Learning from e-commerce: incomplete data and decision support

Abhinanda Sarkar
IBM India Research Lab
Block 1, IIT campus, Hauz Khas
New Delhi 110016, India
sabhinan@in.ibm.com

Rema Ananthanarayanan
IBM India Research Lab
Block 1, IIT campus, Hauz Khas
New Delhi 110016, India
arema@in.ibm.com

1. Introduction

Business or commercial activity on the Internet, commonly referred to as electronic commerce or e-commerce, has been steadily increasing for the last few years. Such activity has by nature generated large volumes of data, both structured and unstructured. This note suggests a way in which statisticians can make use of their training and skills in making use of this data to measure and enhance e-commerce by learning from data it generates.

We shall restrict attention to two topics which, in our opinion, deserve urgent attention, namely decision support and incomplete data. Internet sites are dynamic and the kind of information they require from users varies as the site evolves and often with user. Thus data is not available in the formats most conducive to longitudinal statistical analysis. Once an analysis is performed, an obvious use is to enable users to process the results and use them to reach informed decisions regarding future actions on the same or similar Internet sites.

Instead of a general exposition, we shall present a case study to illustrate the above scenarios. The analysis is simplified for brevity and ease of explanation. More details and extensions will appear elsewhere.

2. Auctioning cars on the Internet

Internet auctions were one of the earliest e-commerce activities. There are a number of popular sites such as Yahoo! and Amazon. We take a closer look at activities on eBay at www.ebay.com, the most active of them all. To restrict attention to a particular product, we chose automobiles as used cars seem to be popular auction items, Internet or otherwise. The

auctions were ascending auctions where bidders bid increasing amounts until the close of the auction where the highest bidder received the car. Lucking-Reilly (2000) gives more details and discusses alternative auction formats. Lucking-Reilly et al. (2000) presents the results of a study based on the auction of collectible coins on eBay and discusses the details of how eBay operates.

We followed the auctioning of 130 cars around June 2000, and then a further 120 cars around December 2000. We chose a popular brand, namely the Honda Accord. For each car, the following were recorded: (1) the age of the car, (2) the mileage of the car, (3) the start price of the auction, (4) whether the seller set a reserve price or not, (5) whether the car got sold or not, (6) the eventual selling price if the car got sold. A reserve price is a price that the seller sets with the understanding that the car will not be sold if this price is not met. This price is not publicly available to buyers, but knowledge of its existence is.

For cars put on auction during June 2000 and earlier the sellers provided information on age and mileage voluntarily. Subsequent to this, the format appears to have been modified and eBay now requires sellers to provide both data items. Hence the June 2000 records have incomplete information on these two variables. As these two variables are closely related, a simple imputation method was deemed adequate as as means to complete the data frame. For the complete data, age was linearly regressed on mileage and vice versa. These regression were used to "predict" missing values if one of the two variables was available. Little and Rubin (1987) provide arguments for the usefulness and validity of such a use of regression for imputation. For example, the model $Mileage = \alpha + \beta\, Age + \epsilon$ is estimated and missing mileage can be predicted to be $\hat{\alpha} + \hat{\beta}\, Age$ if Age is available. If neither age nor mileage is available, we drop the record from the analysis.

Sellers have to decide on two items, whether they want a reserve price and the price at which they want to start the auction. We propose a simplistic decision support system to aid them make this decision. Let Ps denote be the probability that a given car is sold and Es denote the expected selling price. Consider the following generalized linear models with canonical links as described in McCullagh and Nelder (1989).

$$\log\left(\frac{Ps}{1-Ps}\right) = \beta_{10} + \beta_{11}\, Age + \beta_{12}\, Mileage + \beta_{13}\, ReserveIndicator + \beta_{14}\, StartPrice$$

$$Es = \beta_{20} + \beta_{21}\, Age + \beta_{22}\, Mileage + \beta_{23}\, ReserveIndicator + \beta_{24}\, StartPrice$$

Here $ReserveIndicator$ is a qualitative variable (taken to be a dummy variable) indicating whether a reserve price has been opted for or not. While more complicated models were tried, simple linear regression gave satisfactory predictive capability.

