Predictive Analytics for Sales Time Series

Bohdan M. Pavlyshenko

SoftServe, Inc.,
Ivan Franko National University of Lviv Lviv, Ukraine
email: b.pavlyshenko@gmail.com

Abstract—**The use of different approaches for prediction of sales time series has been considered. The case with multilevel stacking model for sales prediction in the store chain has been described.**

Index Terms—**predictive analytics, sales forecasting, times series.**

I. INTRODUCTION

Sales prediction is an important part of modern business intelligence. First approaches one can apply to predict sales time series are such conventional methods of forecasting as ARIMA and Holt-Winters. But there are several challenges while using these methods. They are: multilevel daily/weekly/monthly/yearly seasonality, many exogenous factors which impact sales, complex trends in different time periods. In such cases, it is not easy to apply conventional methods. There is the implementation of SARIMA for time series with seasonality, and SARIMAX for time series with seasonality and exogenous factors. But in these implementations, we can deal with simple seasonality with one time period and exogenous factors which are treated as covariates for residuals. Also, in these implementations, it is not easy to apply different types of regularizations to avoid overfitting or take into account feature interaction. The other main problem is that in some real cases of store sales we do not have an efficient number of historical time series values, e.g. in case when a store has been opened recently or has been acquired recently by a store chain. The lack of historical sales values can appear in case when a new product has been launched recently. In general, sales prediction can be a hard complex problem. One of effective approaches in the predictive analytics is combining linear, machine learning and probabilistic models into one multilevel predictive model [1,2].

II. SALES TIME SERIES DESCRIPTIVE ANALYSIS

We are going to consider several simple cases of approaches in sales times series forecasting. For our analysis, we used store sales historical data from Kaggle [3] competition "Rossmann Store Sales" [4]. These data represent the sales time series of Rossmann stores. For example, figure 1 shows sales time series in stores chain.
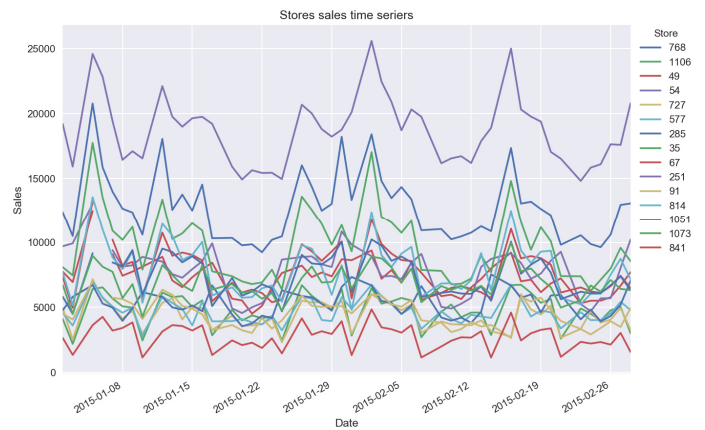

Figure 1. Examples of sales time series.

Before building of the predictive model, we can explore the data and conduct descriptive analytics. Figures 2-5 show different dependencies and distributions of sales time series.
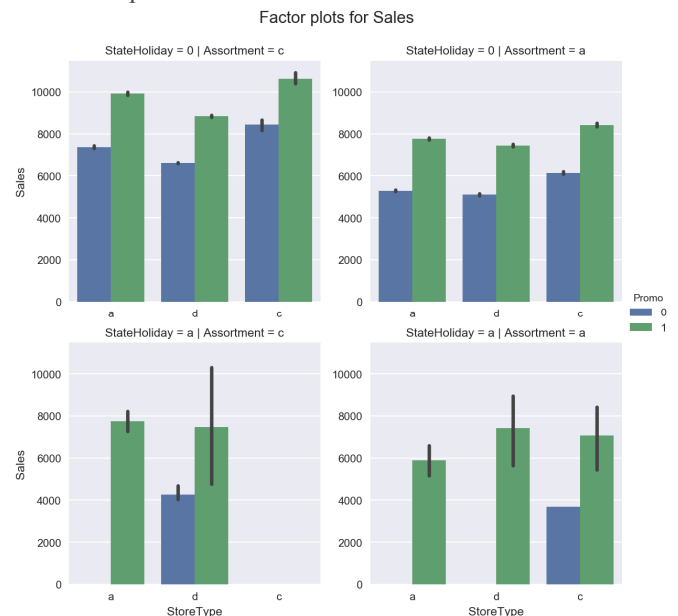

Figure 2. Sales vs different values of categoirical variables.

