Zackery Bradley

BAN 530

20220501

**Milestone 7- Final Project Write-up with Decision Analysis**

*Subjects of Discussion*

1. ***How do you know?***
2. ***Why is it happening?***
3. ***Why do we care?***
4. ***What is your recommendation?***
5. ***Focus of the Scope***
6. ***What are the implications of your recommendation?***
7. ***Where are there areas of concern with this Recommendation and why?***
8. *How do you know?*

Upon examination, Recent studies taken in Thiruvananthapuram, Kerala show a need for air quality improvement. This can be seen by looking at the chart below, displaying the average AQI per quarter for the years 2017 and 2018. Within the last two years, there is only one quarter where AQI levels in Thiruvananthapuram were considered to have minimal impact on citizen health. Levels have predominantly been in the “satisfactory” level of the AQI scale, meaning that air quality may have cause minor breathing discomfort in sensitive people. These levels do not normally pose risk; however, prolonged exposure has certaintly had substantial impacts on human health. Recent analysis throughout India have emphasized the importance of air quality, as life expectancy is thought to be shortened by an average of 5.9 years, and as much as 9.7 years in more densely populated areas where air quality is minimal. This level of air quality is similar to smoking regularly for thirty years. Aside from premature death, studies have shown that poor air quality also carries a host of other health issues. Issues include things such as asthma attacks, cardiovascular diseases, and lung cancer. Poor air quality is also known to stunt lung development in children, harming their health as children while also reducing their lung capacity as adults. These development issues can further impact health, as reduced lung capacity can affect one’s ability to successfully fight off infections and increase symptoms such as wheezing, coughing, and shortness of breath. Ultimately, air quality not only affects the quality of our lives, but further impacts our wallets in areas such as the need for medical costs, and specialized treatments,

*Thiruvananthapuram City Dataset*



*Thiruvananthapuram Day Dataset*

1. *Why is it happening?*

There are many reasons why air quality is subpar in Thiruvananthapuram. Various combustion sites locate throughout the city with much of the locations producing heavy amounts of PM2.5 and PM10 matter have been the main contributing factors into poor air quality. These combustion sites have made releasing a large number of chemical compounds, such as CO, NO, and NH3 into the atmosphere, which all affect air quality detrimentally. These chemical compounds also have a trickle-down effect, causing substantial harm to things such as wildlife, and vegetation, highlighting how much impact human activity can affect the air around them.

1. *Why do we care?*

Earlier, we mentioned that the air quality in India has largely affected the quality of vegetation and wildlife. Within the last ten years, America has increasingly imported goods from India. As recent as 2019. The U.S. spent $2.6 billion on agricultural products stemming from crops and livestock from India. This makes India the 14th largest supplier of agricultural goods in America. Overall, India is the 10th largest supplier of goods imported to the United States. These goods have been produced in areas with dangerous levels of air quality, affecting the health of all those who utilize these products. Furthermore, these products are often placed besides, or even mixed with products from other countries, spreading the particulate matter that may be on these products and cross contaminating a majority of what we use in America today. Even organic material imported from India could subsequently negatively affect the health of consumers, as it is nearly impossible to track exactly where certain products have come from or been in contact with.

1. *What is your recommendation?*

As India’s air quality problem has been a growing concern for years, it is believed that one approach may not be a suitable approach. With this taken into consideration, we will be breaking down a series of recommendations that the country, and its citizens can take to improving air quality both inside and outside of the home.

***Breakdown of Recommendations***

1. ***Plant air quality inducing trees around the country***
2. ***Set stricter home codes in India***
3. ***Implement cleaner ways of transportation***
4. *Plant Air Quality Inducing Trees around the country*

One suggestion that comes to mind when debating how to better India’s air quality issue is to plant trees that are known to be natural filters. Silver Branch and Yew trees are known to be the most effective at capturing particle matter, some suggesting that that reduce particles as much as 79% and 71% respectively. By planting these widespread, this could potentially be a cheap, yet long-term solution to the air quality crisis in India. Additionally, conifers such as pines and cypresses also pose as great natural purifiers. These trees are thought to be particularly good at reduce air pollution in cities, where AQI levels are often heightened. The conifers success in reducing particulate matter comes down to its unique structure, boasting a dense canopy of needle-like leaves, making it remarkably effective in trapping air pollutants. Furthermore, conifers do not normally lose their leaves, making them an excellent choice to reduce air pollutants year-round. The same could be repeated with plants, as the country could plant natural air purifiers such as ferns and lilies to further reduce harmful pollutants.

