Dalhousie University
Department of Industrial Engineering

IENG 8900: MEng Project Report

STORM DAMAGE PREDICTION IN NOVA SCOTIA USING MACHINE LEARNING TECHNIQUES

Prepared By:
Mufaddal Lokhandwala
(B00914329)

TABLE OF CONTENT

ABSTRACT

In the current era, households, businesses, and industries have become extremely dependent on the reliable supply of electric power. Whereas developed nations have well-functioning electricity grids under normal circumstances, these grids experience significant damages during extreme weather conditions like hurricanes and storms, hindering their ability to serve customers. Hence, it is essential for power companies to accurately predict such storm damages to prepare remedial plans to restore power to their customers as soon as possible. Power loss due to storms is a recurring problem in Nova Scotia that affects hundreds of thousands of households every year. Nova Scotia Power Inc. (NSPI), the largest electricity provider in the province, currently uses a simple regression-based tool to predict storm damage based on weather forecasts, which suffers from low accuracy and requires regular manual adjustments. This project aims to develop an advanced tool for storm damage prediction in Nova Scotia based on modern machine learning techniques like Artificial Neural Networks (ANN) and Random Forests (RF). The prediction models are trained using historical data of weather and the damage caused during storms and hurricanes in the past and are found capable of predicting outages with a mean absolute error of 0.05 (i.e., the accuracy of 95%), outperforming by a wide margin the tool currently used by NSPI.

ACKNOWLEDGEMENTS

Dalhousie University is located in Mi'kma'ki, the ancestral and unceded territory of the Mi'kmaq. We are all treaty people. I also acknowledge the histories, contributions, and legacies of the African Nova Scotian people and communities who have been here for over 400 years.

I would like to thank my academic advisor, Dr. Ahmed Saif for providing continuous support and being available for help throughout this project. His knowledge and guidance helped significantly in solving the issues in the model and improving its accuracy. I am also grateful to Dr. Saif for giving me career and academic guidance whenever required.

I would also like to thank Nova Scotia Power Inc. for providing the historical data and sharing their current methodologies. I especially thank Mr. Mike Macmillan for taking out the time to meet us and provide the necessary information. It has been a great experience working with NSPI and I look forward to working with them in the future.

INTRODUCTION

Nova Scotia is one of the three maritime provinces of Canada. It is located on the north Atlantic Coast of North America [2]. It is geographically the second smallest province of Canada with an area of around 53,000 Km^2 and a population of approximately 970 thousand people [3]. This province is mostly surrounded by water and is famous for its high tides. It is prone to severe and unpredictable weather such as storms, hurricanes, blizzards, etc. that leads to power outages [1].

The organization responsible for 95 percent of the power generation, transmission, and distribution in Nova Scotia is Nova Scotia Power Inc. (NSPI), which is privately owned by Emera. At present, it has more than 525,000 customers including residential, commercial, and industrial customers in Nova Scotia. There are around 2000 employees in the company. Their facilities generate over 10,000 gigawatt hours of electricity every year with more than $4.3 billion worth of generation, transmission, and distribution [4]. NSPI is mainly focused on new technologies to improve customer service, reduce emissions, and add renewable energy. They have increased the use of renewable energy from 9 percent to 30 percent in the last decade and intend to achieve 80 percent renewable energy by 2030 [5].

Figure 1: NSPI employees working to restore the power
Source : https://halifax.citynews.ca/local-news

NSPI also encourages modern technologies to be used in predicting power outages based on future weather conditions. This helps them in planning for the allocation of resources to restore the power as soon as possible during and after the event [5]. Their current *storm damage prediction* model is used to predict the location, the extent of damage, and the number of person-hours required for restoring the power considering factors such as wind speed, wind direction, wind gust, precipitation levels, etc. A basic structure of the NSPI damage prediction model (DPM) is given in Figure 2 [6]. It is a regression-based model in Excel that uses Visual Basics for Application (VBA) programming language to handle the input and output of the model [6]. The mathematical equations used for predicting the damages are as follows [6]: -

$$Damages_{s,r} = \left(Q_{s,r} \sum_{j \in J} (-z_{j,r} \cdot Impact_{j,s,r}{}^2 + (z_{j,r} + 1) \cdot Impact_{j,s,r}) \right) + e_{s,r}$$

$$\begin{aligned}
Damages_{s,r} = Q_{s,r} \Bigg(\Bigg(- \Bigg(& y_{Wind,r} + Gg + Ll + \sum_{d \in D} x_{d,r} y'_{d,r} \Bigg) \cdot Impact_{Wind,s,r}{}^2 \\
& + \left(y_{Wind,r} + Gg + Ll + \sum_{d \in D} x_{d,r} y'_{d,r} + 1 \right) \cdot Impact_{Wind,s,r} \Bigg) \\
& + \big(-\big(y_{Rain,r} + 0.25 \cdot Ll\big) \cdot Impact_{Rain,s,r}{}^2 + \big(y_{Rain,r} + 0.25 \cdot Ll + 1\big) \\
& \cdot Impact_{Rain,s,r}\big) \\
& + \big(-(y_{Ice} + Ll) \cdot Impact_{Ice,s,r}{}^2 + (y_{Ice} + Ll + 1) \cdot Impact_{Ice,s,r}\big) \\
& + \big(-(y_{WetSnow} + Ll) \cdot Impact_{WetSnow,s,r}{}^2 + (y_{WetSnow} + Ll + 1) \\
& \cdot Impact_{WetSnow,s,r}\big) \\
& + \big(-(y_{Snow} + 0.5 \cdot Ll) \cdot Impact_{Snow,s,r}{}^2 + (y_{Snow} + 0.5 \cdot Ll + 1) \\
& \cdot Impact_{Snow,s,r}\big) \\
& + \big(-\big(y_{Lightning}\big) \cdot Impact_{Lightning,s,r}{}^2 + \big(y_{Lightning} + 1\big) \\
& \cdot Impact_{Lightning,s,r}\big) \Bigg) + e_{s,r}
\end{aligned}$$

Where,

$y_{j,r}$ = calibration value for weather type j in region r

$y'_{d,r}$ = calibration value for wind direction d in region r

g = calibration value for ground thaw

l = calibration value for leaves on trees

$Q_{s,r}$ = # of units of system s in region r

$Impact_{j,s,r}$ = continous variable indicating the impact on system s of weather j in region r

$x_{d,r}$ = binary variable for if the wind in region r is coming from the direction d

$$\sum_{d \in D} x_{d,r} = 1$$

G = binary variable for if ground is frozen or not (0 = frozen)

L = binary variable for if leaves are on trees or not (1 = leaves on trees)

$z_{j,r}$ = modified sensitivity for weather type j in region r

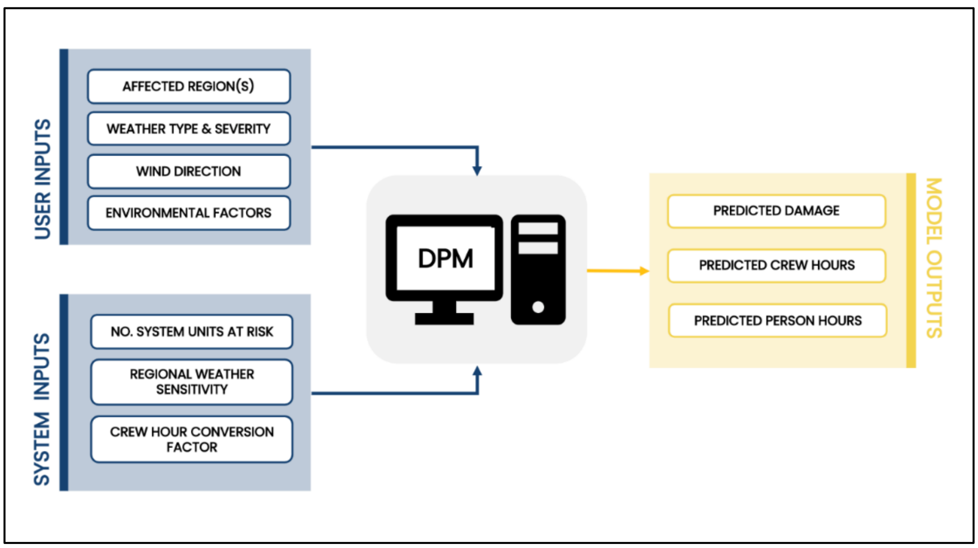

Figure 2: DPM Inputs and Outputs [6]

Nova Scotia is divided into 30 depots for a better understanding of damages in various regions, and there are 8 weather stations covering weather data of those depots [6]. The weather stations associated with the respective depots are given in Table 1. The prediction can be made for all the weather regions individually. Figure 3 represents the division of Nova Scotia into Depots and the locations of the weather stations [6].

Weather Station			*Depots*			
Yarmouth	Bridgewater	Chester	Liverpool	Yarmouth	Barrington	Shelburne
Greenwood	Bridgetown	Digby	Kingston	Clare	Coldbrook	
Halifax	Dartmouth	Halifax	Sackville	Windsor	St Margaret’s Bay	
Sydney	Sydney	Port Hawkesbury	River Bourgeois	Baddeck		
Nappan/ Debert	Parrsboro	Amherst				
Beaver Island	Goshen	Guysborough				
Caribou Pt.	Stellarton	Antigonish	Truro			
Gr. Etang	Ingonish	Cheticamp	Mabou			

