Evgeniy M. Ozhegov[1], *Daria Teterina*[2]

ENSEMBLE METHOD FOR CENSORED DEMAND PREDICTION

Extended Abstract

The grocery retail market has been under the close scrutiny of economists over the past few decades. Prediction of demand and, in particular, sales volume forecasting is widely used for the purposes of customers flow prediction, setting the optimal prices within and between product categories and effective stocks management (Levy & Weitz, 2011). In turn, solving each of the above tasks contributes to improving the financial performance of the company.

For quite a long time, demand prediction in retail was carried out exclusively with the use of econometric methods that seemed to be quite effective for working with small datasets and well interpretable in terms of estimated parameters, including price sensitivity of demand. But with the increased availability of scanner data that contains individual data on purchases, machine learning (ML) methods turns out to outperform econometric models in a demand prediction problem. Methods of machine learning allowed to obtain more precise out-of-sample predictions on large datasets and take into account unobserved consumers' heterogeneity and other non-regularities in sales data (Agrawal & Schorling, 1996; Varian, 2014; Bajari, Nekipelov, Ryan & Yang, 2015a, 2015b). Furthermore, ML methods demonstrate a higher convergence rates compared to non-parametric econometric models which led to the prevalence of their use in cases with a large number of possible predictors.

Despite all the advantages, machine learning methods are efficacious with traditional regression and classification problems only. There is a wide range of econometric models that has been developed for the problem of model estimation on censored data also. Censored demand is a corner solution in demand system observed when the number of product purchases desired by consumers on a certain price is negative, leading to zero purchases. Large fraction of zeros in sales is called the problem of censored demand. Censored data often occur in individual consumption demand models, where the individuals either consume zero (if consumers have not bought anything from the goods available to them), or some positive discrete or continuous amount of good (Ozhegov & Ozhegova, 2018). In the case of data censorship neglect estimation of price parameter are likely to be downward biased because estimation procedures treat all zero

[1] National Research University Higher School of Economics (Perm, Russia). Research fellow, Research Group for Applied Markets and Enterprises Studies. E-mail: tos600@gmail.com

[2] National Research University Higher School of Economics (Perm, Russia).Young research fellow, Research Group for Applied Markets and Enterprises Studies. E-mail: dvteterina@gmail.com

sales as constant even if a price increase substantially. For a retailer, underestimation of the effects of price as well as a bias in promotion or product's characteristic parameters due to the same reasons leads to real financial losses (Levy & Weitz, 2011).

Recent econometric developments for censored data estimation (Chernozhukov, Hong, 2002; Chernozhukov, Fernandez-Val, Kowalski, 2015) use two step approach, splitting an estimation for the steps of discrete part (zero or non-zero sales) estimation and continuous part (strictly positive sales on non-zero sales data) estimation. While machine learning methods manage better with both parts of a problem, including classification to zero and non-zero sales, and prediction of continuous sales data, we construct an algorithm that is based on the econometric idea of dealing with data censorship by problem splitting and apply various machine learning methods for classification and regression problem. The developed estimator is based on the idea of combining several simple predictors (Linear regression, Ridge regression, Lasso regression and Random Forest) into constrained linear ensemble models similar to (Bajari et al., 2015b).

We test the potential capacity of proposed algorithm on a real retail food chain data. The data is provided by the Russian regional grocery retail chain and cover consumer purchases for six years: from January 2009 to December 2014. The analyzed sample size is 800000 daily sales. A unit of observation is a combination of stock keeping unit (SKU), certain store where it was in sale and a certain day. More than 60% of daily observations on SKU sales are equal to zero, one needs to account for demand censorship.

Each model with censorship results to better predictive properties than the same models without censorship accounting. Models combination *via* weighted linear regression, in turn, allows to improve the prediction accuracy in terms of out-of-sample RMSE. Thus, the prediction error for an ensemble model with censoring turned out to be equal to 0.684, while it is 0.781 for the ensemble without censorship, with a statistically significant difference between them. We also test the difference in mean marginal effect of price for the separate ML models and its ensemble with and without accounting for data censorship and show the statistically significant downward bias in models without censorship accounting.

References

Agrawal D., & Schorling C. (1996). Market share forecasting: An empirical comparison of artificial neural networks and multinomial logit model. *Journal of Retailing*, 72(4), 383–407.

Bajari, B. P., Nekipelov D., Ryan S. P., & Yang M. (2015a). Machine Learning Methods for Demand Estimation. *The American Economic Review*, 105(5), 481-485.

Bajari, B. P., Nekipelov D., Ryan S. P., & Yang M. (2015b). Demand estimation with machine learning and model combination. *National Bureau of Economic Research*. (No. w20955).

Chernozhukov V., & Hong H. (2002). Three-step censored quantille regression and extra-marital affairs. *Journal of the American Statistical Association*, 97(459), 872-882.

Chernozhukov V., Fernandez–Val I., Kowalski A. E. (2015). Quantile regression with censoring and endogeneity. *Journal of Econometrics*, 186(1), 201–221.

Levy M.& Weitz. (2011). Retailing Management, 9th Edition. McGraw-Hill Companies

Ozhegov E. M., & Ozhegova A. (2018). Bagging Prediction for Censored Data: Application for Theatre Demand. *International Conference on Analysis of Images, Social Networks and Texts , Springer, Cham.*, 197-209.

Varian H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives,* 28(2), 3–27.