# Comp 309:

Assignment 3:

Keiryn Hart

300428418

2.2 Core:

Exploring and understanding the Data:

The first thing I decided to do when investigating the data and doing some exploratory data analysis was to look at the completeness of individual features. I first looked for variable that contained NA values, in doing this I identified that there was only one variable that seemed to contain NA values, that being twf\_3, it had a total of 108 missing values which accounts for just under 10% of the data, when looking at the distribution of the data it seemed that the values were relatively evenly distributed and therefore I think that imputation of a mean could be a good way to keep the variable without having a major effect on it.

Next I began to notice that there seemed to be some variables that contained a lot of instances of 0 when I looked into it there were multiple variables whose data comprised almost only of 0’s these variables were Southdown, Te Rapa, Whirinaki, Mangahao and tekapo\_a each of these variables had over 1000 instances of 0. I think that the reasons for these variables and others having a lot of instances of 0 is because these power stations may not be in operation all the time, interestingly each of these power stations was either a hydro or gas as their fuel type which indicates that the hydro power plants may only be operational when water levels are high enough, when the dam is open or for seasonal reasons. This seasonal pattern could be the reason that some of these gas power stations are not working very often, they produce a lot of greenhouse gasses and therefore may only be necessary when other power stations such as hydro or wind farms are not producing enough.

A screenshot of a cell phone

Description automatically generated

Correlations:

I then began to look at correlations and more specifically correlations between price and the different power plants, shown above is a correlation matrix for the 5 power stations that had the highest correlations with price, these being Waipori (Hydro), Waipapa (Hydro), Maraetai (Hydro), huntly\_1\_4 (Coal) and Atiamuri (Hydro). It shows that the highest correlation a variable has to price is Waipori at 0.64, it is more interesting to note though that there are some very high correlations between the 5 power stations with Maraetai and Waipapa having a correlation of 0.89 and Atiamuri having correlations of 0.76 and 0.77 with Waipapa and Maraetai respectively.

These higher correlations could be due to the fact that these variables tend to have relatively similar shapes to their data where they hardly appear to be operating at max output at all but are more often operating at half or less of their max.

I think that it will be important to include at least a few of these variables with high correlations to price as they should be good predictors because of their high correlation, Waipori is likely to be one of them as it has the highest correlation with price out of any variables.

Other interesting patterns:

Like the variables with very high instances of 0 there appears to be a similar pattern going on among some of the variables in the sense that a lot of them appear to have a relatively small range where there appears to be a large majority of their data occurring. For example, in Ngawha who as shown below has almost all of its data ranging between 11,000 – 13,000. Throughout the dataset there are other variables with similar distributions, but their output is different in regard to kWh. I think that a combination of variables like this could be a good way for making accurate predictions of price by covering all the various output levels.

A screenshot of a cell phone

Description automatically generated

2.3 Completion:

Initial design:

My initial design is the pipeline that is shown below, I decide to go with a linear regression model due to the fact that price is a continuous variable and linear regression is good for predicting continuous variables.

The attribute selection in the top left corner was used as a reference for the significant correlations between the variables and price as I touched on in the above.

Moving along the pipeline to the remove section, this part of the pipeline was where I chose the variables that I wanted to use in my predictions, at this stage with the research I had done and the information I had identified I decided that a good starting point was firstly to keep the 5 variables with the highest correlations with price (Waipori, Waipapa, Maraetai, huntly\_1\_4 and Atiamuri). Then I also decided to keep the variables that I believed had some form of seasonal pattern through their distributions the variables that I chose from this ended up being Clyde, Aviemore and whakamaru.

I then made the labelled dataset the training set and the unlabeled set the test set and ran it through the linear regression model, appended the predictions, and removed attributes to save the csv for submission.

This initial design was definitely not the most accurate design as I found out once I submitted it to the Kaggle, I believe this to be due to the fact that I did not have very many predictor variables which meant that the model was most likely too simple and did not have enough information to make informed and accurate decisions.

A close up of a map

Description automatically generated

Following on from this in my next attempt I did not change the pipeline at all and everything remained the same except for the variables that I used in my prediction, I kept all the variables used previously in the first attempt but proceeded to add a few more variables. The next variables that I decided to add were firstly the remaining 4 variables out of the 10 variables with the highest correlations with price as I thought that these variables may assist in the prediction given that they had a moderate correlation still, these variables were Arapuni, Ohakuri, Stratford and Karapiro. I also decided to add the variables with a large amount of 0 values that I touched on earlier in my exploratory analysis. My reasoning for this was that I noticed when looking at their non-zero instances that the relative price was either substantially higher than the average or quite a bit lower, I thought that this was interesting and may help to form better predictions.

This resulted in a slightly better model, but it was still not performing exceptionally well, so I decided to use a different algorithm.

Intermediary Stage:

after assessing the results from my initial couple designs on the Kaggle Leaderboard I saw that my design was still not making the best predictions, so, as I said previously, I opted to change my classifier, choosing to go with a random forest classifier as it can also work as a sort of regressor. After switching my classifier to random forest, I noticed a massive difference straight away in my Kaggle results. Although there was a large improvement my results were still not amazing, after this I decided to change up my predictors and to help inform my decision better I used the attribute selector filter with best subset evaluator to select what weka deemed to be the best predictors and with these variables alongside some variables that I still considered to be important. This left me with the variables; Ngatamariki, Glenbrook\_kilns, highbank, huntly\_1\_4, Kapuni, Karapiro, Southdown, Te\_rere\_hau, Waipapa, Waipori and whirinaki. After running the classifier and submitting the dataset to Kaggle I got the lowest mean squared error so far of 1247 which I believe is a step in the right direction for the predictor variables I will be wanting to use for my final design.

From this I tried adding a few more variables to the model to aid in prediction, these variables were tekapo\_a, te rapa and mangahao, this attempt really did not improve the mean squared error too much with a decrease down to 1240. These scores were good, but they were still not placing extremely high on the leaderboard and any minor adjustments I made to my pipeline only resulted in either an increase in my mean squared error or a slight decrease where at best I reached a score of 1143.

I tried adjusting the attribute selector filter to different parameters for their evaluators and search methods such as using the classifier attribute evaluator with ranker and choosing the top attributes that came out of that but again the scores did not change much and in most cases the mean squared error got worse.

A close up of a map

Description automatically generated

This led me to making other small adjustments which eventually ended with my final design.

My final design is much the same as the designs for my intermediary stage but some small adjustments have been made with the most notable being the addition of another stream in the pipeline which included the Attribute Selected Classifier which essentially allowed me to run a best subset evaluation for the dataset based on the algorithm that I was wanting to use which was in this case a random forest. Once the attribute selected classifier had produced a subset of the best predictor attributes for the random forest, I ran those attributes through my random forest classifier. after uploading to Kaggle this technique had produced a score that was far better than any of my previous pipelines and it still contained many of the variables that I deemed to be important in my exploratory analysis so I decided that this model would be my final design.

A close up of a map

Description automatically generated

The variables included in my final design are shown below:

A close up of text on a white background

Description automatically generated

Something interesting about the data in the variables included in this model is that a large majority of the variables have a relatively specific range in regard to the amount of energy they produce meaning their distributions are very tight and the range is not very large, I believe this is an interesting point to bring up as this may be important for building a model that makes good predictions for the price of power as each variable has its own range of output which for the most part is independent of the other variables meaning that together they account for all the possible levels of output in kWh.

Like in my exploratory analysis a few of the variables with high counts of 0 were featured in this model as well as features that shared a high correlation with price, this helps me to believe that my model is making good predictions because the patterns and trends that I initially identified in my exploratory analysis as possibly important did turn out to be variables that assisted in prediction.

2.4 Challenge:

I believe that my model is not the most easily interpretable model, this is because of the fact that the algorithm in use is a random forest which is a combination of multiple decision trees, also given the fact that random forests are usually used for classification the tests on each attribute may not be the easiest to understand given that all the variables are continuous numeric values.

There may be some ethical concerns with the way that my model chose its features for predictions, firstly in regards to the environment a number of the chosen features in my model are power stations that use natural gas to produce their electricity, the problem with these power plants is that they produce a lot of greenhouse gas which is extremely bad for the environment.

There could be societal issues surrounding the power stations included and not included in this model as people may interpret the power stations that are not included in this model as not necessary which could lead to power stations shutting down because they are seen as redundant. If a power station were to shut down, especially in smaller areas with small towns it could lead to a large amount of people losing their jobs where they may not have many other options.