FINAL PROJECT REPORT

Tabassum Nisha and David Josue Molina

Introduction

One can argue that climate change is one of the most pressing issues the world is tackling at the moment. This is why many professionals are invested in how they can utilize data to address issues that arise from climate change. Furthermore, many organizations are invested in protecting natural habitats as it not only may have an impact on ecological systems but also humans' daily lives. One such organization includes the Minnesota Pollution Control Agency (MPCA) that collects data on 25 Sentinel Lakes in the state. Through collaborative efforts, they strive to predict the consequences of climate change on lake habitats in a representative subset of their state's lake.

For the purpose of this report, we have access to a subset of the data that MPCA has specifically on its Trout lakes. The overall goal of this report is to explore unsupervised learning methods detecting climate change by looking for increased variability in temperatures. Our team hopes to figure out ways to measure variability in temperature. Specifically, we want to explore how temperature variability might differ from the same depth , hour, month but for different years.

As is often the case of unsupervised learning, we are hoping to use unsupervised learning methods to explore new possible structures or patterns. The Minnesota Department of Natural Resources cites as one of its goals to look at new ways to process and analyze data and with an unsupervised learning approach, we can provide such insight on such efforts. Previous efforts from this organization may have suggested that water depth may have been the focus. Like previously mentioned, we are hoping to see how temperatures change in response to time. Thus, our hypothesis is that the temperatures of the lake in the dataset is increasing.

Furthermore on the hypothesis, the motivation would be to obtain a more nuanced understanding of how the temperature changes during the hours of a day. For example, it might be interesting to compare the level of variance in the average temperatures of the day over the years may suggest that over the years the temperatures of the lake have been increasing.

Methods

The dataset used was indeed novel but required sufficient pre-processing to be done. The data preprocessing and data mining steps undertaken before the implementation of the machine-learning models are:

1. Data Cleaning
2. Handling missing values
3. Data Transformation/Manipulation
4. Data Exploration/Data Visualization

Data Cleaning:

There was barely any data cleaning required. Both the training and testing datasets had very less noise. We dropped the duplicates from both the training and testing dataset.

Handling missing data values:

Both the training and testing data had missing values. In the training data some dates had data missing some hours of the day whereas the testing data had the values recorded only for the 0th hour of the day. The missing rows were replaced manually with respective dates and given the value 0 for the temperatures at that datapoint. Also, the training data had very less data recorded for the year 2018, i.e, only for the first few months and the testing data had data recorded for the year 2018 and 2019 at 0th hour only for all the depths. Therefore, this made us think that the testing data was not of much importance along with the 2018 data from the training dataset.

Data Transformation/Manipulation:

The data required a lot of manipulation as we needed to check the variation in the temperature over depth/session. We first chose to orient the data depth wise. Each row represented a date and hour and columns consisted of depths. But after attempting multiple times ways for orienting the data this way made us realise that it fails to help accomplish the problem question. We therefore chose to check the temperature variance over the sessions and therefore calculated the average temperature of the day at every hour. We then calculated the average temperature for the day. We oriented the dataframe in the way that each row represented a date and the columns represented hours of that date and the last column being average temperature of that date.

Later for the purpose of implementation of machine learning approaches we oriented the data in a way that each row represented a date from respective months and each column represented respective years. We then filled the cells with the average temperature for the day of that month and date from that particular year.

Data Exploration/Data Visualization:

For data exploration we performed various statistical tests. We isolated the average temperature of a particular month from two different years(2012,2016). We then generated a scatter plot of the same. We also calculated the covariance and Pearson's correlation coefficient for the same.

For data visualization we plotted a regression plot of the average temperatures of the year 2013 and 2017 and faceted boxplot of average temperatures of the days in a month across the years and thus allowing us to visualize the variance of the average temperatures.

Machine Learning Models:

The baseline algorithm that we used for this report was Lasso regression and compared how our other models performed. The other algorithm includes MLP Regressor, XGB Regressor, AdaBoost Regressor,   and GradientBoosting Regressor. The reason we are using MLP is because it provides us some flexibility in terms of dataset setup and additionally since we have time series data, it is generally good practice to at least try this algorithm. Similar to XGB

Regressor, it has been used on time series data and we felt the format of our dataset was appropriate for such regressors. Lastly, we are including GradientBoosting Regressor because it is typically used to produce regressions predict continuous values such as temperatures. Additionally, GradientBoosting Regressor tends to do better in accuracy and thus we wanted to explore that as well.

