Forecast Of Walmart Sales Using Big Data

Kowsik Bhattacharjee, MSc in Big Data Analytics, LYIT, Letterkenny

***Abstract*—The ability to accurately forecast data is critical in a variety of fields, including healthcare, sales, banking, weather, and sports. The study and implementation of big data techniques and the ensemble regression algorithm on Walmart sales data is presented here, which consists of weekly retail sales statistics from different departments of Walmart retail stores across the United States of America over the course of three years, with pre-holiday and holiday data presenting an increase in sales. Random is the model that was used to make the prediction. An evaluation of the model and its ability to predict accurately was completed. The advancement of big data analytics can support a variety of industries. However, only a few researchers have looked into the use of machine learning techniques with large amounts of data. As we all know, ML-based Big Data forecasting approaches are time-consuming. In this paper, we will assess efficiency by performing sales forecasting on a big data platform with a high-dimensional and broad dataset. Artificial neural networks have also been shown to increase efficiency and deliver highly accurate results.**

***Index Terms*—Machine Learning, Data Visualization, Big Data, Data Bricks, Spark, Neural Networks, Random Forest, Sales Forecasting**

I. Introduction

FOrecasting is concerned with accurately predicting the future and it provides critical inputs for many planning processes in business, such as financial planning, inventory management, and capacity planning. There has been considerable interest in both industry and academia in the development of methods that are capable of accurate and reliable forecasting, and many new methods have been proposed each year. In a world where massive quantities of data are gathered on a regular basis, data analysis is a must. Extensive supermarkets, such as Walmart, for example, handle hundreds of millions of transactions each week at thousands of locations around the world.

In today's world, where competition is growing, making business decisions is becoming more difficult. As a result, being able to predict something accurately has become extremely useful. As a result, many industries, such as retail, will make informed decisions based on predictions using big data techniques and machine learning algorithms. Forecasting is crucial in the retail industry, as previously said, because revenue prediction is a more conventional application of forecasting. Overestimation of revenue may result from incorrect forecasting, resulting in substantial business losses and inventory holding costs. Underestimating sales in a forecast, on the other hand, will result in a missed market opportunity and the loss of key stakeholders.

A organization can accurately assess and evaluate revenue, as well as how consumers respond to specific goods or marketing strategies, using big data techniques. Sales forecasting employs patterns or trends gleaned from historical data to accurately predict sales, allowing knowledgeable decisions to be made about inventory management for future production. The issue in this study is focused on a competition on the Kaggle platform to forecast weekly sales for Walmart regional stores [3]. Exploring and analyzing the results using a machine learning model such as Random Forests [4][5], is a supervised learning algorithm that uses an ensemble learning method for classification, regression, and other tasks, that functions by building a large number of decision trees at training time and producing the value that is the mode of the classes (classification) or producing the value that is the mean of the values (regression) of the individual trees.

A. What is Big Data

Big Data is a common concept used to characterize the exponential, structured and unstructured development, availability, and use of knowledge. Big data can be exploited in ways never before possible, and is known as the large amount of both structured and unstructured data that is accessible over the internet. This knowledge has become the backbone of the forecasting field. The oil of the knowledge economy that must be viewed as an economic commodity is Big Data. If not, businesses are doomed to reinforce the old witticism that a skeptic recognizes the price of something and the worth of nothing. Yet Big Data’s importance is not well known [1].

Fig 1: Representation of Big data

Companies are beginning to understand that data is one of their precious assets, no matter what sector they are in. Data will unleash new types of economic value if correctly harnessed.

B. Organizational challenges of big data forecasting

By integrating big data into sales processes at a strategic level, each company must decide when and how much big data technology should be integrated into its planning process. This is based on the relative benefits that the company can potentially reap against the cost of collecting and analyzing such data. Our goal isn’t to address the strategic question of whether and how often to integrate big data into organizational processes; rather, we want to see how quickly we can use a big data system like Data Bricks to handle such a large volume of data and apply supervised machine learning models to get better results. [3].

