Predicting store sales using Scikit-Learn

Amangeldy Sarsembay, Alisher Mussin and Bakdauren Narbayev

[1] *Nur-Sultan, Nazarbayev Universty, School of Engineering and Digital Sciences, Computer Science 3rd year*
[2] *Nur-Sultan, Nazarbayev Universty, School of Engineering and Digital Sciences, Computer Science 3rd year*
[3] *Nur-Sultan, Nazarbayev Universty, School of Engineering and Digital Sciences, Computer Science 3rd year*

December 4, 2020

Abstract - Which is the best way to forecast weekly sales? Effectively predicting data is crucial object of learning in sales. Walmart has released historical sales data for 45 stores located in different regions across the United States. We will use historical markdown data to predict store sales. We have such learning algorithms as Linear Regression, Random Forest Regression, Extra Trees Regression and MLP Regression. From the resulting observations, we concluded that Random Forest Regression best suit for this problem having generalization error smaller than what exists in the literature for the same dataset.

1 Introduction

Current time machine learning gives us the ability to explore the market and predicting sales of our target. Market exploring was a target of the interested people in all time and currently it is also a target of the investors and owners of the markets. Regression models are well suited to solve problems with the predicting and forecasting based on historical data which is given to as from 45 different Walmart stores. Our historical data-set has sequential nature, it takes time range from 2010-02-05 to 2012-11-01, so we used to predict the sales using a bunch of algorithms. Algorithms used in this project are Linear Regression, Random Forest Regression, Extra Trees Regression, MLP Regression. Also, we will identify sales in each department of the stores and affect of markdowns such as holidays on the sales. Machine Learning can help to identify where is minimum or no correlation between sales and factors given as in data. By applying machine learning and listed algorithms we can achieve our challenge to predict store sales using Scikit-Learn.

2 Methods

Walmart market is a good problem for applying machine learning and solve our challenge question. In this project, we have built and evaluated models using Python 3. We have employed the scikit-learn modules for calculating the learning algorithms while using pandas for data management, and matpotlib, seaborn for plotting.

2.1 Software

Tools which we used for implementing our project are python and Jupyter notebook. Python was our choice because we used it before and Python has good libraries and packages suitable for our data processing and training data. Packages and libraries used in our project are Numpy, Pandas, Seaborn, Matplotlib, Sklearn.

2.2 Dataset

We retrieved data from Kaggle which focused on 45 Walmart stores located in different regions. Based on this data-set we need to predict the sales of each department of the stores. In given data-set we have the structure of the files:
1. stores.csv,
2. train.csv,
3. test.csv,

4. features.csv.
The first file has anonymized information regarding the stores and their type and size. The second file is training data with the fields such as Store, Dept, Date, WeeklySales, IsHoliday. Files with the name features.csv include fields Store, Date, Temperature, FuelPrice, MarkDown from 1 to 5, CPI, Unemployment, and IsHoliday. Historical data covers days from 2010-02-05 to 2012-11-01. Data-set is big and it positively affects training our model and also its rarity that such a good data-set is open-source and free because this data might cost a lot of money.

2.3 Data Exploration

In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data [1]. In the broadest sense correlation is any statistical association, though it commonly refers to the degree to which a pair of variables are linearly related. We have checked for correlation this dataset's features.
In figure 1, we can see the visualisation of correlation of features in our dataset as a heatmap. From this figure we can see that features do not correlate much with each other.
In figure 2, there is a visualisation of weeklySales highest correlations with other features. Using this heatmap, we found 9 most correlated features for training and prediction of weekly sales. The dropped columns are: "Store", "Temperature", "Fuel Price", "CPI", "Unemployment", "Day", "Year", "Type B", "Type C".

2.4 Linear Regression

Linear regression is the one of the most popular statistical technique commonly used for predictive modeling. Breaking it down to basics, it comes to providing a linear combination of independent variables, on which our target variable is built upon:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_n x_n,$$

where every obtained coefficient is

$$\beta_i = Cov(x_i, y)/Var(x_i)$$

Using already existing function of Linear Regression from Scikit-learn, we experimented with regression on the dataset.