In terms of our data source, it comes from the Minnesota Department of Natural Resources. The dataset has information about the Trout Lake that has various sensor's temperatures. This is accompanied by timestamps that include dates and times. The dataset was divided into two where was was labeled "Trout_training.csv" and the other one was called "Trout_testing_features.csv". The data spans from 2012-2018

From a high level-description of the code attached, we started our efforts in the pre-processing stage. The pre-processing portion certainly took the majority of our efforts. In this stage we had to modify the training data to prepare for our unsupervised methods. First step was to separate out the timestamps that included dates and times into separate columns such as time, month, date. Next, we addressed duplication issues by dropping those rows. Since we were interested in being able to compare temperatures hours of the day of the same month but different years we transformed the data to have extra columns that represented different hours of the day. We filled in those new columns with average temperatures from the same depth, hour, and month. We ran some summary statistics to see what directions the average temperatures were going. To deepen our understanding on this matter we created a correlation visualization and a pearson analysis to compare two specific dates temperature readings.

Once we have reached the machine learning portion, we decided to set out to predict the values of 2017 as the data from 2018 was less. We did this by using data from 2012-2016. Performed scalar coding on the "x" data and ran our previously mentioned algorithms. We extracted our training, testing, and mean squared error scores.

Evaluation and Analysis

Statistical Analysis:

For statistical analysis we chose to look into the variance in the average temperature of the water for different dates across the years. Upon plotting the scatter plot of average temperatures of May 2012 against the average temperatures of May 2016, we observed that there was a positive correlation. We then plotted the linear regression plot of the average temperatures of the year 2013 and 2017 which also showed positive correlation. The faceted boxplot displays variance in the average temperatures of different dates in a month across the years. We can observe that the average temperatures in the month January gradually increased over the years, hence justifying our hypothesis.

Since the covariance between the average temperatures of May 2012 and May 2016 is 1110.53577269 and 1619.22584516, which indicates that the relationship between the two variables is positive. The NumPy correlation coefficient and Pearson's correlation coefficient between the same variables is 0.974 which again indicates that the correlation is positive. Thus, this indicates that our hypothesis stands true.

Model Analysis:

As mentioned previously, we used Lasso and Ridge as our baseline model to compare how MLP Regressor, XGB Regressor, AdaBoost Regressor,   and GradientBoosting Regressor performed. In short, the GradientBoostingRegressor performed the best out of all the other ones. In terms of our baseline regressors, we observed that lasso regression to have the best performance. We included testing and training scores that also were accompanied by mean squared errors scores.

Specifically, we found that that Lasso model had 0.9999999and 0.000279 for training and testing scores. Additionally the mean squared error score was 0.0096. In contrast, the GradientBoostingRegressor model had a performance of 0.999955 and 0.999896 for training and testing scores respectively. The GradientBoosting Regressor model had a mean squared error score of 1.0612 making it our best performing model. Thus, our choice of model for the given project is GradientBoosting Regressor model. We can surmise that the Lasso model had better performance than our chosen model but GradientBoosting Regressor model performed fairly well.

Related Work

It is not surprising to see that there has been previous efforts on predicting hourly temperatures. For the sake of brevity, we will discuss two papers. The first is about a group of researchers in South Korea who were trying to forecast hourly temperatures using deep neural networks (Lee, Sungjae, Yung-Seop Lee, and Youngdoo Son., 2020). The main takeaways is that the researchers conducted the study on 3 different locations with climates. What the study showed was that hourly temperatures were better less frequently temperatures input. Additionally, researchers used three algorithms which included MLP , Long-Term Short Memory (LTSM), and Convolution Neural Network (CNN).

The second paper of discussion includes researchers that specifically looked at to model hourly river temperature (Hebert, Cindie, et al.,2014). For this paper, researchers used a model called Artificial Neural Network (ANN). The reason for choosing ANN is because ANN has been largely used in hydrology research as many relationships in this field are not linear. Additionally, the paper has pinpointed hourly water temperature models are not common. Lastly, researchers mentioned that water depth is an influential factor on temperatures.

