ANALYZING CLASSIFIED LISTINGS AT AN E-COMMERCE SITE BY USING SURVIVAL ANALYSIS

Ayhan Demiriz
Dept. of Industrial Engineering
Sakarya University
Sakarya, Turkey
ademiriz@gmail.com

Sümeyye Şen
Dept. of Computer Engineering
Sakarya University
Sakarya, Turkey
sumeyye@sakarya.edu.tr

Furkan Gökçegöz
Continuous Improvement
Timay & Tempo
Merzifon, Turkey
furkan.gokcegoz@timay-tempo.com

Abstract:

Sahibinden.com is a leading e-commerce site in Turkey where sellers (buyers) may advertise their goods (needs) with or without a fee. Since it generates a large volume of traffic to the classified car listings, the site plays an important role for determining the market value of the used cars. In this study, we first randomly selected 200 car classifieds from 950 new classified ads on the day of February 22, 2012. We then observed these listings on a daily basis for a month to determine the possible updates and deletions of the ads. We assume that if an ad is taken out it means that the car has been sold. In addition to the cars' features, we observed the posted price and the number of daily views of the ads throughout the data collection. Therefore one can construct survival models to study the effects of the features and price of a car on the life of the ad. In other words, it is possible to study that what features and price levels expedite the sales of used cars.

Keywords: *Survival Analysis, Used Car Sales, e-Commerce, Used Car Price Elasticity, Censored Data.*

1. INTRODUCTION

Automotive industry is one of the leading contributors to GDP in developed countries. Considering that automobiles form 70% of this industry (in terms of number of units), the importance of the automobile trade is obvious (Onat, 2007).

In addition, second hand (used) car (Wikipedia, 2013) sales have far exceeded the new car sales in many countries which shows the importance of the used car sales in world economy (Asilkan, 2009). For example, used car sales have a volume of over twice as much as the new car sales in the U.S.A (Lee, 2006).

Since the most people have access to the internet in developed countries, internet has become very important medium for the second-hand car sales market in developed countries and second-hand car dealers and buyers can reach the other party readily on this environment. As a developing country, the amount of purchases made over the internet is rapidly increasing in Turkey. For instance, there has been an increase of almost 20 percent in January-Februay period of 2013 compared to the same period of previous year which has brought the annual monetary volume of total internet sales in Turkey to 5.2 billion Turkish Lira (Dünya, 2013).

Sahibinden.com is one of the leading e-commerce sites in Turkey with a number of more than 2 million ads. According to the data from Sahibinden.com, the number of vehicles sold or rented within the first three months of 2013 over the same period of the previous year increased by 17 percent to be around 347,000 vehicles.

Considering that one vehicle is sold or rented every 23 seconds, the volume of the used vehicle listings (advertisements or ad) at Sahibinden.com is very significant. Indeed, the automobile listings in the category of Vehicles have a 59% share in terms of number of listings among 16 different types of vehicle categories and the number of listings in the category of vehicles within all categories has a 42% share of listings at Sahibinden.com e-commerce site.

A particular car listing may stay active for a number of days. Indeed, the time that the listing stays active can be considered as a random variable. There might be various reasons that an ad can be taken from the web site. However, it might be acceptable to assume that the particular car might have been sold by the time the ad is removed. The main purpose of this study is to determine the important automobile characteristics that affect directly the automobile sales. To achieve this, we collected data (Gökçegöz, 2012) by observing a random sample of automobile listings at Sahibinden.com for 30 days. We then analyzed the data to determine the impacts of the various car features on the sales. In Section 2, we give the background information on survival analysis. We then give details of the data used in this paper in Section 3. Statistical analysis of the data is given in Section 4. We then conclude the paper in Section 5.

2. 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 observe the object to collect the data. In other words, object 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 (Nelson, 1982). Analysis of such data has been one of the main problems of 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 nonparametric 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 method of analysis are simple calculations and understandability of the outcomes. In nonparametric analysis method, Kaplan-Meier (Kaplan & Meier, 1958) is one of the commonly used calculation 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 utilized in this paper to overcome problems that stem from real life data.

