Model Building Part 2

Team 5

As carbon emissions remain the primary instigator of climate change, methods to predict the amount of greenhouse gas emissions emitted by cities and individual buildings are influential in the design of civil structures. By researching what aspects of a building's infrastructure correlate most to overall greenhouse gas emissions, potential actions to create more energy-efficient systems or discourage excessive carbon pollution will become more feasible.

Our team's main objective is to determine what features of a building influence pollution levels, specifically, their greenhouse gas emission intensity measured by total GHG emissions divided by total area. Additionally, we aim to predict a building's GHG emission intensity given certain attributes. In the previous assignment, our team implemented a decision tree, naive Bayes classification, and quantile regression. The accuracy outcomes of these models led the team to test a neural network in hopes of results that showed a decreased level of overfitting, as well as an overall reduction in noise in the model. Our decision to test a model with both classification and regression stems from our desire to discover which type of prediction would produce more accurate results.

In this most recent round of model improvements, we refine the data that is utilized by the decision tree through Principal Component Analysis (PCA), and apply hyperparameter tuning within our existing decision tree in order to increase accuracy. Previous successes with the decision tree led our investigation to explore several means of improving the overall accuracy of our original model, and we saw an increase of 12.50% as a result of these efforts.

Discrete Prediction Over Continuous Prediction

Based on the previous Model Building Part 1 results, our team decided to continue improving a model that predicts a categorical variable opposed to a numerical variable. This is because the response variable we wish to predict, GHGEmissionIntensity, is highly skewed causing much noise. A sizable amount of data is held in these higher ends of the range, so removing these points may lose some valuable information. Additionally, none of the scatter plots made in the exploratory phase revealed a linear relationship between GHG intensity and other features. This means a linear regression model would not be well-suited for this data.

Choice of Decision Tree

Since no linear relationship seems to exist, logistic regression is also not a good fit. However, decision trees do not assume a linear relationship between the independent and the response variables. Decision trees are also easy to interpret which can provide some insight as to what features are most influential to categorizing GHG intensity as low, medium, high, and extremely high. As our preliminary decision tree we made in the exploratory phase provided a decent accuracy of 75%, this model provides a solid foundation to continue improvements on.

Choice of DBSCAN Clustering

To convert the numerical values of GHGEmissionIntensity to a categorical value, we decided to use DBSCAN. Other techniques to convert continuous to discrete values such as equal width binning would not partition the data into equally sized classes due to the heavy skewed nature of the spread while frequency binning would cause the highest GHG intensity group to span a large range and lower GHG intensity group to be extremely small in range. Because of this, DBSCAN provides a middle ground between these two extremes. DBSCAN is useful to create clusters with differing sizes and when there are many outliers. It is a density based algorithm, so each cluster has a similar density. Unlike k-means, DBSCAN does not require a number of clusters, instead automatically creating an appropriate number of clusters (in this case three). The remaining outliers are then formed to be a fourth class called extreme high. Due to the response variable being highly skewed with severe outliers, predicting a numerical value becomes difficult. Because of this, we determined that the best way to predict the emissions of a building was to predict what tier of emissions the building would be in. From the DBSCAN we were able to sort the data into 4 clusters, low, medium, high, and extreme high. This extreme high cluster perfectly accounts for the intense magnitude of the outliers, and gives us a great base from which to build our future models.

Choice of PCA for New Decision Tree

The decision tree made in the exploratory phase used the techniques explained so far. To possibly improve the results, we used PCA to reduce the dimensions to 3 PCs. PCA serves two purposes: reducing the dimension and indicating what features influence most of the variance (via eigenvector entries). With PCA, we can use the top 3 PCs to rerun the decision tree and compare results with the tree from the exploratory phase.

We performed PCA on all numerical variables excluding categorical variables that were one-hot encoded as these variables decreased the accuracy of the tree immensely as evident in the previous model building phase. The resulting principle components revealed that the top 3 PCs captured nearly 100% of the variance. Hence, we used these three to feed into the decision tree. Once the new tree with PCs was made, it gave the same accuracy as the tree before in the exploratory phase. These matching accuracies suggests that our PCA did a good job of capturing the variance of the original dataset, but it didn't necessarily improve it.