2.5 Random Forest Regressor

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification)

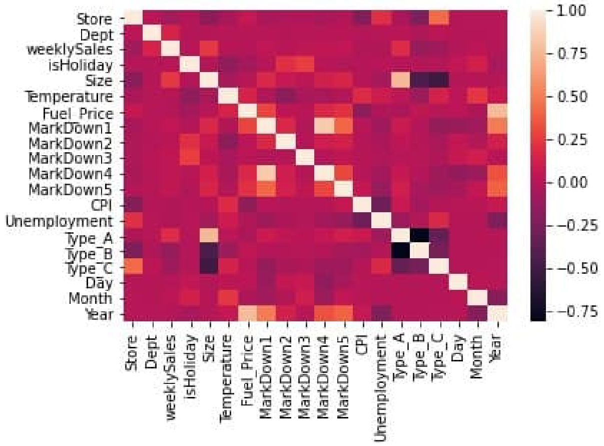


Figure 1: *Visualize correlation of features in our dataset*

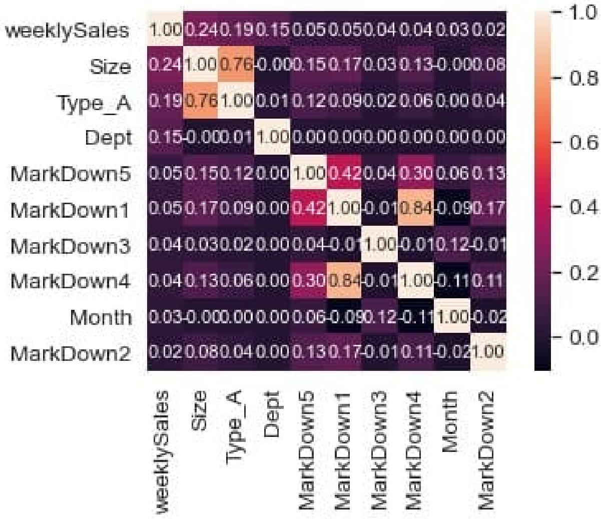


Figure 2: *Visualize weeklySales highest correlations with other features*

or mean/average prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set [2].

2.6 Extra Trees Regressor

ExtraTrees are similar to ordinary random forests in that they are an ensemble of individual trees, there are two main differences: first, each tree is trained using the whole learning sample (rather than a bootstrap sample), and second, the top-down splitting in the tree learner is randomized. Instead of computing the locally optimal cut-point for each feature under consideration (based on, e.g., information gain or the Gini impurity), a random cut-point is selected. This value is selected from a uniform distribution within the feature's empirical range (in the tree's training set). Then, of all the randomly generated splits, the split that yields the highest score is chosen to split the node. Similar to ordinary random forests, the number of randomly selected features to be considered at each node can be specified [2].

2.7 MLPRegressor

A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN) [3]. An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training.Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

3 Results

These efforts have brought us some remarkable results. We have used 4 different models for training our data set. First one was using Linear Regression. In figure 3, we can see the prediction accuracy of this model (red dots - prediction, blue dots - actual values). In the figure, there are only the first 100 results, because it is impossible to show all. The second one was Random Forest Regressor. Its predictions are in figure 4. The third model was Extra Trees Regressor with figure 5 and in figure 6 there is a prediction of the last algorithm MLPRegressor.

This dataset prediction is evaluated on the weighted mean absolute error WMAE (figure 7). From this type of evaluation, we can see the results of 4 learning models in table 1.

