Predicting the price of second-hand vehicles using data mining techniques

Jafari Kang, Masood and Zohoori, Sepideh and Abbasi, Elahe and Li, Yueqing and Hamidi, Maryam

Lamar University

8 November 2019

Predicting the Price of Second-hand Vehicles Using Data Mining Techniques

Masood Jafari Kang[a*], Sepideh Zohoori[a], Elahe Abbasi[a]
Yueqing Li[b], Maryam Hamidi[c]

[a] Doctoral Student, Lamar University
[b] Associate Professor, Lamar University
[c] Assistant Professor, Lamar University

Abstract

The electronic commerce, known as “E-commerce”, has been boosted rapidly in recent years, and makes it possible to record all information such as price, location, customer’s review, search history, discount options, competitor’s price, and so on. Accessing to such rich source of data, companies can analyze their users’ behavior to improve the customer satisfaction as well as the revenue. This study aims to estimate the price of used light vehicles in a commercial website, Divar, which is a popular website in Iran for trading second-handed goods. At first, highlighted features were extracted from the description column using the three methods of Bag of Words (BOW), Latent Dirichlet Allocation (LDA), and Hierarchical Dirichlet Process (HDP). Second, a multiple linear regression model was fit to predict the product price based on its attributes and the highlighted features. The accuracy index of Actuals-Predictions Correlation, the min-max index, and MAPE methods were used to validate the proposed methods. Results showed that the BOW model is the best model with an Adjusted R-square of 0.7841.

Keywords

Text mining, Topic modeling, BOW, LDA, HDP, Linear regression

1. Introduction

The advent of the internet created electronic commerce, known as “E-commerce”, which has been boosted rapidly in recent years. E-commerce can record all transactions and users’ search histories that be considered as a solid source of data. However, it is not easy to deal with such metadata [1]. Some E-commerce companies like Amazon and eBay are willing to analyze this data to provide their users helpful recommendations and information. The information includes various aspects such as price, location, customer’s review, competitor’s price, and so on.

Along with several types of research on human behavior in real-world cases like transportation networks [2], the study of users’ activities on the “virtual world”, the internet, has been the subject of interest for a number of researchers. Van Heijst et al [3] researched price prediction using eBay data in 2008. They took the number of pictures, ratings, and descriptions of a good in their prediction. Greenstein-Messica & Rokach [4] gathered the 6-month transactional data of eBay to model the consumers’ willingness to pay. They deemed discount indication and seller reputation to generate a recommender system. There are some other related papers in this topic namely [5] and [6].

This research carries out an investigation on Divar database. Divar [7] is an e-commerce famous website used for trading second- hand goods in Iran. For the users of this application, both sellers and buyers, knowing about an acceptable and rational price for their item based on its features is crucially important. Therefore, this investigation tries to estimate the price of the merchandise using the historical data of sellers and develop a pricing model for trading vehicles based on Divar dataset. To do so, we, first, extracted useful results from a huge amount of un-cleaned text data using text mining techniques, and second, fitted a model on price of second-hand vehicle based on our achievements at the first step. One of the main challenges was that all text files in our dataset were in Farsi, the national language of Iran, and thus, we utilized several methods and techniques to overcome this problem.

2. Literature Review

Van Heijst et al [3] converted the description part of items to a numerical format, using Bag of Words (BOW) method. They removed some words with high or low frequencies and created a dictionary. After the cleaning step, this method created a vector of frequency for the description column of data. Finally, by extracting new features, they fitted a model for price prediction. Hong Liu et al [8] developed the sequential bag of words method for human action classification. They divided the entire action into sub actions to focus on different parts of actions in details. Kim and Kang [9] used the Labeled Latent Dirichlet Allocation (L-LDA) to investigate the costumers' idea about Korean products. They use the LDA method to clarify the distinguishing product attributes that their users prefer by studying their feedback. Zhao and his colleagues [10] used LDA method to extract risk factors from the news. Park et al [11] presented a new Hierarchical Dirichlet Process (HDP) model to analyze question answering data. They, first, implemented clustering on questions using a certain number of clusters, and then, tried to find the similarity in each cluster. Zavitsanos et al [12] proposed an HDP-based model to analyze the content of collected documents without prior knowledge. They used their model on a real-world database, and their results showed that the proposed model is flexible, well-fitted to data, and not dependent on language.

