Contents

Abstract

An essential daily task in running retail stores is arranging a balanced order, especially for perishable goods. Making wise choices when selecting the proper lot size helps preserve customer satisfaction, boost shop earnings, and decrease perishable food waste. XGB and LGB have been shown to be a successful pattern recognition and time series event forecasting technique, but not much in perishable food sector. This study compared XGB and LGB with Moving Average and Neural Network on perishable food sales prediction by integrating dynamic factors, such as weather, promotion and oil price. The experimental results show that both tree-based models are more accurate and efficient than other two models in perishable food sales forecasting.

Keywords: Perishable food, Sales Forecasting, Machine Learning, Predictive Modelling, FMCG, Decision Support

1. Introduction

Sales prediction is a crucial tool for businesses involved in the production, distribution, or retail of goods in today's competitive and dynamic business climate. Long-term predictions can assist with corporate development decisions, whereas short-term forecasts mostly aid in production planning and inventory management. Because many items in the food industry have short shelf lives and could lose money in the event of shortages or surpluses (J. Gustavsson *et al.*, 2011), sales predictions are particularly crucial for the perishable food sector.

Ordering too little results in missed opportunities, and ordering too much results in wasted merchandise (Corsten and Gruen, 2003; Parfitt, Barthel and MacNaughton, 2010; Peters, 2012; Ehrenthal and Stölzle, 2013; Suban and Bogataj, 2015; Huber and Stuckenschmidt, 2020; Shang *et al.*, 2020). However, the demand for food is also continuously changing because of factors like costs, promotions, shifting customer preferences, and changing weather. With the insufficient application of traditional technologies such as CPFR or EPOS (Dqg *et al.*, no date; Ning *et al.*, 2009; Taylor and Fearne, 2009; Minner and Transchel, 2017), machine learning algorithms have been treated as a solution to increase the prediction accuracy by making use of the extensive resources of sales data and related information. Despite the widespread application in other areas in terms of the tree models (Stamell *et al.*, no date; Marcos Roberto Machado, Salma Karray and Ivaldo Tributino de Sousa, 2019; Lee *et al.*, 2020), the applications on perishable foods sales prediction are fairly limited.

As a result, this research compares the performance of neural networks and moving average with two tree-based models (XGB, LGB) based on a dataset from a large Ecuadorian-based grocery retailer, called Corporacion Favorita. This paper's background discusses the meaning of food sales forecasting and the drawbacks of current systems. By comparing several machine learning models, section 2 emphasizes the significance of using tree models on perishable food sales forecasting. The experiment design and reasons are fully described in sections 3 and 4. The experiment's findings are then examined, and recommendations for future research are made in section 6 and 7.

2 Background

2.1 Importance of perishable food demand forecasting

One of the most difficult aspects of managing the food supply chain is ensuring sustainability. United Nations Food and Agriculture Organization (FAO) estimates that in Europe and the United States, edible product production and food waste per person per year are around 290 and 900 kilograms respectively, while they are 120–170 and 460 kilograms in South and Southeast Asia. Additionally, industrialized nations squander about 40% of the food produced during the retailing and consumer consuming stages, which is nearly equivalent to the total net food output of Sub-Saharan Africa (J. Gustavsson *et al.*, 2011). Although many studies provided different results on food waste in different supply chain stages because of the differences between measurements and approaches, products like fruits, vegetables, and bakery goods combined account for nearly 30% of all food produced for human use that is not wasted, together with the entire supply chain from the point of production to the point of consumption (Parfitt, Barthel and MacNaughton, 2010; J. Gustavsson *et al.*, 2011).

Fresh fruits, vegetables, and bakery products could be categorized as perishable food, as their quality is very easily influenced by the ambient storage environment in a short time, or they have an expiration date that makes restocking problematic (van Donselaar *et al.*, 2006). Perishable food sales are becoming increasingly essential for global grocery merchants, as the sales account for almost 50% of the industry's turnover in Western Europe and North America (Thron, Nagy and Wassan, 2007). Aside from pricing strategy, perishable items are also the major driver via which merchants may build competitive advantages to acquire extra customers. The significance of perishables is further underscored by (Heller, 2002), who concluded that the quality of perishable products is gradually becoming one of the key reasons that customers favor one supermarket over another.

2.2 Demand forecasting mismatching

Retailers seek to provide high levels of service for all products to maintain their competitive advantage. This causes a surplus of leftovers since supply exceeds demand; thus, these wastes must be transported to expensive disposal or recycling facilities rather

than being given to customers (Suban and Bogataj, 2015). According to (Parfitt, Barthel and MacNaughton, 2010), by-products and unsold prepared food products make up a significant component of the UK's 14 megatons (Mt) of garbage, and 5 billion pounds of returned goods are anticipated to wind up in landfills. Other than being disposed of in landfills, others could be given back to producers. For the processing and disposal of consumer returns, the retail industries spend more than $40 billion annually, while OEMs spend two to three times as much on reverse logistics as they do on outbound shipping for comparable goods (Shang *et al.*, 2020).

On the other hand, financial loss of out-of-stock (OOS) due to the demand underestimating is far more difficult to assess since client behavior is unpredictable. According to (Ehrenthal and Stölzle, 2013), products that cannot be purchased by a client at any given time is considered out-of-stock(Corsten and Gruen, 2003). OOS not only results in a 4% immediate revenue loss (Gruen, 2002), but also jeopardizes future sales and harms customer loyalty (Zinn and Liu, 2008). Customers transfer retailers, substitute things, postpone purchases, or do not buy anything if the desired item is not available (Group and Gijsbrechts, 2000). As a result, OOS causes increased customer turnover, lower satisfaction, and lower loyalty (Huber, Gossmann and Stuckenschmidt, 2017). Since the actual contemporary demand of product is hard to be reflected due to the skewed historical sales data, OOS not only obstructs sales planning, but also lowers the precision of forecasting. These consequences are not restricted to the article categories that are directly affected. (Ehrenthal and Stölzle, 2013) report that compared to other categories, OOS of fresh items is the preliminary reason of causing the greatest turnover loss.

The described consequences emphasized the competitive advantage that a retailer could gain by avoiding OOS or demand overestimation. Therefore, understanding the causes is necessary because it identifies difficulties that need to be addressed to attain a higher level of service, and the bullwhip effect is one of the key causes.

2.3 Bullwhip effects and potential solutions

The bullwhip effect was first used to explain the order variance amplification phenomenon between Procter & Gamble (P&G) and its suppliers in the 1990s. (Forrester, 1997) used the “industrial dynamics” technique to first define the variance amplification effect. In

presenting statistics from the beer game spanning two decades, (Sterman, 1989) attributed the amplification to players' propensity to ignore inventory-on-order (things ordered but not yet received), a tendency known as “irrational behavior”. While (Lee, Padmanabhan and Whang, 1997) assumed bullwhip is the outcome of entirely rational conduct combined with four other factors: demand signal processing, rationing game, order batching, and price volatility.

Solutions such as information sharing and Vendor-Manage-Inventory (VMI) have been suggested to mitigate the effect of the bullwhip effect (Lee, Padmanabhan and Whang, 2004). As implied by the name, vendors have access to real-time inventory data from retailers, and vendors choose the proper inventory level for retailers. It is because retailers, rather than manufacturers, are usually able to gather more market demand information as the final component of the supply chain (Aviv, 2007). VMI directly provides competitive advantages for retailers as well as vendors by enabling replenishment planning synchronization (Sari, 2008). However, in most VMI programs, the unique knowledge of the retailer could not be joined into inventory decisions due to outdated or inaccurate data, as well as inadequate information technology and lack of mutual trust (Sari, 2007).

