Eco-Car Final Paper

Aryan Bhatnagar

Abstract:

What would be the best eco-friendly car brand that would fit a person's criteria? We are trying to program software that would help find a person the best car for them while it being friendly toward the environment. This problem is important because a lot of people can't figure out what the best car for them is and even if they find a car, it could be one that could cause a lot of CO2 emissions. Our program allows the person to have the best of two worlds: To find a database with cars that have all the specs on them with the data on the CO2 emissions and then try to convert that data into a way where the person could choose what car would best fit their criteria. We found out that by using machine learning, we could achieve that goal. Using a random forest classifier, we were able to find the accuracy of the program and also figure out that the database was too big to fit all the criteria.

Introduction:

What would be the best eco-friendly car brand that would fit a person's criteria? This is the goal of this project. We are trying to program software that would help find a person the best car for them while it is still being eco-friendly. CO2 emissions produced by cars can seriously impact the environment on a large scale. To prevent that, using the best eco-friendly car that fits all a person's needs can be beneficial to the environment. To do this, we needed a lot of different types of numerical data including Engine Size(L), Cylinders, Fuel Consumption City (L/100 km), Fuel Consumption Hwy (L/100 km), Fuel Consumption Comb (L/100 km), Fuel Consumption, Comb (mpg), CO2 Emissions(g/km), Vehicle_numerical, Transmission_numerical, and FuelType_numerical. To best implement this data, we came up with a variety of algorithms to use, some of them including RandomForestClassifier, MLPClassifier, Scatterplot, LabelEncoder, X train, and Y train. Though not similar, there have

been researches done on a car's CO2 emissions with machine learning. Sites like Hypi.io and dl.acm.org helped me understand what we should try to focus on with my project. The sites talk about how machine learning can be implemented to predict and test CO2 levels from gasoline cars. It gave us a chance to look at our problems in different ways and helped us progress in our research.

Dataset:

A dataset is a collection of data where every column of a table represents a particular variable and each row corresponds to a given record of the dataset. In this project, we used a dataset from a reliable website called Kaggle. The website had a lot of great resources and data on our topic. After looking through the databases, we found one that would best fit the project's needs. The dataset we found can be accessed here. This dataset provided us with a list of the most gasoline cars ever built while giving us detailed research on each car's CO2 emissions. By typing a command called df.shape, we were given the exact columns and rows the dataset possessed. The set had 12 rows and 7385 columns in total filled with valuable information. The 12 rows had: Make, Model, Vehicle Class, Engine Size(L) Cylinders, Transmission, Fuel Type, Fuel Consumption City (L/100 km), Fuel Consumption Hwy (L/100 km), Fuel Consumption Comb (L/100 km), Fuel Consumption Comb (mpg) and CO2 Emissions(g/km). Since some of the rows had information like the model of a car, we had to change it to a numerical value. We changed it to a numerical value because a program can better understand numbers than letters, giving it a little bit less load. We made some of our values numerical by using a code called a label encoder, which can assign different numbers based on the information. After testing the dataset a bit, we ran into a problem. The dataset was a bit too big to carry out some of the tasks, so we had to optimize it by only including the necessary rows like the fuel consumption, model, Co2 consumption, etc. From the dataset, we could also see that this data would be a classification, since each and every column of the dataset was unique, due to the cars having different manufacturers

and parts. In the end, we optimized everything that we could on the dataset to give us the best results.

Methodology/Models:

To further research our question, we needed to use machine learning to enhance our research. The first step to do that was to look back at our database. As we told you before, for the program to fully process the dataset, the world data (alphabetical data) had to be converted to numerical data. After we did that and optimized our dataset, we had to start dividing our data into two categories: The x-train and the y-train. Now you might be wondering what is x-train and y-train. Well, they are kind of like training data sets. The x-train absorbs about 20% of the actual data and the y-train tries to predict the data by using the other 80% of the data it has. Though it might sound simple, it is very complicated. Since our dataset had unique numbers for each row and column, we had to resolve that issue by using classifiers called the random forest classifier and the mlp classifier. A random forest is a meta-estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. A Multilayer perceptron (MLP) is a technique of feed-forward artificial neural networks using a backpropagation learning method to classify the target variable used for supervised learning. Now, why did we use classifiers instead of regressions? The key distinction between Classification vs Regression algorithms is Regression algorithms are used to determine continuous values such as age or price. Classification algorithms are used to forecast or classify distinct values such as a different type of car. This is why we settled on using the MLP and the Random Forest Classifiers.

Results and Discussion:

After we got our x-train and y-train setup, there was only one thing to do and that was to start the comparison and testing between the MLP and the Random Forest Classifiers. Using the help of sklearn, we were easily able to import the classifiers to their full potential. After testing out only the accuracy of the predictions, we found out that the Random Forest Classifier was averaging about 37% and the MLP Classifier was averaging about 58%. Now to get these results we did have to change some of the data in the dataset, first of all, what we learned by using these classifiers for the first time was that they don't have enough capacity to handle such a bit dataset, so what we had to do was merge some of the data together into in group. We merged the Make and the Model of the car together, so we could get a higher and faster result. After carefully testing the model over and over again, we came up with some results. We found out that the classifiers were taking a long time to process and this was due to the fact that these classifiers actually go through the individual data, meaning it would take them longer to process everything. The second thing we found was that the results of the test were quite low. Though it seems like a failure, in our case it isn't. From the beginning, our dataset has just been too big for our code to handle, but due to it being so big, we were able to utilize some unique classifiers to fix that problem. Out of the two classifiers, the MLP classifier was the most successful at predicting the data and this was due to the fact that unlike the random forest, the MLP classifier can process information much faster due to the fact that it is not creating branches for each column.The MLP classifier is able to use its fast multiple layers to predict most things right and in this case, we got close to a 60% prediction rate. In the future, we can try and find a smaller database that fits all our needs. This database and other databases in the future will have the same problem, but we think we can solve it by trying other classifiers that would better fit the database.

Conclusion:

What would be the best eco-friendly car brand that would fit a person's criteria? We started with this question at the beginning of this research, and now we have finally reached the end. We have discovered methods like the rainforest and the mlp classifier to get the best result possible with our database, we have found how to optimize everything to its full potential. In our opinion, even though the model has a percentage between 40%-60%, we would say that it is a massive success, because of all the different values, the classifier could not predict everything right, but it still managed to outperform most models of this size. In the future, we would try to create a simple program that can be released to the public, that would help guide them on a way to become eco-friendly.

Acknowledgments:

In this project, we couldn't do this without my mentor, Matan Gans guiding me every step of the way. As well as my parents, who funded this whole program.

References:

https://dl.acm.org/doi/full/10.1145/3485128

https://iopscience.iop.org/article/10.1088/1748-9326/ab4e55/meta

https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles?resource=download

https://scikit-learn.org/stable/index.html

https://www.ibm.com/cloud/learn/machine-learning

https://www.smarten.com/blog/multilayer-perceptron-classifier-enterprise-analysis/