National College of Ireland

Demand Forecasting based on External Factors using Clustering and Machine learning

MSc Research Project
Data Analytics

Sruthi Prabakaran Paruthipattu
Student ID: x19223269

School of Computing
National College of Ireland

Supervisor: Dr. Barry Haycock

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sruthi Prabakaran Paruthipattu
Student ID:	x19223269
Programme:	Data Analytics
Year	2021-2022
Module:	Research Project
Lecturer:	Dr. Barry Haycock
Submission Due Date:	16/12/21
Project Title:	Demand forecasting based on external factors using clustering and machine learning
Word Count:	6151
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Sruthi Prabakaran Paruthipattu

Date: 16/12/21

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Demand Forecasting based on External Factors using Clustering and Machine Learning

Sruthi Prabakaran Paruthipattu
x19223269

Abstract

Forecasting time series data is challenging when multiple factors are taken into account. The retail industry is one such where multiple factors play an important role. And it is important to maintain the supply chain processes. At different levels of the supply chain, various factors influence the demand creating a dependency. Factors like weather, promotion, holiday, etc create a huge impact on the demand for a few to multiple products. And these factors may influence retail differently based on the location, event, economic crisis, etc. Decision-makers like the CEO, logistics manager, Branch managers, etc will make use of this data in taking impactful decisions and these affect the retailers, production engineers, suppliers, etc indirectly. This research aims to see on what scale some of these factors affect the demand for products in clusters of stores that are related in different aspects using hierarchical clustering. To forecast demand within the identified clusters, ensemble models like random forest and XGBoost is used. A relatively new combinational model of univariate LSTM and RF is also implemented based on previous work. Out of all three models, XGBoost performed well in terms of RMSE and MAE and the results were visualized using the SHAP library. This research is unique in terms of considering partial new store data, macroeconomic factors: oil price, and the influence of these factors along with other factors on different clusters of stores.

1 Introduction

1.1. Background

The number of insights that can be obtained from big data is mindblowing and industries worldwide are now using the same to develop their business. One such is the retail industry that highly depends on the data to balance the demand and sales for an uninterrupted supply chain, balanced sales, and stockout, regional distribution, inventory balancing, etc (Tarallo et al., 2019). In this highly turbulent market, forecasting demand is invaluable to businesses worldwide. It impacts the profitability of the company, because of its direct impact on decision-making for production, demand, supply, stock, logistics, marketing, and finance. So there is a need for robust and accurate models that aligns well with the needs of the stakeholder. As mentioned by Pereira Da Veiga et al. (2021), more accuracy means better decision-making both from a technical and business perspective. This in turn will contribute to better operational efficiency, financial stability, and better customer satisfaction. Multiple things could go wrong in the supply chain. One small misstep can create a ripple causing the bullwhip effect (Carbonneau, Laframboise & Vahidov, 2008), creating unexpected results.

One example would be the famous KFC chicken supply crisis[1] in 2018 and marketing failure[2] back in 2009 (Baryannis, Validi, Dani, and Antoniou, 2018). Both these mistakes created a bullwhip effect disrupting the supply of raw materials, sales, and shortage leading to customer dissatisfaction. To avoid such instances, companies spend a lot and rely heavily on analysts and decision-makers to predict future demand. For any scale of business or supply chain, there are a few factors like weather, promotion, holidays, competition, etc that affect the demand and sales directly and a few other factors like stock prices, GDP, pandemic, etc that affect indirectly. These factors make the data non-linear and this property is often hard to capture using conventional statistical methods. Although there are multiple methods and models, the uncertainties around the best suitable model that can converge for retail-specific factors are still not available. One cannot generalize the factors influencing demand to a wider extent unless it is a multinational retail giant. And yet, there are a few exceptional cases that need to be processed separately like for instance the effect caused by the current pandemic. These factors influence the demand and sales volumes differently based on the location and other aspects of the retail. Often, analysts forget this and generalize the results obtained for the sample of data used.

1.2. Motivation

This research was inspired by the day-to-day activities in a café. Some days, the products were overstocked leading to a lot of waste generation, and some days, the products were understocked causing product unavailability. In retail, it is important to foresee future requirements, consider activities that might tweak the sales, and adjust the supply chain accordingly. Though there have been multiple research works done in this domain, there are still a lot of areas that need to be researched.

1.3. Objective

The main objective of this research is to find out the hidden similarities and sales behaviour between multiple stores using hierarchical clustering based on different measures like an aggregate of unit sales of different products in each store at the day level, transactions of each store at the day level, etc. Based on the outcomes, for a few chosen clusters, demand forecasting is done using multiple ensemble models. This could be implemented in small retail chains to generalize a model within identified clusters and make adjustments on time based on the forecasted demand for individual products.

1.4. Research Question

What kind of external factors influence the sales and demand in retail?

- Are these factors different between groups of stores that are similar in some aspects?
- How well can we predict the demand based on these factors using ensemble models?

[1] www.cips.org/supply-management/analysis/2018/february/five-lessons-from-the-kfc-chicken-crisis/
[2] https://www.nuvonium.com/blog/view/kentucky-fried-chickens-viral-marketing-failure

1.5. Outline

The following sections are structured as follows: Section 2 is the Literature survey describing the previous state-of-the-art research works conducted in this domain, Section 3 is the methodology followed in this with sub-sections on the data, data-pre-processing, and the proposed approach. Section 4 is about design specification. Section 5 details the implementation. Section 6 is Evaluation & discussion and lastly the conclusion and future work.

2 Related Work

Forecasting is not a simple task and multiple factors including the type of data, type of model used, problem domain, etc impact the accuracy of the results obtained. And no one universal model could be generalized till now. But there has been a lot of research done in this area and the further sections describe multiple research works under various domains within this.

2.1 Demand Forecasting in Retail

The application of Artificial Intelligence (AI) in Supply chain and demand forecasting dates back to the early 1990s. For instance, Min (2010) has consolidated all the applications of AI in Supply Chain Management. And one of the research areas is demand planning and forecasting. And the author stresses one particular drawback of forecasting demand for new products. Fast forward a decade later, more research areas have popped up. A similar study was done was by Ni, et al. (2019). Besides addressing major research areas within demand forecasting, the authors have also listed the top most used algorithms used. Among which the most popularly used are the Random Forest, Decision Tree, K-means, etc. Choosing the algorithm depends highly on the nature and the complexity of the data available. This research work points out which model to use with specific kinds of data. Although within the last three years, many advanced models have been implemented.