For example, to address the issue of demand management along the supply chain, many large merchants, particularly those in the food business, have gathered EPOS data and sent it to suppliers via an electronic link. However, in none of the chains examined, EPOS data was ever shared farther upstream than the retailer's immediate supplier. It has been found difficult to gather correct information due to the market's complexity and technological limitations. In addition, due to ineffective algorithms that are incapable of processing vast amounts of data, those data were not completely utilized, as several smaller providers have proven (Taylor and Fearne, 2009).

CPFR (collaborative forecasting and replenishment), on the other hand, has been seen as a solution to overcome the drawbacks of VMI by jointly developing demand forecasts, production, and purchasing plan, which further requires more mutual trust(Sari, 2008). However, due to the separate inventory management and associated tactics (Wang, 2011), retailers prefer alternatives based on predicted order amounts, which simulates FIFO depletion and lost sales; Suppliers, on the other hand, desire higher order numbers and less

variation in merchant orders (Minner and Transchel, 2017). Retailers' need for flexibility and efficiency led wholesalers to increase their stockpiles to have enough stock, resulting in higher purchasing and holding expenses. However, a lack of items to meet store demand will result in shortfall charges (Ning *et al.*, 2009).

In this situation, another possibility to mitigate the effect of the bullwhip effect is to increase the accuracy of demand forecast, and machine learning techniques have been a valuable tool for supply chain management (Dqg *et al.*, no date)

2.4 Literature Review

The current demand forecasting methods could be generally divided into two parts, which are the time-series model (TSM) and machine learning algorithms (MLA) (Song *et al.*, 2016; Ghalehkhondabi *et al.*, 2017). TSMs range from exponential smoothing to ARIMA and its derivation models, which have been widely applied on the future sales prediction by using the historical data. However, due to the limitation of its linear assumption, TSMs are now usually used as univariate prediction baseline without taking external factors.

Instead of TSM, another branch of forecasting is MLA, which includes but is not only limited to Artificial Neural Network (ANN), Long-short term memory (LSTM), Support vectors (SVM) and etc. Many studies have proved the higher efficiency of MLA compared with TSM. Businesses have used statistical methods and tools to calculate product demand, with historical sales series acting as the main data source.(Carbonneau, Laframboise and Vahidov, 2008) compared the performance of advanced machine learning algorithms, including NN, recurrent neural networks (RNN), and support vector machines (SVM), with traditional predicting techniques, such as Moving average (MA) and naïve forecasting techniques by using the datasets from a Canadian Foundries order. The finding implied that RNN and SVM showed a better performance than other techniques. Similar results were also observed by (Chen and Ou, 2008), who proposed a neural network model but further integrated dynamic factors, such as weather and alternative goods promotions, to forecast perishable food sales in convenience stores. The experimental findings demonstrated that the forecasts generated by machine learning algorithms are more accurate than those of the conventional time-series model (MA and ARIMA). The benefits of machine learning techniques have been further confirmed by (Kandananond, 2012), who forecasted the

demand of 6 different consumer products of a Thailand company. It concluded that in terms of MAPE, in comparison to ANN and ARIMA, SVM performs better across all categories, particularly for consumer items and complicated product hierarchies, displaying fewer error deviations. However, (Lee *et al.*, 2012) found that Back-propagation Neural Network (BPNN) only performed best on the prediction of sushi, while Logistic Regression (LR) got the highest value for all others in general. The conclusion is contradicted by the past related literature, and it may be because of the subjective selection of the number of nodes and layers. In addition, random factors have also been excluded from the analysis.

Excepting one, most literatures above proved the efficiency of MLA (ANN, SVM) on prediction compared with TSM. It is worth noting that most literatures not limited to the historical sales data, but also included external factors, and have proved the importance of these variables (Arif *et al.*, 2020). According to (Guo, Wong and Li, 2013), there are various of factors influencing the sales of products, including the inherent property and there is no fixed pattern of time-series data for estimation given the current economic climate, not to mention the linear correlation between these external factors with the target value(Tarallo *et al.*, 2019). This could be one of the reasons why MLA performed better, as it allows other input variables not only limited to the existing time-series data(Tsoumakas, 2019). (Qu *et al.*, 2017) used several sources of data including internal (historical sales and promotion) and external (holiday and regional factor), and the result proved to have better sales prediction.

Realizing the importance of external variables, the dimensionality of data is also increasing, and the limitation of some MLA became to appear. (Stamell *et al.*, no date) pointed out that though NN can process the highly complexed non-linear function, it also has several challenges. First of all, the black-box processing model led to the minimum interpretability; the second is the instability (Lee *et al.*, 2012), and therefore, in order to minimize the negatives brough by the initialization and empirical selection of nodes and layers, substantial training is a necessary, which further lead to a question on model efficiency and computational cost. The cost of training a deep neural network is substantially higher than those of tree-based models. Another model with eye-catching performances that have been mentioned several times in above literatures is SVM. Though one of the advantages of

SVM is to use the relatively smaller training samples to process high-dimensionality data, when the number of features increased, it is very likely for SVM to get stuck in the curse of dimensionality (Melgani and Bruzzone, 2004). In addition, according to (Burges, 1998), if the amount of data is huge, the training time for SVM will take longer. However, tree-based model (Extra Tree) has been proved to have similar prediction performance and better stability but rather lower computational cost compared to SVR (Ahmad, Mourshed and Rezgui, 2018).

Compare to original tree models, such as Decision Tree (DT) or Random Forest (RF), GBDT are more advanced, as each decision tree is not independent. A new tree puts more weight on the misclassified samples that attained by previous trees, and the final result from the average prediction result using the total of all DT outcomes (Liang *et al.*, 2020). In tree-based model family, Extreme Gradient Boost (XGB) and Light Gradient Boost (LGB) are two more advanced ensembled tree-based models because of their strong capacity with well-performed prediction as well as lower computational cost (Chen and Guestrin, 2016), and have been discussed in many studies. (Lee *et al.*, 2020) applied NN and XGB on eddy-current data, and the result proved that XGB demonstrated highest prediction accuracy and Adj-R^2. The similar result has also been achieved by (Stamell *et al.*, no date), who concluded that XGB produced best prediction overall, with low-variance on unseen data, compared with NN and RF. (Marcos Roberto Machado, Salma Karray and Ivaldo Tributino de Sousa, 2019) then compared the LGB with XGB and by evaluating RMSE, and they found that LGB performed better on financial customer loyalty prediction.

Though the high efficiency of XGB and LGB have been proved with many literatures, researches that applied XGB and LGB on perishable food sales forecasting are very rare. As the summarization of literatures presented in table 1, the applications of these two models are most in financial, production industry or meteorology areas(Stamell *et al.*, no date; Marcos Roberto Machado, Salma Karray and Ivaldo Tributino de Sousa, 2019; Lee *et al.*, 2020).

