**Analyzing fuel economy based on automobile features over time**

**IST718: Big Data Analytics**

**Group 12**

**Matt McDonnell, Ji Ma, Zhongwei Chen**

Contents

[1. Abstract 3](#_Toc57569253)

[1.1 High-Level Overview 3](#_Toc57569254)

[1.2 Prediction List 3](#_Toc57569255)

[1.3 Inference List 3](#_Toc57569256)

[1.4 Conclusion Summary 4](#_Toc57569257)

[1.5 Other Goals 4](#_Toc57569258)

[2. Data Collection, Cleaning, and Exploration 4](#_Toc57569259)

[2.1 Expanded Dataset Description and Sourcing 4](#_Toc57569260)

[2.2 Describe the Data Exploration Results, Visualizations, and Details for Training Models 5](#_Toc57569261)

[3. Methodology 5](#_Toc57569262)

[4. Models 6](#_Toc57569263)

[4.1 Linear Regression Predictive Model 7](#_Toc57569264)

[4.2 Random Forest Predictive Model 9](#_Toc57569265)

[4.3 K-Means Cluster Analysis (PCA) for Recommendation system 11](#_Toc57569266)

[5. Conclusion 14](#_Toc57569267)

[Appendix 15](#_Toc57569268)

**1. Abstract**

**1.1 High-Level Overview**

According to the Huffington Post, an average American spends over $2,000 a year on gas for their vehicle; this is a considerable amount for many Americans considering there are other costs associated with driving, such as insurance, registration, and tax. Through this project, we dive into the variables that are responsible for the annualized fuel cost and CO2 tailpipe emissions of certain vehicles, which is ultimately looking to predict a variable called save/spend. Some variables may point to explaining these in laments terms, but we want to provide the most statistically significant and accurate overview possible to allow for feature pairing and creation in aligning consumers and manufacturers respectively to get ahead of the current trend in whatever feature combination is most suitable. This helps car shoppers and manufacturers benefit from a cost savings perspective in the consumer selection process. It would be beneficial if manufacturers could figure out how to best create cars that are desirable yet lower in fuel cost or tailpipe emission. This appeases shifting macroeconomic trends and more progressive consumer taste in renewable energy-based vehicles.

**1.2 Prediction List**

* Prediction 1 (Cost Savings)

What kind of vehicle features can possibly generate lower costs for customers in 5 years after purchasing the vehicle?

We predict that some of the most likely variables to cause lower costs will be displacement, fuel cost, cylinders, and fuel type.

* Prediction 2 (CO2 Tailpipe Emissions)

What kind of vehicle features lead to high CO2 Tailpipe emissions?

We predict that some of the most likely variables to lower emissions will be consumption barrels, cylinders, and fuel type.

**1.3 Inference List**

* Inference 1 (Cost Savings)

Fuel Cost is the most important variable in weighing the first prediction from our findings in both linear regression and random forest model, which boasts an MSE testing of 1.342 in linear model and 2.3931 using random forest.

1. The coefficient of 9.28 is far greater than the others.

2. The next highest coefficient (fuel type has a coefficient of .1)

Fuel type, number of cylinders, and displacement follow in order of greatest to least weight

Transmission, barrels, and class all weigh on this importantly, too

* Inference 2 (CO2 Tailpipe Emissions)

Fuel Cost is the most important variable in weighing the first prediction from our findings in the regression model, which boasts an MSE testing of 1936.808

1. The coefficient of 9.26 is far greater than the others

2. The next highest coefficient (barrels08 has a coefficient of .08)

Barrels, number of cylinders, and transmission follow in order of greatest to least weight

Make and Displacement all weigh on this importantly, too

**1.4 Conclusion Summary**

Based on the linear prediction model accuracy, we believe we have generated relatively accurate prediction model both using linear regression and random forest. We have successfully created a recommender system using Principle Components and K-Means clustering analysis, which can provide customer a list of similar vehicles to a specific car model.