C. Capturing Big Data

Big data brings the opportunity to enhance demand predictions and provide interesting insights into consumer habits with it. However, these possible advantages for demand planners raise immense practical challenges. Second, the sheer data volume can be daunting. Walmart, for instance, gathers more than 2.5 petabytes (1 petabyte = 1 million gigabytes) of data from one million customer transactions every hour [6]. However, only about 0.5 percentage of all collected data is analyzed [7].

A practical question, therefore, is what knowledge should be processed, and for how long? Second, Feng and Shanthikumar [8] points out that while" theoretically more information leads to better forecasts, the challenge, however, comes from dealing with the increased number of variables and their ambiguous relationships." Since large datasets, particularly those used in forecasting, are sparse and non-repetitive, they argue that semi- or non-parametric methods like machine learning are better suited for analyzing them than traditional time series forecasting techniques.

II. Frame Work And Software

We’ll be using the Apache Software Foundation’s Spark big data platform. Apache Spark parallelizes data processing by distributing it throughout the cluster. The SPARK CONTEXT is the Spark’s driver software, and it communicates with the cluster managers, who are linked to the executors on the worker nodes. The staff nodes are the Spark frame’s main processing unit, and they’re in charge of computing the data sets. For ease of implementation, we used the Data-bricks version, which includes an inbuilt Spark Cluster setup in a single node configuration. There are a few benefits to using Apache Spark that should be noted.

1) Since it uses in-memory computation, it is faster in computing bigdata frameworks than Hadoop MapReduce.
2) With respect to the first stage, Spark cannot replace Hadoop, but it can be used in conjunction with it for processing where the former was used for real-time processing and the latter was used for batch processing. [10].
3) RDDs are a form of RDD that is used (Resilient distributed data sets) Spark’s fundamental building block is that it can be run in parallel and spread through clusters, and it is highly accessible and fault tolerant, so it can instantly recover from failures [10].

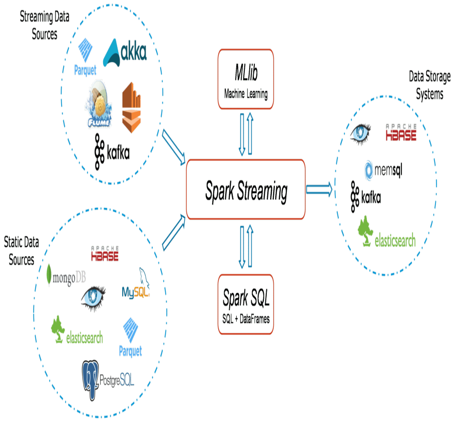


Fig 2: Overview of Spark

4) Spark Streaming is a streaming data network for processing images.
5) It comes with pre-installed Machine Learning libraries.
6) Compatibility with any API, such as Java, Scala, Python, or R, makes programming easy.

III. Dataset

The dataset used for this research comes from the Kaggle Platform [9], which is used for projects and data science related works. The data collection consists of Walmart Inc., an American retail company. It consists of data from 45 centralized Walmart stores around their sales per week. The model used for the study has 421,517 entries that will be used for training. Since no test set is provided, we use 20 percent of the training data provided for crossvalidation, and final review. Details of datasets used and there attributes are as follows:

stores.csv: This file contains anonymized information, indicating the nature and size of the store, concerning the 45 stores.

train.csv: This is the historical data on training spanning the years 2010-02-05 to 2012-11-01. We can find the following fields inside this file:

Store -	store number
Dept -	department number
Date -	week
WeeklySales -	sales for a department in the given store
IsHoliday -	whether the week is a special holiday

features.csv: This file contains additional data for the provided dates related to the shop, department, and regional operation. This includes the following fields:

Store -	store number
Date -	week
Temperature -	average temperature in the region
FuelPrice -	cost of fuel in the region
MarkDown1-5 -	data related to promotional markdowns
CPI -	the consumer price index
Unemployment -	the unemployment rate
IsHoliday -	whether the week is a special holiday

MarkDown data is only available after Nov 2011, and is not available for all stores all the time.