1. *Set Stricter Home Codes in India*

Inside the home, citizens can also play a vital role in reducing air pollution. Implementing stricter codes inside the home, such as requiring citizens to install oven vents and exhaust hoods in their homes could have detrimental effects on reducing the amount of particulate matter in the common household. Gas stoves are particularly known for releasing harmful substances such as carbon monoxide and nitrogen dioxide regularly. These gases regularly escape and can easily be absorbed in one’s bloodstream. By installing kitchen vents and exhaust fans, these gases are further broken down, greatly reducing the impacts of these products on human health.

1. *Implement Cleaner Ways of Transportation*

With an estimated population of 1.4 billion, it can be argued that transportation is likely one of the biggest culprits of India’s air quality problem. While technology on making cleaner more efficient vehicles is growing, there are several things India can do today that would drastically reduce the number of pollutants in the air. One of these ways is by implementing a cable car system for its citizens to travel with. Creating something like this would have substantial benefits such as reducing the number of vehicles in India and could also be profitable as the government could sell passes to these cable cars like they do with bus transportation. Furthermore, this would free up traffic and substantially decrease vehicle idling time, which is known to be one of the worst factors in air pollution.

1. *Focus of the Scope*

Specifically, our major focus was targeted on reducing the air quality in Thiruvananthapuram, Kerala. Several methods were used in the process with a focus on obtaining the best performing model on the dataset. These strategies are listed below:

*Scope of methodology*

1. **Standard Least Squares (OLS)**
2. **Random Forest**
3. **Boosted Tree**
4. **Booster NN**

*Summary of strategies*

1. *OLS*

OLS is typically used as a method for estimating the parameters that are unknown in the model. This method helps us to find the relationships between independent and dependent variables. Stepwise helps us further locate our predictors by adding and dropping one variable at a time in the variable selection method. By employing these two methods, we will be able to determine key factors in our analysis with accuracy. OLS can be described as the “workhorse” method and is a standard operating tool in statistics analysis. On the contrary, OLS tends to produce estimates with a great deal of variance regarding big data. In all, OLS typically generates poor forecasts.

1. *Random Forest*

Random Forest is based on an algorithm-based ensemble decision tree method, which was first developed by Breiman (1996, 2001). This algorithm is considered to be generously user-friendly, as it consists of only two parameters, consisting of the number of variables in the random subset at each node, and the number of trees in the forest. Since its introduction, Random Forest methods are usually highly desired in the statistic world and its creation is considered a modern innovation. Random Forest operates by selecting distinct relevant variables for prediction, without the need to rely on any functional or distributional assumptions Random Forest functions by forming a multitude of decision trees, all of which are independent of each other. Each node in the tree is then split using the best among a subset of predictors that have been randomly chosen at that node. By calculating each tree in random fashion, the Random Forest eliminates the criticisms of decision tree methodology in which individual trees have been known to be high correlated. This functionality gives Random Forest methods immense versatility and flexibility, as it is considered to work well with both categorical and continuous variables. This method is much different from previously explored methods such as penalized regression models, as Random Forest is able to capture nonlinearities in data. In addition to our brief analysis, Random Forest has many additional advantages. One advantage is that Random Forest can automatically manage missing values in the data, as well as being uniquely robust to outliers, as the method oversees them autonomously. However, Random Forest methods do come with some disadvantages as well. One disadvantage worth nothing is that Random Forest usually is a tedious process, as computing the data often takes more time as compared to other methods.

1. *Boosted Tree*

Boosted Tree models are built sequentially, where at each step of the process the model learns from its previous errors. They begin by estimating a model and obtaining residuals in the case of a regression, and misclassification rate in case of classification. This model gives observations with the biggest errors additional weight as opposed to others. The goal for this particular model is to continuously improve model fit by correcting the bad fits given from the previous steps. After eliminating bad fits, Boosted Trees are then averaged to produce a prediction or observation from each observation. By averaging, it is implied that predictions or classifications will not be unstable. Boosted Trees are considered highly efficient on both classification and regression models, and typically provide more accurate predictions than that of a Random Forest. In opposite, Boosted Trees are overly sensitive to outliers, and also tend to overfit data is too many trees are used.

