COMP 309 — Machine Learning Tools and Techniques

Part 1: Evidence related to rental costs in Wellington

Objectives

The goal is to find out the relationships between active rent , rental costs, and the GDP of New Zealand. Also find out does the rent costs effect the enrolment for the University.

Datasets selection

- Enrolment of the University

This is an important indicator for support the motivation. In the enrolment of the university dataset, I did select 4 universities, they are University of Auckland, Victoria University of Wellington, University of Otago, and University of Canterbury. There are the famous university in New Zealand and they campus are all in they own region. That means it can relate to their local rent prices, and it can use to compare with other universities

- Mean rents by region

This is another important indicator for support the motivation. In this data, it directly shows the Mean rents by region in New Zealand. I did select 4 attributes (Wellington, Otago, Auckland and Canterbury) those region is the seat of the Universities. It shows the overall trend of the rents in New Zealand. We can see from the graph, the rent price in New Zealand is continuously increasing. From those informations, it give me a basic concept of rent price in New Zealand. Thus use those dataset to predict the rent price in the future. The rent price of those region will continuously widening.

- Active bounds by region

I thing active bounds is an important attributes for effort the rental costs, the more customer want to rental the house, the homeowners will not worried about rent out problem. Therefore he can increases the rental price. As the quantity demanded increases , the house prices will also increases.

- Population by region

The population in region is an important attributes for effort the quantity demanded.

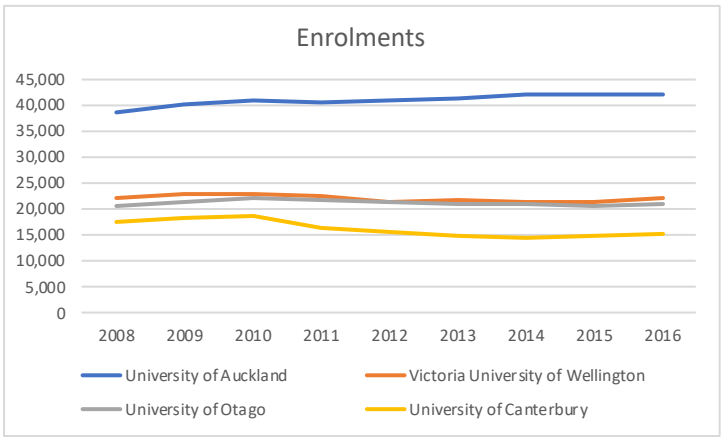
- Gross domestic product in NewZealand those

The gross domestic product can shows the Economic level for new Zealand.

Manipulated in the pipeline

I got different data from different resources, the timeline is also different. I have to clear up the data made them using the same timeline. The time measure unit I design to use is year, this because almost dataset I funded is use this unit, and is hard to find another dataset use month as the measurement unit. The most important reason is the enrolment of university count by monthly it do not exist. In the real world, the enrolment of university only count by trimester or years. If average divide the enrolment into month it will not make any senses. For the dataset that uses month as the timeline measurement (active bounds and Mean rents) , I did average the rents/bounds for each year as the measurement for relative year. Also I find out there is some continues missing data in the population data set, in the year between 1993-1995,1997-2000 and 2002-2005. But from my observation, the population changes is quite little, and it I will not have much effect for the future. So, I did estimate the missing value by regression. Then I can compare the datas for each year.

