Comparison of Machine Learning Algorithms in Restaurant Revenue Prediction

Chapter · February 2020

DOI: 10.1007/978-3-030-39575-9_4

2 authors:

Stepan Gogolev
National Research University Higher School of Economics
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Evgeniy Ozhegov
National Research University Higher School of Economics
26 PUBLICATIONS **8** CITATIONS
SEE PROFILE

Comparison of Machine Learning Algorithms in Restaurant Revenue Prediction*

Stepan Gogolev[1,2][0000−0002−5153−3451] and Evgeniy M. Ozhegov[1,3][0000−0001−6733−4150]

[1] National Research University Higher School of Economics, Research Group for Applied Markets and Enterprises Studies Perm, Russia
[2] s.l.gogolev@gmail.com
[3] tos600@gmail.com

Abstract. In this paper, we address several aspects of applying classical machine learning algorithms to a regression problem. We compare the predictive power to validate our approach on a data about revenue of a large Russian restaurant chain. We pay special attention to solve two problems: data heterogeneity and a high number of correlated features. We describe methods for considering heterogeneity — observations weighting and estimating models on subsamples. We define a weighting function *via* Mahalanobis distance in the space of features and show its predictive properties on following methods: ordinary least squares regression, elastic net, support vector regression, and random forest.

Keywords: Weighted regression · Machine learning · Revenue prediction.

1 Introduction

A global trend of collecting and storing information creates the demand for methods of analyzing and exploiting it. For instance, firms are interested in quantifying some qualitative features, explaining and creating predictions of consumers behavior.

Nowadays, a large number of machine learning methods provides algorithms of model creation to optimize almost any business process. However, some models and methods should be applied to some specific kinds of data only [6]. There is never a unique answer to the question about the choice of models and methods. Therefore, researchers benefit as much from the use of a methodology the most suitable to their dataset and a specific problem.

In the paper, we discuss the advantages of common machine learning methods applied to the problem of restaurant revenue prediction in Russian cities. A

* The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE University) in 2019 — 2020 (grant 19-04-048) and within the framework of the Russian Academic Excellence Project "5-100".

similar problem is studied in paper [7] where decision trees are employed to heterogeneous objects analysis. We follow this paper by applying other methods and show its suitability to a problem of franchise restaurant revenue prediction and choice of the best location for a new restaurant.

In the paper, we use information collected from operating restaurants to predict revenue for potential restaurants in other cities. According to the industrial organization literature (See, for ex., [2], [6], and [13]), revenue of the restaurant depends on a number of characteristics including number of people in the local market, the average wage, a size of different target age groups, the number of direct and indirect competitors, characteristics of nearest competitors, etc. These features are closely related to each other, so the problem of partial feature collinearity occurs.

Heterogeneity of cities is another typical challenge of the location choice problem [14]. The problem is to study cities with a population from 10 thousand up to 12 million people in Russia. The proposed methodology is able to predict revenue for restaurants in cities of various size, taking into account their qualitative differences and variation in population and average wage.

2 Problem Statement

In this paper, we compare different methods for a problem of revenue prediction for franchise restaurants in cities where are no such restaurants. Restaurants belong to a franchise fast-food industry where each restaurant within a franchise is very standardized between cities. There are major distinctions in possible revenue due to the difference in cities, a location of a restaurant within a city and the degree of competition within a city.

Revenue is a relevant factor of success in a franchise as it reflects the number of clients visiting a restaurant and it does not depend on the quality of management, costs at the period and other in a franchise. Franchise system guarantees equal costs and profitability of restaurants in different cities. The reason is in a common technology of production and similar pricing on raw materials. It allows to concentrate efforts on comparison of cities suggesting other factors being equal.

The chosen franchise has about 300 restaurants in 184 cities. We analyze monthly revenue for the last 3 years [4]. From the starting point the data more than half of the restaurants were opened. Some restaurants were opened less than three years ago, so we collected 5889 observations as an unbalanced panel. Maintaining a panel structure instead of aggregating revenue is necessary to avoid seasonal bias.

According to the main goal of the study, we focus on objects' features (characteristics of cities) and creating a prediction for an average restaurant in the city given that there are no restaurants of the chosen franchise. We have three major

[4] Data is available at goo-gl.ru/5vIE

groups of predictors: seasonal factors, specific restaurant characteristics (operating period and part of the revenue from delivery), and market environment features.

