2020 International Conference on Computer Communication and Informatics (ICCCI -2020), Jan. 22-24, 2020, Coimbatore, INDIA

An Accurate Prediction of MPG (Miles Per Gallon) using Linear Regression Model of Machine Learning

Varun Shirbhayye, Deepesh Kurmi

Siddharth Dyavanapalli,Agraharam Sai Hari Prasad, Nidhi Lal

Dept. of Computer Science and Engineering
IIIT Nagpur, India

varunshirbhayye@gmail.com, deekurmi2018@gmail.com,dyavanapallisiddharth784@gmail.com, agraharamroyal@gmail.com,nidhi.2592@gmail.com

Abstract*—Considering the growth of the automobile industry coming across since the past two centuries, we are witnessing theincreasing fuel prices and customers being more particularabout the features, the automobile makers are constantlyoptimising their processes to increase fuel efficiency. But what if you could have a reliable estimator for a car's MPG given some known specifications about the vehicle? Then, you could beat a competitor in the market by both having a more desirable vehicle that is also more efficient, improve the fuel economy and bring more demand and supply to the consumers. We are implementing Machine Learning to design the prediction models and minimise the RMSE(Root Mean Square Error) value in between 1 and 2 of the automobiles made (in MPG) in the past years.

***Keywords*— MPG, Origin, RMSE,Data Ingestion,Data Preprocessing,EDA,Model Training,Linear Regression.**

I. INTRODUCTION

Miles Per Gallon(MPG) is a unit which we use to evaluate the efficiency of a transporting vehicle in terms of the energy produced [1].MPGhas different values based on Origin**.** To check the MPG content in vehicles we have made the graph models which we relate with the current MPG in the vehicles based on the following-Displacement**,** Horsepower, Weight, Acceleration,etc.,[2] Also we have focused on the terms produced in predicting our MPG values so as to develop an update on the functioning of cars which we particularly stressed upon as -

Displacementis the volume of the car's engine, usually expressed in litres or cubic centimetres. Originis a discrete value from 1 to 3. This dataset does not describe it beyond that, but we made assumptions based on the table as- 1 to be American-origin vehicle, 2 is European-origin, 3 is Asia/elsewhere. Model yearis given as a decimal number representing the last two digits of the 4-digit year (e.g. 1970 is model year = 70). Our model in this dataset will be trained on many different cars, and it should give us a good estimate for our unknown car's MPG. Note that some of the values in the dataset are incorrect, so we will be fixing those values as we pre-process the data.

II. RELATED WORK

A.DATA INGESTION

Initial thoughts- According to others using this dataset, some of the MPG values for the cars are incorrect, meaning that some of our predictions will be off by a large amount, but we shouldn't always trust the listed MPG value [3].
There are also unknown MPG values in the dataset, marked with a '?'. We will need to manually replace these with the correct MPG value. While our model is the end result of this work, the data analysis section will be incredibly important in visualising trends without having to use any machine learning techniques.

B.DATA PREPROCESSING

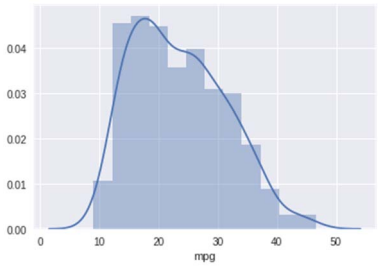

Figure 1. Graph depicting the preprocessed data of the previous MPG model

The purpose of the data pre-processing stage is to minimise potential errors in the model as much as possible. Generally, a model is only as good as the data passed into it, and the data pre-processing we do ensure that the model has as accurate a dataset as possible as shown in Figure 1.

978-1-7281-4514-3/20/$31.00 ©2020 IEEE

While we cannot perfectly clean the dataset, we can at least follow some basics steps to ensure that our dataset has the best possible chance of generating a good model [4]. First, let's check and see null values for this dataset. Null values are empty, useless entries within our dataset that we don't need. If we skip removing null values, our model will be inaccurate as we create "connections" for useless values, rather than focusing all resources onto creating connections for useful values.

C.EXPLORATION DATA ANALYSIS

The purpose of Exploration Data Analysis(EDA) is to enhance our understanding of trends in the dataset without involving complicated machine learning models [5]. Oftentimes, we can see obvious traits using graphs and charts just from plotting columns of the dataset against each other[6].We've completed the necessary pre-processing steps, so let's create a correlation map to see the relations between different features.Acorrelation map (or correlation matrix) is a visual tool that illustrates the relationship between different columns of the dataset. The matrix will be lighter when the columns represented move in the same direction together, and it will be darker when one column decreases while the other increases.[7] Strong spots of light and dark spots in our correlation matrix tell us about the future reliability of the model. We should

import seaborn heat map and plot the correlation map for this dataset.There are some strong correlations between each column. For cylinders, displacement, horse-power, and weight, we had noticed the MPG would be negatively correlated with rising trends in any of the named features.

