TECHNICAL REPORT

Predictive Analysis for Vehicle Transmission Detection

Tanja Adzic

Introduction

The primary goal of this project was to predict vehicle transmission type, if it is a manual transmission type or automatic, by leveraging a dataset that is comprised of 11 features and 32 data instances. Potential implications of this project's results could yield a significant influence to key aspects of vehicle design, fuel efficiency, customer preferences, cost optimisations, as well as research and development, to name a few.

The project consisted of several phases, with the initial phase being dataset exploration, and the final phase being predictive modeling recommendations.

The dataset itself showed no significant anomalies once the relationships between the transmission type, and other attributes were explored using various visualizations. However, they did reveal meaningful relationships between transmission type and attributes that will be further explained in detail.

This particular prediction problem focuses on binary classification, and there are a variety of classification models that I used for prediction. After evaluating multiple classification algorithms, logistic regression emerged as the standout performer, demonstrating strong accuracy. By leveraging its inherent interpretability, I gained in-depth insights into feature significance, enhancing the transparency of my analysis.

In summary, my analysis consisted of data validation, algorithm selection, as well as model optimization. Prominent results of the logistic regression model, reinforced by hyperparameter tuning, underscores its aptitude for precise transmission type classification. My findings provide valuable insights that can inform automotive design and engineering decisions.

Data

The source of the dataset is from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Data interpretation provides a significant part of this project, as it provides me with necessary information for adequate data handling, data validation, and model selection. Here is a brief description of the attributes:

- *am* - transmission type, automatic or manual
- *mpg* - fuel efficiency that indicates how many miles the vehicle can travel on a single gallon of fuel
- *cyl* - the number of cylinders in the vehicle's engine, where the higher number means a more powerful engine
- *disp* - engine displacement which indicates the volume of engine’s cylinders, where a higher displacement usually indicates a more powerful engine
- *hp* - horsepower refers to the power output of the engine
- *drat* - the rear axle ratio represents the ratio of the number of revolutions of the drive shaft to the rear axle's revolutions, where a higher ratio leads to higher acceleration, but lower ratio leads to better fuel efficiency at higher speeds
- *wt* - weight signifies vehicle’s mass, and it influences many aspects of vehicle performance: fuel efficiency, handling, power
- *qsec* - the time it takes for the vehicle to cover a quarter-mile distance from a standstill, measuring acceleration, where lower qsec indicates quicker acceleration
- *vs* - engine types shown as V-shaped (V-engine) or straight (inline) based on the configuration of cylinders

- *gear* - number of gears in the vehicle's transmission
- *carb* - the number of carburetors in the vehicle which can impact engine performance and fuel efficiency

Dataset Exploration and Preprocessing

In this phase, I conducted a comprehensive examination of the dataset in order to identify any potential anomalies that could affect future processes. The initial stage included gaining brief statistical summaries of the data, inspection of the dataset's structure, and a visual representation of the entries in order to identify outliers.

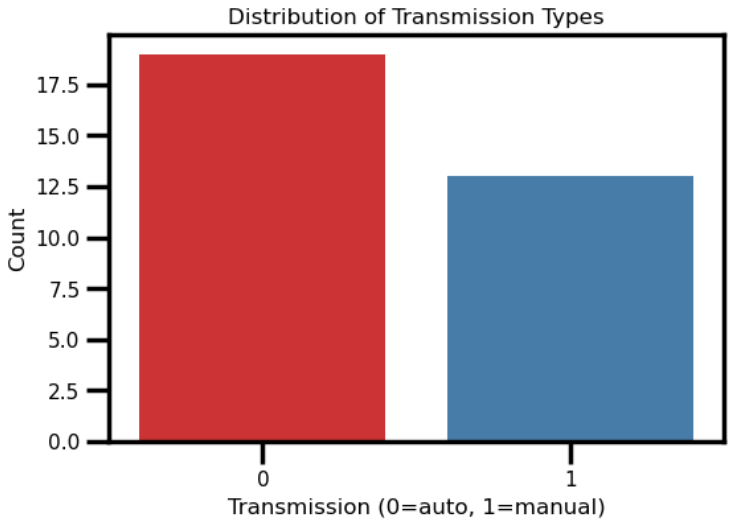
During this exploration process, I checked the dataset for any potential data quality issues, including invalid or anomalous entries. Fortunately, I observed no apparent outliers, missing values, or data quality issues that required further action. The dataset showed a consistent and coherent structure. Having in mind that the dataset consists of only 32 instances, I was able to visually inspect the entire set, and gain insight into the proportions of values for all attributes. This proved to be significant for future data transformation, as I used a scaling method to achieve better model performance.

