Machine Learning Model to Predict a Priority Level of University Admission: Case study for multiclass classification

Summary Report of Capstone Project

by

M. Prabhashrini Dhanushika

Reg.No: 50

Machine Learning Foundation - Batch 04
Data Science Academy
Dialog Axiata PLC

Contents

1 Introduction

The number of students who applies post graduate studies in top ranked universities has increased significantly due to the competition among the young generation for the best knowledge and experience. Therefore, having automated system for solving student admission problem is remarkably important for educational institutes. If there is a mechanism to prioritize the applications they have received, it will speed the university admission procedure. Other than that, it will make student know in advance about their own priority level in the admission procedure when they provide required information.

Therefore, this report summarizes the solution reached through Machine Learning model to predict the priority level of the applicant who applies for post graduate studies in the University. Priority level has been classified as 1st, 2nd, 3rd and 4th in this solution.

2 Data

2.1 Introduction of Data

The data set belongs to the educational domain and it contains 400 number of records. It includes information as follows:

GRE Scores (out of 340)

TOEFL Scores (out of 120)

University Rating (out of 5)

Statement of Purpose (SOP) (out of 5)

Letter of Recommendation (LOR) Strength (out of 5)

Undergraduate GPA (out of 10)

Research Experience (either 0 or 1)

Chance of Admit (ranging from 0 to 1)

2.2 Source for the Dataset

Dataset used in this study was taken from Kaggle database. Link for the dataset is as follows: https://www.kaggle.com/datasets/akshaydattatraykhare/data-for-admission-in-the-university

2.3 Preparation of Dataset

Since the priority of being selected is needed to be predicted, new column 'Priority_Admit' was created by using the values in 'Chance of Admit' column. 'Priority_Admit' has 4 levels as '1st_Priority', '2nd_Priority', '3rd_Priority', 4th Priority'. 4 classes which have been defined in the dataset as class 1 to class 4 are related to 1st priority to 4th priority respectively. These 4 priority levels were decided as per the quartiles in 'Chance of Admit' column.

3 Methodology

Different Machine Learning models were fitted to the data in order to find the solution. In that procedure, as a first step, dataset has been prepared as per the requirement. Column names of the dataset have been prepared by removing the spaces. Then data preprocessing was applied. In data preprocessing, testing was done for the duplicate records and missing values and necessary actions were taken to rectify the data issues.

As a second step, necessary datatype conversions were conducted to optimize the memory usage. After that, data exploration was conducted. There, univariate analysis and bivariate analysis were performed to investigate the nature of data. Finally, correlation matrix was created to inspect the relationship between features and target, as well as among the features.

After identifying feature variables as GRE Scores, TOEFL Scores, University Rating, Statement of Purpose (SOP), Letter of Recommendation (LOR) Strength, Undergraduate GPA (CGPA) and Research Experience and target variable as Priority Level (y_act) which was defined as 1, 2, 3, and 4, well prepared dataset was split to 70:30 ratio to get the train and test datasets. Other than that, pre-processed dataset was saved as 'adm_data_processed.pkl' for later use.

As a next step, multiple models were fitted for training dataset. Logistic regression, Decision Tree, Random Forest, Support Vector Machine, and k-nearest neighbors were selected for this procedure. With logistic regression model, standardization of values of the features was conducted before fitting the model. After applying hyperparameter tuning via Grid Search and Random Search method, for each model, best parameters were found as per the values of the evaluation matrix (accuracy, recall, precision, f1 score). Based on the output given by GridSearchCV, the best classification model could be found. This process was repeated several times with different feature set. When there is a high correlation between two features, one feature was removed and model was trained. As per the output given by feature importance analysis, less important feature was removed and model was trained. However, finally full feature set provided the best model.

After testing model accuracy, Post_processing Function, Score Function, and Prediction Function were defined to develop inference pipeline. As a final step, program was developed to explain the model and prediction given by the best model. For that, LIME (Local Interpretable Model-agnostic Explanation) was used.

4 Results

4.1 Results of Data Exploration

- According to Fig:2 Pie chart (see the python file), there is no significant difference between the applicants who have the research experience and those who do not have the research experience.
- Boxplots for LOR and CGPA indicates that there are some outliers. However, according to the histograms drawn, it seems that there is no any significant impact from them.
- Distribution functions (Density functions) of the quantitative variables drawn by group indicate that:
  - approximately GRE_Score and Toefl_Score are distributed for each priority levels in same way but with different mean values.
  - SOP, LOR, and CGPA have closely distributed functions for each priority level but CGPA has different range than LOR and SOP.
- According to Descriptive Statistics by Group, mean of all the variables increases from less priority level to high priority level but there is no any significant indication that those variables were affected by research experience.
- Correlation values together with heatmap have a clear indication that our target variable(y_act) is highly correlated with all the feature variables except SOP and LOR. Those two variables have moderate correlation with target variable. Strong correlation between feature variables can be seen for the following:
  - TOEFL Score and GRE Score
  - CGPA and GRE Score
  - CGPA and TOEFL score

4.2 Results of hyperparameter tuning with multiple models

Table 1 shows the values for evaluation matrices for the best model generated by hyperparameter tuning for each model. Best values given for the parameters for each model as follows:

Logistics Regression: Tuned Model Parameters: {'C': 0.5}

Decision Tree: Tuned Model Parameters: {'max_depth': 5, 'min_samples_leaf': 10, 'min_samples_split': 2}
Random Forest: Tuned Model Parameters: {'min_samples_split': 20, 'min_samples_leaf': 5, 'max_depth': 100}
SVM: Tuned Model Parameters: {'C': 1, 'gamma': 1, 'kernel': 'linear'}

Table1. values for evaluation matrices for tunned models

Model Type	f1	Recall	Precision	Accuracy
Logistic Regression	0.678918	0.675000	0.687430	0.675000
Decision Tree	0.629361	0.616667	0.673416	0.616667
Random Forest	0.646133	0.633333	0.670059	0.633333
SVC	0.652687	0.641667	0.669670	0.641667

The best model was selected as logistic regression model with C=0.5 and it was saved as model_lgr1_test.pickle for later use.