The supply chain is becoming very data-intensive and Fildes, et al. (2019) have presented a very accurate review of the research works undertaken, some state-of-the-art models, limitations, and areas that need to be improved in forecasting retail demand. Accordingly, retailers need forecasting models to support tactical, strategic, and operational decisions. And these decisions are based on the level of aggregation: market-level, chain-level, store-level, and product-level. In the case of the market-level, the authors have concluded that there is no recent research found which includes macroeconomic factors in forecasting demand. In the case of chain level, new research needs to be done to incorporate both online and offline sales data. And in the case of store-level aggregate, a lot of research has already been done. But, often analysts forget to evaluate data from new stores. Lastly, the authors have listed the dimensions of product level aggregate into time, product, and supply chain. No new research is available for demand forecast with interrelated hierarchies over all three dimensions. This volume is frequently influenced by multiple factors like weather, promotion, intermittence, etc. Few other noteworthy mentions include new product forecast, the impact of product launch, etc. It clearly shows that there are a lot of potential research areas within retail

demand forecasting. Keeping in mind all these, this research work considers the macroeconomic factor, new store, and the external factors.

Two research works considering the influence of factors are by (Patak et al., 2015 & Ramanathan and Muyldermans, 2010). In the first case, the authors argued that there are no good statistical methods to handle the external factors and their influence. They represent the weightage for each factor to find which ones contribute the most to the sales and demand volume. In the second case, the authors consider almost multiple influential factors but stress more on the promotions applied to the products. This shows that the influence of factors varies a lot depending on the retail, place, type of retail, type of customers, etc.

2.2 Demand Forecasting using Traditional Statistical Methods

In the research work by Ren, et al. (2019), the authors have argued why statistical models are better when compared to machine learning models. They have based their argument on the run-time, model, complexity, and ability to extend the model with business decisions. And these capabilities are weak in terms of machine learning models. And that to have a win-win situation, the researcher might go for hybrid models. This research is very opposite to what others have suggested. They concluded that AI and ML are very advanced and will capture the complex features within the data, but it also has its drawbacks and it very much depends on the data and type of model chosen. Where traditional methods are enough, the researcher should not overfit the model by using complex systems. This paper alerts the researchers to be mindful of choosing the best model that fits the problem statement and data used.

It is important to understand the data better before choosing a model. With complex and huge data sets traditional statistical models fail to perform (İşlek and Gündüz Öğüdücü, 2015). The authors worked on a huge warehouse supply dataset using Bayesian Networks. To improve the performance and scaling of the algorithm, the authors proposed a new approach to cluster similar warehouses together using bipartite clustering based on moving average values and then run the models separately for each cluster. This is relatively a new approach to incorporate more data. Although Bayesian networks performed well in this case, more advanced approaches could have been tried to understand the behaviour of models with the clusters. A similar approach was proposed by Chu & Zhang, (2003). They have made a comparative study of linear and non-linear models on aggregated retail sales data. They have argued that the traditional forecasting methods fail to capture the seasonal fluctuations in sales. To reduce this, they have aggregated the data monthly and have deseasonalized it. The choice of linear models includes ARIMA, dummy regression, trigonometric models, and non-linear models including various layers of neural networks. They found neural networks perform well with deseasonalized data compared to other models. They conclude that clustering works well to reduce the overhead involved in training each data point separately. This seems to be reasonable and, in this research, the concept of clustering has been used.

2.3 Univariate Demand Forecasting

Though univariate demand forecasting provides a good understanding and there are multiple statistical and machine learning models to do the same, it often lags in capturing the hidden patterns and effects of other factors. One of the recent innovations for univariate analysis is

the FB prophet created by Taylor & Letham (2018). This model can handle seasonality, trend, outliers, the effect of holidays, and many other use cases. The authors Jha & Pande, (2021) used this model on aggregated sales data and suggested that it outperforms all traditional univariate models by a large index. But, the authors might have used the same on multiple datasets before concluding this. They have suggested using transfer learning with FB Prophet to scale the model to large datasets. Other notable works include the comparison between deep neural networks and gradient boosting by Wanchoo (2019). He argues that not every retailer will have access to the data with supporting variables and that a robust model should be worked upon for univariate data. For comparison, he has used the mean square root error (RMSE) and mean absolute error (MAE). Both the models had their drawbacks. But the conclusion was that the gradient boosting model performed well in terms of the error measures. This seems to be another innovative approach.

In cases where influential factors cause abrupt changes, the performance of univariate forecasting is poor (Guo, Wong & Li, 2013). There are various research works wherein the authors have used multiple ensemble models and deep learning to implement multivariate forecasting as mentioned in the next section.

2.4 Multivariate Demand Forecasting using Ensemble models and Deep learning

Though univariate analysis is easier to implement, its adequacy is not enough when it comes to capturing complex interdependencies. In this section, a few recent and past works on multivariate demand forecasting are discussed.

Ensemble models like Random Forest, XGBoost, LGBM, etc, and deep learning models like CNN, LSTM, etc are now being used popularly to forecast demand. For instance, Pacella and Papadia, (2021) in their recent research proposed multiple LSTM and Bidirectional-LSTM networks to predict demand for 10 different time-series datasets. Out of all the models, BLSTM worked the best, enhancing the accuracy of other LSTM networks. Since the models were tested on different datasets, it is safe to conclude that this is very robust and be scaled to more data.

Multi-channel retail involves much more than a simple forecast. In this research, Punia et al. (2020) have used a combination of random forest and deep LSTM networks to capture the temporal and regression aspects of the data. This combination of the model was suggested to handle the huge volume of data with multiple features. The authors used the residue values from univariate LSTM to train the RF model other features and the result was the sum of the target predicted by LSTM and the residue predicted by RF. The forecasting metrics used were ARME, ARMAE, and ARMSE. Out of all the models used for comparison, the proposed model performed well. This seems to be a robust model and handles the disadvantages of both models very well. Chawla et al., (2018) followed a different approach to forecast demand using Artificial Neural Networks (ANN) using MATLAB. But, this approach is suited only for low-sized datasets and it cannot handle huge datasets. But, this approach could be considered for univariate forecasting in situations where it is only a few hundred or thousand rows of data.