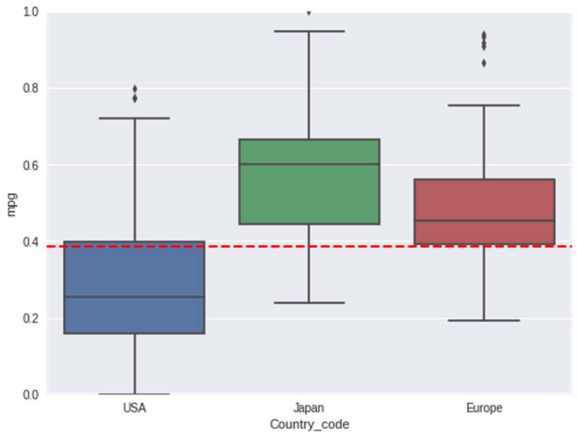

Figure 2. Graph explaining the prediction of MPG in the past years based on the origin

Model year and origin also make sense, since non-American/European countries may contain more fuel-efficient standards due to different fuel prices in those areas. Next, we can plot the number of cars based on their origin (US = 1, Asia = 2, Europe = 3).

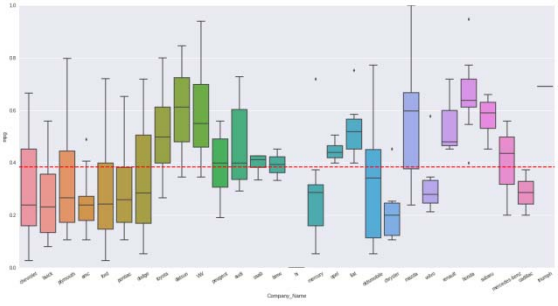

Figure 3. Data representing the MPG values for different car companies

Figure 3. is important to us because we're assuming that different regions have different fuel efficiency priorities, so our model will be skewed towards the region with the most cars in the dataset.[8]

Figure 4. Data representing the scatterplot format of all the factors effecting MPG for cars of different companies and origin

Now that we have looked at the distribution of the data along discrete variables and we sawFigure 4. using the seaborn pairplot. Now let's try to find some logical causation for variations in MPG.

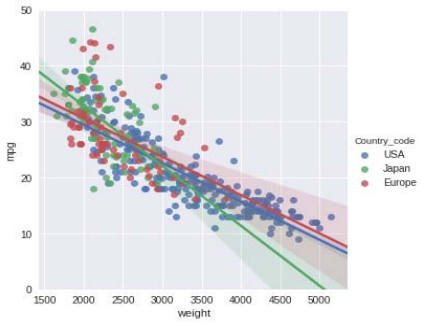

Figure 5. Data representing the weight produced in different cars based on the origin and MPG values

978-1-7281-4514-3/20/$31.00 ©2020 IEEE

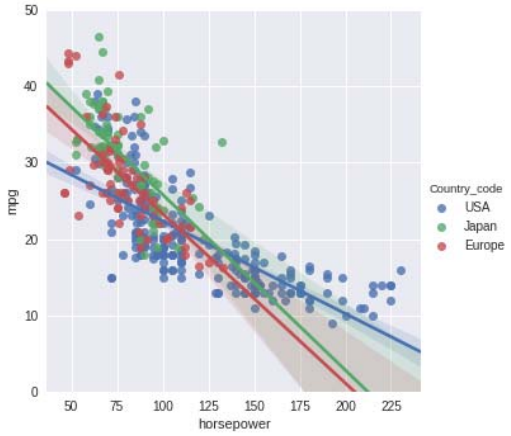

Figure 6. Data representing the horsepower produced in different cars based on the origin and MPG values

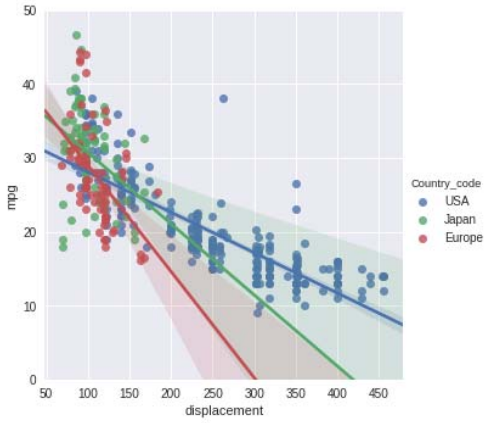

Figure 7. Data representing the displacement produced in different cars based on the origin and MPG values

