ISTANBUL TECHNICAL UNIVERSITY
FACULTY OF MANAGEMENT

Decision Theory (END327E)
2021-2022 Fall

Instructor: Prof. Dr. Burhaneddin Sandıkçı

TERM PROJECT

Car Price Prediction with Decision Tree Algorithm & Determination of Second-Hand Car Buying Preferences with TOPSIS Algorithm

Naz İrem Baz - 070180139
Rıza Semih Koca - 070190704
Berk Özgür - 070180123

Contents

1. The Motivation of The Study and Problem Definition

Purchasing a new car on the pre-owned car market is a difficult decision to make owing to the wide range of operational, physical and technical parameter requirements such as fuel consumption, design, body size, engine capacity, technology employed, and many others. As a result, selection procedure strategies are necessary to address this challenge. Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) is one of the selection procedure strategies utilized in this study as decision-making model. This method serves as a foundation for decision-making procedures in situations where there are number of options each has a high number of qualities (Srikrishna et al., 2014). The automobiles analyzed for this project were rated according to their classification. In such an active and capital market, it is essential to value vehicles correctly in line with the criteria and to determine the optimum values of the cars to trade at the right price (Çelik & Osmanoğlu, 2019).

In this study, up-to-date car ad data which is acquired from https://otoeksper.com.tr with Web Scraping is taken into consideration, the problem of choosing a car according to the customer's preferences has been handled. The purpose of the study; to suggest the most suitable advertisements for the customer's preferences and to decide whether the current price of the advertisement is appropriate by comparing with the predicted prices.

TOPSIS method, one of the multi-criteria decision making techniques, was applied for the selection problem and evaluations were made regarding the results.

2. Problem Solution (Methodology and Application)

a. Data Collection and Preprocessing

i. Web Scraping

Before setting up the predictive model, it is needed to collect the data. We got the data from a used car sales site (https://otoeksper.com.tr) to be more realistic rather than using readymade data. The reason why we chose this site in particular was that it included the appraisal report, so we obtained more data from a single tool. However, the disadvantage of this site was that there were relatively less ads which caused overfit problems in decision tree model. We collected all the advertisements on the site in about 5 hours using a virtual browser using the python language and the selenium library.

ii. Data Preprocessing

Since we pulled the data from the site as strings, we arranged the necessary data types via excel. If we had a large number of data, we could also do this using the pandas library. However, since the number of data is small, we performed this operation via excel. Apart from this, we had to apply the encoding process in order to include categorical data in the model. We

digitized categorical data by label encoding. We did not apply One-hot-encoding so that the matrix size is not large.

b. Decision Tree Algorithm

i. Decision Tree Algorithm, What-How-Why ?

Decision Tree is a regression method that graphically explains the relationship between decisions and opportunities and allows important decisions to be broken down into a series of decisions. To decide which regression to use in the model, we tried randomforest, xgboost, knn, logistic regression and compared error and accuracy values. Finally, we applied cross validation to find the parameters that optimize the model.

ii. Feature Selection

As for the features we use in the model for price estimation:
Brand and Model, Fuel, Usage, Gear, Body Type, Model Year, Mileage, Fuel Consumption, Engine Volume.

iii. The Application

In the application phase, as seen in the *Figure 1* below, when the features of the vehicle are entered, we want to sell to the installed model in numerical form, the model gives us its estimated price. For example, the model we built for the vehicle below estimates the price of the vehicle as 350.950 TL.

```
a = ['Renault Symbol Symbol Collection1.2 16V SL Collection',18,189.000,1,1,1,1,1,75,1.149]
a = [287,18,189.0,1,1,1,1,1,75,1.149]
a = np.array(a) # convert to a numpy array
a = np.expand_dims(a, 0) # change shape from (8,) to (1,8)
dt_model.predict(a) # voila!
```

```
array([350950.])
```

Figure 1

iv. Results

As can be seen in the correlation matrix in *Figure 2* The Correlation Matrixthe features that increased the price the most were the automatic transmission of the vehicle, low mileage, and high model. Apart from this, although the correlation between the variables creates a multicollinearity problem, the predictive ability of the model is considerably high.