In the dataset, there are four holidays:

Superbowl:	12-2-10, 11-2-11, 10-2-12, 8-2-13
LaborDay:	10-9-10, 9-9-11, 7-9-12, 6-9-13
Thanksgiving:	26-11-10, 25-1-11, 23-11-12, 29-11-13
Christmas:	31-12-10, 30-12-11, 28-12-12, 27-12-13

IV. Data Visualization And Evaluation

A. Store Type Vs. Weekly Sales

The revenue distribution by Store Form is depicted in the diagram below. We discovered that, of the three types of stores: Type A, Type B, and Type C, Type A stores have higher medians than the other store types, meaning that weekly sales for Type A stores are higher than the other store types.

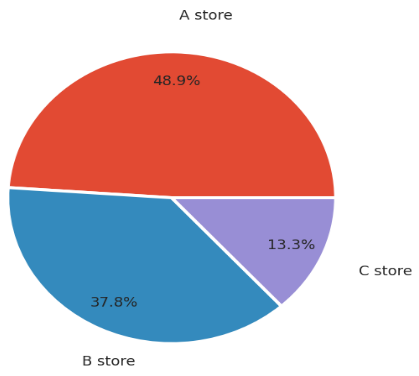


Fig 3: Store Type Vs. Weekly Sales

B. Weekly Sales: Holidays Vs Non-Holidays

Figure 4 shows below typical weekly sales on a holiday and non-holiday period.

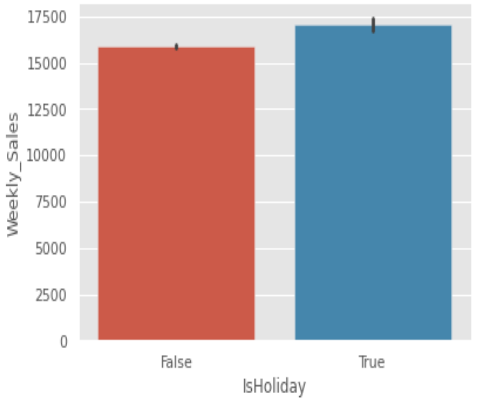


Fig 4: Weekly Sales - Holidays Vs. Non-Holidays

C. Weekly Sales by Department and Holiday

Figure 5 shows Box Plot representation of Weekly Sales by Department and Holiday.

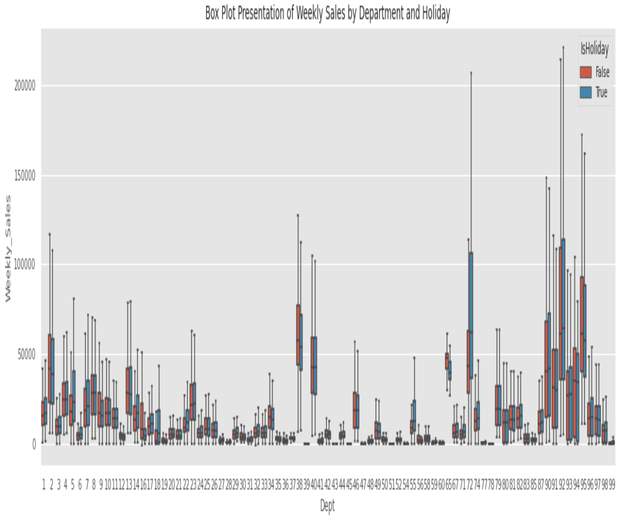


Fig 5: Weekly Sales - Department Vs. Holiday

D. Weekly Sales by Month and Holiday

Figure 6 shows Box Plot representation of Weekly Sales by Month and Holiday.

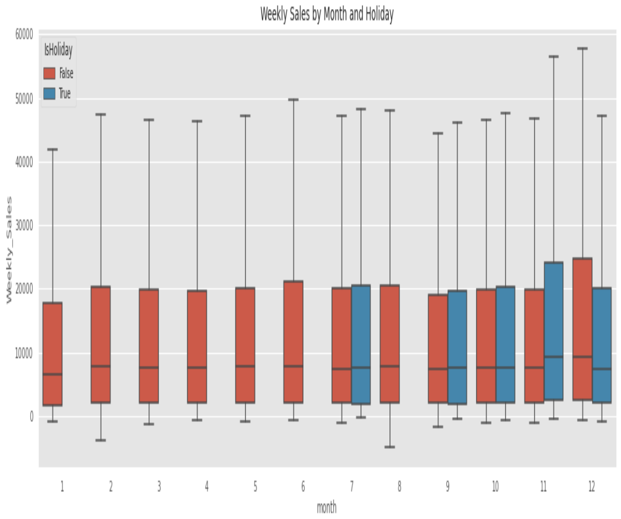


Fig 6: Weekly Sales by Month and Holiday

E. Visualising Weekly Sales

Figure 7 shows how to keep an eye on Weekly Sales so far to predict near future sales even better.

