HOUSEHOLD SPACE HEATING DEMAND MODELLING USING SIMPLIFIED BLACK-BOX MODELS

By

Ankit Dhall (500942040),

A Major Research Project Report

Presented to Ryerson University

in partial fulfilment towards the requirements for the degree of

Master of Science (M. Sc.)

In

Data Science and Analytics

2019-2020

Toronto, Ontario, Canada, 2020

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PROJECT (MRP)

I hereby declare that I am the sole author of this Major Research Paper. This is a true copy of the MRP, including any required final revisions.

I authorize Ryerson University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

Ankit Dhall

HOUSEHOLD SPACE HEATING DEMAND MODELLING USING SIMPLIFIED BLACK-BOX MODELS

Ankit Dhall

Master of Science 2020

Data Science and Analytics

Ryerson University

ABSTRACT

This research aims at applying a novel idea of utilizing an ANN (Artificial Neural Network) black-box model to predict the space heating demand of households in Toronto, Ontario, Canada. The data used is gathered as a part of the Ecobee Donate Your Data program. First, an exploratory analysis is conducted using descriptive analytics and data visualization to try and find patterns, or relationships that could help give insight into the data. Further, multiple approaches and techniques such as data aggregation and inclusion of time-lag information are applied to model and predict the space-heating demands of any house using basic, easy to record features only. In addition, experiments are conducted to gauge the practical viability of the black-box model developed. This research was conducted as a continuation of an ongoing study at the Ryerson Centre for Sustainable Energy Systems (CSES). Despite a few issues faced with the data being modelled, the space-heating demand was successfully predicted using black-box ANN models using simple, easy to observe features and including time-lag information for the past half-hour. In addition, the model was able to portray a practical learning capability as additional data was added. For future studies to predict space-heating using the given data, it is recommended to apply data aggregation techniques and additional feature engineering, as well as filtering out only relevant data using domain knowledge to be able to achieve better prediction results.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude towards **Dr. Alan Fung** for providing me with the opportunity to work and collaborate with his team at the Centre for Sustainable Energy Systems (CSES) at Ryerson University. The experience of working on a real-world problem with the aim of tackling a pressing issue like climate change has helped me gain perspective towards the urgent need of such research.

I am grateful to **Dr. Fung** and **Gülsün Demirezen** for their constant support and assistance towards the progress and completion of this project. As my supervisor and mentor, Dr. Fung and Gülsün have been monumental throughout the term to guide and direct my research and provide valuable feedback for me to improve on.

I am walking away with an invaluable experience and motivation to use my knowledge and skills towards the betterment of the world we live in.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

A. Background

Household energy consumption constitutes a major amount of the total energy consumed by buildings in Canada. Being able to model and predict the space heating needs of a household can help reduce energy consumption, and in turn, reduce the total greenhouse gas emissions from households. This can be done by using such information to increase efficiency of heating systems used.

The space heating demand can be understood as the heating needs associated with any house/building. In Canada, the major amount of the energy demand emerges from the heating systems and appliances due to the colder climates of the region.

With climate change being one of the biggest issues at the hand of the developing world, it is now more important than ever, to understand and deal with the effects of inefficient heating systems. Being able to model and predict the thermal energy demand of a household would greatly help us in reducing the total energy consumed and in turn reduce the production of harmful greenhouse gases.

While there have been several studies performed in this field, there is still little in terms of a finding a method to estimate the thermal energy demands of any household without precise and detailed thermal experiments and data.

B. Research Question

Is it possible to build a generalized black-box model to estimate and predict the space-heating needs of a household?

The aim of the research project would be to work towards building a generalized black-box model that can model and predict the space-heating demand for a household between a given period of time. The model should be able to use basic features like indoor temperature, outdoor temperature, outdoor humidity, and other such easy to obtain publicly available information to make these predictions.

For this study, the data used would be collected from a smart Wi-Fi enabled thermostat system that logs and uploads this data to a cloud server.

Several data analysis and predictive modelling techniques using neural networks would be used to help achieve the desired results. Extra experimentation and revisions can be carried out to try and achieve the best possible results

C. The Dataset

Ecobee, a Canadian home automation company that produces smart thermostats, sensors for temperature & occupancy detection, smart light switches, smart cameras etc. are one of the players in the market for cutting edge smart-home automation services.

Their advanced smart thermostats in tandem with the smart sensors are capable of detecting occupants in a room, set smart comfort schedules, adjust heating/cooling sources, all according to the user's preferences in order to boost efficiency and save energy.

As a part of the Ecobee Donate Your Data program, users can consent to sharing their data with the company for additional research and product improvement purposes. This data, spanning multiple parameters and features is then recorded, anonymized to take care of user privacy, made open-source, and shared by Ecobee with researchers across the globe.

Part of this data is user-reported (metadata including home and occupant characteristics) and in part collected automatically by the Ecobee thermostats (reported in 5-minute intervals). A sample file of these datasets can be found using the link provided in Appendix – B.

The very same data, collected from a few homes in Toronto, Ontario, Canada in the year 2019, would be used for the research project to try and develop a generalized black-box model to predict the space-heating demand for Canadian households.

For the predictive analytics portion of the research, there would be 4 independent variables and 1 dependent (target) variable as listed below:

Independent Variables:

1. *T_ctrl – Averaged Indoor Temperature*
2. *T_stp_heat – Indoor Temperature Setpoint for Heating*
3. *T_out – Outdoor Temperature from Nearest Weather Station*
4. *RH_out – Outdoor Relative Humidity from Nearest Weather Station*

Dependent Variable:

1. *auxHeat1* - Heat source runtime in seconds for the given 5-minute interval. The value of this variable ranges from 0 to 300 seconds (0 minutes – 5 minutes) in 15 second intervals.

2. LITERATURE REVIEW

The Effects of the Socioeconomic Factors on the Household Appliance, Lighting, and Space Cooling Electricity Consumption [1]

Introduction & Aim

The modelling of residential energy consumption helps in understanding and improving the end-use energy efficiency. This in turn helps to reduce energy consumption and hence reduce pollutant emissions associated with it.

Traditional white-box methods like conditional demand analysis (CDA) and the engineering method (EM) have limitations in terms of needing very large amount of data and including external factors like consumer behaviour and other socioeconomic factors associated with prediction of energy consumption respectively.

The aim if the research is to use neural network-based approaches to model and predict the residential energy consumption with inclusion of socioeconomic factors and demonstrates its use for the Canadian residential sector.

Methodology Used

The end-sue energy consumption model consists of 3 sub-models for:

1. Domestic hot-water heating energy consumption
2. Space heating energy consumption
3. Space cooling (ALC – Appliances, Lighting, and Space Cooling) energy consumption

The data used for the ALC NN model consisted of detailed usage information for many kinds of home appliances including the number of hours of air-conditioner being used.

The NN model predictions were extrapolated to estimate the entire Canadian housing energy usage. The results were compared to those of the Engineering model and were found to be about 2% lesser.

Results & Conclusion

The ALC NN model predictions indicate that the ALC energy consumption is lesser in households with electric and wood powered heating systems as compared to that of households using natural gas, oil, or propane for space heating. This can be associated to the fact that houses with natural gas, oil, or propane used for heating have furnace fans or boiler pumps that increase energy consumption.

It can thus be concluded that the neural network model can be used to factor in socioeconomic factors along with other factors to model household energy consumption. It is also successful in capturing the increase of energy consumption by furnace fans and boiler pumps.

Predicting Indoor Temperature from Smart Thermostat and Weather Forecast Data [2]

Introduction & Aim

The indoor temperature of a household hugely affects the energy demand for any internal heating and cooling system. Smart thermostats try to predict indoor temperatures based on different types of strategies that may include many variables such as characteristics of a house, the amount of time and energy it takes to heat/cool the house, occupant behavioural patterns, and outside weather information.

This study applies neural networks to use information about future outdoor weather predictions in tandem with information gathered by smart thermostats and predict the future indoor temperature of a house. This could help in a smooth transition of temperature by optimizing the working of heating and cooling systems leading to lesser energy usage. The Ecobee thermostat data was used to model and compare results.

Methodology Used

The experiment made use of two major machine learning approaches – a generalized regression neural network (GRNN) and an artificial neural network (ANN). Both these approaches were chosen as they did not require any prior information about the characteristics of the house and did not need to know the functional form of the data.

The basic dataset variables used to predict indoor temperature were - cooling setpoint, heating setpoint, indoor humidity, fan runtime, and outdoor temperature. In addition to these, additional features like solar radiation, cumulative fan runtime, and the average indoor temperature from the previous time step were also used. Additional predictors were used where they were available in a house.

The input was scaled normalized where required. Training was carried out each using 1,000, 5,000, and 10,000 training samples with a 80/20 training/validation split. The mean-squared error (MSE) was used as the performance metric for each of the models.

Results & Conclusion

A total of 16 houses were used to test the model. The average MSE of the houses were 0.98 and 5.5, with a maximum deviation of 7.5 and 13.97 from the actual temperatures for the GRNN and ANN algorithms, respectively. The GRNN seemed to have a better fit than that of the ANN model.

In conclusion, the indoor temperatures of households were predicted despite the randomness of user behaviour, changes in outdoor conditions, and difficulty in acquiring real-time high-resolution weather data.

The GRNN results proved that adding solar radiation data along with outdoor and indoor temperature from the previous time step helped achieve significantly better results. This model can be used to implement strategies to help buildings adapt better to changing climatic conditions.

Development and Optimization of Artificial Neural Network Algorithms for the Prediction of Building Specific Local Temperature for HVAC Control [3]

Introduction & Aim

This research was carried out as a part of the development process for a smart duel-fuel switching HVAC system to enable efficient, optimized, and cost-effective switching between a gas and electric heat pump heating system.

The exact outdoor temperature at the site of a house is found to play the most important part in the working of HVAC systems as well as in finding the thermal demand of the house. This information is usually quite different from what is recorded at the nearest weather station. While the exact temperature readings can be gathered using advanced IoT based sensors at the site of the house, the infrastructure and subsequent maintenance can prove to be quite expensive.

This research aims to find the outdoor temperature at a site using a data-driven black-box model using artificial neural networks, which can be used as an alternative to sensors in HVAC systems requiring accurate surrounding temperature information.

Methodology Used

Using white-box models in to model this non-linear time-variant relationship proves to be expensive in terms of time and computation. Hence, this study builds and optimizes artificial neural networks to model and predict the exact outdoor temperatures.

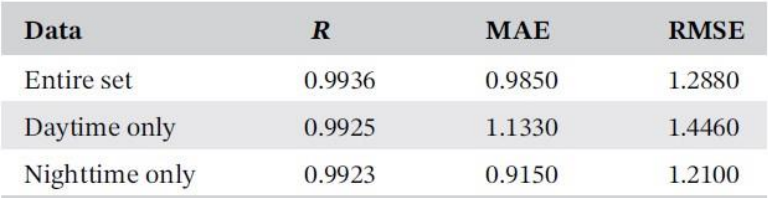
A fully automated data collection system was set up to carefully log readings from multiple sensors installed in a net-zero energy house. This data, in addition to additional weather station data was used in the modelling process. The data was collected in 3 distinct forms – Daytime, Night-time, and a collection of both daytime and night-time datapoints.

Multiple combinations of learning parameters were used to optimize the model. A detailed analysis was for the hidden layer parameters in terms of activation function used, MAE obtained, RMSE, run-time, and the R score.

Results & Conclusion

It was found that the highest R score was obtained using the dataset including both daytime and night-time instances which could be attributed to the fact that there are more datapoints in general. In addition, the MAE and RMSE scores were lowest in the night-time dataset. This result could be attributed to the additional GHI (global horizontal irradiation) data added to the daytime dataset which could add noise. The ANN could successfully be used to predict the exact ambient temperatures at the location of the house.

Table 1: Results Achieved for Research Paper [3]



Data	*R*	MAE	RMSE
Entire set	0.9936	0.9850	1.2880
Daytime only	0.9925	1.1330	1.4460
Nighttime only	0.9923	0.9150	1.2100

The use of data driven ANN models can be used to replace more expensive sensor systems. These systems require minimal upkeep costs and is a promising alternative for home automation systems that require this kind of data.

Moreover, this study proves that the outdoor ambient temperatures cause a huge impact in the operation of HVAC systems and can be crucial in their efficient working. This would then naturally cause a huge impact in the prediction of the thermal demand of any given house/building.

Forecasting Energy Use in Buildings Using Artificial Neural Networks: A Review [4]

Introduction

The energy demand has been increasing at a rapid rate over the past century. This has inevitable resulted in the massive increase in greenhouse gas emissions. Forecasting the energy consumption can be of utmost importance in this situation as it can help with planning, maintaining, optimizing, and conserving the amount of energy being used which in turn can help with conserving the environment. One of the most popular approaches in the recent times to help with this forecasting is the use of neural networks to build data driven models in contrast to more complicated physics-based models.

This paper offers a concise review of some of the popular approaches and studies carried out since the year 2000.

Q1. What are the approaches that have proved to provide good results in terms of energy forecasting?
Ans. The use of machine learning approaches has been popularized over the past 2 decades, however, using neural network-based approaches have only come up over the past decade or so. Both the above approaches have been proven to work better than simple physics based white-box approaches as they can capture non-linear relationships that come up in weather, indoor conditions, and occupancy data. Modelling this information successfully has helped in the high performance of thee AI based approaches.

To further break down the split of models being used, nearly 84% of ANN based models have been black-box based models. Only about 12% have been ensemble-based methods with 75% of these being homogenous models (ANN + ANN) and only 25% being heterogeneous models (ANN + SVM for example). Finally, only 4% of the studies have resorted to using grey-box models.

More specialized neural networks like recurrent neural networks which are more suited towards forecasting time-series data have only been used in about 14% of the studies conducted.

Q2. What is the difference between prediction and forecasting of thermal energy?
Ans. The term 'forecasting' can be understood as determining a value, or thermal energy consumption in this case, for any future time or period. This term has been quite often been used interchangeable with the term 'prediction'. The term 'prediction' can basically be defined as the determining the value of a dependant variable given the independent variable information of the current and/or previous time stamps.
To be able to understand this difference and make an informed decision of the modelling process is a crucial part of the research.

Q3. What kind of subjects has the research been applied to?
Ans. Amongst the papers that have been studied as a part of this review, it is noted that the forecasting for energy consumption has majorly been performed on entire buildings (81%). This may be associated to the easier access of data for an entire building. About 13% of the studies have been performed on the territory level and only about 5% of the studies have been performed on the component level (e.g., chiller, fan).

Within entire buildings, most of the analysis (83%) was concerned with commercial and institutional buildings while only about 17% dealt with residential buildings. However, residential buildings make up a large proportion of the overall buildings and carry huge potential for energy savings and hence, further analysis must be carried out on residential buildings.

Q4. What are the performance metrics popularly used amongst the studies?
Ans. It was found that MAPE (Mean Absolute Percent Error) was used in about 38% of the studies conducted. CV-RMSE and R squared performance metrics were used in about 20% and 17% of the studies, respectively.

Q5. What are some of the disadvantages noticed in using ANNs to model, predict, and forecast energy consumption?
Ans. As with all data driven models, ANNs do not seem to perform well with data outside their training range. For example, a model trained using a summer dataset may not perform well on data concerning the winter season. However, repeated retraining techniques can be used to counter this problem.

Overfitting is another problem that ANNs may be susceptible to. This may be a bigger problem in models that are built to forecast over long range horizons. Many methods exist to help counter the problem of overfitting and it is crucial to make sure that any model developed is not overfitting on the training data.

Another issue with ANN models is that they are black-box models. This means that there is no way of interpreting their internal working and determining what decisions the neural network is taking to produce an output. One solution to this kind of problem is the development of grey-box models that involve the use of external physics based equations that can be understood and made sense of which may help in eliminating some of the 'mystery' from developed models.

3. EXPLORATORY DATA ANALYSIS

D. Overview of Data Variables

Metadata Fields:

The metadata fields are contained in a separate file and contain data variables for home information like Country, Province/State, City, Floor Area in sq. ft., Type of House, Number of Floors, Number of Occupants, Type of Heating and Cooling Sources, any additional equipment, sensors. Sample of the metadata file can be found in Appendix – B.

While this data is not directly used in the modelling process, it helps us select homes in order to proceed with the analysis.

Thermostat Data Fields:

auxHeat1,2,3 *
Runtime (seconds) for any heat source other than a heat pump (where 1,2,3 are the stages of the equipment). All the houses used in the modelling process have only 1 stage of heating.

compCool1,2,3
Runtime (seconds) for any cooling (where 1,2,3 are the stages of the equipment).

compHeat1,2,3
Runtime (seconds) for heat-pumps used in heating.

DateTime
Date and time that the reading was taken.