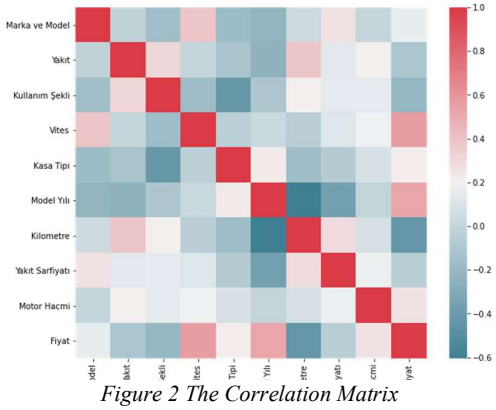
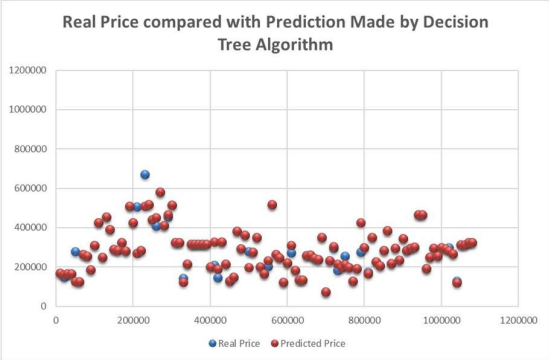


Figure 2 The Correlation Matrix



However, as seen in the results in *Figure 3* Real Price Compared with Prediction made by Decision Tree Algorithm, overfitting occurred in the model due to less data. To overcome this, better results can be obtained with more data.

Figure 3 Real Price Compared with Prediction made by Decision Tree Algorithm

c. Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS)

i. Multi Criteria Decision Making Algorithms

The output of the MCDM problem is a "Decision". This output can be in the form of a best-compromised solution or a ranked (sorted) list of alternatives. The inputs of the problem consist of a sign that tells the decision maker the need to make a decision and initiates the decision-making process, and data that helps explain the decision situation (Ozernoi & Gaft, 1978).

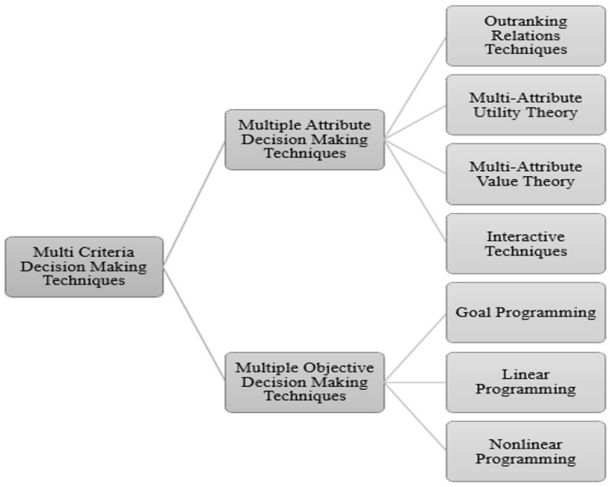


Figure 4 Classification of Multi Criteria Decision Making Techniques

ii. TOPSIS Algorithm, Usage Areas

TOPSIS is a mathematical solution for MCDM problems. It works by finding the ideal and worst solution, then ranking the alternatives starting with the alternative that has the least Euclidean distance to the ideal solution and most distance to the worst solution. TOPSIS is used primarily in ranking problems and choice problems.

iii. Why TOPSIS?

There are a few features that make TOPSIS a unique method and a comparison between the other MCDM methods is given in *Table 6* Comparison of MCDA Methods (Zlaugotne et al., 2020)TOPSIS can take into account both qualitative and quantitative criteria. TOPSIS is a compensatory method which means that the loss in one criteria can be compensated by the gain in another criteria, which creates a more realistic ranking model than outranking and non-compensatory models. (Greene et al., 2011)

iv. Feature Selection and Data Preprocessing

The features which are used for the TOPSIS Algorithm are vehicle age, mileage, fuel consumption, vehicle size, engine capacity, price and derived vehicle general health score. In the selection of these features, weight is given to the features that affect the customers' second-hand vehicle purchase preferences.