The market environment features consist of demographic[5] and competitors' characteristics[6]. The first group of variables includes detailed information about consumers: market size, its specific segments, consumer income. The second one describes firms' behavior on the market: a number of direct and indirect competitors, average restaurant bill, average estate price, and wage. Fast-food restaurants compete simultaneously on several markets: some types of cafes, restaurants, food delivery, etc. These markets are closely linked to each other and have common features. However, employing market and city characteristics raise the challenge of partial features collinearity that may provide a prediction bias.

Heterogeneity is another feature of the data. We analyze heterogeneity through heteroscedasticity of the errors and the presence of outliers. Common White test [15] proves the presence of heteroscedasticity at the 1% significance level. It can be interpreted as follows. Estimating model with linear regression gives different variation of error (degree of model accuracy) for different values of predictors. As for the outliers, the dataset contains some non-representative objects (cities). For instance, Moscow with a population equal 12 million inhabitants is almost 100 times than the average population of cities under consideration. Using the coefficient of variation [3] we check homogeneity of cities reveal the heterogeneity by following features: population, average wage, number of opened cafes, pizzerias and restaurants.

3 Methodology

3.1 Model Comparison Algorithm

In order to overcome the issue with the presence of heteroscedasticity and outliers described above, we follow [16] and use MAPE (mean absolute percentage error) instead of MSE (mean squared error) as a prediction quality criterion. We compare the predictive power of models by MAPE in order to give lower weight to predictions with non-representatively large errors. It does not have the property of underestimation the largest errors like mean squared error or other metrics using squared errors [16]. As we calculate model errors for all objects, MAPE shows the average absolute error of the model in percent of the average value of the target variable in our case. The lower MAPE is, the more predictive power of the model has. It will be useful for further interpretation of the metric.

The next important step is choosing the technique for assessing the prediction power of the models. We compare out-of-sample predictions due to possible overfitting problem by the procedure of leave-one-out cross-validation for model

[5] Data are taken from gks.ru
[6] Data taken from 2GIS

parameters and 10-fold cross-validation for hyperparameters. Now we move to describe the steps of cross-validation in detail.

The main idea is averaging the error of prediction among all available objects through predicting new data that was not used in estimating. For this, we choose one city in leave-one-out and $\frac{1}{10}$ of all cities in 10-fold cross-validation as a test sample and exclude appropriate observations from the dataset, remaining observations from the training sample. Then we train a model on different training objects and choose optimal hyperparameters according to mean absolute error on 10 test samples. Finally, we train the model hyperparameters optimally selected at the previous step on different training samples and create predictions for the test observations related to one object. We repeat these steps for other test cities. As a result, we obtain a vector of out-of-sample predictions for all cities in the dataset. Then we compare mean absolute error between predictions and actual values and test what model gives the best results. To validate our results and estimate the possible overfitting issue, we also calculate the in-sample coefficient of determination (R^2) that shows the proportion of the explained variance in target variable and reflects the goodness of fit of a model on a training data.

3.2 Prediction Models

We follow [6], [7] and use four different methods of regression estimation: linear regression (OLS), elastic net (ELNET), support vector regression (SVR) and random forest regression (RF). The last three methods allow to overcome a feature collinearity problem.

Firstly, we make revenue predictions using naive model and linear regression model to compare other results with these baselines. Naive model is OLS with a constant only. We calculate confidence intervals for models fit *via* bootstrapping with 200 replications test the statistical difference in a fit.

In the linear regression model we minimize the sum of squared errors to obtain optimal parameters values β in a linear index:

$$\hat{\beta}_{OLS} = \underset{\beta}{\operatorname{argmin}}(y - x\beta)'(y - x\beta). \tag{1}$$

where y is monthly revenue of restaurant in the city in the month, x is a vector of features including characteristics of the city and competitors in city , specific characteristics of the restaurant in the period and seasonal factors that depends on the period, and β are parameters to be estimated.

This method may suffer from a high degree of partial multicollinearity. We check the variance inflation factors (VIF) [10] for the group of competitors characteristics. In this group, all 12 factors have VIF more than 20 that signals on the problem of correlated factors.

An elastic net regularization method is one of the solutions to the multicollinearity issue [8]. This method minimizes the sum of squared errors penalized on the absolute and squared values of estimated parameters:

$$\hat{\beta}_{ELNET} = \underset{\beta}{\operatorname{argmin}}(y - x\beta)'(y - x\beta) + \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2. \quad (2)$$

where λ_1 and λ_2 are parameters of regularization to be estimated.