2.1. Survival Function

Survival time is the time interval for a person who is exposed to a specific disease until he heals or dies. Survival time of the individual or the system, indicated by *T*, is a random variable. The probability of an individual to live more than a certain time *t* is called the survival function. The survival function is given by following equation (Wang et al., 2002),

$$S(t) = P(T > t) = 1 - F(t) = \int_t^{\infty} f(u)du, 0 \leq t < \infty$$

where F(t) is the cumulative distribution function of survival time and f(u) is the probability distribution function. Equivalently, hazard function can be defined based on survival function as,

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\ln\big(S(t)\big),$$

which specifies the instantaneous failure rate at time *t* (Wang et al., 2002). Both survival and hazard functions are used extensively for analyzing the survival data in practice.

3. DATA COLLECTION

The data used in this paper were gathered from Sahibinden.com web site. Initially, 200 automobile listings were selected among 950 automobile listings posted on 22 February 2012. Selected advertisements were observed for 30 days from the date of February 22, 2012 until 22 March 2012. The collected raw data were cleaned and prepared for the analyses.

Automobile listings at Sahibinden.com are presented with car features that indicate used or new automobile, price, brand, model, type, mileage, color, engine capacity, engine power, fuel type, gear type, body type, transmission type, warranty status, trade-in options, and the responsible party (owner or dealer). Number of page view (i.e. seen by site visitors) is also shown on listing pages. All of these features were collected by a special software developed to fetch the web pages of 200 random listings used in this paper. The cleaned data were stored in an Excel file for further processing.

The data gathering software was run every day to collect data for each listing to detect the price changes, number of viewers and the status of the listing i.e. whether it was removed or not since the previous day. If a particular listing was no longer accessible, it was then assumed that the car was sold. Then the variable representing the death (failure) is assigned 1. The death time, *t*, was noted for that particular listing. The removal of the listing corresponds to the death or failure in our survival analysis approach.

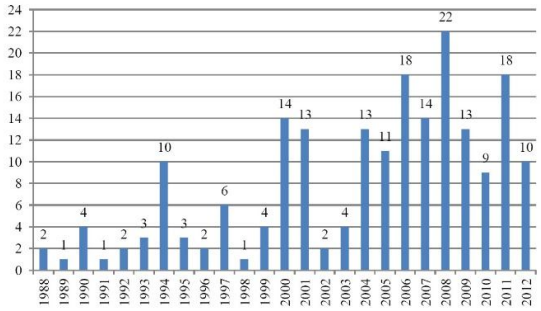


Figure 1 Distribution of Cars by Model Year

Once the 30-day long data collection task was completed, the dataset was preprocessed to convert prices given in foreign currencies to Turkish Lira (TL) based on the exchange rates on the day of original data collection. Some other minor discrepancies were also resolved during the data preprocessing step.

We give the summary charts in Figure 1 and Figure 2 to depict the content of the dataset. Figure 1 summarizes model year of the cars in the dataset. Basically most of the cars are used less than ten years. Figure 2 summarizes the composition of the dataset from the brand point of view. Again "1" represents the sold cars and "0" represents the unsold cars in Figure 2. It is easier to see the distribution of the cars by the brand for the dataset collected.

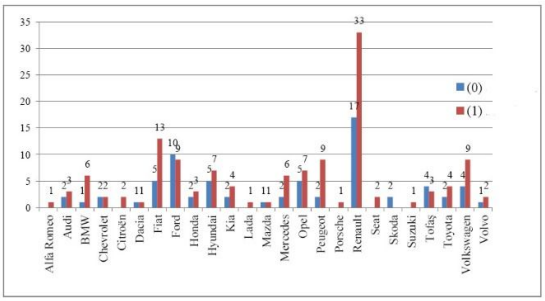


Figure 2 Summary of Data by Car Brand

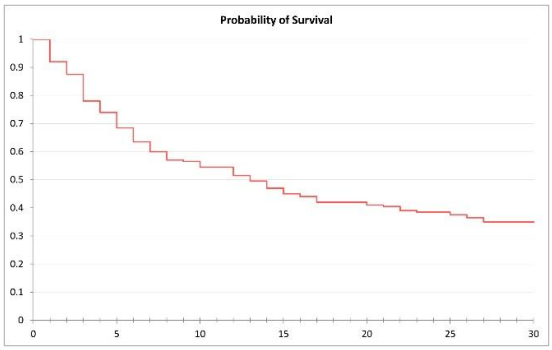


Figure 3 Nonparametric Survival Function

4. ANALYZING THE DATA

In this section we report our results on analyzing the data by survival and regression analyses. Figure 3 depicts the empirical and nonparametric survival function based on 200 observations.

