Design a model to predict the number of international students studying in the UK

I. Introduction

A.

As we all know, British education has a long history, and Britain is the birthplace of modern universities. Currently, the five oldest schools in the world all originate from the UK; enjoy a high reputation. Therefore, the UK has also become a dream destination for many international students. In this report, based on the datasets from HESA (Higher Education Statistics Agency), I will design two machine-learning models by training the model on the dataset to forecast the number of total international students and the main nationalities of the international students which take a huge proportion (China, India, North America, EU) in the next ten years (2021 ~ 2030).

B.

According to the report, "Where Next? What influences the choices international students make?"[1] from the Universities and Colleges Admission Service (UCAS) in May 2022, the survey was conducted on more than 1200 students planning to study abroad in 116 countries and regions. UCAS predicts that one million people could apply to UK universities by 2026 around 27% more than in 2021. Figure 1 shows the prediction of the number of international student applicants. USAS also statistics that the number of Chinese students studying in the UK increased 195%, Indian students increased 222%, and US students increased 165% in the last 10 years as shown in figure 2.

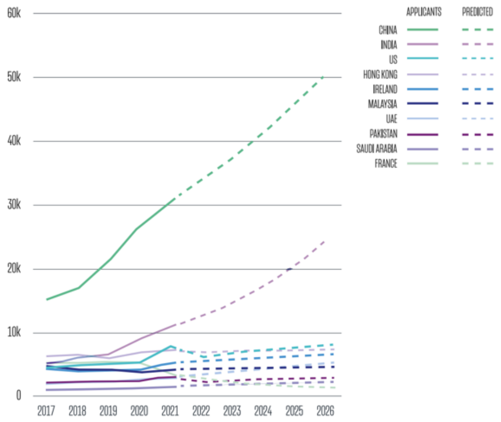
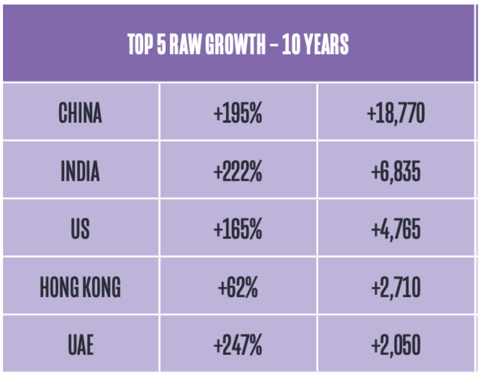


Figure 1: Forecasted UCAS applicants up to 2026 for selected international domiciles [1]



TOP 5 RAW GROWTH – 10 YEARS		
CHINA	+195%	+18,770
INDIA	+222%	+6,835
US	+165%	+4,765
HONG KONG	+62%	+2,710
UAE	+247%	+2,050

Figure 2: The largest increases in applicant numbers10 in the last ten years [1]

The data all over above illustrate that international students are more and more imperative to UK schools; furthermore, international students gradually take up a large proportion of in the UK education system. As a result, an accurate prediction of the number of international students can help the UK government to shape future education policy. Although UCAS has already made the applicant prediction through the survey, in this experiment, I will build two models, polynomial regression and multiple linear regression models and train the dataset from HESA to predict the future ten years of the total number of international students.

II. Methodology and Dataset

C.

Since the composition of international students is a very important key point to predicting the future; therefore, I select different countries as their features and attributes. In my dataset, I mainly focus on those countries or regions which take a huge proportion of all international students, like China, India, North America, and the European Union. All these countries and regions above will have a major impact on the number of future international students in the UK. In addition, using the data from those countries and regions above can make the prediction more accurate.

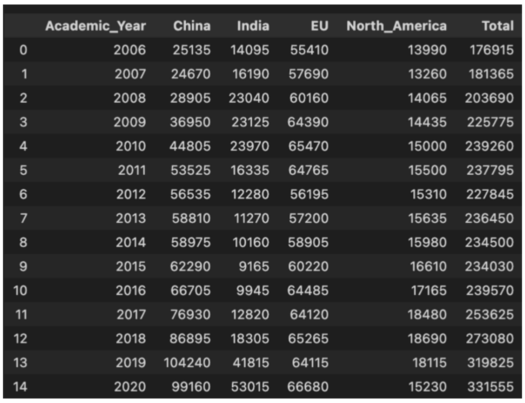
D.

Data exploration phase:

Download all the international student number raw data from the HESA website and divide them into different countries and regions.

