TU Graz

SCIENCE
PASSION
TECHNOLOGY

Prediction Using Regression on Universities Data Set

Group 14:
David Mihola,
Ronald Infanger,
Thomas Sterner

18. 06. 2023

Outline

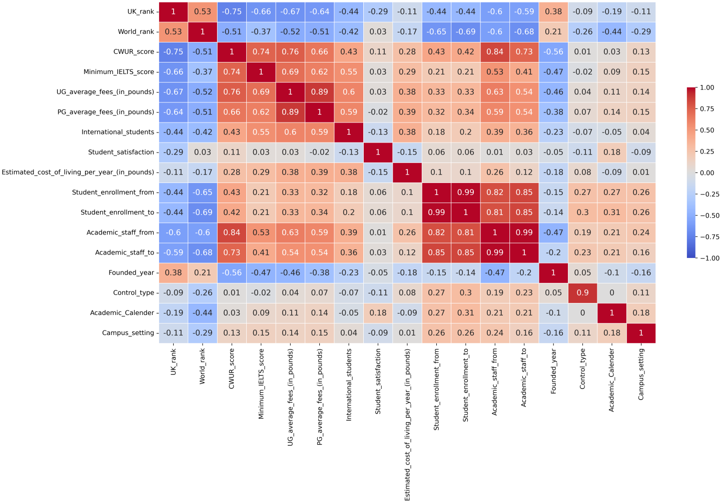
1. Data set analysis,
2. cleaning and pre-processing,
3. training and evaluation,
4. regression models,
5. prediction performance,
6. ensemble.

Data Set Analysis 1

- Universities from the United Kingdom,
- 21 columns and 145 row (131 unique rows).

Idx	Column	Idx	Column
1	University_name	12	Student_satisfaction
2	Region	13	Student_enrollment
3	Founded_year	14	Academic_staff
4	Motto	15	Control_type
5	UK_rank	16	Academic_Calender
6	World_rank	17	Campus_setting
7	CWUR_score	18	Estimated_cost_of_living_per_year_(in_pounds)
8	Minimum_IELTS_score	19	Latitude
9	**UG_average_fees_(in_pounds)**	20	Longitude
10	**PG_average_fees_(in_pounds)**	21	Website
11	International_students		

