NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Evgeniy M. Ozhegov, Daria Teterina

THE ENSEMBLE METHOD FOR CENSORED DEMAND PREDICTION

BASIC RESEARCH PROGRAM
WORKING PAPERS

SERIES: ECONOMICS
WP BRP 200/EC/2018

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

THE ENSEMBLE METHOD FOR CENSORED DEMAND PREDICTION[3]

Many economic applications, including optimal pricing and inventory management, require predictions of demand based on sales data and the estimation of the reaction of sales to price change. There is a wide range of econometric approaches used to correct biases in the estimates of demand parameters on censored sales data. These approaches can also be applied to various classes of machine learning (ML) models to reduce the prediction error of sales volumes. In this study we construct two ensemble models for demand prediction with and without accounting for demand censorship. Accounting for sales censorship is based on a censored quantile regression where the model estimation was split into two separate parts: a) a prediction of zero sales by the classification model; and b) a prediction of non-zero sales by the regression model. Models with and without censorship are based on the prediction aggregations of least squares, Ridge and Lasso regressions and the Random Forest model. Having estimated the predictive properties of both models, we empirically test the best predictive power of the model taking into account the censored nature of demand. We also show that ML with censorship provides bias corrected estimates of demand sensitivity to price change similar to econometric models.

JEL Classification: D12, C24, C53

Keywords: demand, censorship, machine learning, prediction.

[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

[3] The publication was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2018-2019 (grant 18-01-0025) and by the Russian Academic Excellence Project "5-100".

1. Introduction

The grocery retail market has been closely scrutinized by economists over the past few decades. The prediction of demand and, in particular, sales volumes is widely used for the purposes of customer flow forecasting, setting optimal prices within and between product categories and effective stock management (Levy & Weitz, 2011). Solving each of these tasks contributes to improving retail financial performance.

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 were well interpretable in terms of the estimated parameters, including the price sensitivity of demand. With the increased availability of scanner data containing individual data on purchases, machine learning (ML) methods have begun to outperform econometric models in demand prediction. ML gives more precise out-of-sample predictions on large datasets and takes into account unobserved consumer heterogeneity and other non-regularities in sales data (Agrawal & Schorling, 1996; Varian, 2014; Bajari, Nekipelov, Ryan & Yang, 2015a, 2015b). Furthermore, ML methods demonstrate higher convergence rates compared to non-parametric econometric models which led to the prevalence of their use with a large number of possible predictors.

Despite all the advantages, ML methods are efficacious with traditional regression and classification problems only. There is a wide range of econometric models that were developed for the problem of model estimation on censored data. Censored demand is a corner solution in the demand system observed when the number of product purchases desired at a certain price is negative, leading to zero purchases. A 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 any of the goods available to them), or some positive discrete or continuous amount of good (Ozhegov & Ozhegova, 2018). For data censorship a neglected estimation of the price parameter is likely to be downwardly biased because estimation procedures treat all zero sales as constant even if the price increases substantially. For a retailer, an underestimation of the effects of price as well as a bias in the promotion or product characteristic parameters, for the same reasons, leads to financial losses (Levy & Weitz, 2011).

Recent econometric developments for censored data estimation (Chernozhukov, Hong, 2002; Chernozhukov, Fernandez-Val, Kowalski, 2015) use a two-step approach, splitting the estimation into the discrete part (zero or non-zero sales) and the continuous part (strictly positive sales on non-zero sales data). While ML methods better manage both parts of the problem, including the classification of zero and non-zero sales, and the 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 ML methods for the 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 the proposed algorithm on real retail food chain data. The data are provided by a Russian regional grocery retail chain and cover consumer purchases for six years: from January 2009 to December 2014. The analyzed sample size is 800,000 daily sales (purchases). A unit of observation is a combination of the stock keeping unit (SKU), the certain store where it was sold and a certain day. As 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 makes better predictions than the same models without censorship. Model combinations via a weighted linear regression improve the prediction accuracy in terms of out-of-sample RMSE. The prediction error for an ensemble model with censoring is 0.684, while it is 0.781 for the ensemble without censorship, a difference which is statistically significant. We also test the difference in the mean marginal effect of price for separate and combined ML models with and without data censorship and show a statistically significant downward bias in models without censorship.

The remainder of the paper is organized as follows. The second section reviews the literature. The third section describes the data and their preliminary analysis, which explains the motivation for the proposed methodology. The fourth section introduces the demand model and the methodology of its calibration. The fifth section discusses the estimation results. The final section concludes.

2. Literature review

This research draws upon, and contributes to, the literature on demand prediction in retail, ML methods for demand prediction and the demand censorship problem. A surge of interest in demand prediction in retail occurred in the late 1990s when the Nielson and IRI Marketing Research companies began to collect individual data on retail chain purchases (Richards & Bonnet, 2016). Such data are known as scanner data, since they are collected by check-out scanning machines. Scanner datasets usually contain information on the SKU bought by consumers on each shopping trip, and the information such as SKU price, discounts, purchase time (Keane, 2013). The use of scanner data in consumer studies makes it possible to observe and analyze individual choice. The consideration of individual demand allows the construction of richer and more realistic models (Einav, Kuchler, Levin & Sundaresan, 2015).

