Used Car Pricing and Beyond: A Survival Analysis Framework

Ayhan Demiriz
Gebze Technical University
41400, Kocaeli, Turkey
Email: ademiriz@gmail.com

***Abstract*—A significant part of the overall automotive market is derived from the used car trade. Determining correctly the used car market values will certainly help achieving fairer trade in many economies. By using the web listings as a proxy data source, we can create some models for the used car pricing based on the asking prices listed in the web adverts. This type of data acquisition requires a thorough data cleaning process to generate dependable statistical models after all. This paper proposes a survival analysis based approach to study the lifetime of the used car listings that can be found at web sites like Craigslist. Pricing models can be easily built to determine the market values of the used-cars from this type of data. One of the most important assumptions in our approach is to consider the delisting of an advert as a sale event. This is also equivalent to the death in the survival analysis context. Since the collected data have labels in terms of sale or not, we can utilize the predictive models to determine whether a particular car at a certain price will be successfully sold or not.**

Keywords: Used-Car Pricing & Survival Analysis & Cox Regression & Right Censored Data & Lifetime Data & Web Classifieds.

I. Introduction

Automotive industry is an essential part of the world economy. It is very common to observe that the used car sales surpasses the new ones across the globe. For example, according to OICA[1] statistics there were around 17.5M new vehicle sales in 2017 in the USA. On the other hand, used vehicle market had a size of approximately 39.2M in the same year according to an Edmunds report[2]. Therefore it is important to study the used-car market where price is one of the most important determinants of sales [1]. Setting the asking (listing) price high may discourage any potential buyer, which practically hinders the buyer to visit the dealership and/or contacting the private seller. On the other hand, lowering the asking price too much will definitely expedite the sales by lowering the profit of the seller/dealership [2]. Online search practically enables buyers to make better choices over the durable goods. Thus, it reduces the effects of information asymmetry [1] which in turn lowers the prices of the items listed [3]. By reducing information asymmetry, buyers can make more informed decisions. We propose a methodology and introduce a web site as a decision support tool to aid buyers/sellers for determining the used-car prices. The methodology can potentially be also used for pricing many different types of second-hand items including real estate.

Lifespan of web listings may resemble lifespan of any subject that is observed during a survival analysis study. In survival analysis [4], subjects (laboratory animals, people, light bulbs, machines, and so on) are followed during the experiment until a specific event occurs (in most cases failure or death) or until the experiment ends. The data in consideration are primarily event times as well as some characteristics of the subjects that may or may not change over time. The data for subjects that have not experienced the event (such as those who survive) are censored. The exact event time is not known for the censored data, but it is known to have occurred after the censored time. Survival analysis is highly popular in medical sciences, especially in new treatment and related drug studies.

We consider the used-car listings from an e-commerce site as our subjects and we follow them until they are delisted or a 30-day free-listing period expires. We assume that the delisted cars are sold at the time of delisting. Basically, delisting i.e. sale of a car is considered as death (or failure) in survival analysis analogy. If the thirty-day period is expired without delisting (i.e. death), then censoring happens for the web listing. Therefore we can utilize survival analysis to analyze the collected data.

In our work, we analyze the event time (time to delist) via survival analysis. Asking price is certainly a factor for determining the time to delist. A car may be sold at a certain price level for a given listing. It is worth noting that the owners of the listings at e-commerce site (i.e. sellers) may update the asking price any time. In other words, we can claim that the listing will stay on the e-commerce site until the asking price is right for a potential buyer or until it is reduced to a willingness-to-pay level of a potential buyer.

As an extension, once the the data from web listing are collected, we can develop a decision support tool to analyze pricing data and build predictive models of the sale event. This paper also introduces a web-based decision support tool to aid buyers/sellers for assessing their cars' market values (asking prices). This particular tool can also aid to determine the likelihood of selling an advertised car within a predetermined time period (thirty days in our case) given the asking price and the specifics of the car.

[1] http://www.oica.net/category/sales-statistics/
[2] https://bit.ly/2IHLpnr

978-1-5386-9209-7/18/$31.00 ©2018 IEEE
DOI 10.1109/ai4i.2018.00023

The rest of this paper is organized as follows: Section II introduces our methodology. Section III introduces usage of survival analysis model. We extend our statistical models to predict the probability of selling a car given specifics by using logistic regression in Section IV. Finally, Section V concludes our work with a discussion.

II. Methodology

Our methodology is composed of two stages. In the first stage, relevant data from web listings are collected from used-car sections of e-commerce web sites such as Craigslist, Oodle.com, Vast.com, and eBay. In the second stage, we then build pricing models for second hand cars based on the data collected in the first stage. In our paper, we implemented our approach for used-cars but it can be generalized to any items sold through web classified advertisements (listings).

