University/College Data Analysis: Distinguishing Between Private and Public Institutions

Data Science and Analytics

Virtual NLC Semifinals
2020-2021
PID: 205101

Table of Contents

Introduction 2

Methods 2

Exploratory Data Analysis 3

Results 10

Conclusion 11

Introduction

There are thousands of universities and colleges in the United States, and each bears the title of a “public” or “private” institution. While private institutions are often associated with higher tuition and smaller acceptance rates because of research universities and the Ivy League, this analysis seeks to confirm that data on public institutions and private institutions are statistically different enough for a machine learning algorithm, a Decision Tree, to distinguish between private and public schools.

Methods

For this analysis, I used Jupyter Notebook and Python to visualize data and create the machine learning models. Plots that I utilized include line plots, bar plots, histograms, and swarm plots.

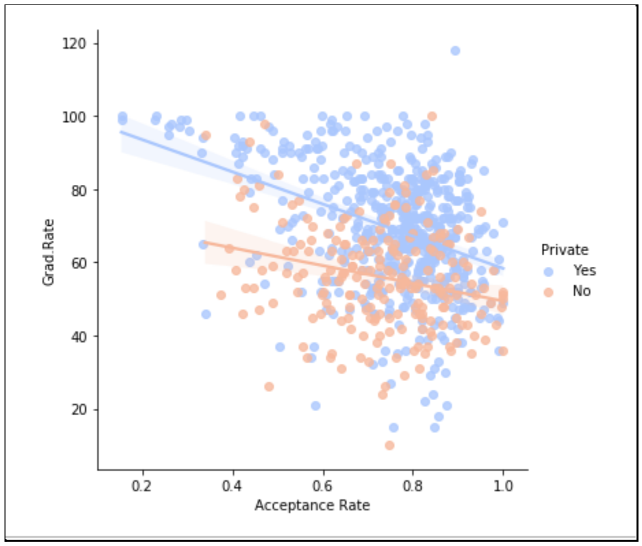
Exploratory Data Analysis

	Unnamed: 0	State	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Termina
0	Abilene Christian University	TX	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78
1	Adelphi University	NY	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30
2	Adrian College	MI	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66
3	Agnes Scott College	GA	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97
4	Alaska Pacific University	AK	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72
5	Albertson College	ID	Yes	587	479	158	38	62	678	41	13500	3335	500	675	67	73
6	Albertus Magnus College	CT	Yes	353	340	103	17	45	416	230	13290	5720	500	1500	90	93
7	Albion College	MI	Yes	1899	1720	489	37	68	1594	32	13868	4826	450	850	89	100
8	Albright College	PA	Yes	1038	839	227	30	63	973	306	15595	4400	300	500	79	84
9	Alderson-Broaddus College	WV	Yes	582	498	172	21	44	799	78	10468	3380	660	1800	40	41
10	Alfred University	NY	Yes	1732	1425	472	37	75	1830	110	16548	5406	500	600	82	88
11	Allegheny College	PA	Yes	2652	1900	484	44	77	1707	44	17080	4440	400	600	73	91
12	Allentown Coll. of St. Francis de Sales	PA	Yes	1179	780	290	38	64	1130	638	9690	4785	600	1000	60	84
13	Alma College	MI	Yes	1267	1080	385	44	73	1306	28	12572	4552	400	400	79	87
14	Alverno College	WI	Yes	494	313	157	23	46	1317	1235	8352	3640	650	2449	36	69
15	American International College	MA	Yes	1420	1093	220	9	22	1018	287	8700	4780	450	1400	78	84
16	Amherst College	MA	Yes	4302	992	418	83	96	1593	5	19760	5300	660	1598	93	98
17	Anderson University	IN	Yes	1216	908	423	19	40	1819	281	10100	3520	550	1100	48	61
18	Andrews University	MI	Yes	1130	704	322	14	23	1586	326	9996	3090	900	1320	62	66
19	Angelo State University	TX	No	3540	2001	1016	24	54	4190	1512	5130	3592	500	2000	60	62

