**Prediction Using Spark MLlib Comparison of prediction time intervals**

Brian Rikshpun

Guy Cohen

Ester Vaknin

Almog Klein

Afeka College of Engineering

29/03/2022

**1    Abstract**

Sales prediction plays a key role in building up a business. It is one of the most important parts of business intelligence. Sales prediction and forecasting give an insight into how a company should manage its workforce, cash flow, and its resources. In simple terms, sales regression analysis is used to understand how certain factors in your sales process affect sales performance and predict how sales would change over time if you continued the same strategy or pivoted to different methods. It is an evaluation tool that uses past and current sales data to predict future performance. Estimating future sales is an important part of the financial planning of any business. Furthermore, with the help of data and gained insights, it becomes easier to understand consumer behavior. Demand and supply actions can be planned by looking at the forecasts. We used Spark and Kafka to maintain high connectivity and processing capacity between the stores and the main processing PC.

This paper consists of a literature review, information regarding the dataset, methodology where it tells how the data has been preprocessed and gives an insight into how feature selection has been done. The results are compared to conclude which regression analysis model is better for sales prediction on the Rossman dataset and preserve the training of this model.

**2    Introduction**

A regression model is used to forecast or predict the value of the dependent variable, based on various independent variables. Where the independent variable in our case contains information about the stores such as promo, customer, holiday, etc., and the dependent variable is sales.

Many types of classifiers can be used to predict sales such as Linear, Logistic, K-nearest neighbor, Support vector, Random Forest, Bayesian classification, Decision Trees and Gradient Boosted Trees regressions [1]. In this work we tested four different type of regression algorithms, Linear, Random Forest, Decision Trees, and Gradient Boosted Trees regressions.

Linear Regression (LR) - Is a linear modeling approach to find the relationship between 1 or more independent variables denoted as X and dependent variable denoted at Y. LR finding the best fitted linear line for the training data as well as test data. The best fitted line can be found by minimizing the distance between all data points and its distance to the regression line, we can find minimize distance.

Random Forest regression (RF) - Essentially a multitude of decision trees. The output obtained from a random forest model is a combination of the outputs obtained from all the decision trees. Nevertheless, it can measure multiple trees with the same datasets and calculate the value of the forecast of every single tree. Can be applied for misuse, anomaly, and hybrid network-based intrusion detection systems. Uses a hybrid detection technique where they employ misuse detection followed by anomaly detection.

Decision Trees regression (DM) - This is a machine learning technique for regression and classification problems. DM is a classifier referred to as a recursive partition of the instant space. It is a powerful form of multiple variable analysis and is a strong data mining tool. Let the objective be denoted as (O) and (Ci) is the ways to follow and let (Mij) the means of action corresponding to these ways, which can be noted by qi, (i= 1 …. n), which meets the relation.

For the means of action (Mij), the important coefficient (aij), includes set of weights, where the sum is equal to 1 for each way

a11+a12+ …. +a1m = 1,

a21+a22+ …. +a2m = 1, 

…..…..,

an1+an2+ …. +anm= 1.

Gradient Boosted Trees regression (GBT) - This is a machine learning technique for regression and classification problems. This approach could ensemble learning method that combines many decision trees to produce a final prediction model. This model is built on a principle that a collection of weak learners combined can produce a strong learner by using boosting process. GBT approach has a strong additive training method, required for adding a new weak learner into the model, the weak learner is the decision tree. Let F(x) is a full model after t-1 round and h(x) is the new tree, added to the model.

F0= 0

Ft (x) = Ft-1(x) +h(x)

Each new function is an attempt to correct the errors of the model built in previous rounds. Hence the new function(x) must be able to predict the residual Ft-1(x). The Figure 8 shows a sample of gradient boosted tree.

Apache Spark is a framework that being increasingly used in distributed data processing. [2]

ML and MLlib machine learning libraries are provided in Spark to make distributed machine learning extensible and easy. At a higher level, Spark provides common ML learning algorithms for classification, regression, clustering, etc.

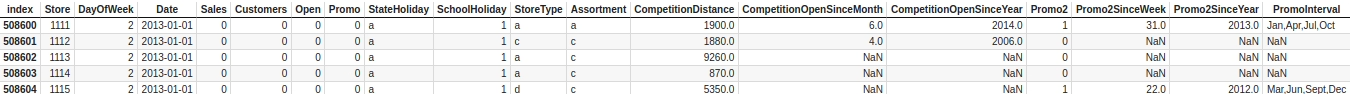
As we have seen in other studies [3], using spark in regression problems has salient features that during our research we visualized in our code. Spark can work with RDD and make calculations much faster if you have the right infrastructure. Spark can handle big data and allows streaming in a very “simple” way. In addition, can integrate with Kafka and use the zookeeper’s pros. the main drawback of spark is limited number of regression models implementation (vs scikit-learn).

