Comp309 Assignment 3
Zhanghao Kong
300432074

2.2 Core: Exploring and understanding the Data [40 marks]

• (20 marks) Highlight the findings of your dataset exploration.

Business understanding:

Data mining goal: The overall aim of this assignment is to predict the price of electricity in Wellington.

Core: The model I build needs to be able to predict the price of each instance.

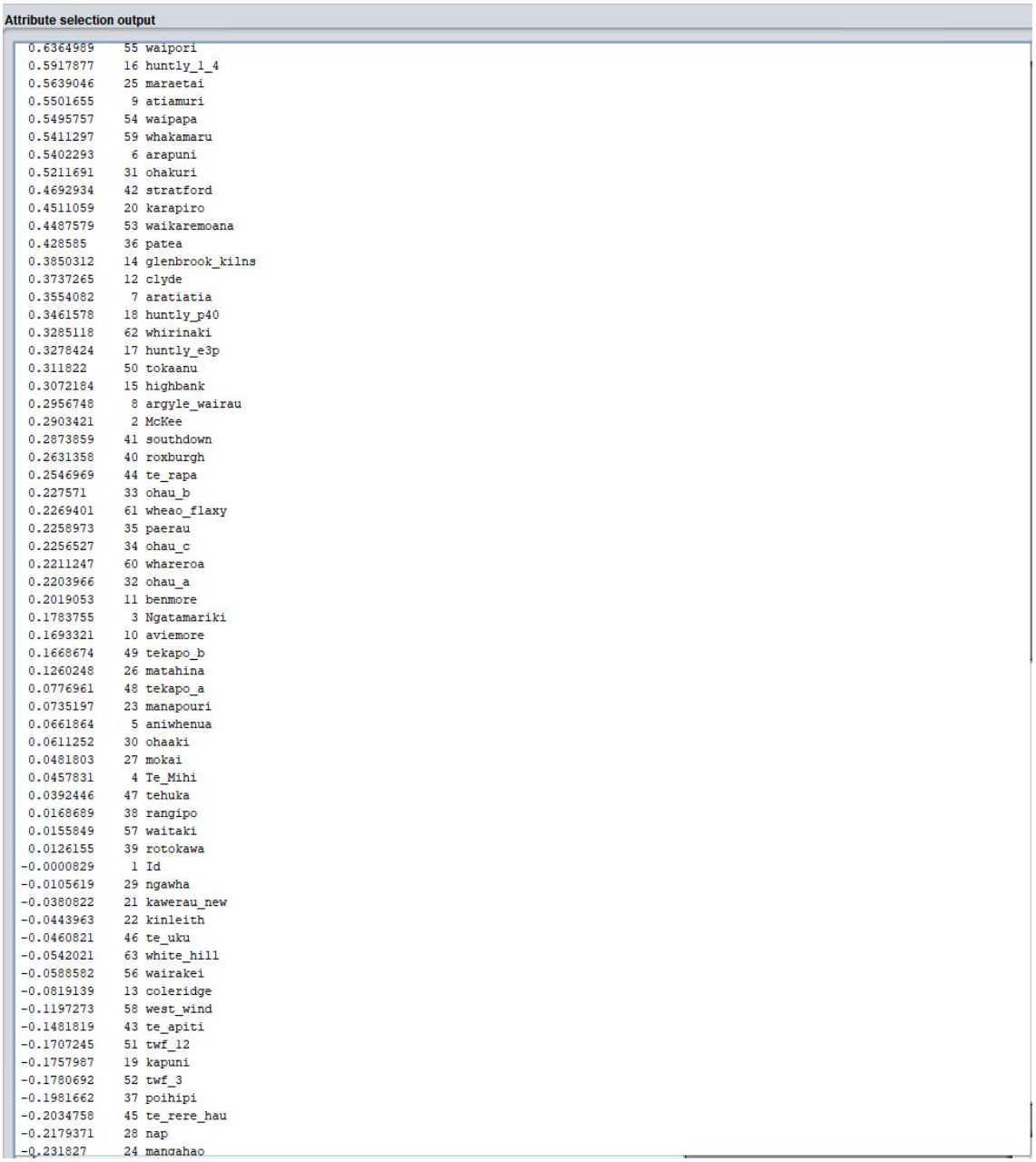
Data understanding:

With the given training set, there are 64 attributes(ID,62 place names and Price). Testing set has 63 attributes excluding the Price attribute. The ID represents the number of each instance. 62 place names represent the generation of a given power station (in kWh) for a given 30-minute time period. The Price attribute represents the price of electricity (measured in $/MWh).