Event
Anything that modifies the schedule (e.g., temperature hold, demand response event, vacation, SmartRecovery feature, eco+ TOU).

Fan
Runtime (seconds) for fan.

Humidity
Indoor humidity (in RH%).

HumidityExpectedHigh
Setpoint (for users who have a Humidifier) (in RH%).

HumidityExpectedLow
Setpoint (for users who have a Humidifier) (in RH%).

HvacMode
Indicates whether the system is off, heating, cooling, or auto.

Remote_Sensor_1,2,3_Motion
Detects motion (binary) at that date/ time at the remote sensor (where 1,2,3 denotes different sensors).

Remote_Sensor_1,2,3_Temperature
Indoor temperature measurement at the remote sensor (where 1,2,3 denotes different sensors).

Schedule
Fields include things like Vacation, Sleep, Away, Nap, etc. which are user-defined descriptors for desired set points against activity/behaviour.

T_Ctrl *
Indoor temperature used by the thermostat to compare against setpoints. It represents a combination

of temperatures across the home. T_ctrl will be equal to Thermostat Temperature if no additional sensors are installed in the house.

Thermostat_Motion
Detects motion (binary) at that date/time.

Thermostat_Temperature
Indoor measurement at the thermostat

T_out *
Outdoor temperate for nearest weather station.

RH_out *
Outdoor relative humidity for nearest weather station.

T_stp_cool
Indoor cool setpoint.

T_stp_heat *
Indoor heat setpoint

*** - Represents features used in the predictive modelling process.**

E. Data Visualization

This section of the report presents some information about heat-usage information and patterns that may exist in the dataset. This information once brought forward can help us formulate and guide experiments to make relevant progress.

While the data visualization is carried out on the entire dataset, only a subset of the dataset for each house will be used for predictive modelling purposes. These subsets are selected using various statistical methods used to help understand the dataset in a better manner.

The following figures show the maximum heat source run-time (300 seconds) by month, day of the week, and hour of the day.

Figures 1-6 show a pattern in heat-source runtimes when plotted against month, day of week, and hour of day. These patterns may be controlled by different reasons such as change in heating setpoint temperatures, outdoor temperatures, and outdoor humidity.

The aim is for the model to capture such patters with the use of such basic features to predict the heat source run-time.

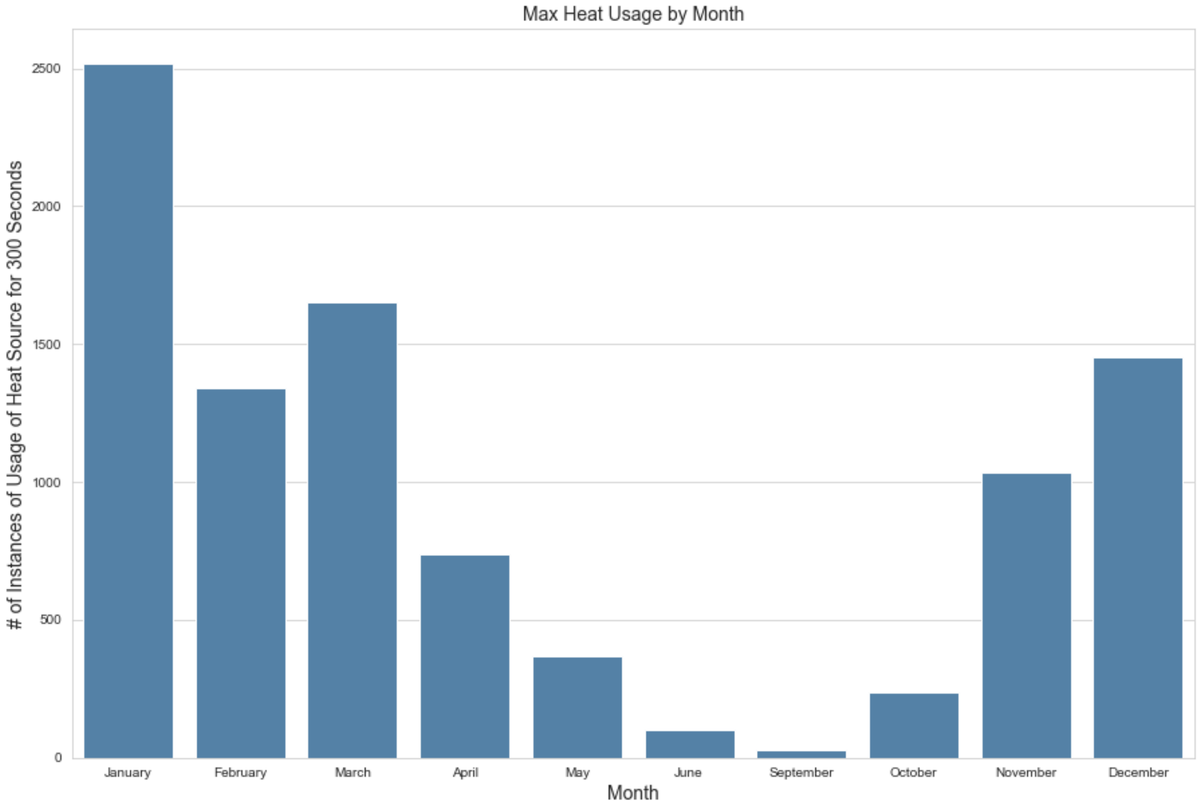


Figure 1: Max Heat Usage by Month

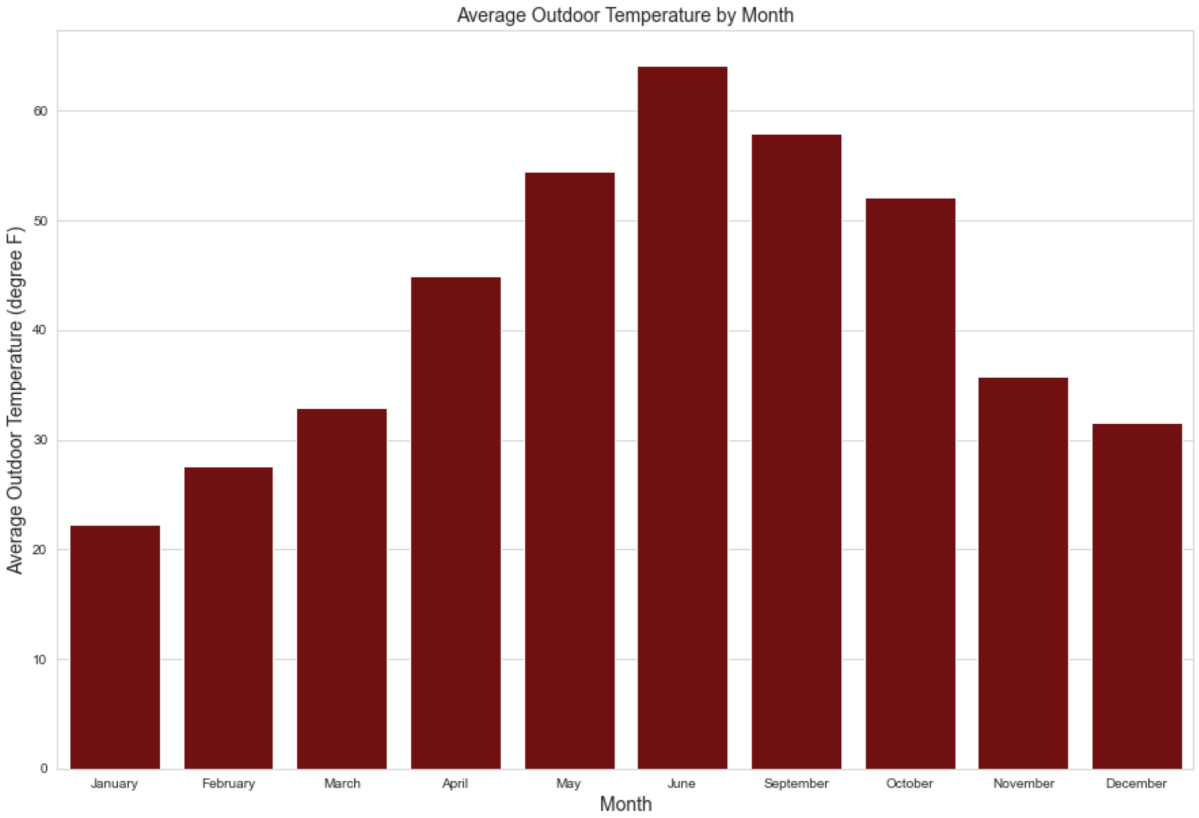


Figure 2: Average Outdoor Temperature by Month

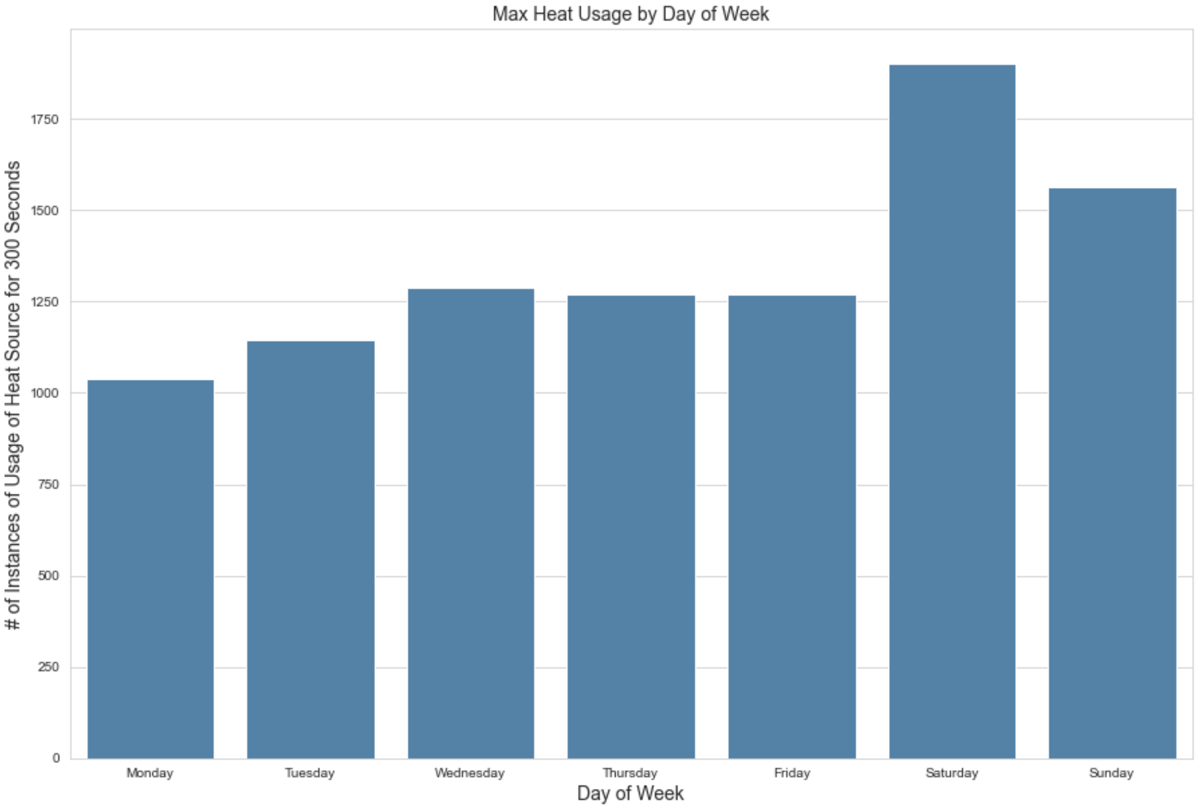


Figure 3: Max Heat Usage by Day of Week

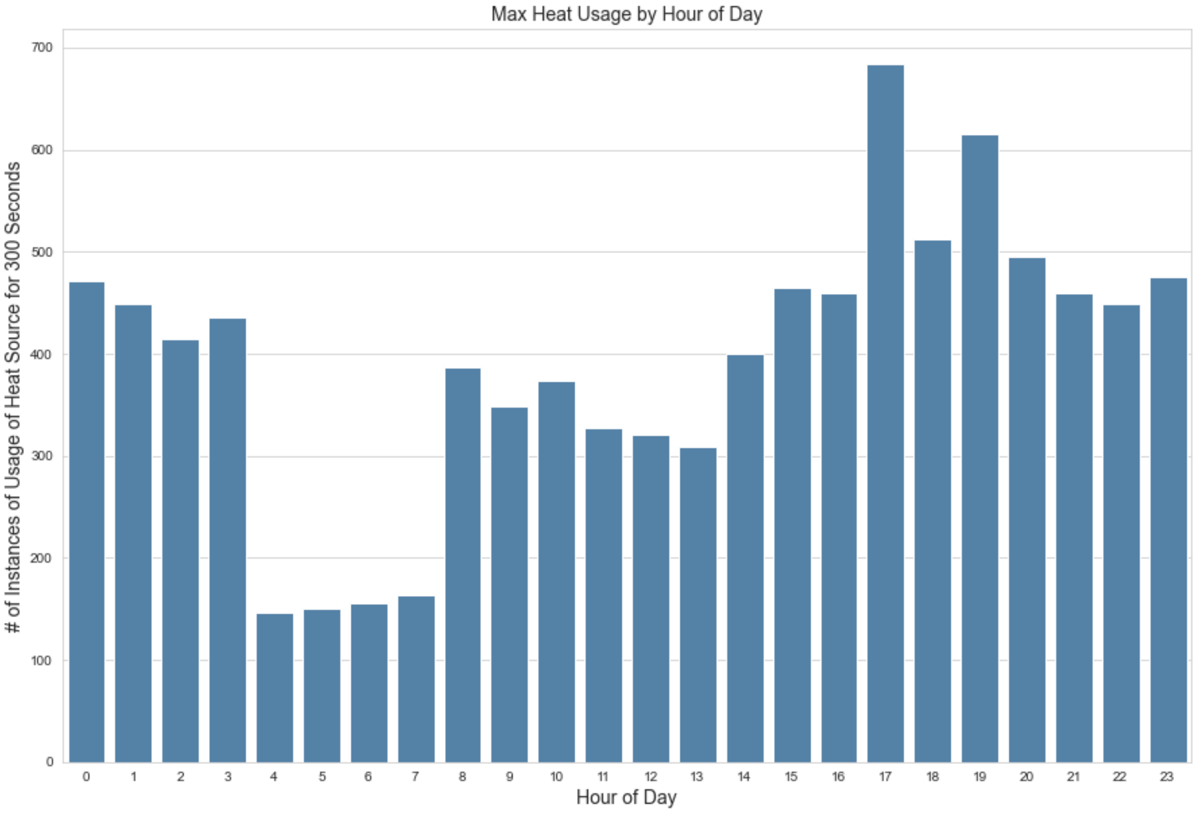


Figure 4: Max Heat Usage by Hour of Day

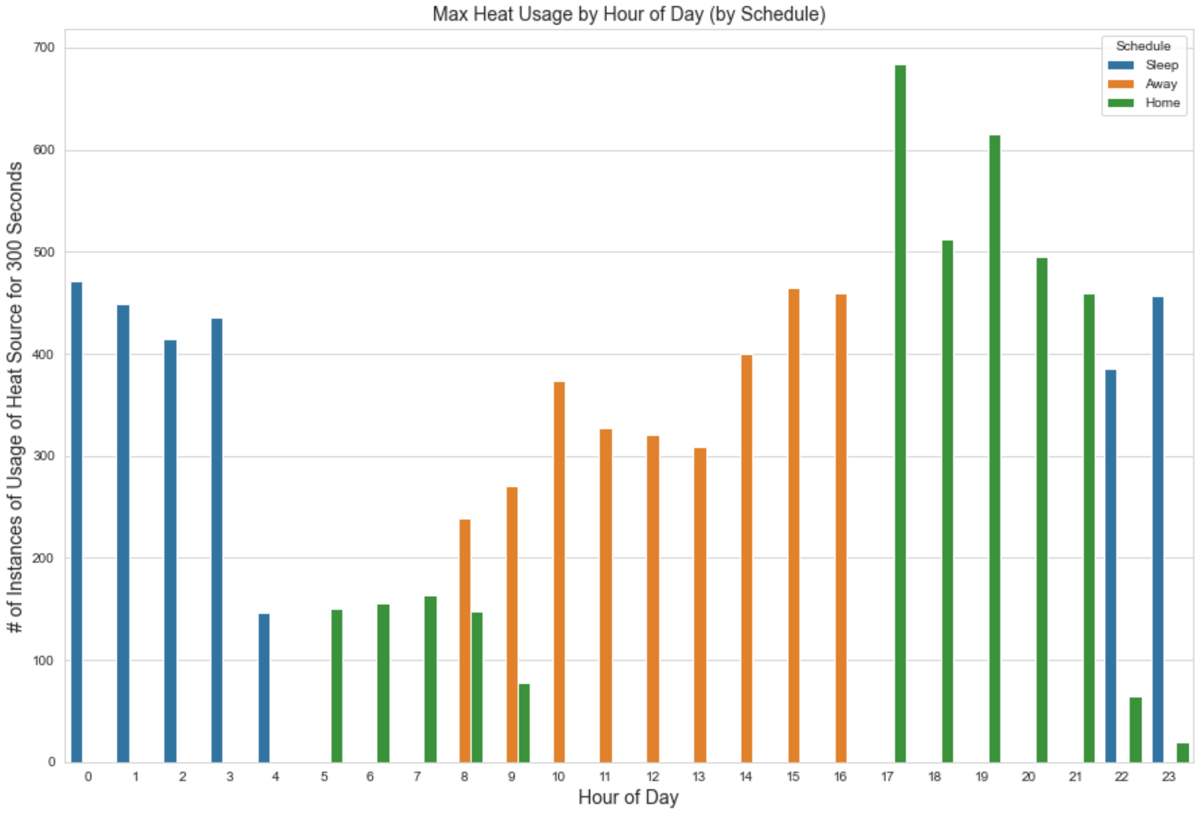


Figure 5: Max Heat Usage by Hour of Day (Categorized by Schedule)