To the best knowledge, the efficiency of XGB and LGB have not been academically discussed on perishable food sales prediction. Therefore, this dissertation aims to fill this gap by applying tree-based models on a dataset from a large Ecuadorian-based grocery

retailer. Since the literatures have also emphasized the importance and significance of external factors for prediction accuracy increasing, factors such as temperature, promotion and oil price, will also be take into consideration in this dissertation. By doing so, this dissertation could provide more choices for business to increase demand forecasting efficiency and therefore better manage inventories and decrease wastes. In addition, business could also realize and focus on the most important external factors that have influence on sales prediction.

Table 1. Methods comparison between literatures

Publish	Objection	Methods	Conclusion	Innovation	Limitation
Carbonneau et al., 2008	Canadian Foundries orders	NN, SVM, RNN, MA, Naïve forecasting	RNN, SVM have better performance	/	Functional object, not fast moving
Chen & Ou, 2008	Fresh foods in Convenient store	NN	NN performed better than MA, ARIMA	Exogenous factors (weather, promotions)	Non-combination approach
W. I. Lee et al., 2012	Fresh foods in Convenient store	BPNN, LR, MA	BPNN performed best on sushi, LR performed best overall	/	Non-combination approach
Arif et al., 2020	Information gathered from a superstore and a verities store about ten distinct products	KNN, SVM, GNB, RF, Decision Tree Classifier and regressions	Geographical factor has an impact on prediction; GNB has the best accuracy.	Exogenous factors (behaviour of the customer, the season, the event, the month, and the type of product)	Focus more on the geographical impact of demand forecasting
Qu et al., 2017	Semi-luxury	regression trees or random forests	The unpredictable revenue due to large variation in demand has been optimized by taking into account external factors like pricing, holiday, discounts, inventory, and other geographical characteristics.	Wider External factors (price, holiday, discounts, stock, and additional regional elements / Heuristic optimization methods	Limited to off-line and costly commodities

Stamell et al., n.d	surface ocean pCO2 from sparse in situ data with full coverage	Feed forward neural network, random forest, XGB	Considering the performance both on seen and unseen data, XGB performed best on limited data	/	Focused on the geoscientific area
K. Lee et al., 2020	eddy-current data	regularized linear regression, SVR, multi-layer neural network, RF, and XGB	ensemble training methods (RF, XGB) demonstrate the accurate prediction ability	/	Focus on the sheet metal materials
Marcos Roberto Machado et al., 2019	Customer loyalty for financial company	XGB, LGB	LGB performed better than XGB in terms of RMSE	Academically applied GBDT model on marketing financial products	Focus on the marketing financial products

3. Data Description

3.1 Sales Data Description

The sales dataset is provided by a large Ecuadorian-based grocery retailer, called Corporacion Favorita, and published on Kaggle, a well-known machine learning and data science community. The sales dataset is ranged from 2013.1.1 to 2017.8.15, with over 125 million sales records. The train dataset includes variables including dates, store, item details, and unit sales. Since Ecuador is a significantly oil-dependent nation and its economic health is extremely subject to the shocks of oil prices, additional files also provide supplemental information, such as holidays and oil prices. The detail information of each feature has been presented in table 2.

Table 2 Data description

Table	Variable	Definition	Types
Train.csv	ID	Unique Key	
	Unit_sales	Integer/float; negative means returns	
	Date	Day of transaction	2013/01/01-2017/8/15
	Store_nbr	code of each store	
	Item_nbr	Code of specific items	
	Onpromition	whether that item_nbr was on promotion for a specified date and store_nbr	True/False/NaN
Stores.csv	City	City of Ecuador	
	State	State of Ecuador	

	Type	Type of stores	A-D
	Cluster	Group of similar stores	1-17
Transactions.csv	Transactions	The count of sales transactions for each date, store_nbr combination	
Items.csv			'BREAD/BAKERY', 'DELI', 'DAIRY', 'EGGS', 'POULTRY', 'MEATS', 'SEAFOOD', 'PRODUCE', 'PREPARED FOODS'
	Family	Family of items belongs to	
	perishable	Items whether perishable or not	1/0
	Class	Class of items	337
Holiday.csv	type	Type of holiday	'Holiday', 'Transfer', 'Additional', 'Bridge', 'Work Day', 'Event'
	Oil_price	Daily oil price (Ecuador is an oil-dependent country)	
	Locale	Whether is regional or locale holiday	'Local', 'Regional', 'National'
	locale_name	Holiday of specific region	
	Description	Description of holiday	
	Transferred	Transferred holiday is more like calendar day, the actually celebrated day is when “type” is transfer	True/False

3.2 Daily Temperature data

The temperature records of each city have been collected from National Oceanic and Atmospheric Administration (NOAA), which is part of the science and technology sector of the US Department of Commerce. According to the sales dataset, retail stores mainly located in 22 cities, while there are only 9 meteorological stations in Ecuador. Therefore, cities that do not have individual station within its jurisdiction have to share the daily temperature records with other cities from a same station, and this sharing is based on the geographical distance between the city and weather station. Table 3 presented the source of temperature data of each city.

Table 3 Weather Station

Weather Station Code on NOAA	**City**
ECM00084036	Ibarra
ECM00084088	Quito
	Cayambe
	Latacunga

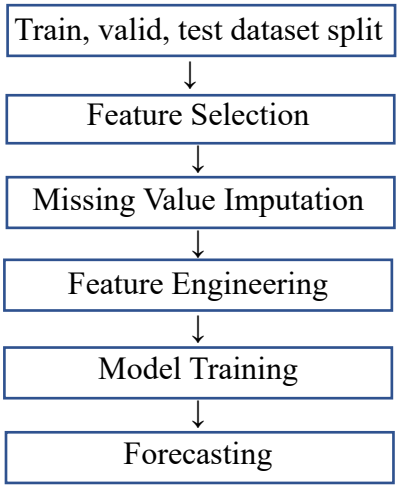
ECM00084270	Loja
	Machala
ECM00084226	Guayaquil
	Playas
	Cuenca
	EI Carmen
EC000000006	Daule
	Babahoyo
	Guaranda
	Quevedo
ECM00084179	Riobamba
	Ambato
	Libertad
	Puyo
ECM00084050	Santo Domingo
	Esmeraldas
ECM00084135	Salinas
	Manta

4. Methodology

4.1 Forecasting Process

The whole forecasting process consists several steps, figure 1 presented the framework.

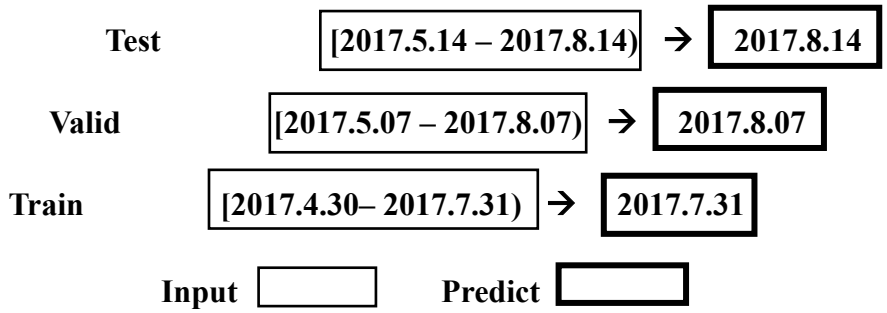
Figure 1. Framework of the forecasting model



4.2 Train, Valid, Test data splitting

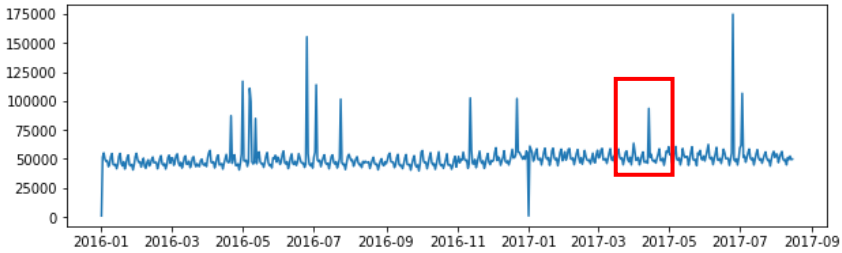
In order to predict the demand of product on 2017/8/14, the whole dataset have been split into three parts: train, valid and test batch, and the splitting have been presented as below:

Figure 2. Train, Valid, Test dataset splitting



As the figure 2 presented, the time span of each batch is 90 days, and it is to avoid the potential noise may be brought by the transaction fluctuation at the middle of April as figure 2 presented. Another point that is different from the most experiment, which usually use the historical data to predict the sales of next day, is that this research not only use the historical data but also include the promotion information on the prediction day.