After the incomplete data has been imputed, we found maximum likelihood estimators $\hat{\beta}_{ij}$ for the parameters. We assumed normality for prices thus leading to multiple linear regression for Es. We used multiple logistic regression for Ps.

The decision support tool that we propose for a car being considered for auction can be thought of in the following steps.

1. Input age and mileage.

2. Take a preliminary decision on reserve price.

3. Taking Age, $Mileage$, $ReserveIndicator$ as inputs, the system uses $\hat{\beta}_{ij}$ to give a predictions for $\hat{Ps}$ and $\hat{Es}$ for a sequence of values for $StartPrice$.

Here $\hat{Ps}$ and $\hat{Es}$ are the estimated probability of sale and estimated final selling price. Based on these estimates, the seller can now decide on a start price that suits his personal objectives. A very high a start price can suggest a high final selling price, but the car is likely to not find a buyer at all. A suggestion may be to choose a start price that maximizes expected revenue as measured by $\hat{Es}\hat{Ps}$. The exercise can be repeated with the alternative choice on the reserve price in step 2 and this incorporates inclusion of that variable in the decision process.

It should be noted that Internet auctions elicit economic behavior different from conventional auctions. For example, high start prices generate very few bids and the final selling prices are often lower than "street" expectations. Thus new software systems, of the above type, have to be designed specifically for Internet auctions.

3. Results

We report on the results obtained for the auctions on eBay. For the 250 cars considered, we had 127 cases of incomplete records which were 77 were imputed and 50 dropped by the linear regression scheme of Section 2. Among other things, we observed that for estimating the probability of a sale, the presence of a reserve price was a significant factor. This was less true for the estimation of expected selling price where age and mileage had more of an effect.

We also present an example of the decision support tools that we suggested in Section 2. For a six year old Honda Accord and with 75,000 miles, an auction with/without a reserve price, Table 1 relates starting price, probability of sale, and expected selling price as per the models discussed.

There seems to be an overall suggestion that low start prices are better, at least for relatively old cars such as this particular one. Setting a reserve price also looks to be counterproductive.

Our results are consistent with the broad findings in Lucking-Reilly et al. (2000) who carried out a more sophisticated analysis and included "book value" information on the auctioned items.

Table 1

Start Price	Exp. Sale Price without reserve	Prob. of Sale without reserve	Exp. Sale Price with reserve	Prob. of Sale with reserve
1,000	7,750	0.85	8,327	0.34
2,000	7,855	0.81	8,433	0.28
3,000	7,961	0.76	8,538	0.22
4,000	8,066	0.69	8,644	0.17
5,000	8,172	0.62	8,750	0.13
6,000	8,278	0.54	8,855	0.10
7,000	8,383	0.46	8,961	0.07
8,000	8,489	0.38	9,066	0.05

4. Conclusion

While decision support and recommendation systems systems such as as the one suggested above have been in existence in the brick-and-mortar world, their extension to the virtual world poses new challenges, statistical and otherwise, as discussed in Schafer, Konstan, and Riedl (1999). In some cases, the data may be voluminous and unstructured and in other cases, the system must be capable of reconfiguring itself as the online environment evolves. While our example is a toy example, we hope that it can serve to motivate the possibilities in statistical learning from e-commerce.

The use of statistical methods to extract information from e-commerce activities is still in its infancy. Data mining, as developed in the computer science community, has been extensively used to enhance customer resource management and online marketing. See Rud (2000) for recent methods. We hope that statisticians will be more active participants in this digital world of the future.

REFERENCES

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* Wiley.

Lucking-Reilly, D. (2000). Auctions on the Internet: What's being Auctioned, and How. *Journal of Industrial Economics.* vol 48, no 3, 227-252.

Lucking-Reilly, D. et al. (2000). Pennies from eBay: the Determinants of Price in Online Auctions. Working paper, Department of Economics, Vanderbilt University.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* Second Edition. Chapman & Hall.

Rud, O.P. (2000). *Data Mining Cookbook: Modeling Data for Marketing, Risk and Customer Relationship Management.* Wiley.

Schafer, J.B., Konstan, J.A., and Riedl, J. (1999). Recommender Systems in E-commerce. *Proceedings of ACM E-Commerce.*