**1.5 Other Goals**

The main goal was to help consumers, but our analysis can benefit car manufacturers

**2. Data Collection, Cleaning, and Exploration**

**2.1 Expanded Dataset Description and Sourcing**

We collect our data from the csv file provided by the U.S. Department of Energy. The dataset includes fuel economy information for 1984 – current model year vehicles and contains informative attributes such as annual petroleum consumption in barrels (barrels08), number of engine cylinders in each vehicle (Cylinders), tailpipe CO2 in grams/mile (CO2TailpipeGPM), and save/spend over 5 years compared to an average car (youSaveSpend). The data set contains a variety of variables involved in measuring the efficiency of a vehicle in regard to fuel consumption and emissions. In total, there are 42,377 rows and 83 columns. There are many columns that were less relevant in exploring this, so we removed them in order to mitigate for this. The key predictor, save/spend, is formatted as either a positive or negative dollar amount indicating a consumer’s cost basis on the vehicle over time. The variables we will look into are involved in the performance aspects of the engine and mechanically engineered portion of the vehicle. So essentially each variable tells a bit of a story about how the car was engineered and we will tie this into how these features relate to costs and emissions. They all paint a greater picture of which what car each row represents.

**2.2 Describe the Data Exploration Results, Visualizations, and Details for Training Models**

After exploring the data through some visualizations, we noticed some linear relationships (Appendix 1). Most relevant to our purposes were negative linear relationships between the predictor, save/spend, annual petroleum consumption in barrels, emissions, displacement and fuel cost. Additionally, Figure 2.2.1 suggests that emissions have a positive linear relationship with displacement. Figure 2.2.2 also indicates that the number of cylinders and engine displacement in liters looked to have the highest potential to affect fuel cost as well as costs incurred on the vehicle within 5 years’ time, which was the most interesting thing we found.

**Chart, histogram, scatter chart

Description automatically generated**In terms of data preparation process, we found 723 NAs, 904 duplicated records and 0 outliers and chose to remove all missing values and duplicates, considering the dataset has 42,377 records in total. Additionally, we only used a subset of 15 columns from the data that are relevant to our problem statement. Save/spend was also found to be unbalanced in the sense that it has about a tenth of rows as positive compared to negative. If a logistic regression was performed, SMOTE would be needed to up sample positive save/spend. We used PCA to reduce the dimensionality for the k-means training.

EDA Plot 2

**Chart, scatter chart

Description automatically generated**

Figure 2.2.2 EDA plot that shows a linear

relationship between displacement and cost/save

Figure 2.2.1 EDA plot that shows a linear

relationship between displacement and emission

**3. Methodology**

In order to reach our project goals, we first determined the workflow of our project using the display in Figure 3.1. By going over the materials we covered this semester, we decided to use topics and algorithms such as linear regression, supervised random forest, and unsupervised k-means clustering with PCA. These helped us shape a recommendation system for consumers We constructed the fundamental methodologies by training these models and continually tuning the parameters.

We started off by applying the above three algorithms into the training data and using MSE to evaluate model accuracy. Once deemed quality, we used these models to find important vehicles and make inferences on methods that can be used to improve Cost/Savings and pipeline emissions by upgrading features. Finally, we applied our model results to a real-world problem case. For instance, we used a recommender system on Ford and BMW to find vehicles that are similar to them.

**Diagram

Description automatically generated**

Figure 3.1 Project Workflow Diagram

**4. Models**

Our approach focused on performing research analysis on two main target values: cost/saving in 5 years and CO2 emission. We used supervised algorithms such as regression modeling and random forest to find out the most important vehicle features that can be used to predict the variables. We also utilized unsupervised modeling such as k-means clustering to classify vehicles based on their features. This generated a recommendation system based on the PC loadings.

