Comp309 Assignment 2

Vincent Yu
300390526

Objectives:

The goal of this assignment is to explore what kind of factors can affect the rental cost in Wellington and explore the relationship of these factors with the number of enrollment of our university.
.

Dataset selection

For this assignment, six datasets were selected to analyze and make predictions. The datasets selected were active-bonds, mean-rents, population, number of enrolments, GDP of NZ and the house value of Wellington. Rental cost should depend on the relationship between market supply and demand. The number of rental bonds can reflect how many houses are available in Wellington. The population of Wellington and the number of enrolment can reflect the demand for the rooms. Normally a large population would lead to higher rental cost. But if the number of available houses is much higher than the demand, it would cause the rental cost decrease. So these two datasets would be good evaluation factors.
GDP shows the primary indicators used to gauge the health of a country's economy. Usually, the GDP rise would indirectly lead to a higher rental cost.
So these datasets are valuable to explore which factor would affect the rental cost of Wellington.
Criteria of selecting datasets are datasets with less missing value, datasets have appropriate number of instances, the datasets should strongly correlate to the class.

Resources & data details:

rental-mean:
https://catalogue.data.govt.nz/dataset/rental-bond-data-by-region/resource/f53e86da-217f-49d9-86cb-cb56ad5cebd3
Active-bonds:
https://catalogue.data.govt.nz/dataset/rental-bond-data-by-region/resource/d477310f-47f1-4aa4-b9ce-84232badfa4b
Details:
These two files include the private bonds and the mean rental-cost starting from January 1993 to July 2018. These two datasets are counted monthly. It contains the data (mean rental cost and number of active bonds) of majorities cities in New Zealand.

Population:
http://nzdotstat.stats.govt.nz/wbos/Index.aspx?DataSetCode=TABLECODE7501&_ga=2.46129063.281699752.1534466641-1077836335.1533857836&_gac=1.218244843.1534213980.CjwKCAjw2MTbBRASE

iwAdYIpsamUN_pJAw8e7FAP27Rg2YNZXG_60uPM7EiyhS0iO6D8ztgBpOgrBhoCUdQQAvD_BwE#

Details:

This data set contains the population by regions of New Zealand. It contains the data of year 1996,2001 and from year 2007 to 2017. It contains 14 years data. The table shows the total population of the New Zealand. It also contains the population of several regions. (Auckland, Waikato, Bay Of Plenty region, Gisborne region, Hawke's Bay Region, Wellington etc).

GDP:

https://www.google.co.nz/publicdata/explore?ds=d5bncppjof8f9_&met_y=ny_gdp_mktp_cd&hl=en&dl=en

Details:

This dataset using the unit of billion US dollar. And it collect the data from year 1993 to year 2017.

Manipulation of Dataset:

Merge datasets:

For this assignment, we can use weka to merge multiple datasets into one and apply to a classifier(linear regression). It can be done by using "simple UCL" in weka. But for this assignment, it was done by excel. The datasets were integrated manually through excel. In the beginning, the dataset was changed counted yearly. But this would cause the number of instances to drop to around 20. So it led to a small dataset. It may cause the final result won't be very accurate.

Pre-processing:

Before import the data into a pipeline, the raw data still needed to be preprocessed. The biggest question was **missing values**. The number of enrollments only collected data from 2008 to 2017. For the missing value in the enrollment dataset, it is not a good idea to use linear regression. Because of the appearance, the number of enrollment has nothing to do with time. Since this, missing data for the number of enrollment were replaced by the nearest value. The house value set only hold five data, and it was hard to predict the missing value. Therefore this dataset was abandoned. The missing values in the GDP dataset were fixed by using linear regression. We can apply a filter called "relplacemissingvalue". But in this assignment, it turned out better use a linear regression in excel because excel can easily illustrate changes by plotting the graph.

The table of datasets before and after manipulation could be found in the appendix.

Feature selection

After the dataset imported into the pipeline through ARFF loader, we could connect

to “data-visualizer” and set the rental cost of Wellington as Y-axis and set the others as X-axis in turn. We desire a linear relationship between rental-cost of Wellington and other relationship. So from the graph, we can roughly find the most correlate feature, which should have a nearly linear slope from the graph. But there is a better way to do this by using explore on the main menu of weka which will be included later.

For the most uncorrelated datasets, we can use a filter called “remove attribute” in Weka, which could remove a specific attribute. In this assignment, it was also applied to another tool in Weka to explore which factor is most correlated. As mentioned before, it is better to use the method in “Explore”. After importing a dataset, we could explore the correlation level by using “correlation-Attribute-Eval”. It showed how likely the attribute correlate to the class.

