Peer-reviewed version available at Data 2019, 4, 15; doi:10.3390/data4010015

Article

Machine Learning Models for Sales Time Series Forecasting

Bohdan M. Pavlyshenko

SoftServe, Inc., Ivan Franko National University of Lviv
* Correspondence: bpavl@softserveinc.com, b.pavlyshenko@gmail.com

Version November 5, 2018 submitted to Preprints

Abstract: In this paper, we study the usage of machine learning models for sales time series forecasting. The effect of machine learning generalization has been considered. A stacking approach for building regression ensemble of single models has been studied. The results show that using stacking technics, we can improve the performance of predictive models for sales time series forecasting.

Keywords: machine learning; stacking; forecasting; regression; sales; time series

1. Introduction

Sales prediction is an important part of modern business intelligence. It can be a complex problem. Especially, in the case of lack of the data, missing data, a lot of outliers. Sales can be treated as a time series. At present time different time series theories and models have been developed. We can mention Holt-Winters model, ARIMA, SARIMA, SARIMAX , GARCH, etc. But their use in case of sales prediction is problematic due to several reasons. Here are several of them:

- We need to have historical data for a long time period to capture seasonality. But often we do not have historical data for a target variable, for example in case when a new product is launched. But we have sales time series for a similar product and we can expect that our new product will have a similar sales pattern.
- Sales can have complicated seasonality - intra-day, intra-week, intra-month, annual.
- Sales data can have a lot of outliers and missing data. We have to clean outliers and interpolate data before using a time series approach.
- We need to take into account a lot of exogenous factors which have impact on sales.

Sales prediction is rather a regression problem than a time series problem. Practice shows that the use of regression approaches can often give us better results comparing with time series methods. Machine learning algorithms make it possible to find patterns in the time series. In papers [1–3], we study different approaches for time series modeling such as using linear models, ARIMA algorithm, XGBoost machine learning algorithm.

One of the main assumptions of regression methods is that the patterns in the past data will be repeated in future. In the sales data, we can observe several types of patterns and effects. They are: trend, seasonality, autocorrelation, patterns caused by the impact of such external factors as promo, pricing, competitors behavior. We also observe noise in the sales. Noise is caused by the factors which are not included into our consideration. In the sales data, we can also observe extreme values - outliers. If we need to perform risk assessment, we need to take into account noise and extreme values. Outliers can be caused by some specific factors, e.g. promo events, price reduction, weather conditions, etc. If these specific events are repeated periodically, we can add a new feature which will indicate these special events and describe the extreme values of the target variable.

In this study, we consider the usage of machine learning models for sales time series forecasting.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 November 2018 doi:10.20944/preprints201811.0096.v1

Peer-reviewed version available at Data 2019, 4, 15; doi:10.3390/data4010015

2. Machine Learning Predictive Models

For our analysis, we used store sales historical data from "Rossmann Store Sales" Kaggle competition [4]. These data represent the sales time series of Rossmann stores. Calculations were conducted in the Python environment using the main packages *pandas, sklearn, numpy, keras, matplotlib, seaborn*. To conduct analysis, *Jupyter Notebook* was used.

Figure 1 shows typical time series for sales. On the first step we conducted the descriptive

