Machine Learning I

Wizardry School Enrollment

Machine Learning I Report

Group 16

December 2023

Maria Rodrigues 20221938

Patrícia Bezerra 20221907

Rita Silva 20221920

Vasco Capão 20221906

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

ABSTRACT

Within an enchanted realm, the various Wizardry Schools are renowned not only for their grandeur but also for their ability to unlock hidden powers within each student. Hence, every witch and wizard eagerly anticipates being admitted into them. With that being said, the necessity for these schools to delve deep into each candidate's qualifications, backgrounds, and magical potential for a more thorough analysis and selection of worthy candidates is crucial.

In this context, building a predictive model to foresee whether a candidate is accepted into the school becomes of great importance. To achieve this model, the provided data—comprising information about students who have previously been accepted or not into the school—underwent preprocessing, with the relevance of each feature carefully assessed. Furthermore, feature selection was performed on both categorical and numerical features, along with encoding categorical values using methods such as Label Encoding and One-Hot Label Encoder, transforming them into a format interpretable by machine learning models. These steps significantly contributed to improved model performance and accuracy, alongside checks for outliers and missing values.

Subsequently, the Random Forest model, constructed through randomized search, was applied, achieving an 81% accuracy on the validation set. With this performance, it can be concluded that this model would make satisfactory predictions regarding the enrollment of the most deserving candidates in the Wizardry Schools.

In summary, we consider this model highly efficient not only for its enhanced ability to correctly identify students who will be admitted but also for suggesting an overall improved discrimination between admitted and non-admitted students.

KEYWORDS

Machine Learning; Predictive Modelling; Supervised Learning; Random Forest Classifier; MLP Classifier; Randomized Search.

INTRODUCTION

In a mystical realm where aspiring witches and wizards dream of mastering the magical arts, the pivotal first step on their extraordinary journey is gaining admission to one of the esteemed Wizarding Schools. With a multitude of talented and motivated young individuals vying for a coveted spot, the task of discerning those deemed worthy of entering the hallowed halls of wisdom and power becomes a formidable challenge.

In light of this, the process of selection involves a meticulous examination of each candidate's qualifications, background, and unique magical potential. This undertaking is not only extensive but also time-consuming. Recognizing the need for efficiency and precision, our exploration into Machine Learning predictive models presents an opportunity to transform this arduous task into a streamlined and effective process.

Within this context, the predictive model assumes a critical role in identifying students who possess the qualifications to enrol in these prestigious wizardry institutions. Its application ensures that only the most deserving applicants gain entry into the realm of magical education, aligning with the broader goal of this project.

To sum up, the core objective of this project is to incorporate a machine learning model, with careful hyperparameter tuning, into the intricate process of selecting candidates for admission to the Wizardry School.

BACKGROUND

To enhance the efficacy of our model, we ventured beyond the confines of classroom instruction, exploring techniques not explicitly covered in our coursework. This section provides an in-depth exploration of these methodologies, offering insights to enhance comprehension of our approach.

WINSORIZATION

Recognizing the prevalence of outliers in nearly every dataset, it becomes imperative to manage their impact, whether through removal or imputation with an appropriate value. In pursuit of this objective, we turned to Winsorization, a statistical technique meticulously crafted to mitigate the influence of outliers. This approach entails substituting extreme values with less extreme counterparts, selecting a specific percentile for outlier imputation rather than relying on conventional measures like mean, median, mode, minimum, or maximum values. Through Winsorization, we seek to enhance the robustness of our data by addressing the challenges posed by outliers in a nuanced and effective manner.[1][2][3]

ONE-HOT LABEL ENCODING

In machine learning projects, the adoption of techniques for managing categorical variables is commonplace, aiming to enhance interpretability and ultimately improve the accuracy and performance of the final model.

One of the chosen techniques, One-Hot Encoding, involves the transformation of categorical variables into a numerical format before fitting and training the model. This process results in a numeric vector with a length equal to the number of categories in the dataset. For a given observation belonging to category 'x', all values in the vector, except for the 'x' category, will be 0, which is represented by a 1.