	Enrolments				Mean rent by region				Active bouns by region				Population by Region				GDP (million)
Time	University of	Victoria Univ	University of	University of	Auckland	Wellington	Otago	Canterbury	Auckland	Wellington	Otago	Canterbury	Auckland	Wellington	Otago	Canterbury	New Zealand
1993					$216.83	$184.25	$161.17	$152.92	$35,455.33	$14,108.58	$6,200.00	$14,373.42					78309
1994					$246.00	$194.00	$168.83	$163.00	$40,354.50	$15,168.42	$6,757.83	$16,044.92					84460
1995					$271.83	$211.25	$169.00	$172.42	$45,762.42	$16,419.58	$7,323.33	$17,474.83					90337
1996					$277.67	$222.50	$163.67	$176.58	$51,359.58	$17,928.42	$7,702.50	$19,231.75	1115800	426900	189300	480400	96235
1997					$269.08	$229.83	$161.00	$174.67	$55,753.00	$19,579.33	$7,894.50	$20,898.42					101101
1998					$263.75	$234.83	$179.08	$171.58	$59,396.17	$21,068.00	$8,191.83	$21,885.17					104814
1999					$266.17	$236.42	$176.58	$175.17	$63,094.83	$22,597.83	$8,437.92	$23,137.50					106827
2000					$275.17	$240.33	$185.75	$178.33	$67,685.00	$24,109.17	$8,751.58	$24,954.25					113229
2001					$300.33	$247.17	$200.83	$191.75	$72,913.00	$25,617.67	$9,372.75	$26,456.50	1218300	440200	188300	496700	119838
2002					$324.92	$256.42	$220.42	$212.92	$76,773.67	$27,389.58	$10,071.08	$28,422.58					128712
2003					$330.83	$267.92	$245.08	$231.42	$81,174.08	$29,473.50	$11,163.50	$30,428.58					135181
2004					$333.58	$277.92	$258.25	$241.58	$85,421.42	$30,899.33	$11,655.83	$31,933.58					144502
2005					$342.08	$295.92	$261.42	$254.00	$89,951.67	$32,116.92	$12,399.75	$32,953.42					154558
2006					$363.08	$323.17	$273.25	$267.92	$93,908.25	$32,788.50	$12,942.00	$34,082.17	1373000	466300	199800	540000	162935
2007					$380.17	$344.50	$283.75	$280.42	$99,477.83	$34,733.92	$13,556.25	$36,311.67	1390400	469300	201000	547400	172112
2008	38,550	21,930	20,645	17,375	$382.08	$351.83	$289.33	$278.00	$105,607.67	$36,905.25	$14,310.08	$38,715.08	1405500	471800	202100	553800	186846
2009	39,935	22,925	21,345	18,320	$394.17	$363.50	$291.92	$284.17	$110,060.08	$38,039.50	$15,252.42	$40,403.42	1421700	475600	203300	560600	189617
2010	40,975	22,880	21,950	18,495	$414.75	$365.25	$310.08	$302.83	$114,515.58	$39,956.00	$15,938.33	$41,038.67	1439600	479400	204600	567700	194251
2011	40,380	22,560	21,640	16,190	$430.42	$371.42	$317.00	$329.58	$117,108.08	$41,320.08	$16,771.92	$41,961.42	1459600	483400	206600	559300	203434
2012	40,785	21,195	21,240	15,545	$444.83	$379.50	$327.00	$366.58	$118,741.08	$42,088.58	$17,361.00	$42,195.92	1476500	485100	207400	556000	213241
2013	41,365	21,480	20,965	14,645	$463.58	$390.17	$341.83	$394.92	$123,745.83	$42,884.67	$17,971.08	$43,065.83	1493200	486700	208800	562900	217472
2014	41,955	21,190	20,935	14,465	$493.67	$400.08	$359.00	$388.17	$129,094.17	$44,070.17	$18,312.33	$44,464.00	1526900	491400	211600	574300	232645
2015	42,100	21,450	20,690	14,570	$514.58	$420.58	$385.17	$376.42	$135,423.25	$45,180.67	$18,803.42	$46,269.08	1569900	496900	215000	586400	242238
2016	41,865	21,950	20,895	15,305	$534.33	$444.67	$415.00	$367.83	$140,927.08	$45,590.42	$19,529.00	$47,830.67	1614500	504900	219200	599900	254704
2017					$549.43	$477.00	$424.29	$374.14	$143,897.29	$45,746.57	$19,812.86	$48,955.00	1657200	513900	224200	612000	270280
2018																	286410



