Predicting a car transmission type using specific car design and performance information

INTRODUCTION

In this paper we investigate the problem of identifying a car transmission type for 32 old car types manufactured in 1973/74. The goal was to classify vehicle transmission as manual or automatic based on different values of other car features as correctly as possible. The data set was small so evaluation of any model would be problematic. After observing data values, descriptive analysis and their correlation to the target value as well as their mutual correlation, several data features were excluded from the data set and models were created and evaluated solely with four features that were kept. Nested cross-validation was used for model selection and evaluation to prevent problems that occur when the data set is too small. The models that were selected for this analysis are Logistic regression, GaussianNB, Linear SVM, kernelized SVM, Random forest and XGBoost. The best results were achieved with Logistic regression with L2 regularization, thus that is our model of choice.

DATA

Our data set consists of 32 instances and each of them has 11 features and one target value that presents an indication whether the transmission type of a car is manual or automatic. The goal is to predict the target value based on different features such as miles per gallon, number of cylinders, gross horsepower, weight, engine type, number of gears etc.

One of the features presents the name and model of a car, and was excluded from our data set due to lack of informativity for our predictions.

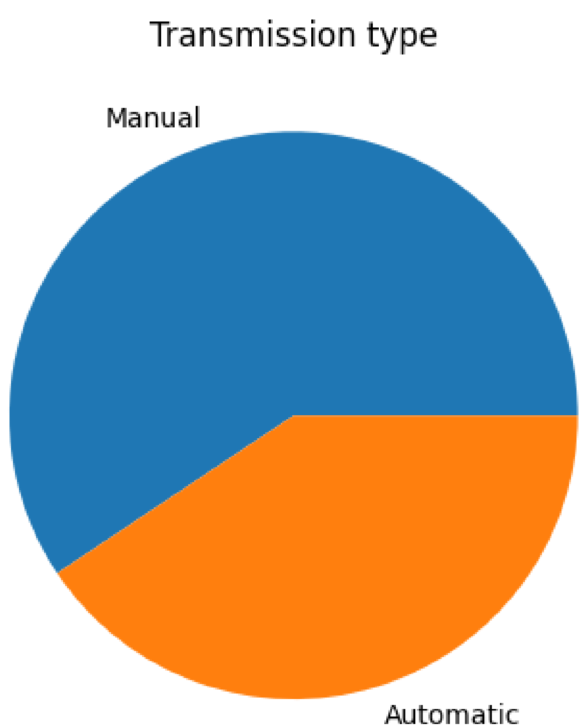
Target values are well balanced so there is no need for oversampling. Descriptive analysis of features shows that features are either continuous or ordinal and contain no missing values.

Feature description	Feature name , count, type		
Miles/(US) gallon	mpg	32 non-null	float64
Number of cylinders	cyl	32 non-null	int64
Displacement (cu.in.)	disp	32 non-null	float64
Gross horsepower	hp	32 non-null	int64
Rear axle ratio	drat	32 non-null	float64
Weight (1000 lbs)	wt	32 non-null	float64
1/4 mile time	qsec	32 non-null	float64
Engine (0 = V-shaped, 1 = straight)	vs	32 non-null	int64
Transmission (0 = automatic, 1 = manual)	am	32 non-null	int64
Number of forward gears	gear	32 non-null	int64
Number of carburetors	carb	32 non-null	int64

