

Methods of Machine Learning for Censored Demand Prediction

Evgeniy M. Ozhegov¹ and Daria Teterina²

¹ National Research University Higher School of Economics, Perm 614070, Russia
tos600@gmail.com

² National Research University Higher School of Economics, Perm 614070, Russia
dvteterina@gmail.com

Abstract. In this paper, we analyze a new approach for demand prediction in retail taking into account data censorship and using machine learning methods. One of the significant gaps in demand prediction by machine learning methods is the unaccounted data censorship. Econometric approaches to modeling censored demand are used to obtain consistent and unbiased estimates of parameters. These approaches can also be transferred to different classes of machine learning models to reduce the prediction error of future prices or sales volumes. In this study we build two ensemble demand models with and without censoring demand, aggregating predictions for machine learning methods such as Ridge regression, LASSO and Random Forest. Having estimated the predictive properties of both models, we empirically prove the best predictive power of the model, taking into account the censored nature of demand.

Keywords: Demand Censorship, Machine Learning, Demand Prediction.

1 Introduction

The grocery retail market has been under the close scrutiny of economists over the past few decades. A surge of interest to this field occurred in the late 90's when the companies Nilson and IRI Marketing Research began to collect individual data on purchases of retail chains visitors. Advances in individual data availability drew the researchers' attention to the methods of machine learning. Investigators of the «big data» analysis have revealed the huge potential of machine learning methods for working with massive data sets, both in terms of the number of observations and predictors [4]. A number of scientists, including Agrawal & Schorling [1], Varian [5], Bajari, Nekipelov, Ryan, & Yang [2], showed greater predictive power of machine learning methods compared to traditional econometric approach. Therefore, today, when solving the problem of demand predicting, analysts' preference is often given to machine learning.

However, despite the significant breakthrough made by scientists in the evaluation of demand due to the methods of machine learning, there are still a lot of gaps, filling

of which can improve the predictive quality of models. One of such white spots is accounting for the censorship of demand. To date, there are a number of works devoted to censored demand prediction using traditional econometric approaches, as well as several studies on demand forecasting (without censoring) using machine learning methods, at the same time, there are no works that combine censorship and machine learning methods. Therefore, in our work, we will try to fill this gap by constructing an ensemble model of censored demand using machine learning methods and empirically checking its predictive properties on the data of the retail food chain.

2 Data

The study is conducted on the data, provided by the regional grocery chain. One product category – pasta – is selected for analysis. The choice of such a product category is justified by the high frequency of purchases of this product and the breadth of the range. The initial data from the grocery chain sales represents the full information on the pasta purchases for 6 years: from December 1, 2009 to January 31, 2014. The size of the analyzed sample, formed on the basis of the initial data, is 800000 observations. An observation reflects a stock keeping unit (henceforth SKU) that was available in a certain store on a specific date. It is known how many units of a single item were purchased every day and at what price they were sold. Also, with the use of the product catalog the number of physical characteristics for each SKU was restored. Thus, for each purchase, not only the price and sales volume are utilized, but also such characteristics as the colour and shape of pasta, the flour type, the volume and type of packaging, the origin country, the brand name. In addition to all of the above, for each observation, the format of the store where the purchase was made is determined and whether the product was a participant of the discount promotion. Note that more than 60% of sales were zero (see Fig.1) – this leads to the necessity of censorship accounting – what we do in our research.

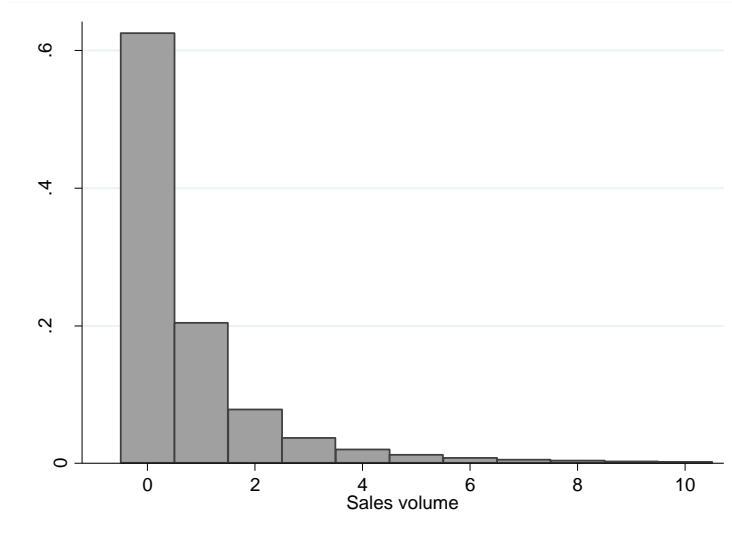


Fig. 1. Frequency histogram of pasta purchases.