Data cleaning phase:

Since there are too many countries in the data and detailed data, which I wouldn't use in my training model; hence, according to our initial goal, data cleaning will be very important during my whole analysis process. After data cleaning, figure 3 shows the number of students studying in the UK from 2006 to 2020 (15 years) in the main countries and regions studying abroad.



	Academic_Year	China	India	EU	North_America	Total
0	2006	25135	14095	55410	13990	176915
1	2007	24670	16190	57690	13260	181365
2	2008	28905	23040	60160	14065	203690
3	2009	36950	23125	64390	14435	225775
4	2010	44805	23970	65470	15000	239260
5	2011	53525	16335	64765	15500	237795
6	2012	56535	12280	56195	15310	227845
7	2013	58810	11270	57200	15635	236450
8	2014	58975	10160	58905	15980	234500
9	2015	62290	9165	60220	16610	234030
10	2016	66705	9945	64485	17165	239570
11	2017	76930	12820	64120	18480	253625
12	2018	86895	18305	65265	18690	273080
13	2019	104240	41815	64115	18115	319825
14	2020	99160	53015	66680	15230	331555

Figure 3: Number of international students from main countries

Modelling and analysis phase:

In my experiment, I follow the requirement of CA to design two machine learning models.

(1) Polynomial regression

In polynomial regression, I trained unitary quadratic linear regression as my model. Using the Normal Equation (Least Square Method)[2] as the loss function to find the lowest cost and build the best regression model, this model can be used to predict the total number and the number of different countries. The regression functions are shown below:

China:

$$Y = 499265008.59 - 501452.83 * X + 125.91 * X^2$$

India:

$$Y = 1862595529.45 - 1851637.09 * X + 460.91 * X^2$$

EU:

$$Y = 73600077.65 - 73537.23 * X + 18.38 * X^2$$

North America:

$$Y = -75915942.87 + 75132.82 * X - 18.59 * X^2$$

Total:

$$Y = 1674061632.42 - 1671488.94 * X + 417.28 * X^2$$

X is the independent variable which represents the academic years, and Y is the dependent variable which represents the predicted number of international students.

(2) Multiple linear regression

In Multiple linear, due to the total number of international students mainly composed of China, India, EU and North America. Figure 4 to figure 7 show the correlation coefficient (PPMCC)[3] between four countries or regions and the total number of international students.

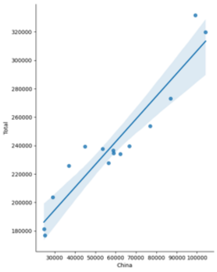


Figure 4: China: 0.9439

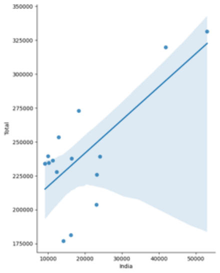
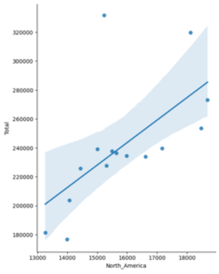


Figure 5: India: 0.7184



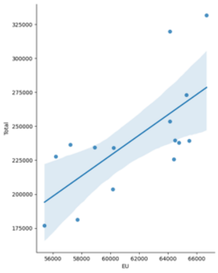


Figure 6: North America: 0.6138   Figure 7: EU: 0.6780

Since all correlation coefficients are bigger than 0.6, which means those factors have a strong correlation with the total number of international students; therefore, I select these four countries or regions as the independent variable to train the model and find the best regression model. In addition, I randomly select 80% of the data as a training dataset and the remaining 20% of data as a testing dataset to make this model have lower bias and variance, as figure 8 shown.

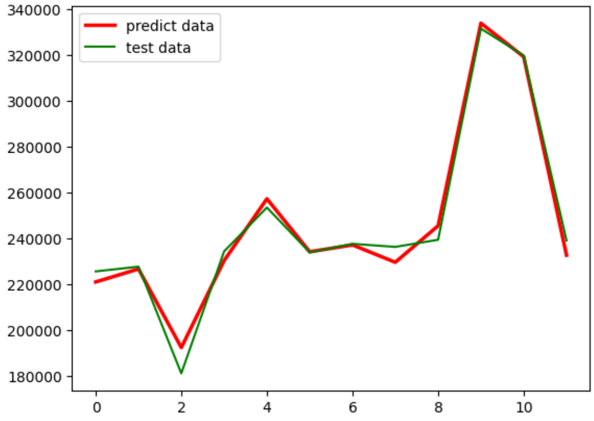


Figure 8

After training the model, we can get the regression function:

$$Y = 107040.47 + 1.46 * X1 + 0.71 * X2 - 3.56 * X3 + 1.48 * X4$$

X1: Student number of China

X2: Student number of India

X3: Student number of North America

X4: Student number of EU

Note: Due to the training data being selected randomly, the parameters of the regression model will change as well.

