Predicting CO2 Emissions of Motor Vehicles by Engine Specifications and Manufacturer

BY

Rubeena Rasheed
Norah Alamri
Connor Mennenoh
Pradyumna Deshpande

1. Executive Summary

Objective

- ➢Create a model that can predict the CO2 emissions of a motor vehicle using its specifications with some modicum of accuracy
- ➢Compare multiple models and determine which perform the best and provide the most inferential value
- ➢Identify avenues of future research

Specific Questions

- ➢What is the impact of each feature on the amount of CO2 emissions?
- ➢What feature has the most effect on the amount of CO2 emissions?
- ➢Can we build a prediction model that predicts the CO2 emissions of cars based on the available attributes?

Future Work

- Linear Models were not the standout performers, but did begin to show promise with shrinkage and variable selection
- There is likely ground to be gained by exploring more nuanced shrinkage methods, as well as attempting other methods of variable selection such as forward or backward selection
- Tree models showed remarkable promise, and methods of improving on a standard decision tree were able to produce the best results among the models tested
- A data set that provides more specific information about the vehicle engine of design could add critical information to enhance the performance of any future models
- Additionally, applying cross validation techniques could assist in refining a given model to a greater degree

2. Overview

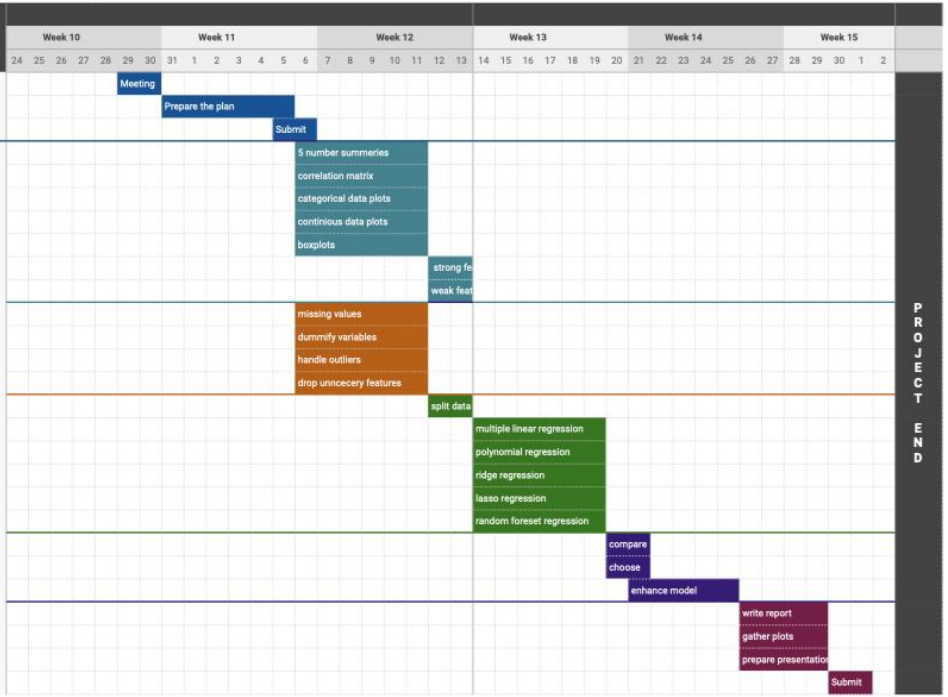
Project Environment Setup

- All data processing, exploration, and analysis was performed in R
- The following packages were used to aid in the processing and analysis of the data
  - caret
  - fastDummies
  - corrplot
  - dplyr
  - gbm
  - glmnet
  - randomForest
  - RColorBrewer

Overview

Timeline and Tasks

	Phase	Tasks	Assigned to
1	Project Plan	1- Meeting	All members
		2- Prepare the plan	
		3- Submit the plan	All members
2	Explantory Data Analysis (EDA)	4- Five numbers summery of each feature.	Norah
		5- Correlation matrix. / insights about it	Connor
		6- Bar charts and Histograms for categorical features.	Norah
		7- Scatterplots and boxplots for continuous features.	Pradyumna
		8- Detect outliers and anomalies via Boxplots.	Pradyumna
		9- Identify features with strong relationship to responce	Connor
		10- Identify features with weak relationship to responce	Connor
3	Data Preparation and Cleaning	11- Replace missing values.	Rubeena
		12- Dummify variables	Rubeena
		13- Handle outliers	Rubeena
		14- Drop unnecessary variables / write justification	Rubeena
4	Model Building and Evaluation	15- Split the dataset to test and train datasets.	Norah
		16- Build and evaluate a multiple linear regression model	Connor
		17- Build and evaluate a polynomial regression model	Pradyumna
		18- Build and evaluate a Ridge regression model.	Rubeena
		19- Build and evaluate Lasso regression model.	Pradyumna
		20- Build and evaluate. Random Forest regression model.	Norah
5	Models Comparison	21- Compare the test results of all models.	Norah
		22- Choose the best model.	Norah
		23- Further, enhance the selected model.	Norah
6	Prepare the final report and presentation	24- Write conclusions and insights.	All members
		25- Gather plots and graphs.	All members
		26- Prepare the final presentation	All members
		27- Submit the project report.	All members