Figure 3. Probability density function of sales with different promo options.
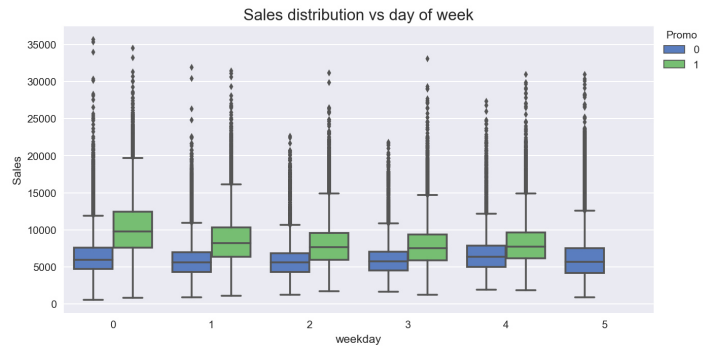

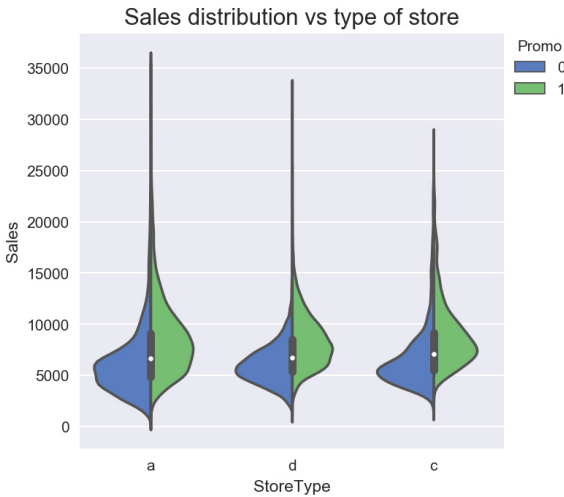Figure 5. Box plots for sales on different days of week.

Figure 4. Probability density function of sales with different store types and different promo options.

III. SALES TIME SERIES FORECASTING

To compare different forecasting approaches, we used two last months of historical data as the validation data for accuracy scoring using root mean squared error (RMSE). We consider sales in the natural logarithmic scale. Figure 6 shows forecasting results and RMSE scores for different approaches applied to the time series under consideration. For the comparison, we used the following methods: ARIMA using R package "forecast" [5], linear regression with LASSO regularization using R package "lars" [6], conditional inference trees with linear regression on the leaf using mob() function from R package "party" [7], gradient boosting XGBoost model using R package "xgboost" [8].
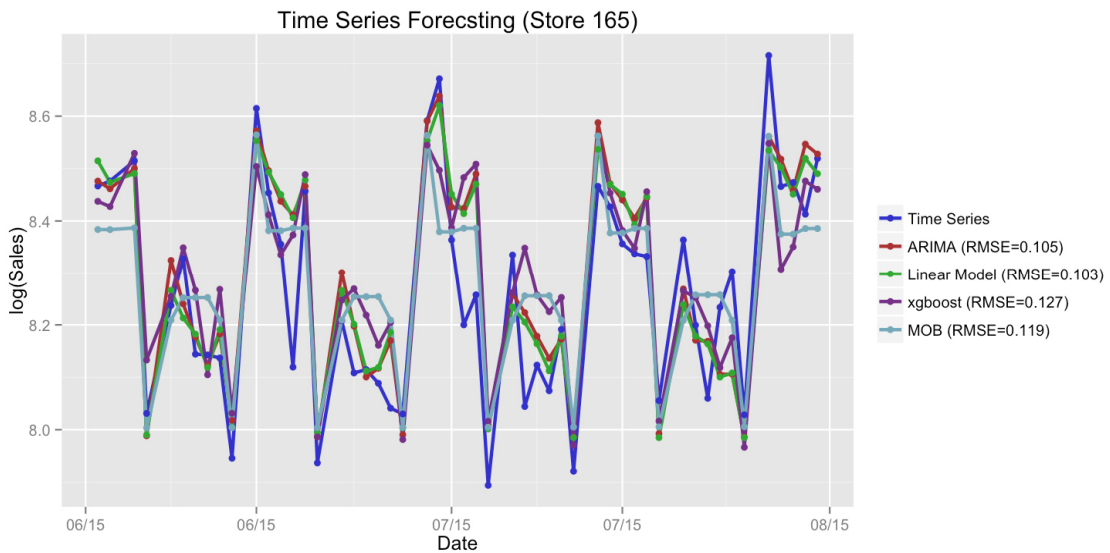

Figure 6. Sales time series forecasting with different methods.