“Vehicle overall health score” and “vehicle size” columns have been created by "digitizing" as seen in the tables below, in order for the algorithm to work properly. After the status information of 11 different parts of the vehicle (hood, left front door, trunk, etc.), 7 different status levels describing the health status of the part (damaged, crushed, fine etc.) are ordinally digitized, the resulting values are averaged and the overall score was calculated by creating a new column named “Vehicle overall health score”.

Table 2 Points for Car Health Status

Status	Point
Damaged	1
Changed	2
Crushed	3
Painted	4
Locally Paint	5
Scratched	6
Fine	7

Table 1 Points for Vehicle Size

Vehicle Size	Point
Vehicle	1
SUV / 4x4	2
Light Commercial	3
Commercial	4

v. The Application

- Application of TOPSIS is started after the features are selected and data preprocessing is completed. The first step is normalizing the data, which is done by using vector normalization.
- Then, normalized data is multiplied by the weight of each feature determined by the user (between 1 and 5 as in *Table 3* Weightage Preference) which gives the weighted normalized data.
- After that, the best and worst alternatives are calculated by finding the minimum and maximum values for each feature.
- Next, the distance to best and worst solutions is calculated for each alternative.

Weightage Preference	
Weight	**Weight Preference**
Not at all important	1
Slightly important	2
Important	3
Fairly Important	4
Very Important	5

Table 3 Weightage Preference

- Lastly, the TOPSIS score for each alternative is calculated, which is the similarity to the worst solution and alternatives are ranked accordingly.

vi. Results

For the results, we applied TOPSIS using Python. Firstly, a hypothetical customer who wants to buy a car was created, and the application was made according to the importance levels determined by the customer. Importance weights and their impacts for the hypothetical customer are as follows:

Vehicle age: 4 and (-), Fuel Consumption: 2 and (-), Mileage: 5 and (-), Vehicle Size: 1 and (+), Engine Capacity: 1 and (+), Price: 5 and (-), General Health: 2 and (+)

In short, the hypothetical customer is someone who gives a lot of importance to low price, low mileage, an underage vehicle, and respectively gives little importance for low fuel consumption, a good general health, and a big vehicle. The top 5 cars that is suggested for this customer using TOPSIS are given at *Table 4* TOPSIS Scores, Top 5 Vehicles below:

Table 4 TOPSIS Scores, Top 5 Vehicles

	Plaka	Kilometre	Yakıt Sarfiyatı	Araç Büyüklüğü	Motor Hacmi	Fiyat	Araç Genel Sağlık Puanı	Model Yaşı	Distance Positive	Distance negative	Topsis Score	Rank
228	35 RN 125	32172	6.6	1	1368.0	100000	7.000000	27	0.100939	1.418178	0.933554	1.0
230	20 AEV 156	3122	5.2	4	1598.0	213950	6.272727	2	0.113281	1.434755	0.926823	2.0
231	20 AEV 971	3089	5.2	4	1598.0	213950	5.909091	2	0.113618	1.434653	0.926616	3.0
563	34 ESJ 876	4	5.7	3	1368.0	208500	6.545455	1	0.114843	1.438860	0.926084	4.0
571	38 AFF 612	8297	5.7	3	1368.0	207950	6.545455	1	0.115229	1.427481	0.925307	5.0

All of these cars have a high TOPSIS score, so the hypothetical customer can further inspect these 5 cars and make a final decision on their purchase.

3. Results

Increasing competition and innovations in the automotive industry have increased the number of criteria to be considered when purchasing a vehicle. In cases where there are many models and various criteria, customers may have difficulty in choosing a car. In this study, a solution proposal to the problem of purchasing a pre-used car is presented with the TOPSIS method.

TOPSIS method is an easy solution approach for decision making problems. The degree of importance that each customer attaches to the criteria may be different. For some consumers, the fuel waste criterion is more important, while for others, price or model age may be more important. Large families may prefer larger vehicles, while younger users may prefer smaller cars. In summary, the selection process is made according to the preferences of each user. For this reason, the weights of the selection criteria may be different for each consumer.