Overview

Dataset

- The data set we chose to use as our primary reference data is the CO2 Emissions by Vehicles data set, sourced from Kaggle but originating from public emissions data collected and made available by the Canadian government.
- This dataset captures the details of how CO2 emissions by a vehicle can vary with the different features.
- This contains data over a period of 7 years. There are total 7385 rows and 12 columns.

Why this dataset?

- The data set has several important features that make it attractive for our purposes.
- The data is consolidated from several years of data, making it more longitudinal than a single data set and allowing the potential for more general impressions to be taken from it.
- There are enough features to provide the potential for meaningful feature selection or shrinkage, without there being so many that the curse of dimensionality becomes a serious concern, and the features themselves cover relevant qualities such as engine size, vehicle size, and fuel type.
- The data is also relatively clean, with very few missing values or unclear records that need to be rectified.

3. Data Preprocessing and Exploratory Data Analysis

Initial Dataset

Make	Model	Vehicle Cl...	Engine Siz...	Cylinders	Transmissi...	Fuel Type	Fuel Cons...	Fuel Cons...	Fuel Cons...	Fuel Cons...
ACURA	ILX	COMPACT	2	4	AS5	Z	9.9	6.7	8.5	33
ACURA	ILX	COMPACT	2.4	4	M6	Z	11.2	7.7	9.6	29
ACURA	ILX HYBRID	COMPACT	1.5	4	AV7	Z	6	5.8	5.9	48
ACURA	MDX 4WD	SUV - SMALL	3.5	6	AS6	Z	12.7	9.1	11.1	25
ACURA	RDX AWD	SUV - SMALL	3.5	6	AS6	Z	12.1	8.7	10.6	27
ACURA	RLX	MID-SIZE	3.5	6	AS6	Z	11.9	7.7	10	28
ACURA	TL	MID-SIZE	3.5	6	AS6	Z	11.8	8.1	10.1	28
ACURA	TL AWD	MID-SIZE	3.7	6	AS6	Z	12.8	9	11.1	25
ACURA	TL AWD	MID-SIZE	3.7	6	M6	Z	13.4	9.5	11.6	24
ACURA	TSX	COMPACT	2.4	4	AS5	Z	10.6	7.5	9.2	31
ACURA	TSX	COMPACT	2.4	4	M6	Z	11.2	8.1	9.8	29
ACURA	TSX	COMPACT	3.5	6	AS5	Z	12.1	8.3	10.4	27
ALFA ROMEO	4C	TWO-SEATER	1.8	4	AM6	Z	9.7	6.9	8.4	34
ASTON MARTIN	DB9	MINICOMPACT	5.9	12	A6	Z	18	12.6	15.6	18
ASTON MARTIN	RAPIDE	SUBCOMPACT	5.9	12	A6	Z	18	12.6	15.6	18
ASTON MARTIN	V8 VANTAGE	TWO-SEATER	4.7	8	AM7	Z	17.4	11.3	14.7	19

Reduction and Reclassification of Features

- Removed “Model” feature as it added no new information not captured in the other features already.
- Three different features measuring fuel consumption and two weighted averages. Removed all but the combined fuel consumption in liters per one hundred kilometers, as it captured the most general information about the fuel consumption using the units most familiar world wide, thus improving potential interpretability later on.
- Reduced “Vehicle Class” feature to general size classifications as many of the categories were subcategories of more general classes such as different sizes of SUVs and trucks and compact cars. The initial 15 categories were reduced to 7 categories as indicated on the right.