**3    Data description**

In our work we used datasets of'ROSSMANN' stores sales. 'ROSSMANN' operates over 3,000 drug stores in 7 European countries.

The datasets we have is extensive and numbers about 1 million records of stores sales. The sales records are of 1,115 stores. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality.

Due to processing power limitations and lack of resources we are required to significantly reduce the size of the training data. First, we split the whole data set into 50/50 train/test. Then we perform the same split on one half of the data, train it, and took a snip of it (100).

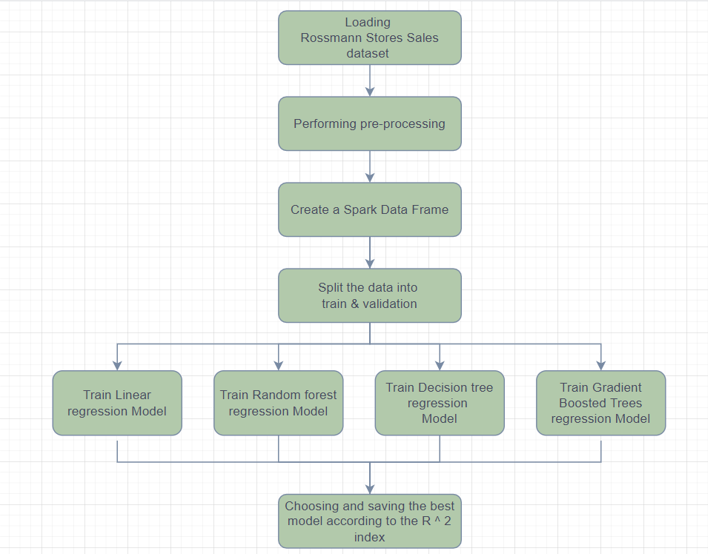


*Figure 1 - 'ROSSMANN' dataset*

**4    Methodology of work**

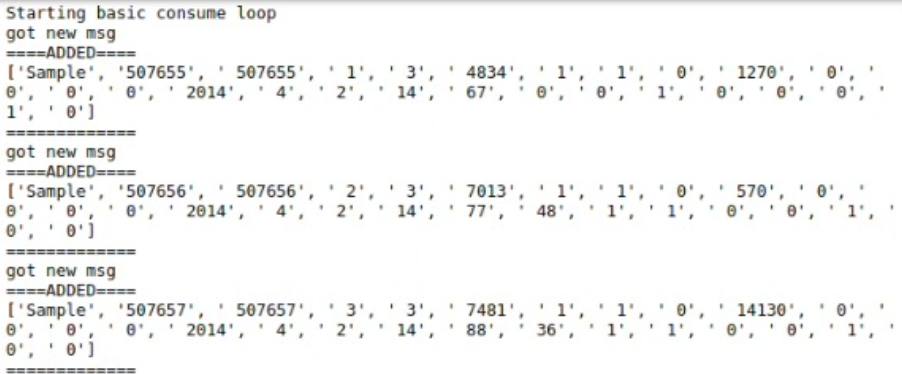
The purpose of this work is to test and produce a regression model (pyspark.ml) which will provide a solution for predicting the value of the daily sales of stores and retrain the model hyper parameters by “streaming” new data using Kafka.

Our methodology includes two main stages. The first stage performs the initial data pre-process, creating a Spark dataframe, training and comparing between the four regression models mentioned above. On the basis of a performance evaluation, a best suited predictive model is suggested for the sales trend forecast. The preferred model is selected by the R^2 index.



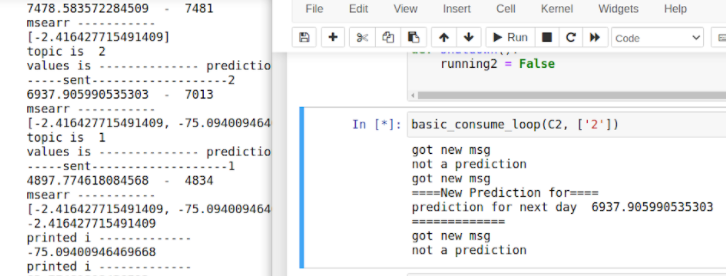
*Figure 2 - methodology, stage 1*

In the second part, we created consumer and producer channels, used by Kafka Library, for the main PC and each one of the stores PC. Transfer the selected model to the main PC and create sales predict for the next day. The transfer is made through the main PC's producer channel p1. Stores receive the prediction through their consumer channel c2.