Demand prediction models are usually used by retailers for solving various problems including optimal price setting (e.g. Bolton & Shankar, 2003; Caro & Gallien, 2010; Shakya, Kern, Owusu & Chin, 2012; Ferreira, Lee & Simchi-Levi, 2015; Qu, Zhang, Chan, Srivastava, Tiwari & Woo-Yong Park, 2017), sales volume forecasting (e.g. Fader & Hardie, 2001; Divakar, Ratchford, Venkatesh & Shankar, 2005; Ali, Sayin, Woensel & Fransoo, 2009; Bajari et. al, 2015b; Pezente, 2018), effective stock management (Agrawal & Smith, 1996; Aburto & Weber, 2007). Solving any of the above tasks is extremely important for retailers because they carry significant financial benefits. The managing director at Conway McKenzie says that when working with one large retailer a 10% increase in forecast accuracy could increase profitability by more than $10 million. That is why retailers are so desperately fighting for any improvement in forecast accuracy. In this research we show an increase in predictive accuracy in daily sales volume models within a product category. The solution to this problem will help to more precisely plan the stock for each individual store of the grocery chain and price the category optimally.

For a relatively long time econometric approaches dominated the field as there was only aggregated sales data. Aggregated datasets consist of market shares aggregated by brand, sales volumes, average prices, etc. An important approach to estimating demand function on aggregated data on the sales of differentiated products was proposed by Berry, Levinsohn and Pakes (1995). They use information on the annual sales volumes of each car model in the US market, the average sale price and the characteristics of the cars, to estimate the parameters of the individual utility function of the average household, as well as the contribution of each vehicle characteristic to the marginal cost function. The further development of multiple choice models on aggregated data is reflected in the introduction of heterogeneity in consumer tastes by observable and unobservable characteristics (Nevo, 2001). Nevo examines the U.S. ready-to-eat cereal market and constructs a more complex utility function. This utility function considers the observable and unobservable characteristics of goods and the heterogeneity of consumers in terms of their tastes, which depended on the observed and unobservable characteristics of consumers. The specification of the utility function is also complemented by the zero alternative, that is, the inclusion of the consumer's ability to buy nothing and utility gained from that.

Despite the great success of Nevo (2001), Berry et al. (1995) and other fundamental studies on econometric approaches to demand estimation, the approaches seem to be quite inflexible, requiring many assumptions on the error or dependent variable distribution while the predictive properties of models were often far from ideal. With advances in the availability of detailed data on purchases, the number of studies using machine learning methods for demand prediction has grown. ML methods have better predictive properties than traditional econometric approaches, which has

been repeatedly proven in a number of papers. For example, Agrawal and Schorling (1996) compare neural networks and multinomial logit models in brand share predictions and find that neural networks predict better; Varian (2014) shows that regression trees work comparatively better than logistic regressions for larger datasets and also demonstrates some advantages of such methods as bagging, bootstrapping and boosting over traditional econometric approaches; Bajari et. al. (2015b) compare the predictive power of a number of traditional econometric models and ML methods in a within-product category prediction problem, and conclude that the latter perform better.

ML methods are widely used for solving the demand prediction problem. The main advantage of ML methods with respect to traditional econometrical ones is their better ability to fit out-of-sample (Richards & Bonnet, 2016). Usually the model with the lowest root mean squared error (RMSE) or another similar prediction accuracy indicator (MSE, MAPE, WMAE etc.) on a cross-validation sample of the data is considered best. Although RMSE minimizing has been solved in computer science for a long time, the application of the models to economic problems with the subsequent possibility of an adequate result interpretation has become widespread only recently. To date, there are a few studies that partially fill the gap between traditional econometric approaches and ML methods in the context of demand prediction (Einav, Liran, Jenkins & Levin, 2012; Varian, 2014; Bajari et al., 2015a, 2015b; Witten et al., 2016; Ruiz, Athey, Blai, 2017). In one of the most recent developments (Bajari et al., 2015b), the authors consider several ML techniques and compare them with traditional econometric models, empirically proving the better predictive power of the former. Further, in order to improve out-of sample prediction accuracy they develop a method of underlying model combination via a constrained linear regression. In our study, we generalize the algorithm described in Bajari et al. (2015b) for censored, dependent data. Our motivation in combining the algorithms of ML and censorship estimation is encouraged by the possibility of increasing the predictive accuracy of well known models.