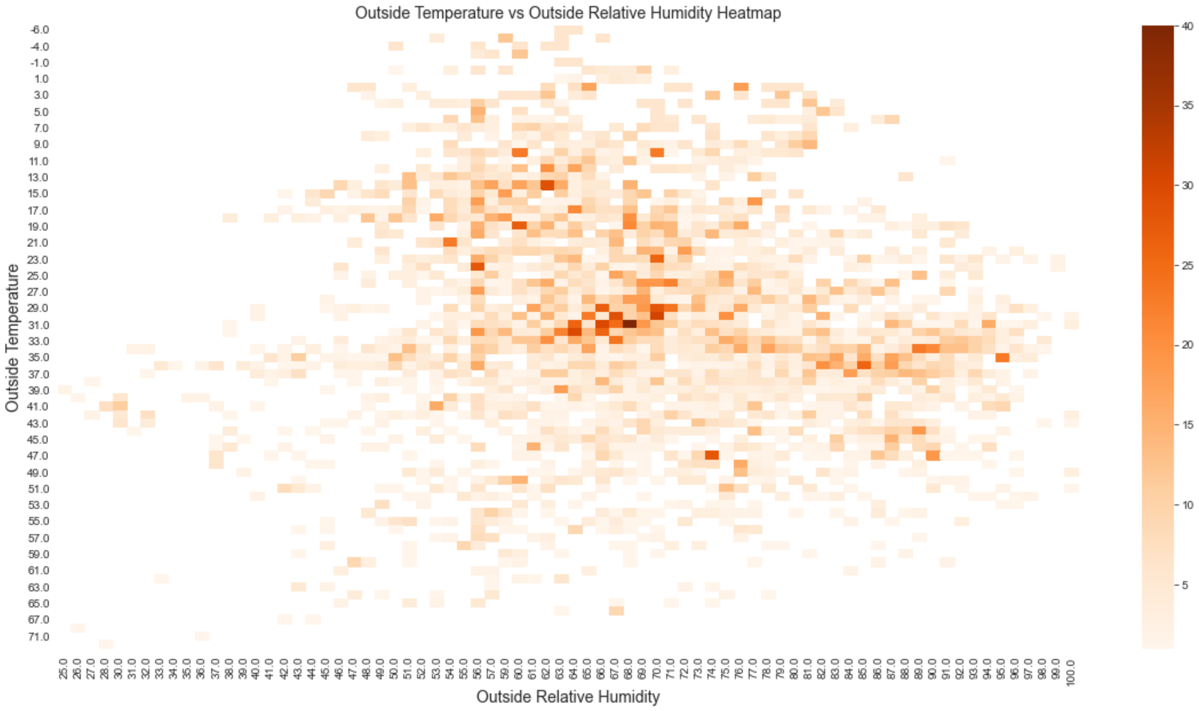


Figure 6: Max Heat Usage Heatmap by Outside Temperature and Outside Humidity

F. Data Cleaning & Selecting Relevant Features:

Since the aim of the project is mainly to be able to predict thermal demand of a given household, we would be focusing mainly on the 'Heating' aspect of the HVAC system. This pretext enables us to select only certain useful variables from the dataset based on intuition and domain knowledge.

Further, we would like to avoid using certain features from the dataset to be able to work on a generalized model. This includes eliminating features that may vary from one household to another.

In addition, certain feature columns in the dataset include all null values and hence have no data to contribute towards the modelling process.

Acting on such features, we would be able to remove a few features from the dataset that would help us build a more concise, efficient, and meaningful model.

G. Feature Engineering

Certain feature engineering was required to use information from some features in the modelling process. A basic summary of the feature engineering performed has been given below:

- **Extracting Month, Day of Week, and Hour of Day from 'DateTime' Feature**

  Since the 'DateTime' feature in the dataset is not of any use in the modelling process in the raw form, we must extract information from the 'DateTime' to create new features like 'Month', 'DayOfWeek', and 'HourOfDay'.

- **Cyclical Feature Engineering**

  Features like hours of the day, months in a year, and days of the week are all examples of cyclical features. This means that, for hours of the day, we must be able to represent the feature in a way that hour 23 and hour 0 are close to each other and not far away. This kind of representation is not possible with a traditional encoding approach and such features must be properly transformed in order for the model to make the right sense.

  To tackle this, we can map and represent such features as co-ordinates of a circle such that the largest and smallest values appear right next to each other [8]

  Keeping this in mind, dataset features like 'Month', 'DayOfWeek', and 'HourOfDay' are converted to their respective sine and cosine projections and stored as a part of the dataset.

Once the above steps are performed, we are left with a cleaner, concise representation of the data at hand. This data is then used for Exploratory Data Analysis.

H. Exploratory Data Analysis

Univariate Analysis:

- **Cyclical Variables:**

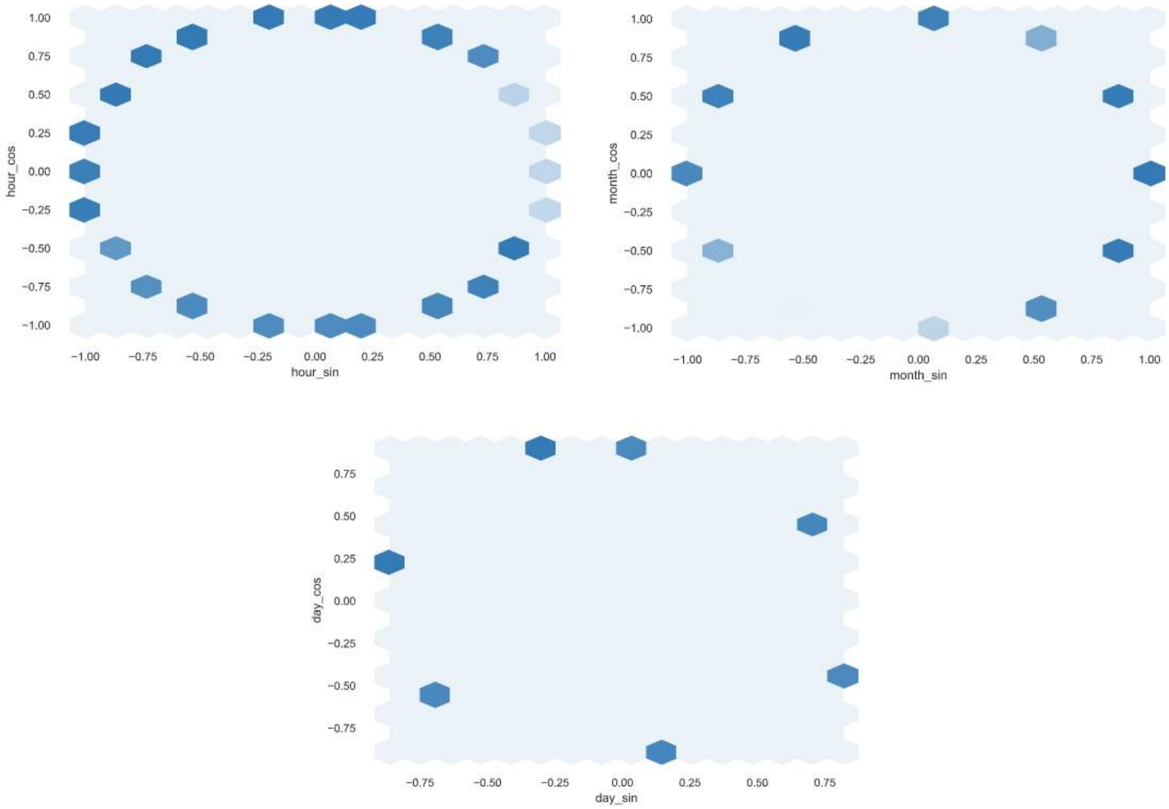
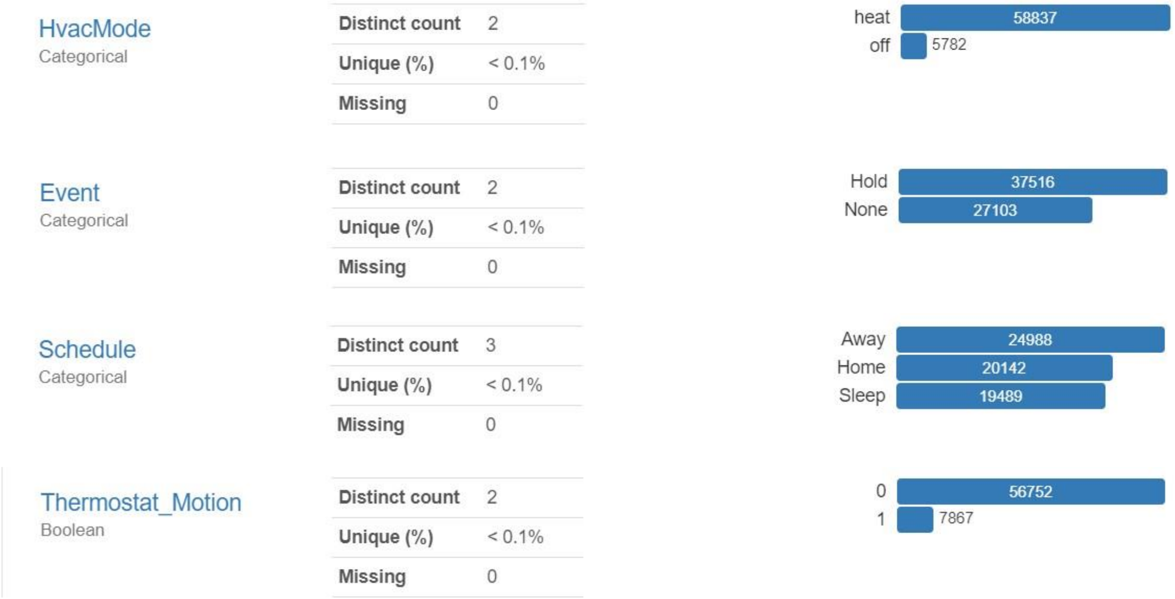


Figure 7: Projection Features for (a) Hour of the Day, (b) Month, and (c) Day of the Week

- **Categorical Variables:**



HvacMode
Categorical

Distinct count	2
Unique (%)	< 0.1%
Missing	0

heat	58837
off	5782

Event
Categorical

Distinct count	2
Unique (%)	< 0.1%
Missing	0

Hold	37516
None	27103

Schedule
Categorical

Distinct count	3
Unique (%)	< 0.1%
Missing	0

Away	24988
Home	20142
Sleep	19489

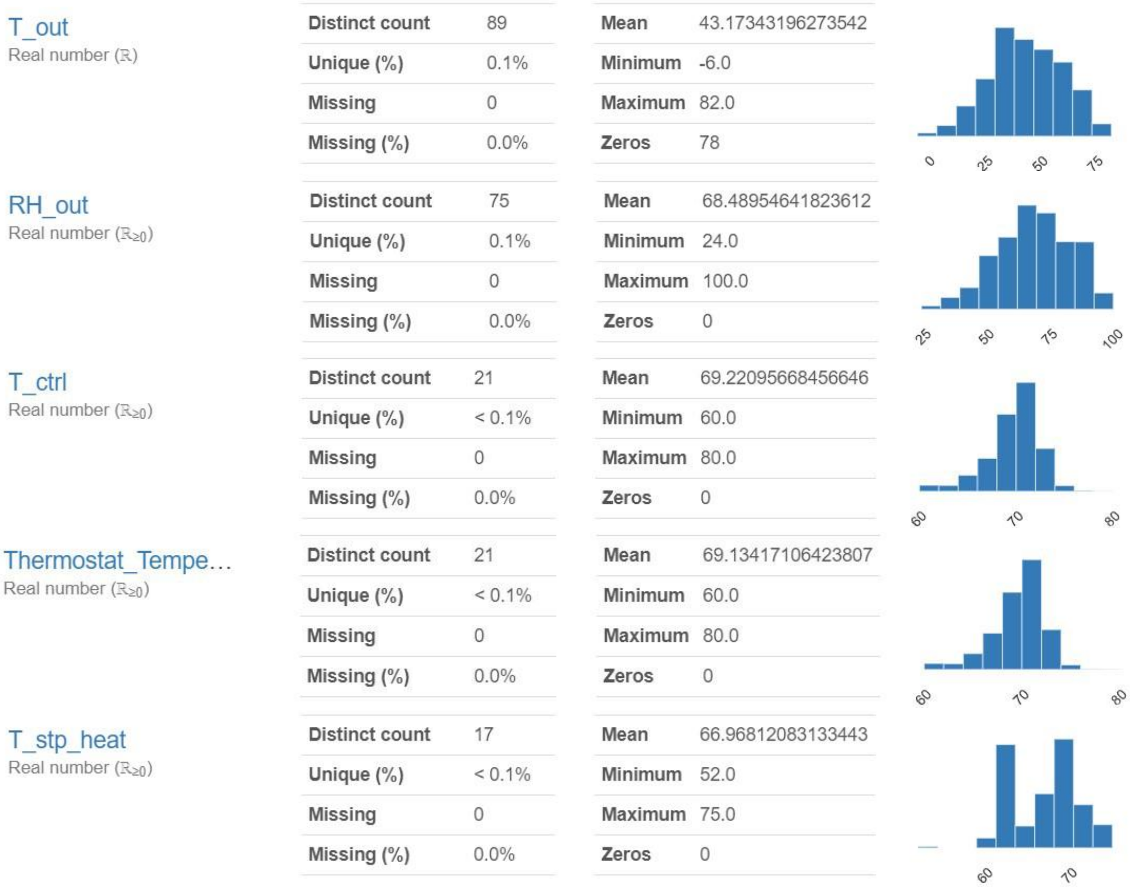
Thermostat_Motion
Boolean

Distinct count	2
Unique (%)	< 0.1%
Missing	0

0	56752
1	7867

Figure 8: Univariate Analysis for Categorical Variables

- **Numerical Variables:**



T_out
Real number (ℝ)

Distinct count	89	Mean	43.17343196273542
Unique (%)	0.1%	Minimum	-6.0
Missing	0	Maximum	82.0
Missing (%)	0.0%	Zeros	78

RH_out
Real number (ℝ≥0)

Distinct count	75	Mean	68.48954641823612
Unique (%)	0.1%	Minimum	24.0
Missing	0	Maximum	100.0
Missing (%)	0.0%	Zeros	0

T_ctrl
Real number (ℝ≥0)

Distinct count	21	Mean	69.22095668456646
Unique (%)	< 0.1%	Minimum	60.0
Missing	0	Maximum	80.0
Missing (%)	0.0%	Zeros	0

Thermostat_Tempe…
Real number (ℝ≥0)

Distinct count	21	Mean	69.13417106423807
Unique (%)	< 0.1%	Minimum	60.0
Missing	0	Maximum	80.0
Missing (%)	0.0%	Zeros	0

T_stp_heat
Real number (ℝ≥0)

Distinct count	17	Mean	66.96812083133443
Unique (%)	< 0.1%	Minimum	52.0
Missing	0	Maximum	75.0
Missing (%)	0.0%	Zeros	0

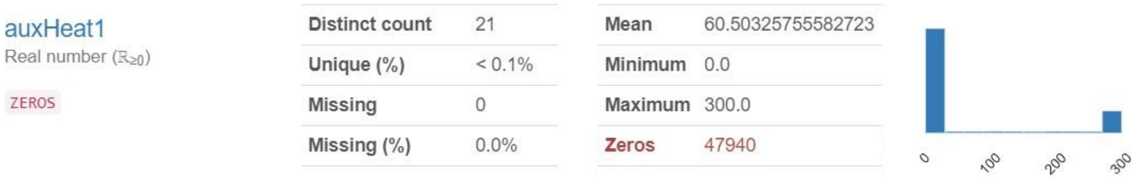
Figure 9: Univariate Analysis for Numerical Variables

It is interesting to note that in the univariate analysis shown above, the mean 'T_ctrl' and 'Thermostat_Temperature' appear to be higher than 'T_stp_heat' (Heating Setpoint). This is as a result of the fact that the analysis includes data for summer months as well as the winter months.