However, it is important to acknowledge that One-Hot Encoding comes with certain drawbacks, with the creation of Dummy Variables being a primary concern. Introducing new columns for each category may lead to a high correlation among attributes, potentially resulting in multicollinearity—a topic that will be further explored in subsequent sections of this report. To mitigate this issue, it is common practice to drop one dummy variable, effectively eliminating a redundant column.

While One-Hot Encoding is a potent method for handling categorical variables, it is not without trade-offs. These include increased dimensionality, sparsity, and susceptibility to overfitting. Consequently, it is imperative to carefully weigh these trade-offs and consider alternative methods, such as binary encoding, based on the specific characteristics of the dataset and the requirements of the model.[4][5]

BINARY LABEL ENCODING

As mentioned in the last topic, Binary Label Encoding was used to develop the final model. Binary Label Encoding is a label transformation technique commonly employed in machine learning projects, particularly in cases of binary classification. In this approach, classes, such as "female" and "male", are represented by binary numbers, typically 0 or 1. Is an effective choice for binary classification problems, providing simplicity in class representation, computational efficiency, and direct compatibility with algorithms designed for binary labels.[6]

RANDOMIZEDSEARCHCV

While GridSearchCV exhaustively explores all possible hyperparameter values and combinations, we initially chose RandomizedSearchCV for our model creation process due to the abundance of hyperparameters.

To elaborate, RandomizedSearchCV efficiently explores a diverse subset of hyperparameter configurations, enhancing computational and time efficiency, especially in extensive search spaces. For example, we specify the parameters to be tested and the desired number of iterations. The algorithm then randomly assesses these parameters for the specified number of iterations.

Therefore, RandomizedSearchCV strikes a balance between exploration and exploitation, potentially identifying superior configurations more rapidly compared to a comprehensive GridSearchCV. [7]

METHODOLOGY

To reach the desired results in this project, certain steps/procedures were made, mainly: data integration, data exploration and modification, feature analysis/selection and, model selection. On top of the latter, we also have the Kaggle submission, which provides a means to assess the model's proficiency, regardless of the accuracy of the validation dataset predictions.

Notably, the entire project was implemented using the Python programming language.

DATA INTEGRATION

For the development of the project, it is first required to import the needed Python libraries and packages, as well as the train dataset provided, which contains information of former students regarding whether they were accepted into the wizardry school. Additionally, the test dataset is also imported.

DATA EXPLORATION AND MODIFICATION

The following tasks of data exploration and modification were made to both the training and test dataset samples.

Firstly, a copy of the training dataset was created to avoid any accidental changes to the original data. Secondly, the “Student ID” was set to an index column in both datasets. Lastly, an understanding of the characteristics and quality of the dataset was crucial to better navigate it and develop the project, which was done by doing an exploratory data analysis (EDA).

FEATURE ANALYSIS

To initialize this step, we compared the target feature “Admitted in School” with the other features and plotted them. Analysing these graphics, we concluded that the feature “Admitted in School” scarcely impacts the other features, since the results when “Admitted in School” is 0 and when it is 1 have minimal changes.

As a starting point, a separation of the data into distinct training and validation sets must be made, to ensure that information from the testing set doesn't inadvertently influence the model during training, meaning that it prevents data leakage, maintaining the integrity of our analysis.

The training data was divided into the predictors (X) and the outcome variable (y) and the observations of the predictors (X) were separated into training and validation sets (X_train, y_train, X_val, y_val) – considering as attributes the test size 20%, and random state equal to 42. All the steps made here on forth that make any modifications to the dataset are always made to both the training and validation sets.

Shifting our analysis’s focus to the exploration of outliers in the dataset using boxplots, provided a visual representation of data distribution, allowing an easy identification of any observations that fall significantly outside the norm. The features that exhibited outliers were Experience Level, Student Siblings, Student Family and Financial Background. However, we specifically addressed and considered only Experience Level and Financial Background. To treat them we used Winsorization.

Transitioning to the strategy for handling missing values, two variables within the dataset presented this issue: "Experience Level" and "School Dormitory." The latter was dropped, while for the former, we employed K-Nearest Neighbours Imputation with four neighbours. This method is esteemed for its

versatility in managing diverse data types, flexibility in accommodating local patterns, and non-parametric attributes. Ultimately, this meticulous approach ensures the datasets are devoid of any missing values.