For many products, particularly for food and beverages, the process of choosing goods by consumers is more correctly described as a discrete-continuous process, rather than simply discrete. Consumers either do not buy anything (zero consumption), or buy some positive quantity of goods, where the positive part of consumption can be both discrete and continuous depending on the type of product. The data, for which the problem of discrete-continuous choice is econometrically solved, often look like a large number of zeroes for non-purchased alternatives and continuous amount for purchased ones (Richards & Bonnet, 2016). Such data is called censored. Not accounting for the censored nature of data results in a biased prediction of consumption. The bias occurs because even when a model is calibrated on uncensored observations only, the transition to

the group of consumers with zero consumption is not taken into account (Ozhegov & Ozhegova, 2018). In this research we subject the models with censorship to scrutiny because the daily SKU sales data are censored on the left (more than 60% of sales observations are equal to zero). Simply dropping zero observations from the sample leads to the endogenous sample selection problem and inconsistent estimates (Heien & Wesseils, 1990). Therefore, to obtain accurate predictions, it is necessary to take into account the data censorship.

The topic of demand censorship is quite well developed in the econometric literature. There are a number of parametric models based on the basic concepts of Tobin (1958) and Heckman (1979). There is also a number of studies where non-parametric (Bester & Hansen, 2009; Hoderlein & White, 2012; Matzkin, 2012) and semi-parametric (Khan & Powell, 2001; Chernozhukov, Fernandez-Val & Kowalski, 2015) models are used to account for censored data with better distributional assumptions. To the best of our knowledge this is the first work that incorporates data censorship into ML algorithms for demand prediction.

3. Data

The study is conducted on data provided by a Russian regional grocery retail chain for a pasta product category. This category has been used for several reasons. First, pasta is included in the list of socially important food products. Secondly, it may be stored for a long time and is characterized by a high number of SKUs and large price variation across SKUs. Therefore, we can take into account a large number of characteristics in our analysis. Thirdly, pasta is a daily demand food product, so its purchase is relatively frequent. Finally, as the pasta category has low substitutability with other product categories, the demand for it is not significantly affected by demand or price variation in other categories.

The initial data from the grocery chain sales systems represent the full information on pasta purchased from 2009 to 2014. The analyzed sample is a random draw of 800,000 observations from the initial sample. One observation reflects a SKU which is displayed in a certain store on a specific date. If any SKU was displayed but not purchased on a certain day, this is shown in the data as a sales volume of zero. All combinations of SKU, store and day where some SKU was not displayed are excluded from the data using additional information on available stock. Preliminary data analysis show that approximately 60% of sales observations among displayed SKUs are zero (See Fig.1). This motivates the necessity to account for censorship.

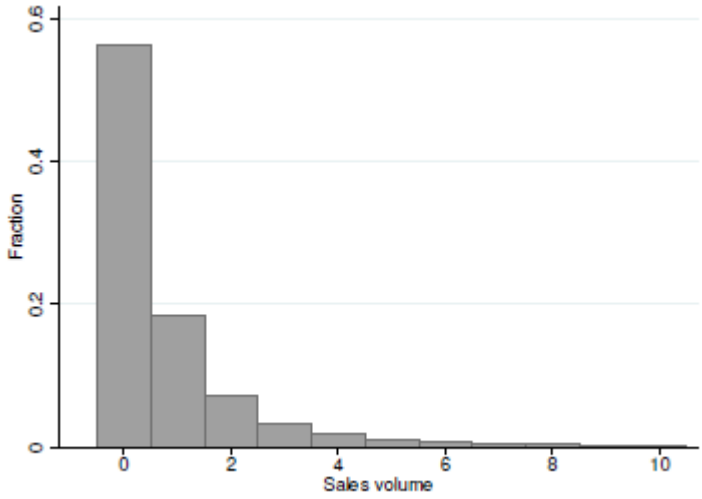

Fig. 1. Fraction histogram of pasta sales

In order to obtain better predictions from the demand model, we use the product catalog to recover product characteristics for each SKU. For each purchase we collect the price, color and shape of the pasta, the flour type, the packaging size and the type of packaging, the country of origin and the brand name. In addition to all of the above, for each observation we trace the format of the store where the purchase was made and any promotion indicator. In total, the data included 38 brands, 6 countries of origin, 13 package weight categories, 5 colors, 22 forms, 8 types of flour, and 5 types of stores where the pasta was bought. Descriptive statistics for each attribute are presented in Appendix 1, while the description of the variables types is presented in Table 1.

Since the dataset we are working with is cross-sectional, and the goal of our study is to predict daily sales, it is necessary to include in the model various time variables, such as year, month, day of the week and an indicator for a holiday, in order to catch intrayear and intraweek seasonality. Descriptive statistics for time features are presented in Appendix 2.

Tab. 1. Types of variables

Variable	Type	Min	Max
Sales volume (number of packs)	Numeric	0	10
Price per package (rub)	Numeric	9	120
Weight (g)	Numeric	150	1000
Promotion	Dummy	0	1
Brand	Categorical (38 categories)		
Country of origin	Categorical (6 categories)		
Color of pasta	Categorical (5 categories)		
Form of pasta	Categorical (22 categories)		
Type of flour	Categorical (8 categories)		

Type of package	Categorical (2 categories)		
Type of stores	Categorical (5 categories)		
Year of purchase	Categorical (5 categories)		
Month of purchase	Categorical (12 categories)		
Day of the week	Categorical (7 categories)		
Holiday	Dummy	0	1

Before the model construction a set of dummies is constructed from all categorical variables, all variables are standardized since some methods of ML work correctly only if this condition is satisfied.

