Microsoft Data Science Institute

Programming in Python for Data Science

Presented: March 2024

SEMOS education

Microsoft Data Science Institute

Programming in Python for Data Science

1. **Introduction**:

   This project aims to predict **the Length of Stay**(LoS)
   of patients in hospital using **Linear Regression,** based on dataset:

   Healthcare.Blueprint-Predicting Length of Stay in Hospitals

2. Sanela Nikodinoska – presenter / contributor

3. Project tasks:

- Load the data from the file: Healthcare.Blueprint-Predicting Length of Stay in Hospitals.
- Perform **Exploratory Data Analysis (EDA)** to understand the distribution of the data, identify patterns, trends, and detect outliers.
- Deal with outliers
- Remove columns which are not significant for the model.
- Preprocess the data by scaling the features, mapping categorical variables, and splitting the data into training and testing sets.
- Train a **Linear Regression** model on the training set.
- Evaluate the performance of the model on the testing set using **R-squared and its adjusted value**.
- Finally, use the trained model to predict the length of stay of new patients.

Microsoft Data Science Institute

Programming in Python for Data Science

4. Conclusion: Predictions based on several experiments:

VER	Project - Predict LoS	R-score	Adjusted r-score	Ridge Regression
1.1	outliers removed quantile / imputed, scaled data	0.683140798227	0.670096851671	67.27%
1.2	outliers removed z_score, quantile/ imputed, scaled data	0.747968512976	0.747889382211	74.75%
1.3	**outliers removed z_score, quantile/ imputed,noscaled data**	**0.747979225732**	**0.747900098330**	**74.75%**
1.4	outliers removed z_score, quantile/ imputed, noscaled data, log tran	0.640173483838	0.640060508480	63.99%
2.1	outliers kept, noscaled data, log transformation	0.642514754627	0.642407475529	64.17%
2.2	outliers kept, noscaled data	0.749246137156	0.749170887481	74.89%
2.3	outliers kept, scaled data	0.749247021539	0.749171772130	74.89%
2.4	outliers kept, scaled data, columns (p_values) removed	0.749247021539	0.749171772130	74.89%

Microsoft Data Science Institute

Programming in Python for Data Science