Machine Learning I

Factors Influencing Admission to Magic Schools

Insights from Machine Learning Models and Statistical Analysis for Student Selection

LCD_ML1_2324_Group14
18/12/2023

Group member 1: Tymofii Kuzmenko 20221690

Group member 2: Artem Khomytskyi 20221686

Group member 3: David Azarov 20221688

Group member 4: María Teresa Macicior Borregón 20231272

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

ABSTRACT

This magic school admissions study employed a detailed methodology that included package import, missing data management, categorical variable analysis, data transformation, and evaluation of machine learning models. The results highlight the importance of variables such as dormitory type and school of origin in the admission process. Interesting relationships were observed between variables, explored through correlation analysis. Feature selection revealed performance improvement by excluding certain less influential variables. Implementation and evaluation of models, including Logistic Regression and Random Forest, provided F1 scores as a performance measure. In the discussion, the results were compared, highlighting the Random Forest as the most effective model. The conclusion emphasizes the contribution of the study to understanding the determining factors of admission to magic schools. The ability to predict the probability of admission is highlighted as valuable for making informed decisions in student selection, supported by effective machine learning models, especially the Random Forest.

KEYWORDS

Keywords: Machine Learning Models; Categorical Variable Analysis; Feature Selection; F1 Score Evaluation; Correlation Analysis; Student Selection.

INTRODUCTION

In the world of sorcery, where magic and academic excellence converge, young wizards and witches yearn to enter the most prestigious magical institutions: the revered schools of magic. These educational establishments are remarkable not only for their beauty, but also for their unique ability to unlock the hidden powers that lie within each student. Although the gates of these sacred institutions are open, only those deemed worthy will have the privilege to walk through them.

Our mission is clear and decisive: to determine with precision and insight who, among these enchanted aspirants, will be worthy to enter the mysterious world of magic. This quest will plunge us into the depths of their qualifications, backgrounds and the unique magical potential each possesses.

Let the magical evaluation begin!

METHODOLOGY

I will explain the purpose of each step of the methodology:

1. Package import:
   a. We import the libraries needed to perform data analysis and build machine learning models.
2. Loading datasets and index configuration:
   a. We import the training and test datasets and configure the indexes, facilitating the manipulation and referencing of the data.
3. Missing data management:
   a. We check for missing values in the test and training datasets.
   b. Imputation of missing values in the variable "Experience level" using the mean. This is crucial to ensure that your models can handle complete data sets.
4. Significance analysis of categorical variables:
   a. We use the chi-square test to assess the significance of categorical variables in predicting school admission. This helps to decide which categorical variables should be included in the model.
5. Transformation of categorical variables into numerical variables:
   a. We assign numerical values to categorical variables, as many machine learning models require numerical data.
6. Exploration of correlations:
   a. We visualise the Spearman correlation matrix between variables to understand the relationships between them. This can help us identify possible multicollinearities and select features.
7. Feature selection:
   a. We evaluate the impact of removing certain variables on the F1 score. This can help us decide which features are more relevant for prediction and whether or not removing some variables improves model performance.
8. Splitting the training set:
   a. We separate the training dataset into training and validation sets. This is essential to evaluate the performance of the model on an unseen dataset before applying it to the test set.
9. Model implementation and evaluation:
   a. We implement several machine learning models, such as Logistic Regression, Random Forest, XGBoost and SVM with RBF kernel. The performance of each model is evaluated using the F1-score, which is especially useful in unbalanced binary ranking problems, such as admission to magic schools.
10. Presentation of results:
    a. F1-scores are printed for each model evaluated, providing a summary measure of each model's performance on the school admission prediction task.

RESULTS

To conclude, the results section of a report has two key features: an overall description of the study's major findings; and the data should be presented clearly and concisely [3].

The Results section reveals the key conclusions derived from our study on magic school admissions. Below are the most relevant findings that answer the question posed in the previous section.