The idea of using survival analysis for used-car related data analysis is not new [2]. It was primarily used for determining time to event in used-car context [2] as well as in new car context [5]. In other words, survival analysis can be used for determining the selling times of the used-cars [2] too.

A. Data Acquisition

As the internet has changed so many things around us, it practically affected used-car markets as well. First of all, the internet reduced the cost of getting information about the used-cars on the market dramatically [6], [3]. On the other hand, the internet offers successful referral services helping buyers to redeem the additional discounts on used-cars [3]. In a traditional setting, car dealers are more informed about the used-car market than buyers which evidentally creates information asymmetry. There is strong empirical evidence that the internet enables the car buyers to be more informed [3], [6] regarding the prospective used-cars on the market which practically alleviates the negative effects of information asymmetry about the used-cars. Thus, the internet has effectively replaced the traditional media such as newspaper and magazine listings around the globe for the used-car classifieds. In contrast to some early work [7], [8] that used newspaper listing as their data source; it is now common to collect used-car related data from the internet-based sources [9]. In order to drive price response function correctly, we need to follow the prices of used-cars while they are on the market [2].

Considering that used-car web sites such as mobile.de, autotrader.com, motors.co.uk are popular destinations around the globe, we have collected data from the most popular e-commerce site for the second-hand items (e.g. cars, computers, real estate and so on) in Turkey, Sahibinden.com (translates as "by owner"). In this study, we observed the used-car listings up to thirty days to collect car related data including sellers' location and price changes. Between January 1, 2018 and May 1, 2018, new listings were selected each day to gather data from the e-commerce site, sahibinden.com. In addition to the car listings each day, we fetched web data to observe the car listings that were previously added to our database. Basically each used-car listing was observed up to thirty days or until they were delisted. Approximately 830K car listings were observed in our data acquisition stage. In addition to some crucial data about the listings, such as make, model, trim line, year, odometer reading, listing date, location and price, buyers can see other information about the cars including engine size and power, transmission type, color, body type, owner type (private, dealer), trade in possibility and warranty at this particular e-commerce web site. There is also a free text area (box) for the owners to give additional information about the car on the web page. We did not fetch data from the text area for this particular study.

In preliminary data acquisition stage, we noticed that dealers have a tendency to pull back their listings arbitrarily and relist them after a while. Therefore we preferred to use the listings of the private sellers. After the completion of the first stage, the collected data were cleaned by removing some unreasonable cases such as the listings with very low mileage but very old cars that were potentially misleading. We also excluded the cars older than ten years. At the end of data preparation step, we had approximately 453K remaining cars.

B. Survival Analysis

Survival Analysis is used for analyzing the data which are obtained at the realization of a predetermined event (such as death, failure etc.) at any time. The main challenge encountered in the analysis of survival data is that by the time predetermined event has occurred we may no longer be able to observe the subject to collect the data. In other words, the subject may survive for a longer period that observations are no longer collected. These cases are called as right censored observations and mostly have longer survival times. Analysis of such data has been one of the main problems for the statisticians. Like the rest of the data, censored observations should be used correctly to achieve better results.

There are various approaches for solving problems related to the survival analysis. In one of these approaches, survival analysis is conducted by using a variety of parametric survival distributions. Another approach is based on the nonparametric distribution analysis which can be used without any prior statistical distribution assumptions. In this study, the outcomes of the analyses are presented by both parametric and non-parametric approaches. Because there are two major analysis methods, the analysis of censored survival data leads to the problem of choices. Of these, the advantages of the non-parametric methods are simple calculations and the clarity of the outcomes. Kaplan- Meier [10] is one of the most commonly used nonparametric analysis methods. On the other hand, parametric models are unbiased even if underlying distribution hypothesis is no longer valid as they are robust methods.

Parametric modeling will yield superior results when the preferred parametric distribution matches with the data. However, censored data particularly may result in poor outcomes when used in conjunction with the parametric methods. In short, best suitable survival analysis methods have been uti-

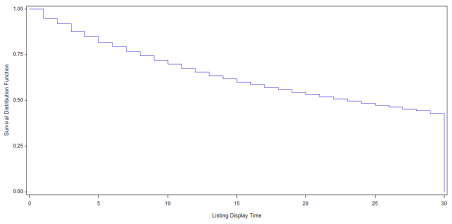

Fig. 1. Survival Plot of Days on Display

lized in this paper to overcome problems that stem from using real life data.

Cox regression model is very popular as a survival analysis tool due to its robustness [4]. The model examines the relationship between survival and a set of covariates i.e. variables where they may be dependent or independent of time. Following model is a sub-class of the proportional hazards family [2] in which different individuals have hazard functions that are proportional to each other. Therefore, the hazard function of t, given covariates, X can be written as follows [4].

$$h(t|X) = h_0(t) \times e^{\sum_i^d \beta_i X_i}$$

where $h_0(t)$ is called baseline hazard function. In Cox regression the hazard function estimates the relative risk of failure (death). The proportional hazard model implies that the effect of the covariates on the relative hazard is constant over time [11]. If there is a time-dependent covariate, proportionality assumption may not hold any more. The extended Cox regression model can be used in this case [4], [11] in which the time-dependent covariate is included as a predictor of $h_0(t)$.