**4.1 Linear Regression Predictive Model**

* Prediction: Cost/Savings

By setting cost/savings in 5 years as our target value, we aimed to find out the significance of each vehicle feature, specifically annual petroleum consumption, engine displacement, fuel cost, number of cylinders, vehicle manufacturer, vehicle model, fuel type, transmission, and vehicle size class. This allowed us to explore which features affect cost/saving most.

We transformed the data to a pipeline to index categorical attributes, assemble attributes vector, and standardize all vectors so that all independent variables have same weight for analysis. The original dataset is randomly split into 70% training data and 30% testing data. We trained the linear regression model by defining Cost/Save as our target label.

When applying linear regression to both training and testing sets, we used Mean Square Error to evaluate model accuracy. Our fit model generates an MSE of 0.739 on the training set and an MSE of 1.342 on the testing set, which indicates that the features we selected for this model are strongly correlated to the cost and spending of a vehicle.

Table 4.1.1 includes the coefficient values of independent features that are used for the linear regression modelling above. While all features are positively related to cost and savings in 5 years, fuel cost apparently plays the most significant role in terms of predicting the target. This makes sense since a vehicle with higher year fuel costs must cost more money annually over time. Fuel type, number of cylinders, and annual petroleum consumption (barrels08) are the top three features other than fuel cost.

Table

Description automatically generated

Table 4.1.1 Features Importance of Linear Regression on Cost/Savings

* Prediction: CO2 Tailpipe Emission

Similar as the previous model, we also wanted to predict CO2 tailpipe emission based on vehicle features. By using the same independent features as above and creating a pipeline, we were able to generate a predictive linear regression on tailpipe emission. More specifically, our pipeline for this model includes a string indexer to index categorical attributes, a feature vector assembler to combine attributes into a vector and standardize features. We used the same training set and testing set as above; however, we changed the target to co2TailpipeGpm.

The linear regression model on tailpipe emission generates a model that produce an MSE of 2380.258 on the training and an MSE of 2015.726 on the testing. Compared to the previous model, the discrepancies between estimated values and actual values are relatively larger, which indicates that the features we selected might not have as strong linearly relations as the previous Cost and Saving model to CO2 tailpipe emission. Nonetheless, we can still manipulate some deep relationships between the dependent variable and independent variables.

Table 4.1.2 shows that the coefficient values of the independent features. We can see the attribute fuelcost08 has a much higher value than the others, but it does not follow the logic of pipe emissions though. From our estimation, fuel cost also represents the amount of fuel consumption for certain vehicles. That being said, more fuel consumption simply means higher fuel cost. Consequently, more fuel cost also indicates more tailpipe emission, which can be associated with the second highest variable, annual petroleum consumption (barrels08). Additionally, our third, fourth, and sixth variables are cylinders, transmission gear box type (trany) and displacement, respectively. We can put them into an actual vehicle powertrain system. Engine and transmission systems are the heart of a vehicle since they provide power. The fact that transmission and cylinders also have high coefficient values indicates that if we want to improve and decrease the tailpipe emission of a vehicle, we should probably start by upgrading its powertrain system.

**Table

Description automatically generated**

Table 4.1.2 Features Importance of Linear Regression on CO2 Tailpipe Emission

**4.2 Random Forest Predictive Model**

Another approach we used to achieve project goals and estimate cost/saving in 5 years as well as tailpipe emission is using the random forest algorithm. Distinct from linear regression, which simply searches for the most important features, random forest finds the most important features by randomly selecting subsets of features and adds more diversity to the final model.

* Prediction: Cost/Savings

From a data transformation aspect, we cast and unified all numeric attributes to double type and use the same nine features as above to predict cost/savings. Feature transformers we used for pipeline include a string indexer, a vector assembler, and a vector indexer that indexes categorical data. Similar to the linear model, we randomly split the dataset into 70% training and 30% test data. In terms of deciding the parameters, we used a grid search and 3 fold cross validation. Specifically, we tried to find the optimal number of trees among 10, 20, 50, 100, 110 and 200 to define an evaluator to assess the RMSE value. This would help to determine the performance of each trial. Some of our attributes, such as vehicle manufacturers (make) and vehicle model (model), have many distinct categorical values. This makes it appear continuous, so we did not want to discretize the data since we want to maintain variability. In order to incorporate more diversity in data, we defined our maximum number of bins to be 4200 for the random forest model.