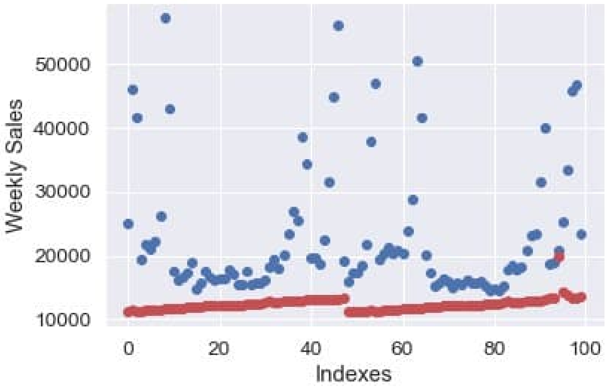


Figure 3: *Predictions for Linear Regression*

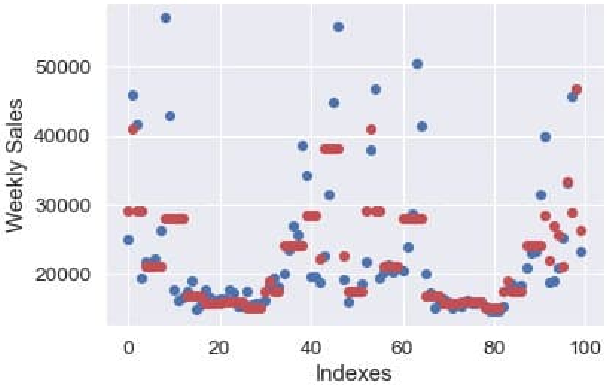


Figure 4: *Predictions for Random Forest Regressor*

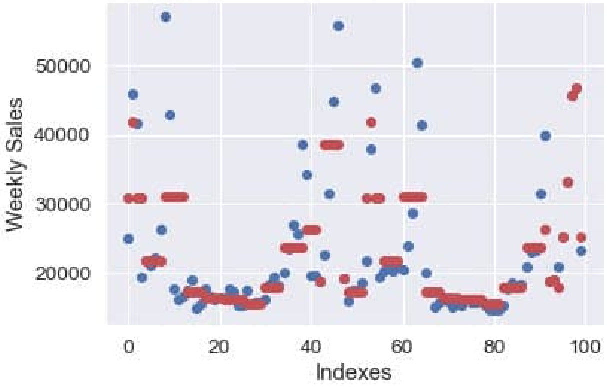


Figure 5: *Predictions for Extra Trees Regressor*

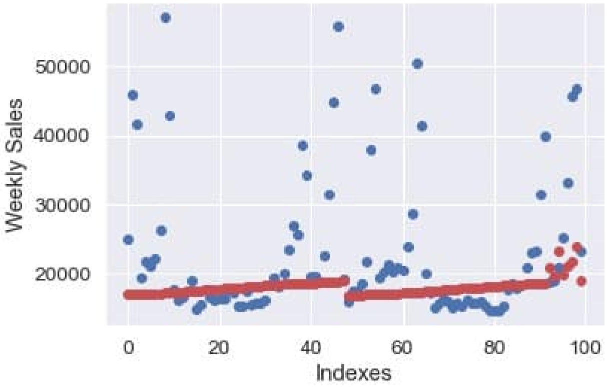
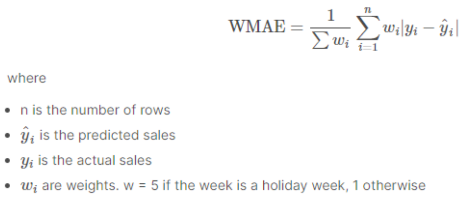


Figure 6: *Predictions for MLPRegressor*



$$\mathrm{WMAE} = \frac{1}{\sum w_i}\sum_{i=1}^{n} w_i|y_i - \hat{y}_i|$$

where

- n is the number of rows
- $\hat{y}_i$ is the predicted sales
- y_i is the actual sales
- w_i are weights. w = 5 if the week is a holiday week, 1 otherwise

Figure 7: *the weighted mean absolute error WMAE*

Table 1. Results of 4 learning models

Learning model	WMAE
Linear Regression	15670.300899184955
Random Forest Regressor	2654.8706877291634
Extra Trees Regressor	2981.036164946908
MLPRegressor	15194.464305684587

4 Conclusion

In our study, we have used Linear Regression, Random Forest, Extra Trees and MLPRegressor for predicting based on historical data, which is given to us from 45 different Walmart stores. Our results show that Random Forest Regressor is best suited for this type of problem among 3 other algorithms. Using this model, we have submitted the predictions on test dataset to kaggle to check the results. It was 6385.58098. In the literature for the same dataset that we have referenced, they had 11517.09 on test dataset [4]. This shows that we obtained generalization error smaller than what exists in the literature for the same dataset.

5 References

1. https://en.wikipedia.org/wiki/Correlation_and_dependence
2. https://en.wikipedia.org/wiki/Random_forest
3. https://en.wikipedia.org/wiki/Multilayer_perceptron
4. https://books.google.com/books?hl=enlr=id=7Y3fAwAAQBAJoi=fndpg=PA135dq=Walmart+Store+Sales+Forecasting+machine+learningots=XyWYoJpbs6sig=IYOXiuJFh2dTZCnXn-zSg5C79_s

Datasets:
19 Free Public Data Sets For Your First Data Science Project https://www.springboard.com/blog/freepublic-data-sets-data-science-project/
17. Walmart
Walmart has released historical sales data for 45 stores located in different regions across the United States. This offers a huge set of data to read and analyze, and many different questions to ask about it—making for a solid resource for data processing projects.

6 Contributions

All students made good contribution to the project.
Alisher Mussin - selecting the best learning machine, implementing a cross validation (CV) strategy.
Amangeldy Sarsembay - data exploration, selecting the best set of hyperparameters.
Bakdauren Narbayev - selecting the best features, selecting the best preprocessing.
All work was shared equally between students and each student explained own part to other students to fill the moments which someone may lose. Also report writing and video presentation was made as teamwork to summarize the project.