Assignment Two

Miles Kent

December 18, 2023

Abstract

This assignment was broken into two parts. The first part was to become familiar with basic machine learning techniques and statistical modeling. The second part of the assignment was to apply the knowledge learned in the first part of the assignment to the NSERC Alliance Missions Air Quality Monitoring project. To elaborate, to determine what environmental factors influence methane emissions from Mildred lake and develop a machine learning model that can accurately make predictions biased off of the data fed to it. I will discuss what I have done to achieve these task and go over the modeling techniques I have implemented. I will also discuss briefly the drawbacks and what needs to be modified to make what I have done more effective.

Comment

All codes and significant outputs that are not included in this write up are included in the Assignment 2 folder. If you have any questions regarding this document or the code within the Assignment 2 folder, please reach out to me.

1 Data

1.1 Generation Of The Set

Air quality monitoring data is from the Mildred Lake Air Quality Monitoring Station provided by the Alberta Data Warehouse (https://www.alberta.ca/alberta-air-data-warehouse). We assume that all measurements are accurate due to the level two status of the data, which means that it has gone through quality control checks implemented by the providing source. Data from the Mildred Lake Air Quality Monitoring Station is supplemented with data from the Mildred Lake Weather Station provided by the Government of Alberta's historical climate data depository (https://acis.alberta.ca/acis/weather-data-viewer.jsp). The Mildred Lake Air Quality Monitoring Station is located about 1.5km from the Mildred Lake Weather Station. The supplementing process is as follows. Both data set rows are matched by date and time. We start our filtering process by discarding all rows in the combined data set where there are no or invalid methane measurements. Both data sets have some of the same variables so we supplement the missing independent variable entries from the warehouse data with the weather station data for these repeated rows. This is not done for any repeated wind parameters as these can change over the distance of 1.5km which can lead to inaccuracy. Once we have supplemented what we can, we drop all repeated parameters (columns) in our combined data set, keeping the columns from the air quality monitoring station out of the repeated columns. Finally, we drop any rows where there is not an entry for all parameters within the row. This leaves us with our finals data set.

1.2 Data Variables And Basic Properties

Our final data contains 24937 hourly measurements. These measurements are within the date range of November 26, 2019 to February 7, 2023. Our data set contains the following parameters which represent the environment around the air quality monitoring station and the local weather around Mildred Lake.

- **Methane ppm**: The dependent variable. This variable measures the average amount of methane detected in the air in parts per million over an hour period.
- **Outdoor Air Temp C**: Measurements of the average Temperature in Celsius over an hour period
- **Relative Humidity Per**: Measurement of average humidity in percent over an hour period. $[0, 100]$
- **Wind Direction deg**: Measurement of average wind direction in degrees relative to geographic north over an hour period. $[0, 360)$
- **Wind Direction Std Dev**: The standard deviation of the wind direction over the hour long measurement period. This variable signifies how often the wind direction changes over an hour.
- **Wind Speed kmhr**: The measurement of the average wind speed in kilometers per hour over an hour period.
- **Wind Speed Std Dev**: The standard deviation of the speed direction over the hour long measurement period. This variable signifies how often the wind speed changes during the hour measurement period.
- **Date Time**; Signifies the date and time of the measurements in the format yyyy-mm-dd hh:mm:ss.
- **Precip Amount mm**: The measurement of the amount of precipitation that has fallen over the hour measurement period in millimeters.

- **Dist To Stat km**: Is the distance to the station from the center of Mildred Lake in kilometers.
- **Ang To Stat deg**: Is the wind direction in degrees relative from the center of Mildred Lake to the Mildred Lake Air Quality Monitoring Station. $(-180, 180]$

The full clean data set can be found in the Assignment two folder as a csv file called "MODIFIED MILDRED.csv"

2 Analysis of Monthly Emission Averages

2.1 Kruskal-Wallis Test

Due to the sample size and the non normality of our distribution of methane measurements (Figure 1), we decide to use a Kruskal-Wallis Test to analyse if there is a difference in emissions by month. The Kruskal-Wallis Test is a non parametric statistical test used to see if there is a difference in the average value of a parameter between populations in a data set with non normal distributed populations. The assumptions of this test are that all of the samples are random, all of the samples are independent of each other, all populations are relatively similar in shape (though not absolutely necessary), and the sample size contains five or more individual measurements. The null hypothesis, H_o, for this test is that there is not a difference in the average value of the parameter being tested between populations. The alternative hypothesis, H_a, is that there is a difference in the average value of the parameter being tested between populations. The way that this test is conducted is we first choose a significance level, preferably .05 or less. Next, assign each of our k samples a rank from lowest to highest and calculate the H critical value using