Furthermore, I verified that the data types of the attributes were aligned with their intended representation and description. Given the absence of missing values and the high-quality nature of the dataset, I determined that no additional data transformations were necessary to proceed with the analysis. However, I do emphasize that it could be worth considering applying additional data transformation techniques in order to achieve better results, and potentially identify additional hidden patterns in the data that might also improve model performance.

Key Findings

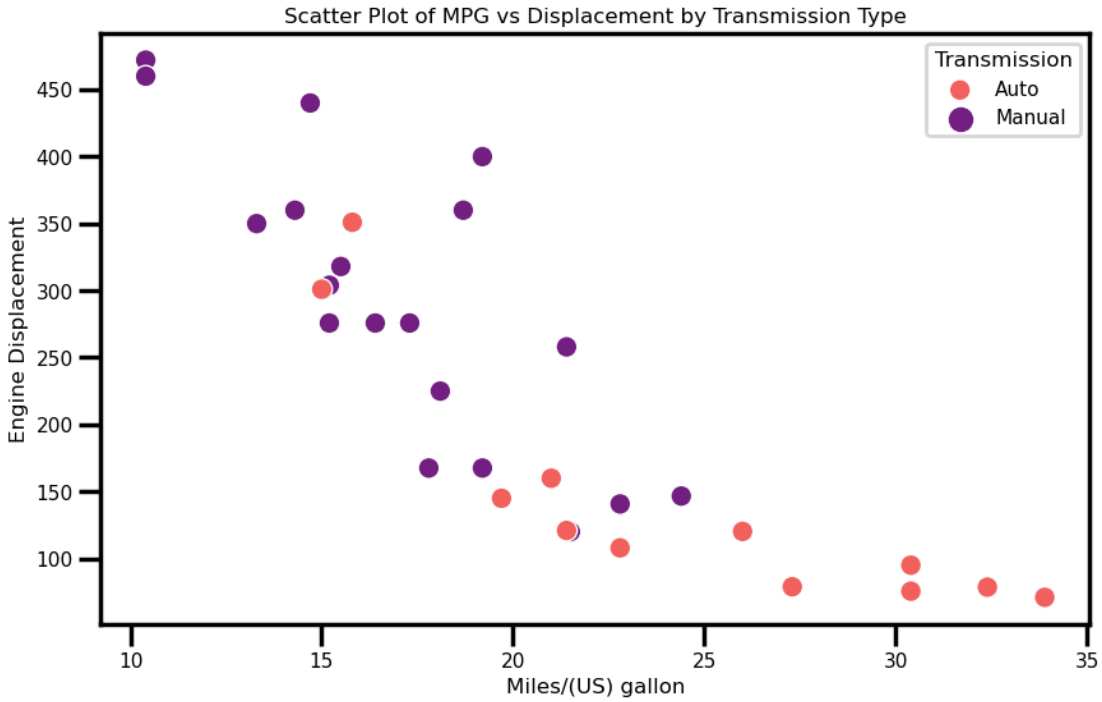
In order to gain a significantly better picture of your data, and the relationship between attributes, I used a couple of visualizations that provided me with important insights.

How many transmission type instances are there?



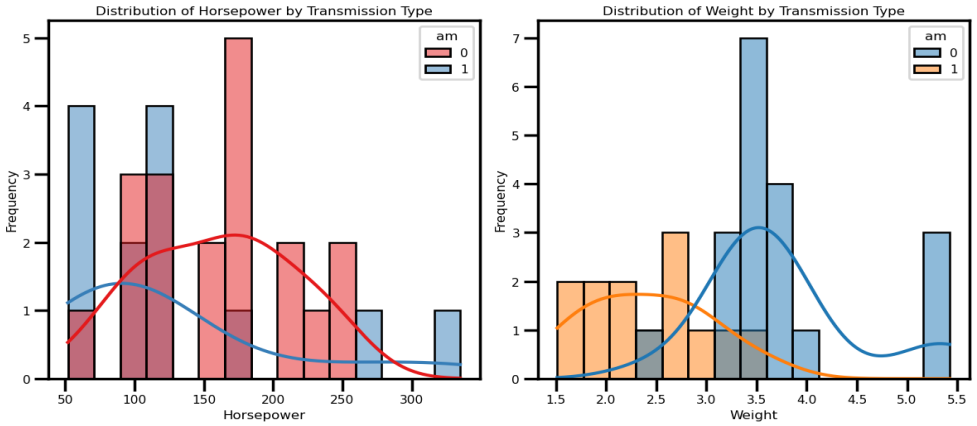
This countplot shows us there are about a third more instances for the automatic transmission type. However, I do not see a significant difference in the instance balance that can potentially have a strong effect on model performance.

