COMP8851 – Internship Report

The Importance of Interpretability in Modelling Urban Water Systems

Water | NSW Department of Planning, Industry and Environment

Francesco Palermo, Student ID: 45539669, Session: 2/2020

Submission Date: 07/11/2020

Macquarie University

2. Acknowledgement

I would like to thank all those who assisted in my placement in my internship and this internship report. Firstly, Mohammad Mortazavi who was my supervisor, and who was one of my main stakeholders for this project.

Also, I would like to thank my professors who have encouraged me and answered my questions throughout this project. I would like to appreciate my fellow students who have supported me as well, by providing peer support as I complete this internship.

I would not be able to accomplish as much without the support of this network.
Dr. Amin Beheshti assisted in placing me in this opportunity and deserves recognition and acknowledgment as well.

3. Executive Summary

The industry-based internship unit is a part-time (20-25 hours per week) opportunity that allows students to gain practical work experience before the end of the Master's program. During this time, I developed professional work skills such as professional workplace communication, self-supervision, time management and ethical behaviour.

The Department of Planning, Industry and Environment gave me the opportunity to apply the theoretical knowledge learned in Macquarie University into a real-world data project, which enabled me to build on previous knowledge as well as build new skills.

The central project that I concentrated on through my internship was providing a report and commentary on **the importance of interpretability in modelling urban water systems**.
The current system used in the Department of Planning, Industry and Environment to estimate yield for the Sydney water supply system can be considered as "black box" model.

The interpretability of a black box model has been accomplished by using **interpretable models** or **model-agnostic interpretation tools**. The extraction of rules and insights from this analysis enables decision makers to better plan for future actions

This paper is organized as follows: **Sections 5 and 6** briefly overviews the history and policy of the organization and describes the organizational structure of the business. This includes the Organizational Hierarchy Chart and the main offices. **Sections 7 and 8** provides information regarding the internship schema and the detailed description of the project assigned. **Section 9** presents a comprehensive report of the project which is divided on a weekly basis of activities and achieved tasks. **Section 10** describes two work samples in detail. **Section 11** shows Critical Analysis, and various challenges faced during the project. **Section 12** illustrates the SWOT Analysis. **Sections 13 and 14** presents the conclusion and future recommendations.

Finally, the analysis has been developed in Python and R. The reader can find some snippets of code throughout the paper. If my supervisor will approve the public sharing of my analysis, the entire work might be shared in GitHub at the end of the internship.

Table of Contents

5. Overview of the Organization

The New South Wales Department of Planning, Industry and Environment (DPIE) is the department of the New South Wales Government responsible for planning and the development of industry in NSW. The former Department of Planning and Environment and the former Department of Industry were dissolved on the same date as the new department was established in July 2019.

The NSW Department of Planning, Industry and Environment – Water group, is responsible for ensuring sustainable, secure, and healthy water resources and services for NSW. They manage surface and groundwater and develop, implement, and manage plans for water security and supply across all of NSW, in both metropolitan and regional areas. They deliver efficient services and focus on innovation, performance, and affordability.

They are responsible for driving delivery of: the **2017 Metropolitan Water Plan** and the **Water Reform Action Plan** (established on recommendations after a compliance review called after allegations of mismanagement of water resources and misconduct).

Since the establishment of the Water Reform Action Plan, in recent years the NSW Department of Planning, Industry and Environment – Water group has established a new Lands and Water division, established a new independent regulator (NRAR), and increased transparency in water management. The goal is to introduce best practice for water management, ensure transparency in how they allocate and manage water, build a compliance framework that ensures regulations and to build capability to support implementation of water reforms.

A separate initiative from the **2017 Metropolitan Water Plan** is the **WaterSmart Cities program**, which is an integrated approach to water planning across both the metropolitan region and expanding communities. The WaterSmart Cities program establishes resilient communities with an integrated approach to providing water, wastewater, and stormwater services.

Another key responsibility of the NSW Department of Planning, Industry and Environment – Water group is managing the **Drought Response Strategy**. As our climate continues to change, resilient water supplies that are resistant to shock are important to keeping communities liveable. The Drought Response Strategy was created to withstand extreme drought and is characterized by supply and demand management measures that are rolled out according to set water level triggers. Some examples of these measures are implementing water restrictions and building new water supplies.

The **Sydney Desalination Plant** is another feature in the **Drought Response Strategy** – without it, Sydney's water supply would depend entirely on rainfall. In January 2019, the Sydney Desalination Plant was returned into service to supplement increasingly low water levels and today total dam levels are 96.5%. The Desalination Plant maximizes the production of drinking water when catchment levels fall below 60% and until they get back up to 70%.

6. Organizational Structure

a. Organizational Hierarchy chart

Our Leadership Team

The department's executive is collectively responsible for leading the organisation, and individually accountable for the strategic and operational activities of their specific areas.

- Jim Betts, Secretary, Department of Planning, Industry and Environment
- Kirstie Allen, Deputy Secretary, Strategy and Reform
- Greg Woodhams, Acting CEO, Greater Sydney Commission
- Paul Grimes, Coordinator-General, Environment, Energy and Science
- James Hebron, Chief Legal Counsel
- Jody Broun, Cluster Lead: Aboriginal Strategy and Outcomes
- Alex O'Mara, Deputy Secretary, Place, Design and Public Spaces
- Marcus Ray, Deputy Secretary, Planning and Assessment
- Alison Frame, Deputy Secretary, Housing and Property
- Jim Bentley, Chief Executive Officer, Water (Deputy Secretary)
- Shaun Smith, Deputy Secretary, Corporate Services
- Kiersten Fishburn, Coordinator-General, Planning Delivery Unit
- Sally Anne Friedlander, Deputy Secretary, People, Performance and Culture

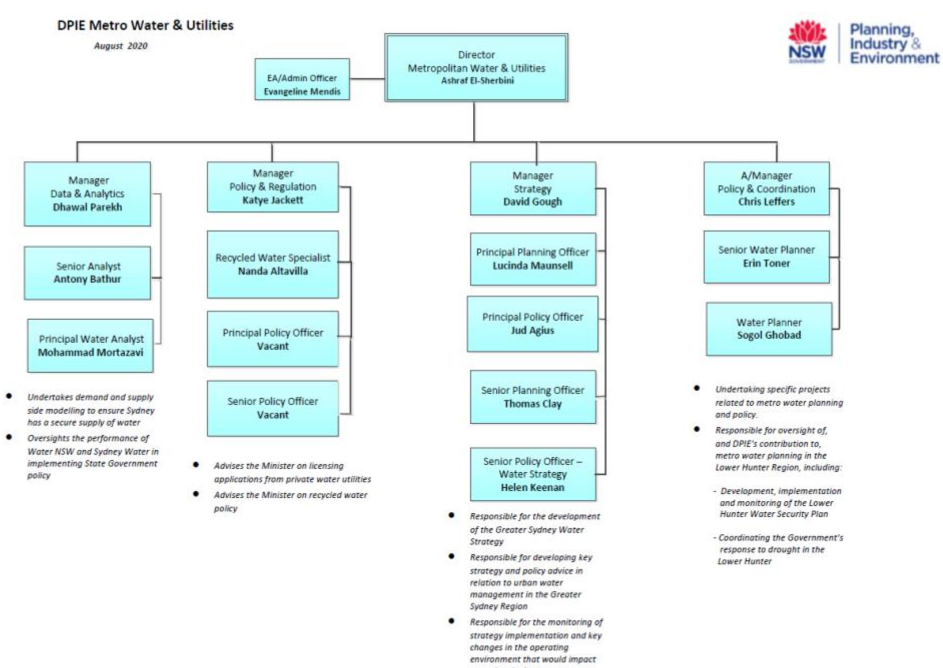


Figure 6.1 – Organizational Chart

b. Number of employees

There are about 16000 employees at the DPIE level and about 450 staff in Water group. There are 14 people in Metro Water & Utilities group.

c. Main offices - Parramatta

d. Introduction of all the Departments

The Department of Planning, Industry and Environment (DPIE)

The Department of Planning, Industry and Environment brings together specialists in urban and regional planning, natural resources, industry, environment, Aboriginal and social housing, and regional New South Wales.

We do this by protecting and improving prosperity, environmental sustainability, safety and security, social inclusion and cohesion, and attractiveness as a place for recreation and relaxation. We bring together the economy, environment, work and play, industry, and recreation. This helps to ensure economic growth and job security considers the environment and character of our neighbourhoods and communities. We make decisions based on advice and analysis that are transparent, efficient, and reflective of the diversity of New South Wales.

Water group

The **Water group** is responsible for ensuring sustainable, secure and healthy water resources and services for NSW.

We manage the surface and groundwater in the state, develop and implement plans for water security in NSW, and manage regional and metropolitan water supply and usage.
We provide transparent stewardship of water resources and deliver services and reforms which support sustainable and healthy environments, economies, and societies.
We are developing an integrated water strategy for NSW that addresses water security and supply for both metropolitan and regional areas. As part of this, we plan and deliver major water infrastructure projects.

We ensure equitable sharing of water resources through secure and tradeable water entitlements and allocations.

We regulate compliance and enforcement of water management in NSW. And we lead government negotiations with the Commonwealth, the Murray-Darling Basin Authority and other jurisdictions.

We drive delivery of the **Metropolitan Water Plan** and the **Water Reform Action Plan** to:

- Meet the needs of a growing population and a growing economy
- Support jobs, economic growth, the environment, and our health and wellbeing
- Support the water rebate scheme and maintain state-wide water regulation and licensing, including enabling competition in water and wastewater servicing
- Improve water efficiency and support communities to adjust to a changing climate plan metropolitan and rural infrastructure that secures water supply and increases drought resilience across the state
- Support the wellbeing of rural and regional communities who enjoy the recreational benefits our regional waterways provide
- Improve the health of waterways and their catchments to support our environmental, social, cultural and economic needs and values

We are preparing a state water strategy that will further support these aims. We aim to deliver efficient services with a focus on innovation, performance and affordability to:

- Introduce best practice for water management
- Ensure transparency in how we share, allocate and manage water
- Build a compliance and enforcement regime that ensures strong and certain regulation
- Build capability to support implementation of water reforms.
- These activities underpin resilient water resources and are critical to NSW community

DPIE - Metro Water and Utilities

A safe, secure and sustainable water supply is essential to maintaining the health, prosperity and wellbeing of Greater Sydney's growing population.