Data Set Analysis 2



	UK_rank	World_rank	CWUR_score	Minimum_IELTS_score	UG_average_fees_(in_pounds)	PG_average_fees_(in_pounds)	International_students	Student_satisfaction	Estimated_cost_of_living_per_year_(in_pounds)	Student_enrollment_from	Student_enrollment_to	Academic_staff_from	Academic_staff_to	Founded_year	Control_type	Academic_Calender	Campus_setting
UK_rank	1	0.53	-0.75	-0.66	-0.67	-0.64	-0.44	-0.29	-0.11	-0.44	-0.44	-0.6	-0.59	0.38	-0.09	-0.19	-0.11
World_rank	0.53	1	-0.51	-0.37	-0.52	-0.51	-0.42	0.03	-0.17	-0.65	-0.69	-0.6	-0.68	0.21	-0.26	-0.44	-0.29
CWUR_score	-0.75	-0.51	1	0.74	0.76	0.66	0.43	0.11	0.28	0.43	0.42	0.84	0.73	-0.56	0.01	0.03	0.13
Minimum_IELTS_score	-0.66	-0.37	0.74	1	0.69	0.62	0.55	0.03	0.29	0.21	0.21	0.53	0.41	-0.47	-0.02	0.09	0.15
UG_average_fees_(in_pounds)	-0.67	-0.52	0.76	0.69	1	0.89	0.6	0.03	0.38	0.33	0.33	0.63	0.54	-0.46	0.04	0.11	0.14
PG_average_fees_(in_pounds)	-0.64	-0.51	0.66	0.62	0.89	1	0.59	-0.02	0.39	0.32	0.34	0.59	0.54	-0.38	0.07	0.14	0.15
International_students	-0.44	-0.42	0.43	0.55	0.6	0.59	1	-0.13	0.38	0.18	0.2	0.39	0.36	-0.23	-0.07	-0.05	0.04
Student_satisfaction	-0.29	0.03	0.11	0.03	0.03	-0.02	-0.13	1	-0.15	0.06	0.06	0.01	0.03	-0.05	-0.11	0.18	-0.09
Estimated_cost_of_living_per_year_(in_pounds)	-0.11	-0.17	0.28	0.29	0.38	0.39	0.38	-0.15	1	0.1	0.1	0.26	0.12	-0.18	0.08	-0.09	0.01
Student_enrollment_from	-0.44	-0.65	0.43	0.21	0.33	0.32	0.18	0.06	0.1	1	0.99	0.82	0.85	-0.15	0.27	0.27	0.26
Student_enrollment_to	-0.44	-0.69	0.42	0.21	0.33	0.34	0.2	0.06	0.1	0.99	1	0.81	0.85	-0.14	0.3	0.31	0.26
Academic_staff_from	-0.6	-0.6	0.84	0.53	0.63	0.59	0.39	0.01	0.26	0.82	0.81	1	0.99	-0.47	0.19	0.21	0.24
Academic_staff_to	-0.59	-0.68	0.73	0.41	0.54	0.54	0.36	0.03	0.12	0.85	0.85	0.99	1	-0.2	0.23	0.21	0.16
Founded_year	0.38	0.21	-0.56	-0.47	-0.46	-0.38	-0.23	-0.05	-0.18	-0.15	-0.14	-0.47	-0.2	1	0.05	-0.1	-0.16
Control_type	-0.09	-0.26	0.01	-0.02	0.04	0.07	-0.07	-0.11	0.08	0.27	0.3	0.19	0.23	0.05	0.9	0	0.11
Academic_Calender	-0.19	-0.44	0.03	0.09	0.11	0.14	-0.05	0.18	-0.09	0.27	0.31	0.21	0.21	-0.1	0	1	0.18
Campus_setting	-0.11	-0.29	0.13	0.15	0.14	0.15	0.04	-0.09	0.01	0.26	0.26	0.24	0.16	-0.16	0.11	0.18	1