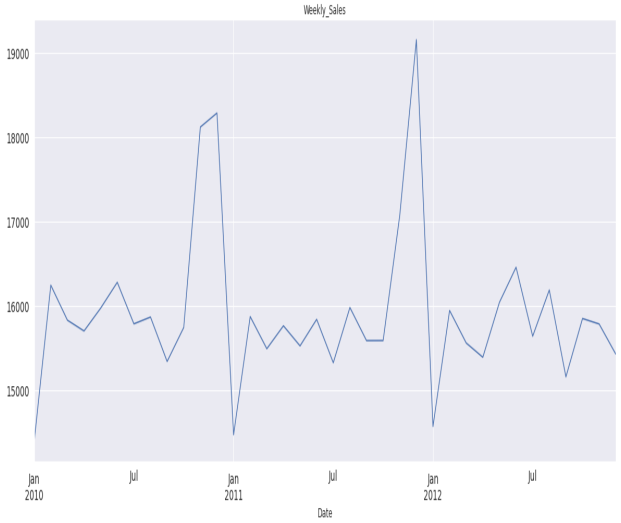


Fig 7: Visualising Weekly Sales

V. Data Processing and Modelling

For this research we used techniques such as:

Data Cleaning: To remove inconsistent data from actual dataset.

Data Integration: Here we will combine multiple data sources.

Data Selection: Relevant data for the research analysis task are retrieved from various datasets.

Data Transformation: Where data is transformed and compiled via description or aggregation operations into forms suitable.

Pattern Evaluation: To define the truly fascinating trends that reflect information centered on measures of interest.

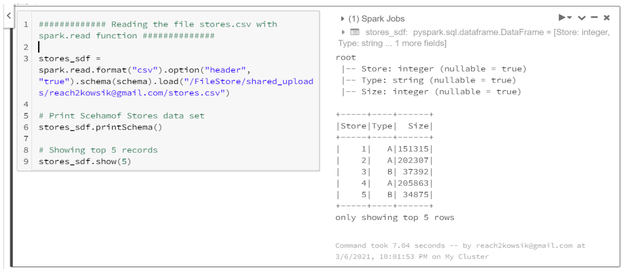
Knowledge Presentation: Where techniques of representation of visualization and information are used for users to present.

We can see that sales are associated with discounts and holidays, as well as higher sales with a larger store. As a result, we can understand that bigger retailers produce more sales, discounts generate higher sales prices, and higher unemployment generates less sales. There appears to be no connection between weekly sales and holidays, temperatures, or fuel prices.

The above mentioned data handling techniques were conducted once we acquired the dataset required to effectively forecast the weekly sales. We first explored the data, for example, to get a sense of its structure and the values embedded within it. We cleaned and extracted conflicting data from the data before transforming it by combining and adding variables that were useful for the analysis and forecast.

We extracted the entire collection of columns from my project paper and filtered out the attributes that were needed. Data extraction can be achieved with a variety of tools, including OpenRefine, Data Wrangler, and Python Pandas, but in my project, we'll be using the Pandas library in conjunction with the Spark read.csv function. As part of the data cleaning process, I discovered a few attributes that aren't important to the data. The benefits of using Spark for parallel processing, as discussed in the preceding section, aid in reducing significant amounts of processing time. Since we used it for research, our dataset isn't really huge, so Spark won't be able to display significant variations in processing capacity, but in terms of industry, it's way beyond the imagination for its high speed when the data is in Zetabytes.