Table 1: Data of weather stations used in various depots

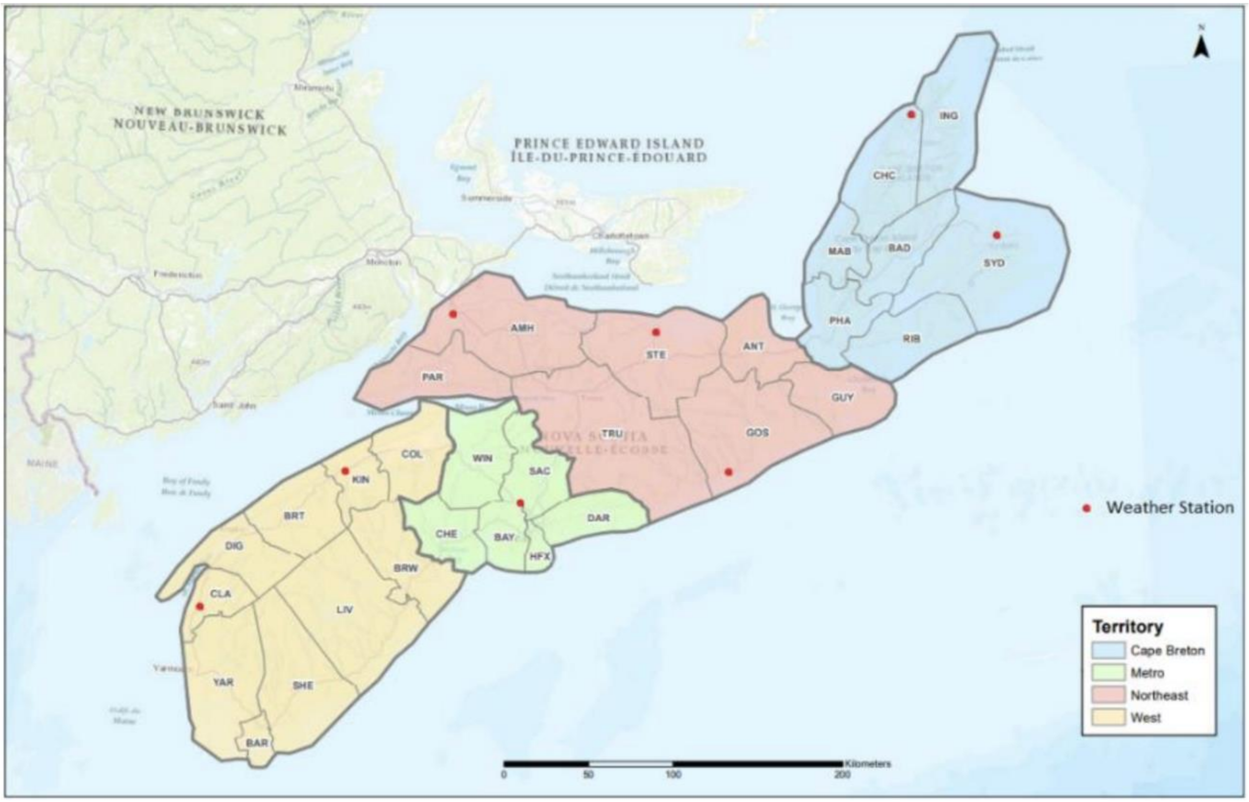

Figure 3: Division of Nova Scotia map in depots and location of weather stations [6]

The current model of NSPI predicts the damages with low accuracy. Moreover, it requires regular manual adjustments in the sensitivity factors in the mathematical model. As visible in the damage prediction equations above, the modified sensitivity of weather type j in region r $(z(j,r))$ is altered by an experienced employee of NSPI according to the requirements [6].

Storm damage prediction is not a new concept. It has been done for many years. Various technologies and programs have been developed for predicting the damages to the power grid due to harsh weather conditions. There are plenty of research papers available to understand more about the modern techniques used for damage prediction. Most of them used machine learning algorithms for creating the model. That was the motivation behind using machine learning for this project.

Machine learning is a branch of computer science that uses Data and algorithms to learn like a human brain and hence increases the accuracy of future predictions [19]. Modern machine learning algorithms like deep learning, reinforcement learning, random forests, etc. are some examples that are capable of achieving highly accurate prediction models. In this project, the concepts of deep learning (in artificial neural networks) and random forests are used. These topics are further explained in the methodology section of the report.

Initially, the most important task was to preprocess and clean the raw data provided by NSPI. It took many trials to reach the optimal dataset to be used for the model. This preprocessed dataset was then divided into training and testing sets used for training and testing the model. It also took plenty of research and trials to design and decide the values of hyperparameters of the model to increase its accuracy. The model is then used to create an executable application (.exe file) for simpler and quicker use for the prediction. This application was tested with high accuracy for the snowstorm in Halifax on 26th January 2023 which might make NSPI interested in using it in the future.

This was the outline of this project work. The Dataset format, Model, Application, and this report will be handed over to NSPI at the end of April 2023.

LITERATURE REVIEW

There is a growing interest in using machine learning algorithms to predict the damage to power systems caused by harsh weather conditions. Modern research in this area is focused on developing machine learning models that can analyze large volumes of data from multiple sources, including weather forecasts, historical weather data, and real-time sensor data from the power grid. These models can identify patterns and correlations that would be difficult for humans to discern and can make accurate predictions about the likelihood and severity of power outages and other disruptions [7].

The research by Seongmun et al. [7] proposes a machine learning-based method for predicting the state of the power grid during heavy-rain hazards. The authors use two sets of historical data - local weather data and power grid outage data - to analyze the correlated characteristics between weather and outages. The study reports that multiple weather effects can cause power outages, even under heavy-rain conditions. The proposed cost-sensitive prediction method using a Support Vector Machine (SVM) model improves the accuracy of the prediction, which is evaluated using G-mean values. The proposed method is verified using actual data from a heavy rain event that occurred in South Korea. One limitation of the paper is the lack of comparison with other prediction methods. The paper's methodology and proposed method are effective in predicting the state of the grid during heavy-rain conditions.

In another article by Swapandeep Kaur et al. [8], the authors highlight the difficulty in analyzing high-volume data of social media and the error-proneness of manual damage detection. The article proposes the use of machine learning and deep learning for automated damage detection, with a special focus on hurricane damage. Overall, the article presents an informative overview of the potential applications of machine learning and deep learning for hurricane damage detection.

Moreover, a research article by Min Li et al. [9] presents a model for predicting power outages for distribution network users during a typhoon disaster. The authors consider twenty-six explanatory variables related to meteorological, geographical, and power grid factors to predict power outages. The correlation between each explanatory variable and the response variable is analyzed, and a global variable model based on the Random Forest (RF) algorithm is established. The most important variables are then extracted to reduce the complexity of the model. The RF-important variable model is compared with other models, including a No-model, Linear Regression (LR), Support Vector Regression (SVR), Decision Tree Regression (DTR), and RF-global variable model. The results show that the RF-important variable model has a better prediction accuracy and is recommended due to its simplified structure and reduced prediction time. Overall, the article presents a useful approach for predicting power outages during typhoon disasters.

The use of machine learning algorithms for predicting damage to power systems due to harsh weather conditions represents a promising area of research that has the potential to significantly improve the resilience and reliability of the power grid in the face of increasing climate-related challenges.

Going through all the above research, a common observation is that all of them are specific for a particular weather condition. The research by Seongmun et al. [7] is specifically for heavy rain hazards and the others are for hurricanes and typhoons respectively. The objective of this project is to design a general damage prediction model that can be applied to every weather condition and can be used on daily basis by the power grids to predict outages. Moreover, there was a need for a prediction model, particularly for Nova Scotia due to the unpredictable and irregular weather conditions in this province.

METHODOLOGIES

DATA PREPROCESSING

Data preprocessing is a crucial step in machine learning that involves transforming raw data into a format that is suitable for analysis and modeling. It involves a range of techniques, including data cleaning, normalization, feature selection, and dimensionality reduction, among others. The goal of data preprocessing is to ensure that the data used for modeling is accurate, complete, and representative of the problem domain. By carefully preparing the data, machine learning models can be developed that are more accurate and effective in making predictions or identifying patterns in the data. In fact, the quality of the data used for training is often more important than the complexity of the machine learning algorithms themselves. Therefore, proper data preprocessing is critical for achieving good performance and generalizability of machine learning models [11].

Visualizing the raw data given by NSPI, the following useful information can be retrieved: -

1. Weather conditions such as wind speed, wind direction, and wind gusts from all the weather stations every hour in the past 30 years.
2. The damages caused to customers of all the depots due to the storms started from February 2020 to April 2022.
3. Storm names along with their start and end timings.
4. The weather data used for various depots from specified weather stations.
5. The data of the power distribution system i.e., number of customers, average customer per km, number of feeder sources, etc.

Several trials were carried out to clean and preprocess the above data. Initially, the hourly weather data was considered along with the name of the weather station as the inputs and in the output, the damage caused in all the cities associated with the mentioned weather station was predicted. This was considered as the problem of multiple output regression. Figure 4 depicts the printout of the excel file for the dataset.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
1	Storm Name	Weather Station	Date	WD	WS	WG	BRW	CHE	LIV	YAR	BAR	SHE	BRT	DIG	KIN	CLA	COL	DAR	HFX	SAC	WI
2	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 2:00	90	24	N/A	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
3	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 3:00	90	23	N/A	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
4	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 4:00	80	22	N/A	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
5	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 5:00	90	22	N/A	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
6	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 6:00	90	23	N/A	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
7	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 7:00	90	18	N/A	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
8	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 8:00	70	17	N/A	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
9	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 9:00	90	17	N/A	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
10	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 10:00	100	25	N/A	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
11	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 11:00	110	24	N/A	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
12	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 12:00	120	22	N/A	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
13	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 13:00	110	27	N/A	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
14	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 14:00	110	16	N/A	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
15	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 15:00	100	28	N/A	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
16	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 16:00	100	18	N/A	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
17	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 17:00	120	23	N/A	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
18	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 18:00	190	41	66	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
19	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 19:00	210	40	64	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
20	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 20:00	200	45	72	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
21	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 21:00	210	47	75	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
22	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 22:00	220	63	101	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
23	Early Feb 2020 Ice Storm	Yarmouth	2020-02-07 23:00	230	74	118	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
24	Early Feb 2020 Ice Storm	Yarmouth	2020-02-08 0:00	250	68	109	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	
25	Early Feb 2020 Ice Storm	Yarmouth	2020-02-08 1:00	270	58	93	11292	185	7114	8255	1872	4254	0	0	0	0	0	0	0	0	

Figure 4: Dataset with inputs as Wind Speed (WS), Wind Direction (WD), Wind Gusts (WG) and 30 output columns each representing damage in a city or a depot of Nova Scotia

The problem with this dataset was that 30 outputs were to be predicted from only 3 inputs. It was becoming extremely difficult with this dataset to train any machine learning model.

Therefore, another dataset was created in which excel files for each weather region were created separately considering the damages only in the depots associated with the respective weather station. The dataset for the Halifax weather station/weather region is shown in Figure 5.