Revising Decision Tree Using PC Top Features

Unfortunately, it is harder to interpret the previous tree when the nodes are splitting on PCs. Because of this loss of interpretability, we made a new tree using the top 5 features in the first principle component. These 5 features are electricity used, the amount of steam used, the amount of natural gas used, the gross floor area of the building, and the source energy use intensity. Doing so proved to be a meaningful improvement as the accuracy of the tree jumped to 86.7% - nearly 10% improvement.

Besides improving our model, which helps directly with our goal of being able to predict a buildings emissions intensity given its features, this step of using top 5 features from PCA

helped us with answering a central question of ours of which attributes of a building are most influential when it comes to pollution. The PCA specifically tells us the most influential features in greenhouse gas emissions, giving us the answer to our question. From our analysis we found that the five most influential attributes are the amount of electricity used, the amount of steam used, the amount of natural gas used, the gross floor area of the building, and the source energy use intensity. Going forward, we now have an answer to our question of which factors contribute the most to a building's emissions, and have a strong understanding of which factors to prioritize in our model.

Hyperparameter Tuning

We tested using more advanced models like random forest, but found that they were inefficient, and prone to overfitting. From this, we concluded that a decision tree was the right level of complexity to model our data. With this in mind, the way to improve our model was to improve our decision tree and maximize its accuracy. At this point we considered different ways to improve our model, like pre-pruning and post-pruning, but these methods can lower interpretability and increase the computational resources needed to run the model, and work best for improving overfitting, which was not a significant issue we were running into. After further analysis, we decided that hyperparameter tuning was a worthwhile strategy . We wanted to be mindful of overfitting, so we also employed 10-fold cross-validation to ensure that we were maximizing the accuracy for the model as a whole, rather than the model for one split of training data.

Hyperparameter Tuning Tree Results

The Hyperparameter tuning was performed using the Scikit-learn grid search function. The grid search was set up to alter max depth, minimum sample split, minimum samples per leaf, maximum leaf nodes, and ccp alpha which is a cost complexity parameter. While a number of parameters were tested, only a select few being different from the default value led to an increase in model accuracy. These were setting the max depth to be 10 which limits the layer of the tree to 10 plus the root node. Max Leaf Nodes were set to 90 which limits the tree to only include 90 or fewer of the leaf nodes with the best reduction in impurity. The minimum samples to form a leaf was left at the default value of 1. The minimum samples for a split was set at 15 which limits the tree growth by requiring at least 15 samples to form a split to form a split. Finally the ccp alpha value was set at 0.001. This value controls the Minimal Cost Complexity Pruning algorithm which helps prevent overfitting of the tree. Higher values of alpha create smaller, more “pruned” decision trees. The selected value is very close to zero which allows the tree to grow large while still retaining predictive power on unseen data. With these selected parameters, we saw a somewhat small increase in accuracy. This is to be expected with hyperparameter tuning, so we considered this to be a successful improvement of the model, and our best iteration yet.

Conclusion

From these three models, several observations can be made that help answer our research question. The first is that the inclusion of categorical variables do not help predict GHG intensity as seen in the tree made in Model Building Part 1. The second is that most of the variance of the numerical data can be summarized in 3 principle components. These 3 PCs do an excellent job of reproducing the same results as the decision tree using these 3 PCs have the same accuracy as the tree made in the exploratory phase. Lastly, the most influential features that predict GHG intensity are the amount of electricity used, the amount of steam used, the amount of natural gas used, the gross floor area of the building, and the source energy use intensity. This implies that building features like the year built and the number of floors are not important in determining the levels of GHG emission intensity. This last observation is possibly the most useful towards our research question as it explicitly states what attributes will contribute most to GHG. With our final iteration of the tree yielding 87.71% accuracy, we feel this model is sufficiently performant at determining what features of a building influence GHG emissions