Data is stored in data frames, which are similar to the tables used in relational database management systems (RDBMS). SqlContext is an object that is specified to allow Spark to handle structured data. For data pre-processing, the Pyspark Spark distribution is used. In comparison to SparkSession, import of the SparkContext is adequate starting with version 2.0.0, as SparkContext contains all and serves as a gateway to Spark functionality. Refer figure-8



```
############# Reading the file stores.csv with spark.read function ##############

stores_sdf = spark.read.format("csv").option("header", "true").schema(schema).load("/FileStore/shared_uploads/reach2kowsik@gmail.com/stores.csv")

# Print Scehamof Stores data set
stores_sdf.printSchema()

# Showing top 5 records
stores_sdf.show(5)
```

```
stores_sdf:  pyspark.sql.dataframe.DataFrame = [Store: integer, Type: string ... 1 more fields]
root
 |-- Store: integer (nullable = true)
 |-- Type: string (nullable = true)
 |-- Size: integer (nullable = true)

+-----+----+------+
|Store|Type|  Size|
+-----+----+------+
|    1|   A|151315|
|    2|   A|202307|
|    3|   B| 37392|
|    4|   A|205863|
|    5|   B| 34875|
+-----+----+------+
only showing top 5 rows

Command took 7.04 seconds -- by reach2kowsik@gmail.com at
3/6/2021, 10:01:53 PM on My Cluster
```

Fig 8: Snippet of Spark Environment

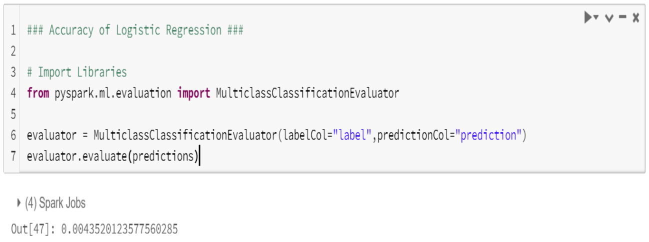
A. Concepts And Techniques

1) Predictive Modelling: Predictive modeling is the foundation of the machine learning field. This model incorporates the use of machine learning algorithms to make predictions based on the training dataset. Regression and classification, also known as pattern classification, are two subcategories of predictive modeling. The first involves making final forecasts based on inferences drawn from the study of variables and patterns. The latter subclass is described as assigning class labels to specific data as a result of a prediction's performance. This pattern classification can also be divided into two types of learning: supervised and unsupervised. In contrast to unsupervised learning, supervised learning is characterized as knowing the class labels and the output. We've outlined some of the main reasons why predictive analysis is important for this study below.
1) It will assist us in gaining an understanding of the sales department's success during holidays and non-holidays, as well as the results of weekly sales.
2) Visualization methods used during the data pre-processing stage may be used to identify potential outliers.
3) The data can be visualized in basic bar graphs and sent as a report to Walmart's Sales Analytics Department, which can then make any necessary adjustments to forecast strategies.

B. Variants Of Predictive Modelling

As previously mentioned, the classification type is the second type of modeling technique we used in our model. Below is a list of the classification techniques we used in our model.

1) Logistic Regression: The dependent variable in this model is either a group or a binary value in statistical terms. It is further divided into binomial logistic regression and multinomial logistic regression in the form of spark.ml. If we need to predict a binary result, we can use binomial logistic regression; however, if we need to predict a multiclass output, we can use multinomial regression [11].



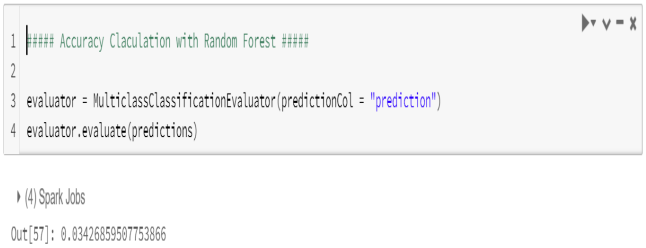
```
### Accuracy of Logistic Regression ###

# Import Libraries
from pyspark.ml.evaluation import MulticlassClassificationEvaluator

evaluator = MulticlassClassificationEvaluator(labelCol="label",predictionCol="prediction")
evaluator.evaluate(predictions)
```

```
▸ (4) Spark Jobs
Out[47]: 0.004352012357756O285
```

Fig 9: Accuracy of Logistic Regression

2) Random Forest: An ensemble machine learning algorithm is a common machine learning algorithm that overcomes the problems of overfitting that decision trees face by combining multiple decision trees. The parameters numTrees, maxDepth, and maxBins must be specified when designing a Random Forest algorithm.