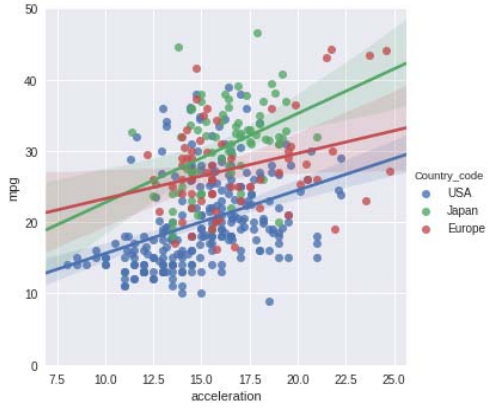

Figure 8. Data representing the acceleration produced in different cars based on the origin and MPG values

We will use the lmplot() function of seaborn with scatter set as true. This will help us in understanding the trends in these relations as shown in Figure 5.,Figure 6., Figure 7. and Figure 8 respectively.The data given helps you represent the trends in weight,horsepower,displacement and acceleration based on the origin . We will split the regressions for different origin countries.

III. PROPOSED WORK

We will use python's scikit learn to train test and tune various regression models on our data and compare the results. We shall use the following regression model-
Linear Regression Model to see how well we can predict the MPG of different vehicles[8]. We should also expect that our predicted MPG values will oftentimes be lower than the actual number because of the number of American cars present within our dataset. We could equalize the distributions of the cars based on region, but doing so would drastically reduce the amount of data points we can use, possibly causing problems in our model due to the lack of training data.

A.MODEL TRAINING

In this section, we will be creating and training our model for predicting what a car's MPG will be[9]. Since there are multiple algorithms, we can use to build our model, we will compare the accuracy scores after testing and pick the most accurate algorithm.

Now, let's initiate by using linear regression algorithm. Then, we train them and check the accuracy on the training set. From this list, we are using Linear Regression Model to perform our predictions [10]. We will then see how the algorithm produces the minimised RMSE value and select it as of a choice for future use. We also want to partition our dataset into training, testing, and validation, so let's add a method for that ability [11]. Let's perform the splitting of our data into test, train, validation using train_test_split. Our testing will take three phases [12]: testing, training, and validation. Training is first, and it's where our model generates "intuition" about how to approach fraudulent and not fraudulent transactions [13]. It is similar to a student studying and developing knowledge about a topic before an exam. The testing phase is where we see how the model performs against data where we know the outcome. This would be the exam if we continue the analogy from before. The algorithms will perform differently, similar to how students will score differently on the exam. From this phase, we generate an accuracy score to compare the different algorithms. The validation testing is how we check that the model isn't over fitting to our specific dataset. Over fitting is when the model starts to develop an intuition that is too specific to the training set.

Over fitting is a problem because our model is no longer flexible. It may work on the initial set, but subsequent uses will cause our model to fail[14]. Continuing the exam

978-1-7281-4514-3/20/$31.00 ©2020 IEEE

analogy, the validation testing phase is like another version of the exam with different questions. If a student happened to cheat on the first exam by knowing the questions, the second exam will give a better representation of performance. Note that verification doesn't completely disprove or prove over fitting, but the testing does give insight into it. We can see how our model performs on the testing set and validation set.

IV. RESULT AND DISCUSSION

Linear Regression is performing with an initial RMSE value of 3.26 and an initial R^2 score of 0.82[2]. While we are picking Linear Regression for future predictions, remember that different algorithms tax different resources if you are prioritising training speed and have limited memory and CPU time. Let's select the most accurate algorithm. The results tell us that our model is decently reliable for the dataset. But we made sure that the RMSE value is minimised and an increased R^2 score as inferred in Figure 9. And Figure 10. To get a better outcome of our tested and fine-tuned training model. Even though some predictions are far away from the actual value, further inspection of the dataset leads me to believe that some of the MPG values are wildly inaccurate. However, we also don't have much data to work with, so we approximated the values to the least possible RMSE value in the range of 1 to 2 and replace the outliers with the actual values of the trained MPG model.