4. Methodology

The general regression task is to predict sales volumes of a product. In linear regression form the model is as follows:

$$y_{jmt} = X_{jmt}\beta + \varepsilon_{jmt} \quad (1)$$

where y_{jmt} is the sales volume of the *j*-th product in store *m* on day *t*; X_{jmt} is the matrix of attributes including the log of the price, product characteristics, promotional indicator, time attributes (dummies for month, year, intra-week seasonality and holidays); ε_{jmt} is an idiosyncratic shock to each product, market and time.

According to the literature (Varian, 2014; Bajari et al., 2015a; Witten et al., 2016), ML methods are better able to cope with demand prediction due to a better out-of-sample fit without the loss of in-sample fit quality. Therefore, to achieve the most accurate predictions, three methods of ML are used in the research. In this study, we generalize the algorithm described in Bajari et al. (2015b) by adding stages estimating censored models similar to Chernozhukov & Hong (2002). Thus, the empirical part of the study can be divided into three stages:

1. the construction of four models (Linear regression, Lasso, Ridge and Random Forest) with censoring;
2. the construction of four models without censoring;
3. the estimation of two ensemble models with and without censoring and a comparison of their predictive power.

Before considering each of the stages in more detail, it is necessary to clarify several features of the original sample division. In our study, following Bajari et al. (2015b), we randomly divide the initial sample into three sets: test (25% of the data), validation (15% of the data) and

training (60% of the data). This is done for the subsequent double cross-validation: on the training sample we train the initial four models; on the validation set we make out-of-sample predictions to choose optimal threshold to classify observations into censored and uncensored and get the weights of the models to their inclusion in the linear combination; on the test sample we obtain an out-of-sample prediction for ensemble models where all parameters are calibrated on the training and validation samples.

The main steps to construct models with censorship are:

1. to construct the indicator variable $d_{jmt} = I\{y_{jmt} = 0\}$ for sales censorship;
2. to train a classification model for censorship dummy *d* using explanatory variables *X*;
3. to classify observations in a training set by probability threshold α into censored when $\mathrm{E}[d \mid X] > \alpha$ and uncensored otherwise;
4. to train a model for the continuous (uncensored) part of training set split by threshold α;
5. to obtain predictions for a validation set and combine predictions from the models of steps (2) and (4). If the predicted dummy for censorship by the classification model is 1 or the prediction on the continuous part of demand in model (5) is below 0 then the predicted demand is 0, otherwise the prediction is equal to the prediction from model (5). Calculate RMSE on the validation set for a given threshold α. Choose the optimal threshold α to split based on the validation set RMSE;
6. to obtain predictions for the validation set using a combination of models from steps (2) and (4) with the optimal threshold α from various classes of prediction models (Linear regression, Lasso, Ridge, Random Forest).

To train the models without censorship we treat all observations as uncensored, skip estimation steps (1-3) and set optimal α as 0.

After training simple models (Linear Regression, Lasso, Ridge and Random Forest) on the training sample, comparing out-of-sample errors and determining the predictive power of each model, we proceed with the construction of the ensemble models. For this, we determine the optimal linear combination of models using a linear regression. The main steps at this stage are:

1. take the validation set. Treat the predicted values of the dependent variable from the four models as regressors and the actual value as the response variable. Assuming that the sum of the coefficients should be equal to one and each individual coefficient must be non-negative, build a constrained linear regression. The coefficients obtained represent the weights with which each of the models should be included in the ensemble;

2. use the fitted models for prediction in the test set and apply the model weights from the previous step, sum them and construct the linearly combined prediction;
3. calculate RMSE on a test set for the final ensemble models.

The empirical part of the study is conducted on *RStudio*, an open resource for data analysis, with the use of programming language *R*. Lasso and Ridge regressions are implemented in the R package *glmnet*, and Random Forest in the package *randomForest*. All hyper parameters for Ridge and Lasso regressions are configured using internal cross-validation. For Random Forest, first, we run the *rfcv*[4] function which implies a *k*-fold cross-validation in order to reveal the optimal number of variables to sample at each tree *mtry*. After that we build the Random Forest model using the optimal value of the *mtry* (in our case *mtry*=35) from the function *rfcv*, and set all other parameters by default. The default value for *nodesize* is 5, *ntree* 50 and *maxnodes* NULL.

5. Results

Since more than 60% of sales are zero, we should check the parameter estimates for the need to use the censored regression model, testing for any bias in the simple linear regression framework (1) versus the censored regression model. The parameter estimates for these two specifications are presented in Table 2.

Tab. 2. Results for linear regression with and without censorship accounting.

Variable	Linear regression	Censored linear regression
Log. of price	-0.742*** (0.006)	-1.442*** (0.018)
N	800000	800000
k	95	95
Test sample RMSE	0.854	0.779

Notes: Parameters estimates are presented in table cells, standard errors in parenthesis. Significance level is p***<0.01, *N* is the number of observations, *k* is the number of parameters. Brands, forms of pasta, package type, colour of pasta, type of flour, time attributes (year, month, day of the week, holiday), promotion indicator, store type are included in the both models as control variables. Some categories of categorical variables are dropped out because of multicollinearity, for ex., a unique combination of country of origin and brand.