Figure 3. Fluctuation of unit sales



4.3 Feature generation

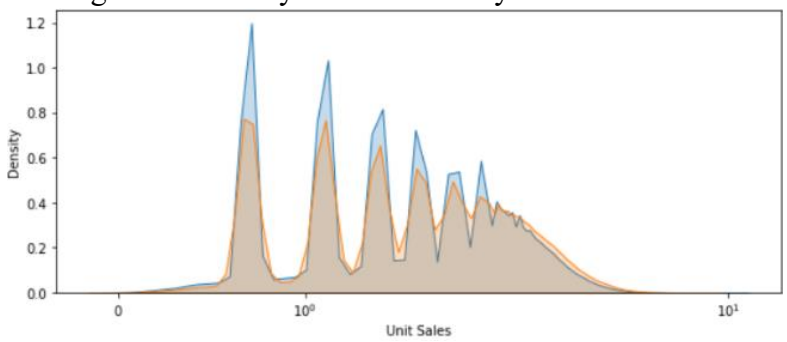
As table 2 have presented, there are only 22 features provided, and most of them are categorical features, such are store type and store cluster, which could not be directly used for model training. Since the target value is to predict the unit sales of each item in each store, the higher level of sales aggregation compared with store-level unit sales would have less direct correlation with the target value prediction, and similarly to the item family and class. In this case, feature engineering is a vital, and it will be based on data exploration.

4.3.1 Promotion

(Srinivasan *et al.*, 2004) noticed that since the 1970s the amount of budget on promotional activities significantly increased. Due to this long history of promotions, much information is available on price promotions. According to (Blattberg, Briesch and Fox, 1995), price promotions are not only used in the FMCG, but are applied in all kind of businesses. Price promotions increasing the store traffic during a promotion week and resulting in increased sales of promoted articles.

With understanding the importance of promotion, data exploration also proved the positive correlation between unit sales and promotion status. As figure 4 presented, unit sales are significantly higher when under promotion. In this case, promotion is considered to have effect on sales of products, and therefore will be included for feature engineering.

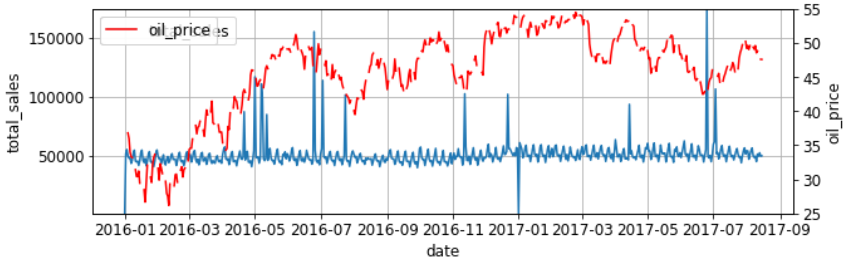
Figure 4. Density of Unit Sales by Promotion Status



4.3.2 Oil price

Ecuador is a highly oil-depending country. According to (Astudillo-Estevez St Edmund Hall, no date), around 57.6% of export income was generated by oil industry, which consisted 13.6% of its GDP. In this case, oil extraction and trading are the most critical industry of this country in terms of the absolute value.

Figure 5. oil & unit sales trend



Though according to figure 6 and 7, there is no strong correlation between oil price and daily unit sales, it does not indicate that oil price will be excluded. By plotting the oil price and promotion together, there does seem to have a joint effect between oil price and whether or not an item is on promotion on the unit sales, and the plot reveals that this interaction is not linear. In this case, oil price will also be included into feature engineering.

Figure 6. Pearson correlation between oil price and unit sales

	total_sales	oil_price
total_sales	1.000000	0.217495
oil_price	0.217495	1.000000

Figure 7. Effect of oil price and promotion status on unit sales

4.3.3 Temperature

Some literatures have also emphasized the importance of temperature. (Agnew and Thornes, 1995) concluded that short-term weather variability will result the change of consumer behavior and this influence will finally affect all the areas of supply chain through supply network, including sales. Similar to the oil price, though there is no strong correlation between temperature and daily sales (figure 8 and figure 9), there is a joint effect between temperature and promotion on the unit sales (figure 10). Therefore, it is also a necessary to include temperature for further feature engineering.

Figure 8. temperature & unit sales trend

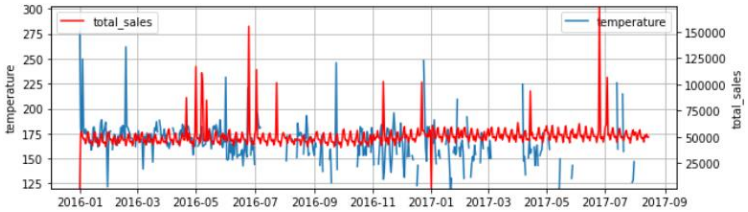
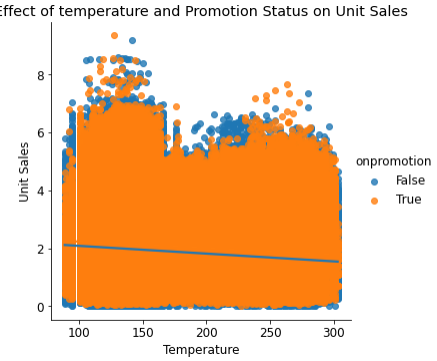


Figure 9. Pearson correlation between oil price and temperature

	temperature	total_sales
temperature	1.000000	-0.203589
total_sales	-0.203589	1.000000

Figure 10. Effect of temperature and promotion status on unit sales



4.4 Missing Value Imputation

4.4.1 Temperature

Based on the missing value matrix (figure 11), there is no direct correlation between the missing of temperature with date or day of week. Therefore, it is reasonable to interpret that the pattern of temperature missing is completely at random, and therefore it will not influence the distribution of target value. Furthermore, since the percentage of missing value is also most equal to 70% (figure 12), missing value imputation may bring more noise instead. Therefore, it is not a good choice to kept this feature, and therefore temperature will not be included for feature generation.