NumTrees:- is the number of trees, which in our case is 100; increasing the number of trees reduces variance in prediction, thereby enhancing model test-time accuracy.
Maxdepth:- Maximum tree depth; the deeper the tree, the more strong it is; however, training takes longer.
MaxBins:- On each node, this is used to break on features. The estimator in the pipe line, the Random Forest classifier, can be exported using the command below [12].



```
##### Accuracy Claculation with Random Forest #####

evaluator = MulticlassClassificationEvaluator(predictionCol = "prediction")
evaluator.evaluate(predictions)
```

```
▸ (4) Spark Jobs
Out[57]: 0.03426859507753866
```

Fig 10: Accuracy of Random Forest

VI. Table 1

Table I

Comparison Of Algorithms

Algorithm	Accuracy
Logistic Regression	0.004352012357756O285
Random Forest	0.03426859507753866

VII. Conclusion And Future Work

We're wrapping up the research report with all of the above dataset analysis. The work primarily entails predicting models using the Spark system and creating visualisations, which are the most straightforward way for a layperson to understand the report.
The visualisation consists of a bar graph, a pie map, and a box plot. Parallel processing cross validation is the focus of our future research. Finally, we are confident that the amount and variety of research performed, when presented to the Walmart Sales Analytics Department in the form of a report, would undoubtedly assist them in streamlining their forecasting method in a comprehensive manner, resulting in more accurate predictions and thus assisting in the rapid growth of the company.

VIII. References

[1] https://search.proquest.com/openview/c1a0851a512acca5fcb0ce0067c1f0fe/1?pq-origsite=gscholarcbl=28144

[2] Ren S., Patrick Hui C., Jason Choi T. (2018) AI-Based Fashion Sales Forecasting Methods in Big Data Era. In: Thomassey S., Zeng X. (eds) Artificial Intelligence for Fashion Industry in the Big Data Era. Springer Series in Fashion Business. Springer, Singapore. https://doi.org/10.1007/978-981-13-0080-6$_2$

[3] https://www.sciencedirect.com/science/article/pii/S0169207018301523

[4] A. Chakure. (2019, June 29). Random Forest Regression, Towards Data Science [Online]. Available: https://towardsdatascience.com/randomforest-and-its-imple mentation-71824ced454f. [Accessed: August 1, 2019].

[5] T. You. (2019, June 12). Understanding Random Forest," Towards Data Science [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d.

[6] https://www.dezyre.com/article/how-big-data-analysis-helped-increase -walmarts-sales-turnover/109

[7] https://www.technologyreview.com/s/514346/the-data-made-me-do-it/.

[8] Feng and Shanthikumar, 2018 Feng Q., Shanthikumar J.G. How research in production and operations management may evolve in the era of big data Production and Operations Management, 27 (9) (2018), pp. 1670-1684

[9] https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/ data.

[10] Z. Han and Y. Zhang, "Spark: A Big Data Processing Platform Based on Memory Computing," 2015 Seventh International Symposium on Parallel Architectures, Algorithms and Programming (PAAP), Nanjing, China, 2015, pp. 172-176, doi: 10.1109/PAAP.2015.41.

[11] “Classification and regression - Spark 2.2.0 Documentation.” [Online]. Available: https://spark.apache.org/docs/2.2.0/ml-classificationregression.htmllogistic-regression. [Accessed: 05-Mar-2021].

[12] .eranga, “Random Forest classifier with Apache Spark,” Medium, 11-Jul2019. [Online]. Available: https://medium.com/rahasak/random-forestclassifier-with-apache-spark-c63b4a23a7cc. [Accessed: 05-Mar-2021].