The Water and Utilities Branch leads the development of strategic water policy and planning for Greater Sydney (including the Blue Mountains and the Illawarra), on behalf of the NSW Government.

This includes:

- Implementing the current Metropolitan Water Plan
- Planning to meet the needs of a growing population
- Maintaining NSW-wide water regulation and licensing under the Water Industry Competition Act 2006
- Drought planning and management
- Supporting the water rebate scheme

More broadly, we also advise on urban water policy and reforms for NSW. We engage with the community, industry, and other government agencies to understand the factors shaping water supply and demand, as well as our stakeholders' values and preferences.

7. Internship Program Plan

I worked within the NSW Department of Planning, Industry and Environment – Water group and Mohammad Mortazavi was my supervisor for the duration of my internship. Because of ongoing restrictions due to the pandemic, I completed my internship remotely. I conducted all communications, from interview to placement using Microsoft Teams, or communicated via email with Mohammad.

Currently the group is working on developing two strategies, namely **Greater Sydney Water Strategy (GSWS**) and **Lower Hunter Water Plan (LHWP)**. These strategies include assessing the future available water resources, demand growth and potential policy changes. The ultimate goal of the strategies is to have a safe, secure and affordable water supply in Sydney and Hunter while keeping healthy environment.

I received training entirely from Mohammad within the NSW Department of Planning, Industry and Environment – Water group and my training was conducted mainly throughout the duration of my internship. The main project that I focused on throughout the duration of my internship was providing a report and commentary on: **The Importance of Interpretability in Modelling Urban Water Systems**
The start date of my internship was Wednesday, 29th July 2020 and the end date was Friday the 30th October 2020. My internship was conducted on a part-time capacity, from 20-25 hours per week.

8. Training Program

a) Overview of the department operations and activities

The Department of Planning, Industry and Environment (DPIE) - The Department of Planning, Industry and Environment brings together specialists in urban and regional planning, natural resources, industry, environment, Aboriginal and social housing, and regional New South Wales. The Water Group is responsible for ensuring sustainable, secure and healthy water resources and services for NSW.

They manage the surface and groundwater in the state, develop and implement plans for water security in NSW, and manage regional and metropolitan water supply and usage. They provide stewardship of water resources and deliver services and reforms which support sustainable and healthy communities.

They are currently developing an integrated water strategy for NSW that addresses water security and supply for both metropolitan and regional areas. As part of this, they plan and deliver major water infrastructure projects and develop, review, and refine water sharing plans and regional water strategies. They regulate compliance and enforcement of water management in NSW. They lead government negotiations with the Commonwealth, the Murray-Darling Basin Authority and other jurisdictions.

b) Detailed Project/Assignment Description

After describing the main activities performed by the department, in this subsection a detailed description of the project is given. The following description was specified by my supervisor at the beginning of the experience.

Water managers are facing a challenging task to plan for future infrastructures and policy changes with deep uncertainties such as climate, population and behavioural changes, to name a few challenges.

Traditionally, the Monte Carlo method or sensitivity analysis has been applied to capture some of these uncertainties in planning. Single or multi-objective optimization have also been used to minimize/maximize the desired outcomes.

The **objectives of this project** are focused on investigating if data science can help to better plan for future water management considering the deep uncertainties. It aims to discover some insights and rules that are easy to interpret, to assist water managers dealing with the challenging task described above.

The combination of inflows, policies and demands represents our **input** to the system.
They are randomly generated based on some domain constraints and a plausible range of values. The **input** is then supplied to the water resources system which is called WATHNET model (Kuzczera et al., 2009). WATHNET has been used to estimate yield for the Sydney water supply system since the mid 1990's [2].

It is a complex model which simulates water supply for the Sydney water. WATHNET primary's goals are founded on some objective functions such as minimize cost and maximizing yield. The latter is the **output** of interest of this analysis.

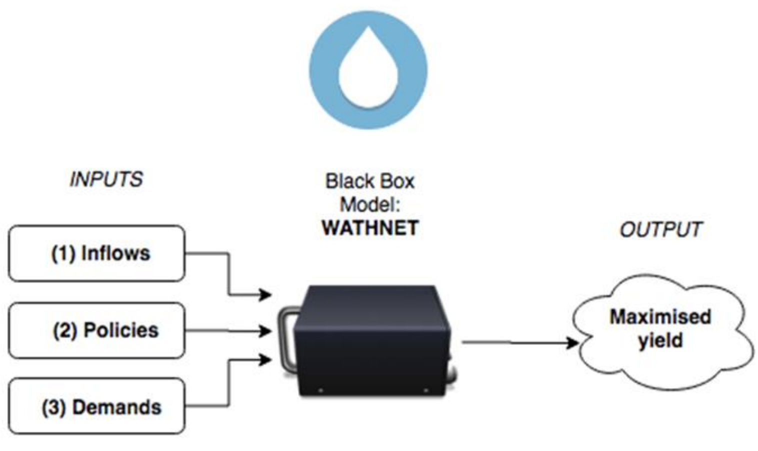


Figure 8.1. *Schematic Flowchart of the water resource system*

One of the disadvantages of WATHNET is that it can be thought as a "black box model". In fact, it is usually difficult to explain in simple terms how the predicted yield values were obtained, and which inputs or interaction of inputs played a role in determining these outputs.

The topic of interpretability in data science is much discussed amongst the researchers and data scientists. There is no formal definition for interpretability. Kim et al. [3] stated that "Interpretability is the degree to which a human can consistently predict the model's result".

If the interpretability of a model is high, everyone can understand why some judgements or forecasts have been made. Interpretability of a model can be done both **locally,** where we are interested in explaining a single prediction, or **globally** where our intent is to find out the general behaviour of the model. The latter approach is the focus of this paper.

Furthermore, interpretability of a black box model can be achieved by using **interpretable models** or **model-agnostic** interpretation tools.

Interpretable models are those that explain themselves**. Linear Regression, Logistic Regression and Decision Trees** are only some examples. For instance, interpreting "learned weights" in linear regression models is moderately easy and this is the reason why they are largely studied in academics. Similarly, Decision Trees provide tree structures which interpretability is arguably comfortable.

Model-agnostic methods [4] work by changing the input of the machine learning model and measuring changes in the prediction output. The outcome of these methods is data instances and the general idea behind these methods is to check how much the predictions of a model are changed by deleting/shuffling one feature variable. If the predictions change drastically, the influence of that feature in determining the output is strong.

9. Reflective Journal Entries

Week 1-2

The focus of the first two weeks was to understand the domain, so I was given the task to review the **2017 Metropolitan Water Plan**, which held great insight into Sydney's water management plans and systems.

The **2017 Metropolitan Water Plan** builds off previous plans to focus on four main outcomes: a secure, affordable and resilient water supply system that also ensure urban communities to remain liveable and healthy.

A key aspect of the **2017 Metropolitan Water Plan** is the **Drought Response Strategy**. Resilient water supplies that are resistant to shock are important to keeping communities liveable in the years to come. The **Drought Response Strategy** was created to be able to predict and manage extreme drought and is characterized by supply and demand management measures that are rolled out according to set water level triggers.

Some examples of these measures are implementing water restrictions and building new water supplies, and the Sydney Desalination Plant (without it, Sydney's water supply would depend entirely on rainfall). The Desalination Plant maximizes the production of drinking water when catchment levels fall below 60% and until they get back up to 70%.

Also featured in the 2017 Metropolitan Water Plan is the **WaterSmart Cities program**, which is an integrated approach to water planning across both the metropolitan region and expanding communities. The WaterSmart Cities program establishes resilient communities with an integrated approach to providing water, wastewater, and stormwater services.

Finally, reading through the 2017 Metropolitan Water Plan was very educational – there was also extensive information about Greater Sydney's water supply catchment area (spanning 16,000 square kilometres to the west and south of Sydney). It includes 11 major dams, and two major river systems, the Hawkesbury-Nepean and the Shoalhaven, as well as the Woronora River. Warragamba Dam is the largest dam, supplying around 80 per cent of Greater Sydney's water.

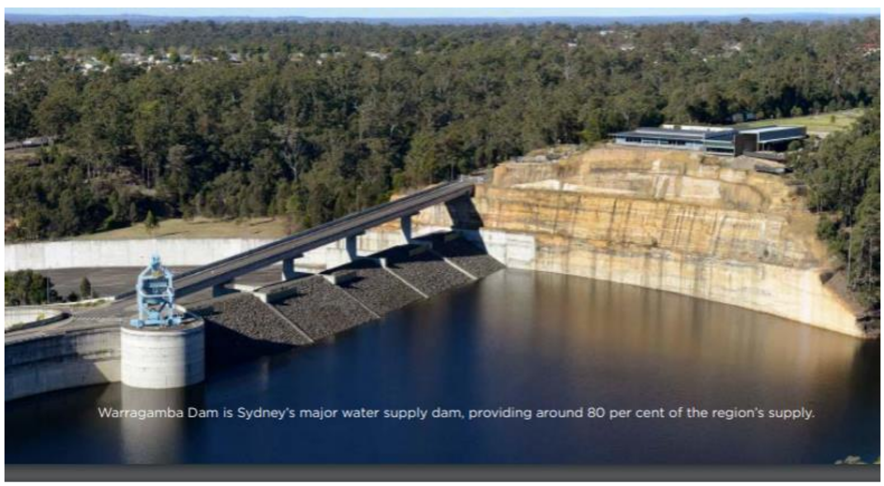


***Figure** 9.1 – Waarragamba Dam*

Week 3- 5

These weeks have been crucial to understand the general behaviour of the water resource system.

The following goals have been set up at the end of Week 2.