****

*Figure 3 - The main computer receives the new lines of each store through its consumer, C1.*

For these operations the main PC arranges the predictions of the stores in an array of topics. The topics allow to transfer predictions to stores from the PC and transfer sales to the PC from the stores.



*Figure 4 - The store received the sales forecast for the next day from the main computer through its consumer. channel*

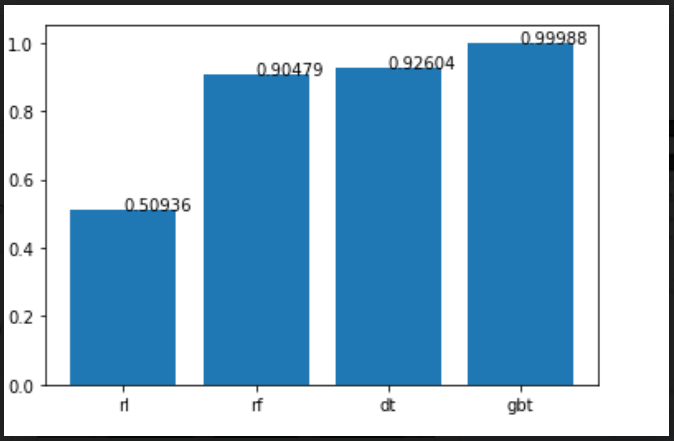
At the end of each day, the results of the daily sales from each store, are record and sent to the main computer in favor of a continuous update of the data. The model re-train and send a new predict for the next period. Handling files in big data scale, distribute them to some users in real time, allowing each user to process, visualize and predict with the data [4].

Diagram

Description automatically generated

*Figure 5 - methodology, stage 2*

**5    Results**

In our study we chose to compare between a few ML models as we mentioned in the first section. The metric we used evaluate the models performance is R squared. The best model was GBT as seen in fig 6.

*Figure 6 – R^2 results of the selected models, stage 1*

A picture containing chart

Description automatically generatedChart

Description automatically generatedFor continues evaluation of the GBT model, we used the matrices MAE and MSE over time.

*Figure 7 - results over time, MAE*

*Figure 8 - results over time, MSE*

As we can see on figure 7 the results are hard to interpret because of the low train volume data for initial training. The low MAE scores in figure 6 can be explained by the fact that our test data has observations with 0 in the predicted value – Sales.

In figure 8 it can be seen that the MSE score of the model is converging over time to the point that at observations 40 and after have MSE smaller then 50,000. These results are great because every sample has 3 observations, in MSE formula the scale of the predicted value is important. Our predicted value scale is between 2,000 to 10,000 so every error of 100 in MAE (~1% of the value) becomes a 10,000 error in MSE.

The results may indicate of an overfitting model based on low volume (100 samples).

**6    Conclusions**

Spark is a very comfortable tool to use ML with wide infrastructure with lots of computers and end stations, although it requires a dedicated server and multiple computers to have the benefit of parallel calculation. Spark MLlib includes most of the needs of a data scientist but it less comprehensive in comparison to scikit-learn (python).

Our research has provided a system that given store sales data from the past it can provide a model that will predict the sales of the stores for the next day while always fine tuning its hyper parameters. All those run on Kafka infrastructure, which provides the user a comfortable way to send the messages he needs (data, prediction) from the stores to the mainframe and the opposite.

It will not be completely true to conclude from the results of our model over time because of the low volume train data we used at the initial training.

If we had the right infrastructure we could train the model on more data, and even retrain the model using few computers to make it even faster using spark.

**7   Reference**

[1] Shreya Kohli, Gracia Tabitha Godwin, and Siddhaling Urolagin. (2021). "Sales Prediction Using Linear and KNN Regression".

[2] Lalitha Agnihotri, Nicholas Lewkow, Shirin Mojarad and Alfred Essa. (2016). "Educational Data Mining with Python and Apache Spark: A Hands-on Tutorial". *ACM. ISBN 978-1-4503-4190-5*.

[3] Mourad Azharia, Abdallah Abardab, Badia Ettakia,c, Jamal Zerouaouia and Mohamed Dakkon. (2020). "Higgs Boson Discovery using Machine Learning Methods with Pyspark".

[4] Zhipeng Gao, Qian Wang, Ting Wang, and Yang Yang. (2018). "Execution Time Prediction for Apache Spark". *Copyright is held by the owner/author(s). Publication rights licensed to ACM.*