****OPEN SOURCE PROJECT DOCUMENTATION****

****UNIVERSITY ADMISSION****

**An open-source project from a data science internship…**

**INTRODUCTION:**

The constant pursuit of a foreign education causes graduates all around the globe to apply for a master's program in foreign universities. This process is not only challenging but also uncertain as most prospective students do not understand the conditions surrounding their chances of getting admitted into their choice schools. The dataset used in this project can be downloaded from [Kaggle](https://www.kaggle.com/mohansacharya/graduate-admissions).

[](https://www.unsplash.com/)

**PROBLEM STATEMENT:**

This project aims to improve the work previously done by fellow interns. This was achieved by:

* Improving the machine learning algorithm used in predicting the chance of admittance (COA),
* Improving the visualisation and creating a real-time dashboard using Tableau,
* Providing a web interface to guide users on how to feed data to the model and get feedback from it. This webpage is available on [Heroku](https://university-admissions.herokuapp.com).

**TECHNOLOGIES USED:**

Python, Jupyter Notebook, Tableau, Github, HTML5, CSS3, Flask.

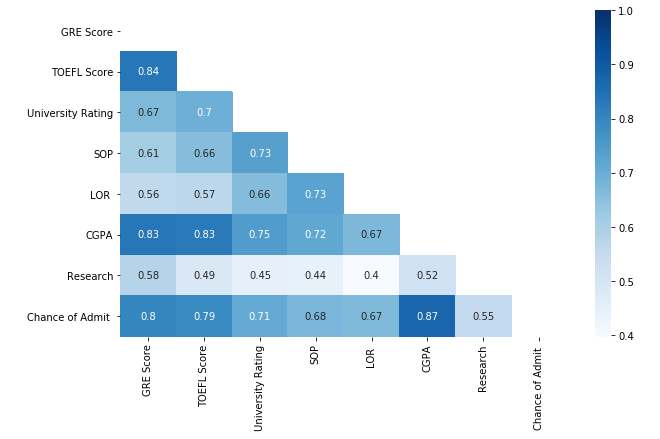
**DESCRIPTION OF THE DATASET:**

The dataset comprises 9 columns and 400 rows with no missing values. The columns include:

* **Serial No** (int 64): Numerical numbering of all data entries. 400 in total.
* **GRE Scor**e (int 64): This represents the GRE Scores of various students, with the maximum possible score being 340.
* **TOEFL Score** (int 64): the maximum possible score is 120.
* **University Rating**: maximum possible rating is 5.
* **SOP** (Statement of Purpose, float 64): The highest possible score for this essay is 5.
* **LOR** (Letter of Recommendation, float 64): with the highest possible score set at 5.
* **CGPA**(Cumulative Grade Point Average, float 64): the highest possible score is 10.
* **Research** (int 64): where research-based applicants are represented by 1 and non-research 0.
* **Chance of Admit** (float 64): The information on this column is based on a scale of 0-1.