Zero values included in T_out (Outside Temperature) do not present a problem as these indicate true zero values indicating that the outside temperature is 0 °F

Analyzing the quantile statistics of numerical features, it can be determined that there are no outliers in the dataset that may affect the performance of the model.

- **Target Variable (Ordinal):**



auxHeat1

Real number ($\mathbb{R}_{\geq 0}$)

ZEROS

Distinct count	21	Mean	60.50325755582723
Unique (%)	< 0.1%	Minimum	0.0
Missing	0	Maximum	300.0
Missing (%)	0.0%	Zeros	47940

Figure 10: Univariate Analysis for Target Variable

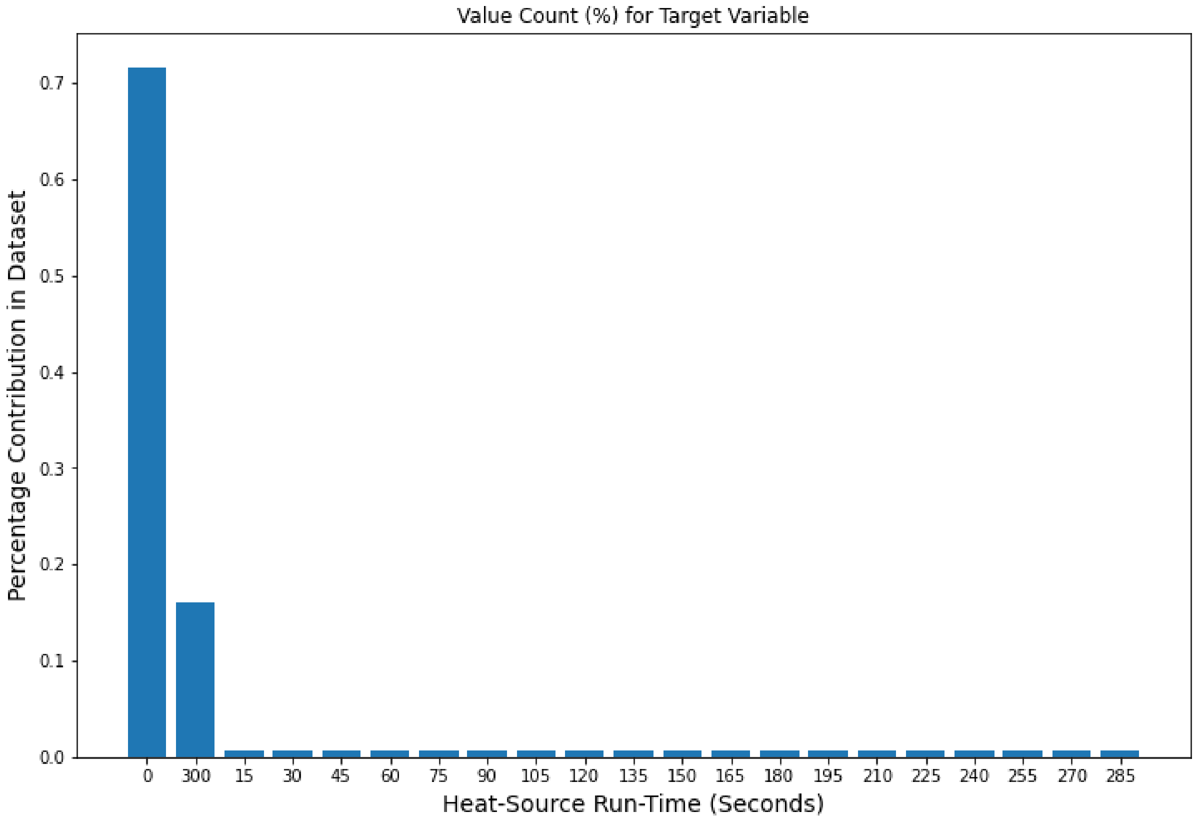


Figure 11: Value Count (%) For Target Variable

Bivariate Analysis:

- **Correlation:**

Spearman’s ranked correlation coefficient is a measure of monotonic correlation between any two variables. It is more effective than detecting non-linear correlation as compared to Pearson’s correlation and also works better for ordinal variables.

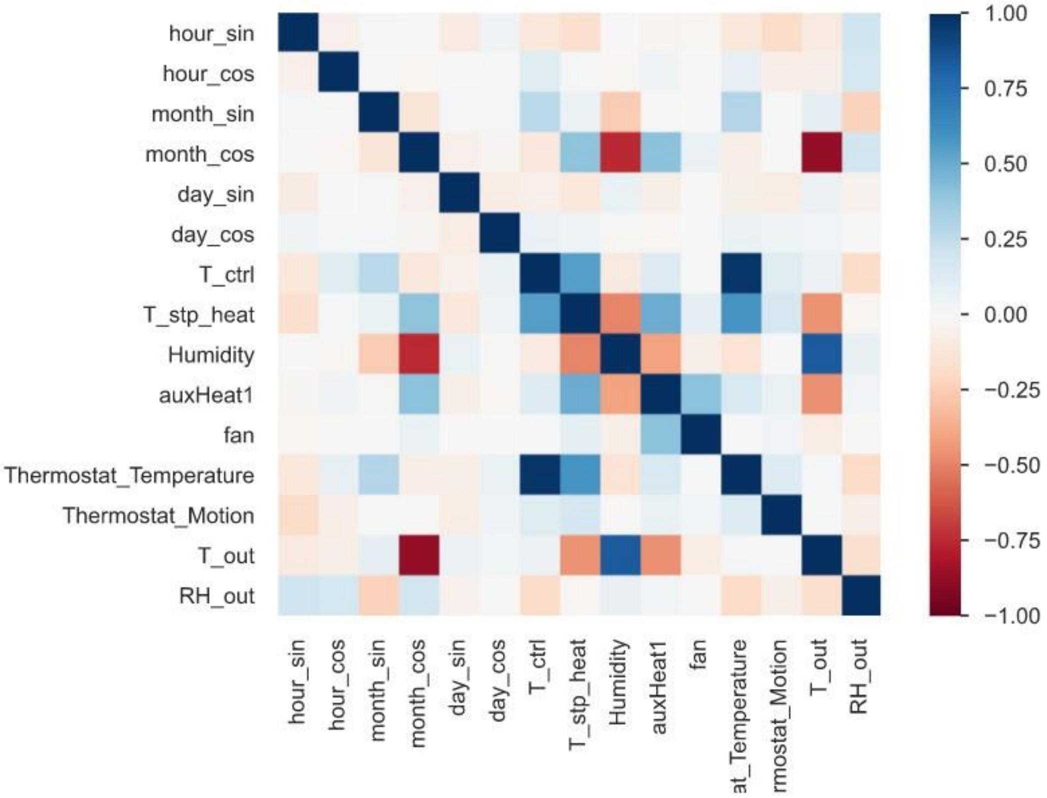


Figure 12: Spearman’s Ranked Correlation

I. *Feature Selection using Embedded Methods:*

Feature selection can be performed by certain algorithms that have their own built-in feature selection methods that indicate feature importance. These algorithms can help us decide which features play a key role in making decisions and can help us narrow down the number of features to be used in the predictive modelling process.

Applying this method using a Decision Tree Classifier, the following is the relative feature importance graph obtained:

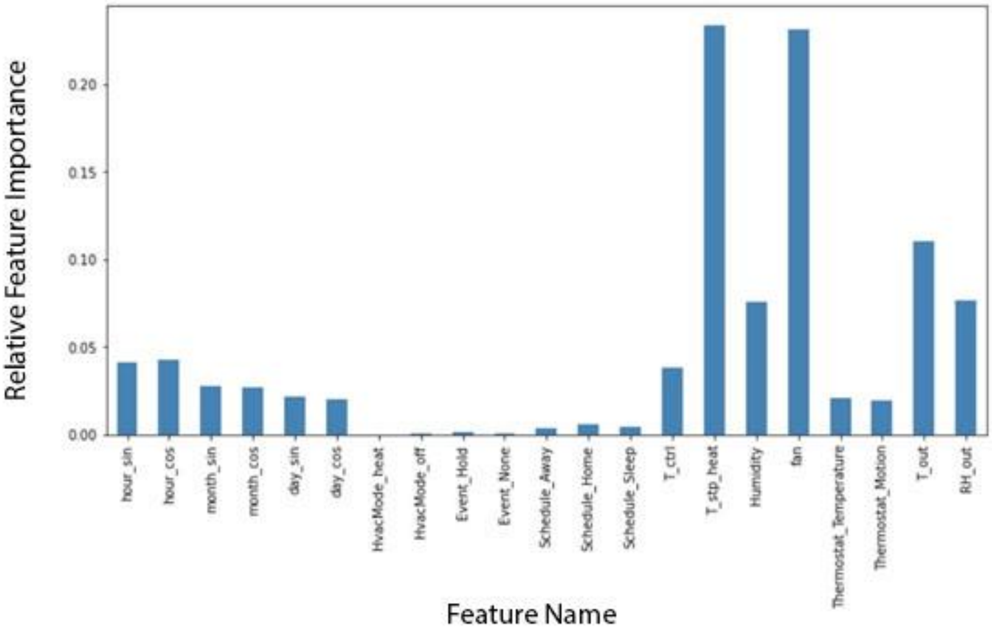


Figure 13: Relative Feature Importance Graph

Figure 12 shows that the most relevant/important features include Hour of the Day, Indoor Control Temperature, Indoor Setpoint Temperature for Heating, Internal Humidity, Fan Runtime, Outdoor Temperature, and Outdoor Relative Humidity. This gives us a good starting point to begin the predictive modelling process.

4. METHODOLOGY AND EXPERIMENTS

J. Aim of Study

The aim of this study is to develop a data-driven model that is able to predict the run-time of the heat source, given very basic features that are easily observable in any given household. Multiple methods and techniques were used to model the data appropriately. The approach that was able to provide us with the best possible results was then used for further experimentation and analysis.

K. Data Sorting & Selection of Subject Houses:

For the purpose of the experimentation process, several parameters were kept in mind while selecting subject houses from the dataset. The basic criteria kept in mind was as follows:

- Type of House – Detached
- Heat-Source Type – Natural Gas Only
- Location of House – Toronto, Ontario, CA
- No additional motion sensors or humidity sensors installed in the house

L. Response (Dependent) and Independent Variable(s):

In order to be able to build a generalized data-driven model, only those features were used from the dataset that can be recorded easily given any household. These would include features such as the control temperature of the house, heating setpoint for the house, current outside temperature, and current outside relative humidity.

Certain extra features like time of day, day of week, month etc., as well as extra weather information would be added towards later stages of the experimentation process in order to see if they improve the results of our prediction model.

Hence, the base model would involve 4 independent variables and 1 dependent (response) variable.

- **Independent Variables:**
    1. **T_ctrl –** Averaged indoor temperature
    2. **T_stp_heat –** Indoor temperature setpoint for heat
    3. **T_out –** Outdoor temperature from nearest weather station
    4. **RH_out –** Outdoor relative humidity from nearest weather station
- **Dependent Variable:**
    1. **auxHeat1 –** Heat source runtime in seconds for the given 5-minute interval. The value of this variable ranges from 0 to 300 seconds (0 minutes – 5 minutes) in 15 second intervals. (0, 15, 30, 45, …, 270, 285, 300 seconds)

M. Approaches Used

Several experimental designs were implemented in order to help the modelling process. These experiments have been listed below:

(a) Hourly Data Aggregation

This approach [10] involved aggregating every 12 consecutive data points (1 hour) from the dataset and averaging the 4 independent variables over this time period. The resultant output variable was added over these 12 time-steps. This problem was then treated as a regression problem in order to predict the total run-time of the heat-source given the aggregated dependent variables. The range of the output variable was from 0-3600 seconds (0 – 60 minutes) The results of this approach were not significant enough for the analysis yielding an RMSE score of 396.5 and a MAE score of 233.06.

(b) Ordinal Classification Problem

In this approach, the data was used as is, i.e., 5 - minute observations as recorded by the thermostat. Since the output variable (heat-source runtime) ranged from 0 – 300 seconds (0 – 5 minutes), in 15 second increments, the data was modeled as an ordinal classification problem. This approach yielded significantly better results in terms of model performance as compared to the hourly aggregation approach.

(c) Introduction of Time-Lag

The independent variables being used (outdoor temperature, outdoor humidity, and internal heating setpoint temperature) do not play an immediate role in the heating needs.

These kinds of factors have a gradual effect on the heating needs. For example, the increasing outdoor temperature may reduce the heating needs of the house, however, this effect will take place gradually as the walls of the house may take some while to heat up. Another example may be the change in heating setpoint of the house. In this case, if there is a difference between the internal household temperature and heating setpoint, the runtime of the heat source will be determined over a few timesteps to bring the temperature to the desired setpoint.

Keeping this effect in mind, a time-lag factor was introduced in the model. This meant that the feature values from the last 30 minutes as well as 1 hour were used in order to make the prediction for the runtime of the heat-source for the current time-step. [5]

It was observed that the model using feature data from the last 30 minutes was as good as the model utilizing feature data from the last 1 hour. To avoid extra noise in the model, it was decided to use the model utilizing data from the past 30 minutes to make a prediction.

(d) Dealing with Bias in Dependent Variable

Since we are now using a classification model, several methods were used in order to treat for the bias in the output variable. It was noted that about 70% of the output variable values were 0. Almost 25% of the put variable values were 300. All the remaining 19 classes cumulatively made up only about 4.5% of the data points.

Several methods were used to deal with imbalanced data [6]. These included Random Undersampling, Oversampling (S.M.O.T.E) and Cost Sensitive Learning. Unfortunately, none of these methods helped to solve the bias. Therefore, to deal with the imbalance, and still be able to model the data successfully, it was decided to use different performance metrics that may help us evaluate the practical working of our model.

N. Measuring Classifier Performance

Since the dependent (response) variable is imbalanced, it is important to make sure that we are able to judge model performance well. Since our modeling did not require giving more importance to either precision or recall, it was decided to use a **weighted f1-score**. Although accuracy was also looked at, it is hard to judge the model based on accuracy scores alone as a high accuracy score when everything gets assigned to the majority class is misleading.

In addition, it was decided to introduce a custom performance metric that would help us judge the practical performance of the data-driven model. The metric that would be used would be the **absolute percentage error** value in prediction over a given time-period. An example working of this kind of metric has been demonstrated below:

Actual runtime of heat source over a period of 3 months = 790,200 seconds

Predicted runtime of heat source over a period of 3 months (by the model) = 770,650 seconds

This would give us an error % in prediction which can be calculated as:

*((|Predicted runtime – Actual runtime|) / (Actual runtime)) * 100*

In the above-mentioned case, the percentage error value would be ***2.474% over a period of 3 months***

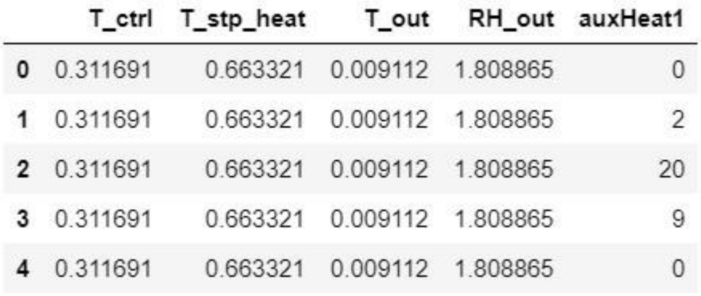
This kind of metric would help us gauge practical performance in terms of predicting the thermal demand over a period of time. A detailed list of all metrics used can be found in Appendix A.

O. Experimental Design – Setting Up the Experiment:

(a) Data Pre-Processing

Since the independent variables used in the modelling process are numerical variables and have values over different scales, we do not want one feature having a larger effect on the model due to a higher range of values.