3. Methodology

3.1 Bag of Words (BOW)

This technique is used to extract the features from the text documents. In this method, a dictionary is created by using all unique words in the text document. To update the dictionary, BOW removes all words with high frequency such as prepositions and words with low frequency. For every description and row in the document, BOW creates a vector with the frequency of all unique words. This frequency can be zero or one, which means whether the word does or does not appear in the text, or it can be numerical, in which the number shows the number of occurrences in the text. These vectors are inputs for topic modeling [13].

3.2 Latent Dirichlet Allocation (LDA)

LDA is a generative method that prepares a model to define configuration of a document database [14]. LDA is a statistical model which is drawn from Dirichlet distribution and expresses the attributes of documents and topics. The Dirichlet distribution is a multivariate distribution that is generated from Beta distribution, and its probability density function is as below [15]:

$$B(\alpha) = \frac{\Gamma \sum_{i=I}^{k} \alpha_i}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1} \qquad \alpha = (\alpha_1, \ldots\ldots, \alpha_k) \tag{1}$$

Where, K is the number of categories (integer), and α_1 to α_k are concentration parameters.

LDA is based on two important principles "Every document is a mixture of topics" and "Every topic is a mixture of words". LDA is going to estimate both mentioned items simultaneously and to figure out the mixture of words that are in association with each topic while demonstrating the mixture of topics explaining each document [15]. This technique categorizes a document, as a corpus of words, into topics that best reflect the corpus's word dependencies. Thus, LDA can be used to extract topics from documents that can be comments, description or any other types of text.

3.3 Hierarchical Dirichlet Process (HDP)

One of the most important parts of machine learning is the selection and adaptation of a model. There are two different approaches toward the model selection and adaptation: parametric and nonparametric models. In the parametric models, the data size does not affect the number of parameters in a model, while nonparametric models make it possible to have more parameters in a model along with increasing the size of data [16]. One of the most popular types of nonparametric models is the Dirichlet Process (DP) mixtures which are unsupervised models considering an infinite number of clusters. Hierarchical Dirichlet Process, known as HDP, is defined as an extension to DPs. HDP considers a document as a corpus of words, or topics, and finds the proportion of each topic [17]. DP depends on two parameters of α_0, scaling parameter, and G_0, base probability measure, and as a result, a random probability G can be presented by the DP distribution of α_0 and G_0 [18].

$$G \sim DP(\alpha_0, G_0) \tag{2}$$

4. Experiment and Result

In this research, transaction data of a commercial website, Divar, is used. Divar is a popular website for trading second-hand goods. This dataset includes users' posts representing attributes of a product in Farsi and its price. Apart from the numerical and categorical attributes, there are two columns that are noticeable: the product price and its description.

Therefore, the aim of the present study at first is to extract highlighted features from the description column using the three different methods of BOW, LDA, and HDP, and second, to fit a multiple linear regression model which predicts a product price based on its attributes and the features obtained from the product description. To do this, after gathering data, the next step is data preprocessing where the data will be read, and redundant characters will be removed from the description column. The next phase goes to text clustering which applies the three methods of topic modeling to find clusters. Finally, in the modeling section, using the clustering output and data attributes, a data frame for modeling the product price is created. Figure 1 shows our stepwise procedure to find the best model. In this investigation, *R-3.5.۳ for Windows* and its packages are utilized.