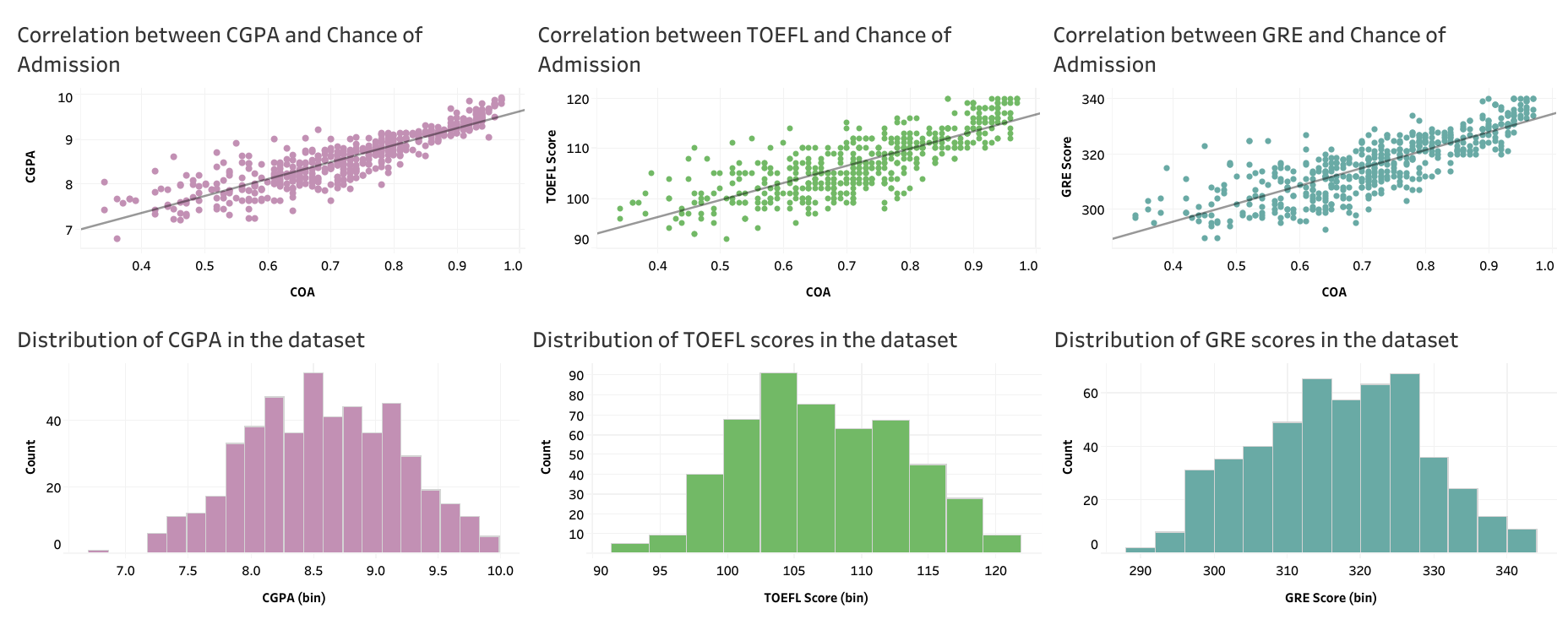
From the above data, all parameters provided are designed to support the prediction of the target variable - **Chance of Admit.**

**EXPLORATORY DATA ANALYSIS:**



From the EDA carried out on the dataset, a heatmap reveals that the feature parameters having the highest correlation to Chance of Admit (the target variable) are: **CGPA**, **GRE Score**, **TOEFL Score**, and **University Rating**, having correlation as high as about **0.8**.

Of these 4 variables, we know that the student can influence the CGPA, GRE, and TOEFL scores. However, the University Rating poses a different problem as it is difficult to explicitly ascertain the rating of a school for each academic year, since the information as to how the ranking was obtained was not documented within the dataset. Also, as a result of this ambiguity, we have decided to train our model on the 3 key features that every student can actually influence, and with that, try to predict the chance of admittance of the student.



The image above shows the correlation and weighted distribution of all of CGPA, TOEFL, and GRE scores. These parameters all show linear relationships with Chance of Admit (COA). From this, we can say that an increase in any of these parameters leads to a significant increase in COA. Also, we observe that these 3 parameters portray a similar behaviour, so we can deduce that a student who has a high CGPA can be expected to produce equally good GRE and TOEFL scores (all things being equal).

We can also observe that the weights of all three parameters are evenly distributed.

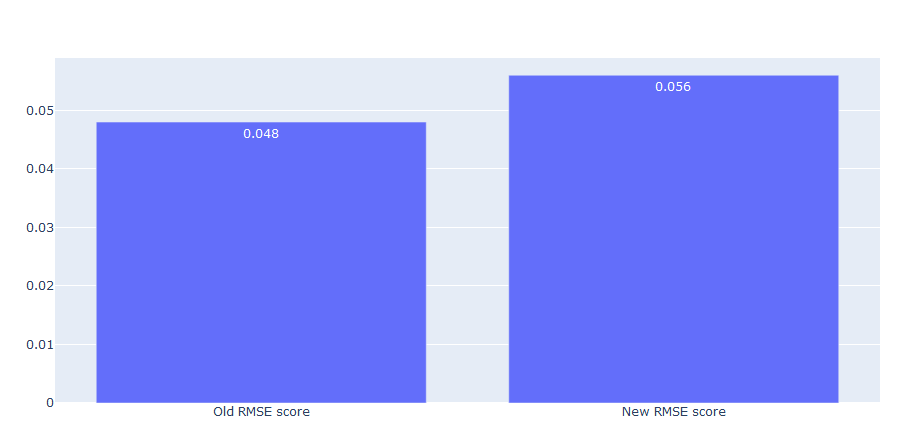
**ASSUMPTION:**

For the purpose of this analysis, we assume that schools require all of CGPA, GRE, and TOEFL scores (among others) to consider a student for a master's program in their university.