Figure 11. Missing value matrix of temperature

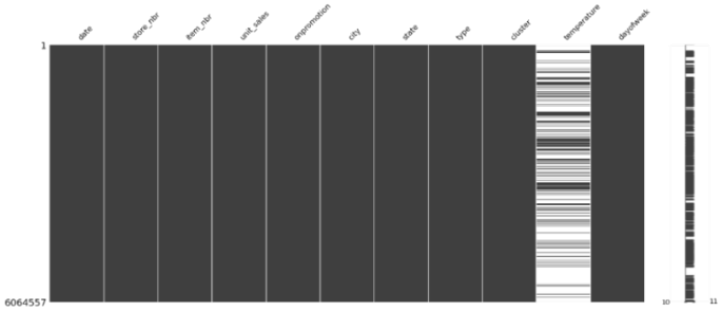


Figure 12. Missing value percentage of temperature

	Missing Values	% of Total Values
temperature	4275154	70.5

4.4.2 Oil

According to the missing matrix as figure 13 presented, it is clear that the day of the week and missing oil prices are correlated, and therefore the missing type of oil price is not at random.

Figure 13. Missing value matrix of oil

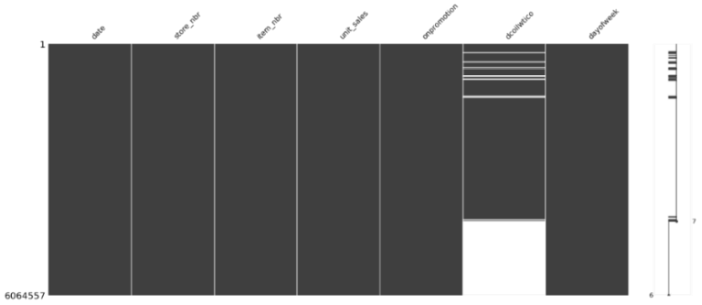


Figure 14. Missing value percentage of oil

	Missing Values	% of Total Values
dcoilwtico	1976024	32.6

Since there is a specific missing pattern and the percentage of missing value of only around 30% (figure 14), the missing oil price could be imputed without introducing more noise. It is known to all that the price of oil is not published during the weekends. Therefore, the imputation methods are pretty intuition, which is remaining the Friday price.

4.5 Feature Engineering

After missing value imputation, the selected features have to be further processed for model training, and the detailed feature engineering process as well as the reasons have been presented in figure 15.

4.5.1 Feature Engineering process

As feature selection process presented, oil price, weekday and promotion have been proved to have correlation with unit sales. In addition to these 3 main features, historical

sales will also be included in feature engineering as this is the most basic source of data to figure out the pattern of sales of each item. As the data splitting sector have explained, the earliest day could be tracked back to 90 days before the standard day for each time-sliding window.

Based on these principles, details of generated features have been presented in table 3.

Figure 15. Time-sliding window

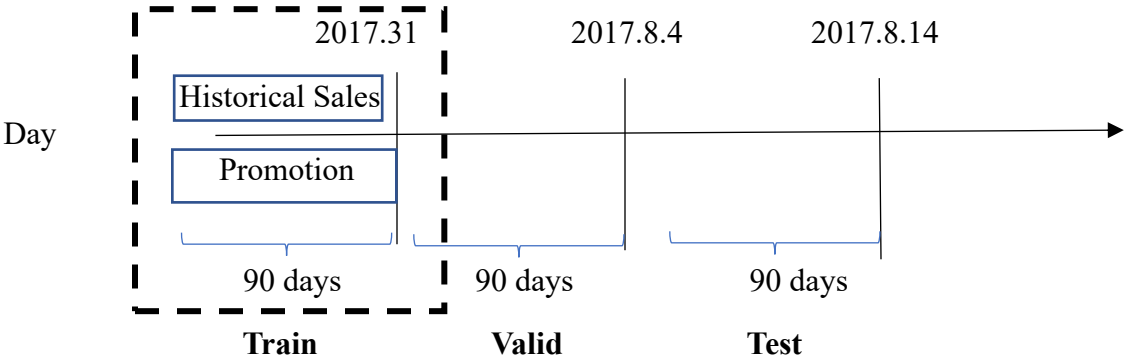


Table 3. time-sliding window feature engineering

Main features	**Before/After**	**Time Window (i days before or on the standard day)**	**Generated Features**	**Feature Name**
Sales	Before	3/7/14/30/60/90	First order difference of the mean sales of each store	diff_i_mean
			Mean, median, min, max and standard deviation(std) of sales in each store	Mean_i, median_i,min_i, max_i, std_i
			Count of days with sales	has_sales_days_in_last_i
			Interval days since the last day with sales	last_has_sales_day_in_last_i
			Interval days since the earliest day with sales	first_has_promo_day_in_last_i
			Mean sales of stores with promotion	has_promo_mean_i
			Mean sales of stores without promotion	no_promo_mean_i

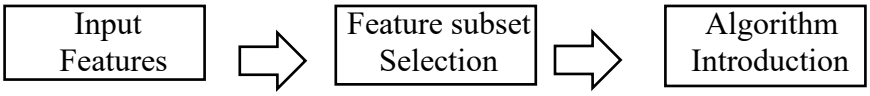
Oil price	Before	3/7/14/30/60/90	Mean oil price	oilmean_i
			First difference of oil price	oil_diff_%s_mean_i
Promotion	Before	3/7/14/30/60/90	count of days with promotion in each store	has_promo_days_in_last_i
			Interval days since last day with promotion within the window	last_has_promo_day_in_last_i
			Interval days since the earliest day with promotion within the window	first_has_promo_day_in_last_i
	Predict day	1	Weather on promotion on the predict day	promo_1
Total		97 features		

4.5.2 Feature Selection

As mentioned in the literature review, NN is very prone to overfitting, and the ensembled tree-based models also have same weakness (Khoshgoftaar and Allen, 2001; Morgan, 2003), though it has already use the ensembled resulted from several individual trees and tend to be very robust. In this case, feature selection is vital(Kalousis, Prados and Hilario, 2007), as it entails the removing of unnecessary, redundant or noisy data by identifying the pertinent features (Li *et al.*, 2017).

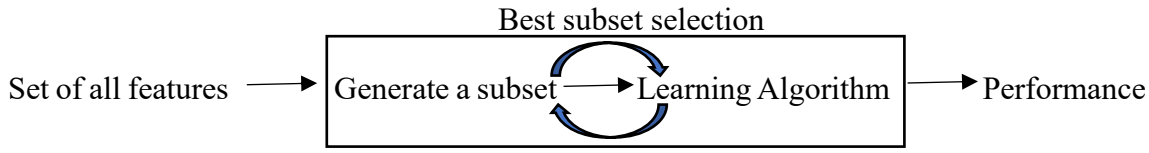
In general, filter methods and wrapper methods have been examined as two types of feature selection approaches in the literature (Zhu *et al.*, 2010). In essence, filter-type procedures are techniques for preprocessing or filtering data (figure 16), and the selection is based on the relevance between features and target groups.

Figure 16. Filter feature model process



In wrapper-type approaches, feature selection is "wrapped" around a learning method; the learning method's estimated accuracy is used to determine how beneficial a feature is. Wrapper methods frequently need to perform a lot of computation to find the optimum features.