1. Reading previous publications written by my supervisor. Papers such as "*Multiobjective optimization of urban water resources: Moving toward more practical solutions"* [5] and "*Robust optimisation of urban water resource systems in the face of known and unknown unknowns*" [6] have been identified for the purpose.

2. Being familiar of the main functionalities of the WATHNET model software.
   My supervisor warned me about the complexity of the overall system and directed me to only to those important aspects of the total system.

I. Task 1

Although the water resource system in place now is quite complex and sophisticated, a high-level overview of the process is given. It is important to note that the above publications consider a *multiobjective optimization*, while the successive analysis only considers a *single optimization* scenario.

The following approach illustrates the steps needed to perform a *multiobjective optimization* of urban water resources.

To measure how the urban headworks system performs, a **simulation model** is usually built in response to a time series of inflows and/or demand inputs.

The lengths of the time series are very important at this stage to operate with a high level of reliability. One approached used to increase the length of the input time series is generating long stochastic input time series by sampling from probability models fitted to historical data [7]. The AR (1) (Autoregressive model) is the preferred model used for this purpose.

The **simulation model** can be mathematically defined as:

$$\min \underbrace{f(Z_n)}_{x}$$

$$\text{subject to } Z_n = \text{M [x, } Q_N, D_N]$$

$$\text{S}f_N = 0$$

where *x* is a vector of **decision variables** that are to be optimized. The latter consists of a set of level of services specified in the latest version of the Metropolitan Water plan.
They are classified as:

- *Operating rules* (which reflects the way the system is operated)
- *Infrastructure* (which indicates to a physical asset)

The function M [x, Q_N, D_N] indicates the headworks simulation model. It takes the input Q_N which represents a matrix of inflows values at different sites for an N year period, and D_N, a matrix of unrestricted demand at multiple sites for the same N year period.

The constraint Sf_N = 0 is vital for the model, as it makes sure that no unplanned demand shortfalls arise throughout the simulation. That would be the scenario where the demand of water cannot be supplied, which would be catastrophic for the community living in NSW.

The optimization model specified above can be solved by using a *single-objective genetic algorithm* (if only one output need to be optimized) or *multiobjective evolutionary algorithm* (if the scope is to optimize conflicting objectives). Once again, this paper will only consider the case of *single optimization* where the objective we want to maximize is the expected yield.

II. Task 2

Once the theoretical aspects of the water resource system have been consolidated, the following weekly task involved becoming familiar with the WATHNET 5 software.

WATHNET was invented by George Kuczera of the University of Newcastle and it has been utilized to predict yields for the Sydney water supply system for more than twenty years. It is a multi-functional water resource system model which incorporates both the simulation model and the optimization algorithms described before.

It includes a nice graphical display of the Sydney headwork systems (Figure 9.2) which includes demand areas, dams, rivers, and transfer links.

The simulation model uses **network linear programming (NLP)** to distribute, in an efficient way, water within the Sydney system. The decision vector *x* (Figure 9.3) is reachable to each script and, thus, can entirely manage the requirement of the network linear program.

NLP has been the preferred approach, rather than applying different pre-defined rules, because it enables the water to move efficiently between different links, based on a set of objective functions (for instance to minimize costs or maximize yield).

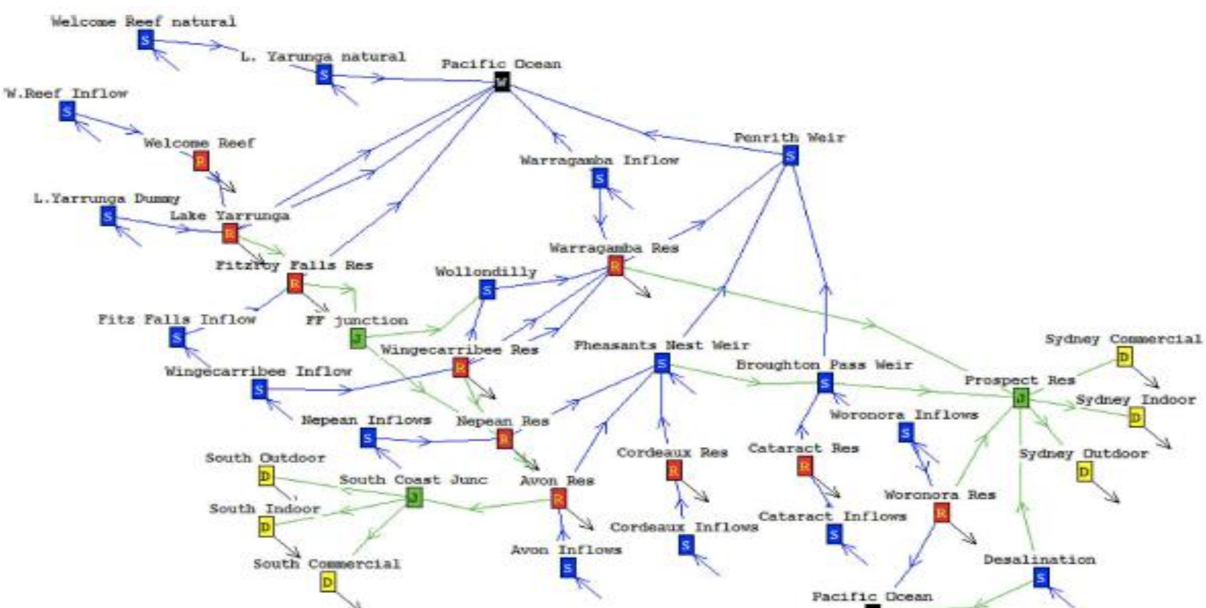


***Figure 9.2**. WATHNET schematic of Sydney water supply headworks system. The nodes labelled R represent dams, S represents stream nodes, D represents demand zones, and W represents waste and sink nodes.*

dec_number	dec_name	fit_value	code	low_range	upper_range	nIntVal
4	DS2 Option Flag (1 or 2) [D]	2.00	I	NaN	NaN	2.0
5	SHT Upgrade Flag 1, 2 or 3 [D]	1.00	I	NaN	NaN	3.0
6	DS Illawa Option Flag (1 or 2) [D]	1.00	I	NaN	NaN	2.0
7	DST Option Flag (1 or 2) [D]	2.00	I	NaN	NaN	2.0
8	GW Kangal Option Flag (1 or 2) [D]	1.00	I	NaN	NaN	2.0
9	Lnay/Wall Option Flag (1 or 2) [D]	1.00	I	NaN	NaN	2.0
11	DS1 b (125ML/d) Storage Trig [D]	0.60	R	0.5	0.99	NaN
12	DS1 c (250ML/d) Stor Trig Frac [D]	1.00	R	0.5	1.00	NaN
13	DS2 Prelm Plnnng Storage Trig [D]	0.45	R	0.3	0.95	NaN
14	DS2 Constrtn Storage Fraction [D]	0.75	R	0.5	1.00	NaN
15	DS2 StateExpn plnng Strge Frac [D]	0.89	R	0.5	1.00	NaN
16	RESTRICTION L1 Storage Trig [D] e	0.50	R	0.3	0.65	NaN
17	RESTRICTION L3 Stor Frac of L1 [D]	0.60	R	0.3	0.60	NaN
18	SHT STOR TRIG Pumping [D]	0.75	R	0.3	1.00	NaN
19	GW Kangal-StorTrig Plan/Constr [D]	0.50	R	0.3	0.90	NaN
20	GW Kangal-StorTrig Supply [D]	0.40	R	0.2	0.60	NaN
21	Lnay/Wall-StorTrig Plan/Constr [D]	0.50	R	0.3	0.90	NaN
22	Lnay/Wall-StorTrig Supply [D]	0.40	R	0.2	0.60	NaN
23	DS Illawa Constn. Storage Trig [D]	0.31	R	0.2	0.90	NaN
24	DS Illawa On Storage Trig [D]	0.80	R	0.2	0.80	NaN
25	DST Constn Storage Trig [D]	0.25	R	0.2	0.80	NaN
26	DST ON Storage Trig [D]	0.40	R	0.2	0.60	NaN
27	DS2 ON Storage Fraction of DS1c[D]	0.73	R	0.0	1.00	NaN

Figure 9.3. *List of decision variables.*

Week 6 and Week 7

The central objective of the internship was introduced at the begin of Week 6.
The research question addressed analysing how different values of decision variables (input) would impact the yield (output).

This can be thought as an easy task for any data analyst, but the real concern was the lack of interpretability of the WATHNET model. However, some **interpretable models** or **model-agnostic interpretation tools** can help to find interesting patterns in the data even when the original model is a *black box*.

Next, the aim is to discuss in details the following units: *Generation Phase*, *Dataset* and *Exploratory data analysis (EDA)*

Generation phase

By looking at **Figure 9.1** (see page 10), the input values that WATHNET model expects can be both **inflows**, **policies** and **demands** or combination of these entities.

My first operational task was to generate 1000 different combinations for each decision variables (policies) in a text file format that WATHNET (software) could read in the system.
The words “*decision variables*” and “*input parameters*” will be used in this paper interchangeably.

The main challenges were:

- The *generation phase*. How do I generate different values for each decision variable and which probability distribution can help achieving this?
- How do I integrate some domain constraints (given by my supervisor) into this *generation phase*?

Figure 9.3 (see page 15) displayed the list of decision variables. The column *code* indicates the data type (**I**: *integer*, **R**: *real*), *low_range* and *upper_range* shows the range of plausible values for *real* input parameters. Finally, *nIntVal* indicates the range of values that an *integer* parameter can take.

After few discussions with Mohammed, it has been agreed that since each parameter has equal probability of having a value within its range, the **Uniform distribution** was the suitable probability distribution for both the discrete and continuous case.

Next, a list of domain constraints needed to be incorporated during the generation phase.

They are restrictions related to the water domain environment. For example, if a parameter assumes a particular value, a list of other parameters needs to be adjusted according to some specific values. They are not going to be shared in this paper mainly because they are not easy to be understood by someone who does not have the domain knowledge.