As a result of the optimization problem solving, some predictors can be excluded from target variable prediction when its parameter value β shrunk to 0. Therefore, the model does not take into account additional information that excluded variables contain.

SVR provides another way of estimating model parameters. Unlike least squares methods, SVR avoids explicit specification of the regression equation [12]. SVR training process depends on the kernel function that defines the relationship between the target variable and predictors. Hence, it is crucial to concentrate on the choice of a kernel function. We check common kernel functions including Gaussian, linear, and polynomial. Generally, type of the kernel function can be chosen based on the type of relation between features if it is known. We use 10-fold cross-validation to select the best kernel function and calibrate its hyperparameters (regularization parameter C, tolerance ε) and degree for polynomial kernel function.

The last method we apply is a RF regression — an ensemble of regression trees. Training of a tree is an iterative process where the input data is split by predictors into smaller groups with different predicted value in each partition group. Combination of such trees is an ensemble that allows to reduce prediction variance and improve out-of-sample prediction power. Another advantage of using regression trees is the revelation of a nonlinear relation between the target variable and predictors [9].

The quality of a RF model mainly depends on the following parameters: the number of trees in an ensemble and the number of predictors randomly sampled in each split. The former should be large enough to reduce the variance of prediction, raised as a result of correlated variables in input data. The last parameter corresponds to the quality of the model. The higher the number of variables used, the better the quality of the model and the higher probability of overfitting. We tune both parameters using the out-of-bag estimation of the model. It is based on the sampling of test observations and calculating prediction error for observations which were not used in the training process of the model. It is proved that out-of-bag error estimations tend to leave-one-out cross-validation estimation what makes them a reliable method for selecting parameters of RF [4].

3.3 Accounting for Heterogeneity

Accounting for a heterogeneity requires the use of specific methods. We use two common ways: weighting of observations and training of model on subsamples through data partition [1]. The first method consists of giving various weights to different observations in the process of model training, while the second way assumes reducing objects in training dataset to the most relevant ones.

Both approaches use implicitly a function that assigns to all objects in the dataset (cities in our case) a value that reflects the proximity of objects. We can define this function as a distance function between two points in the space of objects characteristics. Let us describe steps on implementation of methods accounting for heterogeneity to a problem of revenue prediction.

To create a model that provides a prediction of revenue for the restaurant in a test city i, we should train the model on the remaining dataset $-i = \{j \in N, j \neq i\}$ (observations related to training cities). After that, we calculate distances from each training city in $-i$ to the test city i. In the case of weighting observations, the next step is transforming distances into weights and estimating the model. Naturally, we put higher weight to observation with a lower distance to the test city. Therefore, it is possible to use the inverse function to transform distance into weight. In this work, we use inverse power function. The definition of weight is follows:

$$w_{ij} = {d_{ij}}^{-\gamma} \tag{3}$$

where γ is a parameter to be estimated.

Now we turn to another case of training model on a subsample. We introduce the rule that defines an interval of values of distance that indicates whether to include the city in training dataset or not. We define bounds of the interval so that there are 75% of observations the most similar to the test city. The percent of observations that will be included in the training dataset is chosen according to the size of the overall dataset. That is to say, we include observations related to object j in training dataset if $d_{ij} \leq Q_{0.75}$, where $Q_{0.75}$ is a 75%-th quantile of the distance distribution among all j.

After that, we define a space of characteristics and a distance function between objects. As we can distinguish the most heterogeneous variables, a possible solution is to consider all of them in a distance function. We can construct overall distance as a sum of distances in all dimensions only if dimensions are orthogonal. Otherwise, distances in dimensions responding for correlated predictors would be overfitted. Mahalanobis distance function allows to include values from different dimensions with different weights [11]. It measures the difference between the object and the distribution of other objects in terms of standard deviations. The distance between each training observations and test observation i with the covariance matrix of predictors Ω is follows:

$$d_{ij} = \sqrt{(x_i - x_j)' \Omega^{-1} (x_i - x_j)}. \tag{4}$$

Mahalanobis distance is applicable to correlated variables, hence researcher can choose any combination of variables that forms a space of objects characteristics. In this work, we include three the most heterogeneous variables in weighting function: population, average wage and the number of restaurants competitors in the city.