III. Result

E.

(1) Polynomial regression

In my experiment, I select unitary quadratic linear regression as my first model, and the fitting lines are great.

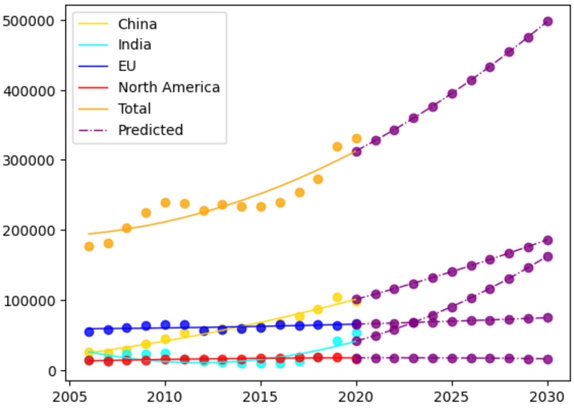
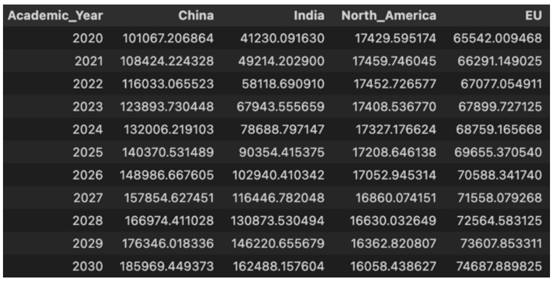


Figure 9: Polynomial regression model

We can observe that figure 9 shows the total number of international students in the UK will continue to increase and especially China and India will drive the increase in the overall number of international students.

(2) Multiple linear regression

The multiple linear regression needs to consider the four main countries and regions; however, since there's no future sub-data for our model to input in this multiple linear regression; hence, I use the prediction data from the polynomial regression model as the input, which is shown as figure 10.



Academic_Year	China	India	North_America	EU
2020	101067.206864	41230.091630	17429.595174	65542.009468
2021	108424.224328	49214.202900	17459.746045	66291.149025
2022	116033.065523	58118.690910	17452.726577	67077.054911
2023	123893.730448	67943.555659	17408.536770	67899.727125
2024	132006.219103	78688.797147	17327.176624	68759.165668
2025	140370.531489	90354.415375	17208.646138	69655.370540
2026	148986.667605	102940.410342	17052.945314	70588.341740
2027	157854.627451	116446.782048	16860.074151	71558.079268
2028	166974.411028	130873.530494	16630.032649	72564.583125
2029	176346.018336	146220.655679	16362.820807	73607.853311
2030	185969.449373	162488.157604	16058.438627	74687.889825

Figure 10: Prediction from polynomial regression

After inputting the data in figure 10, here is the run chart below. Although the original graph should be a multi-dimension graph which can display all variables at the same time. However, I focus on the total number of international students; thus, I use the 2D graph in Figure 11 to show the fitting function and the total number of international student growth in the next ten years.

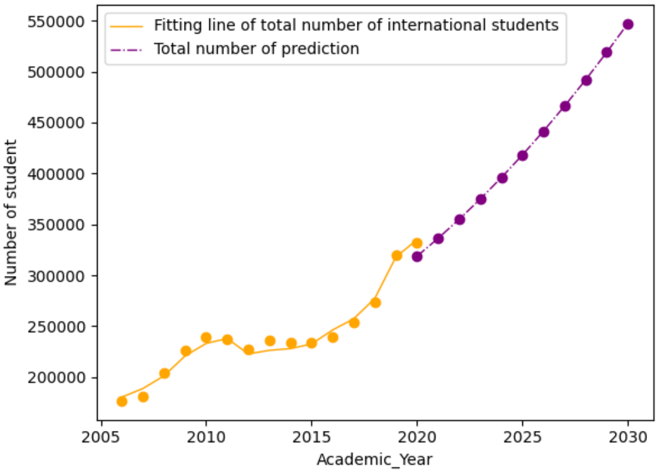


Figure 11: Multiple linear regression model

Furthermore, the multiple linear regression in the predicting phase is more regular, like linear or polynomial regression. The reason I think is that our independent variables are not from the real data but from a predicting function which is polynomial regression. As a result, these two models will basically predict the same consequence and make the multiple linear regression model can perfectly fit the data points in predicting phase.