The above data is my preliminary ML dataset, I design to use liner regression technique to predict the rental costs in wellington. Therefore I have to dealing with the data types and missing values. In liner regression technique, the data type it use is numeric, but some of my dataset is money type, I can simply changes to numeric and match with that. For dealing with the missing value in Population dataset, I find out the population is continuous increasing, so I design to use liner regression to impute the missing values, in this type of imputation it do not have much effort for the whole dataset. But for the enrolment dataset, the gradient is not continuous is hard to estimate the missing values, because there are lots of missing value in enrolment and is not replaceable. The data after imputation shows as follow.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1		Enrolments				Mean rent by region				Active bouns by region				Population by Region				GDP (millio
2	Time	University of	Victoria Univ	University of	University of	Auckland	Wellington	Otago	Canterbury	Auckland	Wellington	Otago	Canterbury	Auckland	Wellington	Otago	Canterbury	New Zealan
3	1993	?	?	?	?	217	184	161	153	35455	14109	6200	14373	1031728	412046	174009	463013	78309
4	1994	?	?	?	?	246	194	169	163	40355	15168	6758	16045	1056251	416006	175872	468699	84460
5	1995	?	?	?	?	272	211	169	172	45762	16420	7323	17475	1080773	419966	177735	474385	90337
6	1996	?	?	?	?	278	223	164	177	51360	17928	7703	19232	1115800	426900	189300	480400	96235
7	1997	?	?	?	?	269	230	161	175	55753	19579	7895	20898	1127816	427319	179614	485695	101101
8	1998	?	?	?	?	264	235	179	172	59396	21068	8192	21885	1152489	431321	181616	491386	104814
9	1999	?	?	?	?	266	236	177	175	63095	22598	8438	23138	1177161	435323	183618	497076	106827
10	2000	?	?	?	?	275	240	186	178	67685	24109	8752	24954	1201833	439326	185620	502767	113229
11	2001	?	?	?	?	300	247	201	192	72913	25618	9373	26457	1218300	440200	188300	496700	119838
12	2002	?	?	?	?	325	256	220	213	76774	27390	10071	28423	1253229	448112	189454	517088	128712
13	2003	?	?	?	?	331	268	245	231	81174	29474	11164	30429	1277696	452037	191473	522485	135181
14	2004	?	?	?	?	334	278	258	242	85421	30899	11656	31934	1302163	455961	193492	527882	144502
15	2005	?	?	?	?	342	296	261	254	89952	32117	12400	32953	1326630	459885	195511	533279	154558
16	2006	?	?	?	?	363	323	273	268	93908	32789	12942	34082	1373000	466300	199800	540000	162935
17	2007	?	?	?	?	380	345	284	280	99478	34734	13556	36312	1390400	469300	201000	547400	172112
18	2008	38550	21930	20645	17375	382	352	289	278	105608	36905	14310	38715	1405500	471800	202100	553800	186846
19	2009	39935	22925	21345	18320	394	364	292	284	110060	38040	15252	40403	1421700	475600	203300	560600	189617
20	2010	40975	22880	21950	18495	415	365	310	303	114516	39956	15938	41039	1439600	479400	204600	567700	194251
21	2011	40380	22560	21640	16190	430	371	317	330	117108	41320	16772	41961	1459600	483400	206600	559300	203434
22	2012	40785	21195	21240	15545	445	380	327	367	118741	42089	17361	42196	1476500	485100	207400	556000	213241
23	2013	41365	21480	20965	14645	464	390	342	395	123746	42885	17971	43066	1493200	486700	208800	562900	217472
24	2014	41955	21190	20935	14465	494	400	359	388	129094	44070	18312	44464	1526900	491400	211600	574300	232645
25	2015	42100	21450	20690	14570	515	421	385	376	135423	45181	18803	46269	1569900	496900	215000	586400	242238
26	2016	41865	21950	20895	15305	534	445	415	368	140927	45590	19529	47831	1614500	504900	219200	599900	254704
27	2017	?	?	?	?	549	477	424	374	143897	45747	19813	48955	1657200	513900	224200	612000	270280

Another way I can think of is only keep the data between year 2008-2016. That because to support the motivation we need the enrolment dataset, and remove all the dataset between 1993-2007 to