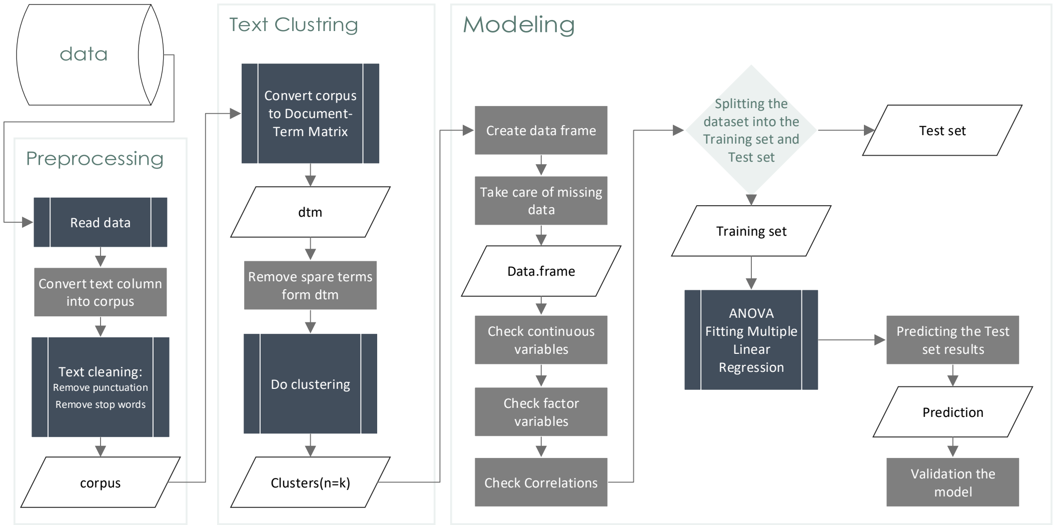
data
Preprocessing
Read data
Convert text column into corpus
Text cleaning:
Remove punctuation
Remove stop words
corpus
Text Clustring
Convert corpus to Document-Term Matrix
dtm
Remove spare terms form dtm
Do clustering
Clusters(n=k)
Modeling
Create data frame
Take care of missing data
Data.frame
Check continuous variables
Check factor variables
Check Correlations
Splitting the dataset into the Training set and Test set
Test set
Training set
ANOVA Fitting Multiple Linear Regression
Predicting the Test set results
Prediction
Validation the model

Figure 1: the flowchart of text clustering and regression modeling

4.1 Data

Divar's dataset, which is reported in January 2019, contains 947635 posts published and archived before 2017. Table 1 represents the data attributes and their definitions with an example based on Divar's information [19]. As it was mentioned before, the sake of simplicity, only light vehicles are investigated in this research.

Table 1: Data attributes and definitions

Attributes	Definition	Example
ID	product unique id	1246772
archive_by_user	a Boolean variable representing if user archived the post or not	True/False
Published_at	the publishing date of posts	Saturday 07 PM
Cat1	the product category at level I	vehicles
Cat2	the product category at level II	cars
Cat3	the product category at level III	light
Title	the post title	پراید هاچ بک مدل ۷۹
city	the city where the post is published	Mashhad
Desc	the product description	لاستیکها نو -موتور سالم-بیمه
Price	The product price in Iran currency (Toman)	9800000
image_count	Number of images of the product	5
platform	The platform used for posting (mobile or web)	mobile/web
Mileage	The product mileage in kilometers (only for cars)	200000
Brand	The product brand in English and Farsi	Pride:پراید هاچبک
Year	The production year based on the Persian calendar. (only for cars)	1379

4.2 Preprocessing

4.2.1 Text cleaning

The first step in text clustering is text cleaning. To do so, a collection of documents, called Corpus, should be created and to clean the corpus, all punctuations, numbers, redundant words or "stopwords" should be removed. Stopwords refer to the words and letters which are frequently used in manuscripts, but however, they do not affect the concept, for example *and, are, do*, etc.