The dataset chosen for this project is very huge and few authors have presented their works using different ensemble and deep learning models. Weng et al., (2019) proposed an innovative combined model using LGBM and LSTM. The time-series features were modeled using LSTM and the statistical features were handled using the LGBM. The result from the LSTM converted as a net feature and was given as an input to the LGBM. The results were compared using weighted RMSLE. The model performed well than other models and to scale its working, it was tested on 2 other datasets. Researchers are now interested in using combinations of different models. Another example is the work by Kechyn et al., (2018). They went ahead one step and used a combination of CNN WaveNet. But due to the model complexity, it was prone to overfitting, and to handle this moving average was used. It is important to handle the complexity of the model based on the type of data in use. There is a lot to be considered, and based on this literature survey, few things have been implemented and few others were left for future work.

3 Research Methodology

This research aims to find similar groups of retail stores based on various aggregate measures and apply ensemble models to forecast and find influential factors affecting the demand volume. The research flow of this research is as presented in Figure 1. The overall methodology followed is KDD. Starting from Data acquisition, data pre-processing, Exploratory Data Analysis, modeling, and Evaluation.

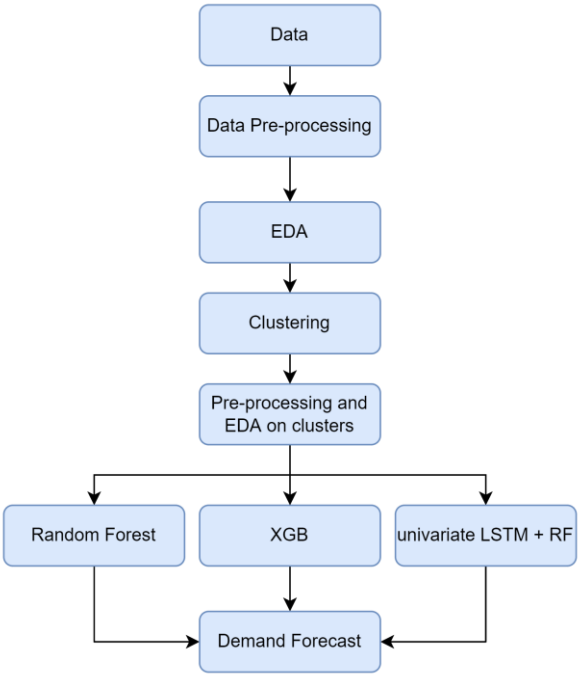


Figure 1: Flow chart of the research

The modeling part consists of two subtasks: Clustering and demand forecasting. The models used to perform demand forecasting are Random Forest, XGBoost, and a combination of univariate LSTM and Random Forest.

3.1 Data Acquisition

The data for this research was acquired from Kaggle. The data contains the sales details of the corporation favorita store in Ecuador. It contains a folder with different files like train, test, transactions, items, stores, holiday events, and oil prices. The description of the files is given in the configuration manual.

3.2 Data Pre-processing

- The data was read in separate files with only the required fields.
- The holiday's data contains a few days where the holiday was transferred to some other day or it was an event, bridge between two holidays, etc. These days were replaced as holidays and the day from which it was originally transferred, was replaced as a workday.
- As a first step, the train data was subsampled with only the sales data for the year 2017 to reduce the memory overhead by using the Datetime and ‘loc’ function.
- The subsampled data was merged with the store's data based on store number
- Next, this merged data was again merged with the holiday's data based on the date column
- This data was used for the clustering process based on unit sales
- For clustering using transaction details, the same process was repeated but with the transaction data in place of the training data.
- For creating separate clusters based on the results obtained from the clustering algorithm, the train data of specific stores in the cluster alone was read and the same steps were repeated as described above. But in this case, the weather data that was pulled from outside and the oil price data, and the items were also merged into the previously merged data.
- The columns with improper names were renamed and the null values were checked and handled
- The categorical variables were encoded using the Label Encoded
- Based on the requirement this data was further modified and used.

4 Design Specification

Four different models were used in this research as mentioned below.

4.1 Clustering

Hierarchical clustering was used in this research to separate and group stores with similar attributes. This process was done with the aggregates of transaction and unit sales. Hierarchical clustering is an agglomerative algorithm that works by generating clusters from each data point and merges them based on the distance between them. This iterative process repeats till all the data points are grouped. The output of the clustering is represented through Dendrogram.

4.2 Random Forest

Random forest is an ensemble model that builds multiple decision trees out of different samples and averages the results of all the trees to produce the outcome, called bagging or bootstrap aggregation. Random Forest has the edge to perform better due to its nature to handle non-temporal aspects of the data.

4.3 XGBoost

XGBoost is another ensemble gradient boosting model which is more straightforward. This has the edge over the random forest in terms of "similarity score", can handle imbalanced datasets, gives more importance to functional space while random forest relies on hyperparameters to optimize the models.

4.4 Long Short-Term Memory (LSTM) & Random Forest (RF)

It is important to understand that demand forecasting involves temporal data and non-temporal data together. LSTM networks are one of the best deep learning models in handling temporal data. Deep learning is a subset within machine learning and is more complex and robust in handling patterns present within the data. In this research, univariate LSTM is used along with random forest to handle the above-mentioned aspects. This approach is used in this research based on the outcomes produced by Punia et al. (2020).

5 Implementation

This research was implemented using Python in Anaconda Navigator. After the necessary data pre-processing was done, as mentioned in section 3.2. Necessary EDA was done to understand the distribution of the data and the statistical features. All the steps were implemented as mentioned in figure 1.

5.1. Clustering

Hierarchical clustering results are represented in the diagrams below. Figure 2 represents the dendrograms of clustering on the aggregated transaction data based on store number and the type of day (whether it was a holiday or not). The first dendrogram was constructed using both standard deviations and the mean of the aggregation. The second dendrogram is based on only the mean of the aggregated values.

The dendrograms are pretty straightforward. But the results do not seem to indicate the separation of clusters. The diagram below is the clustering results based on the aggregation of transactions done store-wise for the days in a week. Yet again, there is no clear separation of the clusters. This is interpreted based on the distance between consecutive vertical lines. Since the transaction data is very generalized, the behaviour of each store cannot be captured clearly by the algorithm.