A sample of one generated file is provided below in **Figure 9.4.** The Jupyter Notebook capable of creating these generated files is called *Decisions_scriptV2*. The *function definition* of the main method is discussed as follow:

```
def decision_script(n_files=1000,n_parameters=8):
```

The method *decision_script* takes as input two keyword parameters:

- *n_files* indicates the number of desired text files the user wants to generate. The default value is 1000.
- *n_parameters* refers to the number of parameters we want to randomly generate for that text file. The value of *fit_value* **(Figure 9.3)** is set for the other parameters. The default value is 8 parameters for each different combination.

Both input parameters will be tweaked repetitively. In particular, increasing the *n_parameters* will result in providing WATHNET a more “variable” set of combinations. It is interesting to see whether different values of *n_files* or *n_parameters* will affect our analysis.

Furthermore, the method returns:

- A folder with all the generated files. These files will be given as input into WATHNET which will generate the maximized yield (output) for each combination.
- A Pandas *dataframe* which includes our INPUT features for our future analysis.

decision_00008.txt

```
Wathnet decision file
26
576000.000000    1.000000    1.000000    2.000000    1.000000    1.000000    2.000000
1.000000    1.000000    0.644978    1.000000    0.843942    0.970903    0.890000
0.500000    0.600000    0.750000    0.500000    0.400000    0.500000    0.400000
0.310000    0.800000    0.441429    0.400000    0.429222
```

Figure 9. 4*. A sample of a generated file. The first 3 parameters could be ignored; The others are randomly generated according to their range*

The Dataset

One important aspect to mention here is the computational cost linked to the WATHNET software. In fact, WATHNET is **computationally expensive,** which means that to produce output values it needs a considerably high amount of time, processing power and memory.

The internal code is run on **parallel computing** and it takes approximately 3 minutes to complete one combination. Hence, running 1000 different combinations may take approximately 42 hours. This is a limitation, since dealing with a larger sample size is always the preferred approach by statisticians and data analysts.

To run some analysis, a Pandas *dataframe* will be constructed. The *dataframe* comprises of **26 input parameters** (8 are discrete/categorical), that have been already produced in the generation phase and **one output**, which is the outcome from the WATHNET software. The variable ***Yield*** will be used interchangeably with **yield** throughout the paper.

The **maximized yield** (output) for each of the 1000 combinations has been provided from my supervisor as 1000 different text files. One Python script takes care of reading each value from the text file and organize it as a column in the *dataframe* (See Section 10 for details).

Exploratory Data Analysis

Before digging into modelling in order to analyse the relationship between **input parameters** and **yield**, a simple **exploratory data analysis (EDA**) will be performed.

Many of the EDA techniques are graphical in nature and the intention is to open-mindedly investigate the dataset to gain some new, often unsuspected, insights from the data. Especially for domain experts, these graphs can be very helpful to describe certain situations with the help of data visualization tools.

In this section, we are mainly focused on checking the relationship between the **input parameters** and the output **yield**, however being aware of the shape of the output distribution can be important for further study.

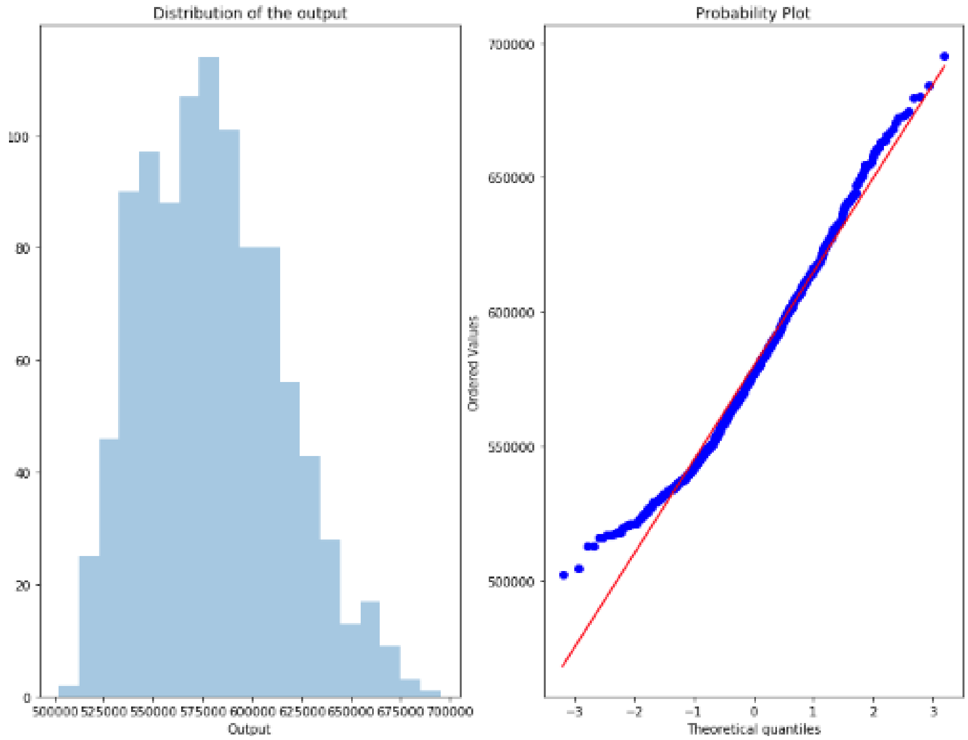


Figure 9.4. *On the left the histogram of yield. On the right the qqplot of yield*

Based on the histogram, the shape of the output **yield** looks slightly right skewed, which means that the mean yield is higher than median yield. The qqplot suggests that the obtained sample distribution can resemble a normal distribution.

Next, a **correlation matrix** was created. It is important to note that only part of the matrix is valuable for our purposes. In fact, the interest here is to determine the strength of the relationship between the **input parameters** and the **yield**.

The method used to calculate the correlation matrix is **Spearman**. **Spearman's correlation** defines the strength and direction of the monotonic relationship between two variables rather than the strength and direction of the linear relationship between two variables, which is what Pearson's correlation determines [8].

This quantity has been preferred because the linearity restriction can be too repressive for our scenario. The following **heatmap** has been obtained by using one row of the **correlation matrix** (covariates vs output).

It can be a valuable source for determining the strength of the relationship between **input parameters** and the **yield.**

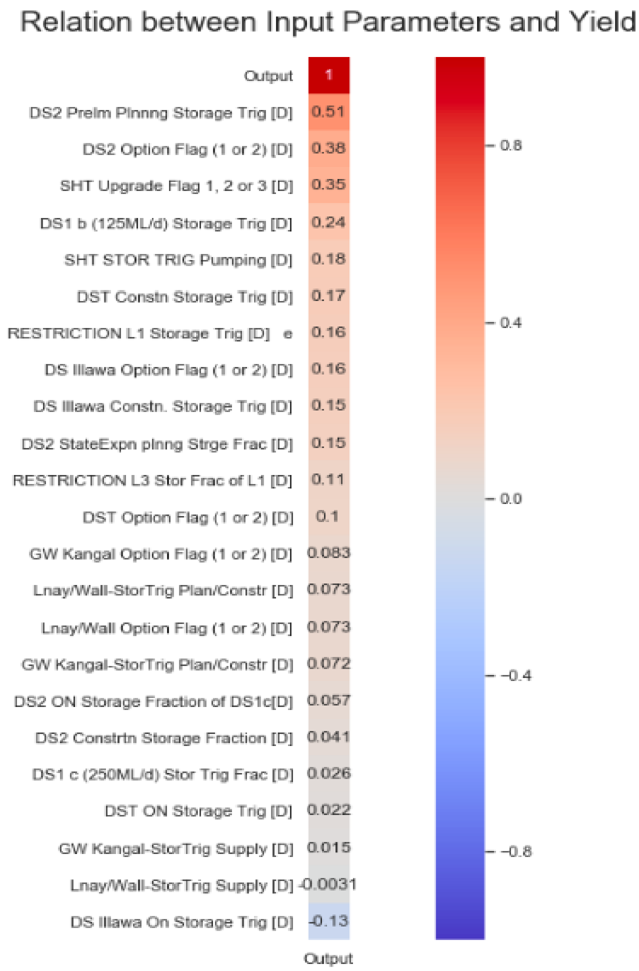


Figure 9.5. *The heatmap shows the relationship between decision variables and yield*

Based on **Figure 9.5,** the following considerations can be highlighted:

- Absolute values of correlation higher than 0.1(this threshold can change) can be interpreted as significant predictors of **yield** and need to be considers later. Since correlation does not mean causation, some high collinearity might be caused by external factor or multicollinearity between covariates. It turns out that ***DS2 Option flag*** is correlated to the **yield** only because of the multicollinearity with ***DS2 Prelm Plnnng Storage* Trig**

- There is a strong positive relation (0.51) between ***DS2 Prelm Plnnng Storage* Trig** and **Yield**. Hence, if this decision variable increases, the Yield increases too.

- There is a slightly negative relationship (-0.13) between **DS Illawa On Storage Trig** and **Yield**. Hence, if this decision variable increases, the Yield slightly decreases.

The last data visualization (others can be seen in *Data Analysis* notebook) is a **comparative boxplot**. The interest is to look at the relationship between ***Yield*** and one of the discrete/categorical variables. The domain expert can change the latter according its preference.

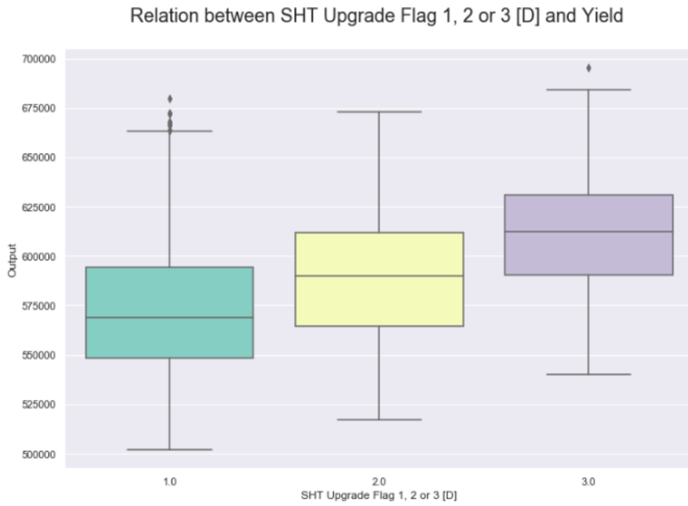


Figure 9.6. *A comparative boxplot between SHT Upgrade (the selected categorical variable) and Yield*

According to the graph above, the median ***Yield*** when *SHT Upgrade Flag* is '3' is higher than the other situations.