As a result of the work carried out to determine the distribution analysis on the dataset with Minitab, it is found that the lognormal distribution is the most suitable distribution for the available data. Therefore, parametric analysis was performed by using lognormal distribution. Figure 4 depicts lognormal survival function. Figure 5 shows nonparametric survival function generated by Kaplan-Meier method.

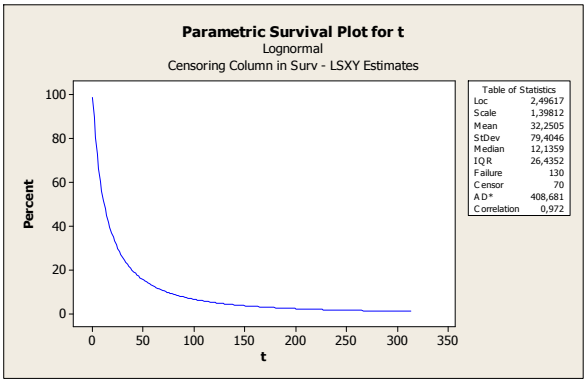


Figure 4 Parametric Lognormal Survival Function

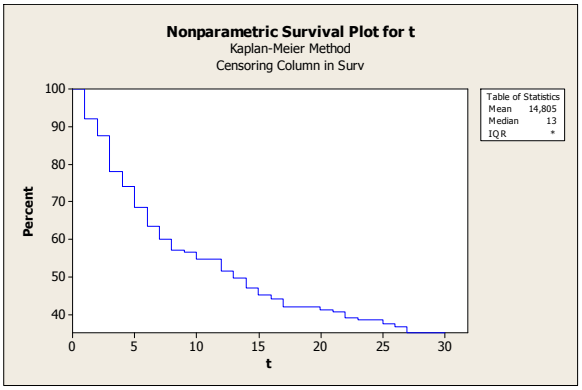
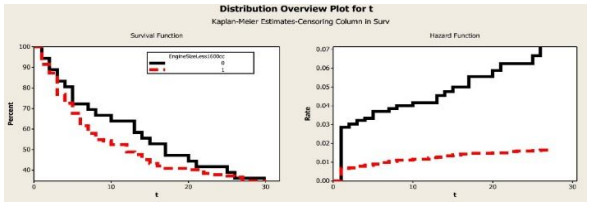