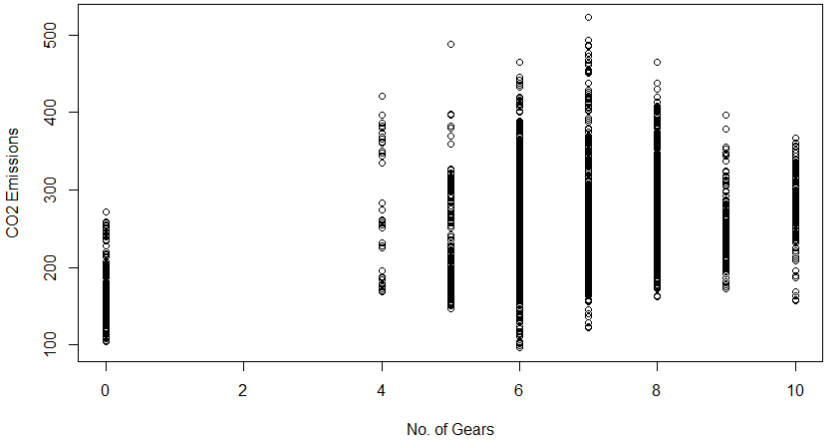
1- SUV - STANDARD 2- SUV - SMALL	1- SUV
3- PICKUP TRUCK - SMALL 4- PICKUP TRUCK - STANDARD	2- PICKUP TRUCK
5- MINIVAN 6- VAN - PASSENGER 7- VAN - CARGO	3- VAN
8- STATION WAGON - SMALL 9- STATION WAGON - MID-SIZE	4- STATION WAGON
10- COMPACT 11- SUBCOMPACT 12- MINICOMPACT	5- COMPACT
13- FULL-SIZE 14- MID-SIZE	6- Sedan
15- TWO-SEATER	7- TWO-SEATER

Challenges

- We encountered large number of unique transmission types.
- We tried to reclassify them in order to reduce the types by trying to group the number of gears which give approximately the same CO2 emission. But based on the CO2 emission vs gear count plot, it was not possible to reclassify the transmission types.

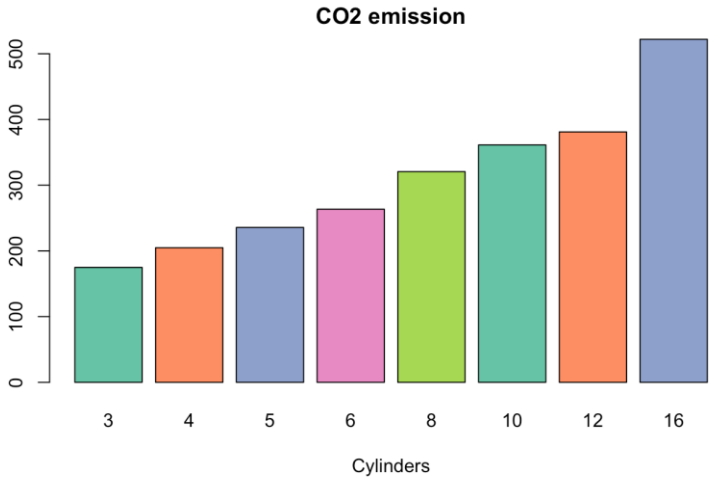
Transmission

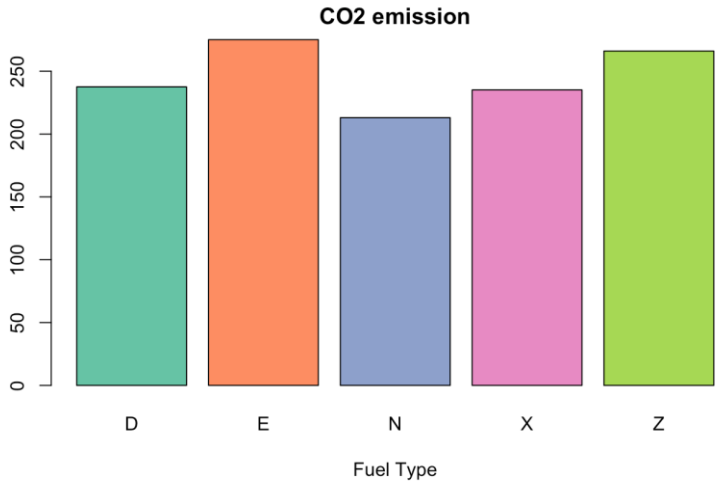
A = Automatic
AM = Automated manual
AS = Automatic with select shift
AV = Continuously variable
M = Manual
3 - 10 = Number of gears



Exploratory Data Analysis (EDA)