Figure 17. Wrapper feature selection process

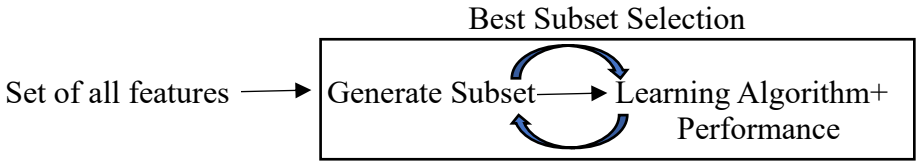


According to (Xing, Jordan and Karp, no date), the key distinctions between the two approaches are:

A wrapper method uses the final classifier's construction algorithm, and compares the performance of the final classifier and searches for an optimal subset using cross validation; whereas a filter method chooses an attribute subset using simple statistics derived from the empirical distribution.

By combining filter and wrapper approaches, embedded methods integrate feature selection into model learning. They are substantially more effective than wrapper methods and inherit the benefits of both wrapper and filter techniques because they do not need to evaluate feature sets sequentially. The most common embedded techniques use regularization models, which minimize fitting errors while concurrently driving feature coefficients to exceptionally low values. The chosen feature sets and regularization model are then returned together with the final findings (Li *et al.*, 2017).

Figure 18. Embedded feature selection process



As the wrapper applied cross validation, which should not be applied on time-series data, this dissertation only applied filter and embedded methods.

In terms of filter method, the generic univariate selection by scoring the mutual information between each feature have been applied. The reason of not applying other metric such as Pearson, is because it is only applicable for linear-correlated features. However, during the feature selection section, features such as promotion and temperature have been proved to have a jointly non-linear correlation with unit sales. Mutual information is based on the

information entropy came by Shannon (Shannon and Weaver, 1949), which means knowing random variable X, the degree to which the uncertainty of the random variable y is reduced. By avoiding the simple linear assumption, mutual information provided more space for feature selection.

Another is random forest model based embedding method. The built-in random forest importance is Gini importance. Similar to the concept information entropy, Gini index is another way to describe the purity information. As this is the embedded function in random forest, this method is more efficiency and convenient for following tree-based model building.

For each feature selection method, only top 30 features have been chosen, and the final features are the union of the selected features by these two methods. Therefore, 41 features have been selected in total, which have been presented as below.

Table 4. Final selected features

Type of features	**Features**	**Number**
Statistical	'mean_3', 'std_3', 'mean_60', 'median_14', 'mean_14', 'mean_90','mean_7','max_90','mean_30', 'no_promo_mean_90', 'has_promo_mean_90', 'no_promo_mean_60', 'no_promo_mean_30', 'no_promo_mean_14', 'no_promo_mean_7', 'no_promo_mean_3', 'median_3', 'median_7', 'min_3', 'max_3', 'median_30', 'min_7', 'max_7', 'max_14', 'has_promo_mean_14', 'max_30', 'median_60', 'max_60', 'has_promo_mean_60', 'diff_60_mean' , 'std_90', 'diff_90_mean', 'std_60', 'std_30', 'diff_30_mean', 'std_14', 'diff_14_mean', 'std_7', 'diff_7_mean', 'diff_3_mean', 'median_90'	41
Categorical	'Item_nbr', 'store_nbr', 'promo_1'	3

It is obvious that over 90% of features are statistical features, and surprisingly, oil price related features have not been selected at all.

4.6. Models

4.6.2 Moving Average

The naïve prediction is one of the most basic forecasting methods, and it is widely used as a standard for gauging how well other methods perform as the linear assumption typically underlies these procedures. As one of the most simple and intuitive prediction methods, Moving Average (MA) is widely preferred because of its simplicity and usability. For experimental comparison, MA has been used as a baseline.

MA a widely used forecasting approach based on historical data. By dividing the sum of observations by the total number of observations, the equation could be written as below:

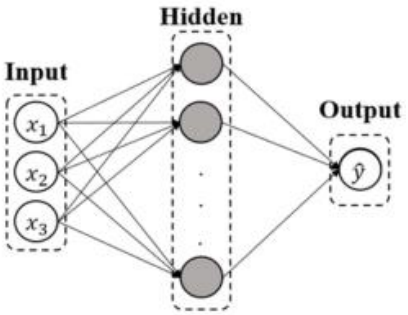
$$MA = X_{n+1} = \sum_{i=1}^{n}[\frac{X_i}{n}] \quad (1)$$

Where n is the day of moving average, and $\sum_{i=1}^{n}[\frac{X_i}{n}]$ is the sum value of recall n days.

4.6.1 Artificial Neural Network

Inspired by the working mechanism of human brain, NN have had remarkable success in a wide range of fields. One of the most basic varieties of predictive models in the category are multi-layer neural networks (NN) (Figure 19). After receiving input values, the "neurons" on the hidden layers compute these values and then propagate those values to the output layer using the weighted sum of the layer before it. And the final result is the sum of these values. To add nonlinearity to prediction, nonlinear mappings between layers known as "activation function" are used (Lee *et al.*, 2020). As the application of NN is very widely used in food sales prediction area as literature review presented, and the strength and weakness have also been extensively discussed, it would be a “mature” comparison model to be compared with tree-based model. Therefore, NN has also been selected.

Figure 19. Final selected features



4.6.2 XGB & LGB

Both XGB and LGB models are based on GBDT. XGB is a tree-based model proposed by (Chen and Guestrin, 2016). Different from GBDT, it has introduced the regularization terms into objective function to prevent overfitting; while LGB was proposed by Microsoft in 2017(Ju *et al.*, 2019), and compared with XGB, whose performance could still be unsatisfied under large data size and high dimensionality have adopted the engineering optimization, LGB used a tree-wise growth algorithm with a maximum depth limit and a histogram-based methodology.

As the research gap have pointed out, though these two tree-based models have been widely applied and discussed in many areas, there is only a limited or even few literatures that academically discussed the performance of XGB and LGB on perishable food sales prediction. For this reason, the performance of these two advanced tree-based models are the two main models to be discussed in this dissertation.

4.7 Hyper-Parameter optimization

4.7.1 Optuna

One of the most challenging tasks in machine learning projects is hyperparameter search. With the rise in popularity of deep learning techniques comes an increase in method complexity, and there is now a greater need than ever for an effective framework for automatic hyperparameter tuning. There are a number of issues with the current optimization frameworks, despite the fact that there is a wide variety in the parameter-sampling algorithms. First, all existing hyperparameter optimization systems to date require that the user statically build the parameter search space for each model in large-scale trials including a number of candidate models of various sorts with big parameter spaces and many condition variables. It might be exceedingly challenging to characterize the search space. Advanced optimization techniques may be useless if the user fails to adequately characterize the parameter space; Second, many existing frameworks lack efficient pruning techniques, despite the fact that both parameter searching methodology and performance estimation strategy are essential for high-performance optimization under constrained resource availability. In terms of these problems, optuna first proposed 1) a define-by-run API that enables dynamic construction of parameter search space; 2) Second,

many existing frameworks lack efficient pruning techniques, despite the fact that both parameter searching methodology and performance estimation strategy are essential for high-performance optimization under constrained resource availability (Akiba *et al.*, 2019). In this dissertation, optuna will be used for XGB and LGB hyperparameter tuning.

4.7.2 Manual tunning

The number of input nodes and the number of hidden nodes affect how large the neural network model's structure is. Without any theoretical foundation to guide the choice, the number of concealed nodes is frequently determined in practice through experimentation or trial and error. Although there is a theory that additional hidden nodes can increase the functional relationship's approximation accuracy, it can also result in overfitting issues. Finding a parsimonious model that fits the data well is the answer to the overfitting issue. In this case, the parameter tunning of neural network starts from a simple setting, and tunning based on the error.