	A	B	C	D	E	F	G	H	I
1	WD	WS	WG	DAR	HFX	SAC	WIN	BAY	
2	70	11	N/A	6719	13423	2347	2234	2753	
3	90	17	N/A	6719	13423	2347	2234	2753	
4	80	19	N/A	6719	13423	2347	2234	2753	
5	80	19	N/A	6719	13423	2347	2234	2753	
6	30	17	N/A	6719	13423	2347	2234	2753	
7	40	17	N/A	6719	13423	2347	2234	2753	
8	40	17	N/A	6719	13423	2347	2234	2753	
9	50	26	N/A	6719	13423	2347	2234	2753	
10	20	20	N/A	6719	13423	2347	2234	2753	
11	50	24	N/A	6719	13423	2347	2234	2753	
12	50	28	28	6719	13423	2347	2234	2753	
13	70	19	N/A	6719	13423	2347	2234	2753	
14	60	20	N/A	6719	13423	2347	2234	2753	
15	70	20	N/A	6719	13423	2347	2234	2753	
16	110	19	N/A	6719	13423	2347	2234	2753	
17	120	24	35	6719	13423	2347	2234	2753	
18	100	20	N/A	6719	13423	2347	2234	2753	
19	120	26	30	6719	13423	2347	2234	2753	
20	110	15	N/A	6719	13423	2347	2234	2753	
21	120	19	N/A	6719	13423	2347	2234	2753	

Halifax

Figure 5: Dataset with inputs as Wind Speed (WS), Wind Direction (WD), Wind Gusts (WG) and outputs as damages in all the depots associated with Halifax

The problem is also with the hourly data that the input data is changing with every hour, but the damage is constant for a specific storm as a single value for damage is given by NSPI for a storm. Therefore, it creates ambiguity for a model to learn as the dataset is giving the same output with a variety of inputs. So, the feasible solution here is to consider the weather data of a specific hour of the storm when the wind speed is maximum as evident theoretically that wind speed is the major cause of damage. It is also proven in the results in Figure 17 that the inputs with the highest importance are wind gusts and wind speed.

In another trial, the nearest four weather stations were considered for predicting damages in a weather region. The hourly data was an issue here as well. The excel snapshot for the same is given in Figure 6.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	WD	WS	WG	WD	WS	WG	WD	WS	WG	WD	WS	WG	WD	WS	WG	DAR	HFX	SAC	WIN	BAY
2	90	24	N/A	60	20	N/A	70	11	N/A	60	6	N/A	30	7	N/A	6719	13423	2347	2234	2753
3	90	23	N/A	70	17	N/A	90	17	N/A	50	6	N/A	40	6	N/A	6719	13423	2347	2234	2753
4	80	22	N/A	60	20	N/A	80	19	N/A	60	7	N/A	30	11	N/A	6719	13423	2347	2234	2753
5	90	22	N/A	60	19	N/A	80	19	N/A	50	7	N/A	40	7	N/A	6719	13423	2347	2234	2753
6	90	23	N/A	80	11	N/A	30	17	N/A	70	9	N/A	40	11	N/A	6719	13423	2347	2234	2753
7	90	18	N/A	70	15	28	40	17	N/A	50	15	N/A	50	13	N/A	6719	13423	2347	2234	2753
8	70	17	N/A	60	19	28	40	17	N/A	40	19	N/A	30	19	N/A	6719	13423	2347	2234	2753
9	90	17	N/A	60	13	N/A	50	26	N/A	50	17	N/A	50	13	N/A	6719	13423	2347	2234	2753
10	100	25	N/A	80	19	N/A	20	20	N/A	40	24	35	40	13	N/A	6719	13423	2347	2234	2753
11	110	24	N/A	60	22	N/A	50	24	N/A	40	19	N/A	50	13	N/A	6719	13423	2347	2234	2753
12	120	22	N/A	60	19	N/A	50	28	28	40	26	N/A	50	15	N/A	6719	13423	2347	2234	2753
13	110	27	N/A	70	20	30	70	19	N/A	60	20	33	50	13	N/A	6719	13423	2347	2234	2753
14	110	16	N/A	60	13	32	60	20	N/A	60	26	35	40	9	N/A	6719	13423	2347	2234	2753
15	100	28	N/A	80	17	30	70	20	N/A	60	22	N/A	40	9	N/A	6719	13423	2347	2234	2753
16	100	18	N/A	50	22	N/A	110	19	N/A	70	24	33	20	15	N/A	6719	13423	2347	2234	2753
17	120	23	N/A	70	13	35	120	24	35	60	20	33	40	13	N/A	6719	13423	2347	2234	2753
18	190	41	66	60	22	N/A	100	20	N/A	70	26	33	40	9	N/A	6719	13423	2347	2234	2753
19	210	40	64	70	19	N/A	120	26	30	70	19	30	50	9	N/A	6719	13423	2347	2234	2753
20	200	45	72	90	26	48	110	15	N/A	50	20	N/A	60	9	N/A	6719	13423	2347	2234	2753
21	210	47	75	50	24	N/A	120	19	N/A	80	24	35	50	9	N/A	6719	13423	2347	2234	2753

Figure 6: Hourly weather data of nearest four weather stations taken as inputs and the damages in five depots of Halifax as outputs

After many more experiments and solving all the above shortcomings, a dataset for this case was prepared through the following preprocessing: -

1. Extract the weather data specifically for storms using start and end time (Figure 7)
2. Choose the point where wind speed was maximum using MAX function (Figure 7)

	A	B	C	D	E	F	G	H	I
1	Yarmouth	Date	WD	WS	WG	AT	DP	WX	
2592		2022-04-19 9:00	120	33	48	5	0	RA BR	
2593		2022-04-19 10:00	120	37	57	5	1	-RA BR	
2594		2022-04-19 11:00	120	39	59	5	1	-RA BR	
2595		2022-04-19 12:00	120	39	67	6	1	RA BR	
2596		2022-04-19 13:00	120	52	74	5	4	-RA BR	
2597		2022-04-19 14:00	120	50	72	5	5	RA BR	
2598		2022-04-19 15:00	130	44	74	6	5	-RA BR	
2599		2022-04-19 16:00	120	59	85	7	6	RA BR	
2600		2022-04-19 17:00	130	48	85	8	7	RA BR	
2601		2022-04-19 18:00	130	48	80	9	8	-RA BR	
2602		2022-04-19 19:00	150	41	74	10	9	RA BR	
2603		2022-04-19 20:00	220	37	94	12	11	RA	
2604		2022-04-19 21:00	190	35	50	11	11	-RA	
2605		2022-04-19 22:00	220	24	39	8	7	-RA	
2606		2022-04-19 23:00	200	24	37	7	7		
2607		2022-04-20 0:00	200	22	33	7	6		
2608		2022-04-20 1:00	200	24	35	6	6		
2609		2022-04-20 2:00	200	24	35	6	6		
2610		2022-04-20 3:00	200	24	39	6	5		
2611		2022-04-20 4:00	200	22	N/A	6	5		
2612		2022-04-20 5:00	230	20	39	5	4		

2022 | 2021 | 2020 | 2019 | 2018 | 2017 | 2016 | 2015 | 2014 | 2013 | 2012

Figure 7: Retrieving data of storm from hourly weather data of an year and selecting the hour with maximum wind speed

3. Include inputs like storm time and month
4. Convert wind direction from degrees to actual direction using several IF conditions in excel (Figure 8)

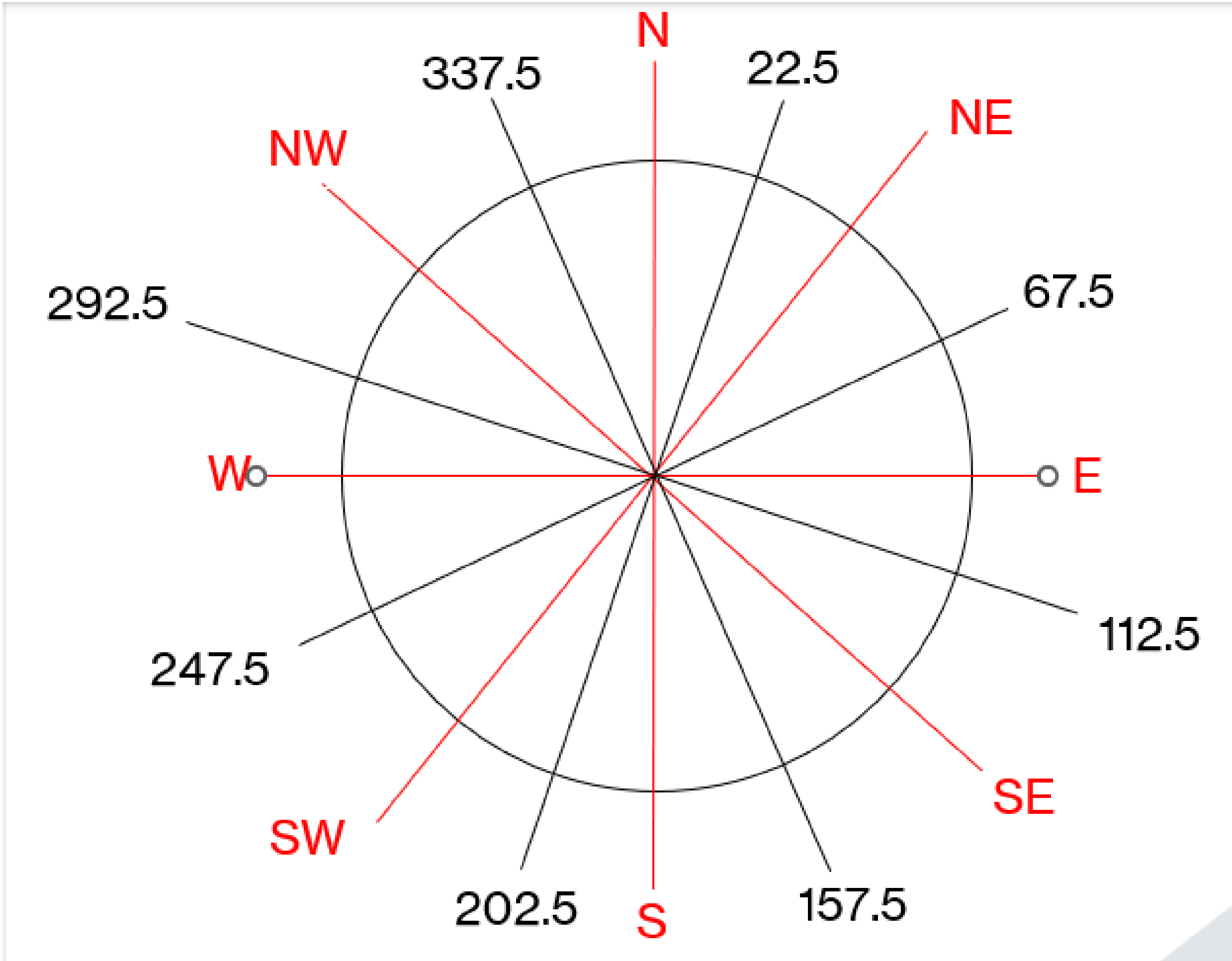

Figure 8: Conversion of wind direction from degrees to actual directions in the range of 45°