Hence, all the independent variables are standardized (Figure 14) in order for all features to have a mean of zero and scaling to have unit variance. This way our model gives an equal importance to all features when the training process begins. A sample file link can be found in Appendix - B



	T_ctrl	T_stp_heat	T_out	RH_out	auxHeat1
0	0.311691	0.663321	0.009112	1.808865	0
1	0.311691	0.663321	0.009112	1.808865	2
2	0.311691	0.663321	0.009112	1.808865	20
3	0.311691	0.663321	0.009112	1.808865	9
4	0.311691	0.663321	0.009112	1.808865	0

Figure 14: Sample of Standardized Features Used

(b) The ANN Structure

- *Input Data Shape:*

Since the input data features from the past 6 time-steps is used, the shape of the input data would include features from these previous time-steps as well. Each timestep in our data concerns a 5-minute interval for a total time-lag of 30 – minutes.

- *Features that Contribute in Previous Timesteps:*

T_ctrl, T_stp_heat, T_out, RH_out, auxHeat1

- *Features that Contribute in the Current Timestep:*

T_ctrl, T_stp_heat, T_out, RH_out

- *Input Layer:*

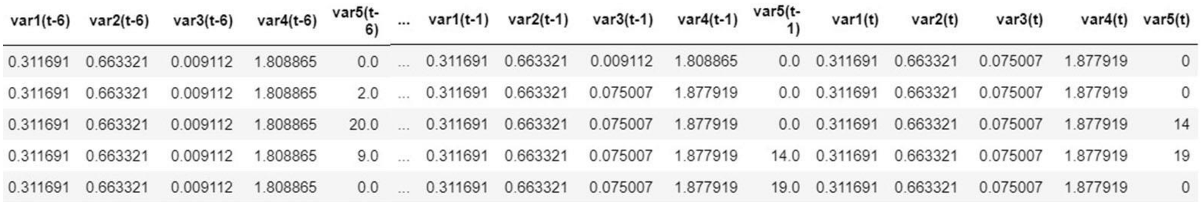
If we consider 6 previous timesteps, the input data would include the following features:

T_ctrl(t-6), T_stp_heat(t-6), T_out(t-6), RH_out(t-6), auxHeat1(t-6),

T_ctrl(t-5), T_stp_heat(t-5), T_out(t-5), RH_out(t-5), auxHeat1(t-5) …

T_ctrl(t-1), T_stp_heat(t-1), T_out(t-1), RH_out(t-1), auxHeat1(t-1),

T_ctrl(t), T_stp_heat(t), T_out(t), RH_out(t)



var1(t-6)	var2(t-6)	var3(t-6)	var4(t-6)	var5(t-6)	...	var1(t-1)	var2(t-1)	var3(t-1)	var4(t-1)	var5(t-1)	var1(t)	var2(t)	var3(t)	var4(t)	var5(t)
0.311691	0.663321	0.009112	1.808865	0.0	...	0.311691	0.663321	0.009112	1.808865	0.0	0.311691	0.663321	0.075007	1.877919	0
0.311691	0.663321	0.009112	1.808865	2.0	...	0.311691	0.663321	0.075007	1.877919	0.0	0.311691	0.663321	0.075007	1.877919	0
0.311691	0.663321	0.009112	1.808865	20.0	...	0.311691	0.663321	0.075007	1.877919	0.0	0.311691	0.663321	0.075007	1.877919	14
0.311691	0.663321	0.009112	1.808865	9.0	...	0.311691	0.663321	0.075007	1.877919	14.0	0.311691	0.663321	0.075007	1.877919	19
0.311691	0.663321	0.009112	1.808865	0.0	...	0.311691	0.663321	0.075007	1.877919	19.0	0.311691	0.663321	0.075007	1.877919	0

Figure 15: Sample of Reframed Dataset as Fed into the ANN. 'var5(t)' is the Target Variable

A total of 34 input features are fed into the ANN. Hence, the ANN consists of 34 nodes in the input layer.

- *Output Layer:*

The output layer consists of 21 nodes, each representing one of the 21 output classes from 0 to 300 seconds (both included) in 15 second intervals.

- *Hidden Layers:*

The number of hidden layers and the number of nodes in each hidden layer can be determined by experimentation and heuristics in order to find the structure that provides us with the best results. The base ANN structure consists of 2 hidden layers consisting of 24 nodes each.

- *Activation Functions Used:*

The best activation functions to be used in the ANN can also be determined by experimentation and heuristic approaches. However, for the base ANN architecture, the following are the activation functions used:

Hidden Layer: ReLu

Output Layer: Softmax

A sample schematic of the resultant ANN structure is shown below:

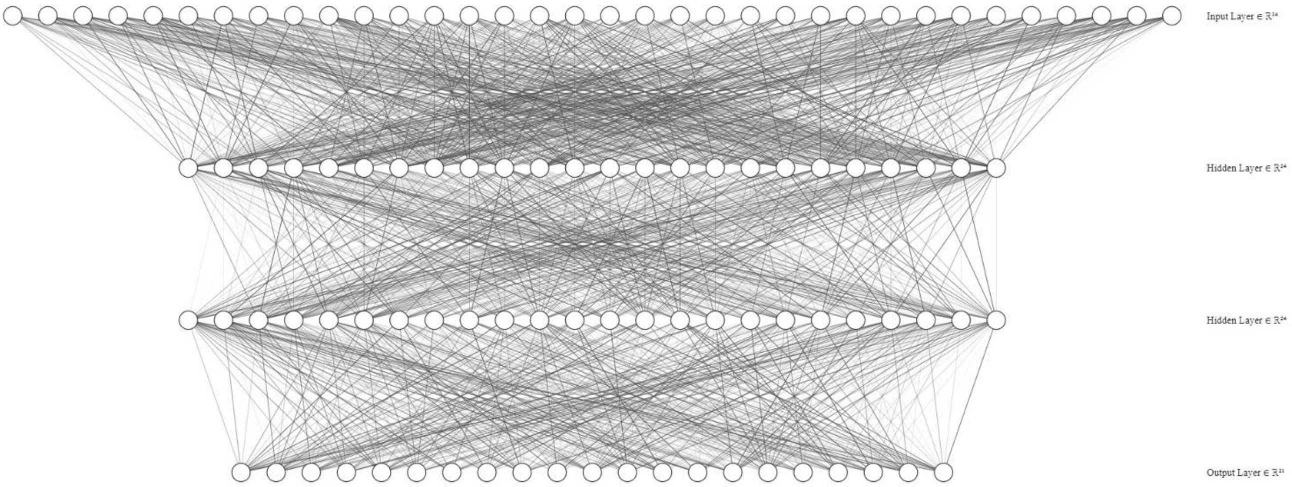

Input Layer ∈ $\mathbb{R}^{34}$
Hidden Layer ∈ $\mathbb{R}^{24}$
Hidden Layer ∈ $\mathbb{R}^{24}$
Output Layer ∈ $\mathbb{R}^{21}$

Figure 16: Structure of ANN Including Input and Output Nodes

(c) Randomization (Train/Test Split)

Once the dataset is transformed to the required format, it is randomly split and shuffled into 2 separate sets, one for training with 80% of the data (about 5 months' worth of data), and the other for testing with 20% of the data (about 2 months' worth of data).

P. Further Experimentation and Revisions:

A series of experiments were performed in order to try and increase model performance. These experiments are further explained below:

(a) Modelling the Basic Features from the Dataset

The first basic modelling experiment included the training and testing of the ANN models on all 4 subject houses from Toronto using only the 4 most important features from the dataset

(b) Including Date-Time Information

Additional features such as month, day of week, and time of day were added to the dataset and used in the modelling process. It was noted that there was no noticeable increase in model performance. Hence, this data will be left out from the modeling process in order to increase model efficiency, prediction time, as well as reduce the amount of noise.

(c) Including Additional Weather Information

Additional weather features for the city of Toronto were added to the dataset. These features included information like Wind Speed (km/h), Station Pressure (kPa), and Wind Chill. Again, it was noted that there was no noticeable increase in model performance. Hence, this data will be left out from the modeling process in order to increase model efficiency, prediction time, as well as reduce the amount of noise.

(d) Model Improvement with the Monthly Addition of Data

This experiment involves using recursive techniques to add data on a monthly basis and then retrain the model on the cumulative data. This experiment will help in determining if the model actually learn and performs better as more data is added to it. This kind of learning capability will act as an important feature if this functionality is added to a smart thermostat system.

Q. Algorithm Selection

While various approaches, techniques, and deep-learning algorithms were used to try and model the data, it was noticed that a simple black-box ANN approach worked the best. The data that was fed into this ANN included data logged in 5-minute time steps and included feature information from the past half hour to predict the heat source run-time for the current time-step.

In addition, the weighted f1-score as well as the percentage error in prediction were used as metrics in order to gauge practical performance of the model.

Further parameter tuning will be performed to try and improve model performance on the dataset.

5. RESULTS AND DISCUSSIONS

R. Exploratory Analysis Results

The initial exploratory data analysis of the Ecobee smart thermostat data presented us with some useful findings that helped us in modelling the data during the further stages. For instance, the exploratory data analysis helped us determine that the majority heating season lasted about 7 months in Toronto, Canada, when the heat usage is considered significant enough to be modeled. These months were January, February, March, April, October, November, and December.

Further, it was observed that certain patters exist in the heat usage of the houses. For example, the heat demand was generally more over the weekends as compared to the weekdays due to the increase in set-point temperatures when the occupants were home. Also, there were certain hours of the day where the heat usage patterns remained consistent. This included low heat usage between the night-time hours of sleep and a sharp increase in heat usage during the evenings between 3pm and 7pm. This again, can be linked to the fact that the amount of heat required increases when the set-point temperature is higher during times of active occupation in the house and is lower when occupant activity is low during the night.

The exploratory data analysis also confirmed that only a few features had a high level of contribution in the decision-making process and that using any additional features may not be very useful. The key features that contributed the most included the Control Temperature, Heating Setpoint Temperature, Outside Temperature, and Outside Relative Humidity and were used to build a simplified black-box model.

Another key observation that was brought to light was the general imbalance in the heat-source run-times for all 4 subject houses in Toronto where the intermediate run-times only contributed to 5-6% of the total data points. This also made it clear that certain techniques would have to be used to help counter the imbalance, either in terms of the modelling process, in terms of metrics used, and/or in terms of increasing/decreasing the amount of data used. The following plots clearly show the distribution of the target variable for all 4 subject houses.

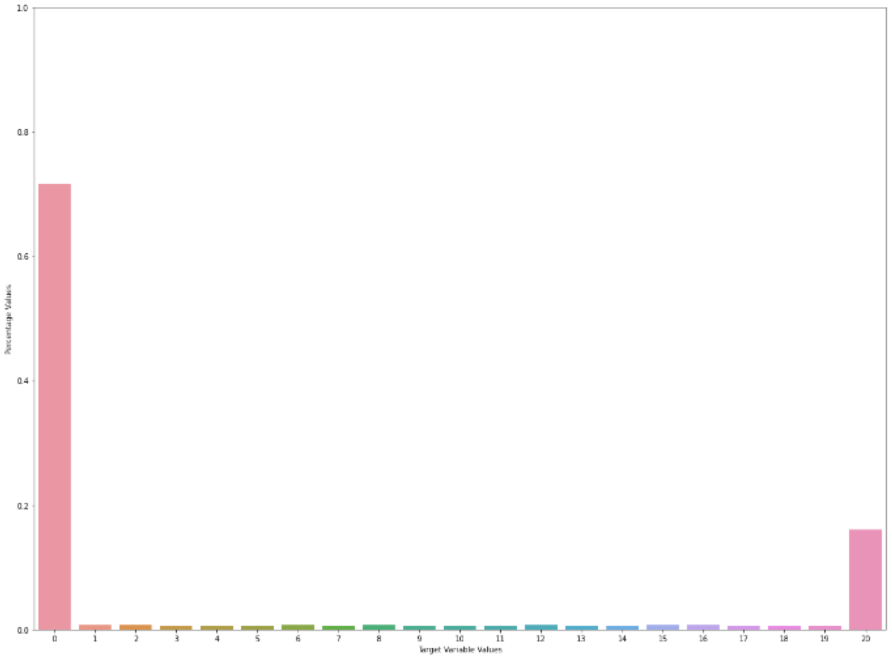


Figure 17: Target Variable Distribution (House 1)

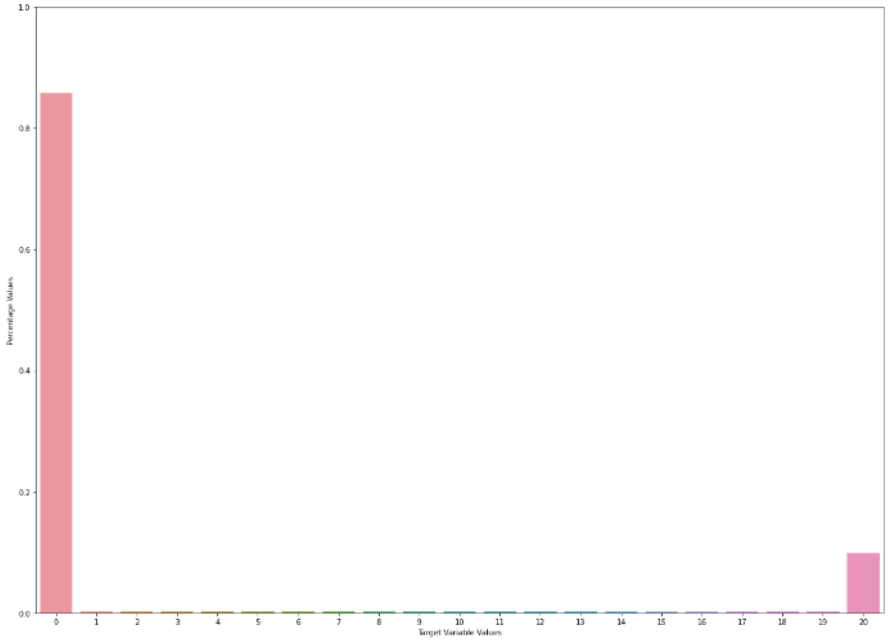


Figure 18: Target Variable Distribution (House 2)

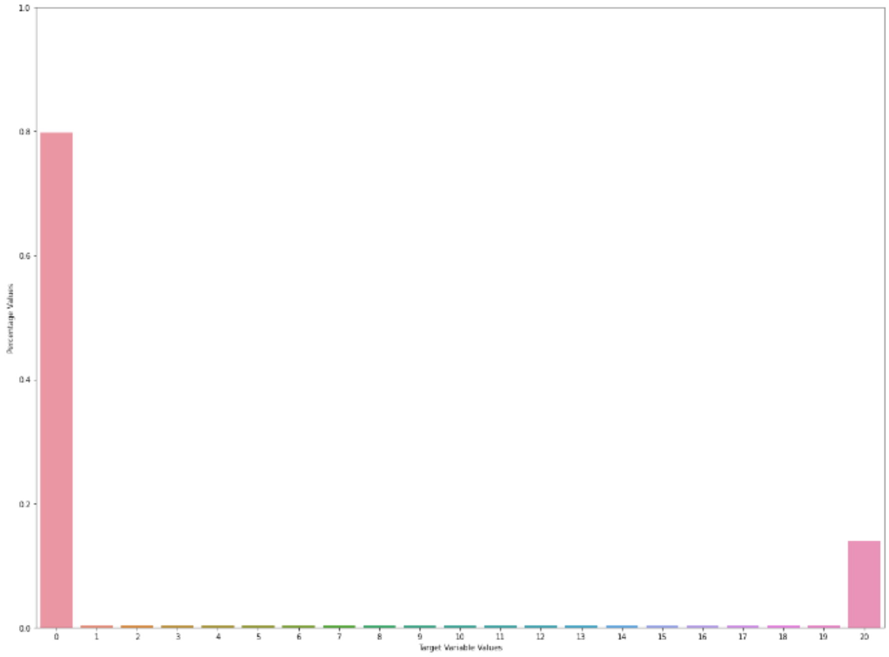


Figure 19: Target Variable Distribution (House 3)

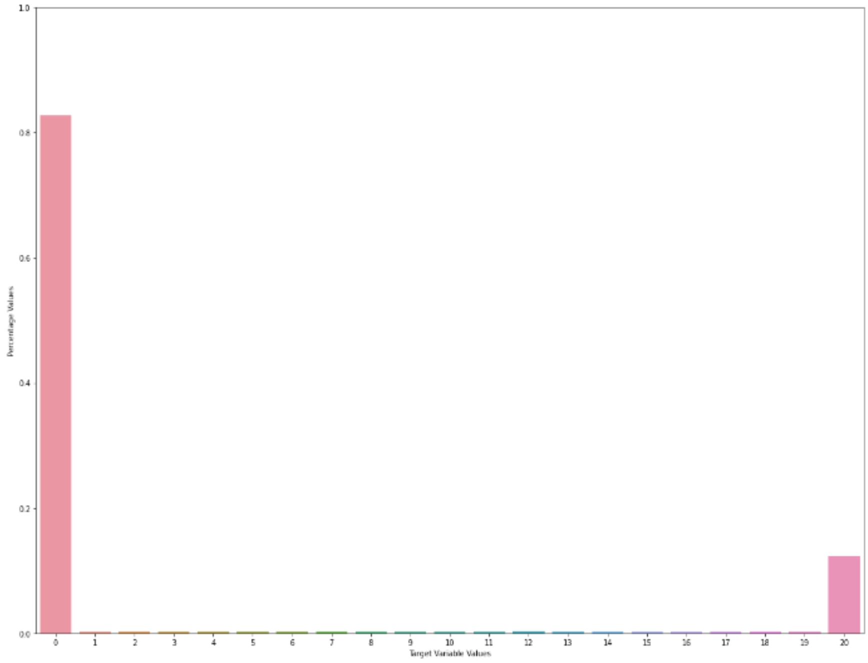


Figure 20: Target Variable Distribution (House 4)

S. Results for Different Approaches Used:

A series of approaches were used to model the data from House 1 appropriately before a single approach was selected to model and experiment on further. The summary of results for these approaches have been outlined below:

(a) Hourly Data Aggregation

This approach helped solve the data imbalance problem to some extent. However, this approach still did not help with achieving a somewhat normal distribution for the target variable. Further, this approach called for us to model the data as a regression problem. The target variable distribution after the removal of outliers has been shown below.