The reported results mean that the effect of price in the model with censorship is greater in absolute value. This supports the theoretical result that models without censorship lead to an underestimation

[4] In *rfcv* function we, following the Bajari et al. (2015b), set *scale*=log, *step*=0.5, which implies the removal of 50% of the variables at each step.

of the parameter estimates. Moreover, the censored linear model has better predictive properties in terms of out-of-sample RMSE.

After evaluating the parameters of the basic linear model, the sales volume variable is fitted in the training set by the four models (Linear regression, Ridge regression, Lasso regression and Random Forest). Then, for every model the measure of the predictive quality (out-of-sample RMSE) is calculated (Table 3). According to the RMSE calculation results, the Random Forest model provides the best predictive power with and without censorship because of its more flexible model structure compared with linear models.

Finally, the models included in the ensemble with linear weights are estimated by using a constrained linear model. The results of the constrained linear regression estimations for both ensembles, with and without censorship, are presented in Table 3 as models weights.

According to the estimation results, both ensemble models with and without censorship have better performance than any of the evaluated models individually. Moreover, the ensemble model with data censorship has better predictive power, which is indicated by the comparatively smaller RMSE. This result confirms our initial hypothesis that the use of ML techniques in conjunction with censorship gives the model better predictive power.

Tab. 3. RMSE for models with and without censorship accounting.

Model	RMSE		Weight in ensemble	
	Without censorship accounting	With censorship accounting	Without censorship accounting	With censorship accounting
Linear regression	0.854	0.779	1%	13%
Ridge regression	0.854	0.781	10%	8%
Lasso regression	0.845	0.765	33%	12%
Random forest	0.796	0.736	56%	67%
Ensemble model	0.781	0.684		
t-stat =3.22	*p*-value=0.01			

Notes: *t*-statistics and its *p*-value correspond to the significance of difference between RMSE in ensemble with and without censorship accounting. Standard error is calculated from panel bootstrap distribution of RMSE difference on 1000 replications with random draws over SKUs.

In order to show the statistically significant downward bias in models without censorship we test the difference in the mean marginal effect of price for the separate ML models and the ensemble with and without data censorship. We calculate the marginal effect via the delta method, randomly perturbing the price and comparing the difference between the predicted values of sales with the actual and perturbed prices. Estimation results are presented in Table 4, which shows that ignoring the censored nature of demand leads to an underestimation of the absolute price effect in all four regression models and the ensemble. Table 4 also shows that there is a substantial difference in the

estimates of the mean marginal effect of price across various models. Least squares, Ridge, Lasso regressions and Random Forest have large differences in the price effect due to different variable selection and the omitted variable problem. However, the omitted variable bias for the purpose of model inference may be corrected using the double lasso method and its various generalizations (see for example Belloni, Chernozhukov & Hansen, 2014). For the purpose of model comparison with and without censorship one may compare estimates presented in Table 4.

Tab. 4. Mean marginal effect of price in various models

	OLS	Ridge	Lasso	Random Forest	Ensemble
Uncensored	-0.742	-0.339	-1.079	-0.472	-0.661
	(0.028)	(0.025)	(0.026)	(0.023)	(0.024)
Censored	-1.440	-0.706	-2.187	-0.619	-0.920
	(0.010)	(0.036)	(0.013)	(0.012)	(0.015)

Notes: Mean marginal effect and its standard error is calculated from 1000 panel bootstrap sample draws and random perturbation of price on [0.01;1] standard deviations.

6. Conclusion

The methods of demand estimation in retail are quite developed in the literature. Previous demand studies report that ML methods have more predictive power (Varian, 2014; Bajari et al., 2015a; Witten et al., 2016;) while allowing for data censorship leads to unbiased estimates of demand parameters (Tobin, 1958; Chernozhukov & Hong, 2002; Chernozhukov et al., 2015). Nevertheless, there are still some gaps in the various methods for demand prediction. In particular, the potential for ML methods for censored demand prediction has not been discussed in the literature. This paper fills this void by introducing a new prediction algorithm dealing with censored demand. We propose an estimator for demand prediction that allows the use of the potential capacity of ML methods as well as accounting for data censorship. The research is based on the idea of comparing the prediction accuracy and parameter estimates of ML methods with and without censorship and combining various estimators with constrained linear ensemble models.

According to the results, two vital conclusions can be drawn. First, we show the better quality of ML method combinations for solving the prediction problem in retail demand. Secondly, we test the better predictive properties of models that take into account the censored nature of retail sales data. We also confirm a statistically significant downward bias of price effect parameter estimates in models without censorship. Since the research is conducted on the basis of real FMCG retail chain data, we can assert that the results have practical significance for retailers to establish

optimal prices for goods with different characteristics and at various time periods, as well as for optimal inventory management.

References

Aburto L. & Weber R. (2007). Improved supply chain management based on hybrid demand forecasts. *Applied Soft Computing*, 7(1), 136-144.