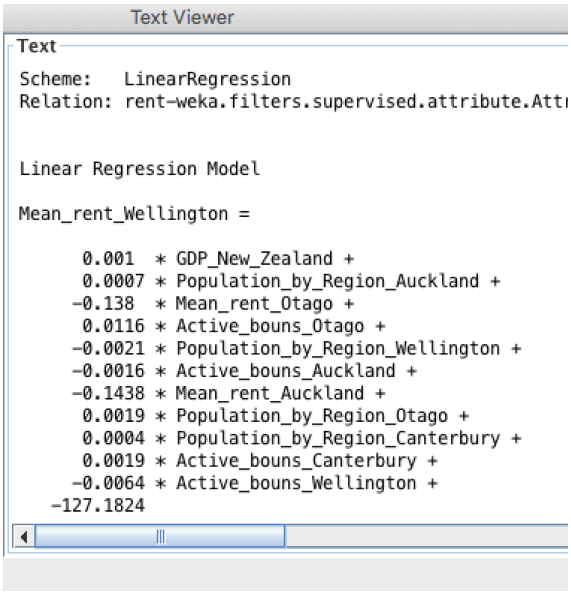
keep the integrity of the data. The reason is their have lots missing values, during that time and it don't have much fluctuation during 1993-2007.

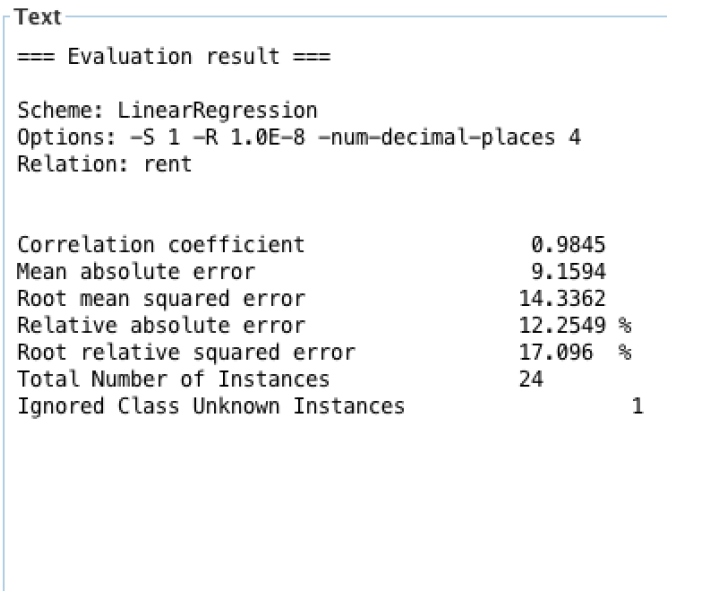
	Enrolments				Mean rent by region				Active bouns by region				Population by Region				GDP (million)
	University of A	Victoria Unive	University of (	University of (	Auckland	Wellington	Otago	Canterbury	Auckland	Wellington	Otago	Canterbury	Auckland	Wellington	Otago	Canterbury	New Zealand
2008	38,550	21,930	20,645	17,375	$382.08	$351.83	$289.33	$278.00	$105,607.67	$36,905.25	$14,310.08	$38,715.08	1405500	471800	202100	553800	186846
2009	39,935	22,925	21,345	18,320	$394.17	$363.50	$291.92	$284.17	$110,060.08	$38,039.50	$15,252.42	$40,403.42	1421700	475600	203300	560600	189617
2010	40,975	22,880	21,950	18,495	$414.75	$365.25	$310.08	$302.83	$114,515.58	$39,956.00	$15,938.33	$41,038.67	1439600	479400	204600	567700	194251
2011	40,380	22,560	21,640	16,190	$430.42	$371.42	$317.00	$329.58	$117,108.08	$41,320.08	$16,771.92	$41,961.42	1459600	483400	206600	559300	203434
2012	40,785	21,195	21,240	15,545	$444.83	$379.50	$327.00	$366.58	$118,741.08	$42,088.58	$17,361.00	$42,195.92	1476500	485100	207400	556000	213241
2013	41,365	21,480	20,965	14,645	$463.58	$390.17	$341.83	$394.92	$123,745.83	$42,884.67	$17,971.08	$43,065.83	1493200	486700	208800	562900	217472
2014	41,955	21,190	20,935	14,465	$493.67	$400.08	$359.00	$388.17	$129,094.17	$44,070.17	$18,312.33	$44,464.00	1526900	491400	211600	574300	232645
2015	42,100	21,450	20,690	14,570	$514.58	$420.58	$385.17	$376.42	$135,423.25	$45,180.67	$18,803.42	$46,269.08	1569900	496900	215000	586400	242238
2016	41,865	21,950	20,895	15,305	$534.33	$444.67	$415.00	$367.83	$140,927.08	$45,590.42	$19,529.00	$47,830.67	1614500	504900	219200	599900	254704