5. All the strings are one hot encoded using get_dummies function of pandas (Figure 9)
6. Convert number of customers to percentage using the total customers in each depot (Figure 9)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	WS	WG	Storm Time (hrs.)	Percentage Customers Affected	AMH	ANT	BAD	BAR	BAY	BRT	BRW	CHC	CHE
2	74	118	61.07	51.21	0	0	0	0	0	0	1	0	0
3	62	98	20.65	56.84	0	0	0	0	0	0	1	0	0
4	48	67	32	1.58	0	0	0	0	0	0	1	0	0
5	67	89	24	0.09	0	0	0	0	0	0	1	0	0
6	56	56	28.5	5.90	0	0	0	0	0	0	1	0	0
7	46	67	21.5	1.32	0	0	0	0	0	0	1	0	0
8	52	76	25.5	4.85	0	0	0	0	0	0	1	0	0
9	56	67	46.5	0.00	0	0	0	0	0	0	1	0	0
10	48	69	16	10.04	0	0	0	0	0	0	1	0	0
11	41	56	35.5	0.00	0	0	0	0	0	0	1	0	0
12	52	70	13.5	8.02	0	0	0	0	0	0	1	0	0
13	61	85	30	16.08	0	0	0	0	0	0	1	0	0
14	61	83	26	47.37	0	0	0	0	0	0	1	0	0
15	57	93	20	6.16	0	0	0	0	0	0	1	0	0
16	43	52	99.5	31.76	0	0	0	0	0	0	1	0	0
17	56	78	16	4.74	0	0	0	0	0	0	1	0	0
18	59	85	19.5	0.52	0	0	0	0	0	0	1	0	0
19	74	118	61.07	2.09	0	0	0	0	0	0	0	0	1
20	62	98	20.65	0.87	0	0	0	0	0	0	0	0	1
21	48	67	32	17.53	0	0	0	0	0	0	0	0	1

Figure 9: All the strings including Wind Directions, Depot Name, and the Storm Month are one hot encoded and the number of customers converted to percentage of customers affected with respesct to total number of customers

7. Use the weather data of weather station at the region and the nearest 3 weather stations (Figure 10) according to Table 2
8. Inputs variables are standardized using SK-Learn
9. Outputs are normalized using SK-Learn

Some of the most efficient machine learning algorithms for the given dataset are deep learning (artificial neural networks) and random forests [19]. Artificial neural networks are one of the best fit for the damage prediction problem because of their capability to learn complex non-linear relationships of the problem. Random forests algorithm is mainly used due to its robustness to overfitting as it uses aggregated results from multiple decision trees. The main advantage of these models is to perform well with large number of input features. The implementation of those algorithms is carried out in this project using python programming and the Scikit-Learn library.

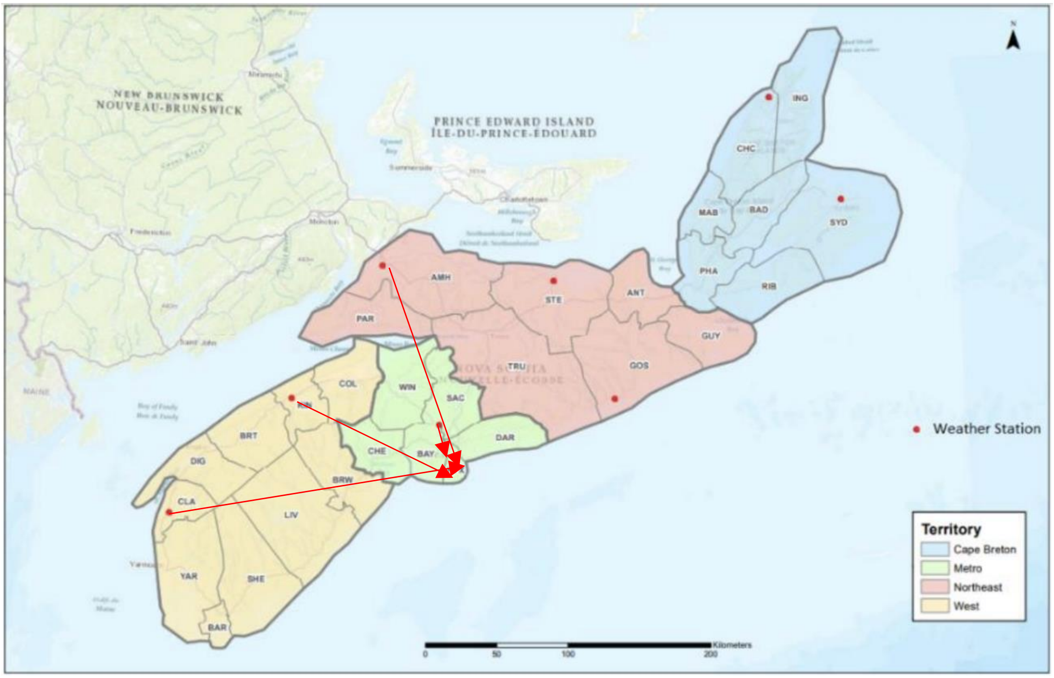

Figure 10: Using weather stations of Yarmouth, Greenwood, and Nappan/Debert for the damage prediction of Halifax

Weather Region	***Weather stations considered for predicting damages***
Yarmouth	Yarmouth, Greenwood, Halifax, Nappan/Debert
Greenwood	Greenwood, Yarmouth, Halifax, Nappan/Debert
Halifax	Halifax, Yarmouth, Greenwood, Nappan/Debert
Sydney	Sydney, Beaver Island, Caribou Pt., Gr. Etang
Nappan/Debert	Nappan/Debert, Yarmouth, Greenwood, Halifax
Beaver Island	Beaver Island, Sydney, Caribou Pt., Gr. Etang
Caribou Pt.	Caribou Pt., Sydney, Beaver Island, Gr. Etang
Gr. Etang	Gr. Etang, Sydney, Caribou Pt., Beaver Island

Table 2: Weather stations used for inputs of weather data for various regions

MACHINE LEARNING WITH PYTHON AND SCIKIT-LEARN

Python programming and Scikit-Learn tools have become popular for machine learning, providing a wide range of advantages and capabilities.

Python is a general-purpose programming language that has become one of the most widely used languages for data science and machine learning. One of the primary advantages of Python is its simplicity and ease of use. Python's syntax is designed to be easy to read and write, making it accessible for beginners and experienced programmers alike [12].

Python also has a large and active community that has developed many powerful libraries and tools for machine learning. These libraries, including NumPy, Pandas, Matplotlib, and Seaborn, provide powerful data manipulation, visualization, and analysis capabilities [12].

Scikit-Learn is a popular machine learning library that is built on top of Python's scientific computing libraries. Scikit-Learn provides a wide range of machine learning algorithms for tasks such as classification, regression, clustering, and dimensionality reduction [13].

One of the primary advantages of Scikit-Learn is its consistency and simplicity. Scikit-Learn provides a consistent API for all of its machine learning algorithms, making it easy to experiment with different algorithms and compare their performance. The library also provides powerful tools for data preprocessing and feature engineering, which are critical steps in any machine learning project [13].

Python and Scikit-Learn are powerful and flexible tools for machine learning, providing a wide range of advantages for data scientists and machine learning engineers. Their simplicity, flexibility, and performance make them the ideal choice for developing machine learning models and solving complex data science problems.

ARTIFICIAL NEURAL NETWORKS

Artificial neural networks are built with the idea of how signals are transmitted in the human brain through layers of neurons. Theoretically, there are 3 types of layers namely, an input layer, an output layer, and a hidden layer. There is one input layer and one output layer with the number of nodes depending upon the number of inputs given and several outputs required respectively. The number of hidden layers can vary according to the complexity of the problem. The model is known as a deep learning model when the number of hidden layers increases significantly. The signals are transmitted from each node of one layer to each node of the next layer. Initially, input to each node is multiplied by a weight factor and the bias factor is added to the whole. This mathematical equation gives an output which is then transmitted to a node of another layer as an activation. The same calculation takes place at this node with the activation received and the process continues till an output is achieved at the output layer. These processes and calculations take place in the entire neural network.

Talking a little bit about the dataset, if the output is known in the dataset, it is known as supervised machine learning. Otherwise, it is known as unsupervised machine learning. In this project the output i.e., the number of customers affected in the past is known. Therefore, it is supervised machine learning. The neural network hereby adjusts its weights and biases according to the given outputs and learns the trend of the data using the algorithm known as backpropagation. Also, formulas of calculus are applied to the model to decrease the error in the predictions by calculating the global minima of the function. This concept is known as gradient descent. This was a brief explanation of how neural networks work [19]. The Figure 11 represents a schematic view of a simple neural network.

A simple mathematical representation of a neural network can be described as follows:

Let x be an input vector of dimension n, and y be the output vector of dimension m. A neural network is a function $f\ (x;\ \theta)$ that maps the input vector x to the output vector y, where θ represents the set of weights and biases in the network [17].

The output of each neuron in the hidden layer is calculated using the following formula [17]:

$$z = w * x + b$$

where w is a weight vector, b is a bias vector, and $*$ denotes the dot product operation.