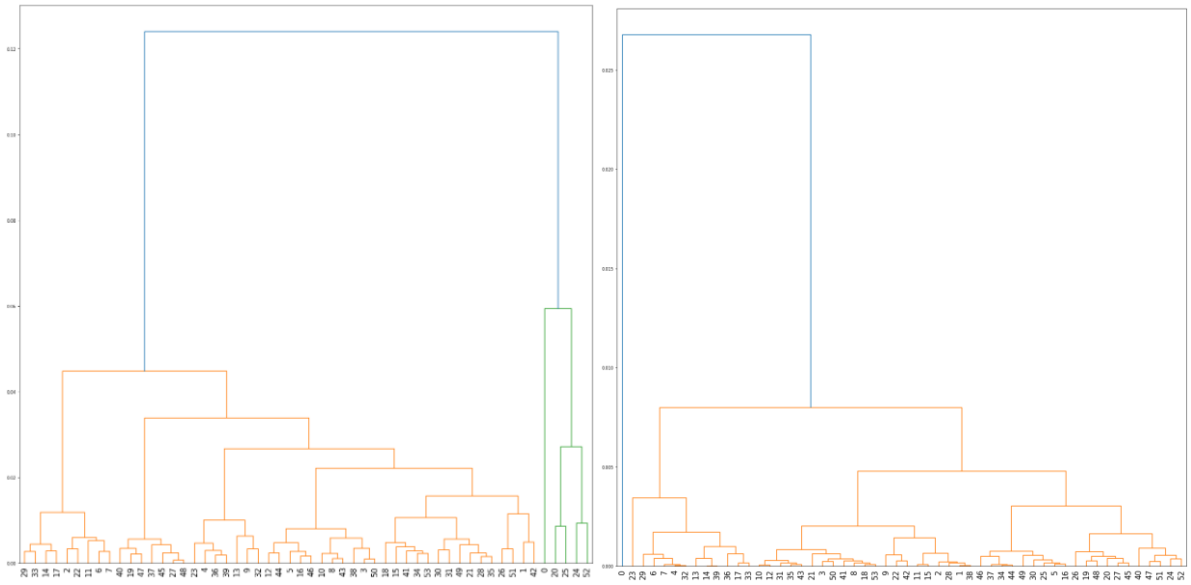


Figure 2: Dendrogram of aggregated transactions based on store and type of day

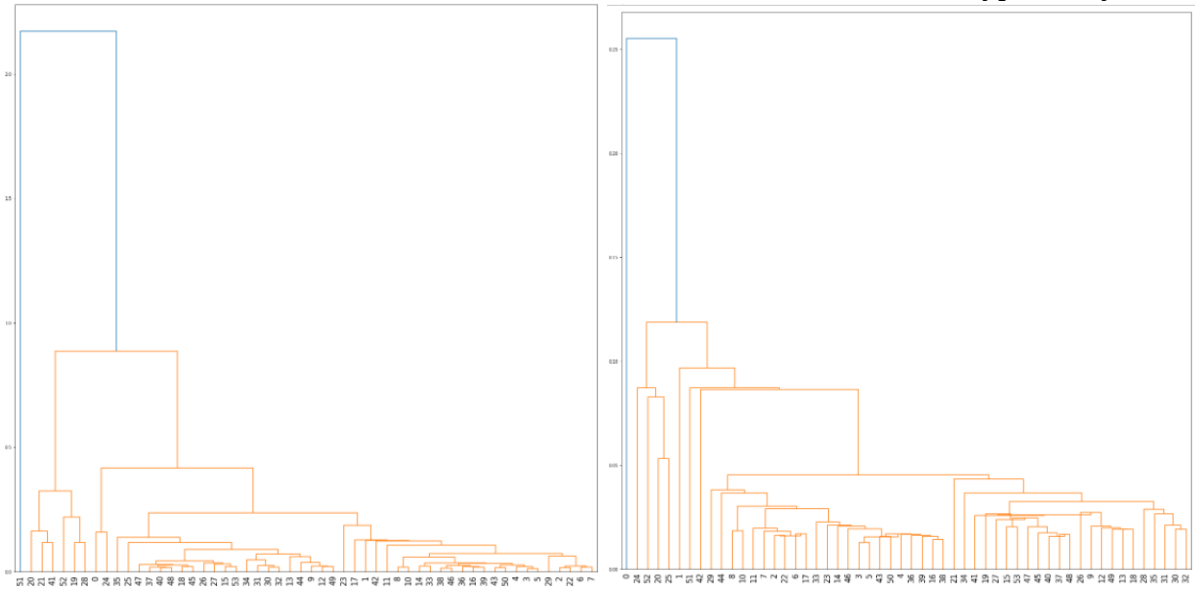


Figure 3: Dendrogram of aggregated transactions based on store and days in a week

It makes more sense to try out clustering based on the unit sales for each product across all the stores. The clustering results in figure 4 are very clear and there is a clear distinction between clusters identified by the algorithm. Too many clusters can disrupt the similarity found by the algorithm and too few clusters can fail to capture the underlying pattern. So based on the dendrogram results, the stores were clustered into 6 distinct clusters as identified by the algorithm. This shows that few stores are similar to other stores based on the sale behaviour of the products which is influenced by factors like promotions, holidays, oil price, weather, etc.