1. *Boosted NN*

Boosted NN models follow this same methodology and is considered a relatively recent and widely popular method in machine learning. Boosting models are built in sequence, learning from its error within each step of the process. Boosting begins by estimating a model and obtaining residuals in the case of a regression and misclassification in the class of classification. From there, observations with the largest errors (where the model performed poorly) are given additional weight. The goal of this is that with each additional step the model fit is improved by correcting the subpar fit of the previous steps. Next, the estimates from these models are averaged to produce a prediction, or classification for each observation. NN also comes with its own list of advantages and shortcomings. As previously mentioned, NN typically has good predictive ability and is capable of capturing complex relationships. However, NN has no variable selection mechanism, so utilizing JMP’s variable importance is vital when identifying the most important variables in the data. Typically, NN is referred to as a “black box” method, meaning that findings are often difficult to interpret.

*Explanation of dataset*

Our dataset consists of information relating to air quality of differing locations in India. Our response variable consists of AQI, which is measured by calculated by using a combination of seven unique variables and are as follows: PM2.5, PM10, NO, NO2, NOx, NH3, CO, S02, and O3 respectfully. Ultimately, through this analysis we are attempting to forecast how AQI levels affect the health of citizens in Thiruvananthapuram, Kerala. Given the data, there are various points of interests that we must identify prior to our analysis. Two variables in particular, PM2.5 and PM10 are of the upmost interest. PM2.5 (fine particulate matter) is an air pollutant that is a concern for people’s health when levels in the air quality reach elevated levels. PM2.5 levels are most likely to be elevated on days where there is little to no wind to break up unhealthy pollutants. PM2.5 is typically measured from 2.5 micrometers and smaller, hence the name. For comparison, the average human hair is about seventy micrometers in diameter – 30 times larger than the largest PM2.5 particle. PM2.5 comes from sources both indoor and outdoor sources. Outside, these fine particles primarily come from vehicle exhausts or other burning fuels such as wood, coal, and natural sources such as forest and grass fires. Indoor sources are made up of tobacco smoke, cooking, burning candles, and operating fireplaces and kerosene heaters. PM10 (particulate matter) similar to PM2.5 except that they are generally measured as ten micrometers and smaller. PM10 has many sources such as dust from construction sites, landfills and agriculture, wildfires and brush/waste burning, wind-blown dust from open lands, pollen, and fragments of bacteria. Both PM2.5 and PM10 have proven to have adverse health effects when prolonged exposure occurs. These health effects include premature mortality, acute or chronic bronchitis, asthma attacks, respiratory symptoms, and several other heart and lung diseases.

*Results for City Dataset*

When examining the data post analysis, our Bootstrap Forest method proved to be superior as opposed to other methodology. Backed with an RSquare of 0.7032, in conjunction with a Rase and AAE of 16.835 and 10.987. Given the results of the model comparison, Bootstrap Forest will be the selected method for this analysis.

| **Validation** | **Predictor** | **Creator** |  | **RSquare** | **RASE** | **AAE** | **Freq** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Test | AQI Predictor DTM | Partition |  | 0.6689 | 17.782 | 11.557 | 1161 |
| Test | AQI Predictor BFM | Bootstrap Forest |  | 0.7032 | 16.835 | 10.987 | 1161 |
| Test | AQI Predictor BTM | Boosted Tree |  | 0.6763 | 17.582 | 11.956 | 1161 |
| Test | Predicted AQI BNN | Neural |  | 0.3619 | 24.688 | 16.873 | 1161 |
| Test | Pred Formula AQI OLS | Fit Least Squares |  | 0.0013 | 30.885 | 22.744 | 1161 |

*Variable Importance*

After analyzing our initial findings, we dove deeper into our data by examining which variables had the most significant impact on model importance. Surprisingly, the biggest impact came from the date variable, as it displayed a main effect of 0.763. Our second most important variable was 02, with a total effect of 0.389. Lastly, our third variable turned out to be hour, having a total effect of 0.027.

**AQI Predictor BFM**

| **Column** | **Main Effect** | **Total Effect** |  |
| --- | --- | --- | --- |
| Date | 0.592 | 0.763 |  |
| O3 | 0.22 | 0.389 |  |
| Hour | 0.009 | 0.027 |  |
| PM10 | 0.001 | 0.003 |  |
| CO | 0.001 | 0.002 |  |
| PM2.5 | 0.001 | 0.002 |  |
| NH3 | 4e-4 | 0.001 |  |
| SO2 | 4e-4 | 0.001 |  |
| NO2 | 2e-4 | 0.001 |  |
| NOx | 1e-4 | 4e-4 |  |
| NO | 8e-5 | 3e-4 |  |
| City | 8e-18 | 2e-17 |  |