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1) \tag{1}$$

where n is the total number of observations, R_j denotes the ranks for the sample data of group j, and n_j denotes the sum of the ranks for the sample data. The H critical value follow a chi squared distribution with degrees of freedom equal to $k-1$, so we can use a chi squared table to identify the p value of our test. If the p value is less then our significance level, we reject the null hypothesis. On the other hand, if the p value greater then the significance level we choose, we cannot reject the null hypothesis

2.2 Applying Kruskal-Wallis Test

We conduct the test at a .05 significance level on our data using R 4.0.5. We obtain H = 4193.5 with 11 degree of freedom and a P value equal to $2.2 * 10^{-16}$. Since our P value is less then our significance level, we can reject H_o and conclude that there is a difference in adverage methane emission by month. With analysis of Figure 1, we see that emission is lower over the summer and higher over the winter.

3 Regression Analysis of Environmental Variables

We analyse the significance of each independent variable on methane emission using regression analysis. We decided to use polynomial regression to analyse this biased off of its ease of implementation and its flexibility. A polynomial regression can be defined as a function that estimates the value of dependent variable taking into account a set of independents. The slopes assigned to the regression model parameters are estimated by fitting the regression model to a data set using least squares. The form of a polynomial regression with no interaction terms is

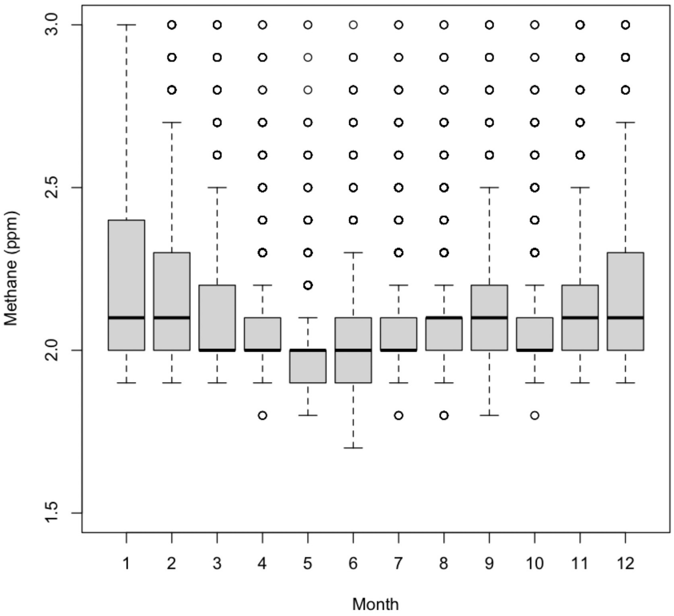

Figure 1: Methane Concentration (ppm) vs Month of Measurement

$$f(x_1, ..., x_n) = \sum_{i=0}^{k} \beta_{1,i} x_1^i + ... + \sum_{j=0}^{l} \beta_{n,j} x_n^j \tag{2}$$

where each β is the slope estimate for the each model parameter dictated by least squares, n signifies the number of independent variables in our regression and often our data set, i and j is the number of degrees we decide to include our regression for each variable, and each x_n is an independent variable. The significance of each term is dictated by its β term, where we can use a basic statistical t-test to determine the terms significance.

3.1 Building Our Regression Models; A Bottom Up Approach

We build our regression using the steps as follows on our data-set. All test and parameters are calculated with R 4.0.5.

1. Transform each independent variable in our data set up to an order of ten. For each transformation, generate a new column on our data set to hold all transformed entries and treat as a new independent variable.

2. Fit a simple linear regression to each independent variable (column) in the set and calculate the AIC score of the fitted model to determine variable significance. Once this is done, rank each model from lowest to highest where the most significant model has the lowest AIC score.

3. Calculate the Pearson's correlation coefficient between the variables/model and remove all highly correlated variables.

4. Fit full multivariate regression.

5. Remove less significant predictors if they exist

We decided to keep the ten most significant variables in our full model. This leaves our full model to be

$$f(x_{ws}, x_{rh}, x_T, x_a, x_{ssd}) = \beta_0 + \beta_1 x_{ssd} + \beta_2 x_{rh} + \beta_3 x_T + \beta_4 x_a + \beta_5 x_T^2 + \beta_6 x_a^2 + \beta_7 x_{ssd}^5 + \beta_8 x_{ws}^7 + \beta_6 x_T^7 + \beta_{10} x_a^7 \quad (3)$$

where x_{ssd} is the standard deviation of wind direction, x_{rh} represents relative humidity in percent, x_a represents the wind direction in degrees relative to the air quality monitoring station from the center of the lake, x_T is the air temperature in Celsius, and x_{ws} is the wind speed in kilometers per hours. From this we derive our additional models to compare which are