4.8 Evaluation

The mean squared error, or MSE, is a metric used to assess the level of accuracy in statistical models. The reason of choosing MSE instead of MAE is because of its derivative ability. Derivation is an essential metric for model optimization to decrease the error. Though one of the disadvantages of MAE is sensitive to outliers, since the dataset have been preprocessed, and therefore outliers will not have huge influence on the evaluation.

MSE calculates the average squared difference between the values of the observed and projected values. The objective is then to minimize MSE, and 0 MSE indicated a perfect prediction of the trend. The following might be represented as the function of MSE:

$$MSE = \frac{\sum(y_i - \widehat{y_i})2}{n} \quad (1)$$

Where y is the value that was seen, $\widehat{y_i}$ is the corresponding predicted value, and n is the total number of observations.

5. Experimental results

This section details the hyper-parameter tuning process and the predicted results obtained with MA, NN, XGB and LGB.

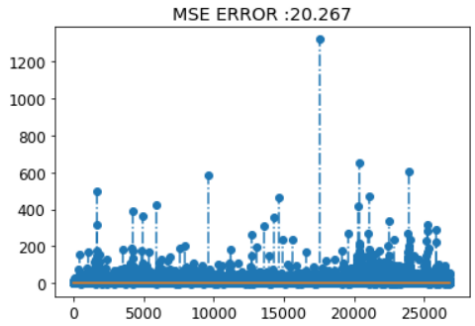
5.1 Moving Average

The model setting process of moving average is very straight. According to the principle of feature engineering, the earliest sales records have also been traced back to 90 days before the prediction day.

5.1.1 Test dataset prediction

By applying moving average to test dataset, the MSE error is over 20. As the literature presented, though moving average works well for the most recent value prediction, since many external factors were included into this experiment, it is reasonable why the model performance is not very satisfying.

Figure 20. MSE error of Moving Average



5.2 Neural Network

5.2.1 Hyper-Parameter Tunning

There are totally 14 hyper parameters of neural network have to be tuned, and the size of epochs, min delta, patience, initializer, optimizer, beta 1&2, as well as the rate of alpha drop out have not been changed during the whole tunning process. Hyper parameter initialization started with 3 dense layers with 10,8, and 1 last neuro in the output layer, as the prediction value is a linear problem. In order to avoid overfitting, early stopping and batch normalization were introduced at the beginning. In order to avoid the saturation of neuros, elu has been selected instead of sigmoid. In addition, this method also enables an average output closer to zero, which increase the learning speed of model. In addition, He

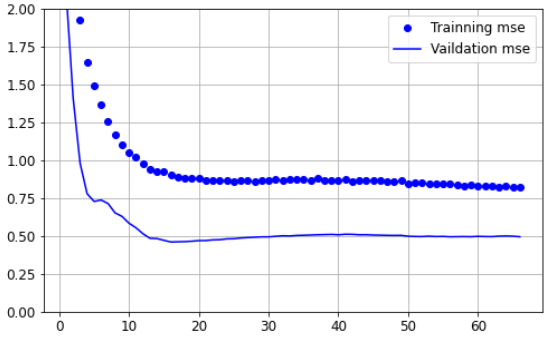
normalization have also been introduced to avoid gradient vanishing or exploding during learning. In terms of the optimizer, Adam have been selected as it is the combination of Momentum & RMSprop optimizer.

Table 4. Hyper-parameter Optimization Process

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Epocchs	400	400	400	400	400	400
Batch size	50	50	100	100	100	100
Min delta	0.001	0.001	0.001	0.001	0.001	0.001
Patience	50	50	50	50	50	50
Alpha Dropout	0.2	0.2	0.2	0.2	0.2	0.2
Dense	10/8/1	6/4	6/4	6/4	6/4	6/4
Activation	elu	elu	elu	relu	relu	relu
Initializer	He_normal	He_normal	He_normal	He_normal	He_normal	He_normal
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam
Batch Normalization	Yes	Yes	Yes	Yes	Yes	Yes
Learning rate	0.00005	0.000055	0. 000056	0.00005	0.000055	0.000055
Beta_1/Beta_2	0.1/0.2	0.2/0.2	0.2/0.2	0.2/0.2	0.2/0.2	0.2/0.2
L1/l2	/	/	/	/	0.002/0.1	0.002/0.2
MSE	**1.2955**	**1.0994**	**1.0063**	**0.9853**	**0.6659**	**0.4948**

After the first initialization, though the MSE on train dataset kept decreasing from 2 to 0. 5, the MSE of validation dataset ended with 1.2955, which significantly higher than that of training dataset, which means the model is overfitting, and therefore the model has to been designed mode simple. Therefore, from the second last training, L1 &2 penalty have been applied on both layers to decrease overfitting, which lead an immediate 0.3 decrease of MSE, and finally it ends with 0.49.

Figure 21. Model Training result of NN



5.2.2 Test dataset prediction

Though the NN model performed better on validation dataset with around 0.5 MSE decreased, the performance on test dataset was worse than that on training dataset. According to the evaluation result, the MSE on test dataset increased to 1.6268. which indicated NN model had a very unsatisfied generalization ability under this specific context.

This result could be explained by several reasons. The first is the similar to the conclusion from (Lee et al., 2012). Since the selection of hyper parameter of neural network highly depends on the past experience and therefore prone to be subjective, it is very likely that though the performance of model has improved after parameter tunning, the range of tunning only covered a small range, which is a big limitation. Another reason could be the model characteristic. As the working process of neural network is a “black box”, compared to tree-based model, the lack of tracking during learning could also make the tunning more difficult, and further influence the final prediction.

5.3 XGB &LGB

5.3.1 Hyper-Parameter Tunning

Compared to NN, the hyper parameter tunning process of XGB and LGB is more efficient and less subjective by applying optuna, and therefore, a wide range of searching scope have been provided from the beginning. Table 4 presents the range of each parameter tunning. Since two models are both based on the GBDT, the range of search scope have been kept the same, which is also convenience for the comparison between two tree-based model.

The optimization of hyper-parameter of both models have been presented in figure 22 and 23, and during optimization processes, both MSE have a significant decrease, which means the model is ready to be applied on test dataset.

5.3.2 Test dataset prediction

By fitting the test dataset into the trained models, it is obvious that the both MSE error of XGB (0.35) and LGB (0.37) decreased compared with the result of validation dataset (figure 24 and figure 25).

