Used Car Prediction Case Study

Anthony Lee 2022-July-01

Table of Content

Sales Manager Presentation

Predicting used car price to improve Discount Motor’s top line

Anthony Lee 2022-July-01

Situation

- Sales declined 18% in recent months
- Problem: Inconsistent pricing of vehicles
- Need: Tool to assist with vehicle pricing

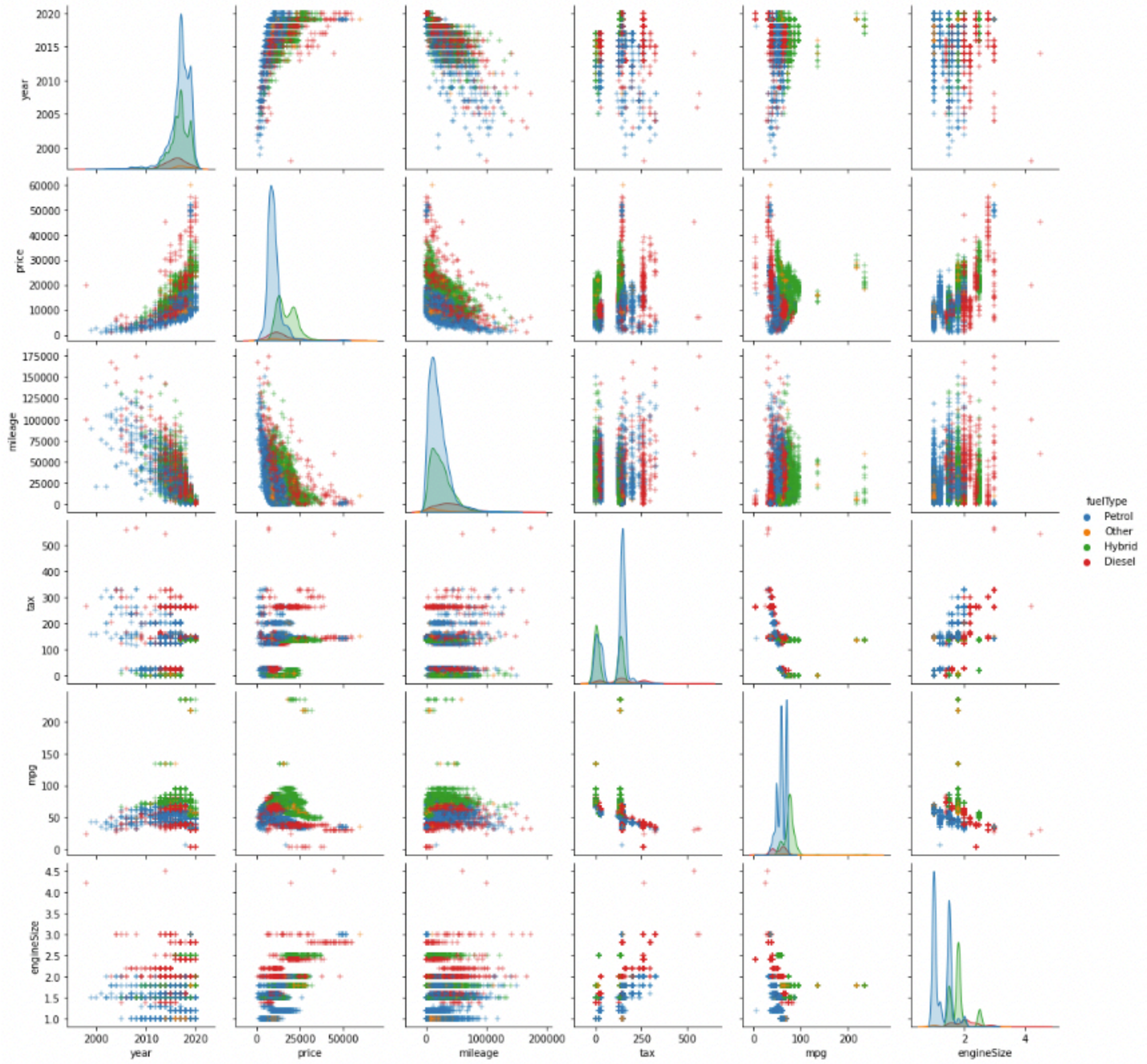
Methods

Data

- 6738 entries of Toyota vehicle attributes and list price of some other retailers.
  - Vehicle attributes: model, year, transmission_type, mileage, fuel_type, tax, mpg, and engine-size
  - Target variable: price (in British Pound)

	model	year	price	transmission	mileage	fuelType	tax	mpg	engineSize
0	GT86	2016	16000	Manual	24089	Petrol	265	36.2	2.0
1	GT86	2017	15995	Manual	18615	Petrol	145	36.2	2.0
2	GT86	2015	13998	Manual	27469	Petrol	265	36.2	2.0
3	GT86	2017	18998	Manual	14736	Petrol	150	36.2	2.0
4	GT86	2017	17498	Manual	36284	Petrol	145	36.2	2.0