$$f(x_{ws}, x_{rh}, x_T, x_a, x_{ssd}) = \beta_0 + \beta_1 x_{ssd} + \beta_2 x_{rh} + \beta_3 x_T + \beta_4 x_a + \beta_8 x_{ws}^7 \quad (4)$$

$$f(x_{rh}, x_T, x_a, x_{ssd}) = \beta_0 + \beta_1 x_{ssd} + \beta_2 x_{rh} + \beta_3 x_T + \beta_4 x_a + \beta_6 x_a^2 + \beta_7 x_{ssd}^5 + \beta_{10} x_a^7 \quad (5)$$

$$f(x_{ws}, x_{rh}, x_T, x_a, x_{ssd}) = \beta_0 + \beta_1 x_{ssd} + \beta_2 x_{rh} + \beta_3 x_T + \beta_4 x_a + \beta_5 x_T^2 + \beta_6 x_a^2 + \beta_7 x_{ssd}^5 + \beta_8 x_{ws}^7 \quad (6)$$

$$f(x_{ws}, x_{rh}, x_T, x_a) = \beta_0 + \beta_2 x_{rh} + \beta_3 x_T + \beta_4 x_a + \beta_5 x_T^2 + \beta_6 x_a^2 + \beta_8 x_{ws}^7 + \beta_6 x_T^7 + \beta_{10} x_a^7 \quad (7)$$

where (4) is the model containing one of each individual parameter, (5) is the model containing parameters where each β has a significance less then 0.001, (6) is a model containing two or less of each parameter, and (7) is a model containing only measurable parameters (no standard deviation).

3.2 Model Assessment

We compare the models using AIC score, VIF score, RMSE, R^2 score, and adjusted R^2 score (Table 1). Below is a brief summary of how each one is calculated and how it helps us assess our models.

- **AIC**: AIC helps us determine which is the best model to use for a set of data when comparing models. It allows us to assess how well a model fits the data by taking into account the number of parameters in the model to prevent over fitting. The lower the AIC score, the better the model is. The AIC score of a model can be calculated with the following formula

  $$AIC = 2K - 2ln(L) \quad (8)$$

  where K is the number of parameters in a model and L is the log likely hood estimate of the model

- **RMSE**: The Root Mean Square Error (RMSE) measures the difference between the true values of a data set and the predicted values of the model (residuals). It is a use-full assessment to assess how well a model fits a data set at face value without accounting for the number of parameters in a model. The formula used to calculate (RMSE) is

  $$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(yest_i - yact_i)^2}{N - P}} \quad (9)$$

  where *yest* is the estimated parameter of the model for observation i, *yobs* is the observed value for observation i, N is the number of observations, and P is the number of parameter estimates, including the constant.

- **VIF** The Variance Inflation Factor (VIF) is a tool used to understand the degree of co-linearity of a models parameters. The smaller the VIF, the less co-linear the variables of a model are. A model parameter is considered to have an acceptably low amount co-linearity if the VIF is less then five. The VIF can be calculated with the following formula

  $$VIF_i = \frac{1}{1 - R_i^2} \quad (10)$$

  where R_i^2 is the adjusted coefficient of determination for variable i in a regression model.

- R^2: The Coefficient Of Determination, or R^2 statistic, tells us how well a variation in the independent variables explains variation in the dependent variable. It can be used to dictate how well a model explains patterns in the data BUT not how well a model fits the data. The R^2 ranges between zero and one. An R^2 score of one means that variation of the dependents is fully explained by the values of the independent while and R^2 score of zero states the contrary. This means the higher the R^2 score, the better. The R^2 score can be calculated by the following formula

$$R^2 = 1 - \frac{RSS}{TSS} \tag{11}$$

where RSS is the sum of squares of the model residuals and TSS is the total sum of squares of the model.

- **Adjusted** R^2: This can simply be described as the Coefficient Of Determination for higher order models. The adjusted R^2 statistic takes into account the number of model parameters and only raises its value if additional parameters within model is significant as apposed to the basic R^2 statistic which will increase with the addition of parameters. This can be calculated with the following formula

$$R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1} \tag{12}$$

where R^2 is the un-adjusted coefficient of determination, N is the total sample seize, and p is the number of predictors.

We calculated all parameters for each of our models and we ended up and we get the following values (Table 1).