Figure 5 Nonparametric Survival Function Computed by Kaplan-Meier Method



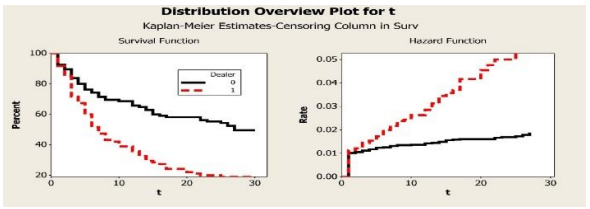


Figure 6 Nonparametric Survival Functions by Engine Size and Seller Type

Table 1 Summary of Regression Models

Model #	Model	R-Sq	F-Val	p Val
I	t = - 0.00590 ViewRatio + 0.0102 Year + 0.000008 KM - 3.3 Gasoline - 5.7 LPG - 2.1 Diesel - 1.16 ManShift - 0.12 AutoShift + 4.30 EngSizeLess1200cc + 4.79 EngSize1201-1400cc + 3.23 EngSize1401-1600cc + 7.48 EngSize1601-1800cc + 8.23 EngSize1801-2000cc + 17.3 EngSize2001-2500cc + 9.13 EngSize2501-3000cc - 1.70 Hatchback_5 + 0.61 Saloon + 3.50 Station Wagon - 9.32 Dealer	0.712	23.55	0.0001
II	t = 1.47 Gasoline - 0.06 LPG + 1.74 Diesel + 2.54 ManShift + 2.85 AutoShift + 18.0 EngSizeLess1200cc + 15.1 EngSize1201-1400cc + 12.5 EngSize1401-1600cc + 15.2 EngSize1601-1800cc + 16.8 EngSize1801-2000cc + 23.8 EngSize2001-2500cc + 15.5 EngSize2501-3000cc - 0.6 EngSize3001-3500cc - 2.39 Hatchback_5 - 1.21 Saloon + 1.59 Station Wagon	0.6537	21.7	0.0001
III	t = - 0.00544 ViewRatio + 0.048 Year + 0.000014 KM + 23.3 Gasoline + 20.6 LPG + 25.0 Diesel - 9.08 Dealer - 4.35 EngSizeLess1200cc - 3.97 EngSize1201-1400cc - 5.00 EngSize1401-1600cc	0.6946	43.2	0.00001
IV	t = - 0.00366 ViewRatio + 0.092 Year - 0.000009 KM - 163 Gasoline - 162 LPG - 161 Diesel + 0.07 ManShift - 0.64 AutoShift - 7.94 EngSizeLess1200cc - 12.5 EngSize1201-1400cc - 10.6 EngSize1401-1600cc - 5.85 EngSize1601-1800cc - 6.63 EngSize1801-2000cc + 5.3 EngSize2001-2500cc - 4.24 EngSize2501-3000cc - 14.2 EngSize3001-3500cc - 2.72 Dealer	0.6515	12.43	0.0001
V	t = 9.64 Gasoline + 9.74 LPG + 12.0 Diesel + 0.57 ManShift - 0.08 AutoShift - 0.95 EngSizeLess1200cc - 4.95 EngSize1201-1400- 2.98 EngSize1401-1600 + 1.36 EngSize1601-1800-0.91 EngSize1801-2000 + 12.4 EngSize2001-2500+ 3.05 EngSize2501-3000 - 3.28 Dealer + 2.14 Station Wagon + 2.09 Hatchback_5 + 1.67 Saloon	0.6313	12.2	0.0001
VI	t = - 0.00390 ViewRatio - 0.105 Year - 0.000008 KM + 15.4 Gasoline + 16.3 LPG + 17.5 Diesel - 2.49 EngSizeLess1200cc - 6.81 EngSize1201-1400cc - 4.91 EngSize1401-1600cc - 2.70 Dealer	0.6354	20.91	0.00001

We also analyzed the effects of the listing (car) features on the survival function. Figure 6 depicts the effect of engine size (whether less than 1600 cc or not) in the top plot and the type of the seller (whether dealer or not) in the bottom plot. The cars with less than 1600 cc are sold quicker than the larger engine sizes. Somehow the car listings posted by dealers were removed earlier than listings by owners which may indicate faster sales by the dealers.

The effects of the engine types on the survival functions are depicted in Figure 7. The top plot shows the effect of diesel engines that first 20 days non-diesel cars are sold quickly. In the last 10 days of the period, diesel cars are sold much faster. Notice that there are also cars with LPG engines. We see very little difference in survival functions of gasoline vs. non-gasoline engines at the bottom plot of Figure 7.

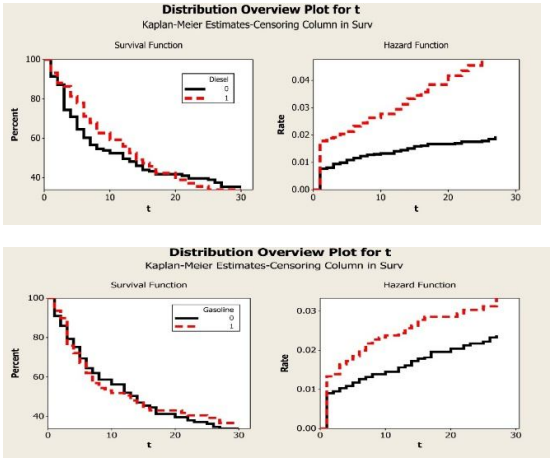


Figure 7 Nonparametric Survival Functions by Engine Types

4.1. Regression Models

Any analysis without proper models to determine the important factors that affect the time to death (in our case time to sell the car) will not be complete (Kleinbaum & Klein, 2005). For this reason we conducted multivariate regression analysis by creating related dummy (indicator) variables that indicate whether a particular car has certain features or not. Notice that car listings already have data of some car features as continuous variables such as mileage (km), price and the number of total listing views.