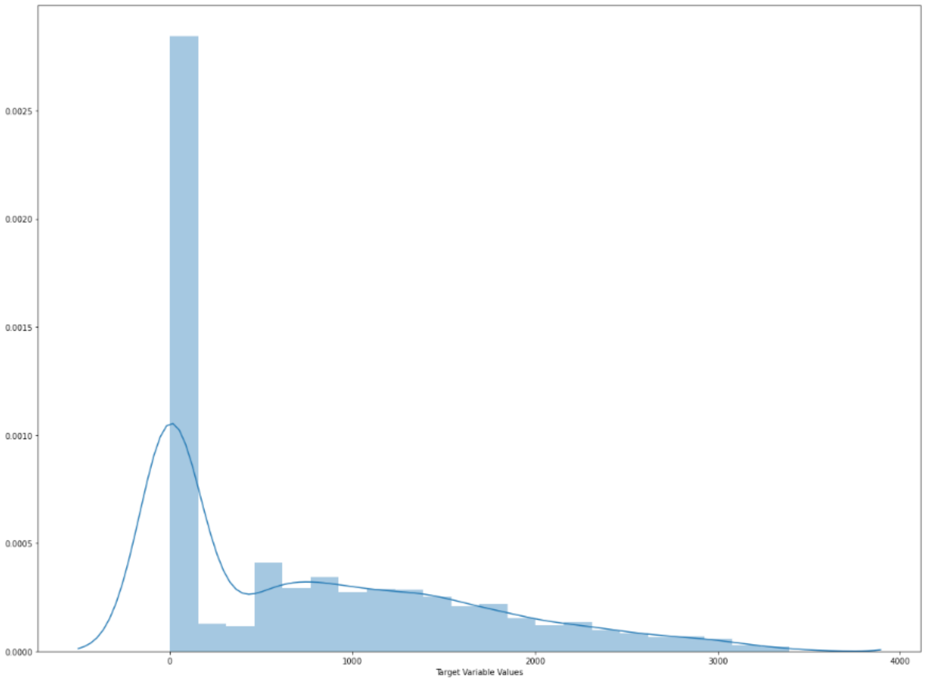


Figure 21: Target Variable Distribution for Hourly Aggregated Data (Approach (a))

In this case, the target variable (heat source run-time) range would be from 0 – 3600 seconds (for the aggregated hourly data) denoting the number of seconds the heat-source ran in the instance. The MSE score achieved on test set was 376.25. This would mean that the heat-source runtime prediction for every hour had an error of about 10.5% which is not a great result and could be improved further. The figure shown below demonstrates the training and testing loss (Mean Squared Error) progression with the number of training epochs.

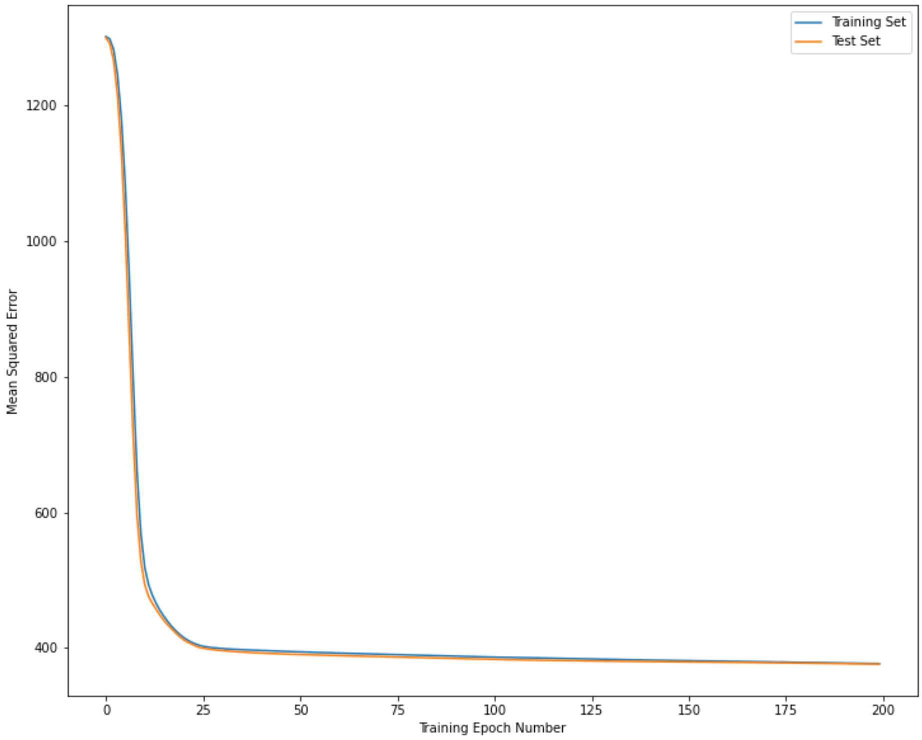


Figure 22: Training and Testing MSE Score for Hourly Aggregation (Approach (a))

(b) Ordinal Classification of Data

Using the default 5-minute interval data and treating it as an ordinal classification approach helped us achieve a better result as compared to the previous approach. This approach gave us a good starting point to improve on. The figures shown below demonstrates the training and testing loss (Categorical Cross-Entropy) and accuracy achieved progression with the number of training epochs.

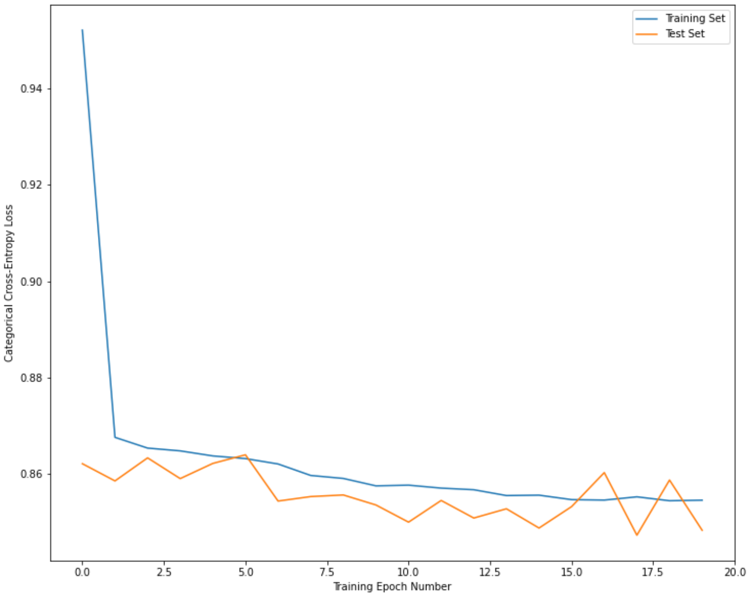


Figure 23: Categorical Cross-Entropy Loss (Approach (b))

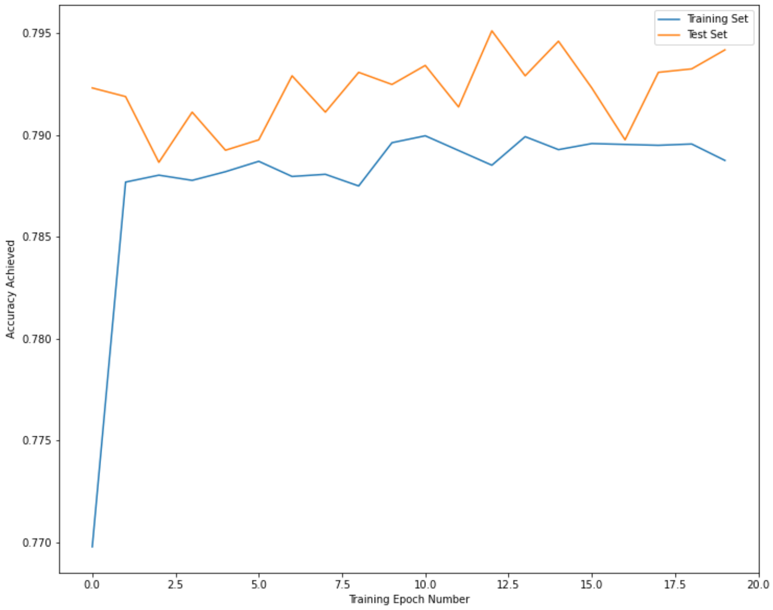


Figure 24: Accuracy Achieved (Approach (b))

Table 2: Results Achieved (Approach (b))

Accuracy on Training Set (About 5 Months)	**Accuracy on Test Set (About 1.5 Months)**	**Weighted F1-Score**	**Error % on Test Data**
78.88%	79.42%	0.738	30.33% over 1.5 months

(c) Introduction of Time-Lag

The introduction of time-lag data proved to be very useful in terms of achieving practically significant results as this approach used implemented the concept that certain features may not play a role in the decision making process immediately but over a period of time. Keeping this in mind, 30-minute and 1-hour time-lagged feature data was also fed into the model and the results achieved have been demonstrated in the figures below:

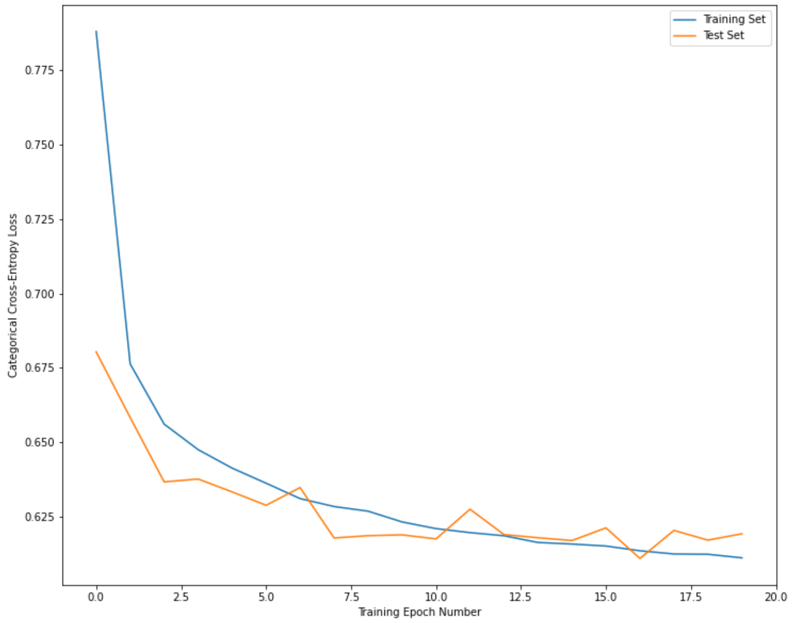


Figure 25: Categorical Cross-Entropy Loss (30 Minute Lag) (Approach (c))

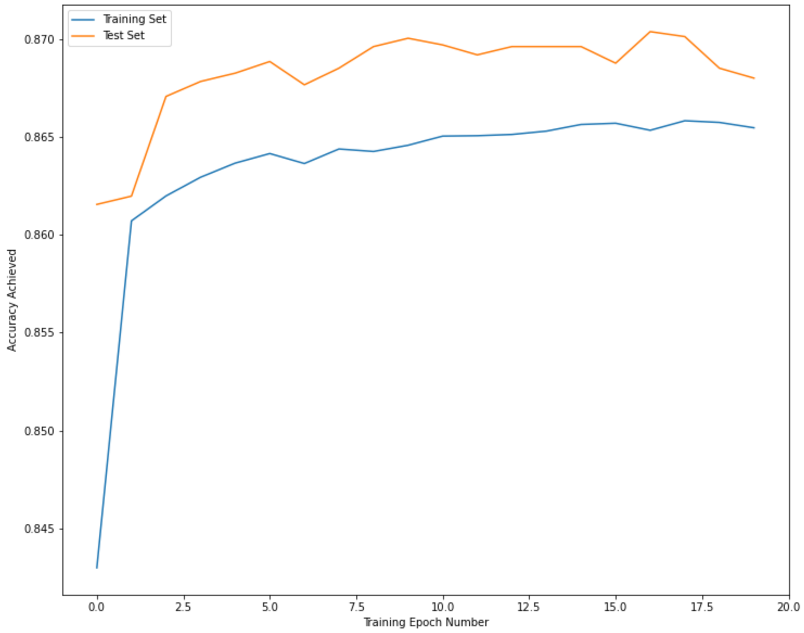


Figure 26: Accuracy Achieved (30 Minute Lag) (Approach (c))

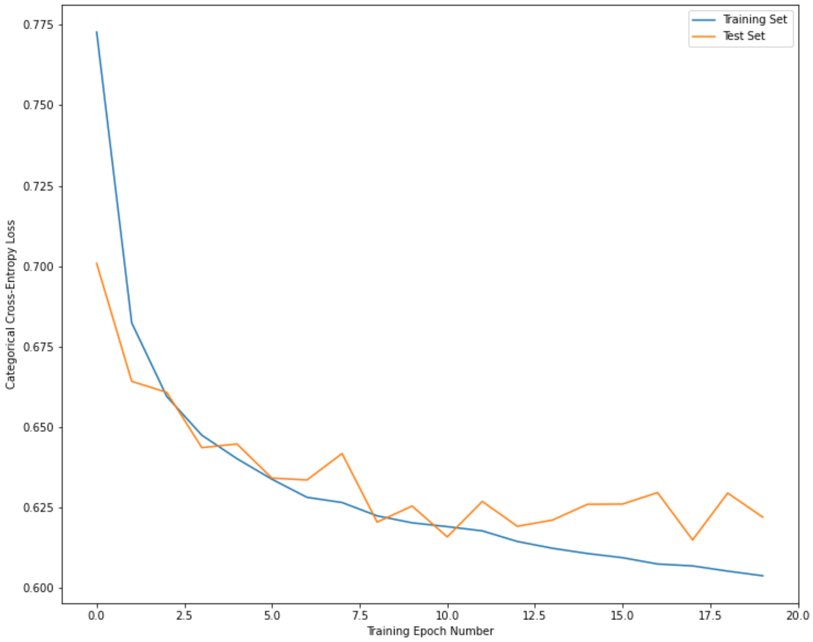


Figure 27: Categorical Cross-Entropy Loss (60 Minute Lag) (Approach (c))

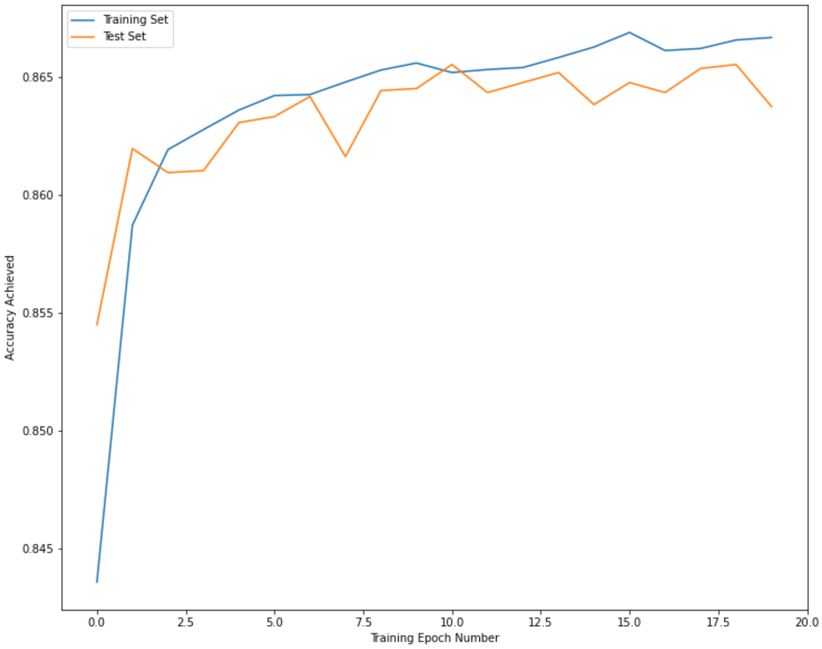


Figure 28: Accuracy Achieved (60 Minute Lag) (Approach (c))

Table 3: Results Achieved (Approach (c))

Time-Lag Duration	Accuracy on Training Set (About 5 Months)	Accuracy on Test Set (About 1.5 Months)	Weighted F1-Score	Error % in Prediction on Test Data
30 Minutes	86.70%	86.67%	0.812	2.33% over 1.5 months
60 Minutes	85.73%	86.48%	0.809	6.11% over 1.5 months

While the training and test set accuracy in both cases were almost similar, the 60-minute time lag case was inferior in terms of the prediction error percentage achieved over the test set.