This kind of graph might help us to better understand which particular category of a variable have a "better"/"worse" impact on the output of interest. If we are interested on minimizing the output ("Cost"), we would look at the smallest median with particular interest, while if we need to maximize the output ("Yield") the opposite comments could be done.

Week 8-12

The final weeks were used to apply more advanced methods and approaches in data science on our dataset. The goal is to keep looking for interesting patterns between **input parameters** and **yield** and consequently trying to interpret the results obtained by the WATHNET model.

Interpretability of a black box model can be achieved by using **interpretable models** or **model-agnostic interpretation tools**.

Interpretable Models

One possible answer for interpreting black box models is to approximate them with an **interpretable model**. **Linear Regression**, **Logistic Regression** and **Decision Trees** are only some well-known examples.

It is very important to highlight that the objective is not building a good model for prediction but look at the **feature importance** to try to identify which predictors have an impact on the dependent variable.

Feature importance specified by machine learning models might benefit domain experts in various ways, for example:

- By having an improved understanding of the model's logic

- By deleting those subsets of variables that are not that significant. This makes the model less complex and can improve performance

Two interpretable models will be analysed: **Decision Tree** and **Multiple linear regression**.

Decision Tree and Random Forest

A **Decision Tree** is a method that can be used to divide up a large collection of data into successively smaller sets of records by applying a sequence of simple decision rules.

Although they are mostly used for classification purposes, **Decision Trees** are also capable of performing regression tasks. The algorithm splits the dataset in a way that minimizes the MSE (mean squared error) if a regression task is required.

The idea here is to fit a tree into the data with no restriction. This is usually a bad habit for "predictive" purposes, as the model *overfits* the data. However, as we want to build some intuitions rather than building a predictive model this approach is accepted.

Decision Trees also play an important role in **Random forest**, which is another well-known supervised ML method. Random forest in an ensemble learning technique which aggregates the prediction of several Decision Trees over different random samples (bagging).

In the next page, the top ten most influential parameters for **Decision Tree** and **Random forest** are shown (Figure 9.7)

Based on the below bar charts, the two most influential **input parameters** for **yield** are:

- *DS2 Prelm Plnnng Storage Trig [D]*

- *SHT Upgrade Flag 1, 2 or 3 [D]*

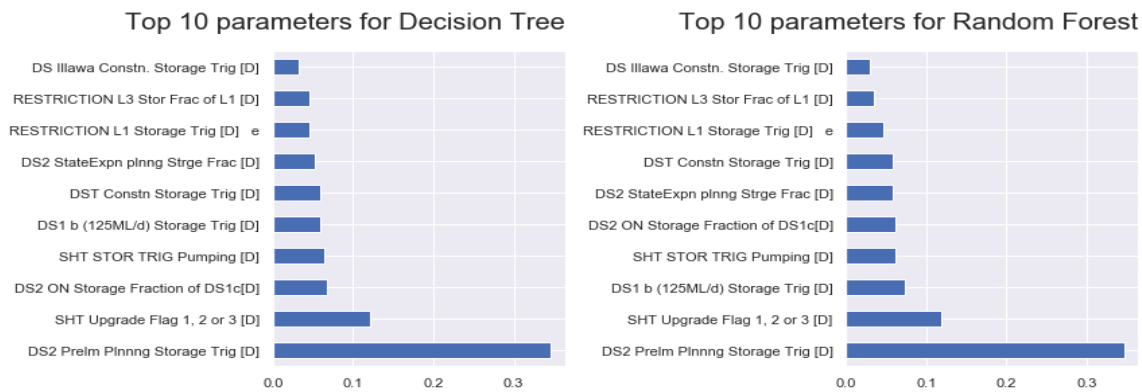


Figure 9.7. *Feature importance for Decision Tree and Random Forest*

The chart provides additional insights. The variability of ***Yield*** is influenced by these top 10 features. This imply that, the remaining parameters are useless, and do not provide any information to the ***Yield***. Hence, the WATHNET model can be run again with almost a third of the initial **input parameters** without losing information. This is a key result, as we can reduce computational resources.

Next, a **Decision Tree** visualization is provided (**APPENDIX A**).
One of the advantages of this ML method is the easier interpretation of the results given by a sequence of *if-then-else* rules that are very intuitive for humans to understand.
For example, we might comment on one specific branch of the tree.

The following chart displays the *if-then-else* rule linked to one of the highest *expected value* of ***Yield.*** It can be found on the right-hand side of the tree **(APPENDIX A**).

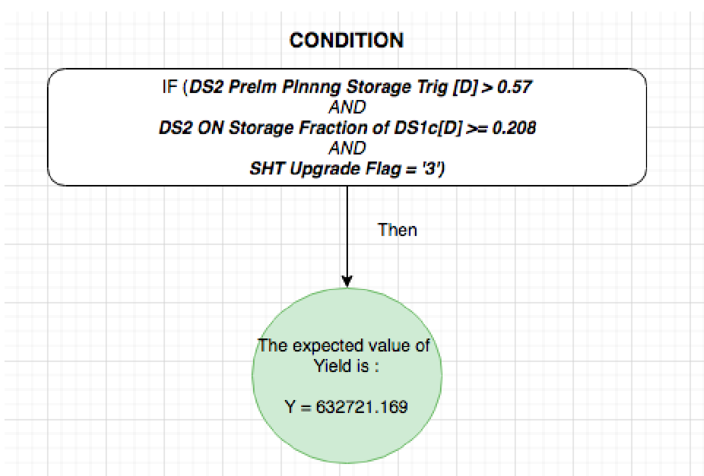


Figure 9.8. *if-then-else rule linked to one of the highest expected value of* ***Yield.***

This extracted decision rule can be difficult to be interpreted by someone who does not have the sector expertise. However, decision makers can use those to better plan future actions.

Multiple linear regression

The **multiple linear regression** predicts the **yield** as a weighted sum of the **input parameters**. The main benefit of **linear regression models** is the fact that is linear: the estimated linear equations have an easy interpretation on a modular level (i.e. the weights). The estimated parameters are easy to comment and might reveal important understandings about the overall data.

Furthermore, estimated weights are often displayed with **confidence intervals**. A **confidence interval** shows a range of values where we believe the “true” population weight can take.

The summary of the most promising linear regression is displayed below. Also, the diagnostic of this model shows that the assumptions of the linear regression are met (not shown).

```
## Call:
## lm(formula = Yield ~ DS2.Prelm.Plnnng.Storage + SHT.Upgrade.Flag +
##     DS1.b125ML + DS2.StateExpn.plnng + RESTRICTION.L1.Storage.Trig +
##     SHT.STOR.TRIG + DS2.ON.storage.Fraction + DST.Constn.Storage.Trig +
##     RESTRICTION.L3.Stor.Frac + DS.Illawa.Constn, data = water)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -57982 -13893  -1097  13720  60428
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   267529       6098   43.87   <2e-16 ***
## DS2.Prelm.Plnnng.Storage      101465       1987   51.06   <2e-16 ***
## SHT.Upgrade.Flag2              12457       1162   10.72   <2e-16 ***
## SHT.Upgrade.Flag3              37440       1134   33.01   <2e-16 ***
## DS1.b125ML                     66368       3369   19.70   <2e-16 ***
## DS2.StateExpn.plnng            50338       3497   14.40   <2e-16 ***
## RESTRICTION.L1.Storage.Trig    96879       6009   16.12   <2e-16 ***
## SHT.STOR.TRIG                  55867       2828   19.75   <2e-16 ***
## DS2.ON.storage.Fraction        20640       1954   10.56   <2e-16 ***
## DST.Constn.Storage.Trig        24306       1966   12.36   <2e-16 ***
## RESTRICTION.L3.Stor.Frac       57251       4172   13.72   <2e-16 ***
## DS.Illawa.Constn               46797       2967   15.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19800 on 2988 degrees of freedom
## Multiple R-squared:  0.6464, Adjusted R-squared:  0.6451
## F-statistic: 496.6 on 11 and 2988 DF,  p-value: < 2.2e-16
```

Figure 9.9. *A summary of the best multiple regression model*

Each parameter is **highly significant** and around 64% of the variability of **Yield** is explained by these selected decision variables.

The interpretation of the two most important features (based on the higher absolute value of their t-statistic) is given below:

- The effect of rising *DS2 Prelm Plnnng Storage Trig [D]* of 0.1 is an increase in expected **Yield** of 10146.5 ML.
- The effect of being in *SHT_Upgrade_Flag=2*, compared with *SHT_Upgrade_Flag=1*, is an increase in expected **Yield** of 12457 ML
- The effect of being in *SHT_Upgrade_Flag=3*, compared with *SHT_Upgrade_Flag=1*, is an increase in expected **Yield** of 37440 ML

The following **weight plot** displays the **confidence intervals** of the selected input parameters.