**MODELLING:**

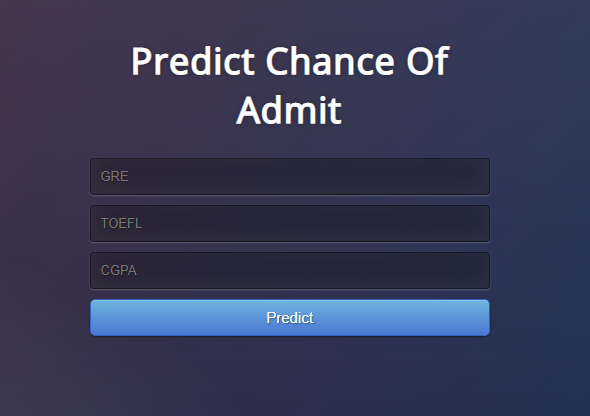
From the EDA carried out earlier, we observed a linear relationship among all 3 feature variables and the target variable. Hence, our model was developed using a simple linear approach. The performance of the model was determined using the root mean squared error (RMSE) metric.

The RMSE score for our model was improved from 0.048 to 0.056 with the line fitness (r2 score) estimated at 0.84 to 2 dp.



**DEPLOYMENT AND PRODUCT TESTING:**

Having conclusively developed a model that best suits the dataset and task at hand, we designed a web-based interface to enable end-users to interact with the model without actually worrying about all the specifics. Here is the[link](https://university-admissions.herokuapp.com/" \t "_blank) to the web page. The web app was also tested using random values of GRE, TOEFL and CGPA, and the results were predicted accordingly.



**RECOMMENDATIONS:**

1. The scope of this project was limited largely due to a small sized dataset (400 data samples), hence the availability of more data from students across the globe would go a long way in increasing the performance of the model.
2. The model is unable to predict a value for COA given the absence of any of the 3 parameters. This is because the current dataset contains no null values, and as such, the model is not optimised to handle missing values.
3. The mock-up design for the web page was not implemented during the course of the project due to time constraint and the lack of a skilled front end web developer to implement the mock-ups.

**BLOCKERS:**

* Agreeing upon the general scope of the project.
* Working within the time stipulated for the project.
* Scheduling meetings at a time suitable with every member of the team.

**CONCLUSION:**

We really hope this project is of some help to someone out there. Please feel free to tag friends who this could be of help to, share, and give this post a thumbs up. Let us know what you think in the comment section below, and of course, take some time out to check out the project in more detail on our [GitHub repository](https://github.com/HamoyeHQ/stage-f-08-university-admissions).

**COLLABORATORS:**

Members of the team that actively contributed to this project are listed below. Their [LinkedIn](http://www.linkedin.com" \t "_blank) profiles are also made available and they would be happy to offer more information as required.

* [Edheba Jerry Efemena](https://linkedin.com/in/jerrylive3) (EDA, Back end web development, Project lead)
* [Sooter Saalu](http://www.linkedin.com/in/ssaalu)  (EDA, Tableau dashboard, Presentation slides)
* [Emmanuel Ogunwede](https://www.linkedin.com/in/emmanuel-ogunwede-665404126/) (EDA, Data modelling)
* [Justus Ilemobayo](http://linkedin.com/in/justusilemobayo) (EDA, Data modelling)
* [Ojo Olawale](https://www.linkedin.com/in/ojo-olawale-a6ab7716b/) (EDA)
* [Temitope Babatola](https://www.linkedin.com/in/temitope-babatola) (EDA, Project documentation, Query analyst)
* [Maybelle Obadoni](https://www.linkedincom/in/mabelobadoni) (EDA, Web-app mock-up designs)
* [Linda Ezeoba](https://www.linkedin.com/in/linda-ezeoba-42a061166/) (EDA, Project documentation)