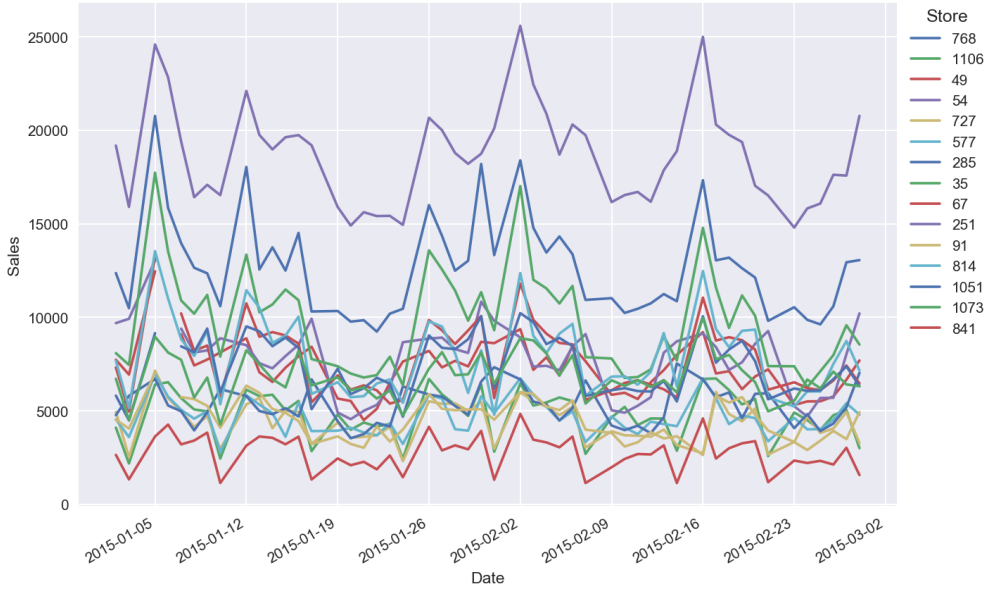

Figure 1. Typical time series for sales

analytics, which is a study of sales distributions, data visualization with different pairplots. It is helpful in finding correlations and sales drivers on which we can concentrate our attention. The figures 2, 3, 4 show the results of the descriptive analytics.

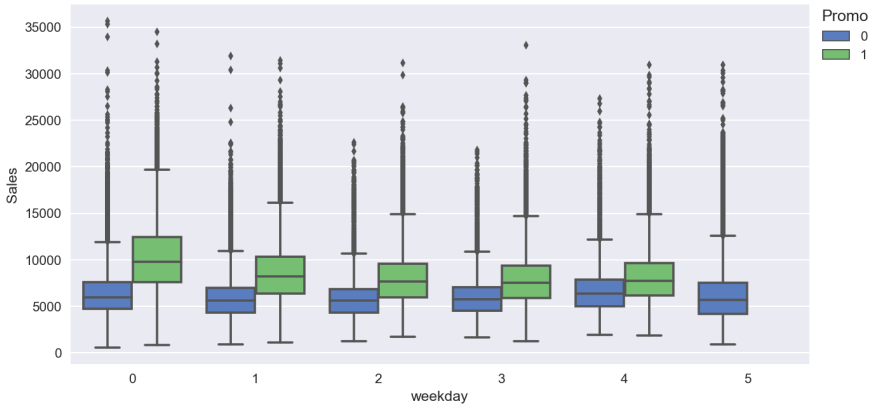

Figure 2. Boxplots for sales distribution vs day of week

Supervised machine learning methods are very popular. Using these methods, we can find complicated patterns in the sales dynamics. Some of the most popular are tree based classifiers, e.g. Random Forest and XGBoost, etc. A specific feature of tree based methods is that they are not sensitive to monotonic transformations of the features. It is very important when we have a set of features with different nature. For example, one feature represents promo, another - competitors' prices, macroeconomic trends, customers' behavior. A specific feature of most machine learning methods is that they can work with stationary data only. We cannot apply machine learning to non-stationary

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 November 2018 doi:10.20944/preprints201811.0096.v1