Today, there are many websites that are built on the purchase and sale of second-hand cars. If the TOPSIS method we applied in the study is added to the sites as a tool, users can make

more effective searches by determining the preference criteria in tool selection. Thus, they provide an ideal service to those who want to buy a car. Additionaly, the price prediction made with Decision Tree Algorithm can guide customers to understand the right price for the sellected model.

4. References

Çelik, Ö., & Osmanoğlu, U. (2019). İkinci El Araba Fiyatlarının Tahmini. *European Journal Of Science And Technology*, 77-83. doi: 10.31590/ejosat.542884

Ozernoi, M., & Gaft, M. (1978). Multicriterion Decision Problems. In Conflicting Objectives in Decisions (p. 18). John Wiley.

Srikrishna, S., Reddy, S., & Vani, S.G. (2014). A New Car Selection in the Market using TOPSIS Technique. *International Journal of Engineering Research and General Science*, 2(4), 177-181

Greene, R., Devillers, R., Luther, J.E. & Eddy, B.G. (2011). GIS-based multi criteria decision analysis. *Geography Compass, 5*(6), 412-432. doi:10.1111/j.1749-8198.2011.00431.x

Zlaugotne, B., Zihare, L., Balode, L., Kalnbalkite, A., Khabdullin, A., & Blumberga, D. (2020). Multi-Criteria Decision Analysis Methods Comparison. *Environmental And Climate Technologies*, 24(1), 454-471. doi: 10.2478/rtuect-2020-0028