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.

Agrawal N., Smith S.A. (1996). Estimating negative binomial demand for retail inventory management with unobservable lost sales. *Naval Research Logistics*, 43, 839–861.

Ali Ö. G., Sayin S., Woensel T., & Fransoo J. (2009). SKU demand forecasting in the presence of promotions. *Expert Systems with Applications*, 36(10), 12340–12348.

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).

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2), 608-650.

Berry S., Levinsohn J. & Pakes A. (1995). Automobile prices in market equilibrium. *Econometrica*, 63(4), 841–890.

Bester C. A., & Hansen C. (2009). Identification of marginal effects in a nonparametric correlated random effects model. *Journal of Business & Economic Statistics*, 27(2), 235–250.

Bolton R. N. & Shankar V. (2003). An empirically derived taxonomy of retailer pricing and promotion strategies. *Journal of Retailing*, 79(4), 213–224.

Caro F. & Gallien J. (2010). Inventory Management of a Fast-Fashion Retail Network. *Operations Research*, 58(2), 257–273.

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.

Divakar S., Ratchford B. T., Shankar V. (2005). Practice Prize Article—CHAN4CAST: A Multichannel, Multiregion Sales Forecasting Model and Decision Support System for Consumer Packaged Goods. *Marketing Science*, 24(3), 305-523.

Einav L., Jenkins M., & Levin J. (2012). Contract Pricing in Consumer Credit Markets. *Econometrica*, 80(4), 1387-1432.

Einav L., Kuchler T., Levin J., & Sundaresan N. (2015). Assessing Sale Strategies in Online Markets Using Matched Listings. *American Economic Journal: Microeconomics*, 7(2), 215-247.

Fader P.S. & Hardie B.G.S. (2001). Forecasting Trial Sales of New Consumer Packaged Goods. In: Armstrong J.S. (eds) *Principles of Forecasting. International Series in Operations Research & Management Science*, vol 30. Springer, Boston, MA

Ferreira K.J., Lee B.H.A., & Simchi-Levi D. (2015). Analytics for an Online Retailer: Demand Forecasting and Price Optimization. *Manufacturing & Service Operations Management*, 18(1).

Heckman J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 48, 153–161.

Heien D., & Wesseils C.R. (1990). Demand Systems Estimation with Microdata: A Censored Regression Approach. *Journal of Business & Economic Statistics*, 8(3), 365–371.

Hoderlein S., & White H. (2012). Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics*, 168, 300–314.

Keane M.P. (2013). Panel Data Discrete Choice Models of Consumer Demand. *Economics Papers 2013-W08*, Economics Group, Nuffield College, University of Oxford.

Khan S., Powell J. L. (2001). Two–step estimation of semiparametric censored regression models. *Journal of Econometrics*, 103, 73–110.

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

Matzkin R. (2012). Identification in nonparametric limited dependent variable models with simultaneity and unobserved heterogeneity. *Journal of Econometrics*, 166(1), 106–115.

Nevo A. (2001). Measuring Market Power in the Ready–to–Eat Cereal Industry. *Econometrica*. 69(2), 307–342.

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

Pezente O.A. (2018). Predictive demand models in the food and agriculture sectors : an analysis of the current models and results of a novel approach using machine learning techniques with retail scanner data. *Master's thesis. MIT, Sloan School of Management, Technology and Policy Program*, Massachusetts Institute of Technology.

Qu T., Zhang J.H., Chan F.T.S., Srivastava R.S., Tiwari M.K., & Woo-Yong Park. (2017) Demand prediction and price optimization for semi-luxury supermarket segment. *Computers & Industrial Engineering*, 113(11), 91–102.

Richards T. J., & Bonnet C. (2016). Models of Consumer Demand for Differentiated Products. *Toulouse School of Economics Working Paper*, 16(741).

Ruiz F.J.R., Athey S., Blai D.M. (2017). SHOPPER: A Probabilistic Model of Consumer Choice with Substitutes and Complements. *Working Paper. Submitted to AOAS.* https://arxiv.org/abs/1711.03560

Shakya S., Kern M., Owusu G., & Chin C.M. (2012). Neural network demand models and evolutionary optimisers for dynamic pricing. *Knowledge-Based Syst*ems, 29, 44-53.

Tobin J. (1958). Estimation of Relationships for Limited Dependent Variables. *Econometrica*, 26(1), 24–36.

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

Witten I. H., Frank E., Hall M. A. & Pal C. J. (2016). *Data mining: practical machine learning tools and techniques*. Fourth edition. Morgan Kaufmann.