The subsequent step in the pre-processing phase involves feature selection, wherein we segregate numerical and categorical features. Although the typical sequence involves scaling the data, we have deferred this step for the following reasons:

- Lasso Tolerance to Scaling: Lasso exhibits insensitivity to feature scaling, rendering it well-suited for our dataset without the need for explicit scaling.
- Future Model Optimization: Our strategy includes exploring diverse scaling techniques tailored to each model after the feature selection phase. This approach is geared towards optimizing overall performance.

This methodology aligns seamlessly with the inherent simplicity of our dataset, ensuring a streamlined feature selection process without unnecessary preprocessing complexities.

In the context of numerical data, we utilized the LassoCV technique as the initial step to identify the most significant features. Subsequently, we complemented this process by employing Recursive Feature Elimination (RFE). This combined approach allowed us to harness the strengths of both methods, with Lasso providing an initial filter for feature importance, and RFE further refining the selection to ensure a comprehensive and nuanced understanding of the numerical features. Meanwhile, for the categorical data, chi-squared was concurrently applied for effective feature selection.

The subsequent step involves handling categorical variables using encoding techniques, such as Label Encoding for binary features like "Student Gender" and One-Hot Encoding for other categorical features. Through these encoding methods, our goal is to transform categorical information into a machine-readable format.

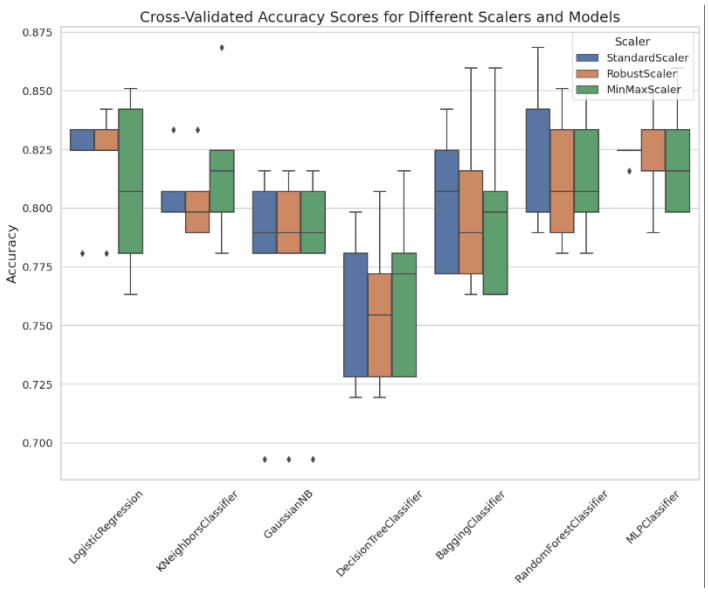
Following the encoding procedures, we delve into understanding feature correlations. This insightful analysis provides a profound understanding of how different variables interrelate. Notably, high correlation between features may suggest redundancy, potentially leading to overfitting, while low correlation implies independence. This understanding is fundamental for ensuring model interpretability, efficiency, and mitigating multicollinearity issues. As a practical outcome of this analysis, the elimination of one of the correlated features serves to streamline our dataset, ultimately contributing to the enhancement of model efficiency. This streamlined dataset, enriched with meaningful features, sets the stage for more accurate and interpretable machine learning outcomes.

MODEL SELECTION

This project's main objective is to determine which machine learning method is most appropriate for the given situation. Hence, the first step should be understanding the type of models that could be applied to this classification problem of predicting the values of the target variable of the dataset. The models and scalers tested to find which combination had a better score were:

- Logistic Regression, K-Neighbors Classifier, GaussianNB, Decision Tree Classifier, Bagging Classifier, Random Forest Classifier and MLP Classifier.
- Standard Scaler, Robust Scaler and MinMax Scaler.

Below, we can identify the Random Forest and MLP classifiers as the top performers, with default parameters, and StandardScaler as the best scaler.