Peer-reviewed version available at Data 2019, 4, 15; doi:10.3390/data4010015

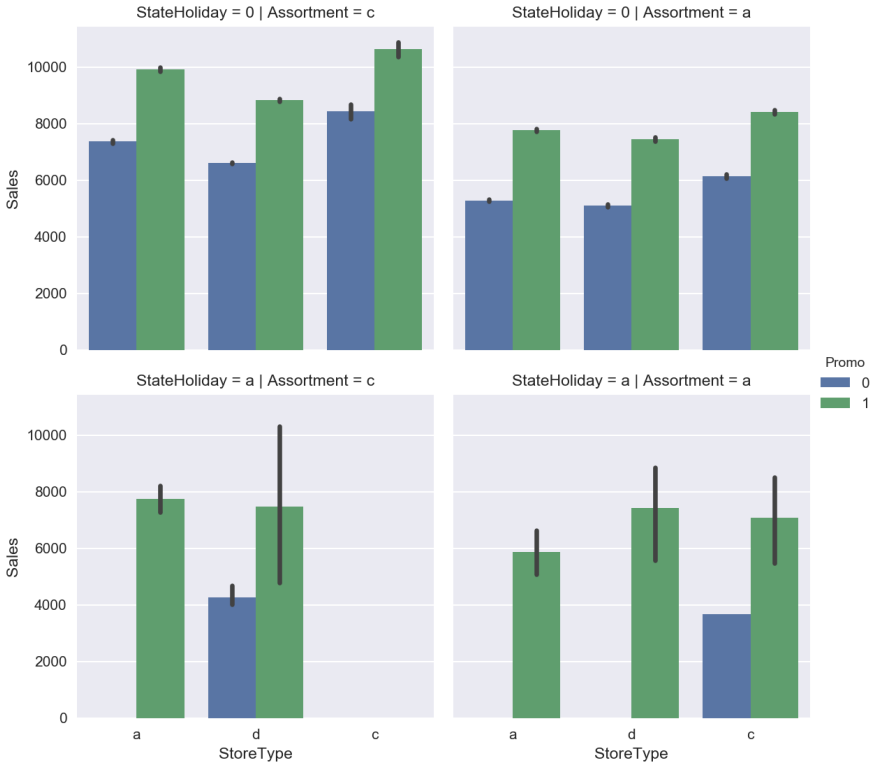

Figure 3. Factor plots for aggregated sales

sales with a trend. We have to do detrending of time series before applying machine learning. In case of small trend, we can find bias using linear regression on the validation set. Let us consider the supervised machine learning approach using sales historical time series. For the case study, we used Random Forest algorithm. As covariates, we used categorical features: promo, day of week, day of month, month. For categorical features, we applied one-hot encoding, when one categorical variable was replaced by n binary variables, where n is the amount of unique values of categorical variables. Figure 5 shows the forecasting for sales time series using of Random Forest algorithm. Figure 6 shows the feature importance. For error estimation, we used a relative mean absolute error (MAE) which is calculated as $error = MAE/mean(Sales) \cdot 100\%$. Figure 7 shows forecast residuals for sales time series, Figure 8 shows the rolling mean of residuals, Figure 9 shows the standard deviation of forecast residuals.

In the forecast, we may observe bias on validation set which is a constant (stable) under- or over-valuation of sales when the forecast is going to be higher or lower with respect to real values. It often appears when we apply ML methods to non-stationary sales. We can conduct the correction of bias using linear regression on the validation set. We have to differentiate the accuracy on a validation set from the accuracy on a training set. On the training set, it can be very high but on the validation set it is low. The accuracy on the validation set is an important indicator for choosing an optimal number of iterations of ML algorithm.

3. Effect of Machine Learning Generalization

The effect of machine learning generalization consists in the fact that a classifier captures the patterns which exist in the whole set of stores or products. If the sales have expressed patterns, then generalization enables us to get more precise results which are resistant to sales noise. It also gives us an ability to make prediction in case of very small number of historical sales data, which is important when we launch a new product or store. In the case study of machine learning generalization, we used the following additional features with regard to the previous case study: mean sales value for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 November 2018 doi:10.20944/preprints201811.0096.v1