The results of multivariate regression models are reported in Table 1. Models I, II, and III comprise the regression models based on full data i.e. they include also the censored observation at 30 days. Models IV, V, and VI, on the other had were constructed by only using the data from sold cars within 30 days. Regression analysis of the all automobiles sold and unsold are compared with regression analysis of the only sold car. Notice that intercept was not fitted in any of the models reported in Table 1

The variable `ViewRatio` is the average number of views per day for a given listing. It is calculated by total number of views divided by *t* (i.e. time to sell the car or the censor time which is 30 at most). The variable `KM` represents the mileage in kilometers. The variable `Year` presents the age of the car in years. The remaining variables in Table 1 are dummy variables to indicate whether the cars have the corresponding features or not. For example, the variable `Diesel` represents whether the engine type is diesel or not.

R-Squared values are reasonable for all the regression models and *p-values* indicate that they are significant. The partial regression coefficients are in line with the expectations. For example, the variable `ViewRatio` has negative signs in the models which indicate the more a car listing is seen on average per day, the sooner it will be sold. The partial regression coefficients of `Diesel` are higher than the other engine types which indicate that it takes longer time to sell diesel cars on the average. As the sign of the partial regression coefficients of `Dealer` are negative, it takes less time to sell the cars listed by the dealers. The variable `KM` has both positive and negative signs in these six models presented in Table 1 which realistically determines that `KM` may not be significant at all. Again small size engines have negative signs which indicates that it is easier to sell cars with smaller engines.

5. DISCUSSION AND CONCLUSION

We presented the statistical analysis results of data collected from an e-commerce site about car listings. We successfully implemented methods from survival analysis to analyze such data. We then implemented regression models to analyze the factors that affect the time to sell the car (or remove the car listing). Survival functions and the regression models agree with each other's outcomes.

Since price data are also collected in our study, we can easily determine the price elasticity of the used cars as prices may vary from day to day for a given car listing. We can also construct classification models to predict that a certain vehicle will be sold within a specified time period or not. We plan to conduct such studies in our forthcoming paper.

REFERENCES

[1] Asilkan Ö., İkinci El Otomobillerin Gelecekteki Fiyatlarının Yapay Sinir Ağları ile Tahmin Edilmesi, Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi Y.2009, C.14, S.2 S.375-391.

[2] Dünya, http://www.dunya.com/e-ticaret-hacmi-yuzde-20-artti-187007h.htm, Accessed on June 6, 2013.

[3] Esen E., Sağ Kalım Analizinde Parametre Tahmin Problemlerine Katkılar, Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, Yüksek Lisans Tezi, Samsun, 2005

[4] Gökçegöz, F., "*Bir İnternet Sitesindeki İkinci El Otomobil İlanları İle Sağ Kalım Analizi Uygulaması*", Senior Thesis, Dept. of Industrial Engineering, Sakarya University. 2012

[5] Kaplan, E., and Meier, P., Nonparametric estimation from incomplete observations. J. Amer. Statist. Assoc. 53, 457-481, 1958.

[6] Kleinbaum, D.G. and Klein, M., Survival Analysis A Self Learning Text, 2nd Ed. Springer, 2005.

[7] Nelson W., Applied Life Data Analysis Canada: John Wiley & Sons, 1982.

[8] Wang, L., Puskorius, G., Nance, B., and Salmeen, I., Application of Survival Analysis for Modeling the Effects of Vehicle Features on Days-on-Lot. In Proceedings of Joint Statistical Meeting, pp. 3585-3590, NY, USA, 2002.

[9] Wikipedia, en.wikipedia.org/wiki/Used_car, Accessed on June 6, 2013