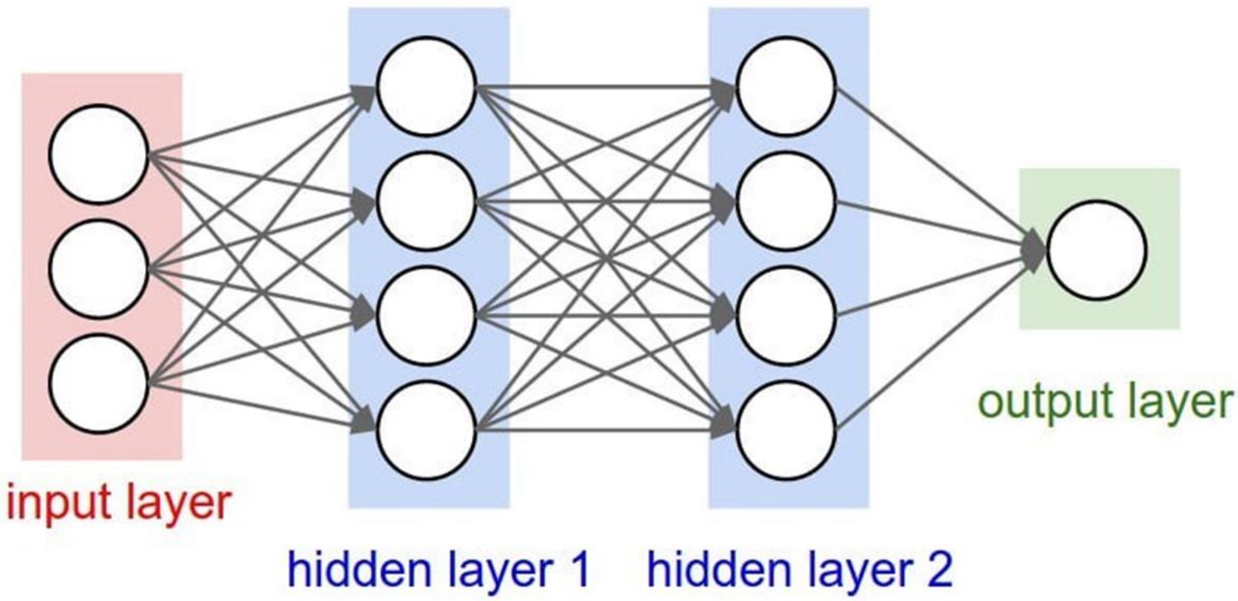

Figure 11: Basic Artificial Neural Network

Source : https://www.skynettoday.com/overviews/neural-net-history

The output of each neuron in the hidden layer is then passed through an activation function, such as the sigmoid or ReLU function, to deal with the non-linearity of the data.

The output of the entire neural network can be represented as [17]:

$$y = f(x;\theta) = g(w_2 * g(w_1 * x + b_1) + b_2)$$

where w_1, w_2 are weight matrices, b_1, b_2 are bias vectors, g is the activation function, and the subscripts denotes the layer number.

The pseudocode for the backpropagation algorithm can be written as follows [16]:

PSEUDOCODE FOR BACKPROPAGATION

```
Input:
X: Input data
Y: Target output
L: Number of layers in the network
W: Weights of the network
b: Biases of the network
alpha: Learning rate
epochs: Number of training epochs

Output:
W: Trained weights of the network

for i in range(epochs):
  # Forward propagation
  Z = []
  A = [X]
  for l in range(1, L):
    z_l = W[l] * A[l − 1] + b[l]
    a_l = activation_function(z_l)
    Z.append(z_l)
    A.append(a_l)

  # Error calculation
  dZ = [None] * L
  dA = (A[−1] − Y) / Y.shape[0]
  for l in reversed(range(L)):
    dZ[l] = dA * activation_function_derivative(Z[l])
    if l != 0:
      dA = W[l].T * dZ[l]

  # Backward propagation
  dW = [None] * L
  db = [None] * L
  for l in range(L):
    dW[l] = A[l].T * dZ[l]
    db[l] = np.sum(dZ[l], axis = 1, keepdims = True)

  # Gradient descent update
  for l in range(L):
    W[l] = W[l] − alpha * dW[l]
    b[l] = b[l] − alpha * db[l]

return W
```

Implementation: -

1. The dataset is split into training and testing sets in a proportion of around 80% and 20% respectively. These proportions are ideal values for any model to be trained and tested as it makes a balance keeping enough data for effective training of the model and enough data to evaluate the performance of the model. Reducing either of them reduces data for either training or testing the model respectively [19].
2. The type of model used here is sequential neural networks. A sequential neural network, also known as a recurrent neural network (RNN), is a type of neural network that is designed to process sequential data such as time series, natural language text, and speech signals. Unlike traditional feedforward neural networks that process a fixed input size, sequential neural networks have a feedback loop that allows them to process sequential inputs of varying lengths [14].
3. The layers are created using an open-source library in python called Keras.
4. After the dataset is one-hot encoded, the number of inputs is 78 in the input layer.
5. The number of hidden layers is 2 with the number of nodes equal to 40 and 20 respectively. It is observed from experiments that, in this case, the model performs satisfactorily while taking the number of nodes equal to half of the nodes in the previous layer. And taking more than 2 layers causes overfitting as it learns to fit the noise rather than predicting the accurate relationship between the inputs and output. Another reason for not exceeding 2 layers is the lack of sufficient data [19].
6. Finally, there is only one output of the percentage of customers affected in the specified depot. Therefore, in the output layer, there is a single node.
7. The activation function used in the input layer and the hidden layers is Rectified Linear Unit (ReLU). It is a popular activation function used in deep-learning neural networks. It is defined as $f(x) = max\ (0, x)$, where x is the input to the function. The ReLU function allows the neural networks to learn complex non-linear relationships between inputs and outputs efficiently, while also avoiding the vanishing gradient problem and improving generalization ability [15]. The graph of the ReLU function is given in Figure 12.

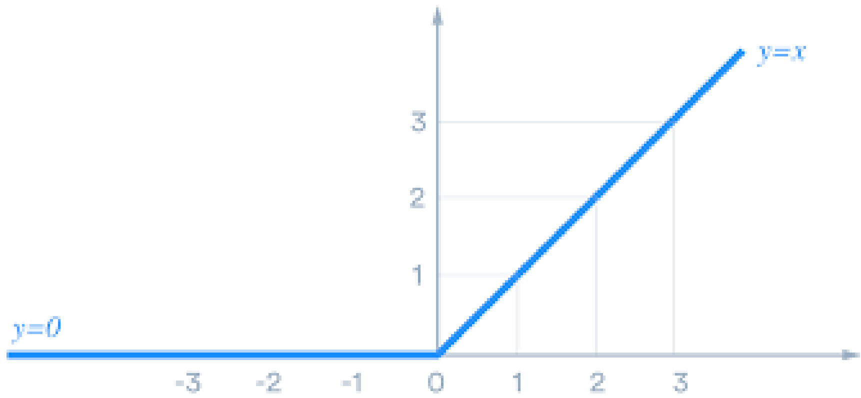

Figure 12: Graph of ReLU function

Source : https://hemanthindala2001.medium.com/types-of-activation-functions

8. The activation function used in the output layer is sigmoid. The sigmoid function is a popular activation function used in machine learning, particularly in binary classification problems. It is a mathematical function that maps any input value to a value between 0 and 1. The sigmoid function is defined as $f(x) = 1 / (1 + exp(-x))$, where x is the input to the function. It has a smooth gradient and is easy to interpret, making it ideal for certain types of problems like this project where output is expected in terms of the percentage of customers affected [15]. The graph of the sigmoid function is represented in Figure 13.

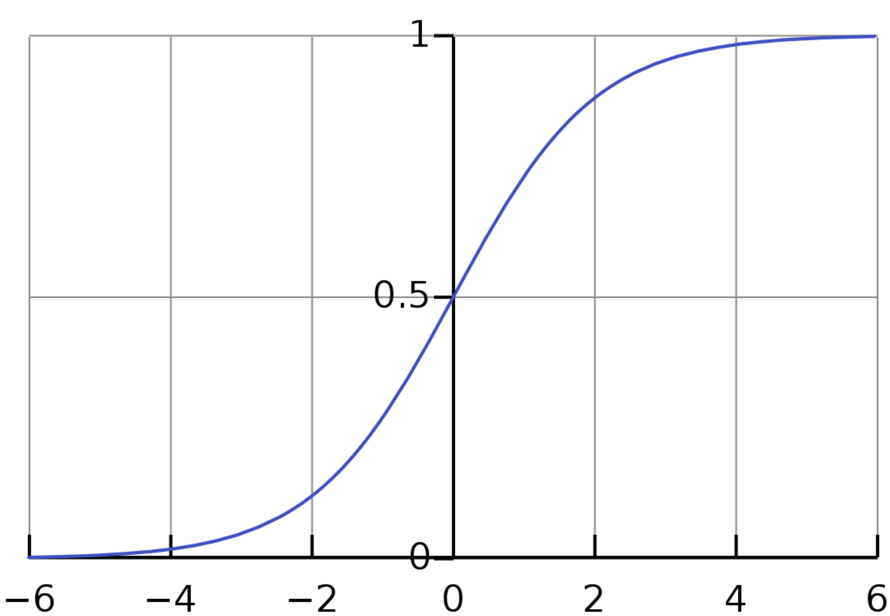

Figure 13: Graph of sigmoid function

Source : https://medium.com/@martinpella/logistic-regression-from-scratch-in-python

9. The following values for the hyperparameters are set for achieving the high accuracy of the model: -

   i) Optimizer = 'adam'

   ii) loss = 'mean_squared_error (MSE)'

   iii) epochs = 50

   iv) batch_size = 64

   These hyperparameter values were selected based on their best match of capabilities with the problem and after experimenting with a wide range of parameter values. Specifically, for the optimizer, 'Adam' is used due to its capability to adapt the learning rate while training based on stochastic gradient descent. MSE loss function is commonly used for the regression problems as it is highly efficient in minimizing distance between actual and predicted values. It ultimately affects the backpropagation algorithm for the training of neural networks [19]. For the number of epochs, values such as 75, 100, 200, etc. were also tried, yet they have not led to significant improvements over 50 epochs and often led to overfitting. Various batch sizes starting from 16 and its multiples were used for experiments. The smaller batch sizes were taking a lot of time to train the model while the larger ones caused the loss in accuracy. Therefore, the optimal value found here is 64.

RANDOM FORESTS

Another machine learning model that is conventionally considered more accurate than neural networks in some cases is random forests [19]. Random forests work by taking majority of the outputs of multiple Decision Trees which constitutes the forest to achieve more accurate results. This is also known as ensemble learning. Decision trees are supervised machine-learning algorithms that consist of branches and nodes. Each node signifies a feature of the model, and the branches are the conditions to be applied for going from one node to another. From each node, the signal transmits to the next node wherever the mathematical condition is satisfied. The model used in this project is the random forests regression model which is also a supervised machine learning. Figure 14 represents a schematic view of a random forest network.