Figure: Correlation heat map

Data Set Analysis 3

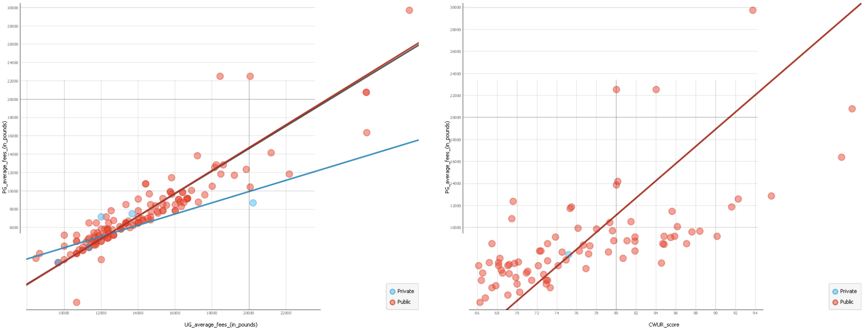


Figure: Linear dependency of PG fees, UG fees and CWUR

Data Set Analysis 4

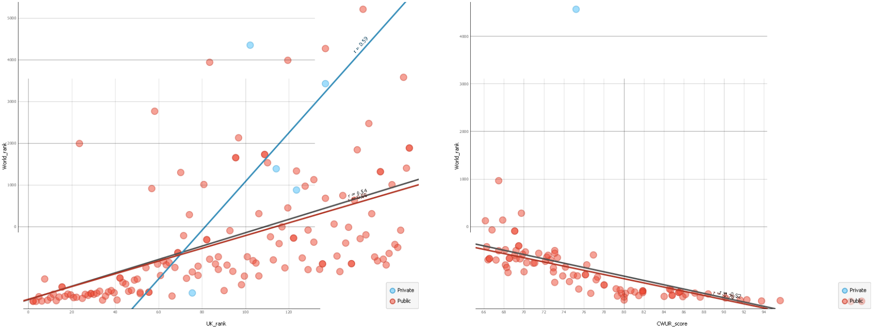


Figure: Linear dependency of World-rank, UK-rank and CWUR

Data Set Analysis 5

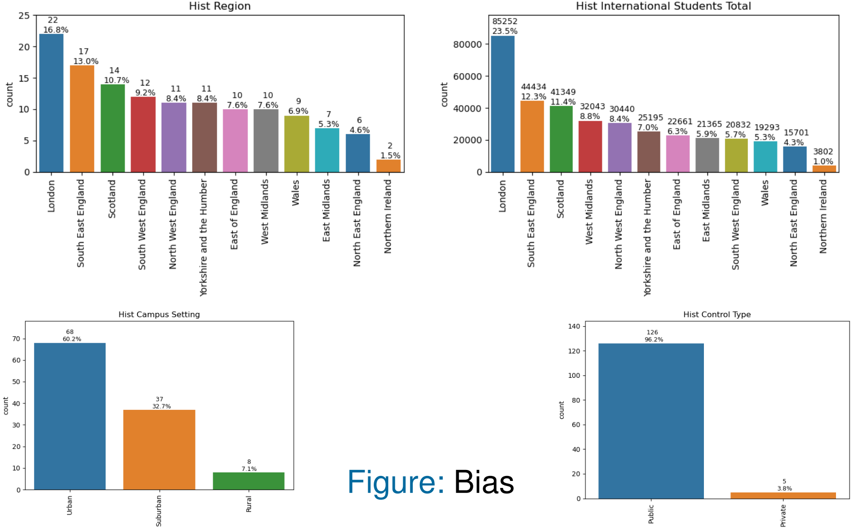


Figure: Bias

Cleaning and Pre-processing

1. Identification of missing values, split of compound columns, deduplication,
2. data set split,
3. missing value imputation,
4. normalization,
5. one-hot encoding,
6. removal of non-numeric columns.

Imputation (mean, median, mixed)

- Missing values (6 out of 21 columns):

Idx	Column	NaN	mean	median	mixed
1	University_name	14	dropped	dropped	dropped
3	Founded_year	14	mean	median	researched online
4	Motto	17	dropped	dropped	dropped
7	CWUR_score	47	mean	median	linear regression on "UK_rank"
16	Academic_Calender	26	mode	mode	mode
17	Campus_setting	18	mean	median	KNN on "Latitude" and "Longitude"

- Suspicious values (3 out of 21 columns):

Idx	Column	Suspicious Value	Count	Imuptation Approach
3	Founded_year	9999	14	researched online
12	Student_satisfaction	0	6	median
14	Academic_staff	over	6	10000

Training and Evaluation

- 5-fold cross validation across seeds $[40, 49]$,
- 3 subsets of the data set:
  - all continuous and categorical columns,
  - only continuous columns excluding `Latitude` and `Longitude`
  - columns with absolute value of correlation higher than 0.5 with the target variables,
- performance evaluation metrics:
  - MSE, MAE, RMSE, R2 score
- average performance across seeds $[40, 49]$.

Models

Linear Regression (LR)

- baseline model to assess performance against,
- cross validation only for column subsets.

Fully Connected Neural Network (FCNN)

- 3 architectures with:
  - increasing number of hidden layers,
  - ReLU hidden activation, linear output activation.

Models

Random Forest (RF)

- grid search parameters:
  - max features $[1, 17]$
  - N estimators $[80, 100]$.

Support Vector Regression (SVR)

- grid search parameters:
  - C (0.005, 0.01, 0.05, 0.1, 0.5, 1, 3, 5)
  - kernel (linear, rbf)
  - epsilon (0.0001, 0.0005, 0.005, 0.01, 0.25, 0.5, 1, 5)
  - gamma (0.0001, 0.001, 0.01, 0.1, 1)

Prediction Performance, Ensemble

- no clear best imputation approach,
- no clear best subset of columns.

	MSE	MAE	RMSE	R2 score	Columns	Imputation
LR	3708162.9	1408.8	1909.6	0.3075	continuous	mixed
FCNN	3389145.4	1304.6	1826.1	0.3659	selected	median
RF	3200786.3	1192.1	1761.7	0.4061	continuous	median
SVR	3643681.6	1219.7	1814.7	0.4625	selected	mixed
Ensemble	2363667.7	1108.6	1537.3	0.5178	——	——

Table: Average prediction performance across random seeds [40, 49]