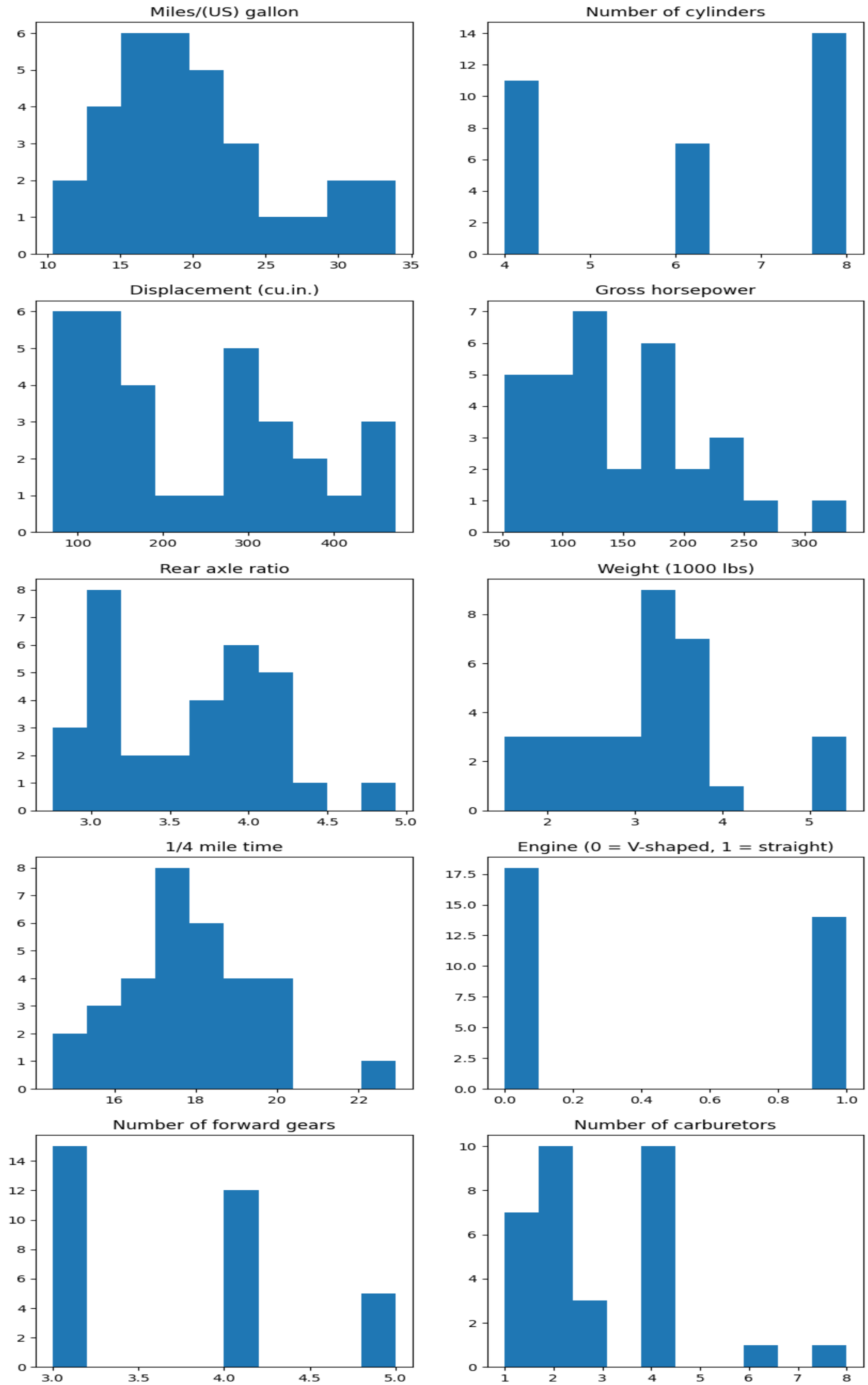
Descriptive analysis of data

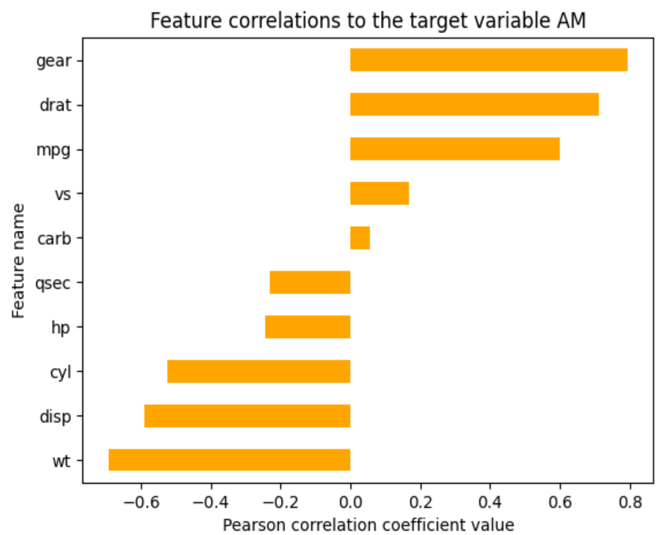
	count	mean	std	min	25%	50%	75%	max
mpg	32.000	20.091	6.027	10.400	15.425	19.200	22.800	33.900
cyl	32.000	6.188	1.786	4.000	4.000	6.000	8.000	8.000
disp	32.000	230.722	123.939	71.100	120.825	196.300	326.000	472.000
hp	32.000	146.688	68.563	52.000	96.500	123.000	180.000	335.000
drat	32.000	3.597	0.535	2.760	3.080	3.695	3.920	4.930
wt	32.000	3.217	0.978	1.513	2.581	3.325	3.610	5.424
qsec	32.000	17.849	1.787	14.500	16.893	17.710	18.900	22.900
vs	32.000	0.438	0.504	0.000	0.000	0.000	1.000	1.000
am	32.000	0.406	0.499	0.000	0.000	0.000	1.000	1.000
gear	32.000	3.688	0.738	3.000	3.000	4.000	4.000	5.000
carb	32.000	2.813	1.615	1.000	2.000	2.000	4.000	8.000