- Pair-wise distribution on a scatterplot on different fuel types
- Correlation shows strong correlation of engineSize with Price (target interest)



Models
Linear Regression

- Models have strong predictive power on the listing price of a vehicle. (Super car vs economy car).
- Vehicle characteristics have strong correlation with the price (Descending order: engineSize, Automatic Transmission, Hybrid, and year)

Descending sort by 'linear regression abs coef'

	variable	linear regression coef	linear regression abs coef	pearson coef
18	model_Supra	23710.063098	23710.063098	0.255997
14	model_Land Cruiser	16146.310113	16146.310113	0.347347
7	model_Aygo	-8126.233966	8126.233966	-0.468763
22	model_Yaris	-7508.988346	7508.988346	-0.211888
15	model_PROACE VERSO	7497.721786	7497.721786	0.123857
13	model_IQ	-5964.973224	5964.973224	-0.032201
5	model_Auris	-5726.498235	5726.498235	0.000589
19	model_Urban Cruiser	-5435.961293	5435.961293	-0.033598
21	model_Verso-S	-5160.480812	5160.480812	-0.024951
20	model_Verso	-4581.344032	4581.344032	-0.008590

Descending sort by 'pearson coef'

	variable	linear regression coef	linear regression abs coef	pearson coef
4	engineSize	3586.290122	3586.290122	0.712259
23	transmission_Automatic	399.816178	399.816178	0.512552
28	fuelType_Hybrid	1230.409226	1230.409226	0.477767
0	year	821.557978	821.557978	0.422690
8	model_C-HR	-228.478014	228.478014	0.352703
14	model_Land Cruiser	16146.310113	16146.310113	0.347347
10	model_Corolla	-902.649576	902.649576	0.263434
18	model_Supra	23710.063098	23710.063098	0.255997
17	model_RAV4	-1322.693821	1322.693821	0.240952
2	tax	-3.106702	3.106702	0.222917

Model
Results

model	train_R2	test_R2	train_CV_R2	test_CV_R2
RandomForest (untuned)	0.994761	0.958548	0.962495	0.928231
AdaBoost (tuned)	0.960136	0.953608	NaN	NaN
DecisionTree (tuned)	0.950412	0.948029	NaN	NaN
RandomForest (tuned)	0.940938	0.942876	NaN	NaN
DecisionTree (untuned)	0.999848	0.935710	0.935955	0.883428
DecisionTree (untuned)	0.999848	0.935063	0.936213	0.885365
LinearRegression	0.927633	0.909811	NaN	NaN
AdaBoost (untuned)	0.514315	0.438232	NaN	NaN
SVM (untuned)	0.007697	0.015125	NaN	NaN
SVM (tuned)	-3.124257	-3.614367	NaN	NaN

- Of the 5x types of models we've fit, Random Forest and AdaBoost performed the best in terms of how accurate of price predictions.
- They are both ensemble algorithms that aggregates multiple decision trees.
- Decision tree is seen as the third best performing model.

Application

- Stage 1:
  - A small program/script that can be installed on the salespeople’s computers to calculate an estimated price after inputing information relevant to the vehicle (model, transmission type, engine size, etc.)
- Stage 2:
  - An API that interfaces with the existing dealership inventory system to automatically calculate the estimated price.
- Stage 3:
  - Including stage 3 and approval feature directory in the inventory system allowing managers to override the sale price if deviates by a certain percentage from the estimated price.

Limitations

- Only has Toyota vehicle data, thus the model may not perform as well on other car brands.
- The data only includes a limited set of car characteristics and should take consideration of attributes that weren’t considered in the model (i.e. road salt corrosion, bumper dents, past repair history, etc.)
- Customer preferences were not taken into consideration, such as the change in color trend preferences.

Further Improvements

- Incorporate data from other car brands.
- Include CarFax history records into the model.
- Improve SVM by adding more attributes to reduce the sparsity of the data matrix.

Data Science Manager Report

Methodology walk through as supplement to ipynb

Anthony Lee 2022-July-01

Situation

- Discount Motors, a UK-based auto dealership is seeing an 18% decline in sales during the recent months after a large number of junior salespeople are hired.
- Management at Discount Motors hypothesized the decrease in sales to be a result of inconsistent pricing from the large number of junior salespeople with limited experiences.
- Thus, Discount Motors is asking for a tool that could assist these junior employees by predicting an estimated price for a vehicle.