We are now refining the hyperparameters of our two best models by systematically exploring various values for each parameter. This process involves fine-tuning each model's configuration to maximize performance. After identifying the best values for each hyperparameter individually, we employ a Randomized Search to further optimize and find the most effective combination. Between Grid Search and Randomized Search, Grid Search systematically explores all combinations within a specified hyperparameter space, while Randomized Search introduces an element of randomness by selecting configurations randomly. This approach strikes a balance between exploration and exploitation, potentially uncovering optimal configurations more efficiently than the exhaustive Grid Search. Consequently, Randomized Search is a less complex yet effective method for hyperparameter optimization.

Following an initial implementation of Randomized Search, both classifiers demonstrated comparable performances. However, there is a possibility of overfitting, indicated by a difference between training and validation set performances. To address this potential concern and further enhance the models' performance, the Grid Search technique was employed for the Random Forest and MLP, exploring a variety of values for their hyperparameters.

Afterwards, the similarity between the training and validation metrics, on the MLP Classifier with Grid Search, suggested the model is likely not overfitting, making it a favourable candidate for consideration.

To achieve the optimal and conclusive predictive model, we thoroughly compared the performance metrics of, what we considered, the two key models developed during our experimentation phase: the Random Forest Classifier, constructed using Randomized Search, and the MLP Classifier fine-tuned through Grid Search. After meticulous evaluation, the preferred choice is the Random Forest Classifier. This decision is grounded in its superior alignment with the project's objectives, presenting a well-balanced trade-off between accuracy, precision, and overall discriminatory power.

RESULTS

In terms of data exploration, the following observations were registered:

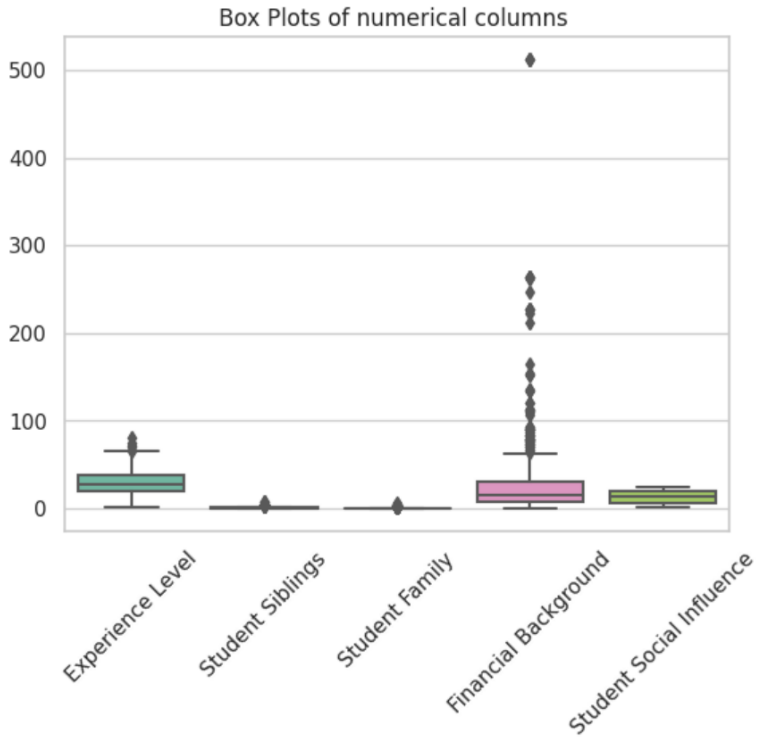
Upon the data exploring (Figure 1, appendix):

- The sample has 713 rows and 11 columns;
- An almost even distribution of numerical and categorical variables.
- “Experience Level” has 146 missing values and “School Dormitory” 560 (to be treated in Feature Analysis);

In the context of descriptive statistics (Figure 2, appendix):

- The mean of "Experience Level" is 29.89, with a standard deviation of 14.60.
- "Financial Background" shows an average of 31.33, but with a significant standard deviation of 50.90, suggesting considerable variability in the numerical values.

For the treatment of outliers and missing values, outliers in the numerical variables “Experience Level”, “Student Siblings”, “Student Family” and “Financial Background” were found. To dispute whether these were actual outliers, we observed their unique values. Even though all these features have suggested outliers, upon closer examination of their unique values, one can infer the existence of outliers solemnly in “Experience Level” and “Financial Background”.