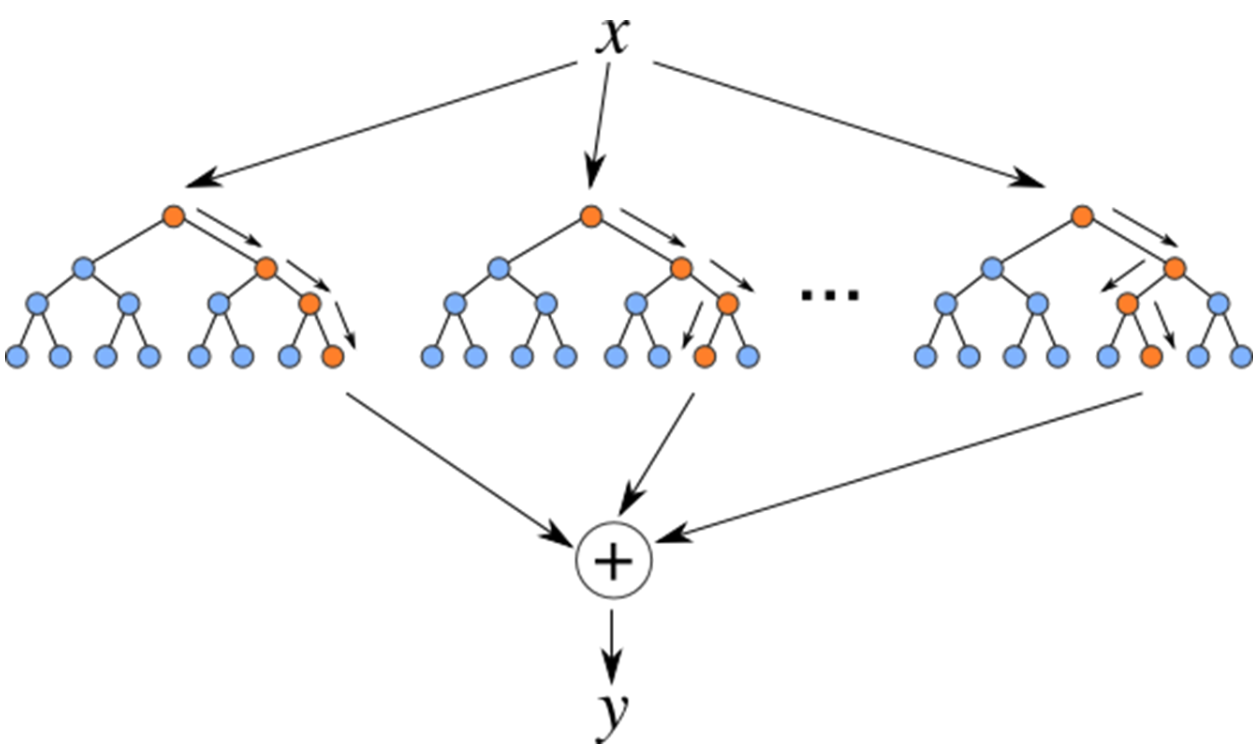

Figure 14: Random Forests Algorithm
Source : https://levelup.gitconnected.com/random-forest-regression

Implementation [18]: -

1. The Random Forest model utilizes a sampling technique to create multiple decision trees. This technique involves selecting a subset of data points and a subset of features from the original dataset. Specifically, a random subset of n records and m features are chosen from the total k records in the dataset to construct each decision tree. By doing so, the model can reduce overfitting and increase the accuracy of its predictions.
2. Decision trees are created for each sample and the output is obtained individually.
3. The predicted value is the average of the predicted values obtained from each decision tree in the case of regression.

To measure the amount of uncertainty in the dataset, random forests use the metric of entropy. In the context of Random Forest Regression, entropy is used to evaluate the quality of a split in a decision tree. The goal is to select the feature and threshold that will result in the greatest reduction in entropy between the parent node and the child nodes. The formula for entropy is:

$$E = -sum(p(i) * log2(p(i)))$$

where E is the entropy, $p(i)$ is the proportion of data points in the ith class, and $log2$ is the base-2 logarithm. The entropy is highest when the data is evenly distributed across all classes, and lowest when all data points belong to a single class.

By using entropy as a measure of feature importance, the Random Forest model can effectively identify the most relevant features for predicting the target variable and create decision trees that generalize well to new data.

Executable applications (.exe file) were also implemented for both the models using pyinstaller. This application can be readily used for damage prediction giving the weather data as inputs.

RESULTS

The mean absolute error is calculated for both the models and is found almost the same around **0.05** which signifies low deviation from the actual outputs. It can also be inferred that the accuracy of models is about **95 percent**. The mean absolute errors from neural networks and random forests are 0.0537 and 0.0497 respectively. It is encircled in Figure 15 and Figure 16 along with the output of the test set respectively.

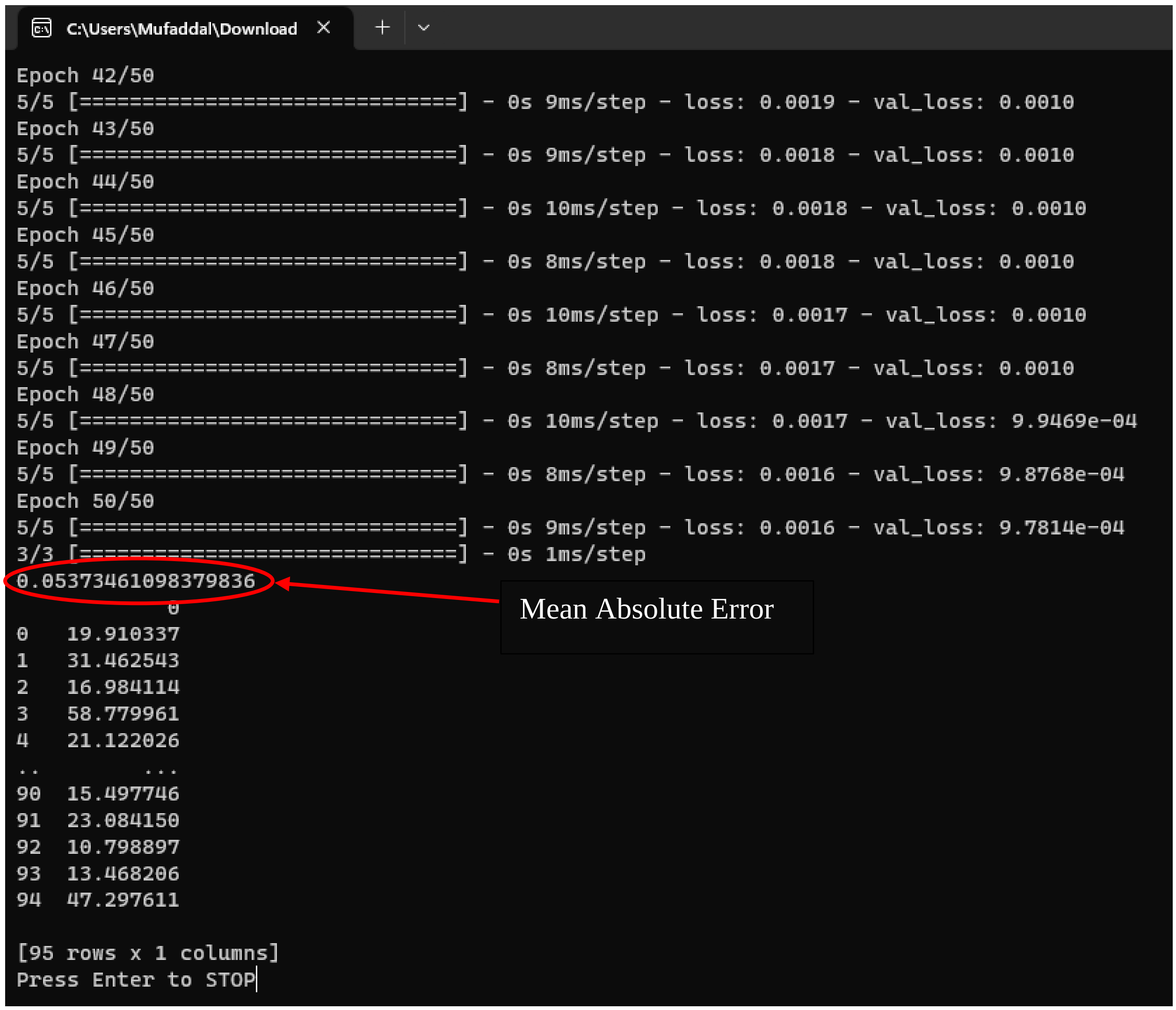

Figure 15: Output of neural network model. Mean absolute error = 0.0537and outputs of test data i.e., damage to percentage of customers