4.2.2 Text Clustering: BOW, LDA and HDP

In BOW method, all words in the whole text document should be recognized and separated. But working on a single word may not lead to remarkable results, because the meaning of the whole text phrase in the description part of the products may be lost. Hence, in this analysis, all pair words in the text document are found and the most frequent ones are used. To do so, the first bag of words including the top 14s frequent words (n> 8500) is generated and all combinations of two words (a pair) in the text document are used. Therefore, a list of pair words as the bag of words is created and finally 8 most frequent pairs as the new features are extracted. These features are included *No Color Damage, Repainted, No Technical Problem, Insurance Discount, Insurance Due, New Brand Tire*, *CNG, Upholstery*. In the next step, the matrix of the frequency of these 8 pairs in all data description columns is generated. The cells of the matrix are including zero and one, showing whether the word appears in the text or not.

To use LDA, first, the corpus of documents should be made and change to a data frame. Next, a Document Term Matrix (DTM) out of single-word token should be created. In this research, six topics are considered for the LDA model and by its result, the frequency of the most common terms within each topic is extracted. Yet, the single-word token analysis may not be very efficient for this study, because the frequency of a single word can have two different meanings. Thus, an n-gram token is tired to analyze the pair of words. According to the n-gram LDA output, 10 features of *No Color Damage, Repainted, Insurance Discount, No Technical Problem, Insurance Due, New Brand Tire, CNG, Fender Painted, No Accident, Upholstery* are founded.

In HDP method, a DTM from the document corpus which is the output of text cleaning part should generate. The object of the text clustering is to find and cluster the terms that are used frequently in the description column. Therefore, the spare terms form the DTM are removed and then the clustering is run. The output of the HDP clustering is 7 different clusters which represent 7 features for a car in this study. These features are *Repainted, Insurance Discount, CNG, Upholstery, No Technical Problem, New Brand Tire, Fender Replacement*. The output of HDP clustering is used to create a matrix of features with 7 columns for each car, and each cell would be 1 if the car has a feature, or otherwise is 0.

4.3 Regression Modeling

4.3.1 Data frame creating and variable checking

To start regression modeling, a data frame of data attributes, including price, location, image, mileage, age, and the feature matrix is created. In the Divar's database, age (year) is presented by the production year of a vehicle in the Persian calendar, but they are converted into the lifetime of the car in a data frame. Besides, to take care of missing variables, missing numeric values are removed but for categorical ones, empty cells are changed to *"other"*. Also, the distribution of both numerical and categorical variables is checked, and continuous variables are standardized.

- Numerical variables:
  Regarding the Divar's database, price, mileage, and age are numerical variables. Price is considered as the response variable, and its distribution can directly affect the model. Figure 2 presents the histogram plot of the price column, and it shows that there are some large prices which need to be removed. Then, a lower and upper bound for the price is defined and the outliers are removed. Figure 3 presents the revised histogram.

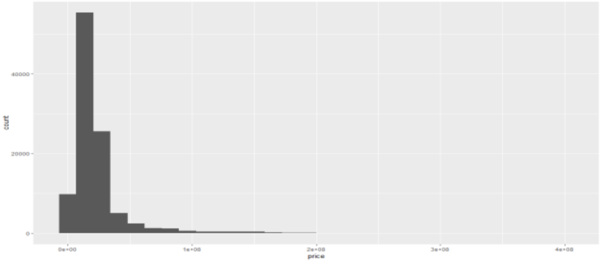

Figure 2: Price histogram before removing outliers

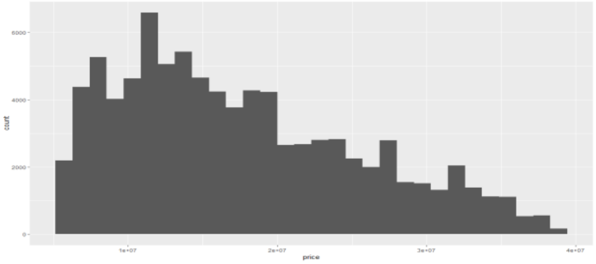

Figure 3: Price histogram after removing outliers

The same procedure is done on mileage and age. After removing all numerical variables outliers, they should become standard which is done by the following equation.

$$\delta_i^{\text{standardaized}} = \frac{\delta_i - min(\delta)}{max(\delta) - min(\delta)} \tag{3}$$

- Categorical variables

  Location and brand are categorical variables in which their frequency plot based on each category should be drawn. As can be seen in Figure 4 and 5, since drawing a conclusion based on a few records is very hard, groups with low frequencies are aggregated with each other.