*Results for Day Dataset*

After running the various methods briefly mentioned above, once again our Bootstrap Forest method proved to be the most significant. Sporting an RSquare of 0.6315 followed by a RASE and AAE of 20.342 and 12.305 respectively, the Bootstrap Forest method won out in convincing fashion in model performance.

| **Predictor** | **Validation** | **Creator** |  | **RSquare** | **RASE** | **AAE** | **Freq** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| AQI Predictor DTM | Test | Partition |  | 0.6222 | 20.599 | 12.652 | 187 |
| AQI Predictor BFM | Test | Bootstrap Forest |  | 0.6315 | 20.342 | 12.305 | 187 |
| AQI Predictor BTM | Test | Boosted Tree |  | 0.6299 | 20.386 | 12.283 | 187 |
| Predicted AQI BNN | Test | Neural |  | 0.5815 | 21.679 | 12.708 | 187 |
| Pred Formula AQI OLS | Test | Fit Least Squares |  | 0.1892 | 30.175 | 21.849 | 187 |

*Variable Importance*

After thoroughly sifting through our data, our chosen model displayed which variables were of most importance. Shown in the visual below, we can see that the variables responsible for the biggest impact of air quality. Our initial hypothesis was differed by our results, as CO had a substantial impact on the pricing of diamonds with a total effect of 0.527. NO2 also seemed to be somewhat significant, it measured to have a total effect of 0.349, Our third most important variable was S02, as its total effect came out to 0.153.

Table

Description automatically generated

*Prediction Profiler*

We further explored out data by analyzing the prediction profilers below. We experimented by placing our most significant variable, CO on the lowest scale and then again on the highest scale to examine the relationship between the other variables evaluated in our dataset. From our experiment, we can see that all of the variables measured seem to have no significant correlation in to each other, meaning that the variables are all predominantly independent amongst the data.

*Prediction Profiler for CO at Minimum*



*Prediction Profiled for CO at Maximum*



1. *What are the Implications with your recommended path?*

Regarding the recommendations above, one could ultimately state that the most effective form of reducing air quality would be the citizens themselves. The air quality problem in India is not going to going aware by the effort of a few people; it has to be collectively tackled as a community with the ideology of making their country a better and safer place to live. With such a massive amount population is India, this would require extreme social efforts, however it is not impossible if the majority of citizens are made aware of the toxins they are breathing in every day. With government aid, and the spirit of the people of India involved, it is believed that the air pollution issue in India can be substantially reduced.

1. *What are the areas of concern for this recommendation and why?*

There are many areas of concern that come with the suggested recommendations mentioned above. Planting trees and plants would not be an immediate solution and would likely require a significant amount of time and labor to successfully plant and tend to. Setting stricter codes in the home may also prove to be difficult as many parts of India is extremely poor and may not be able to afford the new regulations put into place. Many citizens would adhere to the new regulations as they may see the new rules as an unnecessary luxury. Lastly, installing cars would be incredibly expensive, and substantially difficult to engineer as far as creating pathways that are most convenient for the citizens of India. Many of these strategies would likely have to involve outside funding from other governments, as well as an increased commitment by the government of India on combating the air quality problem.

*Sources:*

[7 ways to improve indoor air quality and reduce air pollution in your home. (petro.com)](https://www.petro.com/resource-center/improve-indoor-air-quality)

[The best trees to reduce air pollution - BBC Future](https://www.bbc.com/future/article/20200504-which-trees-reduce-air-pollution-best#:~:text=Silver%20birch%2C%20yew%20and%20elder,still%20captured%20a%20respectable%2032%25.)

[Actions You Can Take to Reduce Air Pollution | Ground-level Ozone | New England | US EPA](https://www3.epa.gov/region1/airquality/reducepollution.html)

[India Population (2022) - Worldometer (worldometers.info)](https://www.worldometers.info/world-population/india-population/)

[Fine Particles (PM 2.5) Questions and Answers (ny.gov)](https://www.health.ny.gov/environmental/indoors/air/pmq_a.htm#:~:text=Fine%20particulate%20matter%20(PM2.5,hazy%20when%20levels%20are%20elevated.)

[Particulate Matter (PM) Basics | US EPA](https://www.epa.gov/pm-pollution/particulate-matter-pm-basics)

[Inhalable Particulate Matter and Health (PM2.5 and PM10) | California Air Resources Board](https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health#:~:text=PM10%20also%20includes%20dust%20from,pollen%20and%20fragments%20of%20bacteria.)