Peer-reviewed version available at Data 2019, 4, 15; doi:10.3390/data4010015

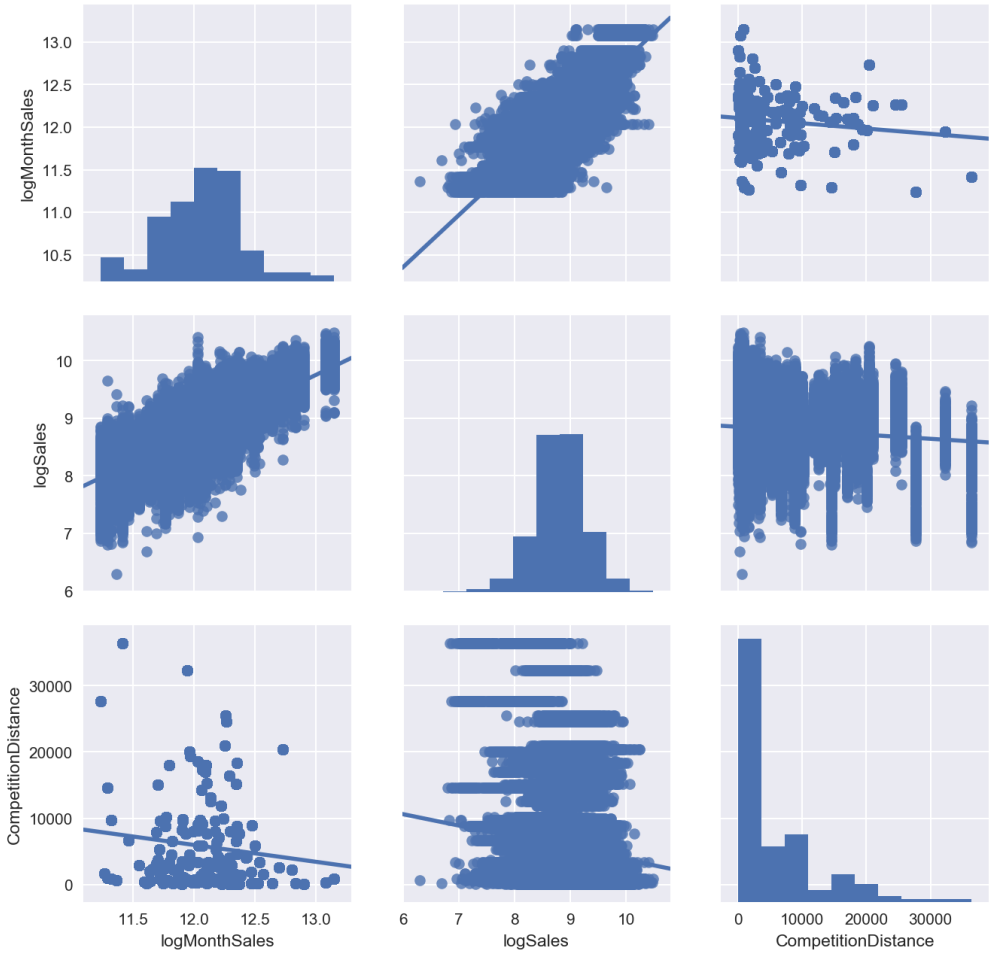

Figure 4. Pair plots with log(MonthSales), log(Sales), CompetitionDistance

for specified time period of historical data, state and school holiday flags, distance from store to competitor's store, store assortment type. Figure 10 shows the forecast in the case of historical data with a long time period (2 years) for a specific store, Figure 11 shows the forecast in the case of historical data with a short time period (3 days) for the same specific store. In case of short time period, we can receive even more precise results. If we are going to predict the sales for new products, we can make expert correction by multiplying the prediction by a time dependent coefficient to take into account the transient processes, e.g. the process of product cannibalization when new products substitute other products.

4. Stacking of Machine Learning Models

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. Forecasting errors of these models will be of weak correlation and these errors will be compensated by each other under the weighted blending. The less is the error correlation of model results, the more precise forecasting result we will receive. Let us consider the stacking technic [5] for building ensemble of predictive models. In such an approach, the results of predictions on the validation set are treated as input regressors for the next level models. As the next level model, we can consider a linear model or another type of a classifier, e.g. Random Forest classifier or Neural Network. It is important to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 November 2018 doi:10.20944/preprints201811.0096.v1

Peer-reviewed version available at Data 2019, 4, 15; doi:10.3390/data4010015

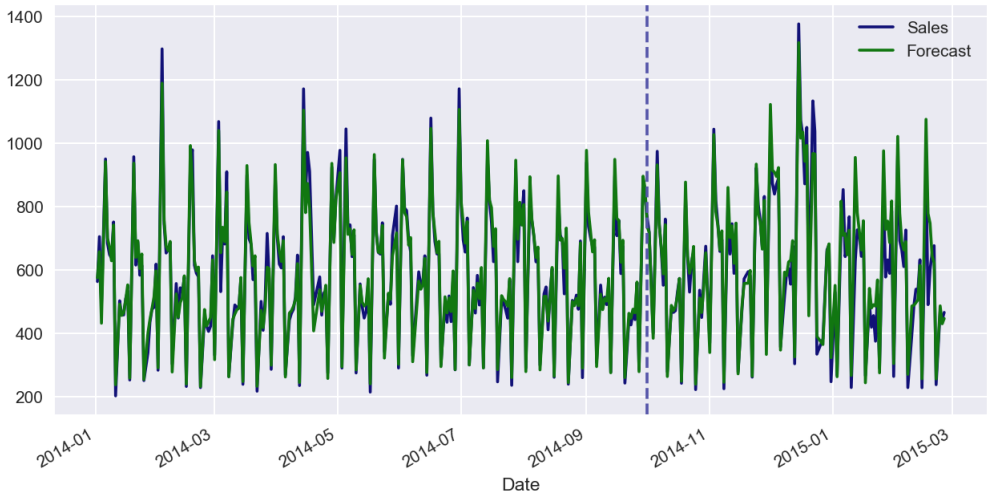

Figure 5. Sales forecasting (train set error:3.9%, validation set error: 11.6%)