Is there a trend between engine displacement and fuel efficiency by transmission type?



In general, there is a downward trend visible on the plot, which signifies that vehicles with higher engine displacements tend to have lower fuel efficiency. This might be due to the fact that larger engines often consume more fuel, since larger engine displacement values are found in more powerful engines. Both transmission types show a downward trend with a visible difference in their distribution. Manual transmission cars seem to cover a wider range, and higher values of engine displacements and fuel efficiencies compared to automatic transmission cars, whereas automatic transmission cars are trending in the lower engine displacement values, and lower fuel efficiency. Some points do overlap indicating that there are manual transmission cars with both higher and lower fuel efficiency for a given engine displacement. Based on the data points available, I assume that engine displacement is negatively correlated with fuel efficiency, and the data points are distributed differently for manual and automatic transmission cars.

How are horsepower and vehicle's weight distributed by transmission type?



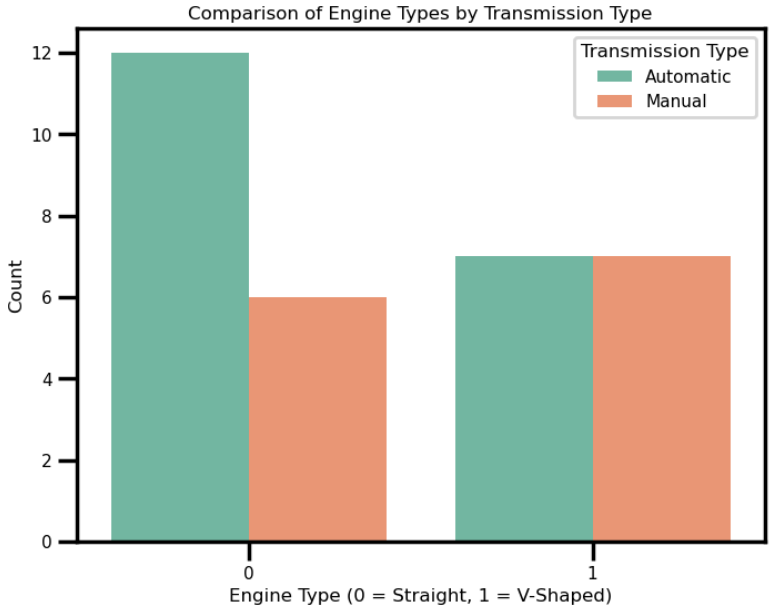
Based on this plot, vehicles with manual transmissions tend to have a higher concentration of horsepower values towards the lower end of the range. On the other hand, vehicles with automatic transmissions are more evenly distributed across different horsepower values. On the other hand, vehicles with manual

transmissions often have lower weights compared to those with automatic transmissions. The distribution for manual transmissions seems to peak towards the lower weight values, while the automatic transmission distribution is more spread out.

How many gears are there by transmission type?

