# 

*Photo:* [*Freekpik; rawpixel.com*](https://www.freepik.com/free-photos-vectors/school)

**HDSC Winter ’22 Capstone Project Presentation: China Scholarship(Economy & Education)**

A project by Team Prophets

Many secured scholarships, but they end up losing them. They do not know the total expenses those scholarships demand. The goal of this project by team Prophets is to help predict the Total Expenses for a scholarship using this [dataset](https://www.kaggle.com/mcmuralishclint96/china-scholarship-data-may-2019).

To follow along, the project was [deployed on Heroku](https://predict-scholarship-fee-app.herokuapp.com/) and the code can be found on [GitHub](https://github.com/drissdunn/Hamoye-Prophet-Team).

Every machine learning project follows a workflow and phases, depending on the project specifications. Generally, these phases were followed by the team to achieve this result. The phases include:

* Data collection
* Exploratory Data Analysis
* Data Preparation(Feature Engineering)
* Modeling
* Deployment.

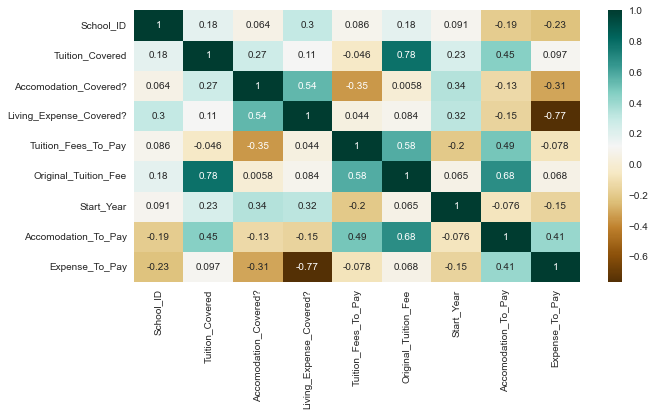
## DATA COLLECTION

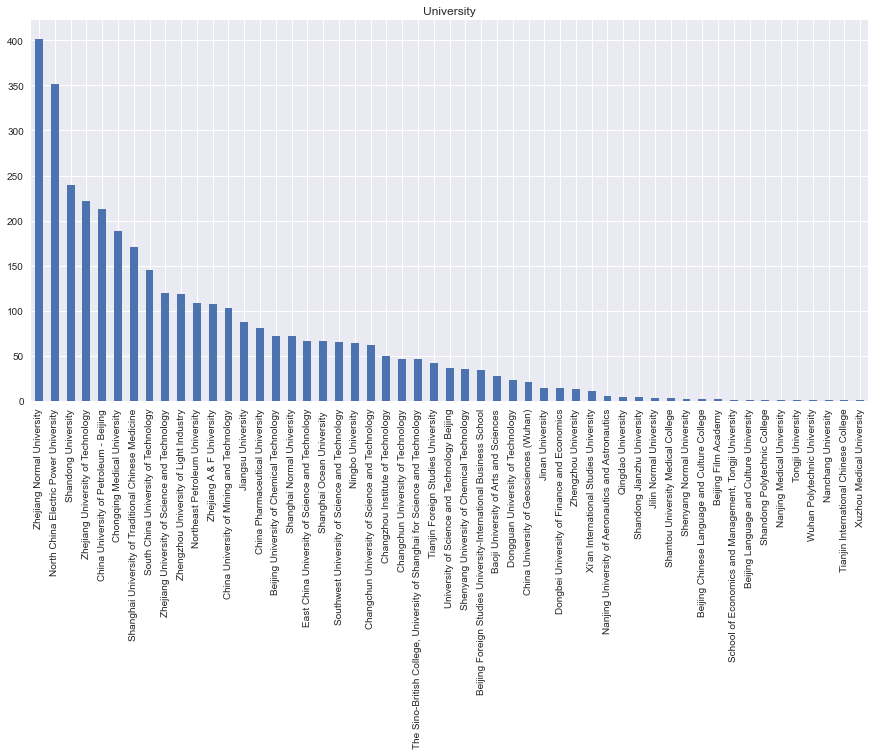
The data used was collected through web-scraping, a process of extracting data from websites, from this [website](https://www.cucas.edu.cn/china_scholarships/). The dataset contains information about the scholarship programs in China as of May 2019. The cleaned data can be easily accessed on [Kaggle](https://www.kaggle.com/datasets/mcmuralishclint96/china-scholarship-data-may-2019).