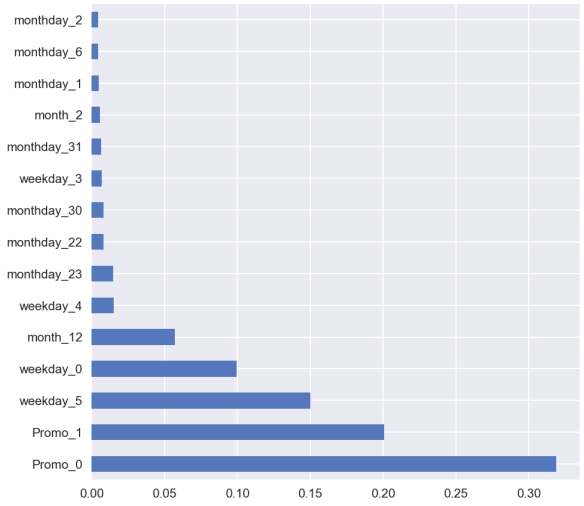

Figure 6. Features importance

mention that in case of time series prediction, we cannot use a conventional cross validation approach, we have to split a historical data set on 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. Figure 12 shows the time series forecasting on the validation sets obtained using different models. Vertical dotted line on the Figure 12 separates validation set and out-of-sample set which is not used in the model training and validation processes. On the out-of-sample set, one can calculate stacking errors. Predictions on the validation sets are treated as regressors for the linear model with Lasso regularization. Figure 13 shows the results obtained on the second-level Lasso regression model. Only three models from the first level (ExtraTree, Lasso, Neural Network) have non zero coefficients for their results. For other cases of sales datasets, the results can be different when the other models can play more essential role in the forecasting. Table 1 shows the errors on the validation and out-of-sample sets. These results show that stacking approach can improve accuracy on the validation and on the out-of-sample sets.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 November 2018 doi:10.20944/preprints201811.0096.v1