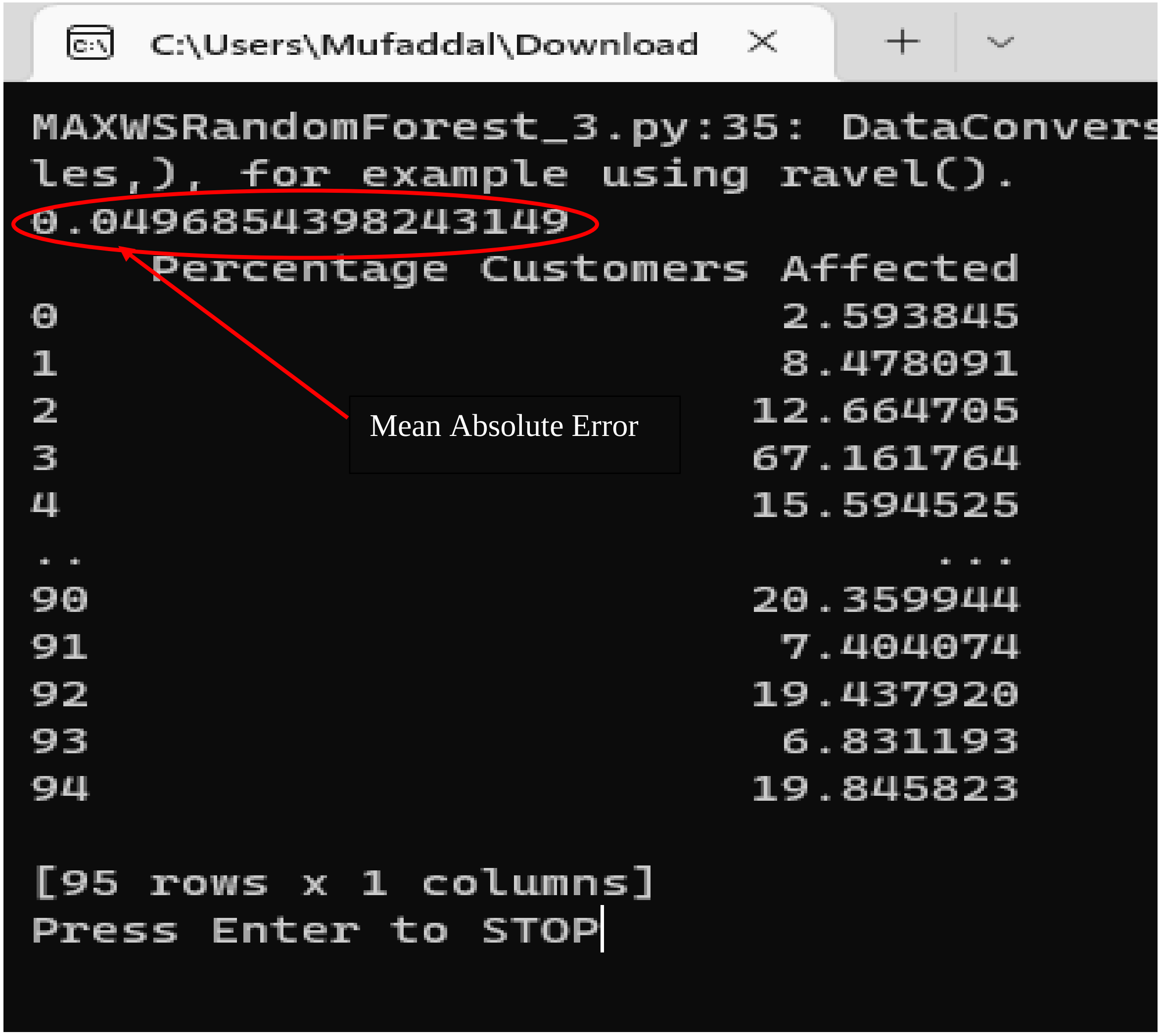

Figure 16: Output of random forests model. Mean absolute error = 0.0497 and outputs of test data

All the features were plotted against their importance (Figure 17) to make the following inferences:

1. The features with the highest importance or the features that are majorly responsible for the number of outages in a city are wind gust of the same city and storm duration.
2. It can also be noted that, in most cases, whenever the wind direction of the first nearest neighbour of the depot will be northwest, it will cause a significant damage to the power grid of that depot.
3. Excessive outages occur in Nova Scotia during the month of April as it has the highest importance among all the months in the bar graph.

4. The depots that are tremendously prone to outages are Antigonish, Chester, Digby, Ingonish, and Windsor.

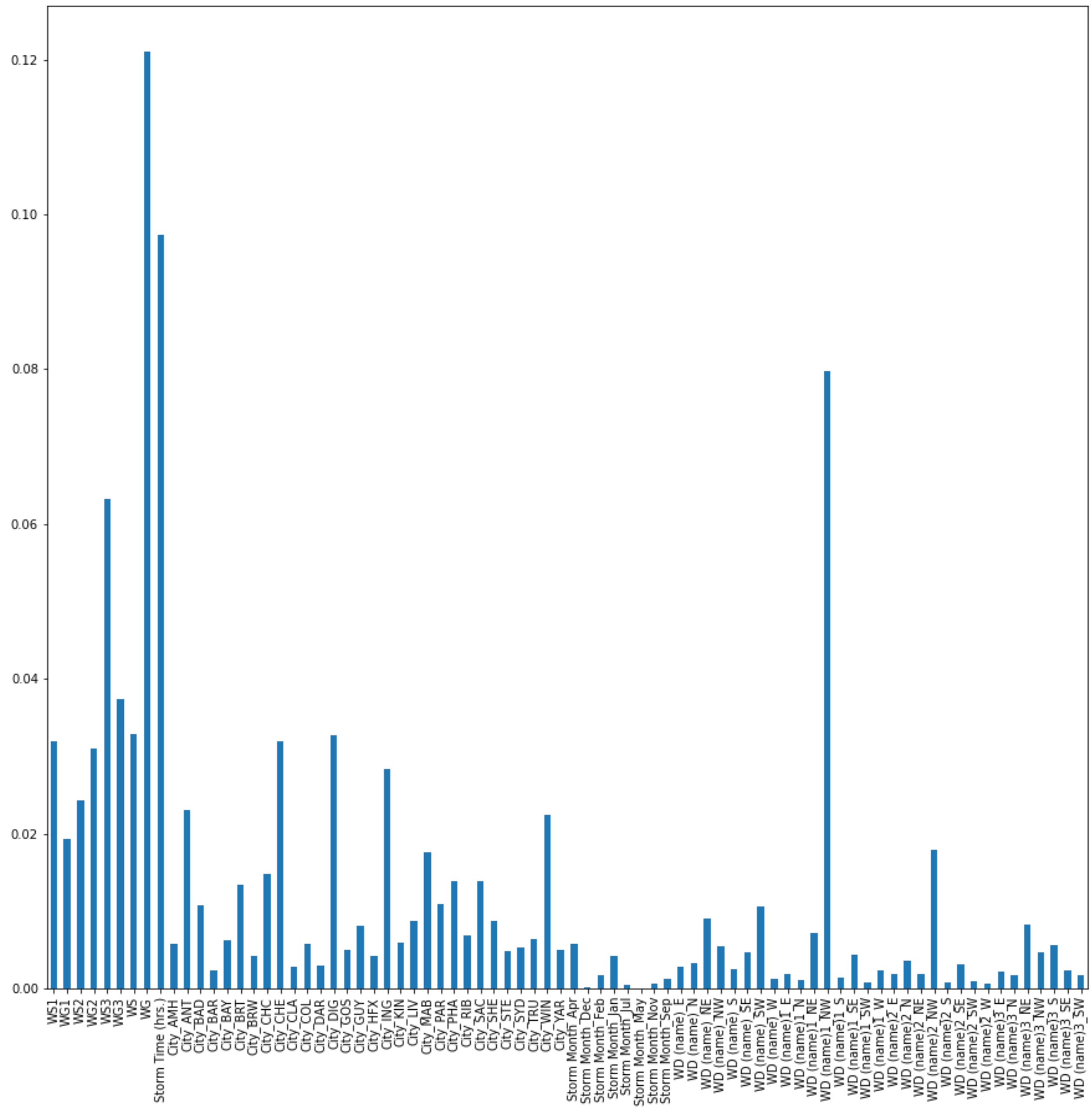

Figure 17: Features importance graph of Random Forests

The EXE application is also tested on a storm in Halifax on 26/01/2023. The total number of NSPI customers in Halifax is around 110,000. The outputs of damages from neural networks and random forests are 38.51 percent and 37.25 percent that signifies damages to about 42,361 and 40,975 customers respectively in Halifax. The actual number of damages is not disclosed by NSPI, but it can be observed from a news channel that the number of outages in Halifax on 26/01/2023 was around 41,000 [10]. Considering this actual output, the accuracy of neural network is around 97 percent and that of random forests is more than 99 percent. The respective outputs can be seen in Figure 18 and Figure 19.

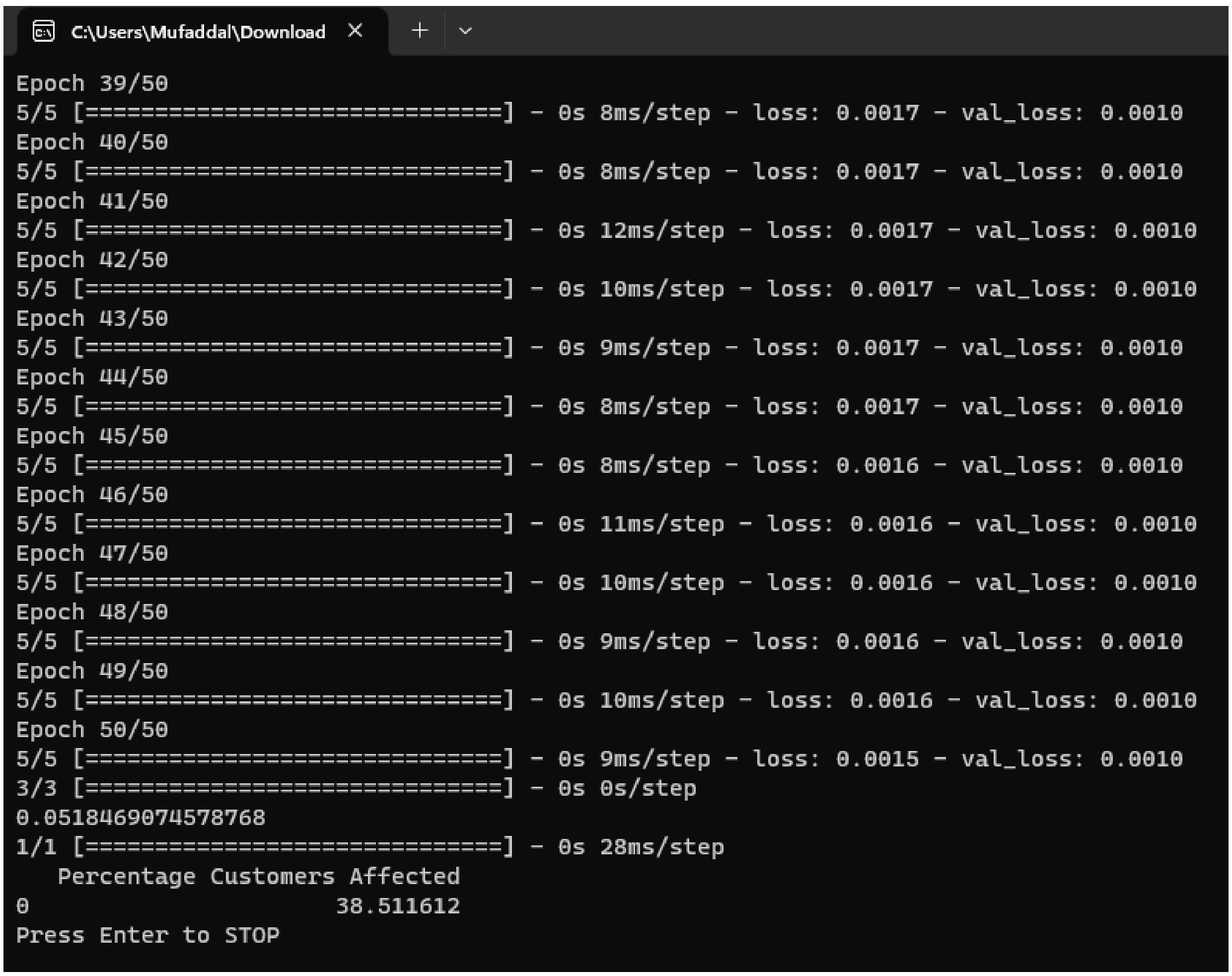

Figure 18: Application of neural networks on the recent storm - 26/01/2023 (Halifax)