3 Methodology

3.1 Econometric models of the demand function

Linear regression. The linear regression seems to be a typical model for demand estimation. It allows approximating the demand through a linear function. In our research the model specification will be the following:

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

where: q_{jmt} - the volume of the j^{th} SKU purchases committed in the store m on the day t .

X_{jmt} - the matrix of attributes including log of the price, product characteristics, promotional indicators, time attributes (dummies for a month, a year, an intra-week seasonality and holidays).

ε_{jmt} - an idiosyncratic shock to each product, market and time.

The model is estimated by the Ordinary Least Squares method. The linear regression model has become the basic specification for such machine learning methods, as: Ridge regression, LASSO and Random Forest.

Ridge regression and LASSO. As the next model specification, it was decided to choose a Ridge and LASSO regressions. Ridge regression refers to the so-called dense models: if we take all coefficients and sort them in descending order, we will note that there are quite a lot factors that strongly affect dependent variable. In our study, we assume the presence of a rather large number of factors (product characteristics, store and time attributes) that affect the demand, so the use of ridge regression seems to be reasonable.

To select a set of important parameters, the following function with penalty subfunction is formed:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_i (q_{imt} - \sum_j x_{jmt} \beta_j)^2 + \lambda \sum_j \beta_j^2 \quad (2)$$

where: $\lambda \sum_j \beta_j^2$ - penalty subfunction.

For λ selection, the cross-validation is used: lambdas are sorting through from infinity up to 0 – thus, overfit in model is rising, bias is reducing and variance is increasing.

As for the LASSO, the algorithm of $\hat{\beta}$ estimation is practically the same – the main difference is in the penalty subfunction:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_i (q_{imt} - \sum_j x_{jmt} \beta_j)^2 + \lambda \sum_j |\beta_j| \quad (3)$$

Random Forest. The random forest is based on tree construction technology. The regression tree represents a set of rules that determine the value of the parameter of the regression function. Tree-based methods divide the characteristic space into a number of hypercubes and adapt the effects to each section, depending on the value of the independent variables [3].

The method works by the following algorithm: let us consider to have the number of regressors . First of all, for each variable X_j , the sample is splitted into two parts $R_1 (z_j < \tilde{z}_j)$ and $R_2 (z_j > \tilde{z}_j)$ (where R_1 and R_2 are areas of the observational space where the entire sample is divided into two parts with respect to the variable X_j in the clipping point \tilde{z}_j). Secondly, $\hat{\beta}_1$ and $\hat{\beta}_2$ are constructed:

$$\begin{aligned} \hat{\beta}_1 &= \underset{\beta}{\operatorname{argmin}} \sum_{i \in R_1} (q_i - x_i \beta)^2 \\ \hat{\beta}_2 &= \underset{\beta}{\operatorname{argmin}} \sum_{i \in R_2} (q_i - x_i \beta)^2 \end{aligned} \quad (4)$$

Then, according to formula (5), the residuals are estimated:

$$\begin{aligned} \hat{\varepsilon}_{1i} &= q_i - x_i \hat{\beta}_1 \\ \hat{\varepsilon}_{2i} &= q_i - x_i \hat{\beta}_2 \end{aligned} \quad (5)$$

Next, the sum of the balances in both areas is minimized:

$$\sum_{i \in R_1} (\hat{\epsilon}_{1i})^2 + \sum_{i \in R_2} (\hat{\epsilon}_{2i})^2 \rightarrow \min_{j, \bar{z}_j} \quad (6)$$

Thus, such variable and cut-off point are chosen, where in each region we build our regression, and these regressions provide the best total prediction for q . Afterwards, we can still partition the regions R_1 and R_2 (or only one of them) in the same way, and then choose the best error partitioning.

The random forest is a complication of the above described model. In the case of a random forest, a lot of trees are built, but for each tree its own subset of factors is selected, and then an average prediction along the trees is constructed.