4.3 Feature Importance

Table 2 shows the less important and more important feature for each model.

Table2. Feature Importance Analysis

Model Type	Less importance	More importance
Logistic Regression	LOR	GRE Score
Decision Tree	Research Experience and University Rating	CGPA
Random Forest	Research Experience	CGPA
SVC	GRE Score	CGPA

4.4 Inference Pipeline

Inference pipeline which contains Post_processing Function, Score Function, and Prediction Function was developed to give predicted class and predicted probabilities for each class when the input data was given as below:

Input data:

	ID	GRE_SCORE	TOEFL_SCORE	UNIVERSITY_RATING	SOP	LOR	CGPA	RESEARCH	CHANCE_OF_ADMIT	PRIORITY_ADMIT_RANGE	PRIORITY_ADMIT	Y_ACT
183	184	314	110	3	4.0	4.0	8.80	0	0.75	(0.73, 0.83]	2nd_Priority	2
146	147	315	105	3	2.0	2.5	8.48	0	0.75	(0.73, 0.83]	2nd_Priority	2

Output data:

For above dataset, 4 new columns (shown in below) will be added.

pred_prob_class1	pred_prob_class2	pred_prob_class3	pred_prob_class4	pred_class
0.251317	0.374074	0.288266	0.086342	2
0.005333	0.111109	0.329080	0.554477	4

4.5 Explain Model with LIME

When one observed record from test data is given, program will generate a chart as shown in figure 1 in order to explain how the prediction happened. The visual representation given by the output well explains the factors for that particular applicant to be in the predicted priority level.

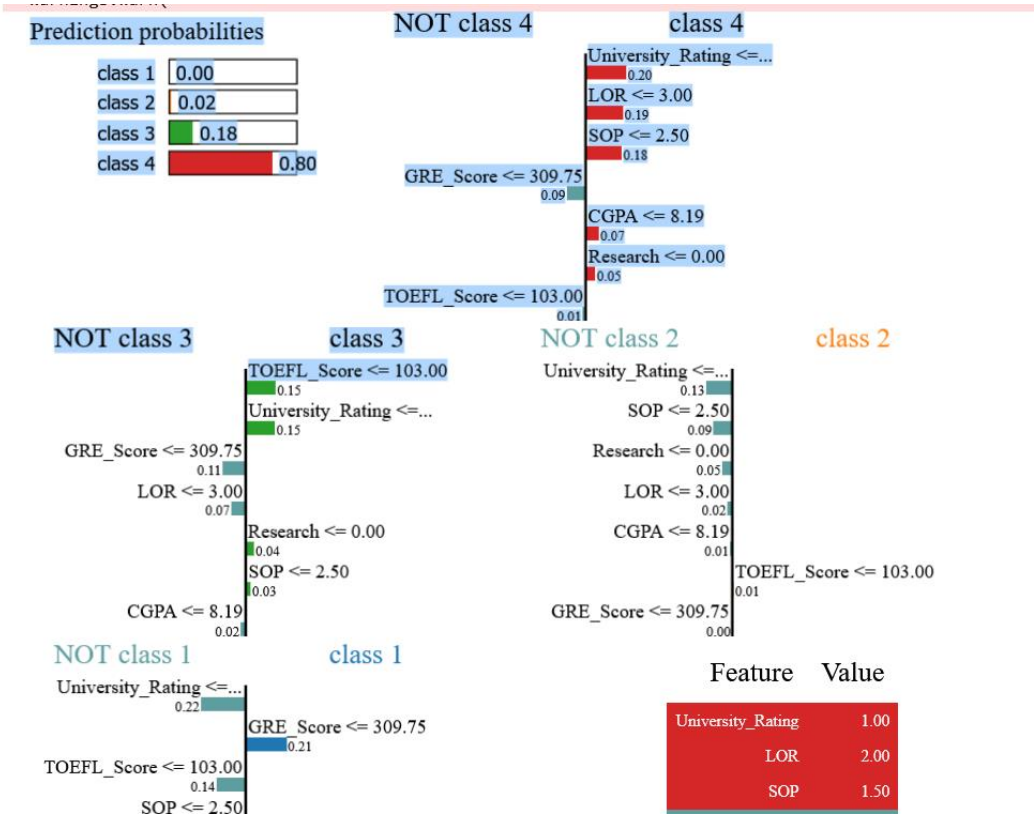

Figure 1. Output given by LIME

5 Conclusion

The best Machine Learning model for this multi class classification problem is Logistics Regression model with parameter C = 0.5. Respective evaluation matrices were given following values:

f1	Recall	Precision	Accuracy
0.678918	0.675000	0.687430	0.675000

6 Discussion

In the process of finding the best model, different feature sets have been used according to the information gained from correlation matrix and feature importance analysis. However, finally best model was given by all the features. Even though this case study indicates that the best model is logistic regression model, both the Receiver Operating characteristic (ROC) curve and Precision Recall curve (PRC) do not provide the well-fitted curves. However further study is needed to find whether there is any other model which will be best fitted with this data.

There is another issue when fitting logistic regression model. Since standardization has already been applied in the program before training the model, fitted logistic regression model always gave 1 for class 1 probability. Since the issue could not be solved, the next best model (SVM) was selected. It also gave 1 always for class 1 probability. I could notice that the issue comes with Standardization. Therefore, standardization was deactivated when training model for all the model except logistic regression model.