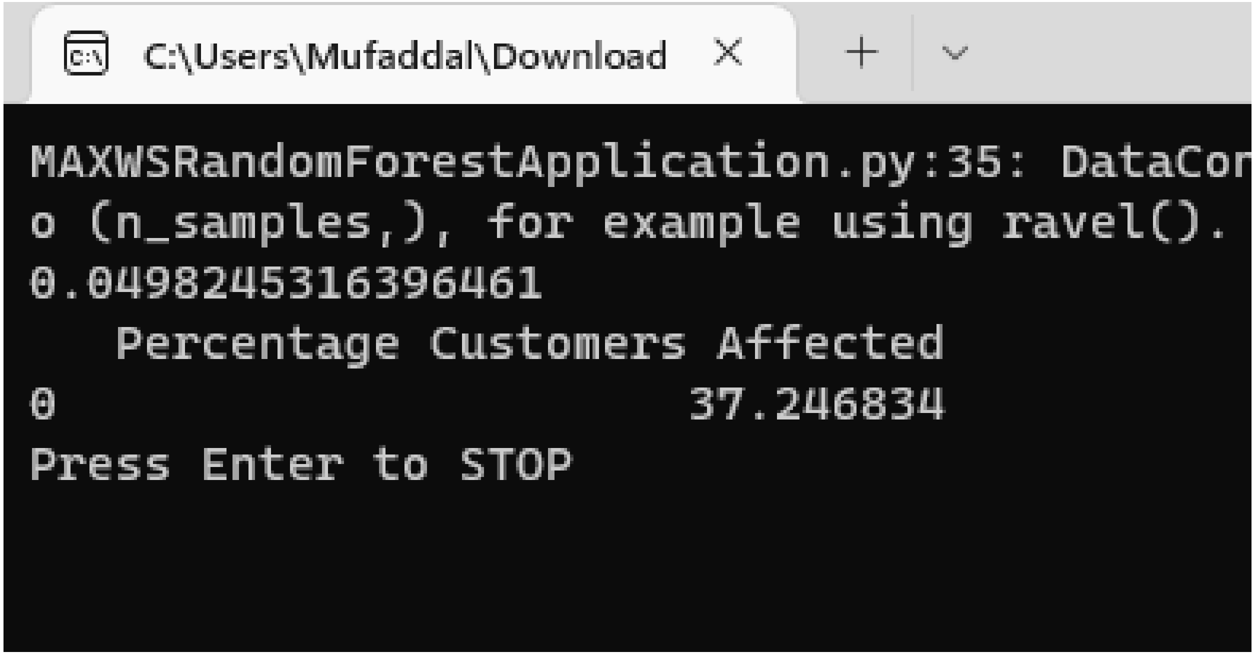

Figure 19: Application of random forests on the recent storm - 26/01/2023 (Halifax)

CONCLUSIONS

After completing this project, it is clear that machine learning models offer significant advantages over conventional regression models, particularly when it comes to damage prediction. In contrast to relying on employee experience, which can be subject to error, machine learning models provide highly accurate predictions that can be easily verified.

This project demonstrates the potential of machine learning models for damage prediction in the power grid industry, and specifically for use by NSPI. The model developed can be readily deployed and utilized for damage prediction, and with modifications, can also be adapted for use in other power grids.

Furthermore, the wide range of applications of modern machine learning algorithms can be used to predict damage to other assets across a province or country, caused by severe weather conditions. These models offer a highly effective means of predicting and mitigating potential damage, thereby saving time and resources, and helping to protect people and infrastructure from the impact of harsh weather.

In summary, this project highlights the power and versatility of machine learning models and their potential to revolutionize the way we approach damage prediction in the power grid industry and beyond. With further development and refinement, these models have the potential to transform the way we predict and manage damage caused by a range of environmental factors, ultimately contributing to a safer and more resilient future for all.

FUTURE WORK

There are several avenues for future work that can be pursued based on the results and findings of this project. Some potential directions for further research and development include:

1. Evaluating the model's performance over time: As weather patterns and environmental conditions change over time, it is important to monitor the model's performance and assess its effectiveness. Future work could involve ongoing validation of the model's predictions using real-world data.
2. Improving feature selection: While this project incorporated historical weather data, there may be other features that can be included to further enhance the model's accuracy. Further exploration of additional features could be performed to identify which ones could be most useful for predicting power grid damage.
3. Getting more data from history: By expanding the dataset used to train the model, the model may be able to identify additional patterns and relationships between weather conditions and power grid damage, resulting in more accurate and reliable predictions.
4. Exploring classification for damage prediction: While this project focused on regression models, future work could explore the use of classification techniques for damage prediction. By categorizing potential damage into different classes, the model could provide more detailed and actionable information to help power grid operators better prepare for and respond to potential damage events.
5. Incorporating external data sources: In addition to weather data, there may be other external data sources that could be useful for predicting power grid damage. For example, data on infrastructure age or maintenance history could be used to identify areas that are more susceptible to damage and prioritize maintenance and repair efforts.
6. Scaling the model to cover larger areas: While this model was designed for the province of Nova Scotia, it could be adapted for use in other regions or countries. By incorporating additional data sources and features, the model could be scaled up to cover larger areas and provide more detailed predictions for power grid damage.

REFERENCES

1. P. S. Canada, “Nova Scotia,” www.getprepared.gc.ca, Dec. 21, 2018. https://www.getprepared.gc.ca/cnt/hzd/rgnl/ns-en.aspx
2. “Some Things You Probably Don’t Know About Nova Scotia,” ThoughtCo. https://www.thoughtco.com/nova-scotia-facts-508579#:~:text=It%20is%20one%20of%20only (accessed Mar. 16, 2023).
3. “Demographics of Nova Scotia,” Wikipedia, Mar. 05, 2023. https://en.wikipedia.org/wiki/Demographics_of_Nova_Scotia#:~:text=Geographically%2C%20Nova%20Scotia%20is%20the (accessed Mar. 16, 2023).
4. “Nova Scotia Power,” Emera. https://www.emera.com/companies/regulated-electric/nova-scotia-power#:~:text=Nova%20Scotia%20Power%20is%20a (accessed Mar. 08, 2023)
5. “Who We Are,” Default. https://www.nspower.ca/about-us/who-we-are#operate (accessed Mar. 08, 2023)
6. IENG 4582: Project in Industry Final Report “IMPROVEMENT OF STORM DAMAGE PREDICTION SYSTEM” (2022) Under the supervision of Dr. Ahmed Saif, prepared by: Ashley Fulmer, Emma Leeshanok, Jacob Locke, and Robert Coakley
7. Seongmun, Oh & Kong, Junhyuk & Choi, Minhee & Jung, Jaesung. (2020). Data-Driven Prediction Method for Power Grid State Subjected to Heavy-Rain Hazards. Applied Sciences. 10. 4693. 10.3390/app10144693.
8. Kaur, Swapandeep & Gupta, Shefali & Singh, Swati. (2021). Hurricane Damage Detection using Machine Learning and Deep Learning Techniques: A Review. IOP Conference Series: Materials Science and Engineering. 1022. 012035. 10.1088/1757-899X/1022/1/012035.
9. Li, Min & Hou, Hui & Yu, Jufang & Geng, Hao & Zhu, Ling & Huang, Yong & Li, Xianqiang. (2021). Prediction of Power Outage Quantity of Distribution Network Users under Typhoon Disaster Based on Random Forest and Important Variables. Mathematical Problems in Engineering. 2021. 1-14. 10.1155/2021/6682242.
10. “Global News at 6 Halifax: Jan. 26 | Watch News Videos Online,” *Global News*. https://globalnews.ca/video/9439991/global-news-at-6-halifax-jan-26-4 (accessed Mar. 18, 2023).

11. "Data Preprocessing - an overview | ScienceDirect Topics," *www.sciencedirect.com*. https://www.sciencedirect.com/topics/engineering/data-preprocessing
12. S. Worsley, "What is Python? - The Most Versatile Programming Language," *www.datacamp.com*, Mar. 2022. https://www.datacamp.com/blog/all-about-python-the-most-versatile-programming-language
13. "15 Most Important Features of Scikit-Learn!," *Analytics Vidhya*, Jul. 15, 2021. https://www.analyticsvidhya.com/blog/2021/07/15-most-important-features-of-scikit-learn/
14. "Recurrent Neural Network - an overview | ScienceDirect Topics," *www.sciencedirect.com*. https://www.sciencedirect.com/topics/engineering/recurrent-neural-network
15. S. SHARMA, "Activation Functions in Neural Networks," *Medium*, Sep. 06, 2017. https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6
16. Forecasting mining capital cost for open-pit mining projects based on artificial neural network approach - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Pseudo-code-of-the-back-propagation-algorithm-in-training-ANN_fig1_335375894 [accessed 26 Mar, 2023]
17. D. S. K, "A Gentle Introduction To Math Behind Neural Networks," Medium, Oct. 30, 2020. https://towardsdatascience.com/introduction-to-math-behind-neural-networks-e8b60dbbdeba
18. S. E R, "Random Forest | Introduction to Random Forest Algorithm," Analytics Vidhya, Jun. 17, 2021. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest
19. "Fundamentals of Machine Learning [1 ed.] 2019945424, 9780198828044," *dokumen.pub*. https://dokumen.pub/qdownload/fundamentals-of-machine-learning-1nbsped-2019945424-9780198828044.html (accessed Apr. 02, 2023)