III. Estimating Event Time by Survival Analysis

In this section, we report experimental results that encompass survival analysis (specifically Cox regression) model and linear regression models as benchmark results. Throughout our study we utilized some SAS procedures such as `PHREG`, `REG`, and `LIFETEST` as needed (see https://bit.ly/2MHoLSd). In studying the analytical models of used-car market, various car characteristics have been used in previous work [2], [9], [12]. In order to use these characteristics in our analyses, we created several dummy variables to indicate them. We also derived some variables e.g. age from the original data.

In the first part of our work, we generated Kaplan-Meier [10] nonparametric survival plot of days on display of the listings as seen in Figures 1. When we analyze any car related data, it is better to group the data by make and model. Each brand and model pair may have very distinctive behavior. Original dataset has 504 make and model pairs and some of the pairs have very few data in them. It is not practical to build some analysis models to explain underlying behavior of very few data points. Therefore we filtered the data to exclude those make-model pairs with less than 100 points. We ended up with 147 make-model pairs after filtering the data and there were around 448K cars left for the statistical modeling.

TABLE I
Summary Results of Survival Analysis

Variable	Count
Price	141
Kilometer	97
Age	89
Manual Shift	89
Petrol (Gas)	81
Annual Kilometer	76
LPG	49

Using best-in-class statistical software SAS, it is possible to conduct analyses by groups within the data. By utilizing this feature we easily conducted our analyses within make-model pairs. Cox regression model can be constructed by SAS `PHREG` procedure and we limit the number of variables to be five. So `PHREG` procedure chooses the best five variables based on Chi-Square. Table I summarizes `PHREG` results for all make-model pairs. Basically, we listed 7 of the most frequent covariates that were included the Cox regression models from 147 make-model pairs the frequency of which can be found in the second column. For instance, "Price" is the most frequently used covariate in Cox regression models; it was included 141 times out of 147 cases. The variable "Annual Kilometer" is a derived variable and is calculated by dividing "Kilometers" i.e. odometer reading by "Age". In [13], it was shown that this particular variable was also a significant predictor for the used-car prices. The other car characteristics in Table I are simply most significant variables for Cox regression models.

`PHREG` generates Chi-Square values for the fittings. The average Chi-Square value is 50.08. The minimum, median and maximum values are 0.80, 24.46, and 330.69 respectively out of 147 models.

IV. A Logistic Regression Based Predictive Model

In the previous section we introduced survival analysis models for predicting display time of web listings as a proxy of the car sale. In this section, we introduce a web site [3] which contains several modules to aid the pricing decisions in used car market of Turkey with various reports and analysis. One feature of the web site is the prediction of sale probability within 30-day period. In certain cases, one may desire to model a binary outcome such as our case i.e. sold cars and unsold cars. In this section, a predictive model based on logistic regression is introduced. Due to the nature of logistic regression, our model can yield a probability of event i.e. a sale of the car.

Logistic regression is used when the computation of a class probability is needed where this probability is a linear function of predictor variables [14]. A binary classification case can be specified by a `logit` (log-odds) transformation as follows:

$$\texttt{logit}\ p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

in which p is the probability of the event, k is the number of predictors (independent variables) and `logit` p is equal

[3] http://www.otokaca.com/

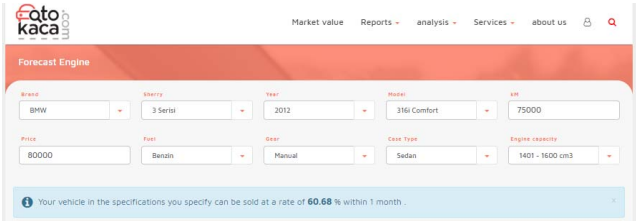

Fig. 2. Representative Result from a Logistic Regression Prediction

to $\log[p/(1-p)]$ (i.e. *log odds*). Notice that the probabilities of event (p) and non-event ($1-p$) sum to one. Parameters of logistic regression can be computed by maximum likelihood estimation. The choice of logistic regression in this study as a predictive model is not arbitrary. First of all, it uses a linear function in predictor variables. In addition, it can give us probability of the event which is an important assessment for a potential car buyer/seller. Once the parameters are estimated, then the probability of the event, p, can be computed as in following:

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}.$$

Figure 2 shows a prediction result of a particular car where the specifics of the car can be seen [4]. In this example, the probability of the sale is almost 0.61. Notice that the specifics of the car can be chosen in this module including asking price and odometer reading. Logistic regression models were run by using `LOGISTIC` procedure of SAS for each make and model pair separately. Estimated parameters were stored in a SQL table to enable predicting for any particular car later on. We believe that this module is also an important component of our decision support tool as seen from Figure 2. A predictive model always needs to be tested for unseen points, but our primary aim in this paper is to show the usability of collecting data from classified web listings.

