Amway Data Science Internship:

Analysis of Indoor Air Filtration System and Indoor Air Quality Trends

Samantha Milano

Data Science and Analytics

Grand Valley State University 2019

Introduction

Amway is a direct-selling company that produces goods for home, health, and beauty. They make over 450 products that people use around the world every day. Amway currently operates in over 100 countries and territories. Founded in 1959 by Jay Van Andel and Richard DeVos, Amway has spent decades investing in Grand Rapids and west Michigan.

Background

The new Atmosphere air filtration unit, Sky, had been released in the Americas, and planned to be released in other regions such as China, Japan, and Korea. In preparation for this, my project focused on exploring the relationship between indoor and outdoor air quality to better identify what factors influence indoor air quality and what areas would have the highest need for an air filtration system. Therefore, the main questions we attempted to answer using the Sky telemetry data were:

- What is the relationship between indoor and outdoor air quality?
- Can we identify the factors with the largest impact on indoor air quality?

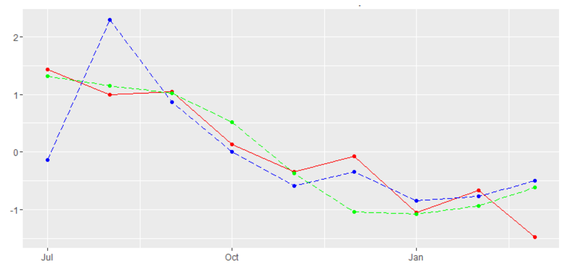
Methodology

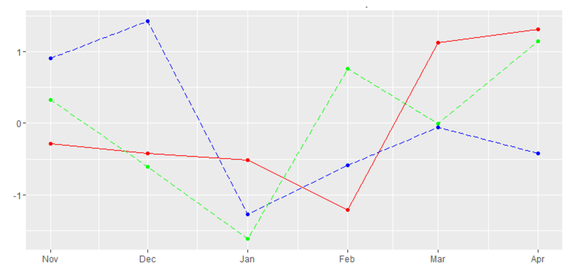
The data was retrieved from the data warehouse Amazon Redshift using SQL. The rest of the analysis was completed using R.

We used the data collected by the Sky sensors to represent the indoor air quality, and AQI to represent outdoor air quality. We brought in outside data sources to supplement information regarding house data, area data, and environmental factors.

We wanted to use a regression model in order to identify what factors were most influential. The response variable summarizes the difference between the indoor air quality trends and outdoor air quality trends, which represents how well the indoor and outdoor trends follow each other. The larger the response value, the more predictable the indoor air quality; the smaller the response value, the less predictable the indoor air quality.

Air Quality Index Levels of Health Concern	Numeric Value	Meaning
Good	0 to 50	Air quality is considered satisfactory, and air pollution posses little or not risk.
Moderate	51 to 100	Air quality is acceptable; however, for some pollutatnt there may be a moderate health conern for a very small number of people who are unusually sensitve to air pollution.
Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general publis is not likely to be affected.
Unhealthy	151 to 200	Everyone may begin to experience health effects; members of sensitive groups may experience more serious helth effects.
Very Unhealthy	201 to 300	Health warnings of emergency conditions. The entire population is more likely to be affected.
Hazerdous	301 to 500	Health alert: Everyone may experience more serious health effects





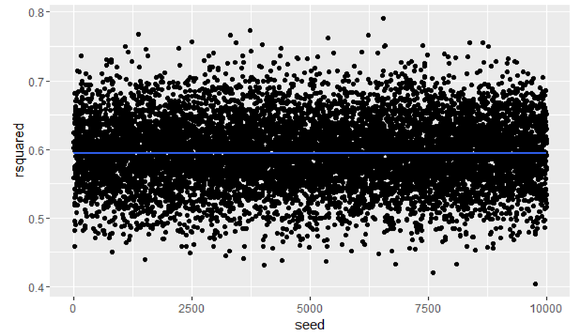
Results

Using forward, backward, and stepwise model selection techniques, we ultimately decided on a model that contained three of the factors from the input variables. For interpretation, the main take-away from the process is that the model was able to identify what factors were useful (ie. in the final model) out of all of the possible factors input to the model selection process.

Factor 1
+
Factor 2 = strength of relationship between indoor and outdoor air quality
+
Factor 3

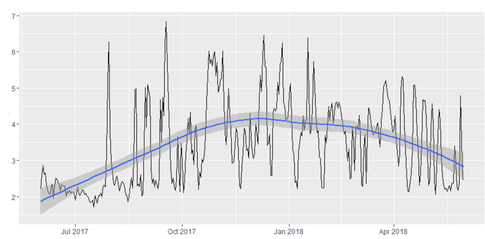
Validation

In order to validate the model, we used cross-validation to test the model against existing data. Since the data set had large variation, there was a large difference in the r-squared value depending on how the data was split and which observations went towards the training set or the test set. To account for this, we ran a 10,000 trial simulation of different splits of the data in order to find the average r-squared value, which was approximately 0.6.



Implementation

Knowing the factors that most influence the relationship between indoor and outdoor air quality can give useful insights to regions that may have a need for indoor air filtration systems. Identifying these places where the relationship between indoor and outdoor air quality is very strong, we can observe historical AQI patterns to predict future trends. For example, using data from a city in the target regions, we know that there is an increase is AQI between October and February. Since the model tells us that there is a strong relationship between indoor and outdoor air quality here, we can extrapolate that indoor air quality will follow the same pattern. Thus, these months would be the optimal time that individuals would have the highest need for an indoor air filtration system.



Acknowledgements

I want to personally thank my supervisor, Josh Schwannecke, and my teammates, Cody Dean and Leslie Walcott, as well as the entire Global Discovery team. Without them, the project would not have been a success and I would not have gained the valuable experience I had this summer.