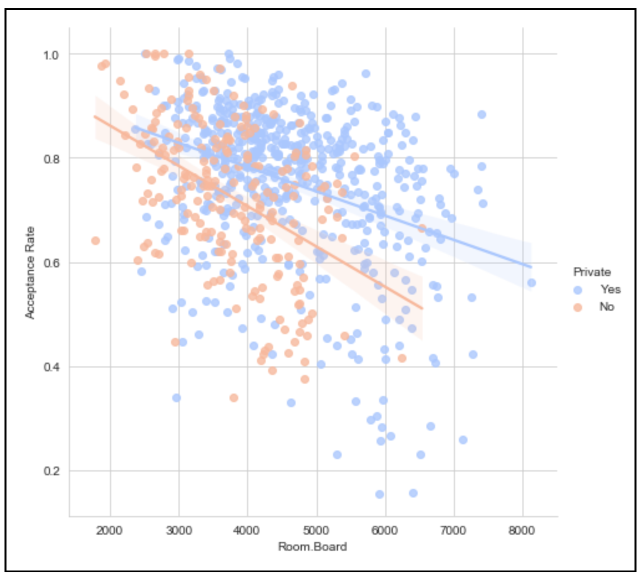
```
df['Acceptance Rate'] = df['Accept']/df['Apps']
```

I started the data analysis by converting the provided dataset into a DataFrame, which can now be manipulated freely. I quickly noticed that the dataset gives information on the number of applicants and the number of acceptances. I divided the acceptances column by the applications column and created a new column in the dataset called "Acceptance Rate."