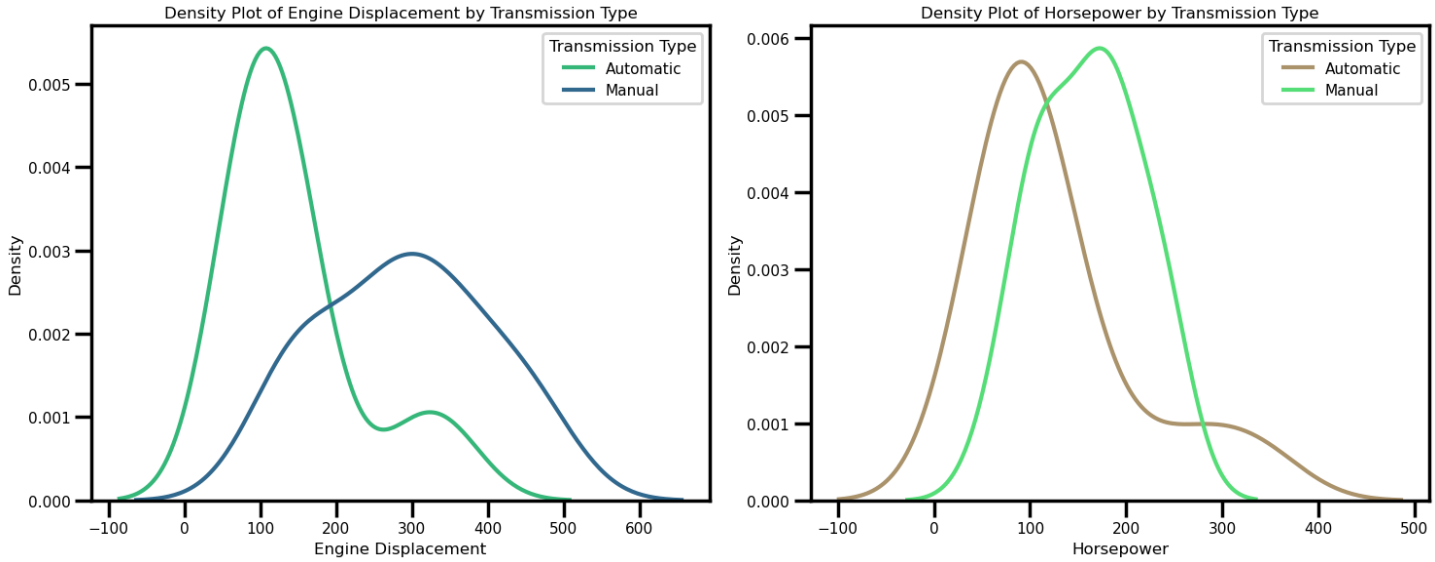
This bar plot allows me to visually compare how the distribution of gear values varies between automatic and manual transmission types, giving insights into the relationship between transmission type and the number of gears in the vehicle's transmission. Vehicles with automatic transmission show a less number of gears in general, whereas number of gears rise above 4 when it comes to manual transmission.

What are the prevalent engine shapes across both transmission types?



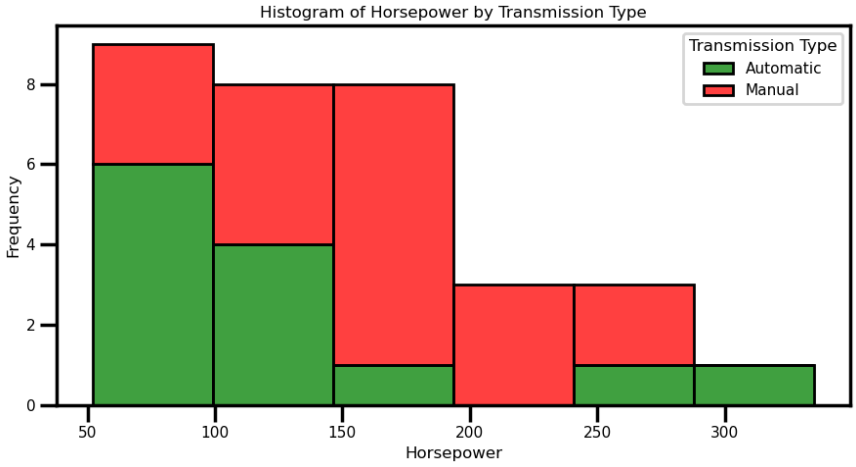
In vehicles with automatic transmissions, the number of V-shaped engines is notably higher than the number of straight engines. On the other hand, in vehicles with manual transmissions, the number of V-shaped engines (vs = 0) is not significantly higher, and the number of straight engines is lower in comparison.

How are horsepower and engine displacement distributed by transmission type?



The density of data points is higher in regions where the curve is taller, indicating where more data points are concentrated. The density curves for engine displacement suggest that vehicles with manual transmissions tend to have higher engine displacements compared to vehicles with automatic transmissions. On the other hand, the density curves for horsepower indicate that both vehicles with manual transmissions and automatic transmission have slightly different distributions, but are evenly distributed with vehicles with manual transmission having slightly more horsepower.

Is horsepower higher or lower in vehicles with manual transmittion?



The frequency of vehicles with manual transmission is higher in vehicles with lower horsepower, and is heavily concentrated in vehicles from 50 to 200 horsepower. Vehicles with automatic transmission seem to be on the higher end when it comes to lower horsepower, however there are data points present in the very high horsepower values over 250.

Methodology

I conducted a comprehensive exploration of classification algorithms in order to analyze essential classification metrics, including accuracy, precision, recall, and F1-score. My intent was not only to identify the top-performing algorithm but to understand the unique contribution and suitability of each model.

Among the array of algorithms I evaluated, my initial choice was Logistic Regression. Despite not achieving the highest accuracy, its transparent interpretability was important for the task at hand. The coefficients generated by Logistic Regression provided direct insights into the relationships between various vehicle attributes and the likelihood of manual transmission. I found this transparency to be crucial in the context of vehicle design and performance prediction.