Data quality:Data quality is quite high because it has many attributes without any missing values. All attributes are related to the goal attribute(except ID for testing). Amount of data for training is big enough.

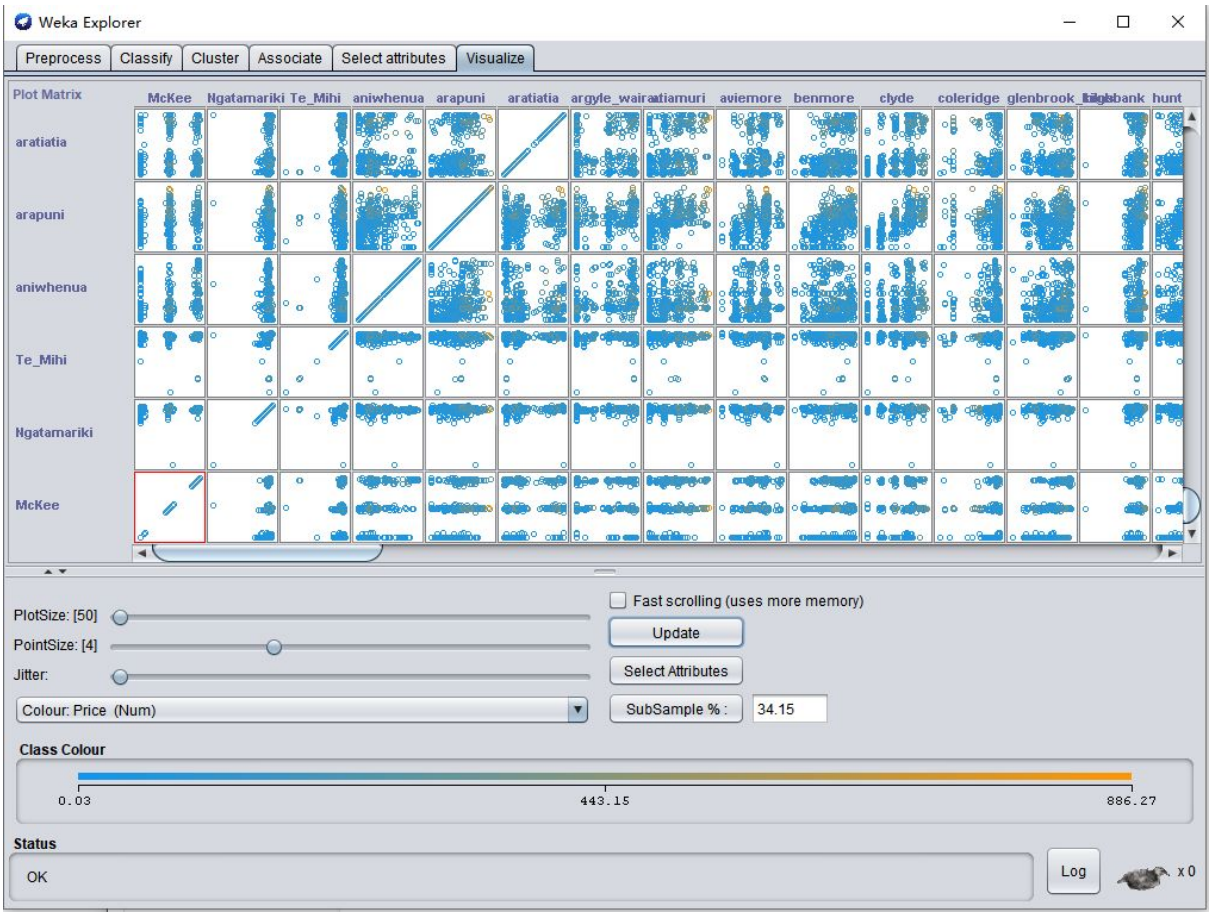
Data completeness: No missing value so have a good data completeness.

Understanding of Features:In the given data sets, attribute “ID” is not related to other attributes, it’s just used for marking instances. All other attributes except “price” and “ID” are not related to each other, they just record the name of the power station. But they all have a relation with “price”. Below is the Correlation Ranking Filter method in Weka. All attributes in the data set have a less or more effect on “price”.



• (20 marks) Visualisation is an important aspect of this task. Please illustrate at least one important finding of your work.

Here is the screenshot of Weka visualisation.



We can easily find that there is a straight line across the whole graph.This represents a normal distribution in the two-dimensional graph. And I found this line comes through the whole graph, which represents that all attributes are related to attribute “price”. This can approve that the data quality is quite good and it’s suitable for being a training set.

2.3 Completion: Developing and testing your machine learning system [50 marks]

• (15 marks) Discuss the initial design of your system

Data preparation:

1. Select Data: I use weka to help rank all attributes.

I choose to use the Correlation Ranking Filter to rank all attributes and the result is below.

I set the attribute “price” as the supervised attribute then run the function.

We can easily find that the only attribute “ID” takes a very few effect on the supervised attribute “price”. So I choose to delete it. But when i want to submit the file, i will add it back. For other attributes, they have more or less influence on “price”, so I keep them for training.

2.Format data: in this data set, all attributes are numeric, so i decide to keep it original state.

3.Build attribute “price” in test data set. Otherwise the date structure is different between training set and testing set, Weka won’t work.

BG	BH	BI
hirinaki	white_hil	Price
0	10215.2	0
0	22490.5	0
0	0	0
0	12070.13	0
0	24409.3	0
0	9356.141	0
0	273.4061	0
0	7520.01	0
0	24439.86	0
0	962.6453	0
0	17220.11	0
0	2714.621	0

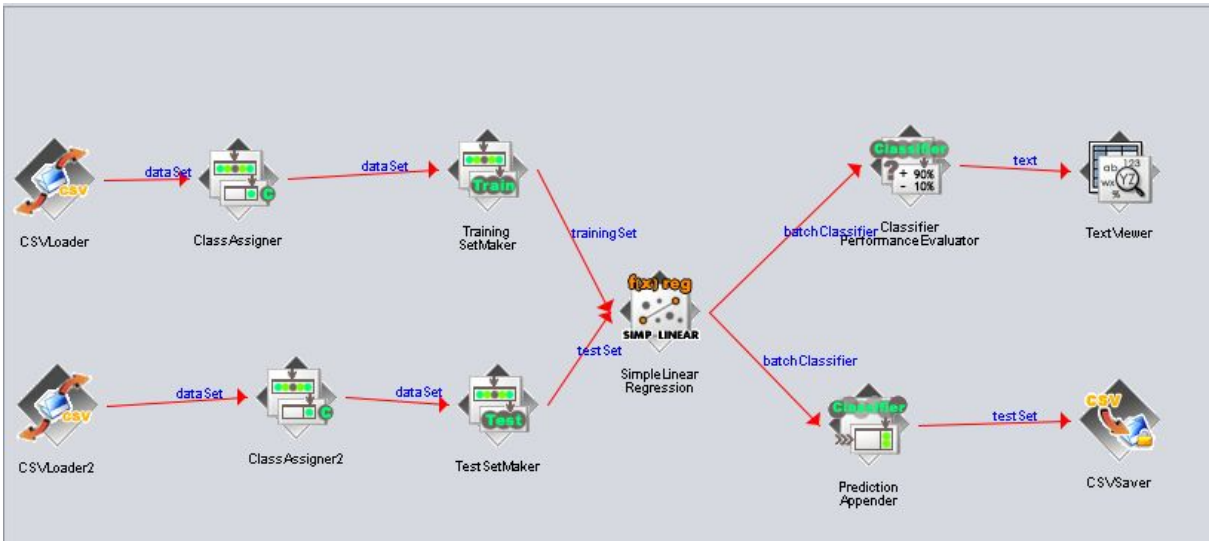
Method Choosing:

I choose Regression because when attempting to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables). In this data set, one dependent variable is “price” and others are independent variables.

In my pineline, I choose Linear Regression as a method to training and testing the data set. Here is pipeline:

As shown in my pipeline, I didn't make any change to the data set because the data set is perfect which has high quality and good completeness. Each independent attribute has a relation with the dependent attribute(“price”). So I use them directly in my pipeline.

Result is bad so i think maybe try another method for testing.



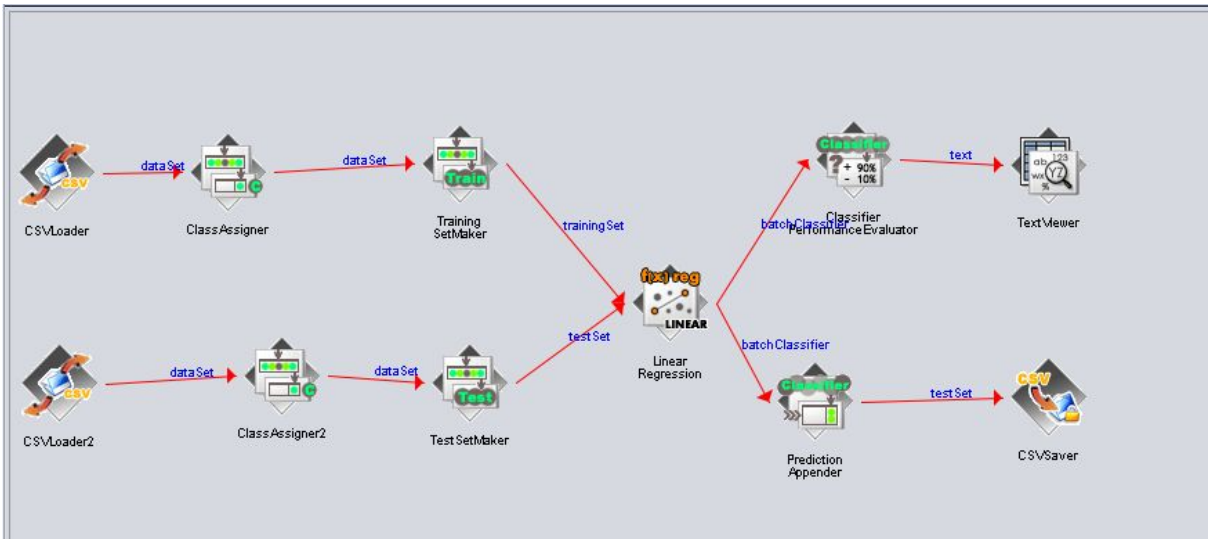
sub4.csv 5269.34666

5 hours ago by Adam Kong

by using SimpleLinearRegression.

• (15 marks) Discuss the initial design of your system

After trying Simple Linear Regression,I realized simple linear regression is suitable for one independent variable and one dependent variable.But in this data set, it contains many independent variables. So I chose Linear Regression as the next method for testing.



sub2.csv 1865.25526

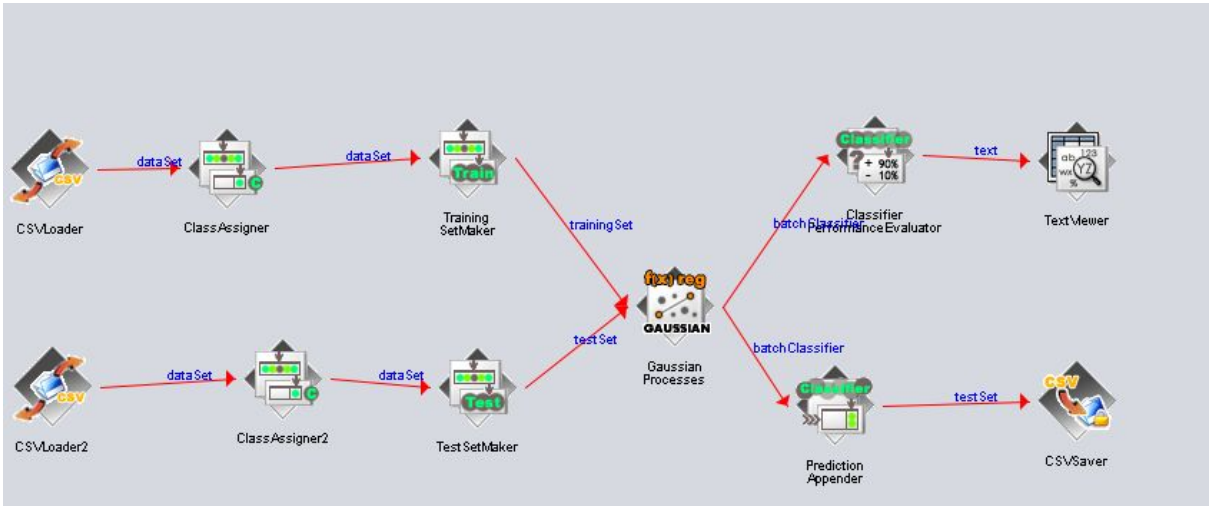
6 hours ago by Adam Kong

submission 2 use linearRegression.

And with this pipeline, I got a much better result.

However the result is still not good enough and there are still some methods I didn't use.

By visualisation, I found the graph has a normal distribution. The linear combination of random variables in the Gaussian process obeys normal distribution, and each finite dimensional distribution is joint normal distribution. So I used the Gaussian process as the next method.



sub6.csv 1839.42415

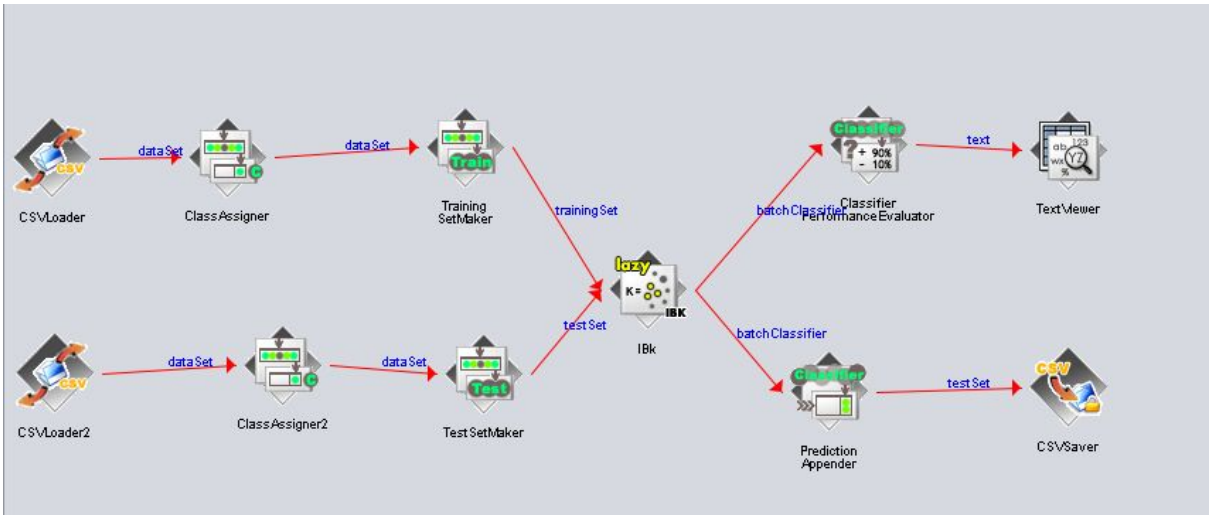
5 hours ago by Adam Kong

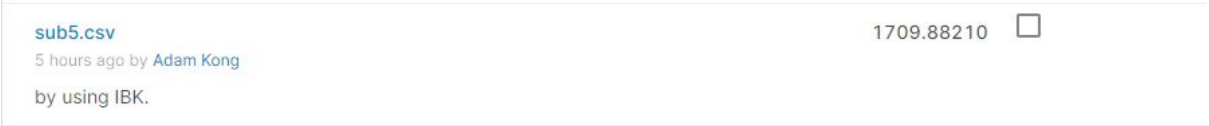
by using gaussian processes.

It improved a little bit but almost the same with linear regression.

• (10 marks). Use your judgement to choose the best system you have developed.

Finally, I chose ibk because in this dataset, there are almost no noisy or irrelevant features. So ibk can be a suitable method for this data set.





But for the KNN algorithm, a good k value can make the result more accurate. So I changed the k value many times and got K = 3 as the most suitable value for this data set.

I chose ibk in my final system because it fit the data set best. It can be used for regression and have a k value which can reduce errors in testing. Besides, this data set has an outstanding quality which all attributes are relevant with goal attribute(“price”) and no missing value. So it could be the perfect method used in my system.

2.4 Challenge: Reflecting on your findings [10 marks]

The model i finally chose is based on ibk, which is easy to explain: For this data set we need to figure out the relationship between “price” and other attributes. Regression is quite suitable to analyze this kind of problem. Ibk can detect most close neighbours with the “price” by k value. It can help calculation reduce errors and noise. Good results come naturally.

Based on my findings, I suggest that the government should promote and support new energy power generation. With the increase of electricity consumption, the demand for electricity is increasing. Promoting new energy power generation (wind power and hydropower) can make the electricity price cheaper and pollution-free. Wind and hydropower in New Zealand are easier to obtain, so they should be promoted to lay a foundation for reducing electricity costs.