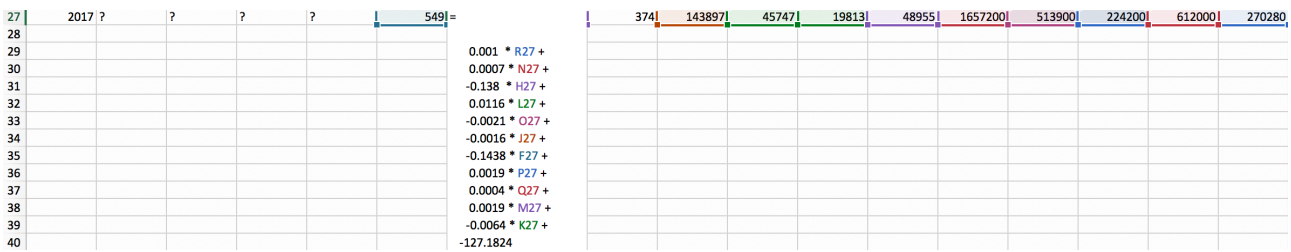
But after I realise that there is not enough instances to support the technique, so I change the data back.

Use liner regression to predict the future rental prices in Wellington

Consider to predict the rental prices in Wellington 2018, we want to know about the relation between the current year attributes and the rental prices in Wellington next year instead of the relation of attributes in the same year. Therefore I move the Wellington rental price back by one year. In this situation, the figure below is the result after I run through the pipeline. And the mean rental prices in Wellington 2018 I predict is $557







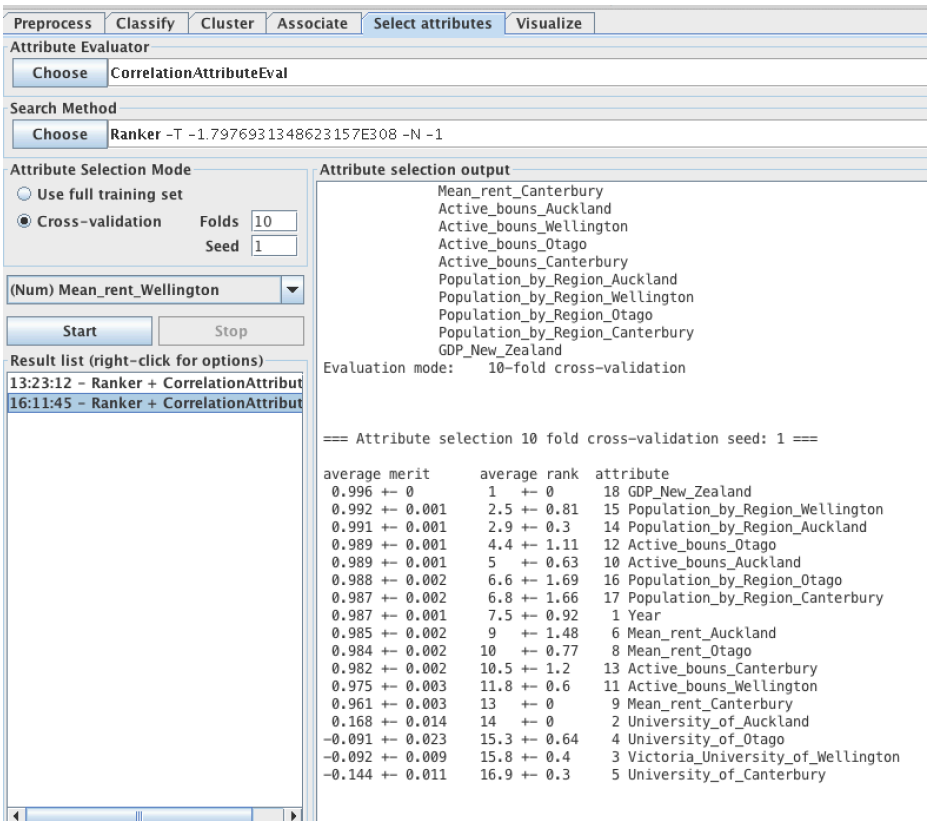
Part 2:

Feature importance to rental costs in Wellington

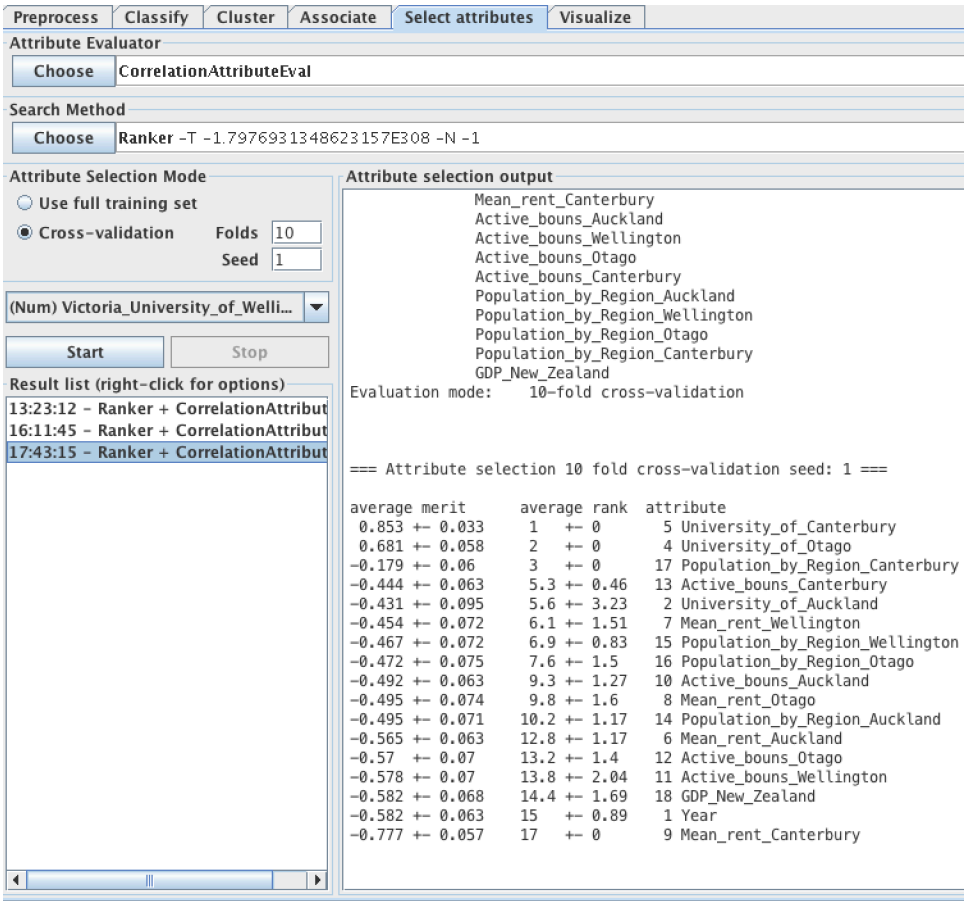
I design to use Weka Select attributes to find out which attributes are irrelevant to the linear Regressing technique. The attribute evaluator method I choose is base on correlation coefficient, to evaluate the worth of an attribute by measuring the correlation between the attributes and the class(Mean_rent_Wellington). Through Cross-validation mode, the result shows the correlation coefficient of last 4 attributes are very close to 0. Those attributes are the enrolment number of 4 universities. That means there are no linear relationship, the enrolment number of university do not have much correlation to mean rent of Wellington. In addition to

those 4 attributes, the correlation coefficient of other attributes are above 0.95. It shows almost perfect uphill (positive) linear relationship, therefore those attributes are important in liner regression. Also the list is order by ranking, we can see the most important attributes are GDP of NewZealand, Population of Wellington, and so on.