Peer-reviewed version available at Data 2019, 4, 15; doi:10.3390/data4010015

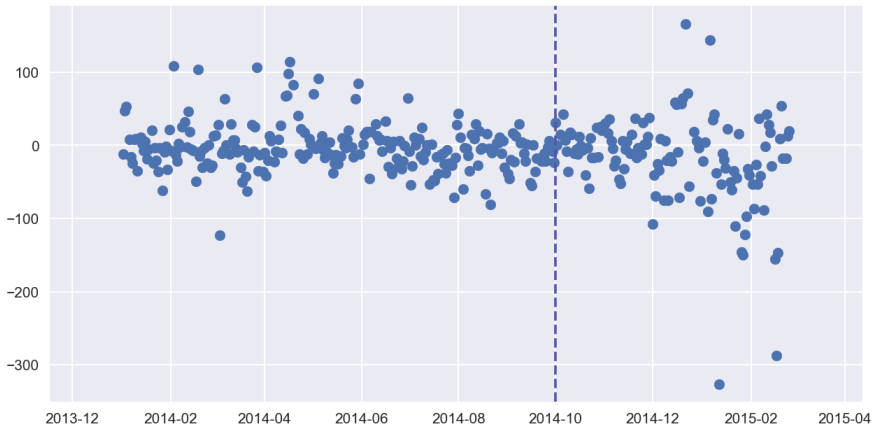

Figure 7. Forecast residuals for sales time series

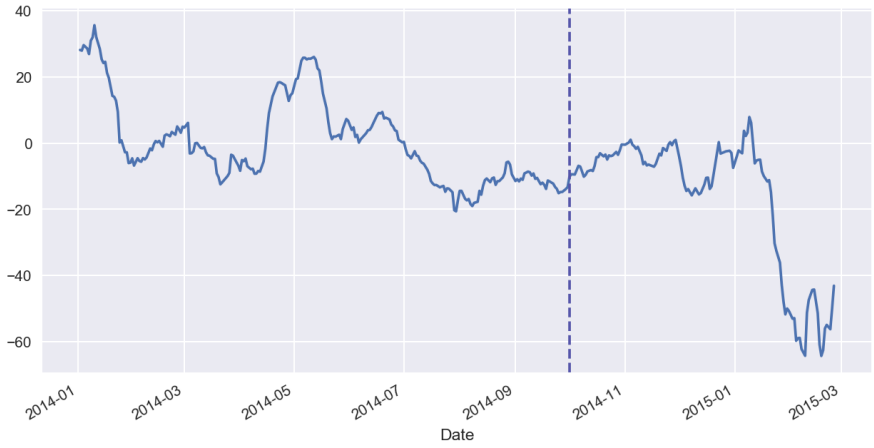

Figure 8. Rolling mean of residuals

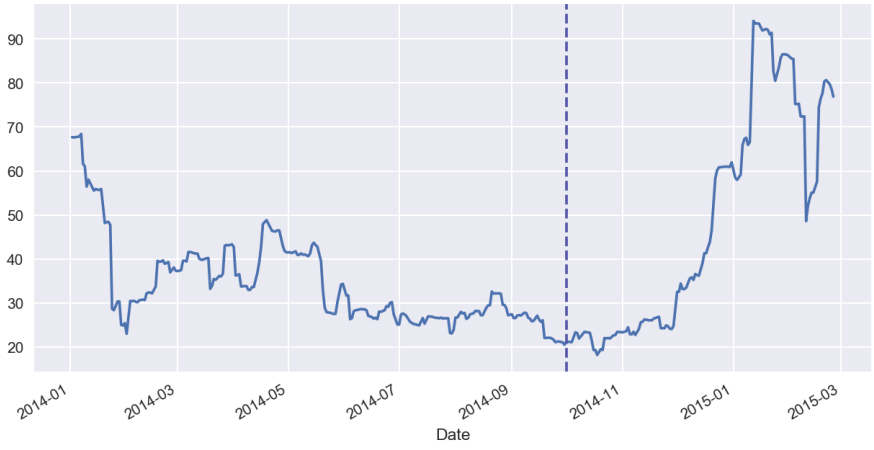

Figure 9. Standard deviation of forecast residuals

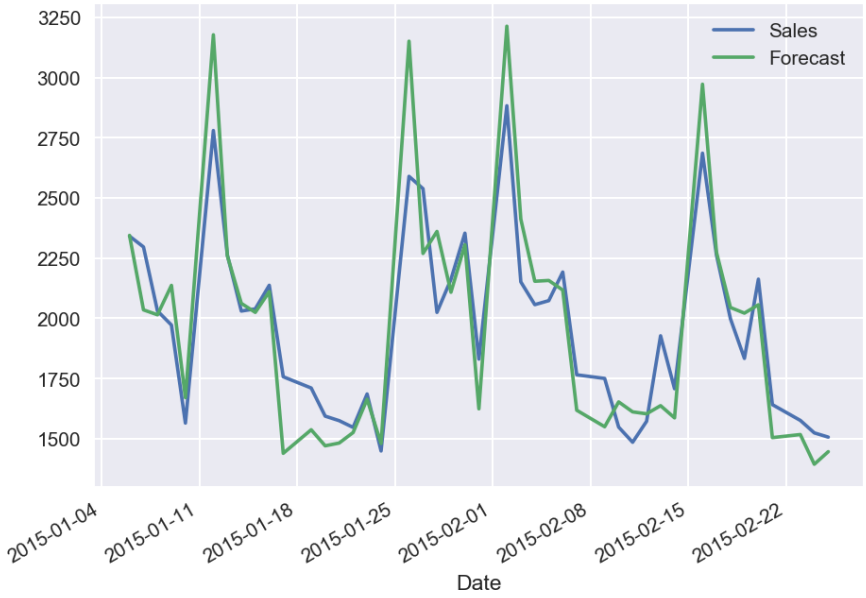

Figure 10. Sales forecasting with long time (2 year) historical data, error=7.1%