By applying random forest algorithm to both sets, we also used MSE to evaluate this model. Our model provides had an MSE of 1.0436 on training set and an MSE of 2.3931 on testing set. Compared to the linear predictive model on cost/saving, random forest generated a slightly higher error, yet we can still conclude that all the features we used to run the model are strongly correlated to cost/saving, which is consistent with linear regression model.

Table 4.2.1 is a summary table that contains coefficient values of all vehicle features that are used to predict the cost/saving in 5 years. Fuel cost still remains to be the most important feature in terms of predicting the cost/saving, however the discrepancy between fuel cost and other attributes is less. The consistency between the linear model and random forest model suggests that fuel cost is certainly correlated to cost/savings. Interestingly, if we look at the second top attribute in both linear regression and random forest, the linear model indicates that fuel type is a factor that contributes to fuel cost. That being said, vehicles using premium fuel are most likely to have higher fuel cost. The random forest model suggests that the amount of petroleum consumption (i.e. amount of fuel used) is as important as fuel cost.

**Table

Description automatically generated**

Table 4.2.1 Features Importance of Random Forest on Cost/Savings

* Prediction: CO2 Tailpipe Emission

In terms of data transformation, we wanted to estimate the value of CO2 tailpipe emissions using random forest. While keeping all attributes and the process the same, we only changed the target label to co2TailpipeGpm for this model.

Our random forest model on tailpipe emission generated an MSE of 2522.2102 on training data and 5201.5870 on the testing set when using 110 trees. Compared to the linear regression model on tailpipe emission, random forest seems to be less accurate since MSE is greater.

In terms of inference, the vehicle model surprisingly becomes the most important factor that is used to predict CO2 tailpipe emission in random forest modeling. This result is completely the opposite of the linear regression model where vehicle model is the least important factor. Other than model, factors such as annual petroleum consumption and fuel cost are still the most important factors that determine CO2 tailpipe emission.

Table

Description automatically generated

Table 4.2.2 Features Importance of Random Forest on CO2 Tailpipe Emission

**4.3 K-Means Cluster Analysis (PCA) for Recommendation system**

To incorporate our analysis to a real issue, we anticipated using a k-means analysis of PCA and generating a recommendation system so that we can use our results to guide customers in terms of their vehicle selection process.

We utilized similar feature transformers, including a string indexer, a vector assembler, a standard scaler, and a normalizer to make sure that all vectors have unit length for k-means clustering. In terms of attribute selection, we maintained consistency and choose the same nine attributes as previous the models. According to Figure 4.3.1, within all nine principle components, our scree plot suggests that our first 2 principle components are the ones that captures the most variation. Based on the scree plot in Figure 4.3.1, having 5 principle components seems to be our cutoff point for bias; however, we do not want to end up with too many principle components since it will affect visualizations. Therefore, we will use 2 PCs for later visualizations. Combined with the cumulative sum plot in Figure 4.3.1, we can see that the first 5 PCs explain around 75% of the data variance.

Chart, line chart

Description automatically generated

Figure 4.3.1 Scree plot and Cumulative Sum of Variance plot based on PCA

Based on the PCA score data, we computed k-means silhouette scores for k ranges from 3 to 9. We used silhouette score as a measure of the variation of our clusters and created a clustering evaluator to decide the optimal number of clusters. According to Figure 4.3.2, our evaluation suggests that having 7 centroids produces the highest silhouette score.