The figure shows as follow:

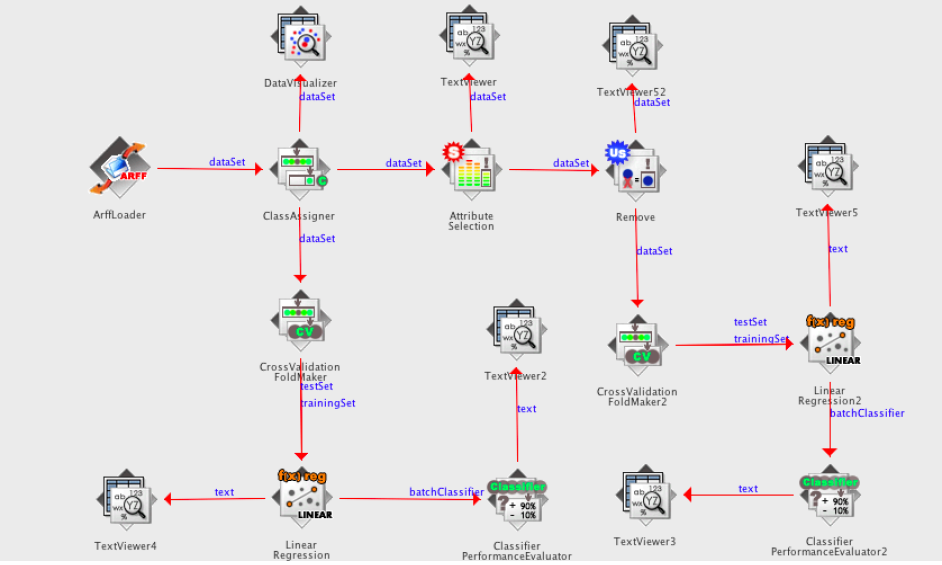


Another issues I interesting about is does the rent costs effect the enrolment for the University?



Feature importance to enrolment of Victoria university

From the graph above it shows most of all attributes have negative correlation coefficient, there only university of Canterbury and university of Otago have positive correlation coefficient, that shows a strong uphill (positive) linear relationship, enrolment of Canterbury and Otago university will have strong effort on the enrolment of Victoria university. The correlation coefficient of the 4th attributes to 2nd to last attributes is around - 0.5. It shows a moderate downhill (negative) relationship, if we decreasing those attribute values, the enrolment of Victoria university will moderately increase. But an interesting thing I find out is the correlation coefficient of mean rent of Canterbury shows A strong downhill (negative) linear relationship. Such as decreasing the rent prices in Canterbury will lead to have more student enrol in Victoria university. But, I cannot guarantee this ratiocination is correct , because I only use 8 years enrolment of Victoria university data to support this. To verify this ratiocination I need more datasets.



Show dimensionality reduction included/excluded from the pipeline

The above figures is use the pipeline to verify which attributes are irrelevant and with attributes are important to liner regression. In the pipeline there is a fork after class Assigner, on one side is directly run through linear regression technique, another side is remove attributes before run through linear regression technique. The attributes used and the result shows as follow.

Before remove irrelevant attributes:

```
@attribute GDP_New_Zealand numeric
@attribute Population_by_Region_Wellington numeric
@attribute Population_by_Region_Auckland numeric
@attribute Active_bouns_Otago numeric
@attribute Active_bouns_Auckland numeric
@attribute Population_by_Region_Otago numeric
@attribute Population_by_Region_Canterbury numeric
@attribute Year numeric
@attribute Mean_rent_Auckland numeric
@attribute Mean_rent_Otago numeric
@attribute Active_bouns_Canterbury numeric
@attribute Active_bouns_Wellington numeric
@attribute Mean_rent_Canterbury numeric
@attribute University_of_Auckland numeric
@attribute University_of_Otago numeric
@attribute Victoria_University_of_Wellington numeric
@attribute University_of_Canterbury numeric
@attribute Mean_rent_Wellington numeric
```