After describing two procedures of heterogeneity eliminating it is necessary to discuss the compatibility of these procedures with 4 ML methods, starting

with the simplest OLS and ELNET methods. The addition of weighting function to them modifies objective functions presented in Eq. (1-2) into weighted errors minimization problems:

$$\hat{\beta}^i_{OLSW} = \underset{\beta}{\operatorname{argmin}}(y_{-i} - x_{-i}\beta)' \operatorname{diag}(w_i)(y_{-i} - x_{-i}\beta). \tag{5}$$

$$\hat{\beta}^i_{ELNETW} = \underset{\beta}{\operatorname{argmin}}(y_{-i} - x_{-i}\beta)' \operatorname{diag}(w_i)(y_{-i} - x_{-i}\beta) + \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2. \tag{6}$$

Estimation on subsamples for these methods is acceptable but has a significant drawback. Subsampling reduces the size of the dataset and estimation efficiency. Strict selecting of observations in training dataset may results in the poor model due to insufficient information in selected data. At the same time, soft selection can keep training dataset unchanged.

Turning to a SVR model, the use of weighting there is not recommended. The algorithm of SVR assumes estimation of the model, based on the training data points nearest to the hyperplane. This means that the model is automatically trained on observations closest to the "average" observation, while outliers are ignored. The most suitable way to train the SVR model for some test object is estimating on a subsample where test object represents average observation. Such a model shows better predictive power despite the small training sample size.

The problem of heterogeneity in RF regression is eliminated automatically due to splitting input data into smaller groups. In this model quality of prediction mostly depends on the number of training objects similar to test ones. If it is large enough, regression trees are able to divide observations into groups better than other methods. However, in the lack of similar objects and observations, RF regression often does not show good fir due to low ability to extrapolate relations.

In the next section, we show the comparison of these algorithms and provide results for an ensemble of simple predictors. We find optimal weights for models in an ensemble using constrained linear regression and explain the resulting weights.

4 Results

Out-of-sample prediction for a city assumes creating a training model on the sample that does not contain any information about the city for what we make a prediction for. Table 1 shows measures of accuracy — MAPE and R^2 — for out-of-sample prediction as an error percentage of mean overall monthly revenue and in-sample coefficient of determination.

The naive model gives the baseline out-of-sample prediction for comparison with other methods as we assumed. 95% confidence interval for MAPE in this

Table 1. Prediction power of models

	Mean	SD	Out-of-sample MAPE	In-sample R^2
y	2 643 306	1 571 221	–	–
Model for $\hat{y}$:				
Naive model	2 643 306	0	59.44%	0.00
OLS	2 367 107	1 476 862	38.17%	0.58
OLS on a subsample	2 332 341	1 622 088	41.26%	0.37
OLS with weighting	2 365 270	1 461 773	36.64%	0.56
ELNET	2 324 616	1 123 046	34.54%	0.45
ELNET on a subsample	2 453 445	1 518 215	37.65%	0.53
ELNET with weighting	2 296 467	1 096 400	33.29%	0.42
SVR	2 387 787	1 271 851	32.76%	0.65
SVR on a subsample	2 353 508	1 301 179	33.37%	0.70
RF	2 379 884	959 417	30.70%	0.93
Ensemble	2 372 320	1 307 328	23.59%	0.97
Number of observations	5 889			
Number of objects	184			
Number of predictors	43			

model is from 46.3% to 72.6%. The highest MAPE in other methods is 41.3% (in OLS estimated on a subsample), so we come to the conclusion that all described methods are statistically significant and provide a better fit than the prediction by mean y. With the improvement model from OLS to ELNET method, MAPE decreases from 38 to 34.5%. It proves the benefits of regularization methods usage in the case of the high number of correlated variables. Modifying the least squares method with weighting function (optimal value of parameter γ is equal to 0.8) also improves the predictive power of the model. It decreases the variability of predictions (SD falls) with error level. Estimating model on a subsample does not improve any model due to the decrease of efficiency of models trained on smaller samples. Similar conclusions are associated with estimating SVR on a subsample. Overall, combining several methods (elastic net method and weighting function) allows to achieve the best quality of out-of-sample prediction for the least squared method.