Appendix

Appendix 1

Tab.1. Descriptive statistics on the stores of pasta purchases

	Mean	St.deviation	Frequency	Share of Total
Hyper	1.765	2.142	27372	3%
Large	1.016	1.639	129894	16%
Middle	0.753	1.317	56367	7%
Small	0.561	1.119	560422	70%
Discounter	3.394	2.830	25945	4%

Tab. 2. Descriptive statistics on the pasta brands

	Mean	St.deviation	Frequency	Share of Total
Makfa	1.304	1.803	151757	18.27%
Granmulino	0.650	1.348	78484	9.35%
Pasta Zara	0.380	0.907	66235	8.12%
Gallina Blanca	0.332	0.762	60449	7.19%
Ameria	0.930	1.601	46863	5.66%
Maltagliati	0.521	1.116	44782	5.44%
Divella	0.281	0.707	34797	4.77%
Uvelka	1.143	1.899	35779	4.35%
Semgarnir	0.866	1.383	33224	3.96%
Soledoro	0.951	1.535	10232	3.64%
Makstory	0.495	0.987	27526	3.27%
Malosemeyka	1.699	2.050	27421	3.26%
Arrighi	1.128	1.768	24153	2.92%
Rummo	0.316	0.891	16278	2.44%
Grand Pasta	0.308	0.677	18737	2.22%
SunBonsai	0.208	0.560	15732	1.87%
Nobrand	0.180	0.496	9900	1.84%
Tomadini	0.596	1.091	12458	1.48%
Rollton	0.652	1.147	10860	1.28%
Shebekenskie	0.767	1.349	9976	1.19%
Dobrodeya	0.332	0.665	7312	0.87%
SenSoy	0.389	0.747	7297	0.86%
Garofalo	0.187	0.561	4580	0.83%
DeCecco	0.478	0.760	521	0.69%
Zalezione	0.312	0.879	4906	0.62%
Smak	4.899	2.902	3190	0.53%
ShaHeNodles	0.215	0.615	3469	0.42%
Barilla	0.693	1.046	2233	0.38%
Vnuk	0.490	0.831	2999	0.37%
Kammy	0.336	0.626	2732	0.33%
Business Lunch	0.952	1.669	1565	0.18%
ExtraM	0.828	1.238	1483	0.17%
Longkou	0.407	0.873	599	0.07%
3Glocken	0.288	0.522	417	0.05%

Makmaster	0.280	0.569	396	0.05%
Saratov	4.606	3.122	340	0.05%
KingLion	2.375	1.944	128	0.02%
Souzpishprom	5.333	3.132	15	0

Tab. 3. Descriptive statistics on the pasta country of origin

	Frequency	Share of Total
Russia	518504	64.81%
Italy	241350	30.17%
China	29700	3.28%
Vietnam	7297	0.91%
Kazakhstan	2732	0.34%
Germany	417	0.05%

Tab. 4. Descriptive statistics on the pasta packages

	Mean	St.deviation	Frequency	Share of Total
Weight (g)				
150	0.39	0.87	599	0.07%
200	0.43	1.16	8698	1.09%
250	0.27	0.63	59347	7.42%
300	0.35	0.86	29145	3.64%
350	0.62	1.18	4858	0.61%
400	1.01	2.31	172607	21.58%
450	1.39	2.79	167011	20.88%
500	0.68	1.67	285274	35.66%
600	0.74	1.78	4967	0.62%
700	0.93	2.27	2472	0.31%
800	1.39	2.33	58454	7.31%
950	0.71	1.08	6520	0.82%
1000	0.60	0.86	48	0.01%
Type of package				
Packet	0.91	2.10	776880	97.11%
Box	0.50	1.41	23120	2.89%

Tab. 5. Descriptive statistics on the pasta colour, form and type of flour

	Frequency	Share of Total
Colour of pasta		
Without colour	780801	97.60%
Multi	16083	2.01%
Green	2480	0.31%
Black	402	0.05%
Red	319	0.04%

Form of pasta		
Penne	94001	11.75%
Fusilli	91920	11.49%
Spaghetti	91600	11.45%
Stringozzi	86320	10.79%
Fettuccine	68805	8.60%
Sedani	65766	8.22%
Lumache	41200	5.15%
Conchiglie	34237	4.28%
Tortiglioni	31521	3.94%
Tagliatelle	31532	3.94%
Rotelle	28880	3.61%
Boccoli	28321	3.54%
Radiatori	21760	2.72%
Bucatini	13682	1.71%
Farfalle	13358	1.67%
Fettuccine	12400	1.55%
Fiori	12004	1.50%
Lasagna	10401	1.30%
Lagane	8488	1.06%
Scialatelli	6163	0.77%
Canestrini	6001	0.75%
Alfabeto	1679	0.21%
Type of flour		
Wheat	766241	95.78%
White rice	22321	2.79%
Bean	5278	0.66%
Buckwheat	3276	0.41%
Brown rice	1603	0.20%
Starch	562	0.07%
Rye	400	0.05%
Soybeans	317	0.04%

Appendix 2

Tab. 1. Descriptive statistics on the time of pasta purchases

	Mean	St.deviation	Frequency	Share of Total
2009	1.078	1.807	98023	12%
2010	0.865	1.545	123058	15%
2011	0.681	1.339	136610	17%
2012	0.626	1.287	159507	20%
2013	0.786	1.437	135197	17%
2014	0.772	1.433	147605	19%