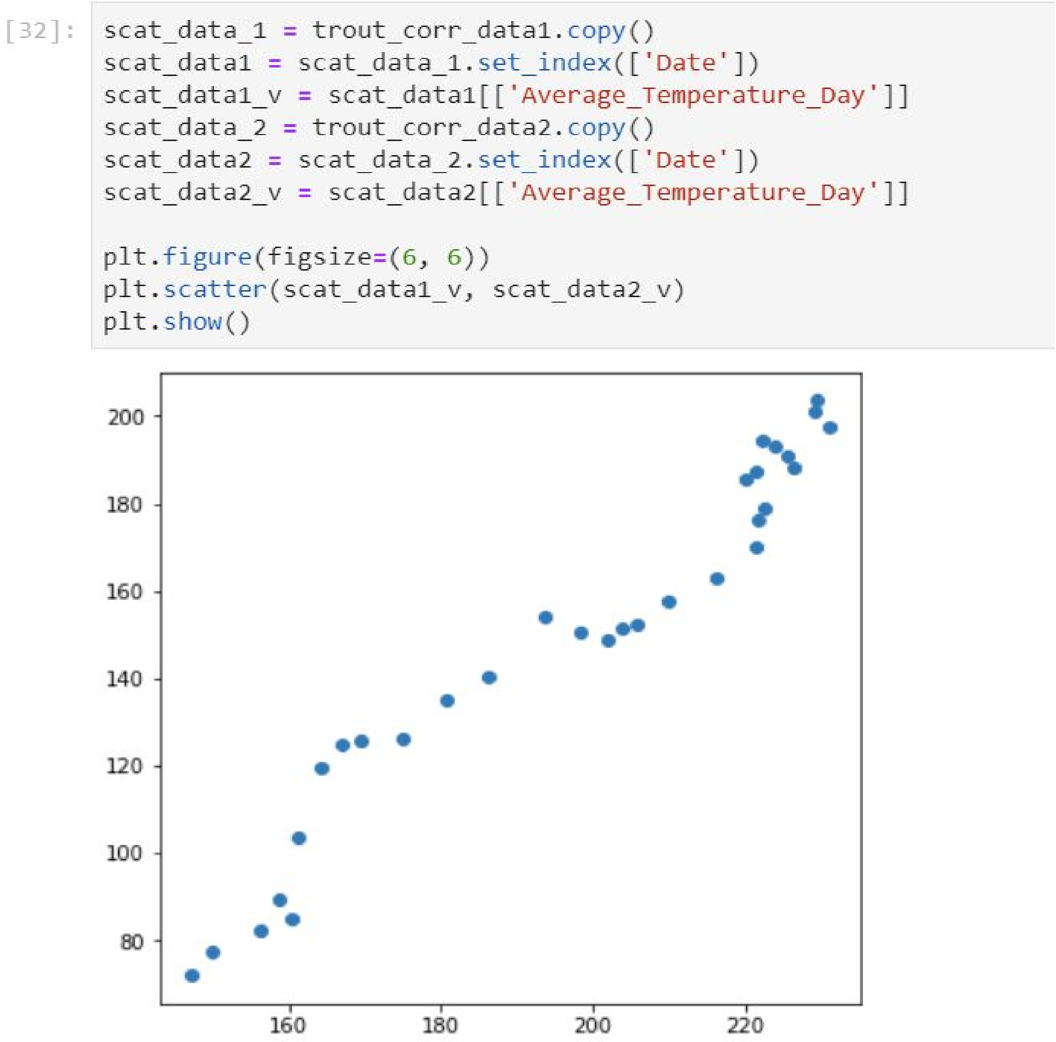
Discussion and Conclusion

Overall, our findings showed that our hypothesis was supported. Additionally, many of our models performed well in training and testing scores but it was the mean squared errors that help differentiate each model from each other.

There were definitely some learning moments throughout this final project. First, there were critical decisions of how to transform data into a format that was appropriate for unsupervised machine learning. This is in part because we were working with time series and continuous which gave us options on such transformations. Secondly, we gained a deeper knowledge of the different metrics to understand variance in the data.

If we worked on this project further, we would have liked to create predictive models to predict temperature's variance. Additionally, it would have been fruitful to explore other algorithms as the research suggests that there might be other algorithms that handle time series and hydrology relationships better. Lastly, it may be helpful to include other evaluation metrics to compare models.