Chart, line chart

Description automatically generated

Figure 4.3.2 K-Means Clustering Silhouette Scores Summary Plot

Following the analysis on principle components and k-means clustering, we are able to provide some interesting insights. First, we infer the importance of each vehicle feature using absolute values of PC1 and PC2 loading vector coefficients and the phi coefficients. Table 4.3.1 suggests that the loading coefficient for each vehicle feature is in the first principle component. Specifically, annual petroleum consumption (barrels08), fuel cost, and fuel type are the top 3 important factors that determine PC1. Table 4.3.2 indicates that displacement, number of cylinders, and vehicle manufacturer are the 3 most important factors that affect PC2.

Graphical user interface, application, table

Description automatically generated Graphical user interface, application

Description automatically generated

Table 4.3.2 Loading Vector Coefficients on PC2

Table 4.3.1 Loading Vector Coefficients on PC1

Using PC1 and PC2 as our indicators, we put k-means clustering into visualizations and create a recommender system that can actually help customers find similar vehicle models based on the features. Since one vehicle manufacturer can make different vehicle models with completely distinctive features, we do not want to create a recommender system that only provides results based on vehicle manufacturer. Therefore, we manipulate our attributes and concatenated the “make” and “model” columns into one attribute called Brand/Model. Figure 4.3.3 provides the visualizations of k-means using 7 clusters.

Chart, scatter chart

Description automatically generated

Figure 4.3.3 K-Means Clustering Plot

A picture containing chart

Description automatically generatedAlthough there seems to be some overlap, if we zoom into cluster 4 and cluster 6 (Figure 4.3.4), we can see that similar vehicles are being discretized into one cluster. For instance, we have Lexus LC, Audi A3, and BMW 335 gathered in one cluster together, which makes sense.

Figure 4.3.4 Zoomed in K-Means Clustering

We created a User Defined Function that calculates the distance between two input column vectors and generates a recommender system based on PC1 and PC2 as another key inference from the machine learning. For example, if a customer wants to know what kind of vehicles are similar to Ford, then our function will provide recommendations such as Nissan Truck, Ford Ranger, and Hyundai Veracruz. Similarly, if a customer wants to find vehicles similar to BMW, then our function will recommend him Audi A3, Infinity QX, and Mercedes GLC300. (see Appendix 2 and Appendix 3 for plots)

**5. Conclusion**

We use linear regression and random forest models to determine features that are crucial to predicting Cost/Saving and CO2 tailpipe emissions. However, both linear and random forest models provide higher accuracy when predicting for Cost/Saving. While the MSE generated from the tailpipe emission model is relatively higher. Both supervised models point out that annual petroleum consumption goes along with the fuel cost. That being said, petroleum consumption significantly impacts fuel cost and thus affects Cost/Saving. Incorporating this inference into a real scenario, if a manufacture wants to persuade customers to purchase a vehicle that can save money in 5 years, he will need to decrease petroleum consumption by reducing the amount of displacement and number of cylinders.

Moreover, although the model results for CO2 tailpipe emission indicate a larger discrepancy between estimate and actual value, we can still make the conclusion that the transmission, number of cylinders, and displacement, which can be summarized as powertrain, are the major impacts on the tailpipe emission. Only fuel consumption is excluded from the powertrain attributes. Therefore, if a customer wants to purchase a vehicle with less CO2 emissions, then they should select vehicles that have relatively small powertrain size.

The inference for the k-means clustering analysis used two samples: Ford and BMW. This was to successfully illustrate the idea of clustering analysis by using principle components. Our model can also detect vehicles with the same models, but different features. For instance, Porsche 911 and Porsche 911 SC can be detected like this since we aim to maintain the variety within the dataset. In this case, customers can search for vehicles with different brands or models. They can also select from the same model with different features.

Overall, we have successfully solved our project goals by using different machine learning algorithms and applied the results to solve our problem statement.

# A close up of a screen Description automatically generatedAppendix

Appendix 1. Summary of EDA on numeric attributes

A picture containing diagram

Description automatically generated

Appendix 2. Recommendation System on Vehicle Brands similar to Ford

Timeline

Description automatically generated

Appendix 3. Recommendation System on Vehicle Brands similar to BMW