The missing values found in “Experience Level” and “School Dormitory” were dealt with by filling them in with KNNImputer and removing the variable from the dataset, respectively, owning to the fact that in the former only 146 out of the 713 values were missing and the latter had approximately 80% of the total values in that specific column missing.

In feature selection our finding led us to:

- Keeping all numerical features, except “Favourite Study Element” (Figure 3, appendix);
- Only discarding the categorical feature “Student Family” (Figure 4, appendix).

The Binary Label Encoding was used in the “Student Gender” column, where “1” represents Male and “0” Female, whilst the One-Hot Encoding converted each category of the referred feature to a new column with binary values indicating the presence or absence of the category in each observation.

In the assessment of performing the spearman correlation between the variables, a substantial correlation of -0.77 was observed between two features: "School of Origin_Eldertree Enclave" and "School of Origin_Eldertree Academy" (Figure 5, appendix). Since such high correlation indicates a strong linear relationship between these features, and retaining both might introduce redundancy to our model, the former was dropped.

After the data treatment, the final features (Table 1) included in our dataset, and subsequently in the predictive model, are represented on the tables below, as well as their scores.

“Student Gender”	“Student Social Influence”
“Experience Level”	“Program_Sorcery School”
“Student Siblings”	“Program_Witchcraft Institute”
“Financial Background”	“School of Origin_Mystic Academy”

Table 1

Model	Accuracy (f1-score) in TD	Accuracy (f1-score) in VD
RandomForestClassifier, with RandomForestSearch	0.89	0.81
MLPClassifier, with RandomForestSearch	0.84	0.82
RandomForestClassifier, with GridSearch	0.97	0.83
MLPClassifier, with GridSearch	0.82	0.79

Table 2

Upon scrutinizing the various models developed during our experimentation phase, the two final contenders are: the Random Forest, constructed through randomized search, and the MLP, fine-tuned using grid search.

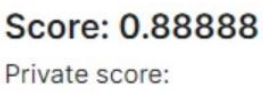
After examining the metrics for both models in the context under study, we have decided to proceed with the Random Forest Classifier as our final model. The decision is based on the following key observations:

- The Random Forest Classifier demonstrates superior performance in terms of accuracy, precision, recall, and AUC compared to the MLP Classifier.
- It achieves a higher accuracy and precision for predicting admission (class 1), indicating a better ability to correctly identify students who will be admitted to the school.
- The higher AUC also suggests improved overall discrimination between admitted and non-admitted students.

Here we can see the observations stated above:

```
Random Forest Classifier:
              precision    recall  f1-score   support

           0       0.82      0.91      0.86        92
           1       0.80      0.63      0.70        51

    accuracy                           0.81       143
   macro avg       0.81      0.77      0.78       143
weighted avg       0.81      0.81      0.81       143

Accuracy: 0.8111888111888111
AUC: 0.8810741687979541

MLP Classifier:
              precision    recall  f1-score   support

           0       0.82      0.86      0.84        92
           1       0.72      0.67      0.69        51

    accuracy                           0.79       143
   macro avg       0.77      0.76      0.77       143
weighted avg       0.79      0.79      0.79       143

Accuracy: 0.7902097902097902
AUC: 0.8503836317135549
```

Ultimately, resulted with an accuracy of 0.89 and 0.81 for the training and validation datasets, and even though it appears to show quite a bit of an overfitting, it was still the model that showed a higher score in terms of the Kaggle submission.



DISCUSSION

Within this discussion section, we delve into the outcomes and implications of the project in the realm of machine learning applications for educational admissions, weaving a narrative that highlights the significance of each aspect. The project's foremost achievement lies in its contribution to overcoming the time-consuming challenges of admissions processes in Wizardry Schools. By introducing a machine learning model, the project takes a significant step towards streamlining candidate selection. This aligns with the overarching trend of integrating technology into educational workflows, promising enhanced efficiency.

An integral component ensuring the reliability and accuracy of the predictive model is the approach to data preprocessing. The utilization of Winsorization for outlier management and K-Nearest Neighbors Imputation for handling missing values underscores the importance of thorough data cleaning.