Diagrams

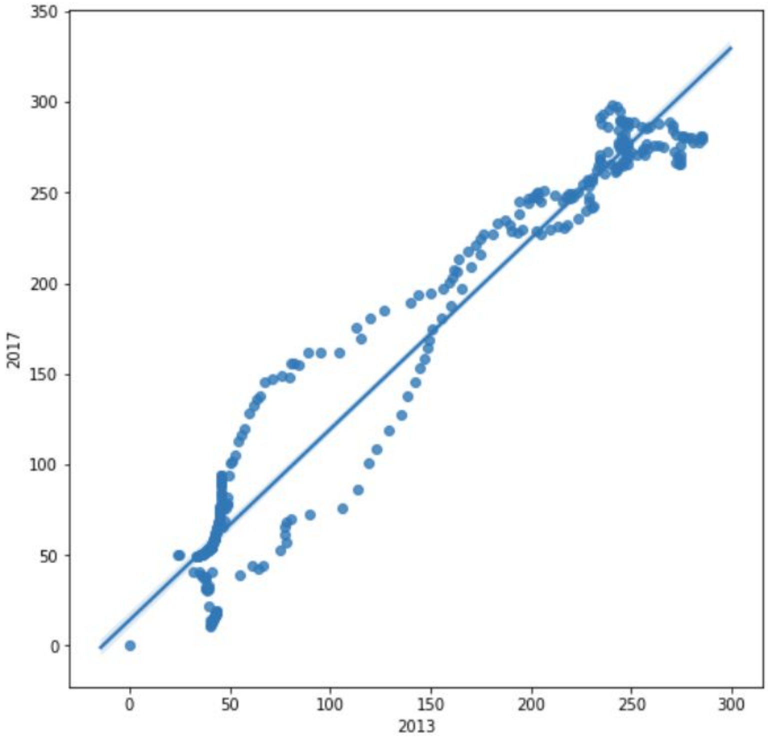


```
[32]: scat_data_1 = trout_corr_data1.copy()
      scat_data1 = scat_data_1.set_index(['Date'])
      scat_data1_v = scat_data1[['Average_Temperature_Day']]
      scat_data_2 = trout_corr_data2.copy()
      scat_data2 = scat_data_2.set_index(['Date'])
      scat_data2_v = scat_data2[['Average_Temperature_Day']]

      plt.figure(figsize=(6, 6))
      plt.scatter(scat_data1_v, scat_data2_v)
      plt.show()
```

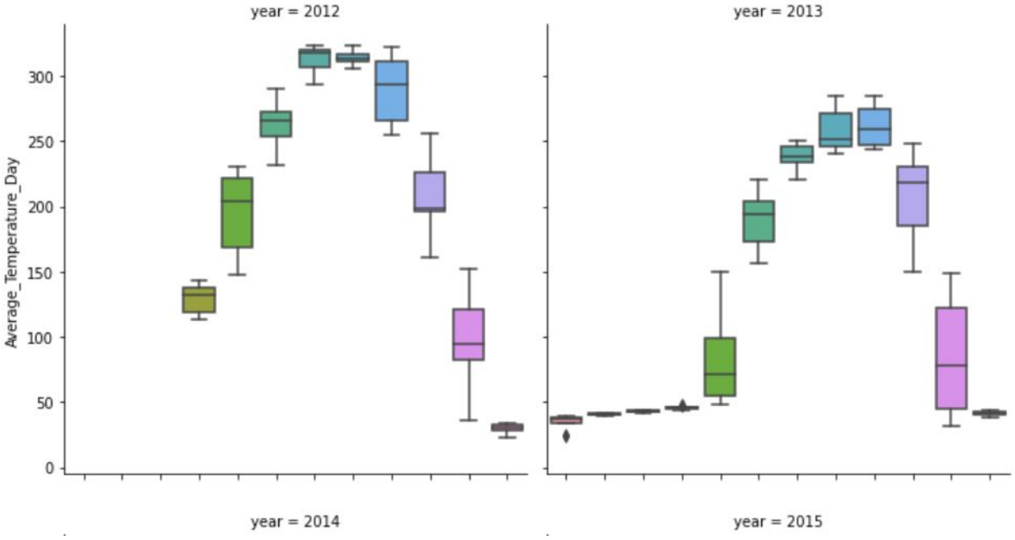
```
trout_reg = sns.regplot(x="2013", y="2017", data=analysis_dataset_train)
trout_reg
```