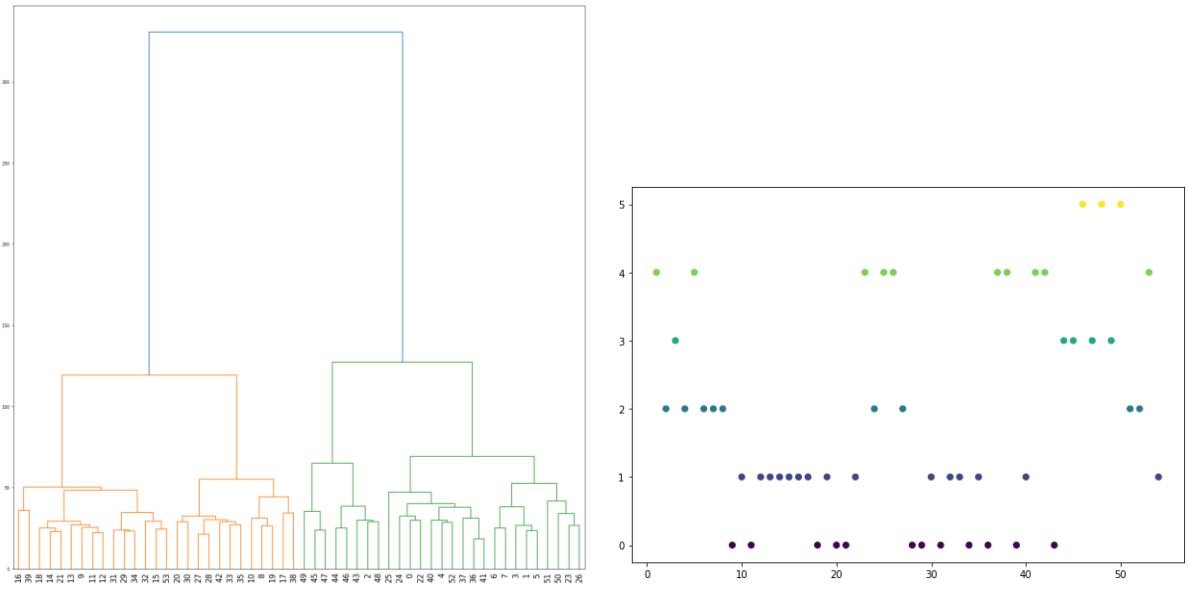


Figure 4: Dendrogram of clustering based on unit sales in each store

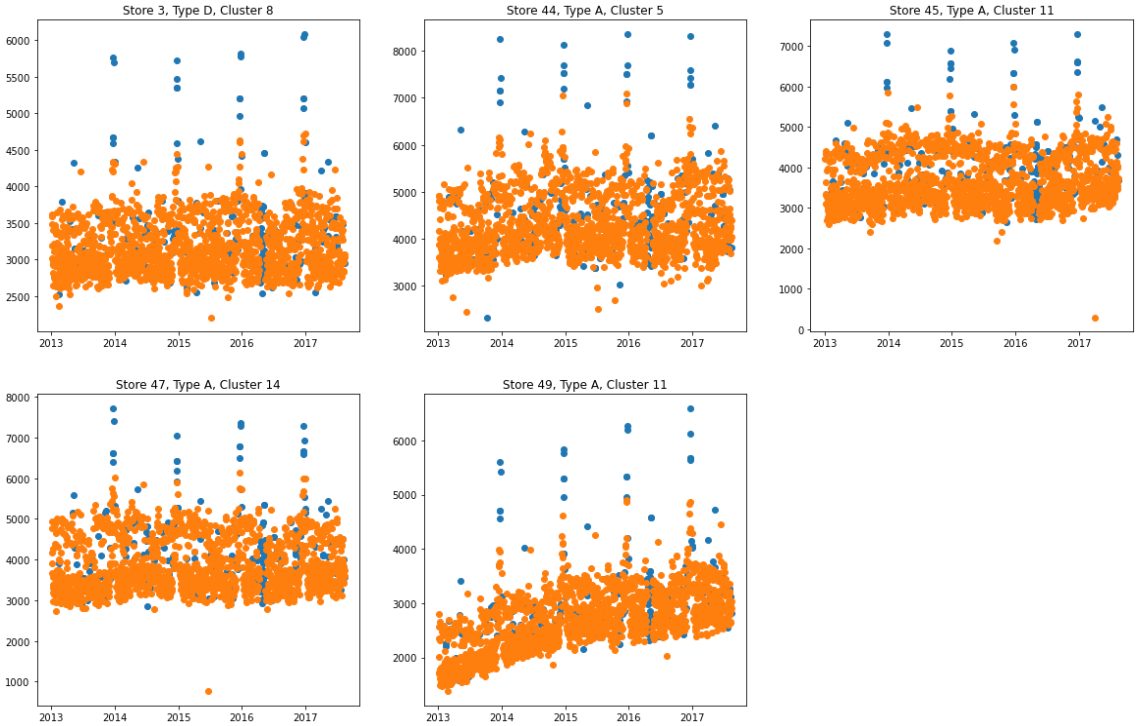


Figure 5: Cluster 3

The clusters identified are visualized in the diagrams above. In this research, two clusters, cluster 3 with the highest number of transactions (shop 45) and cluster 5 were used for comparison of factors influencing the demand. In Figure 5, the transaction behaviour seems to be either stable or increased linearly since 2013. Also, there seem to be gaps during the new year. But this property is visible in other clusters as well. But another reason for this clustering might be the range of transactions. All the stores have more transactions between 2000 and 5000.