(d) Dealing with Bias in Dependent Variable

As it is evident that all the houses have a similar imbalanced distribution in terms of the target variable, i.e., the run-time of the heat-source, a few techniques were used to try and treat the imbalance. Techniques such as oversampling (S.M.O.T.E), random under-sampling, and cost sensitive learning were used.

However, all these approaches presented their own issues while training:

S.M.O.T.E – Caused the total number of instances to increase largely. Unfortunately, it was very difficult to train a fully connected ANN network on the resultant large dataset due to hardware constraints.

Random Under-Sampling – While this approach made it easy to use the data to train our model, it reduced the dataset to only a very few instances. This data was then not sufficient for the model to learn from and resulted in a very poor generalization score.

Cost Sensitive Learning – This approach aims at increasing the cost of misclassification of the minority classes. This approach worked the best for us to help us achieve the best possible results.

In addition to these approaches, the metrics being looked at would be the weighted f1-score along with the error percentage in prediction to gauge the practical working of the model. While accuracy is also looked at, it would not be the key metric used.

Looking at the results achieved from the above approaches, it was decided to take-up and improve upon the model that uses a 30-minute time lag to predict the heat-source run-time for the current timestep. Cost Sensitive Learning would be used to help treat imbalance in the dataset to the best ability. Additional modelling experiments are carried out on this model.

T. Results for Model Experiments:

A series of experiments were carried out on the model approach selected. These experiments were carried out on all 4 subject houses from Toronto. The results obtained from each of these experiments have been demonstrated below:

(a) Modelling the Basic Features from the Dataset

As a part of this experiment, training and testing of the ANN models was carried out on all 4 subject houses to see how the resultant model performs using only the 4 most basic and important features. The results achieved have been tabulated in Table 4:

Table 4: Results Achieved (Experiment (a))

Subject House Number	Time-Lag Duration	Accuracy on Training Set (About 5 Months)	Accuracy on Test Set (About 1.5 Months)	Weighted F1-Score	Error % in Prediction on Test Set	Error % in Prediction on Random 1 Hour of Data
House 1	30 Mins	86.70%	86.67%	0.812	2.33% over 1.5 months	9.58% over 1 hour
House 2	30 Mins	95.68%	95.39%	0.939	3.79% over 1.5 months	1.85% over 1 hour
House 3	30 Mins	93.05%	92.84%	0.898	3.46% over 1.5 months	6.25% over 1 hour
House 4	30 Mins	94.81%	94.55%	0.923	0.40% over 1.5 months	6.38% over 1 hour

The analysis helped us understand what results these basic features help us achieve. A further analysis of feature importance was also carried out to see what the most important features were according to the model, including the time-lagged features. The feature importance results have been demonstrated using Figure 28 and Table 5.

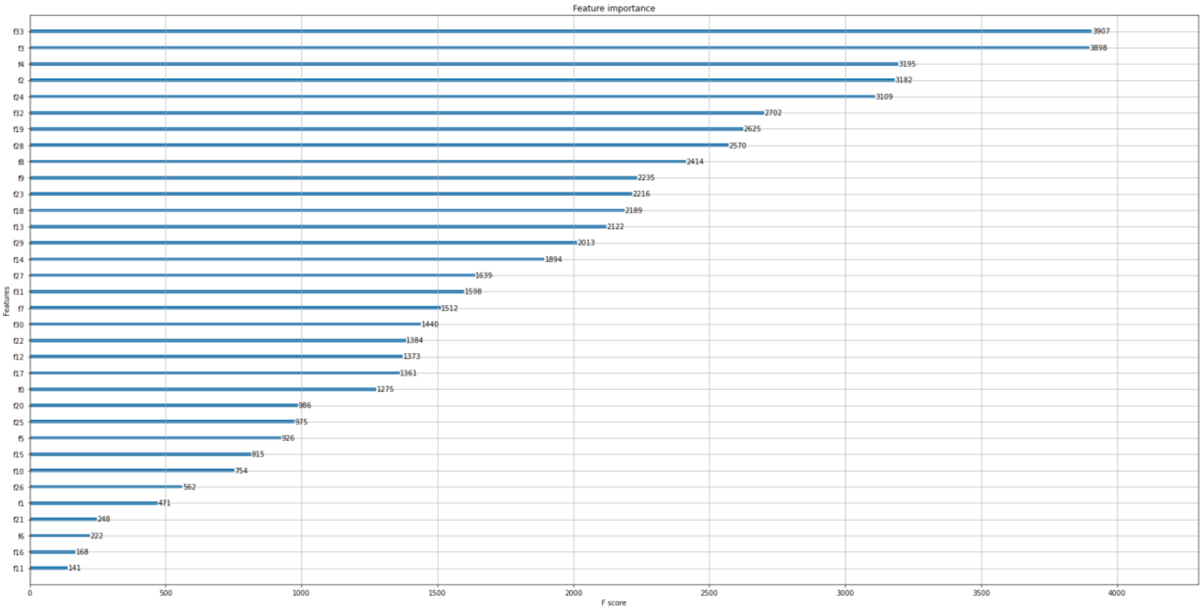


Figure 29: Feature Importance Graph (Experiment (a))

Table 5: Most Important and Least Important Features (Experiment (a))

10 Most Important Features **(From Highest to Lowest Importance)**	**10 Least Important Features** **(From Lowest to Highest Importance)**
1. RH_out (T)	1. T_stp_heat (T-4)
2. RH_out (T-6)	2. T_stp_heat (T-3)
3. auxHeat1 (T-6)	3. T_stp_heat (T-5)
4. T_out (T-6)	4. T_stp_heat (T-2)
5. auxHeat1 (T-2)	5. T_stp_heat (T-6)
6. T_out (T)	6. T_stp_heat (T-1)
7. auxHeat1 (T-3)	7. T_ctrl (T-4)
8. RH_out (T-1)	8. T_ctrl (T-3)
9. RH_out (T-5)	9. T_ctrl (T-4)
10. auxHeat1 (T-5)	10. T_ctrl (T-1)

This analysis makes it very clear that the most important features that help to make a prediction are the outside temperature, outside relative humidity, and heat-source run-time from the previous time steps.

(b) Including Date-Time Information

The main purpose of this experiment was to try and include additional date-time features like month, day of the week, and time of day in order to help the model try and capture patterns that may be linked to certain times, days, or months of the year. The results achieved have been shown in Table 6.

Table 6: Results Achieved with Date-Time Information (Experiment (b))

Subject House Number	**Date Time Features Included**	**Accuracy on Training Set (About 5 Months)**	**Accuracy on Test Set (About 1.5 Months)**	**Weighted F1-Score**	**Error % in Prediction on Test Set**	**Error % in Prediction on Random 1 Hour of Data**
House 1	No	86.70%	86.67%	0.812	2.33% over 1.5 months	9.58% over 1 hour
House 1	Yes	86.58%	86.69%	0.812	3.41% over 1.5 months	9.58% over 1 hour
House 2	No	95.68%	95.39%	0.939	3.79% over 1.5 months	1.85% over 1 hour
House 2	Yes	95.52%	95.43%	0.934	8.644% over 1.5 months	1.85% over 1 hour
House 3	No	93.05%	92.84%	0.898	3.46% over 1.5 months	6.25% over 1 hour
House 3	Yes	93.01%	92.69%	0.896	1.23% over 1.5 months	6.25% over 1 hour
House 4	No	94.81%	94.55%	0.923	0.40% over 1.5 months	6.38% over 1 hour
House 4	Yes	94.67%	94.71%	0.924	5.08% over 1.5 months	6.38% over 1 hour

Table 6 demonstrates that the experiment carried out does not help us achieve any significant improvement in results as compared to the model with only the 4 basic features. As a result, these features can be left out to increase model efficiency and decrease training and prediction times.

(c) Including Additional Weather Information

The main purpose of this experiment was to see if there are any other weather factors including Wind Speed (km/h), Station Pressure (kPa), and Wind Chill that may be contributing to the heat-source run-time. This weather information was gathered from the Canadian Weather and Climate database. The results achieved have been shown in Table 7.

Table 7: Results Achieved with Additional Weather Information (Experiment (c))

Subject House Number	Additional Weather Information Included	Accuracy on Training Set (About 5 Months)	Accuracy on Test Set (About 1.5 Months)	Weighted F1-Score	Error % in Prediction on Test Set	Error % in Prediction on Random 1 Hour of Data
House 1	No	86.70%	86.67%	0.813	2.33% over 1.5 months	9.58% over 1 hour
House 1	Yes	86.6%	87.13%	0.819	2.88% over 1.5 months	9.58% over 1 hour
House 2	No	95.68%	95.39%	0.939	3.79% over 1.5 months	1.85% over 1 hour
House 2	Yes	95.68%	95.5%	0.939	6.7% over 1.5 months	1.85% over 1 hour
House 3	No	93.05%	92.84%	0.898	3.46% over 1.5 months	6.25% over 1 hour
House 3	Yes	92.96%	93.12%	0.903	0.20% over 1.5 months	6.25% over 1 hour
House 4	No	94.81%	94.55%	0.923	0.40% over 1.5 months	6.38% over 1 hour
House 4	Yes	94.81%	94.65%	0.924	2.74% over 1.5 months	6.38% over 1 hour

Table 7 demonstrates that the experiment carried out does not help us achieve any significant improvement in results as compared to the model with only the 4 basic features. As a result, these features are also left out to increase model efficiency and decrease training and prediction times.

(d) Model Improvement with the Monthly Addition of Data

This experiment was aimed at the practical usage and functionality of our data model. It was tried to see if the model performance improves as more data is fed into it. In a real-world scenario, if this model would be deployed into the smart thermostat system, it is important that the model can demonstrate a learning capability and tailor its predictions to the usage patterns of the house it has been installed in. Figures 29, 30, 31, and 32 show the resultant f1-scores as monthly data is added to the models for each house.

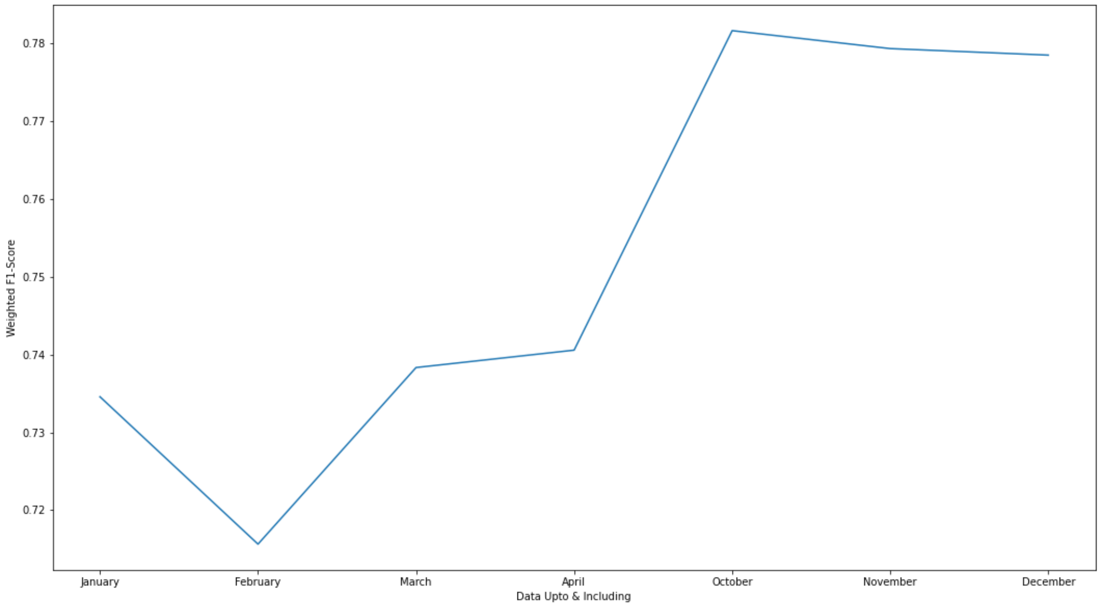


Figure 30: Monthly Data Addition vs Weighted F1-Score for House 1 (Experiment (d))

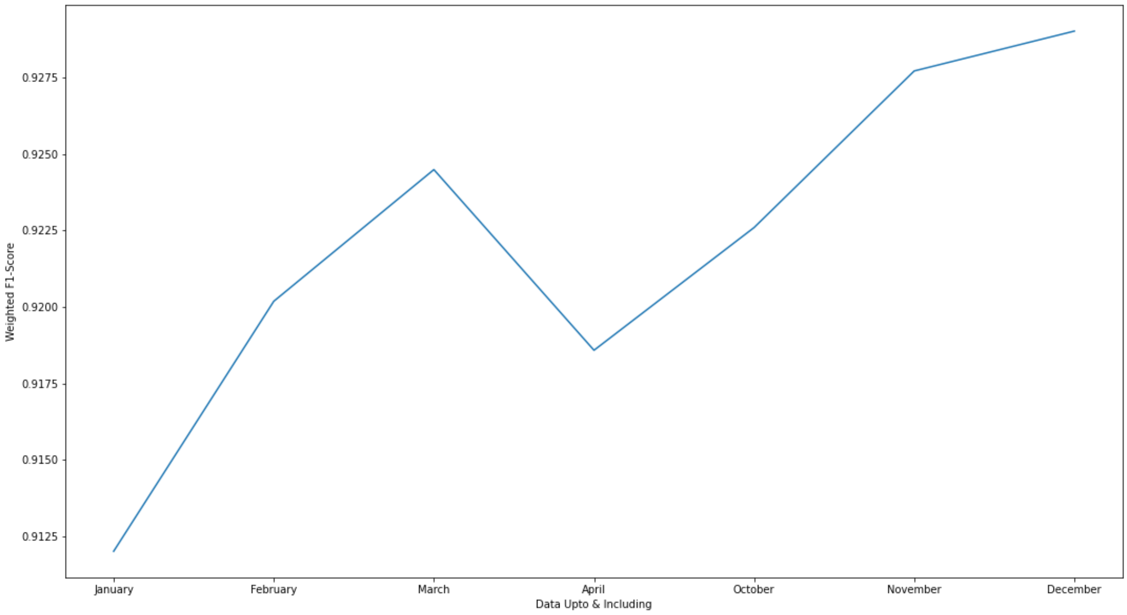


Figure 31: Monthly Data Addition vs Weighted F1-Score for House 2 (Experiment (d))

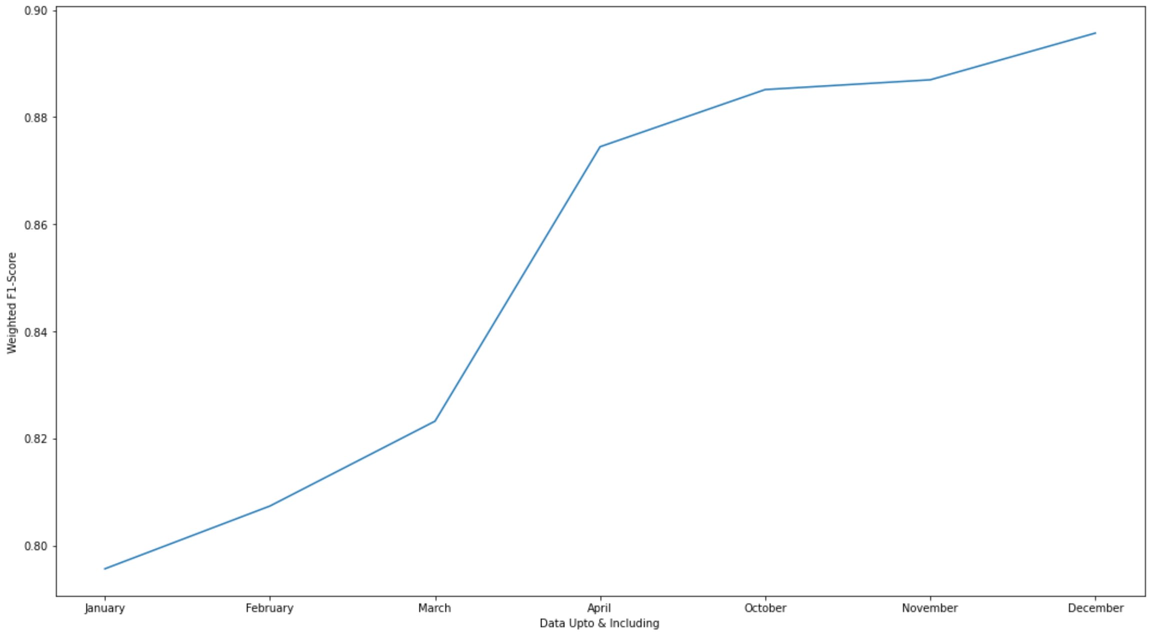


Figure 32: Monthly Data Addition vs Weighted F1-Score for House 3 (Experiment (d))