3.2 Estimation algorithm

According to the literature review, machine learning methods are better able to cope with demand predicting (in particular, in grocery retailing), because they produce better out-of-sample fits than linear models without loss of in-sample fit quality [2]. Therefore, in order to achieve the most accurate prediction, we assume to use three methods of machine learning and only one traditional econometric approach - linear regression (as a basic model). In our study, we assume to partially follow the algorithm described in the [2] research, modernizing it somewhat by adding the stages of estimating censored models. The main steps of the empirical part of the study are supposed to look like this:

1. Theoretical construction of models: Linear regression, Ridge regression, LASSO regression and Random Forest.
2. Splitting the data randomly into three groups for the subsequent cross-validation, where 25% of the data falls into the test sample, 15% - in the validation, and 60% - in the training.
3. Build models on the training sample, compare out-of-sample errors and determine the predictive power of each model.
4. Building a prediction on the validation sample and obtaining weights with which each model should be included in the final ensemble model.
5. Formation of an ensemble model of demand forecasting, its application to a test sample, determination of predictive power.
6. Implementation of paragraphs 1-5 for models with censoring.
7. Comparison of the predictive power of ensemble models with and without censoring.

Steps 6 and 7 are new for the algorithm proposed by the Bajery et. al [2]. They allow us to take into account the censored nature of the analyzed data. Therefore, to reveal the essence of the 6th and 7th steps of the previous algorithm, we propose to use the following steps for censorship accounting:

1. Construct dummy for observation censorship: $d = 1$ if the sales volume of the j -th SKU purchases committed in the store m on the day t is greater than zero and $d = 0$ - otherwise.
2. Build probit model, where the created in the previous step dummy variable is taken as the dependent variable, and independent variables of all abovementioned models (Linear, Ridge, Lasso Regressions and Random Forest) used as regressors.
3. Classify observations into censored and uncensored $\hat{d} \in \{0;1\}$ based on the probit estimates.
4. Train the Linear Regression on the dataset created from observations classified as uncensored ($\hat{d} = 1$).
5. To build out-of-sample errors and determine the predictive power of the model.
6. Repeat the 4th and 5th steps for the Ridge, Lasso and Random Forest models.

3.3 Ensemble model construction

After building models (Linear regression, Ridge, LASSO 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 model. The main steps of this phase are described in [2] and look as follows:

1. Take the validation dataset. Treat the predicted values of the dependent variables 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.
2. Take the test dataset. Use the fitted models for prediction in the test set, after that apply the model weights from the previous step, sum them up and construct the linearly combined prediction.

The above algorithm is repeated for both model classes: with and without censorship.

4 Results

Since more than 60 percent of sales are zero, we should check the parameter estimates for the need to use the censored regression model, testing for a bias in the multiple regression model (1) versus the censored regression model. For this, the parameter estimates for two abovementioned specifications were calculated (Table 2).

Table 1. Estimation results for linear regressions with and without censorship of dependent variable

Variable	Linear regression	Censored linear regression
Log(price)	-5.581*** (0.058)	-7.795*** (0.078)
Size weight	0.006*** (0.0001)	0.008*** (0.0002)
Promotion	0.099*** (0.028)	-0.120*** (0.034)
Price*Size weight	0.001*** (0.000)	0.001*** (0.000)
Country of origin		
China	1.233* (0.673)	2.302*** (0.703)
Colour		
Black	5.076*** (0.480)	8.589*** (0.511)
Green	-0.006 (0.107)	0.476*** (0.129)
Multi	1.033*** (0.069)	1.937*** (0.087)
Red	-0.634*** (0.217)	-0.056 (0.236)
Flour		
Bean	1.600*** (0.122)	2.685*** (0.147)
Brown rice	0.015 (0.115)	0.663*** (0.149)
Buckwheat	1.332*** (0.174)	2.838*** (0.264)
Starch	0.676*** (0.246)	1.153*** (0.303)
Soybeans	-1.743*** (0.348)	-1.308*** (0.381)
Time attributes		
Holiday	-0.178*** (0.048)	-0.314*** (0.062)
Sunday	0.658*** (0.033)	0.833*** (0.041)
Monday	0.074** (0.034)	0.118*** (0.042)
Wednesday	0.034 (0.033)	0.045 (0.042)
Thursday	0.086** (0.034)	0.158*** (0.042)
Friday	0.454*** (0.033)	0.605*** (0.041)
Saturday	0.846*** (0.034)	1.137*** (0.041)