Text

```
=== Evaluation result ===

Scheme: LinearRegression
Options: -S 1 -R 1.0E-8 -num-decimal-places 4
Relation: rent


Correlation coefficient                  0.991
Mean absolute error                      7.8519
Root mean squared error                 11.2984
Relative absolute error                 10.1655 %
Root relative squared error             13.0297 %
Total Number of Instances               25
```

After remove irrelevant attributes:

```
@attribute GDP_New_Zealand numeric
@attribute Population_by_Region_Wellington numeric
@attribute Population_by_Region_Auckland numeric
@attribute Active_bouns_Otago numeric
@attribute Active_bouns_Auckland numeric
@attribute Population_by_Region_Otago numeric
@attribute Population_by_Region_Canterbury numeric
@attribute Year numeric
@attribute Mean_rent_Auckland numeric
@attribute Mean_rent_Otago numeric
@attribute Active_bouns_Canterbury numeric
@attribute Active_bouns_Wellington numeric
@attribute Mean_rent_Wellington numeric
```

Text

```
=== Evaluation result ===

Scheme: LinearRegression2 : LinearRegression
Options: -S 1 -R 1.0E-8 -num-decimal-places 4
Relation: rent-weka.filters.supervised.attribute.A


Correlation coefficient                  0.9972
Mean absolute error                      5.0808
Root mean squared error                  6.246
Relative absolute error                  6.5779 %
Root relative squared error              7.203  %
Total Number of Instances               25
```

I did run through cross Validation to get this result, it will be fair for using to comparing. The attribute list is order by ranking, I did remove the last 5 attributes from the original dataset, the correlation coefficient improve 0.0062. The correlation coefficient is already very high before, so it don't have much improvement. But I am convinced of the error values are all decreases. That verifies the attributes are irrelevant to liner regression.

Part 4: Consider the consequences and ethics of reporting your findings

From Part 2, it shows about the correlation coefficient between class and attributes. Therefrom, we know about when correlation coefficient close to 1 means there are have a perfect positive linear relationship, if close to -1 there are have a perfect native linear relationship, and 0 means no liner relationships at all. According to this knowledge, we can come to the conclusion: if some attributes, which have high correlation coefficient, changes, The output of the class will also change. Therefore we can have a hypothesis, like chain reaction:

If the New Zealand's GDP increases 10% higher than alternatives , then the rental cost in Wellington will also increases 5%. Than the consequences is the total enrolment of Victoria university will drop 2%.

To support my hypothesis, I increases the NewZealand's GDP in 2017 by 10%(change form 270280 to 297308), feed into the existing model. The rental cost of Wellington in 2018 will predict as $584. That means the rental cost increase $27, through the simple calculation((584-557)/557= 0.048) it increase around 5%. In terms of enrolment of Victoria university, the predicted enrolment in 2018 is around 22528, after the rental prices increase 5%, the predicted enrolment in 2018 is around 21217. Through the simple calculation ((22528-21217)/22528= 0.048) it decrease around 5%.

The hypothesis is correct at the first half part, but the decreasing rate is out of my predict range this relation is indirect. Thus, cannot say increase GDP will leads to decrease enrolment of Victoria university. React to this information, rental sector react should control the rental prices in Wellington, as far as possible to keep the increase rate of the rental prices under half of the GDP increase rate.

Data reference:

Active bonds by region:
https://catalogue.data.govt.nz/dataset/rental-bond-data-by-region/resource/d477310f-47f1-4aa4-b9ce-84232badfa4b

Mean rents by region:
https://catalogue.data.govt.nz/dataset/rental-bond-data-by-region/resource/f53e86da-217f-49d9-86cb-cb56ad5cebd3

Enrolment of the University:
https://www.educationcounts.govt.nz/statistics/tertiary-education/summary_tables

GDP of NewZealand:
https://www.stats.govt.nz/information-releases/gross-domestic-product-march-2018-quarter

Population by region:
https://www.stats.govt.nz/large-datasets/csv-files-for-download/