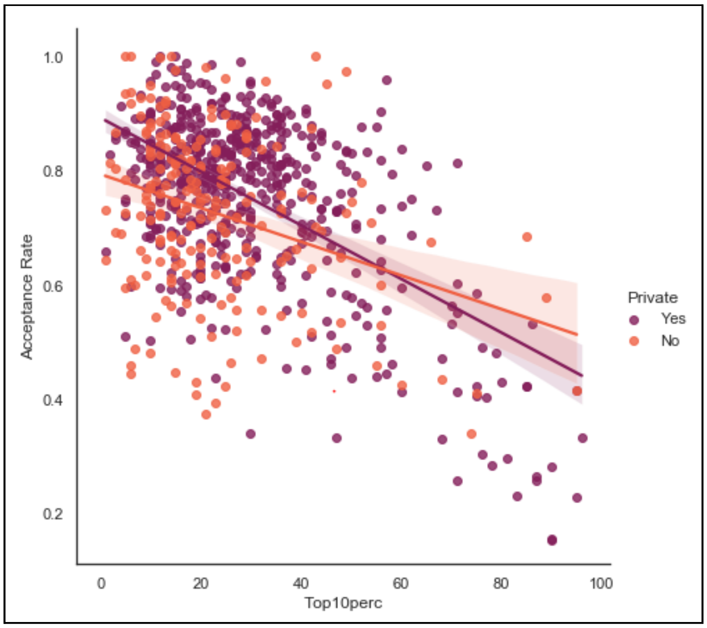
Grad.Rate Vs. Acceptance Rate



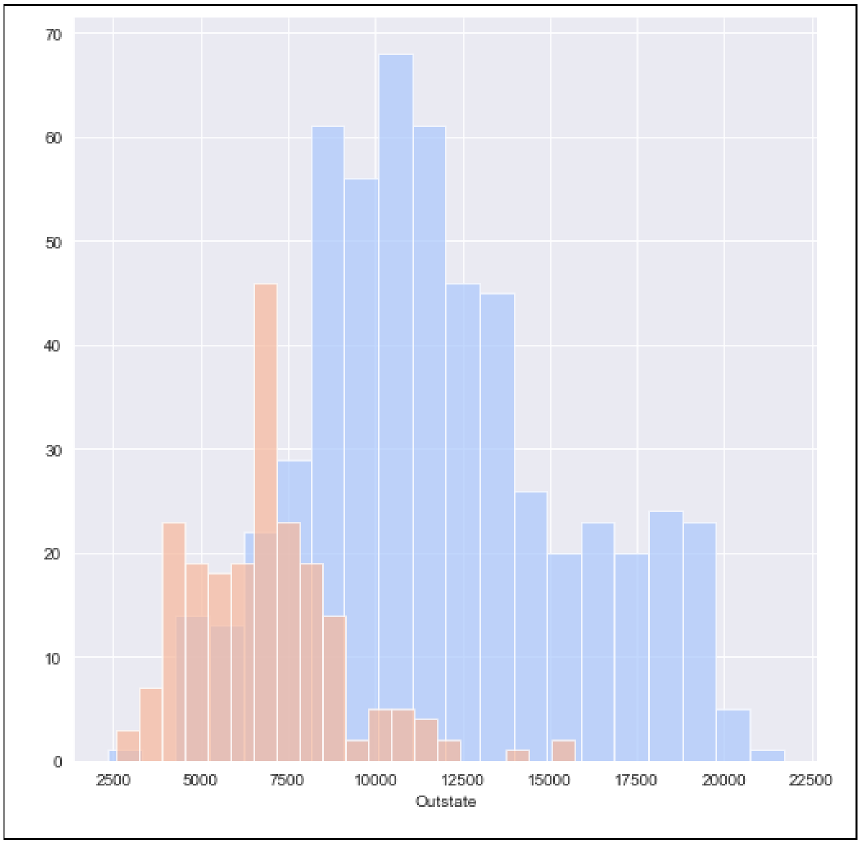
Acceptance Rate Vs. Room.Board



Acceptance Rate Vs. Top10perc



I started the exploratory data analysis with three line plots (fitted with regression lines) observing the relationship between acceptance rate and other variables in the dataset, as well as how private schools are represented in this data versus public schools. Across all three plots, private and public schools share similar trends. Both types of schools tend to have lower graduation rates with higher acceptance rates, lower acceptance rates with higher room + board costs, and lower acceptance rates with a higher proportion of new students that were top 10% of their high school class. However, there are some notable differences. As depicted in the first and second plots, public schools tend to have lower graduation rates overall and lower room + board costs. These quantitative differences can help a machine learning algorithm determine whether a school is public or private.



Next, I plotted a histogram of private school (blue) and public school (pink) out-of-state tuition to observe the distribution. Evidently, private schools have a significantly further reaching distribution and higher average out-of-state tuition. On the other hand, public schools appear to have smaller average tuition. This noticeable difference between public out-of-state tuition distribution and private out-of-state distribution tuition bolsters the theory that a machine learning model can identify the difference between a public and private school.

