Predictive Modeling based on the College_Admission_data

Final Data Dive

The first day of data dive, your group will be focused on performing exploratory data analysis and identifying the most helpful classification and regression problem that can help make a better admission decision.

In the second day of the data dive your group will be focused on the modeling part and optimize the implemented models.

Each group will do a PowerPoint presentation summarizing their discoveries on the final day. Your group will focus on the preparation of the presentation and present. The presentation should consist of 5-8 slides and cover the motivation and objectives, problem statement, methods employed, findings, and conclusion. Please ensure that your group prepares the presentation accordingly.

Step 1

*Read the data from the file
* Explore the data and create some useful visualization
* Change the categorical variables to numerical variables using the built-in function in sklearn or by writing the function
* What would be an interesting/helpful classification problem based on your data? Identify/create the possible target variable for the classification problem.
* Explore the relationship of predictors and also with the target variable
* Look for missing values, impute or drop as needed, and complete the data preprocessing step.
* Perform feature engineering to create better features that are more informative and helpful in modeling.
* Take a final look at the data and perform the trail test split (70/30)

You have data ready and the problem ready (i.e., you know what you are trying to learn/ what problem you will solve) utilizing the data you have! Let's start building appropriate models.

Step 2

* Run the Logistic regression on the preprocessed data using all appropriate predictors with the chosen target variable
* Build the model on 70 percent of the data and evaluate the model on the remaining 30 percent of the dataset
* Explain the level of importance of each predictor
* Write the logistic function
* Explore the various accuracy measures such as Accuracy, Sensitivity, Specificity, and ROC AUC. What do those values mean in terms of your problem? Explain.
* Discuss with your group and provide a summary of the result.

Step 3

Take your notebook at the end of the previous step and do the following.

* Run the Decision Tree Classifier on the preprocessed data using all appropriate predictors, with the chosen target variable

* Build the model on training data and evaluate the model on the test dataset

* Explain the level of importance of each predictor

* Plot the decision tree with the max_depth = 5

* Prune the decision tree and build the final decision tree model and name it dt_best

* Plot the final decision tree. Did you notice any difference?

* Explore the various accuracy measures such as Accuracy, Sensitivity, Specificity, and ROC AUC. What do those values mean in terms of your problem?

Step 4

* Run the Random Forest Classifier on the preprocessed data using all appropriate predictors with the chosen target variable

* Build the model on training data and evaluate the model on the test dataset

* Explore various hyperparameters including class_weight, n_estimators, and max_depth in building the trees and find the optimal combination

* Fit the final random forest model and name it rf_best

* Explain the level of importance of each predictor on your classification model

* Explore the various accuracy measures such as Accuracy, Sensitivity, Specificity, and ROC AUC. What do those values mean in terms of your problem?

Step 5

*From the prior work, take the final random logistic regression model and name it logi_best

* Prepare the table of all the performance scores of all the models experimented in one table and name it df_scores

* Create visualizations to gain further insights into the results

* Write the summary of your work.

Step 6

You will work on the regression modeling part based on the useful target variables identified by your group.

Similarly, as in the classification modeling, you will implement LASSO, Ridge, and Random Forest models for the regression analysis and perform the comparative analysis of the results.

Provide a summary of the work.

Step 7

* Kindly ensure that you list the names of all the members in your group on the jupyter notebook you will be submitting.

* Plan for the 5-minute presentation of your findings to the class. Your presentation should clearly mention why the chosen target variable is important and how its accurate prediction helps the college to make informed decisions.

* Each group will present the findings to the class, and One of the members of the group will email the final jupyter notebook to me

Note:

* Based on your power point presentations and your submitted notebook, and your attendance in the data dive, you will receive the grades for the data dive.