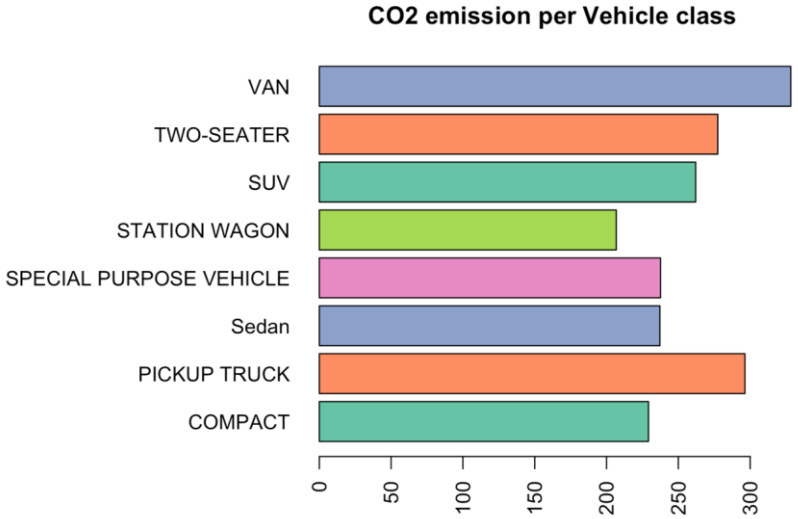
Categorical Data Analysis

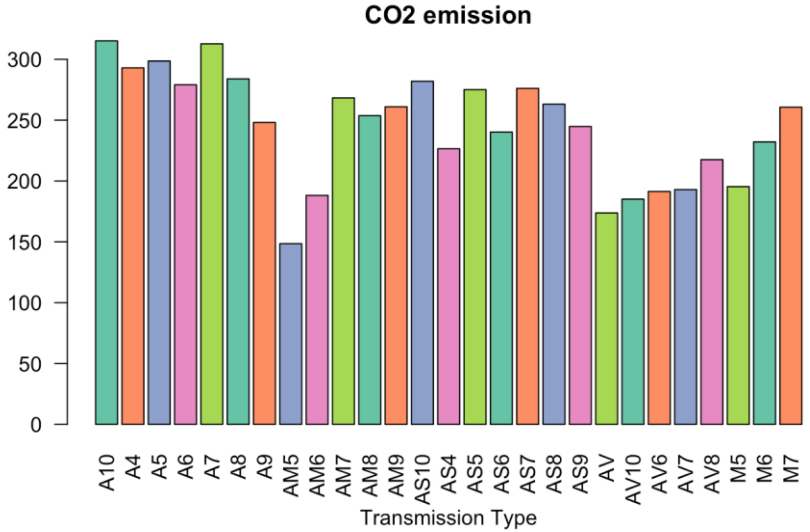




Exploratory Data Analysis (EDA)