	TX	IA	NE	AL	LA	KS	FL	VT	DC	MO	...	MI	VA	MN	TN	IL
0	0.742169	0.895408	0.955848	0.899709	0.876768	0.802326	0.791919	0.630058	0.835234	0.876095	...	0.768207	0.886762	0.774924	0.833603	0.882384
1	0.565254	0.800000	0.956349	0.730435	0.688689	0.981013	0.698663	0.747812	0.259199	0.503185	...	0.905740	0.803103	0.725993	0.798361	0.735632
2	0.841772	0.867498	0.894073	0.855263	0.523100	0.910891	0.779887	0.836879	0.765182	0.704315	...	0.852407	0.900000	0.586117	0.867701	0.672000
3	0.880494	0.737575	0.866585	0.728435	0.894451	1.000000	0.820057	0.805128	NaN	0.928571	...	0.623009	0.888889	0.921109	0.590909	0.906291
4	0.899736	0.930464	0.714693	0.830357	0.728708	0.693027	0.504594	0.803121	NaN	0.841371	...	0.847534	0.867497	0.804305	0.792965	0.759871
5	0.907923	0.919257	0.956349	0.339828	0.913209	0.976231	0.811505	0.938318	NaN	0.750382	...	0.658902	0.814224	0.886905	0.819820	0.758824
6	0.738442	0.745946	NaN	0.805911	NaN	0.697619	0.628857	0.932301	NaN	0.757781	...	0.809783	0.436420	0.946619	0.612291	0.919275
7	0.711073	0.681216	NaN	0.701169	NaN	1.000000	0.753077	0.722513	NaN	0.928912	...	0.866238	0.905350	0.810415	0.845070	0.880313
8	0.578856	0.880978	NaN	0.732301	NaN	0.712062	0.745258	0.833333	NaN	0.917647	...	0.833389	0.844444	0.820080	0.647413	0.774487
9	0.928898	0.781840	NaN	0.660252	NaN	0.648211	0.647721	0.784027	NaN	0.866727	...	0.873950	0.826736	0.509017	0.714104	0.771225
10	0.707192	0.898644	NaN	NaN	NaN	NaN	0.788054	NaN	NaN	0.819816	...	0.848734	0.765257	0.869509	0.606557	0.440000
11	0.803302	0.909556	NaN	NaN	NaN	NaN	0.423561	NaN	NaN	1.000000	...	0.903017	0.788250	0.832536	0.755738	0.812500
12	0.895893	0.861413	NaN	NaN	NaN	NaN	0.710004	NaN	NaN	0.858777	...	0.973118	0.523126	0.915525	0.795395	0.651685
13	0.990654	0.873846	NaN	NaN	NaN	NaN	0.756248	NaN	NaN	0.900000	...	0.788565	0.827243	0.744217	0.784483	0.909457
14	0.733119	0.943023	NaN	NaN	NaN	NaN	0.696111	NaN	NaN	0.705354	...	0.675647	0.470908	0.745706	0.594249	0.909879
15	0.820595	0.800446	NaN	NaN	NaN	NaN	0.616155	NaN	NaN	0.972829	...	1.000000	0.680743	0.599451	0.718855	0.826767
16	0.726751	0.858268	NaN	NaN	NaN	NaN	0.804878	NaN	NaN	0.705192	...	0.784444	0.854214	0.578705	0.668512	0.754054
17	0.909263	0.968750	NaN	NaN	NaN	NaN	0.510714	NaN	NaN	0.687092	...	0.615639	0.883768	0.878464	0.601977	0.553892
18	0.751893	0.815396	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.694737	...	NaN	0.500690	NaN	NaN	0.873269
19	0.851107	0.681750	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.835347	...	NaN	0.820404	NaN	NaN	0.836938
20	0.746835	0.872461	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.825038	...	NaN	0.858295	NaN	NaN	0.676413
21	0.846284	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.927505	...	NaN	0.748165	NaN	NaN	0.776344
22	0.749691	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	0.885057	NaN	NaN	0.674295
23	0.863436	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	0.803772	NaN	NaN	0.423143
24	0.619511	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	0.892473	NaN	NaN	0.689756
25	0.748663	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	0.870130	NaN	NaN	0.739631
26	0.779945	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	0.461303	NaN	NaN	0.824204
27	0.648861	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	0.339706	NaN	NaN	0.864173
28	0.735120	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	0.704614	NaN	NaN	0.472432