The value we want to predict is the value of ‘am’ column and it has values 0 or 1 indicating manual or automatic transmission and it is not imbalanced



Feature distribution histograms

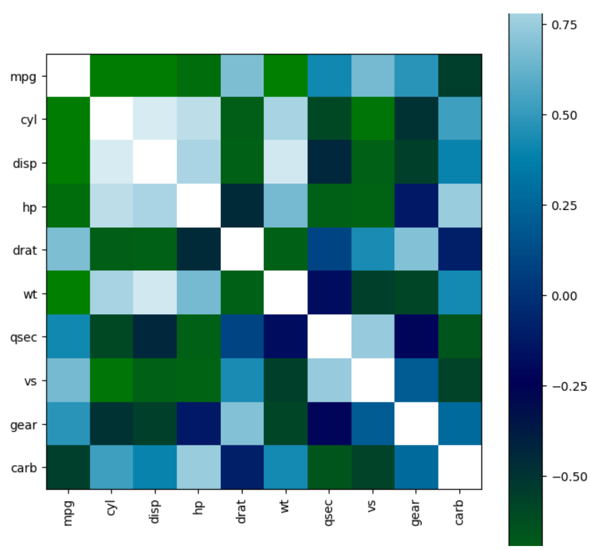




Feature correlation to target value

```
mpg 0.599832429454648
cyl -0.5226070469006754
disp -0.5912270400639476
hp -0.24320425718585106
drat 0.7127111272262697
wt -0.6924952588394844
qsec -0.22986086218488297
vs 0.16834512458535864
gear 0.7940587602563435
carb 0.057534351070504114
```

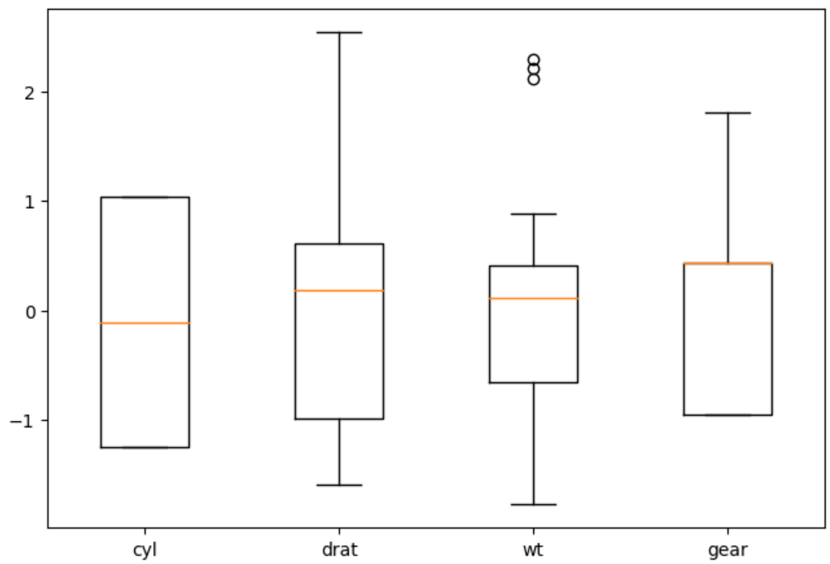
Mutual feature correlation shows that there are several correlated features and some of them should be excluded from further analysis either manualy or by using PCA analysis to reduce the number of used features.