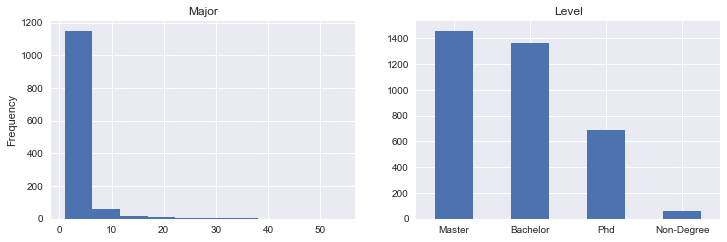
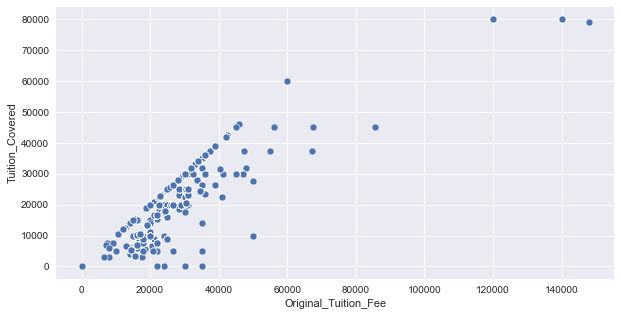
## EXPLORATORY DATA ANALYSIS(EDA)

This phase involves processes of understanding the data using visualizations and statistical tools. It is the process of digging into the data to get intuitions that will be needed in the feature engineering phase. Some of the EDA we carried out include checking for missing values, distribution of the features, multicollinearity of variables and performing univariate and bivariate analysis.

From the analysis, the dataset has 3576 observations with no duplicates and 16 columns. The dataset contains 7 categorical and 9 Numerical data types columns. The *Tuition\_Covered* and *Original\_Tuition\_Fee* columns have 123 null values each.

* *Original\_Tuition\_Fee, Accomodation\_To\_Pay* and *Tuition\_Covered* are highly positively correlated while *Living\_Expense\_Covered* and *Expense\_To\_Pay* are highly negatively correlated as shown in the heatmap diagram.
* The dataset contains 53 universities in 38 districts and the Zhejiang Normal University has the highest frequency.



* Most Universities give scholarships at the Master's level followed by the Bachelor's level and, Computer Science and Technology is the highest major for scholarships granted.
* The most common language is Chinese, which is expected as it is a Chinese dataset. The start month of the scholarship is usually September and the *Accomodation\_Duration* and *Expense\_Duration* are usually monthly.
* The majority of the scholarships do not cover accommodation(*Accomodation\_Covered*) and living expenses (*Living\_Expense\_Covered*).
* The *Tuition\_Fees\_To\_Pay* increases as *Original\_Tuition\_Fee* and *Tuition\_Covered increase.*

## DATA PREPARATION(Feature Engineering)

This phase makes the data ready for modeling. It requires a numeric representation of raw data to be fitted to models. It is also a process of generating more features using domain knowledge to improve the performance of the model.

After much deliberation and consideration, only 8 features were used for modeling.

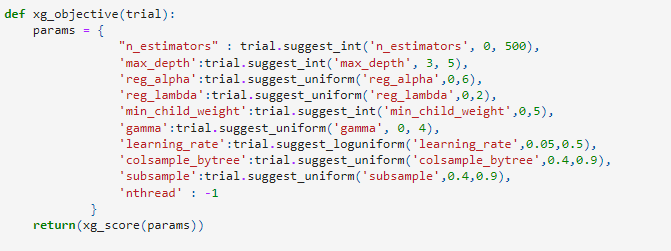
* Numerical features missing values were filled using Mean with *SimpleImputer* function and scaled with *MinMaxScaler().*
* Categorical features were encoded using *OrdinalEncoder* and also scaled.

## MODELING

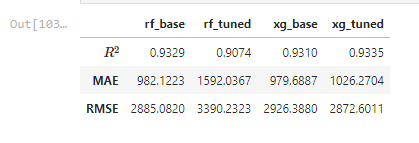
Now the data is well prepared. This phase is the Supervised learning stage where an algorithm learns from data and develops weights and biases that will be used to predict the unseen data.

Random Forest and Xgoost, popular tree-based models with track records of good performance on data and their simplicity were the models used.

Xgboost performed better after Hyperparameter tuning with an R2 score of 0.933. The goal of hyperparameter tuning is to get optimal parameters that will produce the best result and this was done using a package called Optuna.



Other metrics were used to check the performance of the models, as shown in the table below.



## DEPLOYMENT

This is the final phase of a machine learning project, although it is an iterative process. The goal is to implement and make available the model to the end-user.

The team deployed the model as a WebApp:

* with Heroku, a platform as a service (PaaS) that enables developers to build, run, and operate applications entirely in the cloud.
* and the Web page was designed using Html and CSS.

Here is the [link](https://predict-scholarship-fee-app.herokuapp.com/) to the WebApp.

## CONCLUSION

The team has been able to help scholars predict the *Total Expenses* that they should budget for in the scholarships they have been granted.

The performance of the model can also be improved with more Hyperparameter tuning, feature engineering and blending/stacking the results of multiple models. Also, with more data, we can extend the solution to the whole world and help more scholars.