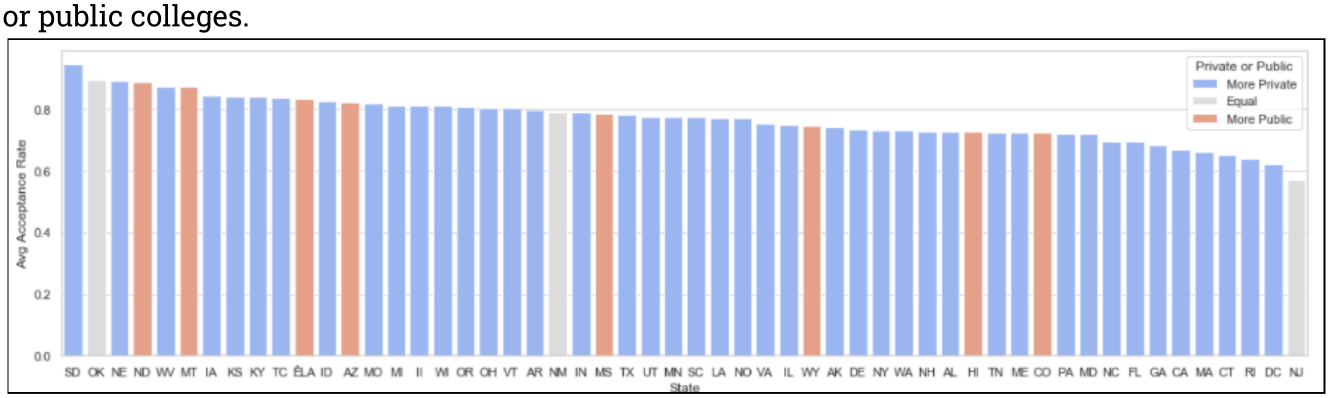
Afterwards, I decided to explore the average acceptance rate by state. I started by putting each college's acceptance rate under the column of its respective state. Since some states had more or less institutions than others, empty values were filled with "NaN."

	NE	AL	LA	KS	FL	VT	DC	MO	MS	PA	...	MN	TN	IL	UT	OR	AR	ND	DE	IA
0	More Private	More Private	More Private	More Private	More Private	More Private	More Private	More Private	More Public	More Private	...	More Private	More Private	More Private	More Private	More Private	More Private	More Public	More Private	More Private

In addition to collecting states' acceptance rates, I designed an algorithm to check whether each state has more private or public colleges. If they had the same number on both sides, they were labeled "equal."

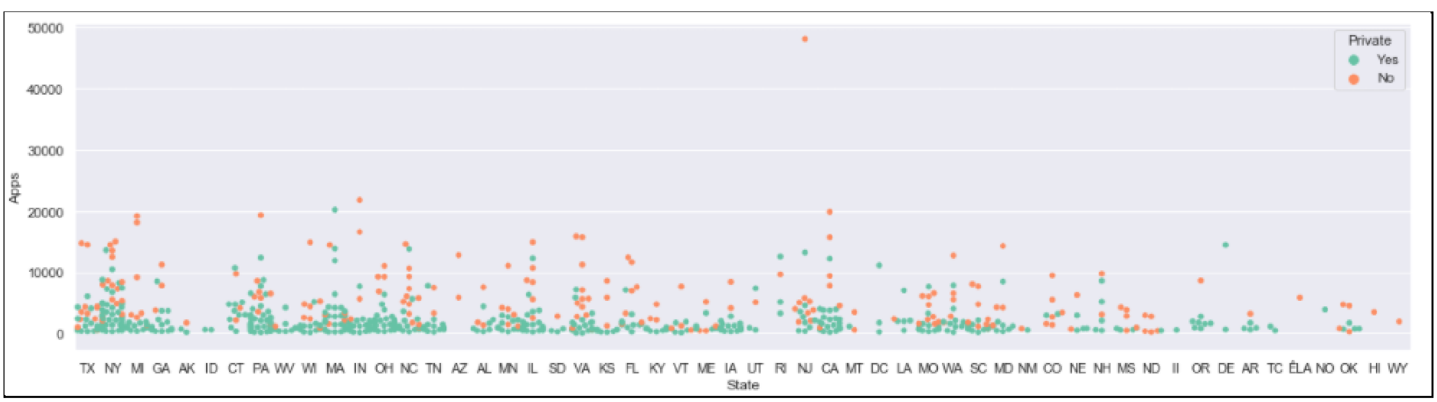
	State	Avg Acceptance Rate	Private or Public
12	SD	0.944902	More Private
35	OK	0.895828	Equal
0	NE	0.890649	More Private
50	ND	0.889421	More Public
24	WV	0.875295	More Private
36	MT	0.871882	More Public
52	IA	0.843987	More Private
3	KS	0.842138	More Private
17	KY	0.840658	More Private
22	TC	0.835645	More Private
27	ÈLA	0.832722	More Public
33	ID	0.825215	More Private
23	AZ	0.820893	More Public
7	MO	0.818485	More Private
42	MI	0.812685	More Private
39	II	0.810714	More Private
14	WI	0.810340	More Private
48	OR	0.808548	More Private
40	OH	0.805270	More Private
5	VT	0.803349	More Private
49	AR	0.795249	More Private
38	NM	0.791170	Equal
15	IN	0.790487	More Private
8	MS	0.786531	More Public
53	TX	0.783542	More Private
47	UT	0.775023	More Private
44	MN	0.774978	More Private
41	SC	0.773915	More Private
2	LA	0.770821	More Private
29	NO	0.769627	More Private
43	VA	0.753501	More Private