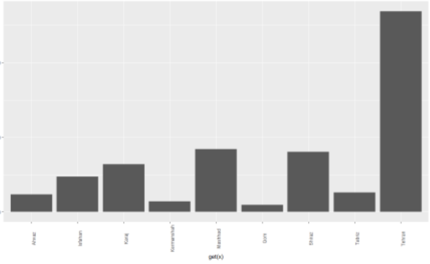

Figure 4: Location frequency plot before aggregation

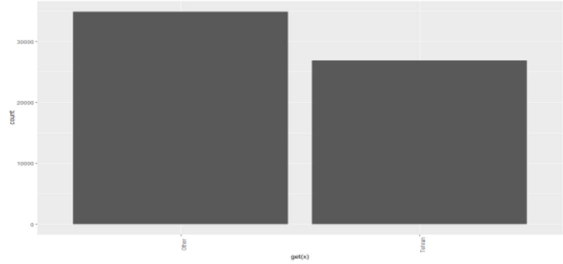

Figure 5: Location frequency plot with new categories

4.3.2 Regression and model validation

In this study, 80 percent of data are considered as a training set and are used to fit a multiple linear regression model and the rest of them deemed as the test set. As it has indicated in equation 4, multiple linear regression is employed to model the relationship between price as a dependent variable, and other predictor variables of both data frame and description part.

$$\begin{gathered} price \sim age + age^2 + age^3 + mileage + mileage^2 + mileage^3 + age.mileage + age^2.mileage^2 + \\ age^3.mileage^3 + brand + location + image + \\ X_{Repainted} + X_{InsuranceDiscount} + X_{CNG} + X_{Upholstery} + X_{NoTechnicalProblem} + X_{NewBrandTire} + X_{FenderReplacement} \end{gathered} \tag{4}$$

Then, gradually, ANOVA is used to reduce the factors which do not have a meaningful effect on price. Table 2 represents the summary of final models which is developed by using three methods of BOW, LDA, and HDP.

Table 2: Regression model summary for BOW, LDA and HDP methods

Methods	Residual Standard Error	R^2	Adjusted R^2	P-value
BOW/Regression	0.1027	0.7842	**0.7841**	< 2.2e-16
LDA/Regression	0.1035	0.7809	0.7808	< 2.2e-16
HDP/Regression	0.1032	0.7824	0.7823	< 2.2e-16

As the final step, to validate the model, the regression models are used to predict the price of the test subset. In other words, prediction results are compared with the actual prices in the test subset. To so doing, there are various methods like Actuals-Predictions Correlation, Min-Max Accuracy and MAPE (Mean Absolute Percentage Error), which Actuals-Predictions Correlation is used here. This method finds the correlation between the actual values and the model predictions. It declares the closer to 1 the correlation gets, the more validate the model becomes. Table 3 shows the outputs for three methods which shows the HDP method has the highest amount of Actuals-Predictions Correlation which means the actual values and model prediction is best describes by HDP regression model.

Table 3: Model validation for BOW, LDA and HDP methods

Methods	Actuals-Predictions Correlation
BOW/Regression	0.8828113
LDA/Regression	0.8803158
HDP/Regression	**0.8846973**

5. Conclusion and Discussion

In the present study, three different models were developed to estimate the price of second-hand light vehicles using the database of Divar's website. The models were built in two steps: first, text modeling methods including BOW, LDA, and HDP were used to extract car features based on sellers' description, and next, a multiple linear regression model was fitted on a data frame including vehicle attributes from the database and obtained features from the text mining. The first contribution of this papers was adapting topic modeling techniques for Farsi, as a foreign language. Thus, we encoded all text files to "UTF-8" format, defined the list of "stopwords" for Farsi, and moreover, analyzed the semantic of both single words and their combinations. Scoendly, our results showed that BOW model was the best model based on adjusted R^2. But the accuracy index of Actuals-Predictions Correlation showed that the regression model based on HDP, was more valid. Considering that the adjusted R^2 was very close, it can be concluded that HDP regression method was more robust for future prediction of light vehicle's price based on Divar's information.

