**Summary**

**Title: Machine-Learning Models for Sales Time**

**Series Forecasting**

**Author: Bohdan M. Pavlyshenko**

Link to research paper: <https://www.researchgate.net/publication/330484523_Machine-Learning_Models_for_Sales_Time_Series_Forecasting>

Brief summary of paper 2 (include the title, source and year the paper was published); (i) what does it relate to? (problem, data or technique you plan to use) (ii) any assumptions in the paper? (iii) what are the main claims? (iv) what is the takeaway from this paper for you?

Abstract:

This paper deals with the usage of machine learning models for sales predictive analysis. Also, the effect of machine-learning generalization is demonstrated. Lastly, a stacking approach is demonstrated for building a regression ensemble of many models. Stacking is shown to improve performance.

Introduction:

Sales prediction is important for modern business intelligence. This paper deals with providing a solution to “Rossman Stores” dataset on Kaggle. It is similar to the dataset we have chosen, and the approaches shown in this paper can be taken as an inspiration to arrive at predictions for our dataset.

Various machine learning approaches are shown and critical points are made. Some limitations of time series approach(Holt-Winters, ARIMA, SARIMA, SARIMAX, GARCH) for sales forecasting are listed. To summarize them,

* Time Series analysis required long period of historic data to capture seasonality. However, when new products are launched lots of historical data may not be available.
* Sales data can have a lot of outliers. Must clean outliers and interpolate data.
* Lot of exogenous factors need to be taken into account.

The paper says that sales prediction is rather a regression problem rather than a time series problem. “Practice shows that use of regression approaches can often give us better results compared to time series methods”

Complicated patterns can be found using supervised ML Models. Some of the most popular models are tree-based ML Models like Random Forest and Gradient Boosting Machine.

Sales data can have several types of patterns and effects, namely trend, seasonality, autocorrelation. Noise is also observed in sales. Risk assessment can be performed by taking into account noise and extreme values.

Assumptions:

One of the main assumptions of regression methods is that the patterns in the past data will be repeated in the future.

Claims:

First descriptive analytics is conducted. Descriptive analysis helps finding correlations and sales drivers on which there is focus.

For the case study, Random Forest Algorithm is used. Train set error of 3.9% and validation set error of 11.6% is obtained. Feature importance, residuals are also plotted.

Sometimes bias may be observed on the validation set. Correction of bias can be done using linear regression on the validation set.

The effect of machine learning generalization is discussed. If the sales have patterns then generalization helps in getting more precise results which are resistant to noise. 2 cases are demonstrated for a specific store, forecast in case of historical data with a long time period(2 years) and also forecast in the case of historical data with a short time period(3 days). It is shown that in case of a short time period even more precise results are obtained.

Lastly stacking of machine learning models is discussed. Results of different predictive models with different sets of features are combined into 1. Predictions on the validation sets are treated as regressors for the linear model with Lasso regularization. It is shown that stacking approach can improve accuracy on the validation and on the out-of-sample sets.

Main Takeaways:

Sales prediction is more of a regression problem rather than a time series problem. Regression methods can give us better results. The effect of machine learning generalization consists in the fact of capturing the patterns in the whole dataset. This is helpful when a small amount of historical data is available.

Stacking approach can be used to improve accuracy for out of sample cases. The results of multiple model predictions on the validation set are passed as input to next level models. Lasso Regression can be used as the next level model. Stacking helps to take into account the difference in the results in multiple models.