Moving forward, the project adopts a comprehensive strategy for feature selection, incorporating techniques such as LassoCV, Recursive Feature Elimination, and chi-squared methods. This not only enhances model performance but also provides insights into the key factors influencing admissions outcomes.

The exploration of different machine learning models and a comparison of Randomized Search and Grid Search methods offer valuable insights into effective model selection strategies. Understanding the delicate balance between exploration and exploitation in hyperparameter optimization is essential for developing models that generalize well to unseen data.

Addressing the common challenge of overfitting, the project's systematic handling of potential overfitting in the Random Forest model reflects a thoughtful approach to model development. This contributes to the ongoing discourse on balancing model complexity and performance.

The emphasis on the practical implications of the predictive model for admissions underscores the broader impact of machine learning in educational decision-making. Aligning with the narrative of leveraging technology to improve decision quality, efficiency, and overall effectiveness in educational institutions, the project reinforces the practicality of its findings.

Adding a practical dimension, the inclusion of Kaggle submission results demonstrates the project's applicability in real-world scenarios. Benchmarking models in diverse contexts is crucial for assessing their generalizability and performance beyond the training and validation datasets.

Looking ahead, the discussion highlights potential areas for future research, including further optimization of the predictive model, exploration of alternative algorithms, and investigations into the generalizability of the model to different educational contexts or datasets. These considerations contribute to the ongoing evolution of machine learning applications in education.

CONCLUSION

Creating a prediction model for a classification problem, more precisely, a model that can predict if a student would be admitted to this school or not, was our aim. After careful consideration of the evaluation metrics for the models we created, we have decided that the best solution to the problem presented to us was using the Random Forest Classifier as our final model. This model achieved an accuracy of 81%, which means that 81% of the students in the test set had their admission status accurately predicted by it. Also, this model has a precision of 82% for students not admitted to school, and for the students admitted, a precision of 80%.

In general, the model exhibited satisfactory performance and demonstrated effective generalization to the test set. Nevertheless, there exists an opportunity for enhancement. One plausible direction for further refinement involves experimenting with diverse feature selection methods or fine-tuning the hyperparameters of the model. To sum up, the Random Forest Classifier emerges as a promising candidate for predicting if a student is admitted into the school or not.

REFERENCES

[1] Zulmuthi, H. (2022, May 12). Out with the outliers. Medium. https://medium.com/@haniszulaikha/out-with-the-outliers-fc39c2bcacd7#:~:text=Winsorization%20is%20essentially%20similar%20to,bottom%205%25%20of%20the%20data.

[2] Author links open overlay panelJoe H. Sullivan a, a, b, c, AbstractRecent research in leading business journals has varied widely in how statistical outliers are identified and handled; many techniques were reported. But most articles with empirical data have not mentioned outliers; many others simply referred to, Abbey, J. D., Frazer, L., Grice, J. S., MacKenzie, S. B., Oppenheimer, D. M., Aguinis, H., Andrews, D. F., Andrich, D., Barnett, V., Casella, G., Chambers, R., Chatterjee, S., Cohen, J., Eisend, M., … Kerlinger, F. N. (2021, May 4). So many ways for assessing outliers: What really works and does it matter?. Journal of Business Research. https://www.sciencedirect.com/science/article/abs/pii/S0148296321002290

[3] Demir, S., & Sahin, E. K. (2023, July 24). Application of state-of-the-art machine learning algorithms for slope stability prediction by handling outliers of the dataset - earth science informatics. SpringerLink. https://link.springer.com/article/10.1007/s12145-023-01059-8

[4] Jodha, R. (2023, February 21). One hot encoding. Scaler Topics. https://www.scaler.com/topics/data-science/one-hot-encoding/

[5] GeeksforGeeks. (2023, April 18). One hot encoding in machine learning. GeeksforGeeks. https://www.geeksforgeeks.org/ml-one-hot-encoding-of-datasets-in-python/

