for the patient
Key Complaints Code	Codes given to the key complaints faced by the patient
Total Cost to Hospital	Actual cost incurred by the hospital
Total Length of Stay	Number of days patient stayed in the hospital
Length of Stay - ICU	Number of days patient stayed in the ICU
Length of Stay - Ward	Number of days patient stayed in the ward
Implant used (Y/N)	Any implant done on the patient
Cost of Implant	Total cost of all the implants done on the patient, if any