V. Conclusion and Discussion

We successfully applied survival analysis to analyze the used-car listing data collected from an e-commerce site. It is also possible to use survival analysis to mine and predict the certain trigger events [15]. We only worked on building some statistical models to understand the underlying process behind used-car pricing based on data collected from the leading e-commerce site in Turkey. In the light of some recent research on online pricing [16], one possible future research topic could be to incorporate various price data from other sources to understand the price differences that may exist in online and offline markets better.

In addition to statistical pricing models, providing the users with decision support tools may improve better determining the used-cars' market values. We introduced some new web-based tools that may be very innovative and useful in terms of the value to the decision makers (buyers/sellers). Collecting listing data continuously may certainly help in building better predictive models that are proven to be good for the future unseen data. Our tools show the importance of a predictive model for pricing decisions. We can certainly deploy other predictive machine learning models in place of logistic regression models. However, caution should be taken to manage and maintain hundreds of such models.

[4]Note that the original web page in Turkish. The Google translation of the web page is shown in the figure without any image editing.

Acknowledgment

The web site `www.otokaca.com` is maintained by Verikar Software which is founded by the author.

References

[1] S. Singh, B. T. Ratchford, and A. Prasad, "Offline and online search in used durables markets," *Journal of Retailing*, vol. 90, no. 3, pp. 301 – 320, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022435914000232

[2] A. Jerenz, "Survival analysis: Estimation of the price-response function," in *Revenue Management and Survival Analysis in the Automobile Industry*. Gabler, 2008, pp. 63–96. [Online]. Available: http://dx.doi.org/10.1007/978-3-8349-9840-8_4

[3] F. Zettelmeyer, F. S. Morton, and J. Silva-Risso, "How the internet lowers prices: Evidence from matched survey and automobile transaction data," *Journal of Marketing Research*, vol. 43, no. 2, pp. 168–181, May 2006.

[4] D. G. Kleinbaum and M. Klein, *Survival Analysis A Self-Learning Text*, 2nd ed. New York: Springer-Verlag, 2005.

[5] L. Wang, G. Puskorius, B. Nance, and I. Salmeen, "Application of survival analysis for modeling the effects of vehicle features on days-on-lot," in *Proceedings of Joint Statistical Meeting*, NY, USA, 2002, pp. pp.3585–3590.

[6] F. S. Morton, F. Zettelmeyer, and J. Silva-Risso, "Internet car retailing," *The Journal of Industrial Economics*, vol. 49, no. 4, pp. pp. 501–519, 2001. [Online]. Available: http://www.jstor.org/stable/3569793

[7] J. Odink and E. van Imhoff, "Prices of used cars in west-germany before and after the second energy crisis," *De Economist*, vol. 130, no. 3, pp. 381–396, 1982. [Online]. Available: http://dx.doi.org/10.1007/BF02371748

[8] P. Kooreman and M. Haan, "Price anomalies in the used car market," *De Economist*, vol. 154, no. 1, pp. 41–62, 2006. [Online]. Available: http://dx.doi.org/10.1007/s10645-006-6485-z

[9] J.-D. Wu, C.-C. Hsu, and H.-C. Chen, "An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference," *Expert Systems with Applications*, vol. 36, no. 4, pp. 7809 – 7817, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417408008634

[10] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, pp. 457–481, 1958.

[11] R. Meijer and S. Bhulai, "Optimal pricing in retail: a cox regression approach," *International Journal of Retail & Distribution Management*, vol. 41, no. 4, pp. 311–320, 2013.

[12] E. A. Gilmore and L. B. Lave, "Comparing resale prices and total cost of ownership for gasoline, hybrid and diesel passenger cars and trucks," *Transport Policy*, vol. 27, no. 0, pp. 200 – 208, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0967070X13000036

[13] M. Engers, M. Hartmann, and S. Stern, "Annual miles drive used car prices," *Journal of Applied Econometrics*, vol. 24, no. 1, pp. pp. 1–33, 2009. [Online]. Available: http://www.jstor.org/stable/40206260

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2001.

[15] E. Malthouse, "Mining for trigger events with survival analysis," *Data Mining and Knowledge Discovery*, vol. 15, no. 3, pp. 383–402, 2007. [Online]. Available: http://dx.doi.org/10.1007/s10618-007-0074-x

[16] B. T. Ratchford, "Online pricing: Review and directions for research," *Journal of Interactive Marketing*, vol. 23, no. 1, pp. 82 – 90, 2009, anniversary Issue. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S109499680800008X