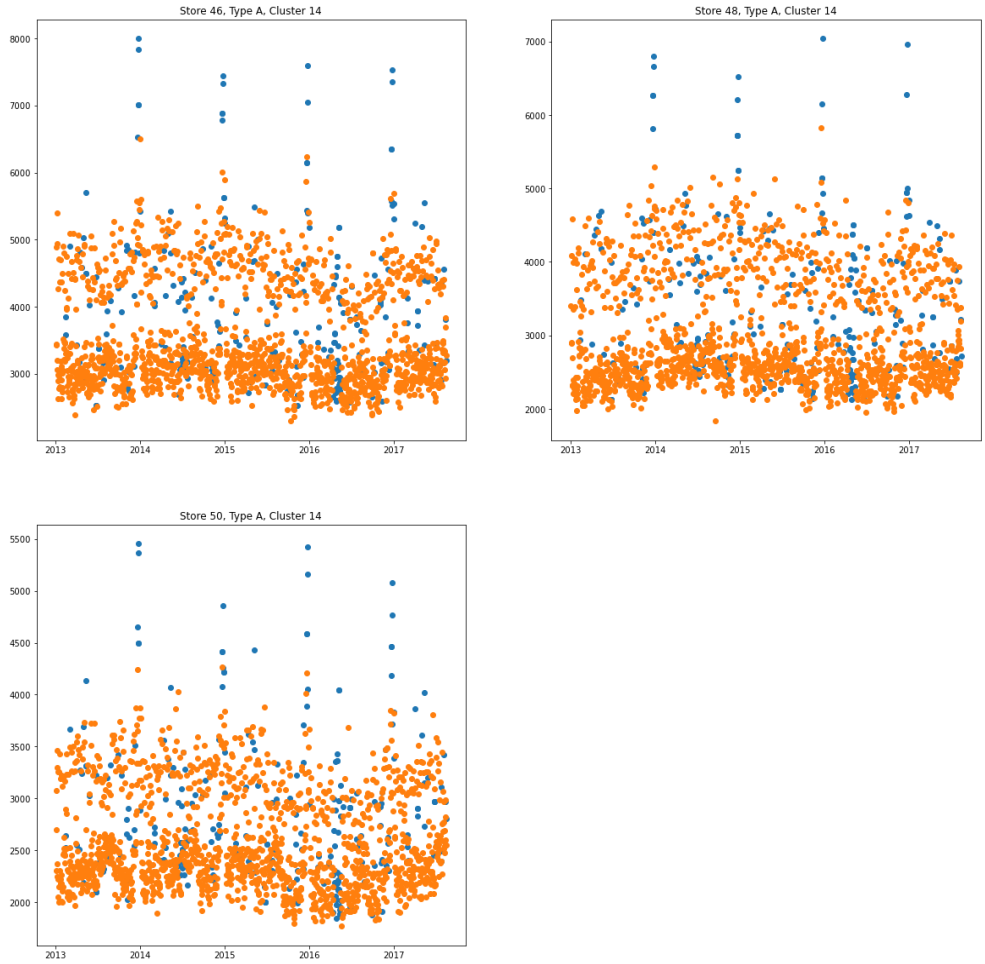


Figure 6: Cluster 5

In the case of cluster 5, three seems to be very little transaction and there is a visible break in between the levels. This might indicate that these shops have more sales during a particular time of the week compared to other days. In both the clusters, one similarity is that the transaction volume is more during the holidays represented by the blue dots compared to work days represented by orange dots. Though cluster 3 had linearly increasing transaction behavior, cluster 5 has some kind of a tilde (~) pattern. This drop in sales could be due to the earthquake that hit Ecuador in mid-2016. These shops could have been located in the most affected areas, explaining the sudden drop. It took several weeks for the country to gain back the pace.

5.2. Random Forest

The Random Forest model was instantiated with the parameters mentioned in Figure 7. It was kept minimal to reduce the training time. But even with this simple parameter, the model was able to produce very good results.

```
rf = RandomForestRegressor(n_jobs = -1, n_estimators = 15,min_samples_split = 10,random_state=0)
y = rf.fit(Xg_train, Yg_train)
print('Train accuracy',rf.score(Xg_train,Yg_train))
```

Figure 7: Random Forest Regressor

5.3. XGBoost

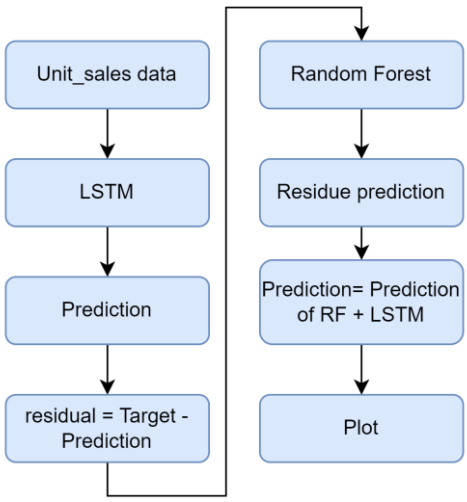
The XGBoost model was improved with better parameters. The gbtree booster was used to train the model with an eta of 0.3, maximum depth of 25, boost rounds of 20, and subsample of 0.9.

```
params = {"objective": "reg:linear",
          "booster" : "gbtree",
          "eta": 0.3,
          "max_depth": 25,
          "subsample": 0.9,
          "colsample_bytree": 0.7,
          "silent": 1,
          "seed": 1301
          }
num_boost_round = 20
watchlist = [(dtrain, 'train'), (dvalid, 'eval')]
```

```
gbm = xgb.train(params, dtrain, num_boost_round, evals = watchlist,
  early_stopping_rounds = 20, feval = rmspe_xg, verbose_eval = True)
```

Figure 8: XGBoost

5.4. LSTM & RF



Model: "sequential_27"

Layer (type)	Output Shape	Param #
lstm_27 (LSTM)	(None, 16)	1536
dense_139 (Dense)	(None, 8)	136
dropout_114 (Dropout)	(None, 8)	0
dense_140 (Dense)	(None, 8)	72
dropout_115 (Dropout)	(None, 8)	0
dense_141 (Dense)	(None, 4)	36
dropout_116 (Dropout)	(None, 4)	0
dense_142 (Dense)	(None, 2)	10
dropout_117 (Dropout)	(None, 2)	0
dense_143 (Dense)	(None, 1)	3

Total params: 1,793
Trainable params: 1,793
Non-trainable params: 0

Figure 9: Flow Chart of LSTM+RF and LSTM network

The LSTM+RF model is combinational. The LSTM model is composed of the LSTM layer followed by a combination of dense and drop-out layers with different neurons. The initial LSTM layer contains 16 neurons, meaning 16 input nodes, and is reduced to half in the consecutive layers with a dense output layer with 1 neuron as the final output.

- This model is used to perform univariate demand prediction($\hat{X}^1_t$).
- The predicted forecast is compared with the actual demand value to calculate the residuals ($r_t = X_t - \hat{X}^1_t$)

- These residuals are considered as the dependent variable for the random forest model and the final forecast ($\hat{X}^2_{t,r}$) is a sum of the forecast made by LSTM and the prediction on a residue by the random forest ($\hat{X}_t = \hat{X}^1_t + \hat{X}^2_{t,r}$).

6 Evaluation

6.1 Cluster 3

The Evaluation of the models was done using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) as mentioned in Table 1. Out of all the models, XGBoost seems to perform well in terms of error rate with a low RMSE rate of 0.41 and MAE of 0.31. And predicts well for the test set as well. The results of prediction by the models are represented in figure 10. Out of all the models, XGBoost seems to have little residual value compared to other models.