	mpg	cyl	disp	hp	drat	wt	qsec	vs	gear	carb
mpg	1.00	0.85	0.85	0.00	0.00	0.87	0.00	0.00	0.00	0.00
cyl	0.85	1.00	0.90	0.83	0.00	0.00	0.00	0.81	0.00	0.00
disp	0.85	0.90	1.00	0.00	0.00	0.89	0.00	0.00	0.00	0.00
hp	0.00	0.83	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
drat	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
wt	0.87	0.00	0.89	0.00	0.00	1.00	0.00	0.00	0.00	0.00
qsec	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
vs	0.00	0.81	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
gear	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
carb	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Coefficients with absolute value higher than 0.8

From features fairly correlated to target value additionaly we decided to exclude 'mpg' and 'disp' because of their mutual correlation to other features. Thus the features that are kept for further analysis are: **cyl** -0.52, **drat** 0.71, **wt** -0.69, **gear** 0.79



Kept feature scaled values boxplot

MODEL SELECTION AND EVALUATION

Nested cross-validation is used on small datasets when testing is problematic and danger of overfitting and it allows us to find the best model and estimate its generalization error correctly.

The data set is very small and thus nested cross validation will be used for evaluation of models. Models that were trained on data are Logistic Regression with L2 regularization, Linear SVM classifier, SVM classifier with gaussian and linear kernel, Gaussian naïve Bayes algorithm for classification, Random forests for classification, XGBoost for classification with different number of classifiers and depth.

Hyperparameters used to select models:

XGBoost – max_depth = [2, 3, 4]; n_estimators = [10, 20, 30, 40, 50]

SVM classifier- gamma = [0.001, 0.01, 0.1, 1, 10], C=[10, 100, 1000], kernel=['rbf', 'linear']

Random Forest - max_depth = [2, 3, 4]; n_estimators = [10, 20, 30, 40, 50]

Logistic Regression with L2 regularization was used with fixed regularization type and strength while Linear SVM classifier, Gaussian naïve Bayes algorithm were trained with no hyperparameters.

RESULTS

Model	*Best score*	*Average score*	*Parameters*	*Scores*
SVM	0.90909	0.77272	C = 10	0.90909
			gamma = 0.001	0.90909
			kernel = linear	0.50000
Logistic regression	1.00000	0.88888		0.66667
				1.00000
				1.00000
Linear SVM	1.00000	0.86904		0.85714
				0.75000
				1.00000
Gaussian NB	0.88889	0.74074		0.88889
				0.83333
				0.50000
Random Forest	0.85714	0.73015	max_depth = 2	0.66667
			n_estimators = 40	0.66667
				0.85714
XGBoost	1.00000	0.70909	max_depth = 2	0.72727
			n_estimators = 10	1.00000
				0.40000

The model that gives best results from all of the trained models is Logistical regresion with L2 regularization, so this model will be trained on data and used as estimator.

```
Classification report
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         4
           1       1.00      1.00      1.00         6

    accuracy                           1.00        10
   macro avg       1.00      1.00      1.00        10
weighted avg       1.00      1.00      1.00        10

Confusion matrix
[[4 0]
 [0 6]]

Coefficients = ([[ 0.03051147,  0.71571355, -1.50259361,  1.58427969]])
```

```
Intercept = ([-4.66197341])
```

The highest influence on transmission prediction has the *number of forward gears* (gear feature).

CONCLUSION

Small data sets always present challenge for creating a model that would generalize well.

Additionally, the evaluation of such models is problematic because it has been tested on insufficient amount of data, so the results are not very reliable.