PROJECT WRITE-UP

DESCRIPTION

Background and Objective:
Every year thousands of applications are being submitted by international students for admission in colleges of the USA. It becomes an iterative task for the Education Department to know the total number of applications received and then compare that data with the total number of applications successfully accepted and visas processed. Hence to make the entire process easy, the education department in the US analyze the factors that influence the admission of a student into colleges. The objective of this exercise is to analyse the same.

Domain: Education

Dataset Description:

Attribute	Description
GRE	Graduate Record Exam Scores
GPA	Grade Point Average
Rank	It refers to the prestige of the undergraduate institution. The variable rank takes on the values 1 through 4. Institutions with a rank of 1 have the highest while those with a rank of 4 have the lowest.
Admit	It is a response variable; admit/don't admit is a binary variable where 1 indicates that student is 0 indicates that student is not admitted.
SES	SES refers to socioeconomic status: 1 - low, 2 - medium, 3 - high.
Gender_male	Gender_male (0, 1) = 0 -> Female, 1 -> Male
Race	Race – 1, 2, and 3 represent Hispanic, Asian, and African-American

SUMMARY:

- Upon performing exploratory data analysis, the dependent or target variable is found to be **admit**. And the independent variables are: gre, gpa, rank, gender, ses, race.
- The dataset was checked for any missing value using is.na() and the result returned 0.
- Outlier analysis was performed by visualizing the continuous variables gre and gpa using boxplot.Outliers found were successfully removed by dropping the respective record.
- The structure of the dataset was found using str() function and datatype conversion was erformed using as.factor() function.
- Normality of the data set was determined using histogram and density plot for continuous variable gre and gpa.
- The gre and gpa was scaled for better processing thereafter.
- Logistic regression was performed. Observing the summary of model, it was clear that the significant variables were gre, gpa, rank.insignificant variables were dropped.
- The dataset was split using split() function into train and test dataset in the ration 7:3.
- Confusion matrix was build. Which states the following
    - Accuracy : 74.79%
    - Specificity: 59.09%
    - Sensitivity: 78.35%

- Other modeling techniques used are as follows:
- Suppor vector machine : used svm(). The model yielded the following results:
    - Accuracy : 73.11%
    - Sensitivity : 88.2%
    - Specificity : 35.29%
- Decision tree : used rpart().following are the evaluations:
    - Accuracy : 70.59%
    - Sensitivity : 80%
    - Specificity : 47.06%
- Random forest : used Randomforest().following are the evaluations:
    - Accuracy : 65.5%
    - Sensitivity : 82.35%
    - Specificity : 23.53%
- Upon comparison , the model with highest accuracy was found to be Logistic Regression. Hence it was selected as the champion model.
- The descriptive part is left behind because, I dont remember similar method taught during the sessions and after raising a ticket, I was given a video explaining the project, but the descriptive part is not explained in the same.