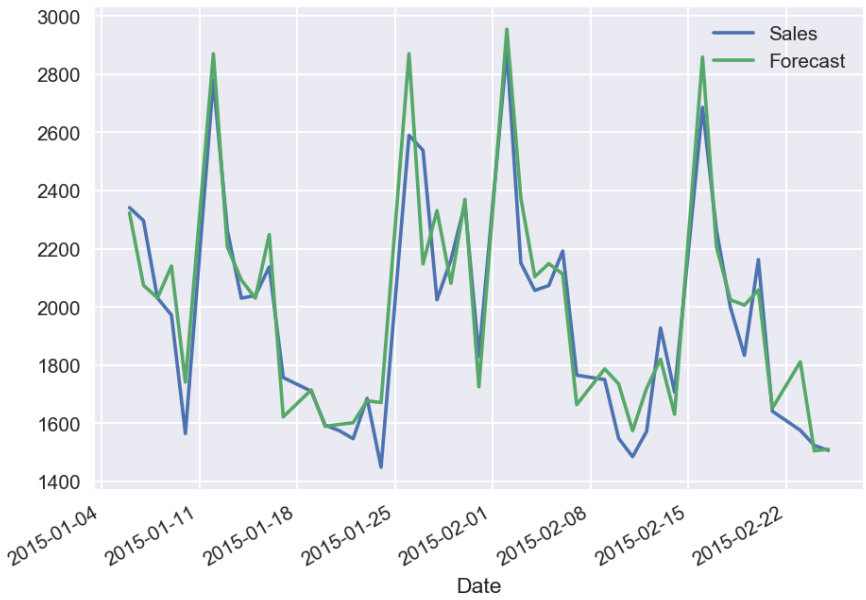

Figure 11. Sales forecasting with short time (3 days), historical data, error=5.3%

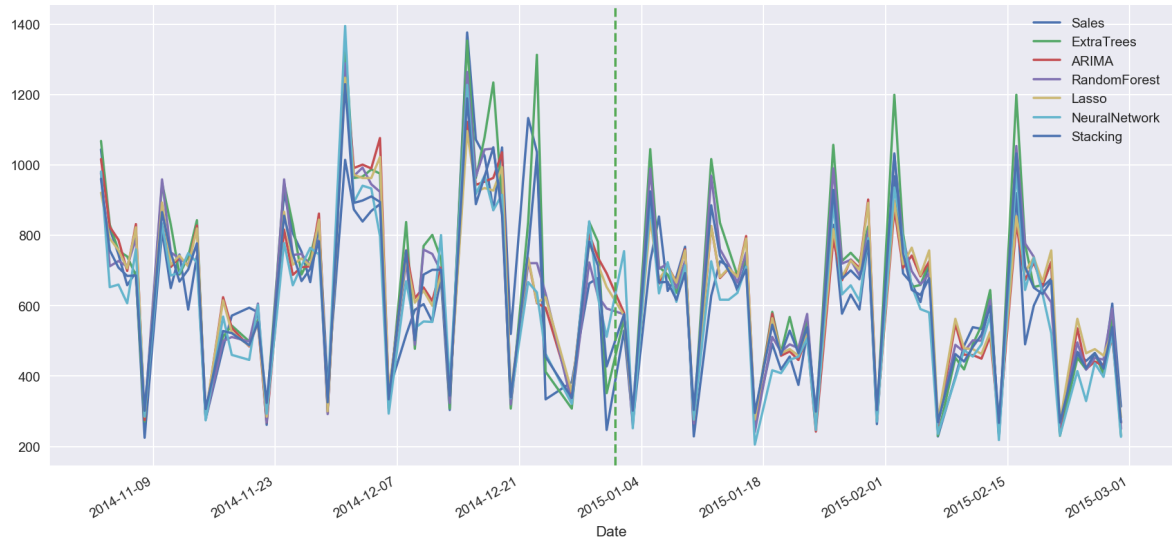

Figure 12. Time series forecasting on the validation sets obtained using different models

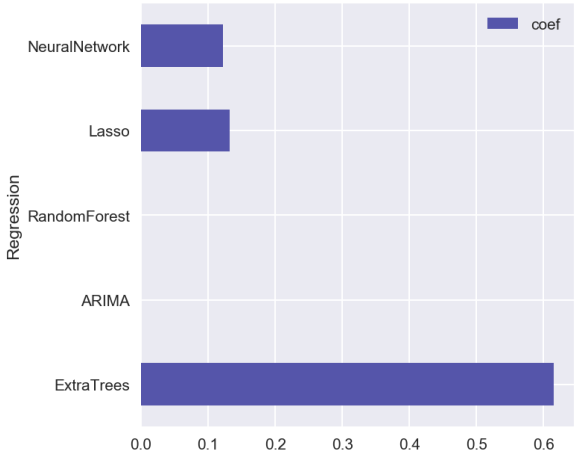

Figure 13. Stacking weights for regressors

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 November 2018 doi:10.20944/preprints201811.0096.v1