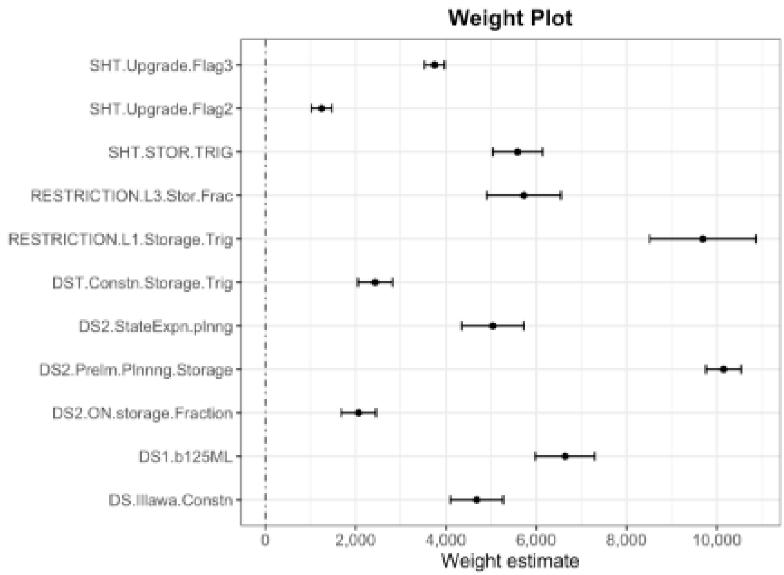


Figure 9.10*. Weight plot of the weights*

The weights are displayed as estimate points and the 95% confidence intervals as lines. For example, the effect of rising *DS2 Prelm Plnnng Storage Trig [D]* of 0.1 is an increase in expected **Yield** of 10146.5 ML (point estimate) with a 95% confidence range of [9756.993,10535.996].

Model Agnostic Method

One alternative approach for interpreting black box models is to use **model-agnostic interpretation tools**. The greatest benefit of **model-agnostic interpretation methods** comparing to **model-specific ones** is their elasticity. In fact, data scientists are allowed to use any machine learning algorithms they like.

The Model-Agnostic methods used are **Feature Interaction**, **Permutation Feature Importance** and **Partial dependence plot**. The last two approaches will be discussed in this section as it provided good insights, while **Feature Interaction** will be reviewed in Section 10.

Permutation Feature Importance

We measure the **feature importance** by computing the increase in the model's prediction error after permuting or deleting the feature. An **input parameter** is then "important" if shuffling or deleting its values upsurges the model error. This signifies that the model significantly relies on that feature for the prediction. On the other hand, a feature is "unimportant" if the model error does not change drastically.

The ideal setting would be using the WATHNET model to generate 26 different model's predictions (by excluding/shuffling one variable at the time) and then look at how much MSE would change comparing to the "observed" initial output. However, as suggested by my supervisor, this approach would be too computation expensive.

The following solution can be still implemented though. The *eli5* is a *scikit-le*arn library, used for computing permutation importance. It offers a method that compute **feature importance** for any black-box estimator by calculating how the score reductions when a feature is not available in the model.

We assume that a **Random Forest** model can be a good approximation of our data.

Next, the top ten most influential parameters for **permutation feature importance** using a random forest model is shown (Figure 9.9).

Once again, *DS2 Prelm Plnnng Storage Trig [D] and SHT Upgrade Flag 1, 2 or 3 [D]* are the two most influential **input parameters** for **yield.**

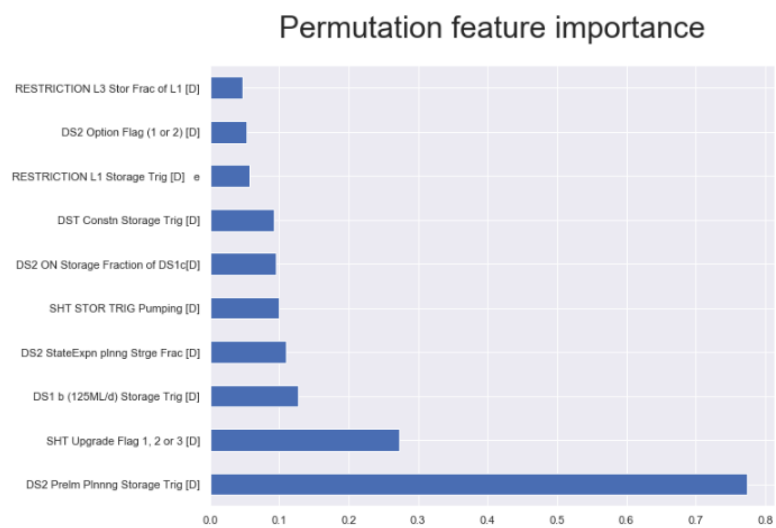


Figure 9.11. *Permutation feature importance bar chart*

Partial dependence plot and Individual Conditional Expectation

Partial dependence plot (**PDP**) displays the average marginal effect that one or two input parameters have on the predicted yield of a specified machine learning model.

To approximate the WATHNET model, we used a Random Forest which can reach a 96% accuracy. Next, only the PDPs plots on the best continuous predictors will be displayed, but this can be extended to other variables (Figure 9.10).

All of these plots shows a quite positive marginal effect with **Yield**.

By looking at *DS2 Prelm Plnnng Storage Trig [D]*, it looks like that between 0.3 and 0.6 the **Yield** increment is quite linear, while after 0.6 the curve seems almost vertical which means that for that range the increasing rate of Yield is almost doubled.

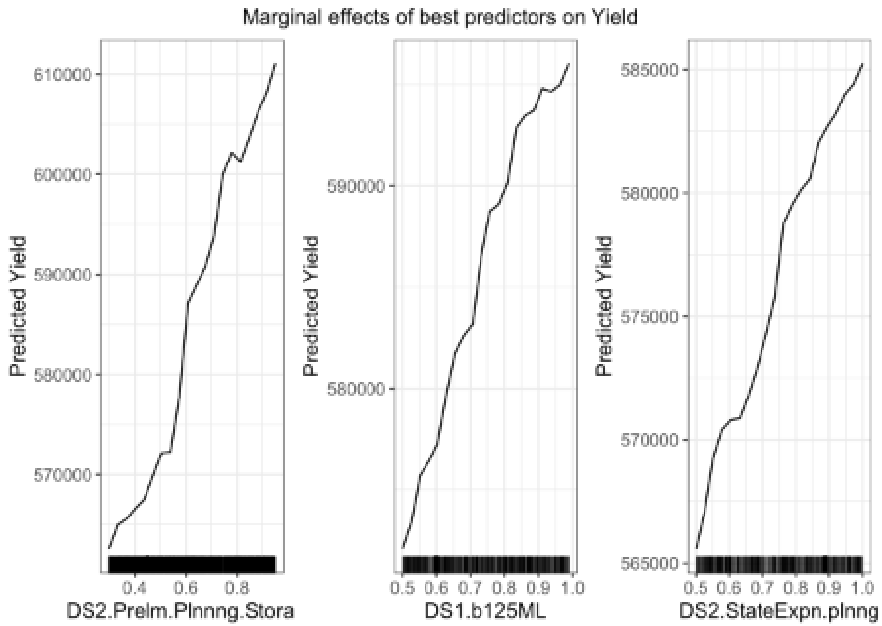


Figure 9.12. *Individual PDPs of the best 3 continuous predictors of yield*

Finally, we can also visualize the partial dependence of two features at once. For example, we might be interested on the interaction between *DS2 Prelm Plnnng Storage Trig [D]* and *DS2**.**StateExpn.Plnng*.

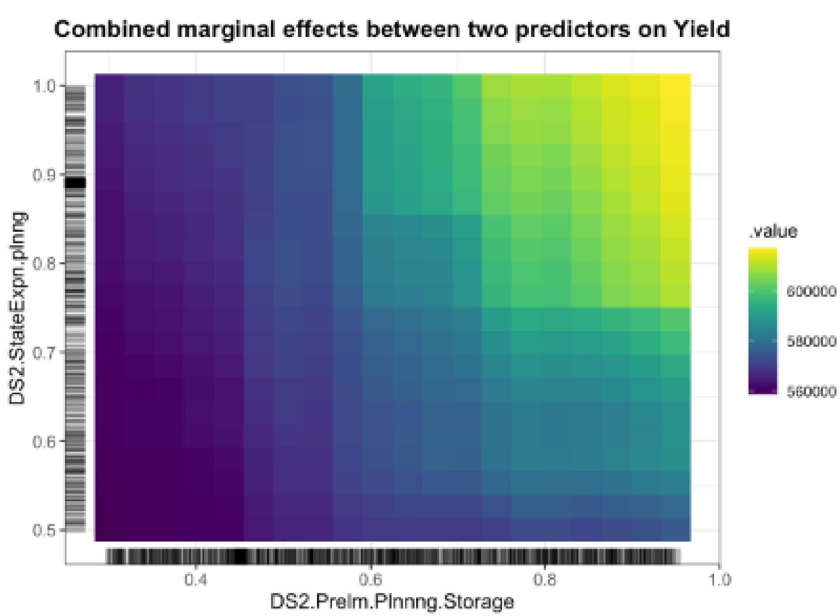


Figure 9.13. *2-D Partial dependence of two features at once*

Based on Figure 9.13, the PDP of **yield** is plot by considering the interaction between *DS2 Prelm Plnnng Storage Trig [D]* and *DS2***.***StateExpn.Plnng*. Clearly, the yellow color shows combination of features where the **yield** is very high, while the dark blue color reflects the opposite scenario.

Individual Conditional Expectation (ICE) plots are similar to PDPs, but present one line per observation. Hence, PDPs display the average effect of a feature on the output of interest, while ICE focus on specific instances.