SVR and RF regressions outperform results of linear regressions: MAPE is 32.8 and 30.7 % respectively. The random forest model works better than other models at the regression problem with heterogeneity by construction. RF does not reveal averaged relations and does not extrapolate relations between variables to uncommon values of them. That is the reason why it is useless for predictions revenue in atypical cities. However, it is the best among the considered method for predicting. A higher value of in-sample R^2 indicates on a possible overfitting problem there. Using the ensemble of models improves results in terms of the coefficient of determination and MAPE because of overcoming overfitting problem and combining advantages of considered methods. Out-of-sample error in the ensemble is 2.5 times fewer that error in naive model and the lowest among all methods.

Although we show a statistically significant difference in accuracy by comparison errors of prediction with errors of the naive model, there is a lot of ways to improve results. As we solve problem of cities comparison and not the problem of optimal location within a city, we cannot include in the model some specific features (for instance, spatial characteristics about competitors inside the city).

To sum up, we compare the predictive power of some models on the dataset with heterogeneity and correlated predictors. Results show that the ensemble has properties to overcome both problems and has the lowest mean absolute error and the highest coefficient of determination. Moreover, we show the advantages of weighting observations in the estimating process and possible drawbacks of estimating models on subsamples.

5 Conclusion

In this paper, we summarize the methodology of constructing a model with the best predictive power to forecast revenue in the restaurant in the out-of-sample city. We describe methods of heterogeneity elimination in the model: observations weighting and data partition with the following estimation on subsamples. Additionally, we suggest some ways of dealing with collinearity problem: an elastic net method, support vector and random forest regressions. We show advantages of those methods under different assumptions and validate these statements at the problem of revenue prediction.

Basically, the paper can be extended in two ways. First of all, it is possible to consider other methods of solving problems: for instance, principal component analysis for reducing the number of correlated predictors or more detailed analysis of ensemble trees algorithms (bagging, boosting, etc.). The second way is to use a more accurate approach to compare model prediction power. For each model we can calculate a MAPE confidence interval using the bootstrap. Computing confidence intervals allows to compare the predictive power of models with more certainty.

References

1. Athey, S., Imbens, G.: Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, **113(27)**, 7353–7360 (2016)
2. Berthon, P., Holbrook, M., Hulbert, J.: Beyond market orientation. A conceptualization of market evolution. Journal of Interactive Marketing, **14, 3**, 50—66 (2000)
3. Bennett, B.: On an approximate test for homogeneity of coefficients of variation. In Contribution to Applied Statistics, 169–171 (1976)
4. Breiman, L.: Heuristics of instability and stabilization in model selection. The annals of statistics. **24(6)**, 2350–2383 (1996)
5. Chang, M., Lin, C.: Leave-one-out bounds for support vector regression model selection. Neural Computation. **17(5)**, 1188–1222 (2005)
6. Chiang, W., Chen, J.., Xu, X.: An overview of research on revenue management: current issues and future research". Int. J. Revenue Management, **1, 1**, 97—128 (2007)

7. Kim, S., Upneja, A.: Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. Economic Modelling, **36**, 354–362 (2014)
8. Lee, D., Lee, W., Lee, Y., Pawitan, Y.: Sparse partial least-squares regression and its applications to high-throughput data analysis. Chemometrics and Intelligent Laboratory Systems. **109(1)**, 1–8 (2011)
9. Liu, S., Dissanayake, S., Patel, S., Dang, X., Mlsna, T., Chen, Y., Wilkins, D.: Learning accurate and interpretable models based on regularized random forests regression. BMC systems biology. **8(3)** (2014)
10. Mansfield, E., Helms, B.: Detecting multicollinearity. The American Statistician, **36(3a)**, 158–160 (1982)
11. Neale, M. C.: Individual fit, heterogeneity, and missing data in multigroup sem. Modeling longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples. Hillsdale, NJ: Lawrence Erlbaum Associates (2000)
12. Cristianini, N., Shawe-Taylor, J.: An introduction to support vector machines and other kernel-based learning methods. Cambridge university press (2000)
13. Walzer, N., Blanke, A., Evans, M.: Factors affecting retail sales in small and mid-size cities. Community Development. **49:4**, 69—484 (2018)
14. Wang, K., Wai, K., Liping, L., Xiaowen, F.: Entry patterns of low-cost carriers in Hong Kong and implications to the regional market. Journal of Air Transport Management. **64B**, 101—112 (2017)
15. White, H.: A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica, **48(4)**, 817–838 (1980)
16. Willmott, C., Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research. **30(1)**, 79–82 (2005)