Model	AIC	RMSE	MAX Variable VIF	R^2	Adjusted R^2
1	8548.43	0.29	2.39	0.16	0.16
2	9905.85	0.29	1.84	0.15	0.15
3	8580.29	0.29	1.72	0.16	0.16
4	8548.43	0.29	2.39	0.16	0.16
5	9299.33	0.29	1.64	0.13	0.13

Table 1: Model statistics

All statistics indicate that (4) and (1) are the strongest models, but not by much as all of them have very similar statistical scores. Very little of the variation on the dependent parameter of each model is explained by the independent parameters which is signified by a low R^2 and adjusted R^2 scores. This means that non model is very precise in predicting dependent variable values from the independents. Our models do have low co linearity between independent variables with the max variable VIF amongst all models is 2.39 which is less then 5. To make our models more effective, we need to do the following. First. we need to include interactions between the model parameters as methane transmission is directly effected by the interaction of several parameters in our data set. For example, humidity and sunlight both affect the transmission of methane as high humidity and high amounts of sunlight lead to the dissipation of methane emissions, while they are a lot less effective on there own. Second, we need to include negative exponential terms as well within the regression model to account for negative relationships in the data. Finally, we don't need to go up to as high of an order of terms as degree of ten is very steep and may not accurately represent the data, especially if we include negative and interacting terms. Our model should represent data much better if these modifications are included which will give us more significant results and a much stronger model.

4 Predicting Emissions Using Classification Trees

A Classification Tree is a machine learning model that labels, records, and assigns variables into discreet classes. These classes are build by recursive partitioning, which is an iterative process of splitting the data into partition and then splitting it up further on each of its branches. The easiest way to build a model using this method is to feed an algorithm a set of training data, such as the CART algorithm, and then test it with the actual data or the test data.

4.1 Model Assessment

A classification tree can be simply assessed using a confusion matrix. A confusion matrix is a $n * n$ matrix that signifies weather a machine learning model has correctly classified an event. Without loss of generality, the rows of the matrix signify the correct classification of the event while the columns signify the classification of the model. This means that every entry on the matrix diagonal is a correct classification by the event while the entries not on the diagonal are incorrect classifications. The accuracy of the model can then be dictated by

$$acc(A) = \frac{\sum_{i=1}^{n} a_{i,i}}{\sum_{j=1}^{n}(\sum_{i=1}^{n} a_{j,i})} \tag{13}$$

where $a_{j,i}$ and $a_{i,i}$ are elements of confusion matrix A.

4.2 Experimental Setup

We decided to test a classification trees effectiveness by splitting up our data into six groups (Table 2). For each binned group (not RAW), we take into account that the air has a base concentration of methane so we consider every measurement less then 2.1ppm to be a base concentration and bin accordingly. For the raw data, the bins are of width 0.1ppm as this is the maximum precision of the measurements.

Group	Test/Training Split	Bin Width
1	1/9	0.7ppm
2	1/6	0.7ppm
3	1/9	0.4ppm
4	1/6	0.4ppm
5	1/9	RAW
6	1/6	RAW

Table 2: Experimental setup

We train a total of six models for each group using 'ctree' in R 4.0.5, where we very the amount of training data the bin size containing the dependent measurements fed to the algorithm. This was to test models ability to make predictions with more limited amount of data the precision and accuracy that the model can make predictions at. The overall goal was to generate a model with the highest degree of accuracy that could operate effectively with the least amount of training data.

4.3 Results and Analysis

We test each model on there respective test data sets and we get the following accuracy for each group (Table 3). We instantly see that the models fit to the binned groups are significantly more accurate in predicting emissions within their assigned groups. The classification tree models maintain more then 90% accuracy for bin with up to 0.4ppm. We also see that the results are very similar for each training and test data split for each bin group which shows that our model is fairly robust and can be fed less then 90% of the data to be accurate. We can conclude that if we take

Group	Model Accuracy
1	99.8%
2	98.6%
3	91.8%
4	92.1%
5	0%
6	0%

Table 3: Model accuracy

into account a base concentration methane in the air and bin our data into bins of width 0.4ppm or more, our model will be able to make predictions with +90% accuracy.

When looking at the confusion matrices for the binned groups, we also see that even when our model miss-classifies an estimate, the model is not "out to lunch", and makes a estimation close to the actual measurement. We also see that the model is 100% accurate in predicting base concentration cases as well. This means that there is a definite parameter difference between base measurements of methane concentration and increased measurements of methane concentration. Binned confusion matricies are located in the Assignment 2 folder

References

[1] Department of Statistics. (2018). *10.7 Detecting Multicollinearity Using Variance Inflation Factors*. The Pennsylvania State University

[2] FrontlineSolvers inc. (2022). *Classification Trees*. FrontlineSolvers; A leader in analytics for spreadsheets and the web. https://www.solver.com/classification-tree

[3] Wagner, G. (2022). *Notes-Topic 5 Mutable Linear Regression*. University of Alberta

[4] Wagner, G. (2022). *Notes-Topic 3 Inference For Several Populations*. University of Alberta

[5] Wood Buffalo Environmental Association. https://wbea.org/data/network-map-station-data/

[6] Zach. 2021, *What is Considered a Good AIC score*, Statology. https://www.statology.org/what-is-a-good-aic-value/