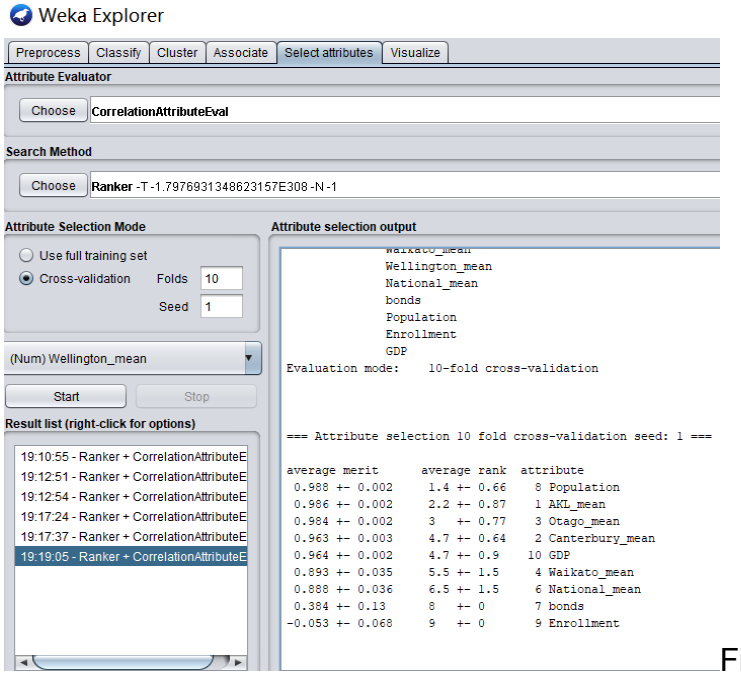

Fig1.1

The average merit represents the correlation coefficient, the amplitude closer to 1, the better. And a positive number means a positive linear relationship, negative coefficient means a negative linear relationship.

As fig 1.1 shown above the most correlate datasets were the population of Wellington, mean-rental-cost of Auckland, mean-rental-cost of Otago, mean-rental-cost of Canterbury, GDP, mean-rental-cost of Waikato and mean-rental-cost of New Zealand is the most correlate feature. The enrollment has no linear relationship with the rental cost of Wellington. And the bonds of Wellington has a weak linear relationship.

From this table, the enrollment feature was filtered out, due to low correlation. But in order to explore which feature could affect the number of enrollment, this feature was set to class later. Weka is a very good tool for data mining, it has pre-implemented “normalize” filter which could help normalize the dataset which could reduce even eliminate data redundancy. It was applied to the pipeline used in this assignment, but it doesn’t vary the result much due to the small size of instances.

Explore importance of attributes:

In this part, used pipeline in weka to explore which feature is the most important, and the combination of attributes to have the best performance. At this stage we have two main goals to achieve, one is the smallest root mean squared error and the other one is the highest the correlation coefficient. Small root mean squared error means less difference between values predicted by the model and the values in the data, which means the model is more reliable. High correlation coefficient means the model is more correlate to the data.

Minimize squared error

$$\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)^2$$

$$0=\frac{\partial\sum\varepsilon_i^2}{\partial\beta_0}=\frac{\partial\sum\left(y_i-\beta_0-\beta_1 x_i\right)^2}{\partial\beta_0}$$

$$=-2\left(n\bar{y}-n\beta_0-n\beta_1\bar{x}\right)$$

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$$

Fig 2

As fig 2 shown above, it illustrates the formula for calculating squared error. The root mean error can be calculated by taking the square root of the squared error.