5. Appendix

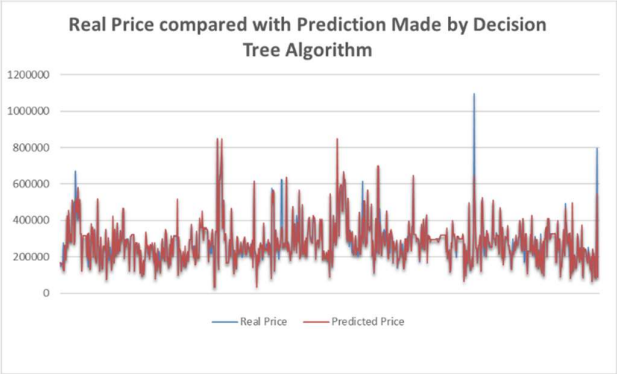


Figure 9 Price Prediction with Decision Tree Algorithm

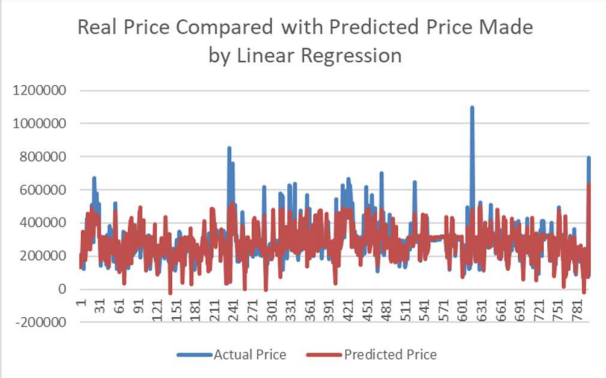


Figure 10 Price Prediction with Linear Regression

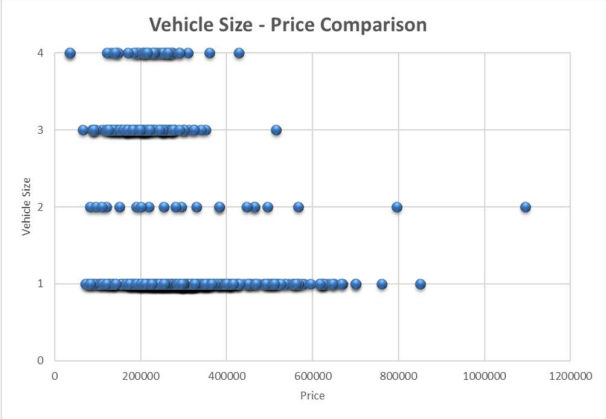


Figure 5 Vehicle Size - Price Comparison

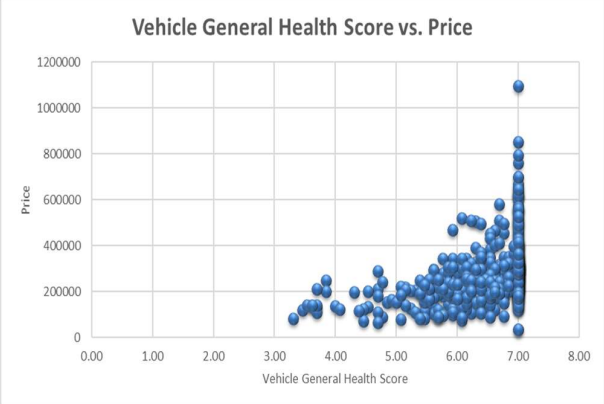


Figure 6 Relation between Vehicle General Health Score and Price

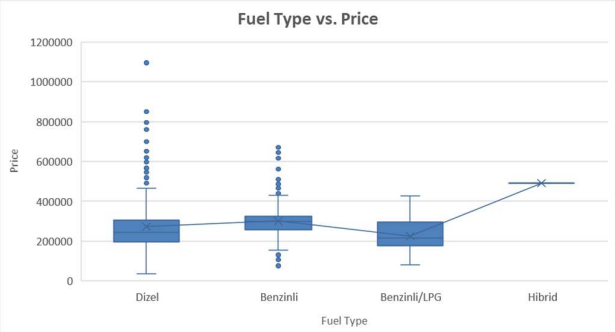


Figure 8 Relation between Fuel Type and Price

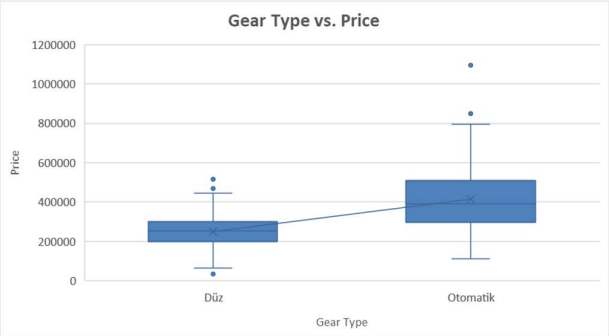


Figure 7 Relation Between Gear Type and Price

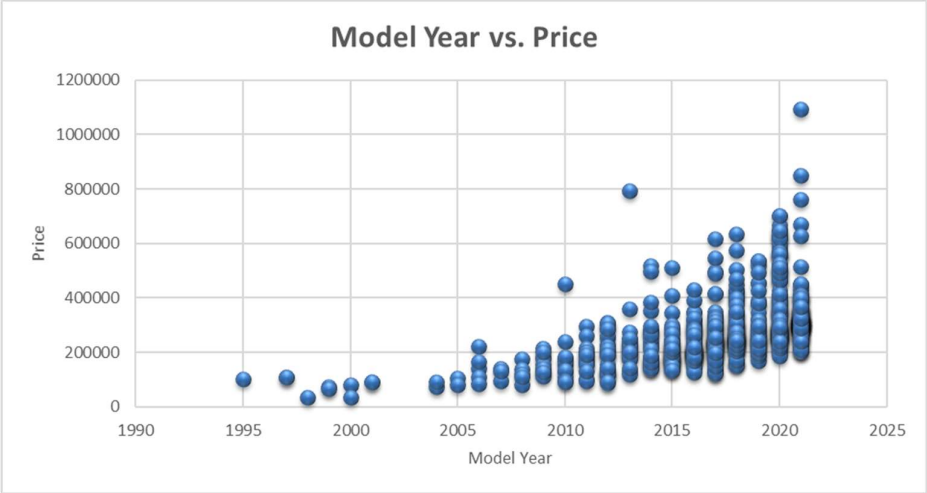


Figure 11 Relation Between Model Year and Price

Table 5 TOPSIS Outputs Top 20 Vehicles

	Licence Plate	Mileage	Fuel Consumption	Vehicle Size	Engine Capacity	Price	Vehicle General Health Point	Model Age	Topsis Score	Rank
228	35 RN 125	32172	6.6	1	1368	₺ 100,000.00	7.00	27	0.934	1
230	20 AEV 156	3122	5.2	4	1598	₺ 213,950.00	6.27	2	0.927	2
231	20 AEV 971	3089	5.2	4	1598	₺ 213,950.00	5.91	2	0.927	3
563	34 ESJ 876	4	5.7	3	1368	₺ 208,500.00	6.55	1	0.926	4
571	38 AFF 612	8297	5.7	3	1368	₺ 207,950.00	6.55	1	0.925	5
570	35 AYK 103	10356	5.7	3	1368	₺ 210,000.00	6.09	1	0.924	6
566	06 CSK 189	790	5.7	3	1248	₺ 220,000.00	7.00	1	0.922	7
532	06 AFR 609	25648	5.7	3	1248	₺ 209,500.00	7.00	5	0.920	8
557	34 DLY 931	35675	5.7	3	1248	₺ 199,900.00	7.00	2	0.918	9
545	34 BNK 175	18363	5.7	1	1368	₺ 197,950.00	6.91	4	0.918	10
126	35 GG 640	38365	7.6	1	1598	₺ 158,000.00	7.00	4	0.918	11
225	34 EDS 658	4559	6.1	4	1598	₺ 229,950.00	4.27	1	0.918	12
454	35 AVP 147	5396	5.2	1	1368	₺ 210,000.00	7.00	1	0.918	13
548	42 ABC 609	26600	5.7	3	1248	₺ 215,000.00	7.00	4	0.918	14
562	09 ABG 905	25129	5.7	3	1368	₺ 219,900.00	7.00	2	0.917	15
137	27 C 8616	20301	5.1	1	1229	₺ 201,950.00	7.00	6	0.917	16
564	34 ETM 282	45	5.7	3	1.36	₺ 205,000.00	5.45	1	0.916	17
538	06 AGR 258	41410	5.7	3	1248	₺ 197,500.00	7.00	5	0.916	18
555	42 AFM 109	36122	5.7	3	1248	₺ 207,000.00	6.45	3	0.916	19
