**Literature Review and Exploratory Data Analysis of Predicting**

**Name: Peter Yu, 501137007**

**Date: June 20th, 2022**

**Supervised by: Dr. Alan Fung**

**Background**

The purpose of this study is to explore some of the models used for forecasting household temperatures using local weather station data, as well as examine the types of features that could potentially help with the related task of time series forecasting future household temperatures, as well as the adjacent task of feature engineering. Since there is no formal structuring of the literature review, the plan is to go through several recent academic papers, and summarize their methodologies, results, and findings.

**Literature Review**

**Modelling residential house electricity demand profile and analysis of peaksaver program using ANN: Case study for Toronto, Canada - 2016**

Energy management has been a growing concern of the 21st century given the recent national agreements where various countries have agreed upon aiming for greener solutions as well as the increasing energy prices around the world. This is a two-sided problem as on the demand management side, electricity cannot be easily stored in large quantities, thus electricity companies need to estimate the number of electricity generators needed to meet customer power demands at varying times during the day. Since there is varying demand in electricity, utility companies cannot simply just build more power generators, as sudden drops of electricity demand would result in reduced revenue. Instead, one could investigate machine learning to try to predict energy demand. Essentially Poulad, et al., created an artificial neural network trained on Toronto weather data labels including weather condition, date, temperature, humidity, wind direction, wind speed, visibility, and the average power demand as the ground truth value. They used various data from 2008-2011 along with several customers to average their energy demand, and then they evaluated the artificial neural network performance on various days in the summer, and winter, where the actual energy demand was compared with the predicted throughout the day, and its error was evaluated with Pearson’s coefficient. Having done that, they then evaluated their trained model by backtesting during certain periods, and calculating the cost of electricity, to see whether their model would save customers money. Their findings were that the ANN was able to predict electricity demand in summer as well as winter very accurately (R2 was roughly around 0.99) and that energy savings were easily achieved. One thing to note is since different models use different time periods for training and testing, for simpler generalizability, it might be worth using the same dataset for comparing between other models, as well as looking into other error metrics rather than R2, as Pearson’s coefficient loses value with increases in the number of features.

**Predicting Indoor Temperature From Smart Thermostat And Weather Forecast Data**

In a similar paper, weather station data was then coupled with thermostat data, to try to predict indoor household temperatures. By predicting and forecasting indoor household temperatures, smart thermostats can optimize residential HVAC systems and make buildings more “comfortable” for its tenants. Essentially Yu et al., coupled data from 16 Ecobee thermostats with solar radiation data (namely average indoor temperature, heating/cooling setpoints, humidity, fan runtime, outdoor temperature, and solar radiation) from 2015 and 2016, and trained both a generalized linear regression neural network (GRNN) and artificial neural network (ANN) individually on an 80/20 data split. They used mean squared error as an error metric and found that the MSEs were 0.79 and 11.52 for the GRNN and ANN respectively. Their findings were that by adding additional features outside of thermostat data, the model performance was improved upon.

**Forecasting The Air Temperature at a Weather Station Using Deep Neural Networks**

In a similar topic, apart from the papers published alongside Dr. Fung, was a paper on forecasting air temperature at a weather station using deep neural networks. This paper was chosen as it seemed like the current work being conducted (i.e., using a weather station as a mock representation of a house, and using its weather data to forecast for future temperatures). The author proposed three different neural network architectures from a baseline multi-layer perceptron neural network (MLP) to a Long Short Term Memory Network (LSTM), to finally a Convolutional Neural Network with LSTM (CNN + LSTM) model. By using average wind speed, precipitation, snowfall, snow depth, average temperature, maximum temperature, and minimum temperature from JFK International Airport with data from Jan 1st 2009 to Jan 1st 2019, three neural networks were evaluated for their RMSE and MAPE scores. They found that with the CNN+LSTM model, a 97.42% MAPE was obtained, which was nearly 2% better than the LSTM model and 8% better than the MLP model when forecasting one day ahead. The models suffer when looking forecasting ten days ahead, where the CNN+LSTM model drops to a MAPE of 71%. For future studies, the authors proposed adding more weather variables.