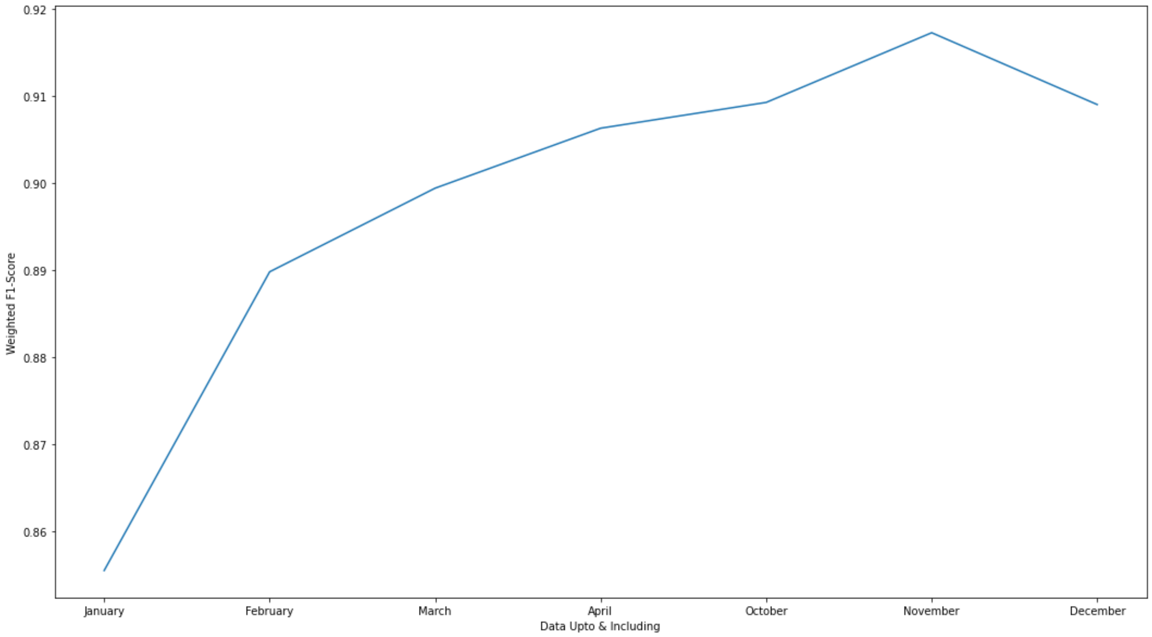


Figure 33: Monthly Data Addition vs Weighted F1-Score for House 4 (Experiment (d))

The trend graphs (Figure 29, 30, 31, and 32) for the weighted f1-scores clearly show an improvement in predictions for each house as more data is added to the model in a monthly manner. This indicates that the model is learns and can make better predictions as more data is logged from each house.

U. General Hyperparameter Tuning Results

The approach to choosing model parameters for training is generally considered to be heuristic. Given the heuristic manner of choosing these parameters, certain techniques can be used to help and tune hyperparameters to try and achieve the best possible results

A grid search approach was used to help find a good choice of hyperparameters including batch size, number of epochs, and optimizer functions. Due to hardware limitations and the amount of time taken, the grid search was applied using only a few parameter choices.

A summary of the detailed results of the grid search have been demonstrated below in Table 8.

Table 8: Results Achieved with Grid Search

Rank	**Batch Size**	**Number of Epochs**	**Optimizer Used**	**Mean Fit Time (Training) (Seconds)**	**Mean Prediction Time (Seconds)**	**Mean Test Accuracy** **(10 Fold Cross-Validation)**
1	64	50	Adam	108.4	0.12	0.866
2	10	50	Adam	700.2	0.42	0.865
3	10	10	Adam	151.9	0.43	0.864
4	64	50	Rmsprop	131.3	0.11	0.862
5	64	10	Adam	22.3	0.12	0.859
6	64	10	Rmsprop	26.9	0.11	0.849
7	10	10	Rmsprop	176.7	0.40	0.770
8	10	50	Rmsprop	832.0	0.42	0.760

The best hyperparameter choices for the model were found to be as follows:

Batch Size = 64, Number of Epochs = 50, Optimizer = Adam

Using these hyperparameter choices, a detailed analysis for the activation function to be used was carried out. Figures 33 and 34 compare 3 different types of activation functions in terms of 2 key areas – Weighted F1 Score and Time Taken for Prediction.

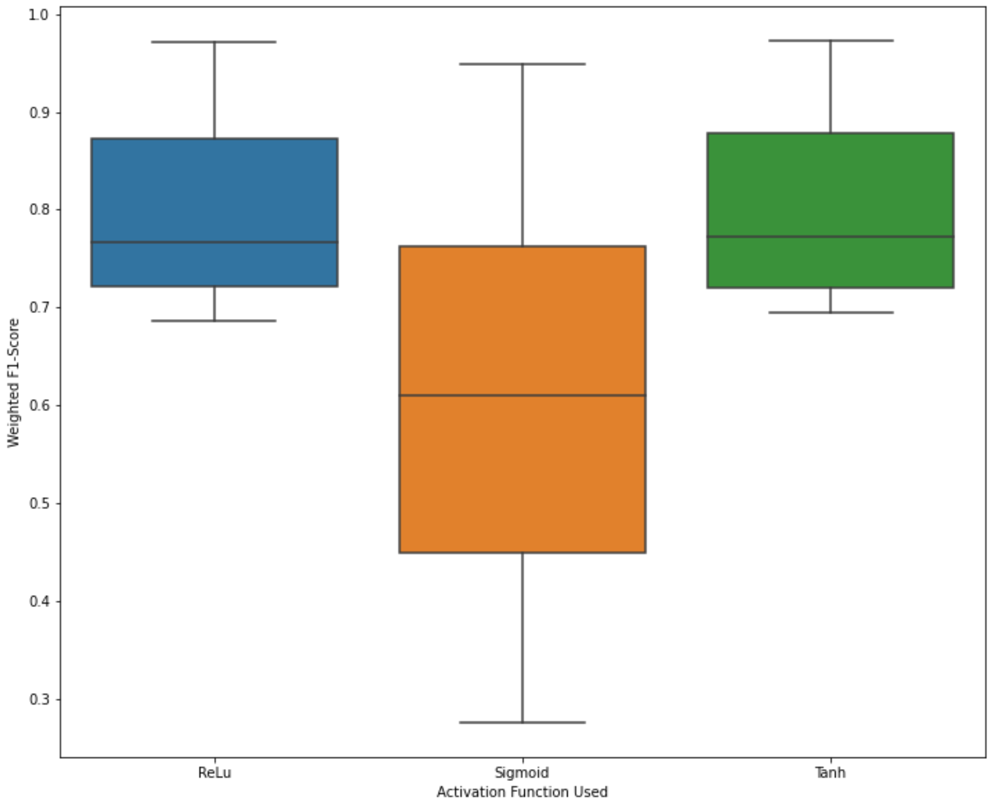


Figure 34: Comparison of Activation Functions in terms of Weighted F1-Score

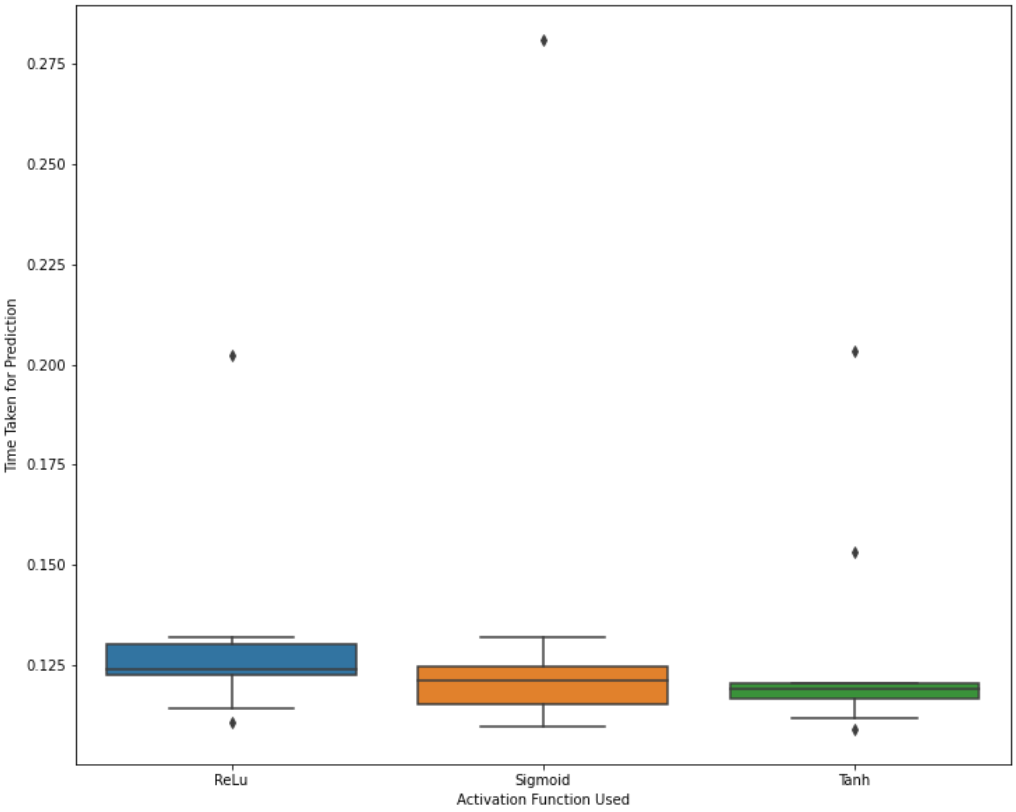


Figure 35: Comparison of Activation Functions in terms of Time Taken for Prediction

V. Discussion and Conclusion

This project put to practice a novel idea of using a black-box approach to build a generalized data-driven model to predict the space-heating demand of a household. The aim of the project was to develop a model that would use a small number of easily observed features to be able to make this prediction. Efforts were made to use different techniques and approaches to achieve the best possible results using the open source dataset as acquired as a part of the Ecobee Donate Your Data program.

While the dataset included data from thousands of houses from across the world, 4 main subject houses were chosen to be studied from Toronto, Ontario, Canada. This decision was made keeping in mind the heating needs due to the cold climate of the region.

As a part of the experimentation process, the biggest problem faced was the imbalance in the target variable. It was observed that the distribution of the target variable for all houses showed a similar bias towards the extreme values of heat source run-time. Keeping this in mind, many different techniques were applied to try and balance the dataset. Unfortunately, all the approaches presented a different set of problems that made the training process harder. Finally, through multiple series of experimentations, it was decided to use the Cost Sensitive Learning approach to help the training process. In addition to this, it was decided to use alternate metrics such as a weighted f1-score and percentage error in prediction to be able to gauge model performance as accuracy would not be able to give a clear picture.

Despite the issues faced, many interesting results were observed as a part of the experimentation process. It was observed that time-lag information from the past 30 minutes played a huge role in determining the space-heating needs of the given time-step. It can also be noted that the ANN model using the 4 most basic features to make predictions worked as good as the models that used either additional weather information or date-time information to make predictions.

Another interesting observation is the ability of the ANN model to learn as data from additional months was added to train on. This demonstrates a key practical usage feature that could be used on smart thermostats.

While notable results were observed, it would be beneficial to have more instances in the data for intermediate classes for heat source run-time. This would help the prediction model gain more instances for minority classes to learn from and hence, be able to make better predictions in terms of net space-heating demand prediction.

6. FUTURE WORK

For further exploration of this study, it would be useful to explore different kinds of patterns between features. It would also be useful to manipulate and aggregate the data in a way that may bring out correlations between features such as outside temperatures and heat source run-time.

It would also be interesting to engineer additional features like 'outside temperature – inside temperature' to try and see if any additional hidden relationships can be established. It may also be useful to try and gather the exact locations of the houses to be able to gain more location-specific weather information like solar radiation and exact temperatures at the location of the house[3] instead of using general information gathered from the nearest weather station.

As a further step, information about the type of heat-source can also be gathered so that the total heat source run-time information can be used to calculate the total thermal energy demand of the house. This information can further be used to determine efficiency of the heat source and quality of insulation of the house.

This kind of study has an end goal of trying to help achieve cleaner, greener, and more efficient households that help reduce energy consumption and in turn reduce the amount of greenhouse gases being produced. [9]

While white-box and grey-box models work great for many control studies, obtaining such detailed, high-resolution, and good quality data to be used in such models is very expensive. Therefore, the use of simplified black-box models and data collected from inexpensive high frequency IoT sensors is needed.

7. APPENDIX – A | COMPARISON OF METRICS USED

Mean Absolute Error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Root Mean Square Error:

RMSE = $\sqrt{MSE}$

Absolute Percentage Error:

APE = $\left|\frac{|\hat{y} - y|}{y}\right|$

Precision:

Precision = $\frac{\text{True Positive}}{\text{True Poitive+False Positive}}$

Recall:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Poitive} + \text{False Negative}}$$

F1-Score:

F1 = 2 $* \frac{Precision*Recall}{Precision+Recall}$

F1 Score is a better measure to use in case there is a need to seek a balance between Precision and Recall and there is an uneven class distribution (with a large number of actual negatives) [7]

8. APPENDIX – B | GITHUB LINK

W. Github Link

https://github.com/ankitdhall97/major-research-project

X. Sample Dataset Files

Metadata File:

https://github.com/ankitdhall97/major-research-project/blob/master/sample_files/meta_data.csv

House 1, Toronto, Raw Data:

https://github.com/ankitdhall97/major-research-project/blob/master/sample_files/toronto_house1.csv

Standardized Data Sample:

https://github.com/ankitdhall97/major-research-project/blob/master/sample_files/std_data.csv

9. REFERENCES

[1] Aydinalp, M., Ugursal, V. I., & Fung, A. S. (2003). Effects of socioeconomic factors on household appliance, lighting, and space cooling electricity consumption. *International Journal of Global Energy Issues, 20*(3), 302. doi: 10.1504/ijgei.2003.003969

[2] Yu, D., Abhari, A., Fung, A. S., Raahemifar, K., & Mohammadi, F. (2017). Predicting Indoor Temperature from Smart Thermostat and Weather Forecast Data. *Communications and Networking Symposium (CNS 2018)*. doi: 10.22360/springsim.2018.cns.012

[3] Demirezen, G., Fung, A. S., & Deprez, M. (2020). Development and optimization of artificial neural network algorithms for the prediction of building specific local temperature for HVAC control. *International Journal of Energy Research*. doi: 10.1002/er.5537

[4] Runge, J., & Zmeureanu, R. (2019). Forecasting Energy Use in Buildings Using Artificial Neural Networks: A Review. *Energies, 12*(17), 3254. doi: 10.3390/en12173254

[5] Luk, K., Ball, J., & Sharma, A. (2000). A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting. *Journal of Hydrology, 227*(1-4), 56-65. doi:10.1016/s0022-1694(99)00165-1

[6] Boyle, T. (2019, February 04). Methods for Dealing with Imbalanced Data. Retrieved August 13, 2020, from https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18

[7] Shung, K. (2020, April 10). Accuracy, Precision, Recall or F1? Retrieved August 13, 2020, from https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

[8] Feature Engineering - Handling Cyclical Features. (2017, October 30). Retrieved August 13, 2020, from http://blog.davidkaleko.com/feature-engineering-cyclical-features.html

[9] Tabatabaei, S., Ham, W. V., Klein, M. C., & Treur, J. (2017). A Data Analysis Technique to Estimate the Thermal Characteristics of a House. *Energies, 10*(9), 1358. doi:10.3390/en10091358

[10] Kipping, A., & Trømborg, E. (2017). Modeling Aggregate Hourly Energy Consumption in a Regional Building Stock. *Energies, 11*(1), 78. doi:10.3390/en11010078