788	38 ABS 862	48992	6.4	3	1498	₺ 186,950.00	7.00	2	0.915	20

Table 6 Comparison of MCDA Methods (Zlaugotne et al., 2020)

	TOPSIS	VIKOR	COPRAS	MULTIMOORA	PROMETHEE-GAIA	AHP
Type of normalization	Vector normalisation (square root of sum (L2 normalization)	Linear normalization (L1 normalization)	Vector normalization (sum)	Vector normalization (square root of sum)	Normalization is performed automatically	Vector normalisation (sum)
Suitability	Choice problems, ranking problems	Choice problems, ranking problems	Choice problems, ranking problems	Choice problems, ranking problems	Choice problems, ranking problems, description problems (GAIA)	Choice problems, ranking problems, sorting problems (AHPsort)
Inputs	Ideal and anti-ideal option weights	Best and worst option weights	Best and worst option weights	Best and worst option weights	Indifference and preference thresholds weights	Pairwise comparison on ratio scale (1–9)
Outputs	Complete ranking with closeness score to ideal and distance to anti-ideal	Complete ranking with closeness score to best option	Complete ranking	Complete ranking	Partial and complete ranking (pairwise outranking degrees)	Complete ranking with scores
Preference function	Distance metric (Euclidean distance, Manhattan distance, Tchebycheff distance)	Distance metric (Manhattan distance)	Min Max	Min Max	Usual, Linear, U-shape, V-shape, Level, Gaussian	
Approach	Qualitative and/or quantitative	Quantitative	Quantitative	Quantitative	Qualitative and/or quantitative	Qualitative
Ranking scale	0 to 1	Positive values	Positive values	Positive values	−1 to 1	0 to 1
Best alternative	Max value	Min value	Max value	Max value	Max value	Max value
Consistency levels	no restrictions	no restrictions	no restrictions	no restrictions	7±2	9
Software	MS Excel, Matlab, Decerns	MS Excel	MS Excel	MS Excel	Visual Promethee, Decision Lab, D-Sight, Smart Picker Pro	MS Excel, MakeItRational, ExpertChoice, Decision Lens, HIPRE 3+, RightChoiceDSS, Criterium, EasyMind, Questfox, ChoiceResults, 123AHP, DECERNS