2009	-0.179*** (0.041)	0.204*** (0.048)
2010	-0.546*** (0.038)	-0.346*** (0.045)
2011	-3.152*** (0.061)	-
2012	-0.965*** (0.55)	-1.373*** (0.089)
2013	-0.593*** (0.034)	-0.975*** (0.041)
January	-0.226*** (0.044)	-0.311*** (0.056)
February	-0.152*** (0.044)	-0.207*** (0.055)
March	-0.076* (0.043)	-0.091* (0.053)
May	-0.093** (0.044)	-0.099* (0.054)
June	-0.048 (0.045)	-0.052 (0.054)
July	-0.161*** (0.045)	-0.152*** (0.053)
August	-0.071 (0.045)	-0.096* (0.053)
September	-0.075* (0.044)	-0.118** (0.053)
October	-0.203*** (0.044)	-0.229*** (0.054)
November	-0.116** (0.045)	-0.252*** (0.058)
December	0.072 (0.046)	0.041 (0.056)
Store type		
Discounter	2.278*** (0.012)	4.413*** (0.147)
Middle	0.272*** (0.007)	1.084*** (0.005)
Large	0.578*** (0.005)	1.785*** (0.007)
Hyper	1.289*** (0.001)	3.259*** (0.014)
Const	15.127*** (0.760)	19.739*** (0.818)
R^2_{adj}	0.229	0.297
N	800000	800000
K	95	94
	1.256	1.220

Notation: Parameters' estimates are presented in the cells of the table, standard errors – in the brackets.
Significance levels: $p^* < 0.1$, $p^{**} < 0.05$, $p^{***} < 0.01$

N – the number of observations, K – the number of parameters.

Brands and forms of pasta are also considered in the model as control variables.

Basic category: Brand – «Maltagliati», country of origin – Italy, colour – without colour, flour – wheat, day of the week – Tuesday, year – 2014, month – April, store type – small.

As can be seen, the effects of explanatory variables slightly vary over specifications, but the sign of parameter estimates is practically the same in both models. In the second specification, the effects value of almost all parameters is larger in modulus. This can be explained by the fact that linear regression without censoring underestimates the values of the parameters.

What is more vital to notice this is the better explanatory properties of the model accounting censorship. Thus, the value of the adjusted R^2 for censored linear regression is higher than for model without censorship.

After evaluating the parameters of the basic linear model, the actual dependent variable is fitted in the training set on each of the four models (Linear regression, Ridge regression, Lasso regression and Random Forest). Then, for every model the measure of the prediction quality, expressed by Root Mean Square Error (RMSE), is calculated (Table 2).

The next step that is taken after RMSE estimation is the determination of the weights of each model for their inclusion in the final ensemble model. To do that the validation set is used: predicted values of the dependent variable from four models are treated as regressors into constrained linear regression, where the actual value of sales volume is used as dependent variable. Constrains for the linear regression are as follows: firstly, the sum of the estimates of the model parameters must be equal to one (since in the future the parameter estimates will be used as weights); secondly, the value of each parameter should be positive (for the same above-described reason). The results of constrained linear regressions estimation are presented in Table 2 as models weights in the combined model.

Table 2. Root Mean Square Error (RMSE) for model specifications with and without censorship accounting and models' weights in ensemble model.

	RMSE		Weight in the linear combined model	
	Without censorship accounting	With censorship accounting	Without censorship accounting	With censorship accounting
Linear regression	1.256	1.220	1%	3%
Ridge regression	1.255	1.218	13%	11%
Lasso regression	1.244	1.203	42%	39%
Random Forest	1.198	1.164	44%	47%
Total RMSE for ensemble model	1.163	1.116		

According to the estimation results, both ensemble models with and without censorship accounting have better performance than any of the evaluated models individ-

ually. Moreover, the ensemble model accounting censorship of the data, has a better predictive power, which is indicated by the comparatively smaller RMSE.

All in all, two vital conclusions can be drawn from the results of this study: firstly, we showed the real strength of machine learning methods combination for solving the prediction problem in retail demand. Secondly, we partially filled the gap associated with the demand censorship, showing the best predictive properties of models that take into account the censored nature of the retail data.

References

1. Agrawal, D., Schorling, C.: Market share forecasting: An empirical comparison of artificial neural networks and multinomial logit model. *Journal of Retailing* 72(4), 383–408 (1996).
2. Bajari, B. P., Nekipelov, D., Ryan, S. P., Yang, M.: Machine Learning Methods for Demand Estimation. *The American Economic Review* 105(5), 481–485 (2015) .
3. Ozhegov, E. M., Ozhegova, A. Regression tree model for analysis of demand with heterogeneity and censorship. Working Paper
4. Richards, T. J., Bonnet, C.: Models of Consumer Demand for Differentiated Products. TSE Working Papers 16-741 (2016), Toulouse School of Economics (TSE).
5. Varian, H. R.: Big data: New tricks for econometrics. *The Journal of Economic Perspectives* 28(2), 3–27 (2014).