An empirical comparison of machine learning and time-series models for sales forecasting in the automotive sector.

Manju Kiran[1], Pauline Sherly Jeba[1], Amit Kumar Sharma[1], Divakar Venkatesh[2]

[1]*Engineering Data Science Department, RBEI, Robert Bosch GmbH.*
[2]*RBEI, Robert Bosch GmbH.*

---***---

Abstract – The business unit approached us to help build a tool to automate monthly sales forecast activity for a series of automobile parts. Initiallythe automation was performed manually using excel sheets involving a tedious process of aggregating the data from various sources and systems onto a single excel file, followed by the time-consuming activity of normalizing and preparing data for forecasting using excel formulas. We proposed an empirical comparison of machine learning and time-series models for a vast combination of stock-keeping units,forecasting sales in the automotive sector.

Key Words: Automotive, Machine learning, Time-series, Sales forecasting, ensemble.

Introduction

The tool developed can forecast sales from any time series data with reliable accuracy (Cooper et al. 1999). A sales forecast can be made monthly, yearly, weekly, or daily depending on the time interval provided on the input data (Trapero et al. 2013). For our approach, we selected monthly sales data (Chang, Liu, and Fan 2009). The dataset for the model includes baseline and external data. The Historical sales data is considered as Baseline data which contains sales information for about three years. The external data includes factors that can impact sales such as Price change, Demand of the product, its Secondary sales, GDP, etc. (Joseph, Larrain, and Turner 2014),

As mentioned earlier, the tool reduces the time consumption in data preparation using excel formulas. The input data undergoes two basic data preparation steps – Missing Date/Value imputation and Discarding outliers in the data (Guo, Wong, and Li 2013).

Furthermore, the data is clustered based on its characteristics like Seasonality, Trend, and Sparsity. The data is checked for stationarity, and lag is determined (P, D, Q parameters) (Stover and Ulm 2013). The automation process helps in identifying Seasonality and Trend in the historical database. Based on the characteristics, the dataset is grouped under different clusters. This segregation helps in reducing computational time for model selection.

Data to be trained is further split into Train and validation set on a ratio of 7:3. A total of about 21 Models were used on the tool. It includes 8 Time series and 13 Machine learning models. From the bag of 21 models, the models to be run are selected based on the data profiling/data cluster. The Hyperparameters of the models are auto-tuned using Grid search and random search techniques (Stover and Ulm 2013).

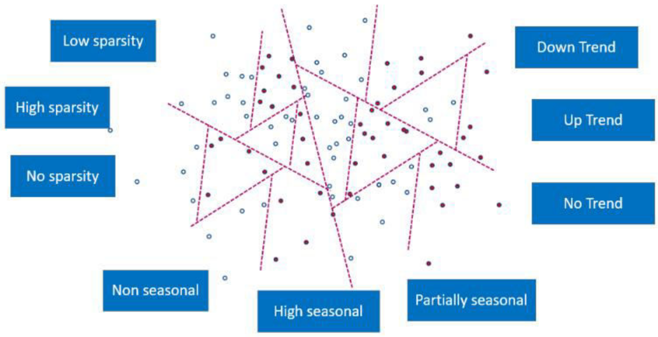

Figure 1: Sales characteristics.

From the TS and ML model run for the Validation set data, the best three models are selected based on the minimum RMSE calculated. The best three models from the Time series and Machine learning are identified. The model with the least RMSE from the Time series (TS) and Machine Learning (ML) is used for the ensemble approach. The errors calculated from selected models are used for Weight calculation to perform an ensemble approach. These estimated weights are multiplied with respective machine learning and time series forecast results, and by ensembling, the final forecasts for the validation set are calculated.

A comparison is made on the results obtained from the Time series, Machine Learning and Ensemble approach. The method with minimum Root Mean Square Error (RMSE) is opted to Forecast the data for the future period. To understand the performance of the model, the performance metrics such as FACC (Forecast Accuracy) and MAPE (Mean absolute percentage error) (Sa-Ngasoongsong et al. 2012). Around 60% of the data showed higher FACC and lowered MAPE with Machine Learning Models, and 30% of the data had better FACC with the ensemble approach.

Recommendation

During the project's complete experiment, it was found that the machine learning models outperform time series and traditional forecasting models provided we have good quality historical data (2-years and above) with minimum no missing values.

References

Chang, Pei Chann, Chen Hao Liu, and Chin Yuan Fan. 2009. "Data Clustering and Fuzzy Neural Network for Sales Forecasting: A Case Study in Printed Circuit Board Industry." *Knowledge-Based Systems* 22(5): 344–55.

Cooper, Lee G. et al. 1999. "PromoCast^{TM}: A New Forecasting Method for Promotion Planning." *Marketing Science* 18(3): 301–16.

Guo, Z. X., W. K. Wong, and Min Li. 2013. "A Multivariate Intelligent Decision-Making Model for Retail Sales Forecasting." *Decision Support Systems* 55(1): 247–55.

Joseph, Anthony, Maurice Larrain, and Claude Turner. 2014. "The Treasury Bill Rate, the Great Recession, and Neural Networks Estimates of Real Business Sales." *Procedia Computer Science* 36(C): 227–33.

Sa-Ngasoongsong, Akkarapol et al. 2012. "Multi-Step Sales Forecasting in Automotive Industry Based on Structural Relationship Identification." *International Journal of Production Economics* 140(2): 875–87.

Stover, Jason H., and Matthew C. Ulm. 2013. "Hyperparameter Estimation and Plug-in Kernel Density Estimates for Maximum a Posteriori Land-Cover Classification with Multiband Satellite Data." *Computational Statistics and Data Analysis* 57(1): 82–94. http://dx.doi.org/10.1016/j.csda.2012.06.010.

Trapero, Juan R., Diego J. Pedregal, R. Fildes, and N. Kourentzes. 2013. "Analysis of Judgmental Adjustments in the Presence of Promotions." *International Journal of Forecasting* 29(2): 234–43.