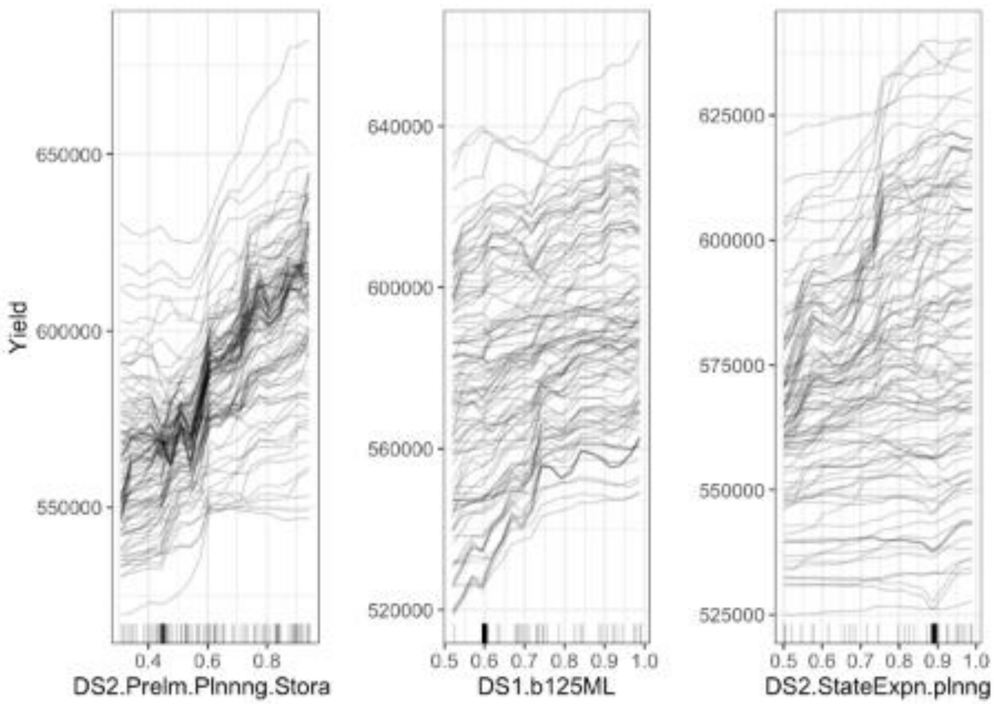


Figure 9.14. *Individual conditional expectation (ICE) of the best 3 continuous predictors of yield*

The Individual Conditional Expectation (ICE) plots above illustrate the best 3 continuous predictors of **yield** based on a random sample of size 100.

Based on the left chart, we can confirm that there is a steep increase of **yield** when *DS2 Prelm Plnnng Storage Trig [D]* is higher than around 0.6.

The ICE plot for *DS2***.***StateExpn.Plnng* (right) shows that this decision variable has no effect on Yield for value lower than 550000 ML. After this threshold, *DS2***.***StateExpn.Plnng* seems to have a positive relationship with **yield**.

10. Work Samples

In this section two samples of my work during the internship will be discussed. The first one has been developed in the first half of the internship, while the second one on the final weeks.

First Work Sample

The first work sample inspects the Python function which aim is to read some values from text files into a *dataframe*. Being familiar with *file input and output* and *string manipulation* in Python helped me to achieve this task in a fairly small amount of time.

Figure 10.1 illustrates an example of an output text file obtained by the WATHNET model.

```
Root finder search summary

Target variable: "STORAGE Security Constraint [Ct]"
   target value:    0.000000

Fitted decision variable: "DEMAND Annual [D]"
            search range:    400000.0       800000.0
                    root:    614219.6
                residual:  -0.2826434E-03

Search converged in   5 iterations

Convergence settings: Max iterations =  20
```

Figure 10.1 *A sample of a generated output file from WATHNET*

The numeric value next to the word "root:" indicates the **maximized yield** obtained by that combination of input parameters. The goal of this task was to read all the "root" values from the 1000 generated output files into a *dataframe* column.

Two python methods in the *Data Analysis* Notebook accomplished that task. The first one (*load_output_file*) reads the output.txt and returns the root value as float number, while the second function (*load_all_files*) simply loops over the directory where all the output.txt files are stored and returns a list of "root" values.

Of particular interest is the *load_output_file* function.

```
def load_output_file (filename):
    'This function reads the output.txt file and return the root value as float number'
    outputfile=open(filename,"r")
    lst_strings=outputfile.readlines()
    string=','.join(lst_strings)
    pos=string.find('root:')
    return float(string[pos+9:pos+17])
```

As the comment suggests this function reads the output.txt file (filename) and return the root value as float number.

Based on some experiments, the more successful approach for returning the correct value was by using to the "find" string method. This method guaranteed to return the **maximized yield** back to the user in a correct manner for each of the 1000 text files.

Second Work Sample

The second work sample examines another **model-agnostic interpretation tool**, which is named **Feature interaction**.

When the **input parameters** interrelate with each other in a prediction model, this effect needs to be taken into account. The interaction between variables adds extra information in the variability of the output variable **yield**.

The **interaction** amongst features is the alteration in the forecast that happens by varying the features after the individual feature effects has been considered.

One method to assess the **interaction strength** between features is to quantity the amount of the prediction's variation that relies on the **interaction** of the features. This measurement is called **H-statistic**, introduced by Friedman and Popescu (2008) [9].

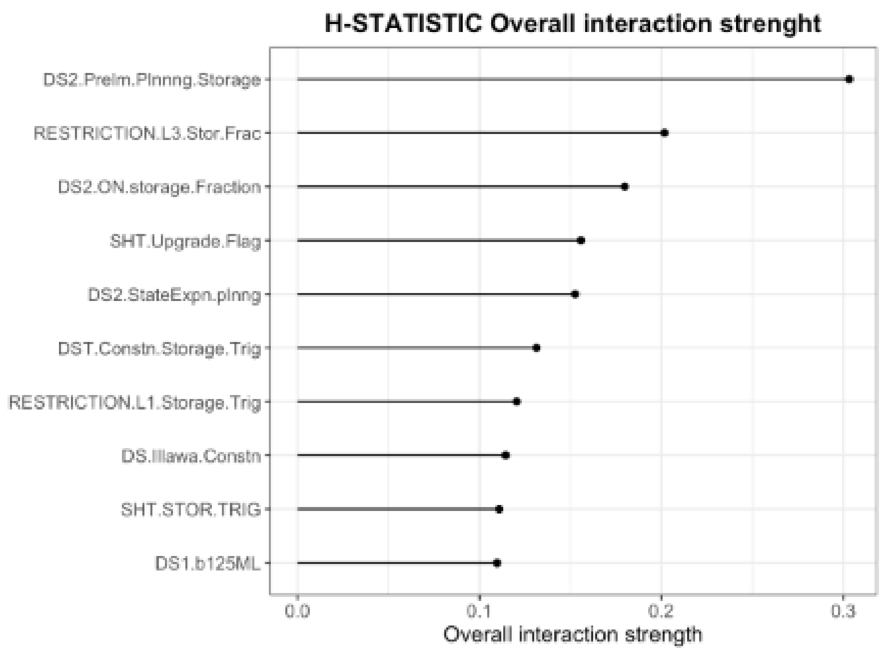


Figure 10.2 *H-Statistic for each important feature with all other features*

The above figure shows the **interaction strength** (**H-statistic**) for some of the most promising input parameters with all other features for a random forest predicting **yield**. It looks like that, *DS2 Prelm Plnnng Storage Trig [D]* has the highest relative interaction effect with all other features.

Once we observed at the **feature interactions** of each feature with all other features, it might be interesting to choose one of the input parameter with high interaction and deeply scrutinize all the 2-way interactions between the selected input parameter and the other features.

For example, we might be interested to look at the 2-way interaction strengths (**H-statistic**) between *DS2 Prelm Plnnng Storage Trig [D]* and each other feature.

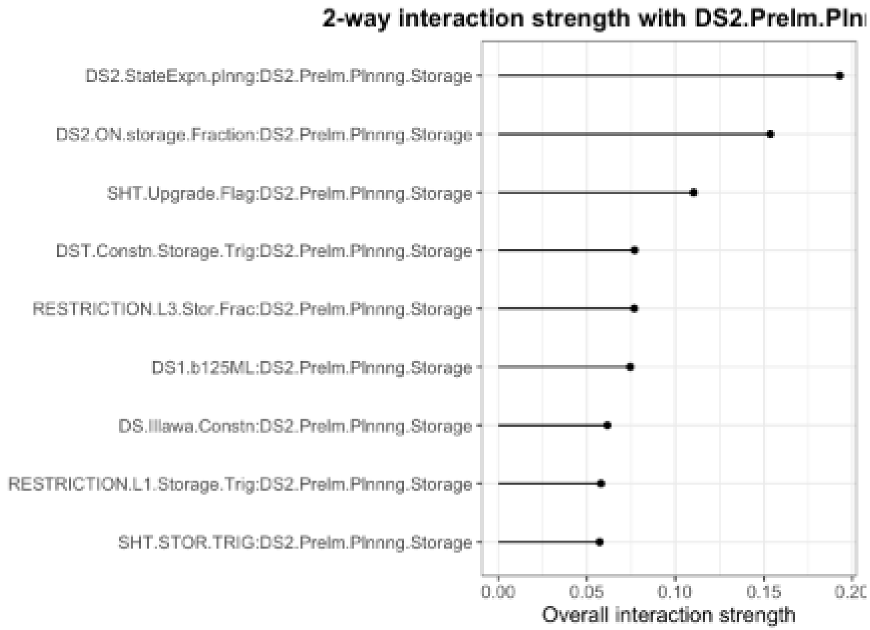


Figure 10.3 *The 2-way interaction strengths between DS2 Prelm Plnnng Storage Trig and each other feature*

Based on the Figure 10.3, an **H-statistic** of around 0.2 between *DS2 Prelm Plnnng Storage Trig [D]* and *DS2***.***StateExpn.Plnng* signify that there is a significant effect on the prediction of **Yield** given by this interaction.

Nevertheless, the **H-statistic** indicates the strength of interactions, but it does say how the interactions look like. The 2-D partial dependence plot at figure 9.13 can help visualize this interaction better.

11. Critical Analysis

In this section the focus is to relate the theoretical aspects learned during the master of Data Science to the practical understanding learned during the internship. In particular, two situations will be discussed and some suggestions are given on how to re-structure the course content in order to provide some additional knowledge for future students.

Machine Learning Algorithms

Three units covered machine learning algorithms: **Data mining**, **Data science** and **Machine learning**. They covered fundamental techniques and tools of data science, such as *data visualization*, *supervised learning algorithms* (both conventional and not), and *unsupervised methods* such as clustering analysis.

After the completion of these units, skills such as recognizing which data science techniques are most suitable for a real-world dataset or being able to interpret the outcomes of the analysis were only some of the acquired abilities.