[6] Menear, K. (2023, February 8). Comparing label encoding, one-hot encoding, and binary encoding for handling categorical variables... Medium. https://medium.com/@kevin.menear/comparing-label-encoding-one-hot-encoding-and-binary-encoding-for-handling-categorical-variables-933544ccbd02#:~:text=Binary%20encoding%20is%20a%20technique,the%20category%20in%20the%20data

[7] Saeed, N. (2023, February 4). What is RANDOMIZEDSEARCHCV in machine learning. LinkedIn. https://www.linkedin.com/pulse/what-randomizedsearchcv-machine-learning-noor-saeed/

APPENDIX

```
DATA TYPES of Dataset ----------
Program                      object
Student Gender               object
Experience Level            float64
Student Siblings              int64
Student Family                int64
Financial Background        float64
School Dormitory             object
School of Origin             object
Student Social Influence      int64
Favourite Study Element      object
Admitted in School            int64
dtype: object
-----------------------------------
MISSING VALUES in Dataset ------------
Program                        0
Student Gender                 0
Experience Level             146
Student Siblings               0
Student Family                 0
Financial Background           0
School Dormitory             560
School of Origin               0
Student Social Influence       0
Favourite Study Element        0
Admitted in School             0
dtype: int64
```

Figure 1

```
DESCRIPTIVE Info about Dataset ---------------------------------------------------
                          count unique              top freq
Program                     713      3    Sorcery School  391
Student Gender              713      2              male  469
School Dormitory            153      6  Mystical Chamber   51
School of Origin            713      3    Mystic Academy  524
Favourite Study Element     713      4             Earth  184
-----------------------------------------------------------------------------------
                          count       mean        std   min     25%   50%  \
Experience Level          567.0  29.890952  14.599272  0.42  20.750  28.0
Student Siblings          713.0   0.521739   1.057287  0.00   0.000   0.0
Student Family            713.0   0.354839   0.770985  0.00   0.000   0.0
Financial Background      713.0  31.327238  50.903034  0.00   7.925  14.4
Student Social Influence  713.0  12.719495   6.949648  1.00   7.000  13.0
Admitted in School        713.0   0.353436   0.478372  0.00   0.000   0.0

                           75%       max
Experience Level          39.0   80.0000
Student Siblings           1.0    8.0000
Student Family             0.0    6.0000
Financial Background      30.0  512.3292
Student Social Influence  19.0   24.0000
Admitted in School         1.0    1.0000
```

Figure 2

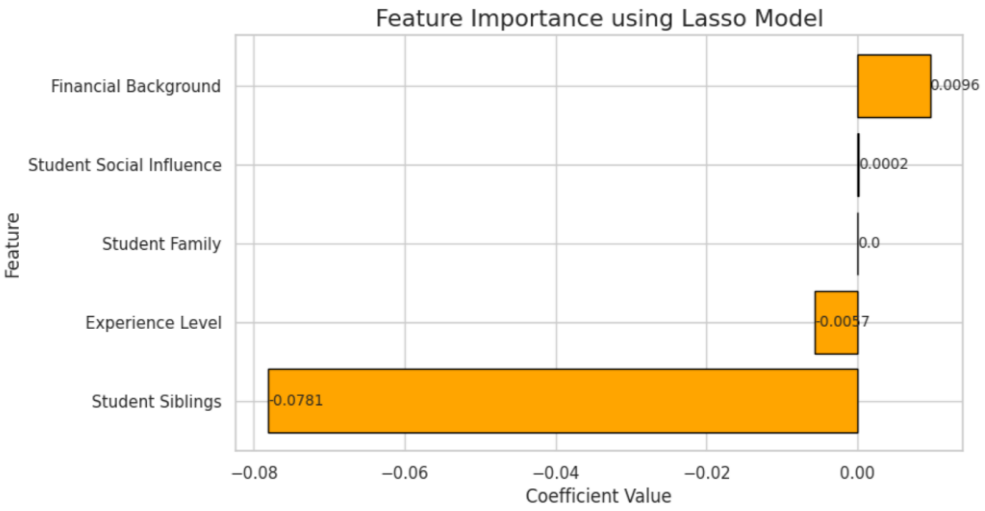


Figure 3

```
Program is IMPORTANT for Prediction
Student Gender is IMPORTANT for Prediction
School of Origin is IMPORTANT for Prediction
Favourite Study Element is NOT an important predictor. (Discard Favourite Study Element from the model)
```