Table 1: Cluster 3 Results

Model	Train RMSE	Test RMSE	Train MAE	Test MAE
Random Forest	0.4400	0.5499	0.3315	0.4192
XGBoost	0.4198	0.5327	0.3099	0.4054
LSTM+RF	1.1881	1.0465	0.5245	0.4654

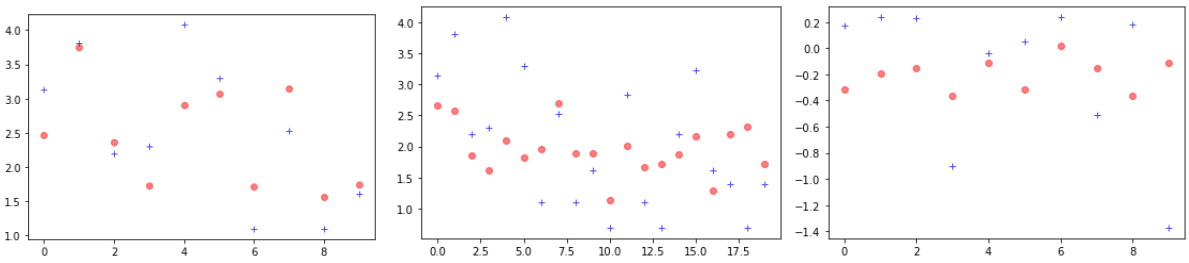


Figure 10: Prediction Vs Actual values of Random Forest, XGBoost and LSTM+RF

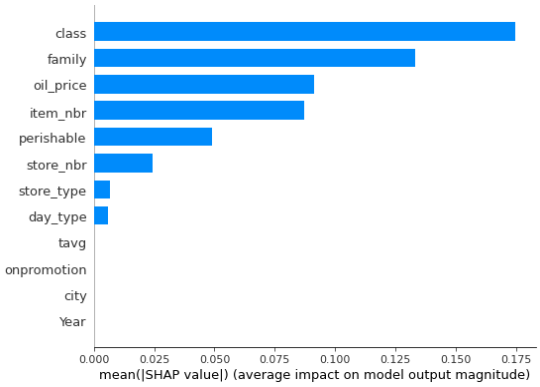


Figure 11: SHAP plot for cluster 3

In terms of the factors influencing the sales, for the stores in cluster 3, the class and family of the product, oil price, the item number (representing the actual product), and

whether or not the item is perishable are the top 5 factors influencing the demand. This is represented very clearly in Figure 11. More details on how each factor is disturbing the sales value are provided in the configuration manual.

6.2 Cluster 5

The results for cluster 5 are quite similar to cluster 3. XGBoost performs well in terms of both RMSE and MAE. The prediction Vs Actual forecast graph in figure 12 shows the prediction by random forest is more accurate. But this is just for the first 10 product sales and when the whole 6 lakh data points are considered, XGBoost seems to perform the best with an RMSE of 0.54 and MAE of 0.41 on the test data.

Table 2: Cluster 5 Results

Model	Train RMSE	Test RMSE	Train MAE	Test MAE
Random Forest	0.4460	0.5540	0.3374	0.4246
XGBoost	0.4411	0.5412	0.3306	0.4172
LSTM+RF	1.1915	1.0485	0.4917	0.4508

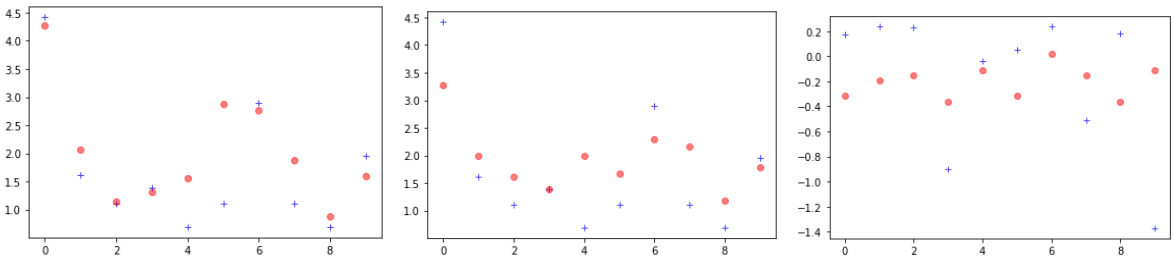


Figure 12: Prediction Vs Actual values of Random Forest, XGBoost and LSTM+RF

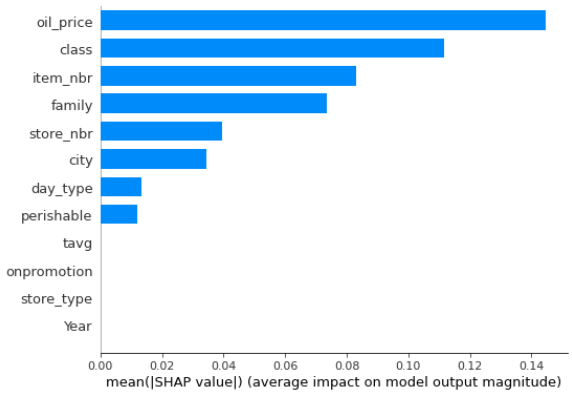


Figure 13: SHAP plot for cluster 5

In terms of the most influential factors (Figure 13), the oil price, class of the product, item number and family of the product, and the store number are the most contributing. This might be due to the presence of an oil factory nearby since Ecuador is an oil-dependent country.

6.3 Discussion

Multiple machine learning methods were used in this research. More specifically:

- Hierarchical Clustering on transactions of all stores for every day and the unit sales of multiple products in each store for every day in the year 2017.
- Two clusters: Cluster 3 and Cluster 5 were chosen to check for the influence of the external factors
- Random forest is best in considering the non-temporal aspects of the data. Considering this reason as mentioned by Punia et al. (2020), the random forest ensemble regressor was first implemented on both clusters. The results were pretty good on both the train and test sets. On average, the RMSE was obtained around 0.55 and MAE of 0.42 on the test data for both the clusters.
- XGBoost is more robust than random forest as mentioned in section 4.3. In terms of error rate, it performed well than random forest. The XGBoost model also has the added advantage to run faster with the SHAP library. The average RMSE obtained was around 0.53 and MAE of 0.41 on the test set. The most influencing factors were visualized using the SHAP library in the previous section. A more elaborate description of how each factor is affecting the demand is provided in the configuration manual.
- The extended combinational model of LSTM and RF by Punia et al. (2020) is very effective in terms of handling huge datasets, handling both temporal and non-temporal aspects of the data. Due to the lack of advanced systems, only one iteration with 5 epochs was run. Even with this, the model produced good results. An average RMSE of 1.04 and MAE of 0.45 were obtained which is not very large than the other two models.