XGBoost served as my second choice and its selection wasn't based on standout metrics, but on its potential for improvement. XGBoost exhibited consistent results across metrics. My decision to proceed with XGBoost originated in its ensemble nature, and the capacity for enhanced performance through hyperparameter tuning. The approach recognized the value of models that offer a balance between performance, transparency, and optimization potential.

In conclusion, both methodological approaches emphasized the thoughtful selection of models that align with the context of the problem. While one approach focused on identifying transparent insights from coefficients, the other focused on potential through optimization.

Data Modeling and Evaluation

Here are the results of various classification algorithms. These results were obtained using cross validation.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.847619	0.850000	0.766667	0.771429
Random Forest	0.728571	0.785714	0.666667	0.653333
Decision Tree	0.700000	0.585714	0.600000	0.553333
SVM	0.785714	0.785714	0.800000	0.753333
KNN	0.695238	0.785714	0.600000	0.613333
Naive Bayes	0.785714	0.785714	0.800000	0.753333
Gradient Boosting	0.733333	0.619048	0.700000	0.613333
Neural Network	0.819048	0.650000	0.700000	0.671429
XGBoost	0.785714	0.785714	0.800000	0.753333

As previously stated, I chose both Logistic Regression and XGBoost as promising models for further performance enhancement. However, XGBoost performed not as expected, and the results were not promising:

XGBoost Classification Report:

	precision	recall	f1-score	support
0	0.60	0.75	0.67	4
1	0.50	0.33	0.40	3
accuracy			0.57	7
macro avg	0.55	0.54	0.53	7
weighted avg	0.56	0.57	0.55	7

I assume this is due to the fact there are only 32 data points in the dataset, and the test set sample is very small.

On the other hand, Logistic Regression shows well enough results when a cross validation performed, however it seems that it generalized to perfect separation according to evaluation metrics:

- *Mean Cross-Validation Score*: 0.9199999999999999
- *Standard Deviation of CV Scores*: 0.09797958971132709
-
- *Coefficients:* [0.7127164 0.17747573 -0.09659108 0.1300007 0.12033337 -0.11563555 -0.97829085 -0.24767974 0.40665683 0.41313098]
- *Intercept:* -0.009911206525683855

```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         4
           1       1.00      1.00      1.00         3

    accuracy                           1.00         7
   macro avg       1.00      1.00      1.00         7
weighted avg       1.00      1.00      1.00         7

ROC AUC: 1.0
Precision-Recall AUC: 1.0
```

Model Limitations and Future Recommendations

In the context of a small dataset comprising 32 data points, I was able to discern certain limitations from the model analyses. Firstly, the logistic regression model exhibited instances of perfect scoring, indicating the presence of perfect separation. This phenomenon happens when certain combinations of predictor variables precisely predict the outcome, resulting in extreme coefficient estimates as seen in the results, and instability in predictions. Additionally, the XGBoost model exhibited suboptimal performance, manifesting in low accuracy metrics such as precision, recall, and F1-score. This underperformance might stem from a small dataset size, which may not provide sufficient data for the complex model to generalize effectively. Considering these limitations, it is also important to consider overfitting when interpreting the results.

To address these issues and enhance future analyses, I suggest acquiring a larger dataset to improve model generalizability and apply techniques such as regularization for logistic regression which might yield more reliable insights. Additionally, exploring alternative ensemble techniques and optimizing hyperparameters could help enhance the predictive capacity of the XGBoost model.

Conclusion

The primary objective of this project was to construct a classification model capable of precisely categorizing vehicle transmission types—manual and automatic. This was achieved through a systematic process of dataset exploration, algorithm selection, and model optimization. The dataset comprised 32 instances and 11 attributes. Initial analysis revealed intricate relationships between transmission types and attributes, forming the basis for subsequent examination.

However, I encountered different challenges during this project. The presence of perfect separation, a condition where a predictor variable entirely distinguishes the outcome, hindered the application of logistic regression, leading to inconclusive outcomes. Additionally, the XGBoost model exhibited suboptimal performance, possibly due to the dataset's limited dimensionality and size. The project emphasized the challenge between algorithmic choices and real-world dataset characteristics. While logistic regression encountered limitations, this exploration of methodologies offered valuable insights into feature significance.

In conclusion, this project highlights the intricate nature of predictive modeling.. The insights derived from this analysis, while navigating these intricacies, contribute meaningfully to the domain of vehicle engineering and design, with potential applications for informed decision-making in the future.