**Exploratory Data Analysis**

As this is an ongoing research project, currently the data on hand is from four different weather stations, including stations in Toronto city, Toronto City Center, Toronto International Airport, and Buttonville. This results in three neighbouring weather stations, with Toronto City being the weather station situated in between the other three. When loading our data in, we get the following dataset:

Table

Description automatically generated

with the following data types:

Table

Description automatically generated

When looking at the above data, my two concerns are having using too many correlated features for forecasting (which could potentially result in a curse of dimensionality scenario) as well as dealing with the NaN values, which seemed to occur quite commonly. This was the same case for the other three dataframes which contained the other weather station data. Using Pandas API, the NaN values were quickly calculated, and the below image showed that there were quite a large number of NaN values for certain features, which resulted in my decision to drop any of the features with >50,000 NaN row values (as my dataset had only ~55,000 rows).

Table

Description automatically generated

In addition to that, I also dropped a number of the columns present in the initial dataframe based on my *a priori* knowledge of which features were very correlated with one another, while keeping in mind that for future analysis, some of these features may be included again. To give context to the EDA process, I also converted the wind direction (in 10s degree) to compass rose bearings, and converted them into dummy columns, as well as added wind chill as a feature to my dataset. Humidex was omitted for now as the following figure showed that the mean was 8.7, while the humidex was only really useful for Humidex values ranging from 20-45+.Text

Description automatically generated

By using the .describe() method, my data examination process could be expedited as shown below:

Graphical user interface, application, table

Description automatically generatedwhich showed that given my *a priori* knowledge, there weren’t any extreme data values in my dataset (i.e., temperature of 50 Celcius which would indicate sensor failure). When generating a heatmap, we then see some good news, where most of the features are not too correlated with one another, which is good because that means less feature engineering is required later on.

A picture containing chart

Description automatically generated

Table

Description automatically generated with medium confidence

The outliers were then scrubbed by removing values that were outside the 1.5\*IQR+Q3 and Q1-1.5\*IQR range. Having cleaned up my final dataframe, the final steps were adding lags to the features as the plan was to forecast for the next hourly periods using 1-hour lagged data, as well as converting the weather labels to something manageable, as there were 80 labels for weather. This procedure was done by converting labels like “Heavy rain”, or “Light rain” into one generalizable label titled “rain”, and simplifying many of the labels. In the end, the distribution of labels was: Chart, bar chart

Description automatically generated

As a final check, some of the dataframe values were also plotted for to confirm that the data did not contain any sizeable outliers.

Plot of the ground truth temperature

Chart, histogram

Description automatically generated

Histogram of the dew point distribution, which should be very similar to the temperature as they are very correlated with one another:

Chart, histogram

Description automatically generated

Histogram of the actual temperature distribution which aligns with the dew point being correlated.

Chart, histogram

Description automatically generated

**References**

1. Poulad, M. E., Department, M. and I. E., Fung, A. S., He, L., Colpan, C. O., & Engineering, D. of M. (2016, July 25). *Modelling Residential House Electricity Demand Profile and analysis of peaksaver program using ANN: Case study for Toronto, Canada*. International Journal of Global Warming. Retrieved June 20, 2022, from https://www.inderscienceonline.com/doi/abs/10.1504/IJGW.2016.077911
2. University, D. Y. R., Yu, D., University, R., University, A. A. R., Abhari, A., University, A. S. F. R., Fung, A. S., University, K. R. R., Raahemifar, K., University, F. M. R., Mohammadi, F., University, S., & Metrics, O. M. V. A. (2018, April 1). *Predicting indoor temperature from Smart Thermostat and weather forecast data: Proceedings of the communications and networking symposium*. ACM Conferences. Retrieved June 20, 2022, from https://dl.acm.org/doi/abs/10.5555/3213200.3213209
3. Roy, D. S. (2020, December 7). *Forecasting the air temperature at a weather station using Deep Neural Networks*. Procedia Computer Science. Retrieved June 20, 2022, from https://www.sciencedirect.com/science/article/pii/S1877050920323784