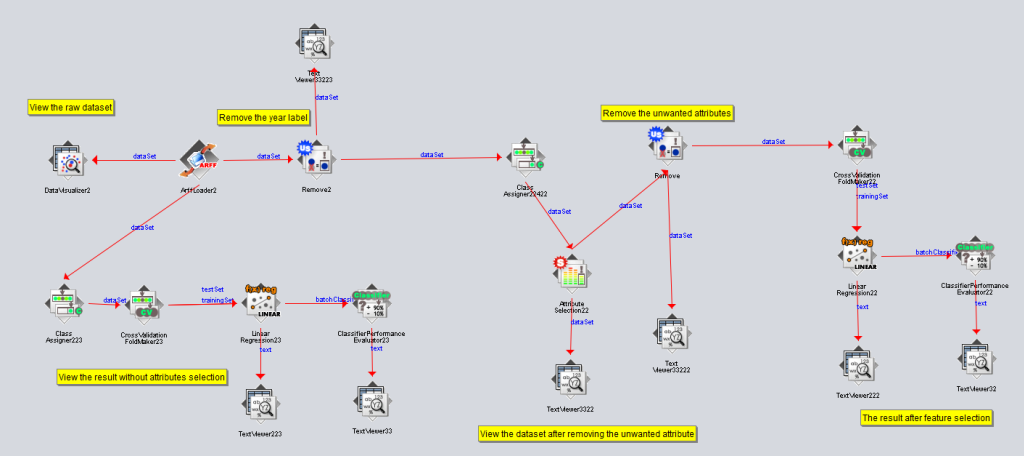

Fig 3

As fig 3 shown above, the pipeline is illustrated. After exploring by several trails, the best combination of datasets was the combination of the population of Wellington, GDP of New Zealand, the mean rental cost across New Zealand, and the active rental bonds in Wellington.

Result:

```
=== Evaluation result ===

Scheme: LinearRegression22 : LinearRegression
Options: -S 0 -R 1.0E-8 -num-decimal-places 4
Relation: RentalCost-weka.filters.unsupervised.attribute.Remove-R1-weka.f


Correlation coefficient                  0.9959
Mean absolute error                      6.5905
Root mean squared error                  7.7304
Relative absolute error                  8.3702 %
Root relative squared error              8.7354 %
Total Number of Instances               26
```

Fig 4

The result for the best combination of attributes is shown in Fig 4. The less correlation attributes were filtered out. This process made the correlation coefficient increase. And it has the smallest root mean squared error.
The correlation coefficient is 0.9959, it is very close to 1. It means the model is very suit this classification. And the RMSE is only 7.7, it means the values predicted by the model is very close to the dataset.

```
Scheme:   LinearRegression
Relation: RentalCost-weka.filters.unsupervised.attribute.Remove-R1-weka.filters


Linear Regression Model

Wellington_mean =

      0.0008 * Population +
     -0.3714 * Otago_mean +
      0.9452 * Waikato_mean +
      0.3106 * National_mean +
   -269.8059
```

Fig 5

Fig 5 illustrates the model found by Weka. The coefficient of population is very small because the values in population attributes are very large, and by linear regression it limits the coefficient to a very small number. We can predict the rental cost with this model. But maybe it won't be very accurate, because there are only four variables. Say we have a very large outlier in the mean-rental cost Waikato attribute, the rental cost of Wellington will be extremely large. And the result will be ridiculous. So if we want to have a accurate prediction we should collect more data and more correlated attributes.

Appendix:

Unfixed dataset:

Year	AKL	Canterbury	Otago	Waikato	Wellington	National total	WellingtonBonds	Wellington Population	Wellington Enrollment	GDP
1993	199.0833	147.5	154.5833	134.1667	177.75	161.9166667	12750.33333	?	?	46.78
1994	216.8333	152.9166667	161.1667	138.6667	184.25	171.25	14108.58333	?	?	55.32
1995	246	163	168.8333	150.5833	194	185.3333333	15168.41667	?	?	63.92
1996	271.8333	172.4166667	169	162.4167	211.25	200.5	12750.33333	426900	?	70.14
1997	277.6667	176.5833333	163.6667	169.5833	222.5	207.0833333	14108.58333	?	?	66.07
1998	269.0833	174.6666667	161	171.1667	229.8333333	206.8333333	15168.41667	?	?	56.23
1999	263.75	171.5833333	179.0833	172	234.8333333	206.3333333	12750.33333	?	?	58.76
2000	266.1667	175.1666667	176.5833	170.3333	236.4166667	208.25	14108.58333	?	?	52.62
2001	275.1667	178.3333333	185.75	171.25	240.3333333	214.0833333	15168.41667	440200	?	53.87
2002	300.3333	191.75	200.8333	178.75	247.1666667	228.0833333	12750.33333	?	?	66.63
2003	324.9167	212.9166667	220.4167	194.1667	256.4166667	246.5833333	14108.58333	?	?	88.25
2004	330.8333	231.4166667	245.0833	210.6667	267.9166667	259.1666667	15168.41667	?	?	103.9
2005	333.5833	241.5833333	258.25	224.0833	277.9166667	268.4166667	12750.33333	?	?	114.7
2006	342.0833	254	261.4167	237.75	295.9166667	280.4166667	14108.58333	466300	?	111.6
2007	363.0833	267.9166667	273.25	253.9167	323.1666667	298.9166667	15168.41667	469300	?	137.3
2008	380.1667	280.4166667	283.75	265.3333	344.5	313	12750.33333	471800	21930	133
2009	382.0833	278	289.3333	268.9167	351.8333333	314.9166667	14108.58333	475600	22925	121.3
2010	394.1667	284.1666667	291.9167	272.6667	363.5	323.1666667	15168.41667	479400	22880	146.6
2011	414.75	302.8333333	310.0833	278.3333	365.25	334.4166667	12750.33333	483400	22560	168.5
2012	430.4167	329.5833333	317	283.5833	371.4166667	347.4166667	14108.58333	485100	21195	176.2
2013	444.8333	366.5833333	327	290.3333	379.5	361.6666667	15168.41667	486700	21480	190.5
2014	463.5833	394.9166667	341.8333	297.5	390.1666667	376.5833333	12750.33333	491400	21190	200.7
2015	493.6667	388.1666667	359	310.1667	400.0833333	394.6666667	14108.58333	496900	21450	175.6
2016	514.5833	385.1666667	385.1667	333.6667	420.5833333	412.9166667	15168.41667	504900	21950	185
2017	534.3333	367.8333333	415	357.3333	444.6666667	431.75	12750.33333	513900	?	?
2018	549.4286	374.1428571	424.2857	216.6667	477	261.5833333	45746.57143	510898.4848	?	?