6. References

[1] J. Guo, Z. Gao, N. Liu and Y. Wu, " Recommend products with consideration of multi-category inter-purchase time and price," *Future Generation Computer Systems,* vol. 78, pp. 451-461, 2018.

[2] M. E. M. M. Ilbeigi, "Statistical Forecasting of Bridge Deterioration Condition," *Journal of Performance of Constructed Facilitie,* vol. 34, p. 04019104, 2020.

[3] D. Van Heijst, R. Potharst and M. Van Wezel, "A support system for predicting eBay end prices," *Decision Support Systems,* vol. 44, no. 4, p. 970–982, 2008.

[4] A. Greenstein-Messica and L. Rokach, "Personal price aware multi-seller recommender system: Evidence from eBay.," *Knowledge-Based Systems,* vol. 150, pp. 14-26, 2018.

[5] M. Gorgoglione, U. Panniello and A. Tuzhilin, "Recommendation strategies in personalization applications," *Information & Management,* 2019.

[6] H. Hwangbo, Y. S. Kim and K. J. Cha, "Recommendation system development for fashion retail e-commerce," *Electronic Commerce Research and Applications,* vol. 28, pp. 94-101, 2018.

[7] "Divar," [Online]. Available: https://divar.ir/.

[8] H. Liu, T. Hao, X. Wei, G. ZiYi, T. Lu and G. Yuan, "Sequential Bag-of-Words model for human action classification," *CAAI Transactions on Intelligence Technology,* vol. 1, no. 2, pp. 125-136, 2016.

[9] S. Kim and J. Kang, "Analyzing the discriminative attributes of products using text mining focused on cosmetic reviews.," *Information Processing & Management,* vol. 54, no. 6, pp. 938-957, 2018.

[10] L.-T. Zhao, S.-Q. Guo and W. Yi, "Oil market risk factor identification based on text mining technology," *Energy Procedia,* vol. 158, pp. 3589-3595, 2019.

[11] S. Park, D. Lee, J. Choi, S. Ryu, Y. Kim, S. Kown, B. Kim and G. G. Lee, "Hierarchical Dirichlet Process Topic Modeling for Large Number of Answer Types Classification in Open domain Question Answering," *Information Retrieval Technology,* pp. 418-428, 2014.

[12] E. Zavitsanos, G. Paliouras and G. A. Vouros, "Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes," *Journal of Machine Learning Research,* vol. 12, pp. 2749-2775, 2011.

[13] L. Lin, L. T. Wen Dong, S. Yao and Z. Wei, "An overview of topic modeling and its current applications in bioinformatics," *SpringerPlus,* vol. 5, no. 1: 1608, 2016.

[14] D. M. Blei, A. Y. Ng and M. I. Jorda, "Latent dirichlet allocation," *Journal of machine Learning research,* vol. 3, no. Jan:, pp. 993-1022, 2003.

[15] . M. Ponweiser, Latent Dirichlet Allocation in R. Theses / Institute for Statistics and Mathematics, Vienna.: WU Vienna University of Economics and Business, 2012.

[16] P. Orbanz and Y. W. Teh, "Bayesian nonparametric models," in *Encyclopedia of Machine Learning*, Springer Link, 2010, pp. 81-89.

[17] C. Wang, J. Paisley and D. Blei, "Online variational inference for the hierarchical Dirichlet process," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011.

[18] Y. W. Teh, M. I. Jordan, M. J. Beal and D. Blei, Hierarchical Dirichlet Processes., 2005.

[19] "Cafebazaar," [Online]. Available: https://research.cafebazaar.ir/visage/divar_datasets/.