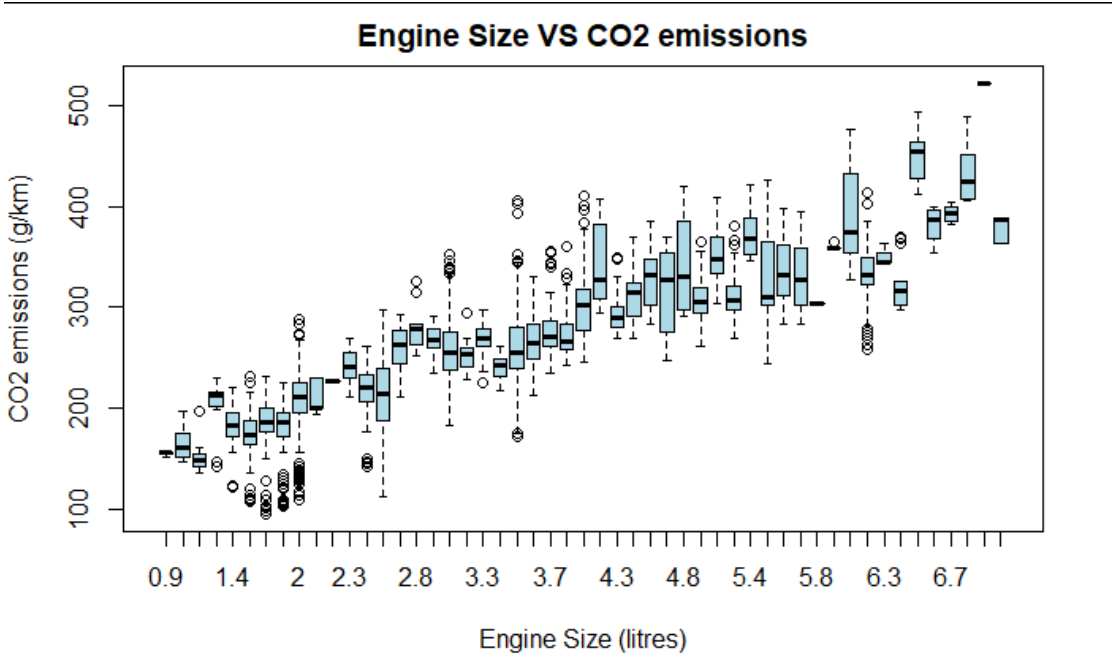
Categorical Data Analysis





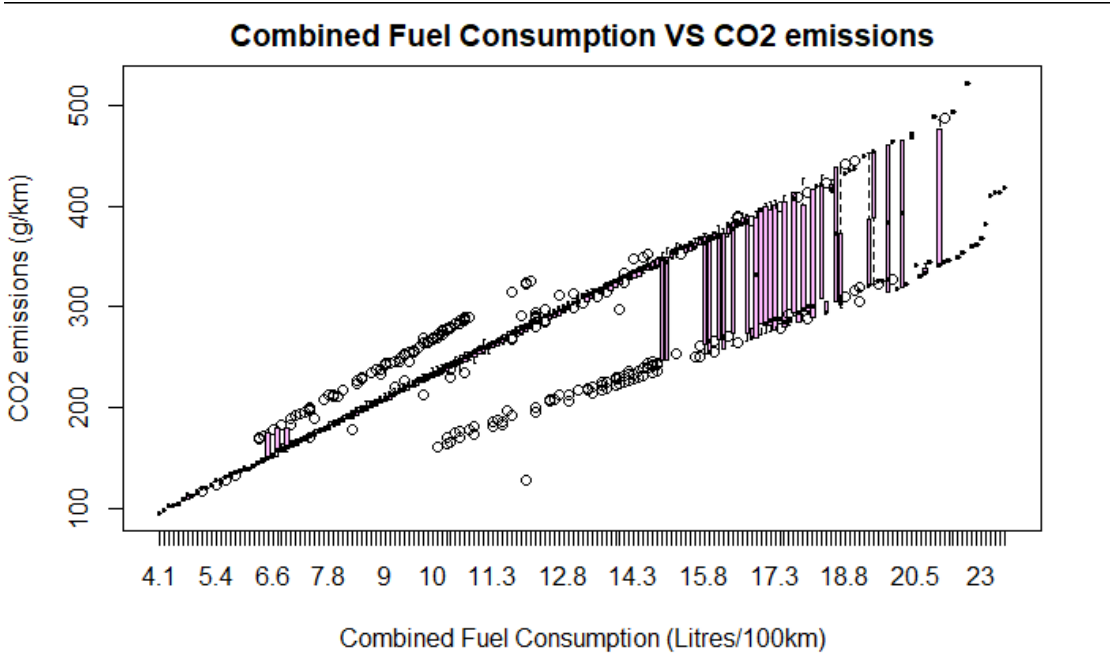
Exploratory Data Analysis (EDA)

Numerical Data Analysis



Exploratory Data Analysis (EDA)

Numerical Data Analysis



4. Modeling

Model Results

#	Model	R2	MSE	RMSE	MAE
1	Linear Regression	-	6775.53	-	-
2	Lasso - Linear Regression	0.99	31.36	5.60	3.03
3	Ridge - Linear Regression	0.972	103.71	10.18	7.58
4	Random Forest	0.994	19.25	4.39	2.32
5	Gradient Boosting	0.996	14.25	3.77	2.04

5. Conclusion

- From the EDA, we have learned that the engine size, cylinders number, and fuel consumption of the vehicle contributes directly to the amount of CO2 emissions.
- From the data pre-processing phase, we have successfully reduced the number of categories in the vehicle class variable. From 15 category down to 7 categories.
- From the Modeling phase, across all 5 models, the best model was a Decision Tree Model which is Gradient Boosting model. This model averages the results of 10,000 trees, therefore it had the best performance out of all models .
- In the future we plan to apply cross-validation techniques to improve the quality of training and increase model performance.
- From this project, we have learned a lot, mostly on: working on a tight schedule, acquiring teamwork skills, solving new technical problems.