F.

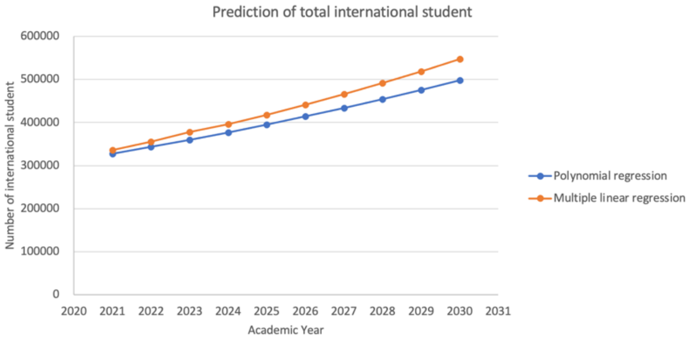


Figure 12: Prediction of two models

Figure 12 shows that both models have a similar prediction for the next ten years; however, the difference between the two predictions will gradually widen over time. In my opinion, the multiple linear regression will be more accurate since this model references more variables and all these variables happen to be important components that mainly affect the total number of international students; the cost of multiple linear regression is lower than the cost of polynomial regression as well. On the other hand, the polynomial regression in my model is a quadratic regression equation in one variable; therefore, the trend of the prediction will always be in one direction, increasing or decreasing, and it won't truly reflect the reliability and accuracy of the data in some special situation like pandemic or war. In addition, the polynomial regression only references one variable which is the academic years but not the student source structure; this issue makes the prediction of the inaccuracy and reliability of prediction is greater than Multiple linear regression.

IV. Discussion

G.

In the next 10 years, we can observe in figure 12 that the number of international students will increase massively. By 2030, the number of international students will be 1.5 times that of 2020, reaching about 500,000 students. China and Indian students will continually take the main part of the total number and drive the growth trend stable, as figure 9 shows.

H.

Looking back on my experiments, there are still some points which can be improved. For example, polynomial regression can not only be quadratic but can be designed to a higher degree. A higher degree can make the fitting line more suitable to the model, and it can let the model not only have a single growth trend like quadratic but have other possibilities to deal with some special periods. I have tried to build the higher degree model at the end, as figure 13 shows, but this prediction is far from my main design model. The prediction is almost three times higher than the prediction of previous models in 2030. I think I still need to modify the parameters and functions of this higher-degree model.

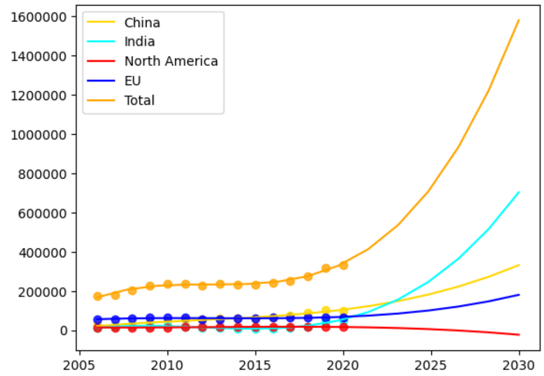


Figure 13: Polynomial regression (Degree = 4)

Besides, the amount of data is only 15 years; hence, it is harder to train my models enough, test the accuracy, and make the models perfect. I need to reference more variables and use the loss function to build a combined model which can reference more variables like countries, ethnicities, economic indexes etc., and predict the total number at the end.

Reference

[1] " Where next? what influences the choices international students make?", *UCAS,* 2022. [online]. Available: https://www.ucas.com/. [2022/11/30]

[2] E. Ostertagová, "Modelling using polynomial regression," *Procedia Engineering,* vol. 48, pp. 500-506, 2012.

[3] M. Tranmer and M. Elliot, "Multiple linear regression," *The Cathie Marsh Centre for Census and Survey Research (CCSR),* vol. 5, no. 5, pp. 1-5, 2008.

\