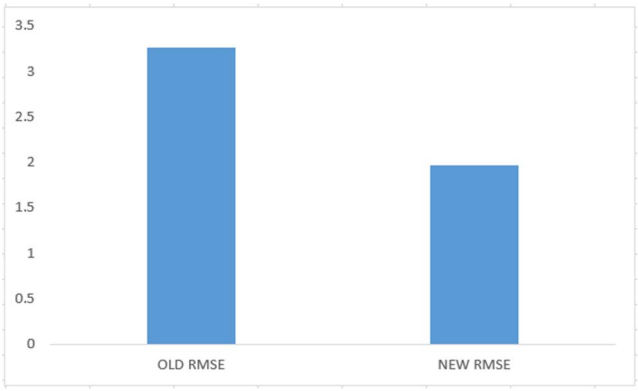

Figure 9. Graph of RMSE values

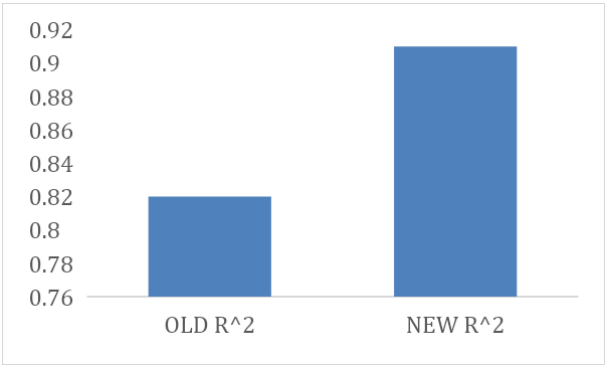

Figure 10. Graph of R^2 Score

V. CONCLUSIONS

During this process, we built a model that could reliably predict a car's MPG given some information about the car and deduce the code then made it possible by obtaining the modified RMSE value of 1.97 from a larger value of 3.26 as the initial RMSE score. This model could be trained with newer car data and be used to predict even the R2 score as we were able to code and found out a steady increase in the value from 0.82 to 0.91 which shows that the model is much reliable in use and also is useful for our competitor's future MPG ratings for upcoming cars, allowing companies to potentially resources currently used today on making more efficient, more popular vehicles that outshine competitors. While our model may be inaccurate in some cases, we talked about how our dataset can contain inaccurate values for the MPG, and oftentimes, our predictions are more accurate than the values in the dataset. For newer cars, the collected data is significantly more reliable, so our model will be able to perform better with different, more accurate dataset.

VI.REFERENCES

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Quinlan, J. Ross. "Combining instance-based and model-based learning." In *Proceedings of the tenth international conference on machine learning*, pp. 236-243. 1993.

[3] Pelleg, Dan. "Scalable and practical probability density estimators for scientific anomaly detection." PhD diss., PhD thesis, Carnegie-Mellon University, 2004. Tech Report CMU-CS-04-134, 2004.

[4] Tao, Qingping. "Making efficient learning algorithms with exponentially many features." PhD diss., University of Nebraska--Lincoln, 2004.

[5] Palmer, Christopher R., and Christos Faloutsos. "Electricity based external similarity of categorical attributes." In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 486-500. Springer, Berlin, Heidelberg, 2003.

[6] Inui, Kentaro, Jing Jiang, Vincent Ng, and Xiaojun Wan. "Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

[7] Pelleg, Dan, and Andrew Moore. "Mixtures of rectangles: Interpretable soft clustering." In *ICML*, pp. 401-408. 2001.

[8] Li, Jinyan, Kotagiri Ramamohanarao, and Guozhu Dong. "Combining the strength of pattern frequency and distance for classification." In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 455-466. Springer, Berlin, Heidelberg, 2001.

978-1-7281-4514-3/20/$31.00 ©2020 IEEE

[9] Melluish, Thomas, Craig Saunders, Ilia Nouretdinov, and Vladimir Vovk. "The typicalness framework: a comparison with the Bayesian approach." *University of London, Royal Holloway* (2001).

[10] Zhou, Zhi-Hua, Shi-Fu Chen, and Zhao-Qian Chen. "A statistics based approach for extracting priority rules from trained neural networks." In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, vol. 3, pp. 401-406. IEEE, 2000.

[11] Birattari, Mauro, Gianluca Bontempi, and Hugues Bersini. "Lazy learning meets the recursive least squares algorithm." In *Advances in neural information processing systems*, pp. 375-381. 1999.

[12] Greig, D., T. Siegelmann, and M. Zibulevsky. "A New Class of Sigmoid Activation Functions That Don't Saturate." (1997).

[13] Fürnkranz, Johannes. "Pairwise classification as an ensemble technique." In *European Conference on Machine Learning*, pp. 97-110. Springer, Berlin, Heidelberg, 2002.

[14] Brown, Christian T., Harry W. Bullen, Sean P. Kelly, Robert K. Xiao, Steven G. Satterfield, and John G. Hagedorn. "Visualization and Data Mining in an 3D Immersive Environment: Summer Project 2003." *US National Institute of Standards and Technology* (2003).

978-1-7281-4514-3/20/$31.00 ©2020 IEEE