The factors indeed influence the demand differently for a different group of stores. This shows care must be taken before generalizing the results obtained for a subset of data. At different levels, the stakeholders can use the model within the cluster to forecast demand and adjust the discrepancies based on how each factor is disturbing the sales.

Though it is hard to build a common or generalized model to work for all data, clustering can help in identifying stores that behave similarly as mentioned in the research by İşlek & Gündüz Öğüdücü, (2015) and Chu & Zhang, (2003). Other note-worthy research gaps were the consideration of new stores, macroeconomic factors like GDP, oil price, population, etc, and the influence of these factors on both offline and online data as mentioned by Fildes, et al. (2019). In this research, the new store and macro-economic aspects were included. Relatively store 52 was a newly opened store with very little data, but its behaviour in terms of sales and transactions is similar to other stores in cluster 2. The macroeconomic factor considered here is the oil price since Ecuador is an oil-dependent country. There have been very good models proposed to handle the complexity of the multivariate data. Along with clustering, the combinational LGBM and LSTM model proposed by Weng et al., (2019) and the CNN WaveNet proposed by Kechyn et al., (2018) could be implemented to improve the accuracy.

7 Conclusion and Future Work

This research aimed to analyze the influence of external factors in the sales and demand of products in retail and whether these differ for groups of stores that exhibit similar behaviour. The main idea was to first group the stores into different clusters using hierarchical clustering and then use multiple ensemble and combinational models to check how the external factors influence the demand. This is quite important for the stakeholders in terms of maintaining the availability of products, avoiding over or understocking, attain a demand-sales balance. The Hierarchical clustering was done on various aggregations on the transaction data and unit sales data. Two of the identified clusters were used as the data for the models. Out of all the models: Random Forest, XGBoost, and combination of LSTM & RF, the XGBoost regressor performed the best with a very low average RMSE of 0.536 and MAE of 0.411. The limitations of the proposed approach is that it is not able to scale properly to the range of data points and the predictions are distributed around a constant range. This might be caused due to underfitting of the data which is a direct effect of time and system constraints.

Trying different standardization and normalization methods could be a potential future work to improve the accuracy of the proposed approach. Also, a better and more complex LSTM network could be used to train the univariate series data along with hyperparameter tuning to identify the best set of parameters.

References

Carbonneau, R., Laframboise, K., & Vahidov, R., (2008), 'Application of machine learning techniques for supply chain demand forecasting', *European Journal Of Operational Research, 184*(3), pp. 1140-1154.

Chawla, A., Singh, A., Lamba, A., Gangwani, N. and Soni, U., 2018. Demand Forecasting Using Artificial Neural Networks—A Case Study of American Retail Corporation. *Advances in Intelligent Systems and Computing*, pp.79-89.

Chu, C. and Zhang, G., (2003), ' A comparative study of linear and nonlinear models for aggregate retail sales forecasting', *International Journal of Production Economics*, 86(3), pp.217-231.

Fildes, R., Ma, S. and Kolassa, S., (2019), 'Retail forecasting: Research and practice', *International Journal of Forecasting*, pp.1-36.

Guo, Z., Wong, W., & Li, M., (2013), 'A multivariate intelligent decision-making model for retail sales forecasting', *Decision Support Systems, 55*(1), pp. 247-255.

İşlek, İ., and Gündüz Öğüdücü, Ş., (2015), 'A retail demand forecasting model based on data mining techniques ', *2015 IEEE 24th International Symposium on Industrial Electronics (ISIE)*, pp. 55-60.

Jha, B. and Pande, S., (2021),' Time Series Forecasting Model for Supermarket Sales using FB-Prophet', *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 547-554.

Kechyn, G., Yu, L., Zang, Y., & Kechyn, S, (2018), 'Sales forecasting using WaveNet within the framework of the Kaggle competition', *ArXiv, abs/1803.04037*.

Min, H.,(2010), ‘Artificial intelligence in supply chain management: theory and applications’, *International Journal of Logistics Research and Applications: A Leading Journal of Supply Chain Management*, 13(1), pp. 13-39.

Ni, D., Xiao, Z., & Lim, M., (2019), ‘ A systematic review of the research trends of machine learning in supply chain management ’, *International Journal Of Machine Learning And Cybernetics*, *11*(7), pp. 1463-1482.

Pacella, M. and Papadia, G., 2021. Evaluation of deep learning with long short-term memory networks for time series forecasting in supply chain management. *Procedia CIRP*, 99, pp.604-609.

Patak, M., Branska, L. & Pecinova, Z., (2015), ‘Demand Forecasting in Retail Grocery Stores in the Czech Republic’, *2nd International Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM2015*.

Pereira Da Veiga, C., Rita Pereira Da Veiga, C., Catapan, A., Tortato, U., & Vieira Da Silva, W., (2021), ‘Demand forecasting in food retail: a comparison between the HoltWinters and ARIMA models’, *WSEAS Transactions On Business And Economics*, *11*(1), pp. 608-614.

Punia, S., Nikolopoulos, K., Singh, S., Madaan, J., & Litsiou, K., (2020), ‘ Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail’, *International Journal Of Production Research*, *58*(16), pp. 4964-4979.

Ramanathan, U. and Muyldermans, L., (2010), ‘Identifying demand factors for promotional planning and forecasting: A case of a soft drink company in the UK’, *International Journal of Production Economics*, 128(2), pp.538-545.

Ren, S., Chan, H., & Siqin, T., (2019), ‘ Demand forecasting in retail operations for fashionable products: methods, practices, and real case study ’, *Annals Of Operations Research*, *291*(1-2), pp. 761-777.

Tarallo, E., Akabane, G., Shimabukuro, C., Mello, J. and Amancio, D., 2019. Machine Learning in Predicting Demand for Fast-Moving Consumer Goods: An Exploratory Research. *IFAC-PapersOnLine*, 52(13), pp.737-742.

Taylor, S., & Letham, B., (2018), ‘ Forecasting at Scale ’, *The American Statistician*, *71*(1), pp. 37-45.

Wanchoo, K., (2019),‘ Retail Demand Forecasting: a Comparison between Deep Neural Network and Gradient Boosting Method for Univariate Time Series’, *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, pp.1-5.

Weng, T., Liu, W. and Xiao, J., (2019), ‘Supply chain sales forecasting based on lightGBM and LSTM combination model ’, *Industrial Management & Data Systems*, 120(2), pp.265-279.