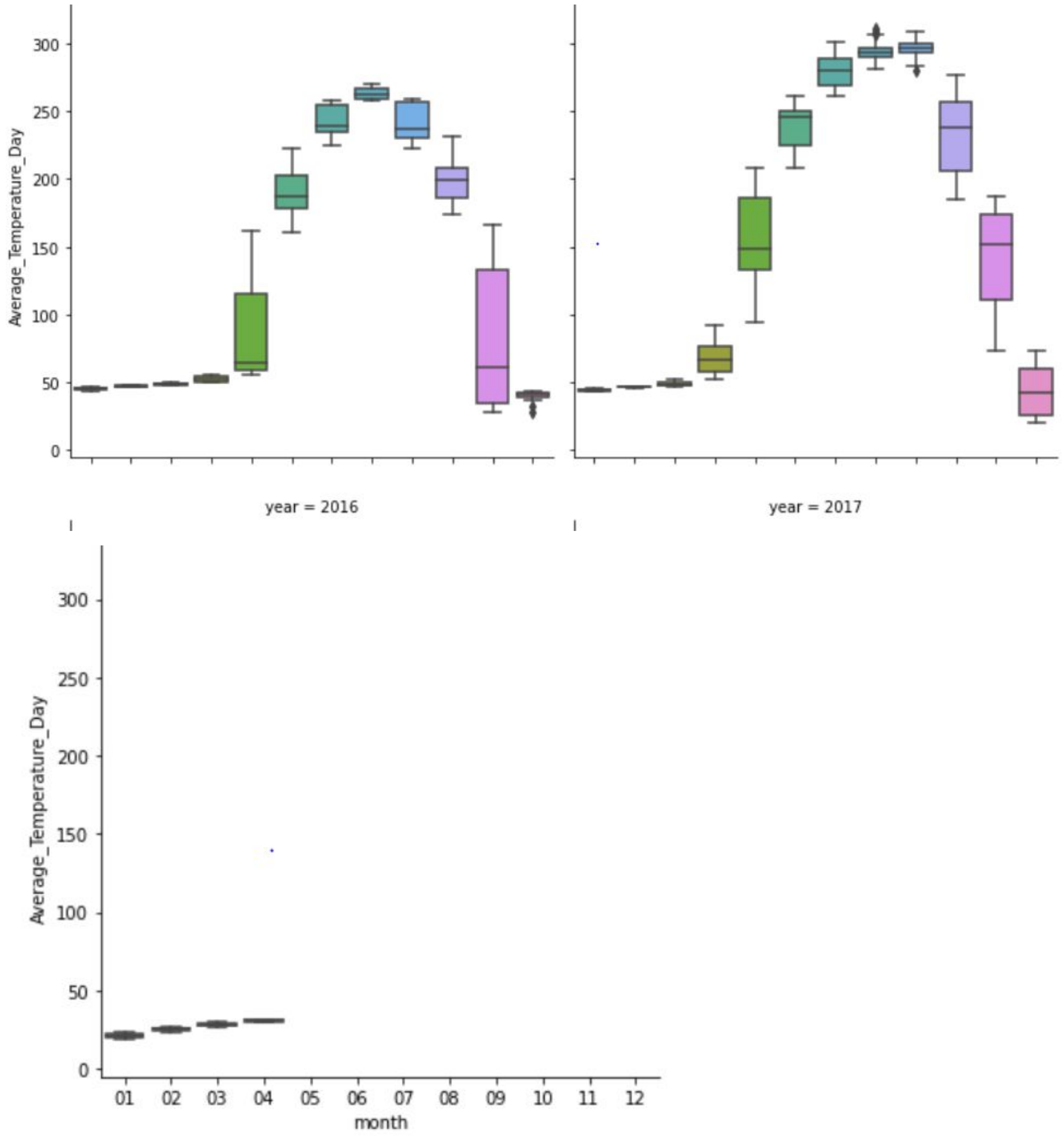
```
<matplotlib.axes._subplots.AxesSubplot at 0x180e0fcd448>
```



```
trout_boxplot = sns.catplot(x="month", y="Average_Temperature_Day",col="year", kind='box',col_wrap=2, data=trout_ready)
trout_boxplot
```

```
<seaborn.axisgrid.FacetGrid at 0x180e3d2fb08>
```





References

1. Lee, Sungjae, Yung-Seop Lee, and Youngdoo Son. "Forecasting daily temperatures with different time interval data using deep neural networks." *Applied Sciences* 10.5 (2020): 1609.
2. Hebert, Cindie, et al. "Modeling of hourly river water temperatures using artificial neural networks." *Water Quality Research Journal of Canada* 49.2 (2014): 144-162.