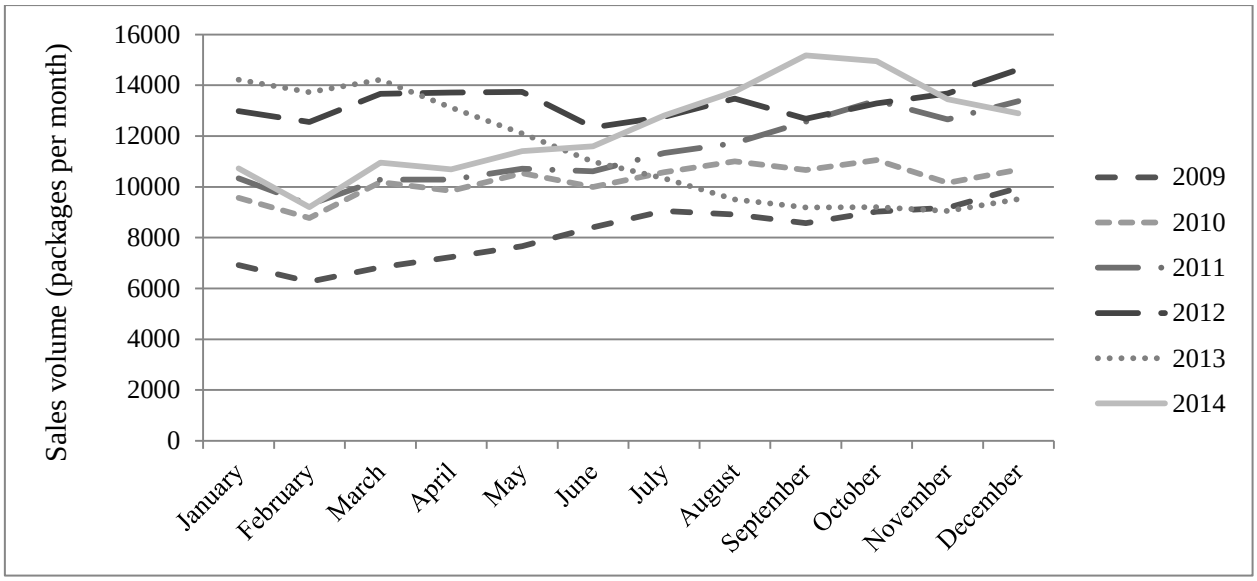

Fig. 1. Dynamics of pasta purchases in grocery chain stores by years

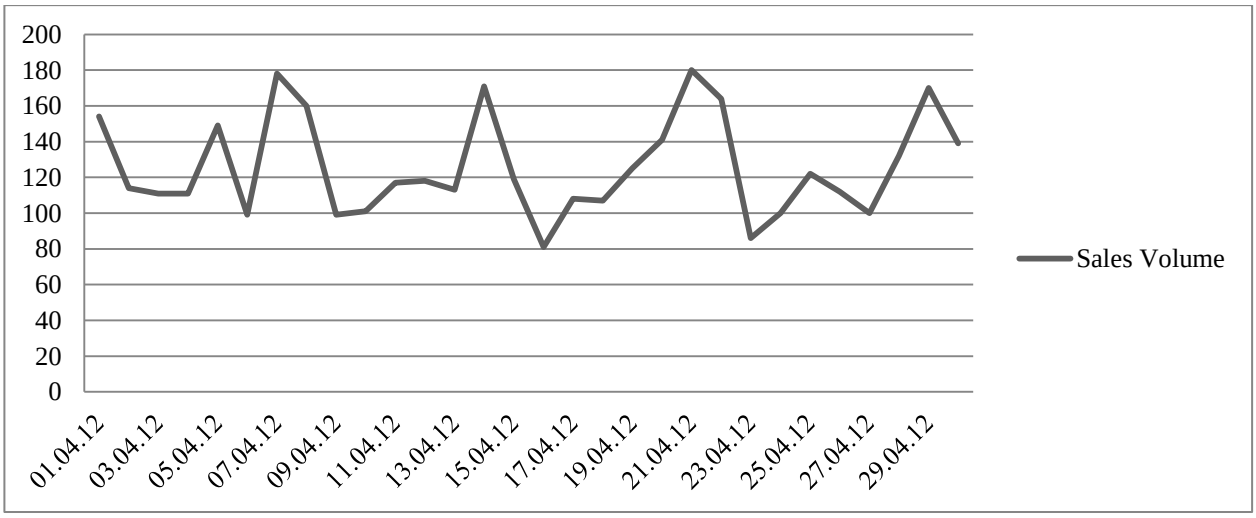

Fig. 2. Intramonth dynamics of pasta purchases (Example: April 2012)

Authors:
Evgeniy M. Ozhegov
National Research University Higher School of Economics (Perm, Russia).
Research Group for Applied Markets and Enterprises Studies. Research Fellow;
E-mail: tos600@gmail.com

Daria Teterina
National Research University Higher School of Economics (Perm, Russia).
Research Group for Applied Markets and Enterprises Studies. Young Research Fellow;
E-mail: dvteterina@gmail.com

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.