Peer-reviewed version available at Data 2019, 4, 15; doi:10.3390/data4010015

Table 1. Forecasting errors of different models

Model	Validation error	Out-of-sample error
ExtraTree	14.6%	13.9%
ARIMA	13.8%	11.4%
RandomForest	13.6%	11.9%
Lasso	13.4%	11.5%
Neural Network	13.6%	11.3%
Stacking	12.6%	10.2%

To get insights and to find new approaches, some companies propose their analytical problems for data science competitions, e.g. at Kaggle [6]. The company Grupo Bimbo organized a Kaggle competition Grupo Bimbo Inventory Demand [7]. 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 at [8]. To build our final

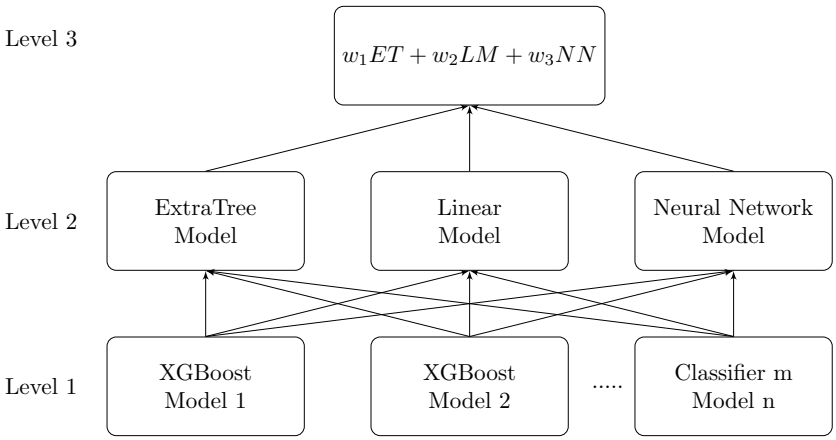

Figure 14. Mulitlievel machine learning model for sales time series forecasting

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 (Figure 14). We built a lot of models on the 1st level. The training method of most 1st level models was XGBoost. On the second level, we used a 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 the target variable grouped by factors and their combinations, frequency features of factors variables. One of the main ideas in our approach is that it is important to know what were the previous week sales. If during the previous week too many products were supplied and they were not sold, next week this product amount, supplied to the same store, will be decreased. So, it is very important to include lagged values of the target variable as a feature to predict next sales. More details about our team's winer solution are at [8]. The simplified version of the R script is at [9].

5. Conclusion

In our case study, we considered different machine learning approaches for time series forecasting. The accuracy on the validation set is an important indicator for choosing an optimal number of iterations of ML algorithm. The effect of machine learning generalization consists in the fact that a

classifier captures the patterns which exist in the whole set of stores or products. It can be used for sales forecasting when there is a small number of historical data for specific sales time series in the case when a new product or store is launched. Using stacking model on the second level with the covariates that are predicted by machine learning models on the first level, makes it possible to take into account the differences in the results for machine learning models received for different sets of parameters and subsets of samples. For stacking machine learning models the Lasso regression can be used. Using multilevel stacking models, one can receive more precise results in comparison with single models.

References

1. Pavlyshenko, B. M. 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. Pavlyshenko, B. 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. Pavlyshenko, B. Using Stacking Approaches for Machine Learning Models. In 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP) , Lviv, Ukraine, pp. 255-258, August 21-25, 2018.
4. 'Rossmann Store Sales', Kaggle.Com, URL: http://www.kaggle.com/c/rossmann-store-sales .
5. Wolpert, D. H. Stacked generalization. Neural networks, 5(2), pp. 241-259,1992.
6. Kaggle: Your Home for Data Science. URL: http://kaggle.com
7. Kaggle competition 'Grupo Bimbo Inventory Demand ' URL: https://www.kaggle.com/c/grupo-bimbo-inventory-demand
8. 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
9. 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