Finally, I calculated the average acceptance rate per state and made a new DataFrame containing the states, their average acceptance rate, and whether they have more private or public colleges.



This graph represents the DataFrame mentioned above. I plotted a bar graph representing each state's average acceptance rate. On top of this graph I applied a hue which distinguishes the state's status of having more private or public colleges. In most states, private colleges champion in number regardless of the acceptance rate trend. Therefore, including the "State" column in our ML model's input data will affect the results because private colleges are overrepresented in most states and therefore the state could infer to the ML model of private or public college bias.

To examine the distribution of acceptance rates per state more closely, I plotted this swarm plot. For the vast majority of data points, private and public colleges appear to have similar acceptance rates. However, for ranges with notably lower acceptance rates, there is an overrepresentation of private colleges. These small differences between public colleges and private colleges eventually become significant as the ML model finds several of them.



For the last plot, I examined the distribution of applications per state to determine if there are any more observable differences between public and private colleges. Throughout most states, public colleges (orange data points) have more applicants than the state's private colleges. This pattern further supports the capability of a model to differentiate between private and public institutions.

Results

Classification Report #1 ("State" Column Included in Input Data)

	precision	recall	f1-score	support
0	0.81	0.86	0.84	51
1	0.96	0.94	0.95	180
accuracy			0.93	231
macro avg	0.89	0.90	0.90	231
weighted avg	0.93	0.93	0.93	231

Finally, I started the machine learning portion of this analysis. For this analysis, I used a supervised learning algorithm: the Decision Tree. This algorithm is simple and efficient, making it a suitable choice for this analysis. After training, I printed the classification report of the model displaying its accuracy. The model has a weighted accuracy of 93%, communicating that the model can differentiate between a public and private institution 93% of the time based on the data. This high accuracy solidifies that the trends I observed in the exploratory data analysis portion were significant enough to allow the model to accurately classify institutions as private or public.

Classification Report #2 ("State" Column Not Included in Input Data)

	precision	recall	f1-score	support
0	0.81	0.90	0.85	51
1	0.97	0.94	0.95	180
accuracy			0.93	231
macro avg	0.89	0.92	0.90	231
weighted avg	0.94	0.93	0.93	231

In many datasets, labels like the "State" label, a string, will often be omitted and the luxury of having this data is nonexistent. To ensure that this factor was not overwhelmingly affecting the model's accuracy, I ran the model without the "State" labels. The model performed similarly well, ensuring that the "State" label was providing a significant advantage to the algorithm and that the numerical data alone is enough for the Decision Tree to classify institutions as private or public.

Conclusion

Though private colleges and public colleges share similar trends, their differences in terms of their acceptance rates, tuition/fees, and more are significant enough for a machine learning algorithm (Decision Tree) to accurately determine which colleges are private and which ones are public. This analysis serves the purpose of highlighting the idea that college applicants should carefully consider factors such as tuition, acceptance rate, number of applicants, and more when applying to private and public institutions because oftentimes, these differences are significant.