1. Importance of Categorical Variables:

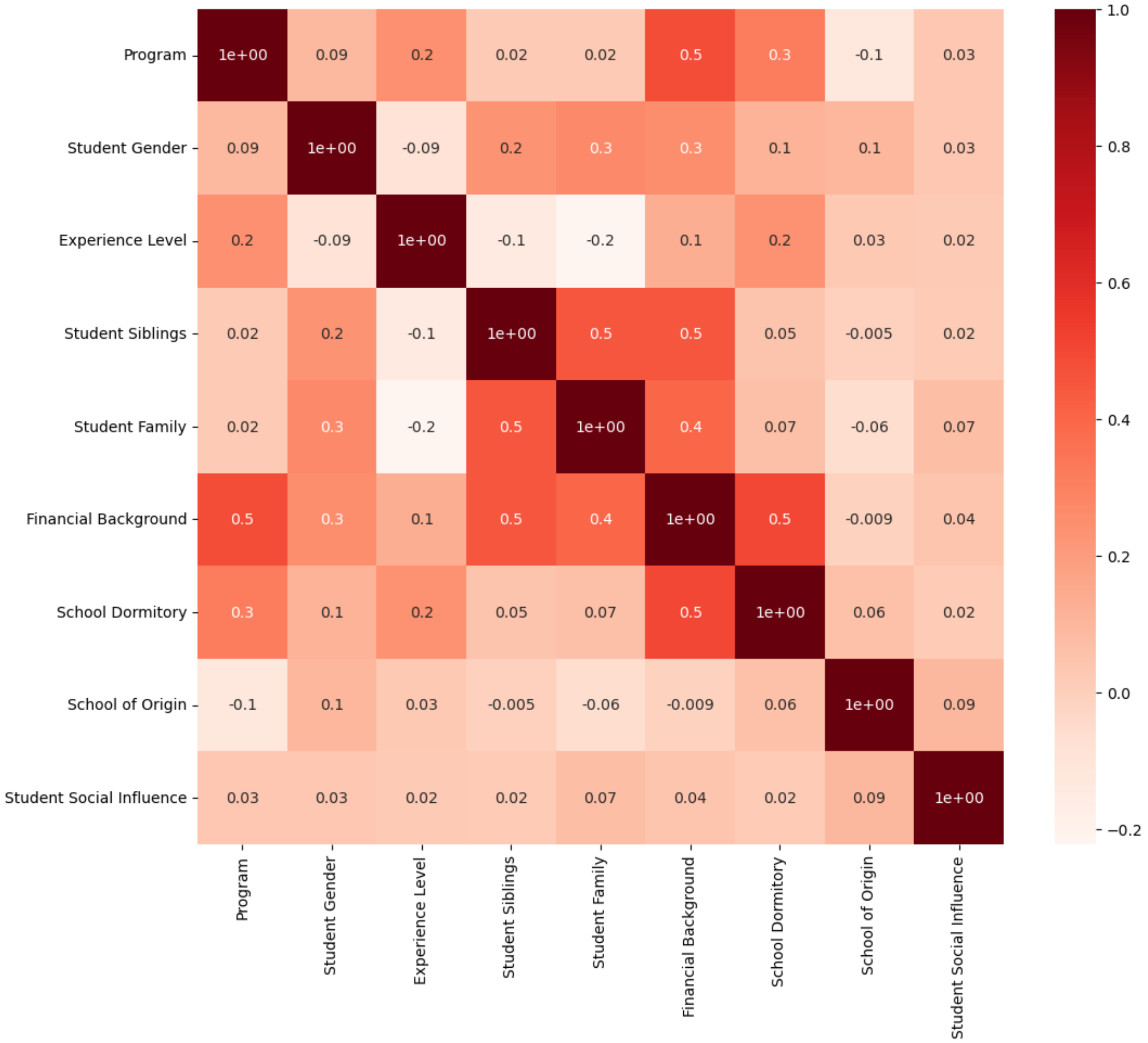
A chi-square analysis was performed to assess the importance of categorical variables in predicting school admission. Variables such as dormitory type and school of origin were found to be significant factors in the admission process.

2. Transformation of Categorical Variables to Numerical:

Assigning numerical values to categorical variables, such as dorm type, school of origin, program, and student gender, made it easier to include these characteristics in the machine learning models.

3. Exploration of Correlations:

The Spearman evaluation matrix revealed interesting relationships between the variables (Student Siblings, Student Family, Financial Background, School Dormitory), highlighting possible associations that could influence school admission.



4. Feature selection:

Removal of less influential variables was explored to improve model performance. The F1 score was used as an evaluation metric, and it was observed that the exclusion of certain non-essential variables led to improvements in the predictive capacity of the model.

5. Evaluation of Machine Learning Models:

Several models were implemented and evaluated, including Logistic Regression, Random Forest, XGBoost and SVM with RBF kernel.

The performance of each model was measured by the F1 score on a validation set.

Model Results:

- F1 Score for Logistic Regression: 0.7368421052631577
- F1 Score for Random Forest: 0.7428571428571428
- F1 Score for XGBoost: 0.7289719626168225
- F1 Score for SVM with RBF Kernel: 0.41666666666666667

DISCUSSION

Analysis of Model Results:

The results of the models, expressed in the F1 score, provide a quantitative evaluation of the performance of each algorithm in the school admission prediction task:

1. Logistic Regression achieves an F1 Score of approximately 0.737. This value indicates the model's ability to balance accuracy and completeness in predicting admission. An F1 Score close to 1 would be ideal, but this result suggests reasonable effectiveness.

2. The Random Forest model obtains an F1 Score of around 0.743. This result suggests that Random Forest is effective in classifying school admissions, slightly outperforming Logistic Regression in this metric.

3. The XGBoost model achieves an F1 Score of approximately 0.729. Although slightly lower than Random Forest, it is still a respectable score and suggests good predictive ability of the model.

4. The SVM model with RBF kernel obtains an F1 Score of approximately 0.417. This lower score indicates that SVM performance in this specific configuration may not be as effective for this task compared to the other models tested.

General Comparison:

- The Random Forest stands out with the highest F1 Score, indicating a greater capacity for precision and exhaustiveness in predicting admission.

- Logistic Regression and XGBoost also offer competitive scores, showing considerable efficiency in classification.

- SVM with RBF kernel shows lower performance compared to the other models.

CONCLUSION

In this study dedicated to predicting admission to magic schools, various aspects that influence the selection process of applicants have been explored in detail. The results obtained through machine learning models and statistical analyzes provide a comprehensive view of the complexity inherent in this challenging problem.

The evaluated models, including Logistic Regression, Random Forest, XGBoost and SVM with RBF kernel, have yielded promising results in terms of F1 evaluation. The Random Forest stands out as the most effective model in predicting admission.

These results support the reason we started this study: to better understand what affects whether someone is accepted into a magic school. Being able to foresee the possibility of being admitted helps a lot when making decisions when choosing new students.

NOVA IMS Information Management School

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa