April 2020

Written test for the position of Data Scientist

Thank you for applying to the Data Scientist position at Vector. Your profile passed the 1st screening and we would like to invite you to participate in the written test.

The next phase of the recruitment process will involve a technical interview and a human resources interview.

Before starting with the written test, please read the following instructions carefully before starting:

- Please respond to the questions in this document and return it along with an analysis script, while including your first and last name to file/document name. You ask you to please return your response by Monday 4th May midnight to marnie.simmonds@vector.co.nz
- This written test has two parts: 1) analytical thinking and 2) technical analysis
- For the technical analysis part, you have also received 2 data files (more information in the document below)
- You are expected to do the test on your own without the assistance from another person.
- As in the position of data scientist, you are allowed and encouraged to find and learn from the best by using internet searches, but referencing this material as needed is seen as essential.
- We ask you to not ask for clarifications to the questions asked and respond based on your understanding and knowledge base. Clear and concise explanations will ensure we can follow your reasoning. Good data-driven observations and comments in your response are encouraged.
- Please do get in touch if you can't access the 2 data files

Applicant Name: Jake Cherrie

1. Analytical thinking

Please choose your preferred question among the following options and respond in words only (and possibly schematics/equations) of how you would design your analysis. Your written response should be roughly half a page only (about 250 words), but strictly less than 1 page if schematics are used.

Option 1

Vector wants to carry out a peak-time rebate trial, which consists of sending customers a notification to reduce energy usage on a day where Auckland's energy use is expected to be high. If a customer manages to reduce energy usage, they receive a payment as reward, otherwise they receive nothing. The project lead has a certain budget at their disposal to pay as rewards. The project lead is looking to you to provide a written description of a pragmatic approach to identify if customers have responded to the notification and by how much, so that they can divide up the reward payment in such a way that it incentivises a customer to continue to reduce energy use when these events are called. The project lead has smart meter data (electricity consumption (kWh) in 30-min intervals) for each customer at your disposal, but if you need additional data then please highlight that too.

Option 2

Electric vehicles are a key technology to decarbonise the energy system, but the network needs to plan to integrate this new and relatively large load. The network planning manager provides you with the latest government forecast for electric vehicles in Auckland for 2030 and wonders if that total number will be evenly spread across residential customers. The manager asks you to prepare a plan for a new spatial forecast of where EVs are more likely to be adopted in the future. You are provided with a data set of annual EV adoption by suburb and some socio-demographic information, but if you need additional data then please highlight that too.

Option 2: Development of a spatial forecast of EV adoption

Step 1: Digest the problem statement and investigate existing methodologies.

- Talk to the network planning manager to make sure you lock the requirements down (do they need a full 10 year forecast?, do we care about seasonal variations?, time constraints?).
- Reach out to the team behind the government forecast to see if they have looked into spatial dependence. If they have not looked into any spatial dependence can their methodology be leveraged, extended and validated to come up with a spatially dependent model.
- Research other methods and sources; there are papers that address almost this exact problem such as (Heymann et al., 2017) which suggests the use of diffusion theory.

Step 2: Obtain additional data and clean and merge the datasets.

- EV price data, competing vehicle price data and adoption, will be helpful (I suspect EV uptake will be closely tied to free income and price).
- Make sure you have all relevant socio-demographic data in the dataset supplied.
- Explore and clean the data; see what useful features can be engineered.

Step 3: Explore and test algorithms and methodologies.

- Split out a testing and a validation set (making sure they are chronologically ordered).
- Decide on an appropriate accuracy metric.
- Test some algorithms. Depending on your research maybe ARIMA, MCMC, recurrent neural networks and other methods identified in step 1.
- Test and compare the models on the testing set before selecting a methodology.

Step 4: Refine and improve the models.

- Further feature engineering (not required for NN).
- Clustering the suburbs by socio-demographic profiles if data is scarce.
- Apply seasonal adjustments, autocorrelation.
- Hyper parameter tuning.

Step 5: Validate the model.

- Validate the model on the validation set.
- Hand the model over for peer review and validation.
- If everyone is happy, move to step 6 otherwise repeat steps 1-5.

Step 6: Productionise the model.

- Create adequate documentation and ensure the analysis is source controlled and recreatable (Docker).

2. Technical analysis

We will ask you to do a short piece of analysis and document and visualise your results in this document, while also returning a script with your analysis (Python / R script or Excel spreadsheet)

Residential electricity load profiles

You have been provided an historical sample of residential smart meter load data for 50 residential customers and temperature data for Auckland for the full year 2015. Smart meters measure electricity consumption (kWh) in 30-min intervals. You have also received hourly temperature data for Auckland in 2015.

You are asked to please complete the following 2 parts of analysis:

Part 1: The customer connections team is helping a property developer to size infrastructure for a new residential development and wants to know what the load as a group may look like and what variations may be expected. They ask you to determine a typical daily load profile(s) (24 hours from midnight to midnight) for winter and summer at aggregate level. Please comment in the script on how you chose to define typical and why.

Part 2: The customer connections team also considers asking the developer to put additional insulation into certain houses but given the current budget constraints due to COVID-19, it knows that the developer can only do this for the houses with the highest gain. Please determine the 10% of customers who will benefit the most.

The analysis can be found in the attached notebook and the full project can be found in the github repo here.

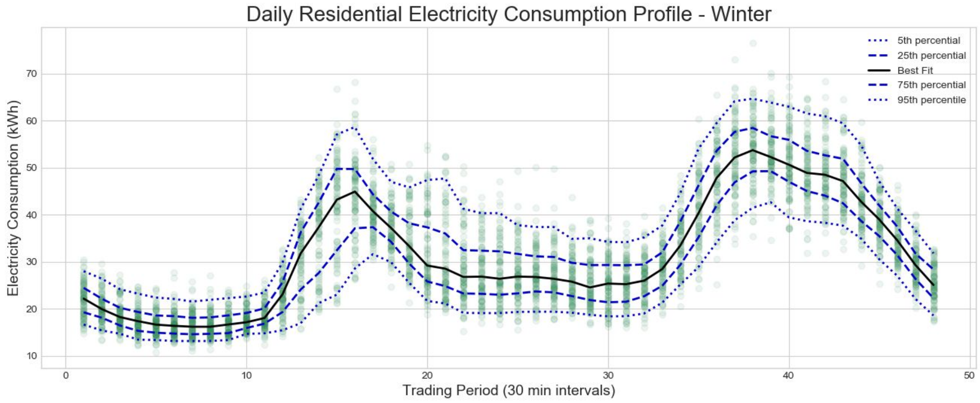
Part 1: Determining Typical Electricity Daily Load Profiles

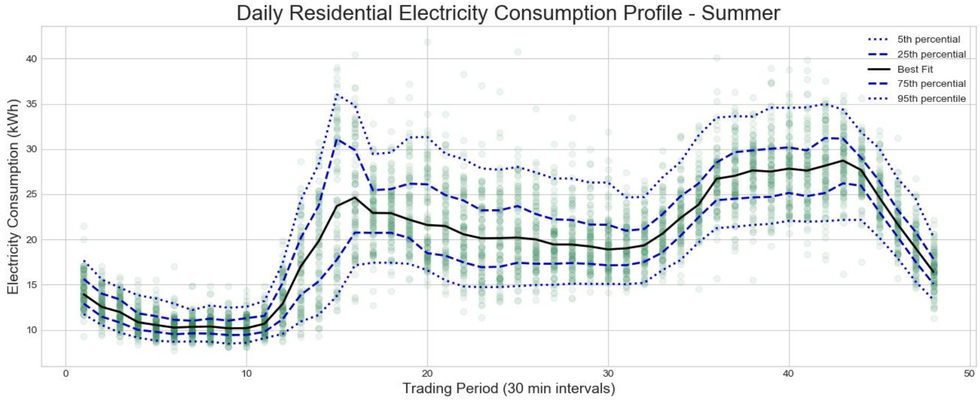
In order to determine a "typical" daily load profile of the aggregated sample (the consumption of all 50 sampled ICP Keys) a gradient boosted regressor was used to find the best fit for the daily electricity consumption profile over the summer and winter periods (further optimisation would be beneficial).

The expected variations in the profiles are captured with quantile gradient boosted regressions at the 5th, 25th, 75th and 95th percentiles.

Some exploration was done on removing outliers such as weekends or public holidays but it was not included in the final analysis because the outliers will inform the variations captured in the quantile envelopes.

The daily load profiles can be seen in the figures below:





Part 2: Determine The Most Beneficial Customers For Insulation

In order to find the "houses with the highest gain" or "who will benefit the most" a definition of gain and benefit in this context needs to be established.

While it may be debatable I will assert that the reduction of the overall electricity consumption is not as important as the number of people and specifically how many vulnerable people (young and elderly, etc.) would be positively impacted by having insulation. I also suspect that the less wealthy households (maybe the ones with more vulnerable people or dependents) have tried to save money on their electricity bills and, therefore, the historical meter data will not show the benefits that those households would get.

A thorough and holistic approach would be to collect additional data such as, internal temperature and socio-demographic information (number of people, age, income, expenses, etc). I expect that socio-demographic data would be obtainable from sources like Statistics NZ; including data on households whose power can't be turned off due to medical or other reasons. The household's temperature dependent electricity consumption could then be modelled (with non-linear feature dependence) and decomposed to find the electricity consumption required to maintain a comfortable temperature while removing the impacts of socio-demographic factors like income and expenses. The benefits could then be weighted to account for the number of people or more specifically the number of vulnerable people.

Given that the data provided (or readily obtainable) does not contain this information I've reframed the problem to identifying which households have the most temperature dependent electricity consumption (this is what the insulation will theoretically reduce).

To address this reframed problem with the data available the following method was applied:

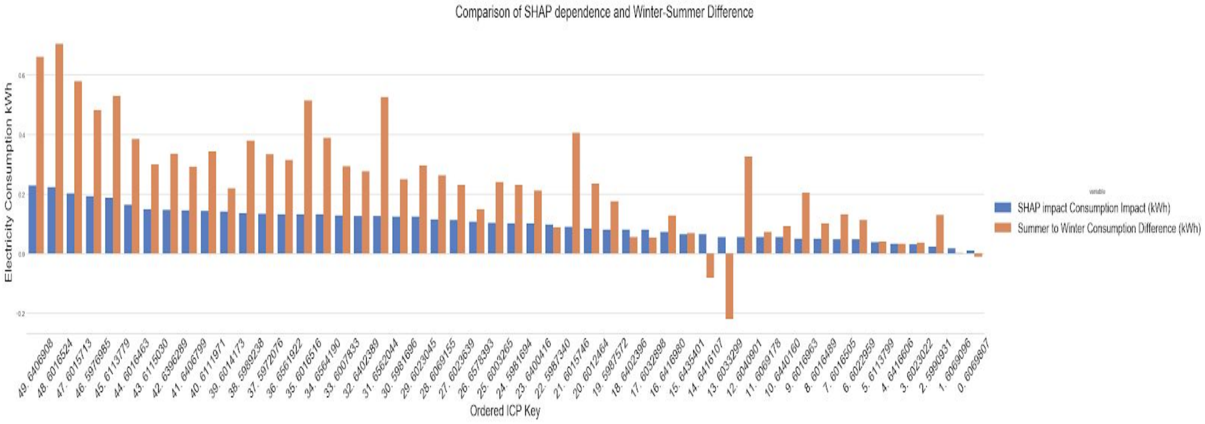
To align the time scale and metric types in the meter and temperature datasets a pchip interpolation and mean aggregation was applied to the temperature data to find the average temperature for each trading period before merging them into a single dataframe.

From there two separate approaches were tried:

- Firstly, a very simple calculation of the difference in consumption between summer and winter periods for each household (this is a very naive solution as discussed in the attached notebook).

- Secondly, a model of the electricity consumption was constructed for each household, using a random forest regressor, before applying SHAP analysis (read about SHAP here: https://github.com/slundberg/shap) to decompose the impact of temperature. This is a more complete method and goes some way in addressing the flaws of the simple calculation as discussed in the attached notebook)

Both methods returned 4 out of 5 of the same households (5 being 10% of the 50 households sampled) which was reassuring but due to the shortcomings in the first method I would recommend that the property developer insulate the households with the closest similarities to the ones with the ICP Keys returned by the SHAP dependence (6023045, 6016524, 6113779, 6015713 and 6406908).


Comparison of SHAP dependence and Winter-Summer Difference
Electricity Consumption kWh
variable
SHAP impact Consumption Impact (kWh)
Summer to Winter Consumption Difference (kWh)
Ordered ICP Key