Data

- 6738 entries of Toyota vehicle attributes and list price of some other retailers.
  - Vehicle attributes: model, year, transmission_type, mileage, fuel_type, tax, mpg, and engine-size
  - Target variable: price (in British Pound)

Methods
Outline

- Clean data
  - Impute/Remove missing data
  - Check data veracity and fix inaccurate data
- EDA:
  - Pairplot to observe the pair-wise scatter plot
  - Pair-wise correlation heatmap
  - Observe data types
  - Dummify categorical variables (model, transmission, and fueltType)
- Model Training:
  - Linear Regression
  - Decision Tree
  - Ensemble (AdaBoost, Random Forest)
  - Support Vector Regressor (SVR)

Methods
Clean Data

- The data is read in from a locally stored CSV using the pandas library.
- The library automatically labels the data type, and it manually verified using `df.info()`
- Missing data is first checked using `pd.DataFrame.isna().sum()` to missing values per column, then reviewed using `pd.DataFrame.value_counts()` for missing values not recognized by the pandas functions or inaccurate values.
- 6x entries were recorded with engineSize of 0, which is inaccurate as all these vehicles should have engineSize greater than one (unless they are electric vehicles using motors.)
- Models names were also found to have unnecessary whitespaces that were stripped using `pd.Series.str.strip()`

Methods

EDA

- `seaborn.pairplot()` was used to contruct a pairwise distribution plot for each variable and further color coded by fuelType for more granular view of the distributions.
- Pair-wise correlations between variables were found using `pandas.DataFrame.corr()` and presented as a heatmap using `sns.heatmap()`.
- Observations and results are written in the ipynb notebook next to the plots.
- To ready the variables for model trainings, categorical variables (e.g. model, transmission, and fuelType) are converted to dummy variables using `pd.get_dummies()`.
- Test data set was created using `sklearn.model_selection.train_test_split()` to hold out 20% of data for testing.

Model Training

Model 1 - Linear Regression

- Residual Squared (R2) was used as a model performance metric as the data is a regression data.
- The training set was fit to an `sklearn.linear_model.LinearRegression()` model and cross-validated using `sklearn.model_selection.cross_val_score()`
- Pearson coefficient is calculated using `sklearn.feature_selection.r_regression()`
- The linear regression coefficient for each variable and the Pearson correlation coefficient to the price variable are concatenated into a dataframe to be sorted.
- We observe variables with large absolute regression coefficient and use the +/- to determine the direction of influence of a variable on the listing price (e.g. increase/decrease the listing price per unit change of a variable.)
- We also observe variables with the strongest correlation with the price variable.
- The linear regression was able to achieve R2 = ~0.927 for the test data set, relatively high for a simple model.

Model Training

Model 2 - Decision Tree Regressor

- Utilizing the `sklearn.tree.DecisionTreeRegressor()`, the data is fit to the default hyperparameters and later tuned.
- The DT was tuned using `sklearn.model_selection.RandomizedSearchCV()` and with the best estimator, the R2-score improved from 0.885 to 0.959 after tuning

Model Training

Model 3 - (Ensemble) AdaBoost

- Because of the amount of information captured by the DT model, I am interested in further improving the DT using a forward learning algorithm, AdaBoost.
- The AdaBoost with hyperparameters tuned has an R2-score of ~0.954 on the test dataset, not significantly different from the performance of the tuned decision tree model.
-

Model Training

Model 4 - Random Forest Regressor

- An alternative ensemble method of a decision tree is the Random Forest algorithm.
- The R2-score of the tuned Random Forest on the test dataset is ~0.970, an improvement from the performance of the decision tree model.

Model Training

Model 5 - SVM (Support Vector Machine)

- Lastly, I attempt to train a support vector machine (regressor), however, the performance is very low (R2-score on training set ~0.011; R2-score on testing set ~0.005).
- The low performance of a SVM is most likely due to the sparsity of the dataset after dummifying the categorical variable.

Application / Deliverable

- For our limited purpose, we can wrap the prediction model in a function that accepts model, year, transmission_type, mileage, fuel_type, tax, mpg, and engine-size as inputs.
- The function can be wrapped as a pickle file that can be imported to any computer using a BASH script.
- The BASH script will set up a local price predictor for each junior salesperson as the base-line price and prohibits sale price to be more than £1500 above the predicted price without manager approval.

Limitations

- The model is only trained on a limited data set of Toyota vehicles only and thus may not be as accurate on vehicles of other makers.
- The model does not consider geographical information and only considers a limited set of vehicle characteristics.

Future Improvements

- Incorporate data from other car manufacturers as car brands are known to be a major factor of a customer’s willingness to pay.
- Include CarFax information on the vehicle so to consider more tailored attributes of a vehicle’s price.
- Package the model as an API allowing the dealership’s inventory system to auto query an estimated price for each vehicle.