(*Note: the cell filled with "?" are the missing values*)

Fixed dataset:

Year	AKL	Canterbury	Otago	Waikato	Wellington	National total	WellingtonBonds	Wellington Population	Wellington Enrollment	GDP
1993	199.0833	147.5	154.5833	134.1667	177.75	161.9166667	12750.33333	412103.7251	21930	46.78
1994	216.8333	152.9166667	161.1667	138.6667	184.25	171.25	14108.58333	416059.7951	21930	55.32
1995	246	163	168.8333	150.5833	194	185.3333333	15168.41667	420015.8651	21930	63.92
1996	271.8333	172.4166667	169	162.4167	211.25	200.5	12750.33333	426900	21930	70.14
1997	277.6667	176.5833333	163.6667	169.5833	222.5	207.0833333	14108.58333	427395.6297	21930	66.07
1998	269.0833	174.6666667	161	171.1667	229.8333333	206.8333333	15168.41667	431389.7266	21930	56.23
1999	263.75	171.5833333	179.0833	172	234.8333333	206.3333333	12750.33333	435383.8234	21930	58.76
2000	266.1667	175.1666667	176.5833	170.3333	236.4166667	208.25	14108.58333	439377.9202	21930	52.62
2001	275.1667	178.3333333	185.75	171.25	240.3333333	214.0833333	15168.41667	440200	21930	53.87
2002	300.3333	191.75	200.8333	178.75	247.1666667	228.0833333	12750.33333	448112.4709	21930	66.63
2003	324.9167	212.9166667	220.4167	194.1667	256.4166667	246.5833333	14108.58333	452036.5967	21930	88.25
2004	330.8333	231.4166667	245.0833	210.6667	267.9166667	259.1666667	15168.41667	455960.7226	21930	103.9
2005	333.5833	241.5833333	258.25	224.0833	277.9166667	268.4166667	12750.33333	459884.8485	21930	114.7
2006	342.0833	254	261.4167	237.75	295.9166667	280.4166667	14108.58333	466300	21930	111.6
2007	363.0833	267.9166667	273.25	253.9167	323.1666667	298.9166667	15168.41667	469300	21930	137.3
2008	380.1667	280.4166667	283.75	265.3333	344.5	313	12750.33333	471800	21930	133
2009	382.0833	278	289.3333	268.9167	351.8333333	314.9166667	14108.58333	475600	22925	121.3
2010	394.1667	284.1666667	291.9167	272.6667	363.5	323.1666667	15168.41667	479400	22880	146.6
2011	414.75	302.8333333	310.0833	278.3333	365.25	334.4166667	12750.33333	483400	22560	168.5
2012	430.4167	329.5833333	317	283.5833	371.4166667	347.4166667	14108.58333	485100	21195	176.2
2013	444.8333	366.5833333	327	290.3333	379.5	361.6666667	15168.41667	486700	21480	190.5
2014	463.5833	394.9166667	341.8333	297.5	390.1666667	376.5833333	12750.33333	491400	21190	200.7
2015	493.6667	388.1666667	359	310.1667	400.0833333	394.6666667	14108.58333	496900	21450	175.6
2016	514.5833	385.1666667	385.1667	333.6667	420.5833333	412.9166667	15168.41667	504900	21950	185
2017	534.3333	367.8333333	415	357.3333	444.6666667	431.75	12750.33333	513900	21950	205.6
2018	549.4286	374.1428571	424.2857	216.6667	477	261.5833333	45746.57143	510898.4848	21950	213.26

(*Note: the data in red are fixed by linear regression, the data in blue are fixed by choosing the nearest value*)