Having different predictive models with different sets of features, it is useful to combine all these results into one. There are two main approaches for such a purpose – bagging and stacking. Bagging is a simple approach when we make weighted blending of different model predictions. Such models use different types of classifiers with different sets of features and meta parameters, then forecasting errors will have a weak correlation and they will compensate each other under the weighted blending. In the stacking approach, the results of predictions on the validation set are treated as input regressors for the next level models. As the next level model, one can use a linear model or another type of a classifier, e.g. Random Forest classifier or Neural Network. The target values of a validation set on this level are treated as a target variable for the classifier. It is important to mention that in case of time series prediction, we cannot use conventional cross validation approach, we have to split a historical data set into the training set and validation set by using time splitting, so the training data will lie in the first time period and the validation set – in the next one.

To get insights and to find new approaches, some companies propose such type of problems for data science competitions, e.g. at Kaggle [3]. The company Grupo Bimbo organized Kaggle competition Grupo Bimbo Inventory Demand [9]. In this competition, Grupo Bimbo invited Kagglers to develop a model to forecast accurately the inventory demand based on historical sales data. I had a pleasure to be a teammate of a great team "The Slippery Appraisals" which won this competition among nearly two thousand teams. We proposed the best scored solution for sales prediction in more than 800,000 stores for more than 1000 products. Our first place solution is described at [10] . The simple R script which based on the Xboost [8] classifier described is at [11]. To build our final multilevel model, we exploited AWS server with 128 cores and 2Tb RAM. For our solution, we used a multilevel model, which consists of three levels and has the following structure:
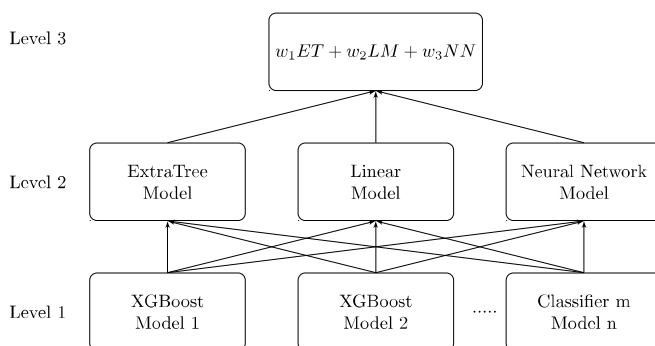
We built a lot of models on the 1st level. The training method of most 1st level models was XGBoost [8]. On the second level, we used the stacking approach when the results from the first level classifiers were treated as the features for the classifiers on the second level. For the second level, we used ExtraTrees classifier, the linear model from Python scikit-learn and Neural Networks. On the third level, we applied a weighted average to the second level results. The most important features are based on the lags of the target variable grouped by factors and their combinations, aggregated features (min, max, mean, sum) of target variable grouped by factors and their combinations, frequency features of factors variables.

IV. CONCLUSION

In our case studies, we showed different modern approaches for sales predictive analytics. The keystone is the historical data structure. Relying on it, one can select and construct new features, choose different technics and methods for the analysis. Creating ensembles of different models can give us better accuracy in comparison with single models.

REFERENCES

[1] B. M. Pavlyshenko. "Linear, machine learning and probabilistic approaches for time series analysis," in IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, pp. 377- 381, August 23-27, 2016.

[2] B. Pavlyshenko. "Machine learning, linear and bayesian models for logistic regression in failure detection problems.,' in IEEE International Conference on Big Data (Big Data), Washington D.C., USA, pp. 2046-2050, December 5-8, 2016. □

[3] Kaggle: Your Home for Data Science. URL: http://kaggle.com

[4] "Rossmann Store Sales", Kaggle.Com, URL: http://www.kaggle.com/c/rossmann-store-sales .

[5] Hyndman, R., & Khandakar, Y. (2008). "Automatic Time Series Forecasting: The forecast Package for R." Journal of Statistical Software, 27(3), pp.1 – 22, 2008

[6] Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. "Least angle regression." The Annals of statistics 32, no. 2,pp.407-499, 2004

[7] Zeileis, Achim, Torsten Hothorn, and Kurt Hornik. "Model-based recursive partitioning." Journal of Computational and Graphical Statistics 17.2, pp.492-514, 2008

[8] T. Chen and C. Guestrin. "Xgboost: A scalable tree boosting system." In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM. 2016, pp. 785-794. □

[9] Kaggle competition "Grupo Bimbo Inventory Demand " URL: https://www.kaggle.com/c/grupo-bimbo-inventory-demand □

[10] Kaggle competition "Grupo Bimbo Inventory Demand ". #1 Place Solution of The Slippery Appraisals team. URL:https://www.kaggle.com/c/grupo-bimbo-inventory-demand/discussion/23863

[11] Kaggle competition "Grupo Bimbo Inventory Demand". Bimbo XGBoost R script LB:0.457. URL: https://www.kaggle.com/bpavlyshenko/bimbo-xgboost-r-script-lb-0-457

Figure 7. Multilevel stacking model for sales prediction.