Figure 4

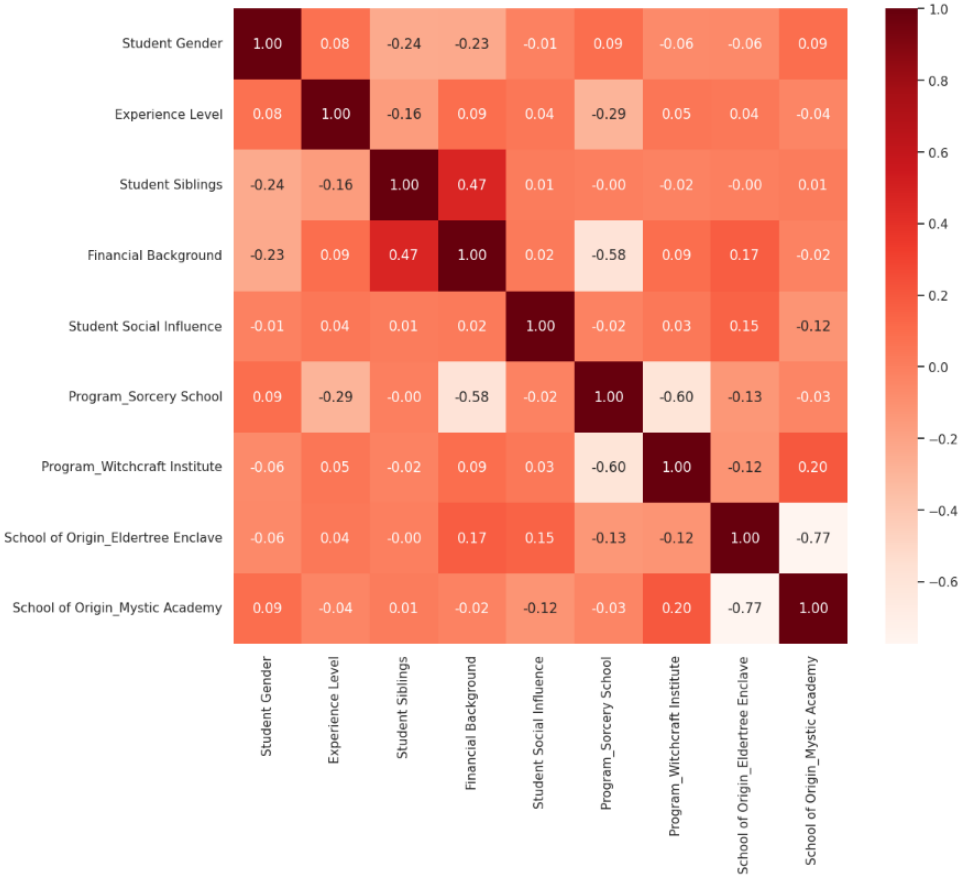

Student Gender
Experience Level
Student Siblings
Financial Background
Student Social Influence
Program_Sorcery School
Program_Witchcraft Institute
School of Origin_Eldertree Enclave
School of Origin_Mystic Academy
1.00 0.08 -0.24 -0.23 -0.01 0.09 -0.06 -0.06 0.09
0.08 1.00 -0.16 0.09 0.04 -0.29 0.05 0.04 -0.04
-0.24 -0.16 1.00 0.47 0.01 -0.00 -0.02 -0.00 0.01
-0.23 0.09 0.47 1.00 0.02 -0.58 0.09 0.17 -0.02
-0.01 0.04 0.01 0.02 1.00 -0.02 0.03 0.15 -0.12
0.09 -0.29 -0.00 -0.58 -0.02 1.00 -0.60 -0.13 -0.03
-0.06 0.05 -0.02 0.09 0.03 -0.60 1.00 -0.12 0.20
-0.06 0.04 -0.00 0.17 0.15 -0.13 -0.12 1.00 -0.77
0.09 -0.04 0.01 -0.02 -0.12 -0.03 0.20 -0.77 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Figure 5

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

Universidade Nova de Lisboa