The theoretical aspects learned from university helped me to solve real-world tasks with the right mindset.

Data visualization tools such as **scatterplot** or **heatmaps** were used, and they were capable of visually summarize information in a way that could identify patterns and trends.

Supervised learning algorithms were applied. For instance, **Decision Trees** provided a very comprehensive structure that can be easily interpreted by anyone who do not have any statistical background.

Finally, being proficient with R and Python helped me to develop scripts in a fast way.

Useful Suggestions for Future Students

The critical analysis section ends with two suggestions that in my opinion can help future Data Science master students.

The first suggestion is to include a mandatory unit in **Python programming**. The actual course structure does not include a unit like that.

Fundamental of Computer Science (COMP6010) provides a study of algorithms, data structures and programming techniques but the used software is *Java*. Although *Java* is the second most popular language in the world (TIOBE Index for September 2020), data scientists' favourite language is Python.

Changing the language can benefit students dramatically as most of the succeeding core subjects (a.e Machine Learning) are taught in Python.

The second suggestion is to include a mandatory unit in **Data Analysis with Python and R**.
I struggled a lot with manipulating *dataframe* in Python especially at the begin of the master and I had to enrol to additional external units to fill this gap.

The inclusion of this subject would enable students to learn features such as:

- Acquire basic and advanced features in *NumPy*
- Be familiar with data analysis tools such as *Pandas* library for Python, or *tidyverse* for R

12. SWOT Analysis

A SWOT analysis breaks down the strengths, weaknesses, opportunities, and threats of an organization, please read below for the SWOT analysis on the Department of Planning, Industry and Environment – Water group (DPIE – Water group).

The strengths of the Department of Planning, Industry and Environment – Water group are many. Being a government department, there is an allocated budget in place, so strategies and actions can be planned well in advance due to having that knowledge. Not only that, there is infrastructure in place and extensive resources (data being one of them, and most useful in the context of this internship), for employees to work with. From an employees' perspective, there is strong central planning and structure and policy in place, so employees know what is expected of them and can perform at a higher capacity. All of these (allocated budget and business planning, extensive resources, and stable infrastructure, established human resources policies and procedures) are some internal strengths of the DPIE – Water group.

On the other side, there are weaknesses of the Department of Planning, Industry and Environment – Water group. Firstly, centralized decision making can be a strength in terms of having a common message, however processes to receive approval can be time consuming and take many steps (this is commonly referred to as 'red tape'). In this case, infrastructure and policies can get in the way of completing tasks as quickly as possible.

Another weakness of the DPIE – Water group is again its size, when you have a large number of employees working, and especially in this case when any are geographically dispersed and working remotely, it can be difficult to monitor behaviours and quality of work can suffer. Also, it can have a negative impact on workplace culture.

There are opportunities of the Department of Planning, Industry and Environment – Water group, first and foremost the increasing threat of climate change means there will be more and more opportunity to obtain funding and budget allocations to do the very important work that they do. There will be more and more external support in the future as water supply is an important topic that will receive a lot of attention in the coming years. This is also an opportunity to make a significant difference in improving the lives of those living in NSW.

The biggest threat to the Department of Planning, Industry and Environment – Water group is of course, climate change and drought, and the ongoing pandemic. To address climate change and drought – firstly Australia has just experienced one of the worst bushfire disasters in recent years.

There is a lot of debate on the topic, however whether climate change is the cause of the bushfire seasons getting worse or not, it is a direct threat to the water supply and management that is the responsibility of the DPIE – Water group. There are large regions over Australia that have recently experienced drought or are experiencing drought and managing the water supply and ensuring best practice water management is extremely difficult in those instances. The Murray-Darling Basin is an example of the difficulty in managing water supply in regional areas.

In regard to the pandemic, there is no denying that the entire world has changed this year. Increasing pressure on the economy is a direct threat to funding and budget allocation for the DPIE – Water group. Also, the pressure of recent events has changed the way that people work and can have a significant impact on workplace culture and employee wellbeing. In turn, poor mental health and wellbeing can have an impact on leave, causing employees to take more annual or personal leave, and can also have significant impact of productivity as well.

To summarize, some of the strengths of the DPIE – Water group are its resources and infrastructure. On the other hand, some weaknesses are that bureaucracy and centralized decision making can slow down work performance and cause dissatisfaction in employees.

Also, the size of the department can make it difficult to monitor and manage performance. Some opportunities are that the support the DPIE – Water group provides is crucial and needed and guaranteed to secure support in the years to come and will likely grow as their services are needed. However recent events such as bushfires, climate change and drought are a real threat to the department.

13. Conclusion

This paper has shown the **importance the interpretability of a black box model in an urban water system**. Planning future infrastructures and policy changes under deep uncertainties of climate and population changes are a very challenging mission for water managers.

The advent of WATHNET model in the mid 90's was used to estimate yield for the Sydney water supply system, and it is still in use nowadays. One possible drawback of this model can be the lack of interpretability of its outputs.

The paper aimed to demonstrate how data science tools can help to better plan for future water management. The interpretability of a black box model has been accomplished by using **interpretable models** or **model-agnostic interpretation tools**.

Interpretable models revealed how **Decision Tree** are able to provide a very straightforward graphical display which enable non-data scientists to extract useful decision rules, while **multiple linear regression** offered a very easy interpretation of the parameter estimates. Overall, *DS2 Prelm Plnnng Storage Trig [D] SHT* and *Upgrade Flag 1, 2 or 3 [D]* were considered the most influential **input parameters** for **yield.**

Model-agnostic interpretation tools have shown how *DS2 Prelm Plnnng Storage Trig [D]* has the highest relative **interaction effect** with all other features. Also, the PDPs have shown a quite positive marginal effect between the most influential features and **Yield**.

14. Recommendation

Some recommendations on future works are briefly discussed.

Unsupervised learning can be valuable to the analysis. For instance, it can be interesting to apply *cluster analysis* in the data to detect some pattern or obvious groups. We might discover that, for example, the combination of input parameters and output always fall into 3 to 4 groups with similar characteristics.

Local interpretability for a single Prediction. The paper only discussed **global interpretability** as requested by my supervisor. However, interesting the reasons behind a particular prediction of **yield** can be noteworthy. For example, examine the impact of the decision variables for the maximum or minimum value predicted by WATHNET.

Finally, a possible future work can be the **automatization** of the overall process. At this stage, the procedure needed to obtain outputs from models is not smooth yet.
In fact, input parameters need to be sent to WATHNET and only when the last one has finished a *dataframe* can be constructed and the analysis can begin. One possible solution is building Python pipelines. This will ensure to collect all Python functions into a single pipeline capable of producing results within one single run.

References

[1] Administrative Arrangements (Administrative Changes—Public Service Agencies) Order 2019 [NSW] (159)" *(PDF). Government Gazette of the State of New South Wales. 2 April 2019. p. 7-8. Retrieved 4 April 2019.*

[2] Ian Varley and Pavel Kozarovski, June 2016, "*WATHNET Water Supply System Model Independent*"

[3] Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo, 2016. *"Examples are not enough, learn to criticize! Criticism for interpretability."* Advances in Neural Information Processing Systems

[4] Molnar, Christoph, 2019. *"Interpretable machine learning. A Guide for Making Black Box Models Explainable".*

[5] Mohammad Mortazavi et al., 2011. "*Multiobjective optimization of urban water resources: Moving toward more practical solutions".*

[6] Mohammad Mortazavi et al., 2013*, " Robust optimisation of urban water resource systems in the face of known and unknown unknowns".*

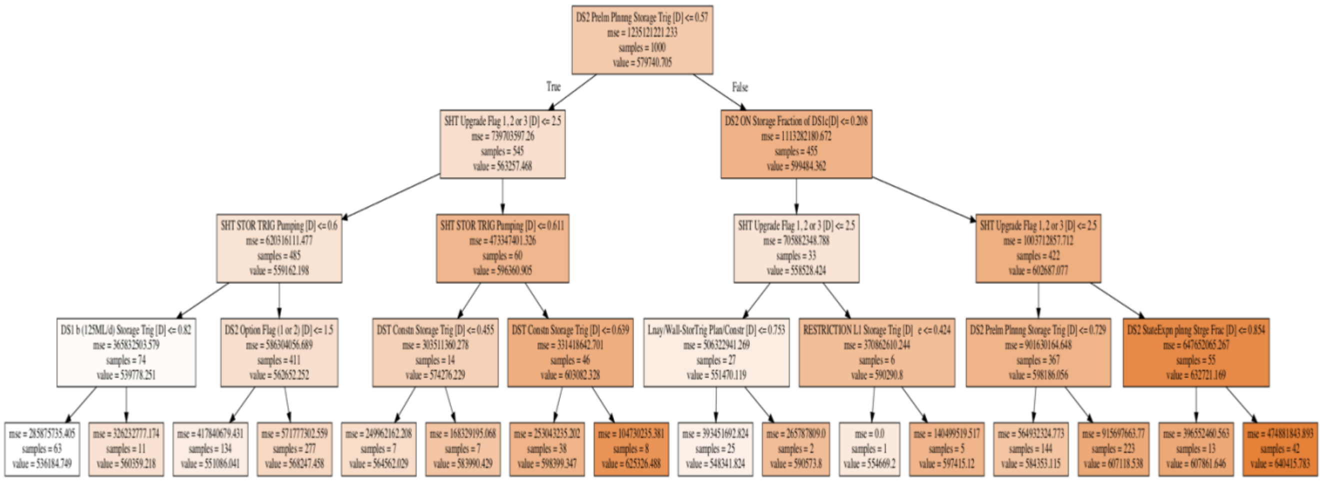
[7] Salas, J. D. ,1992, *"Analysis and modeling of hydrologic time series, in Handbook of Hydrology, edited by D. Maidment".*

[8] Spearman's Rank-Order Correlation, https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

[9] Friedman, Jerome H, and Bogdan E Popescu, 2008, "Predictive learning via rule ensembles". The Annals of Applied Statistics.

Appendix

APPENDIX A.



***Appendix A. Decision Tree** visualization*