Table 4. Range of tree-based hyper-parameter and final parameter

Parameter	Searching scope of hyper-parameter	XGB final parameter	LGB final parameter
estimators	[50,300)	222	223
feature fraction	[0.01,0.08)	0.037	0.069
Bagging fraction	[0.01,0.08)	0.012	0.053
Bagging freq	2	2	2
Reg alpha	[1e-3, 10.0)	1.88	0.184
Reg lambda	[1e-3, 10.0)	0.003	0.023
Colsample bytree	[0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0]	0.7	1
Subsample	[0.4,0.5,0.6,0.7,0.8,1.0]	1	0.7
Learning rate	[0.05,0.1,0.01]	0.05	0.05
Max depth	[5, 7, 9, 11, 13, 15, 17, 20, 50]	7	9
Leaves	[2,100)	32	71
Min child samples	[1,300)	153	74
Min child weight	[0.001,0.01)	0.0076	0.003
MSE		**0.35**	**0.37**

Figure 22. Optimization history of XGB

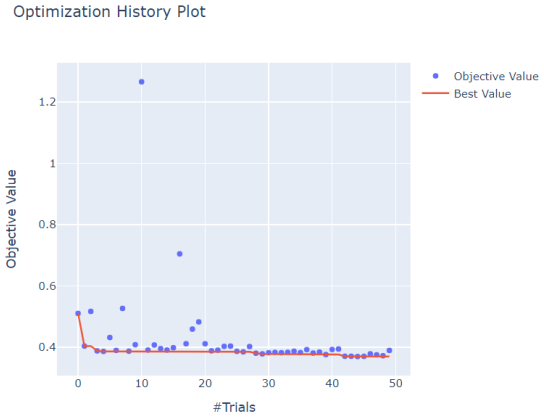


Figure 23. Optimization history of LGB

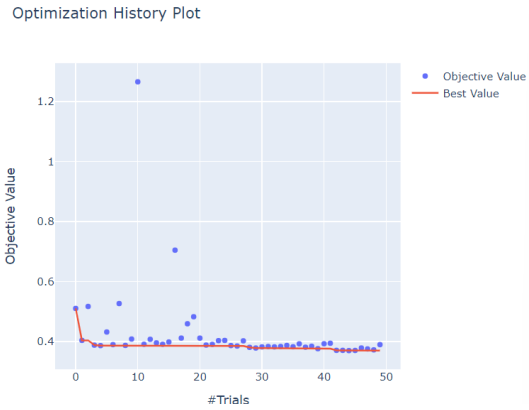


Figure 24. XGB MSE error (0.334)

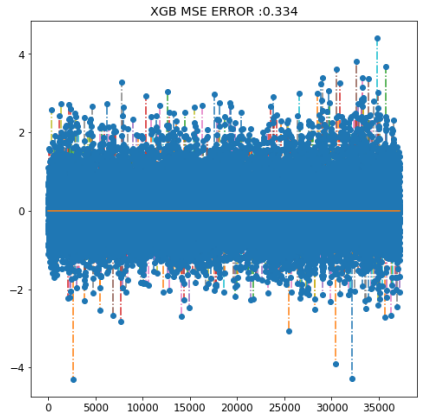
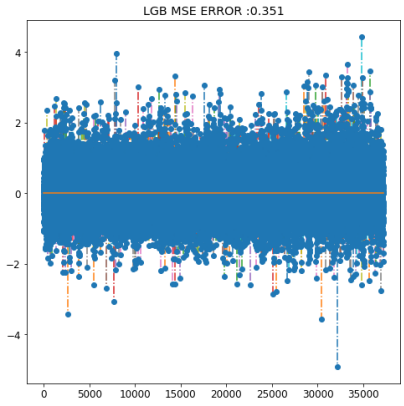


Figure 25. LGB MSE error (0.351)



6. Discussion

The result presents that tree-based model performed the best, while MA has the highest error among these 4 models. This finding is corresponding to the most literatures (Carbonneau, Laframboise and Vahidov, 2008), as MA works based on the simple-linear assumption. Compare to MA, there is a significantly decrease on NN's MSE error, which is around 90%. This finding is also consistent with the conclusion of mentioned literatures, that NN can process the highly complexed non-linear function (Stamell *et al.*, no date), as most external factors have non-linear correlation with sales. In addition, compared to MA, which works best on the most recent dataset (Carbonneau, Laframboise and Vahidov, 2008), NN seems to be more robust on larger dataset with longer time-period. In terms of XGB and LGB, the error further decreased around 80% compared to that of NN. This finding filled the gap by proving the high accuracy and efficiency of tree-based model on perishable food sales prediction with hyper-parameter tunning methods optuna. Within two tree-based models, though XGB had a better performance with lower MSE error, LGB had a significantly higher efficiency on model training. With only 0.02 differences on MSE error, LGB is still considered to be the best choice in this dissertation.

Compare to the forecasting result of NN (Chen and Ou, 2008), which was 0.01799 in MSE, 1.63 is a significantly higher error in this research. However, both the scale of dataset or the complexity of experiment designing in (Chen and Ou, 2008) are more straightforward compared to that of this dissertation: only 326 data point have been used to build the NN model and time lag is also simply ranged from 3 to 8. With larger scaled of dataset and more external factors included, 1.63 seems to be reasonable and acceptable. The situation is relatively different in terms of the tree-based models. Compare to the 5.66 (Stamell *et al.*, no date) RMSE error of XGB and 3.6 of LGB (Marcos Roberto Machado, Salma Karray and Ivaldo Tributino de Sousa, 2019), 0.334 and 0.351 MSE are much satisfying, and it may because of several reasons. Firstly, both these two researches are not time-series problem, and therefore the research designing is relatively different; In addition, the range of dataset collected by (Stamell *et al.*, no date) is 34, which also introduced more noise with large scale of samples. In terms of the theoretical contribution, by expanding the comparison models to NN and MA which suggested by (Marcos Roberto Machado, Salma

Karray and Ivaldo Tributino de Sousa, 2019), the conclusion of this dissertation provides a more rigorous proof of the validity of the tree models. More importantly, the background of this dissertation is based on a new sector, perishable food sales prediction, and therefore filled the research gap.

This finding also provides more choices for businesses practical usage. XGB and LGB could provide corporates with more accurate prediction results but less computation cost. And by applying these two models, both product waste and financial loss could be highly decreased. Another interesting point is the selected external features. In terms of the types of features, it is obvious that the statistical features, such as mean, median, max or min sales, have appeared most often; In terms of the category of features, promotion seems to have higher influence of sales prediction compared with daily oil price, which is an unexpected finding as Ecuador is an oil-dependent country. It may because that the fluctuation of oil price is relatively stable during the observation period, and therefore, have limited influence on sales. Therefore, when business need to manage the stock of perishable food, they could use LGB with more statistical and promotional information of items and stores.

7. Conclusion

Perishable food sales prediction is one of the most important tasks for modern corporations, as it could help companies decrease the probability of OOS or demand overestimation, and therefore avoid wastes and increase customer loyalty. However, though many traditional and deep learning algorithms have been proved to be helpful for sales prediction, the efficiency of tree-based model on perishable food sales prediction have not been academically discussed yet. Therefore, this dissertation compared two tree-based model (XGB and LGB) with NN and MA to fill the gap. And the whole dataset is provided by an Ecuadorian-based grocery retailer, called Corporacion Favorita, and published on Kaggle.

Not only limited to the time-series prediction question, this research also includes some other interesting challenges. The first is to include the external factors, such as temperature, oil price and promotion into feature engineering, in order to increase the

prediction accuracy; and another is the multi-tasking prediction task, as the model should be able to predict the future sales for each item in each store. Therefore, models should have well-generalized ability.

Therefore, the whole prediction process could be divided into 6 parts, starting from dataset splitting, feature selection and engineering, missing value imputation and finally model training and prediction. The number of generated features according to literatures and data exploration have been further downsized based on filter and embedded feature selection models.

This dissertation proved the efficiency of tree-based model on perishable food sales prediction by comparing XGB and LGB with neural network as well as Moving Average. And among 4 models, XGB performed best among 4 models, with 0.02 lower MSE than that of LGB. However, by considering the computational cost and efficiency, it is believed that LGB is more robust to even larger dataset in the future practice. It is also worth noting that promotion and basic statistical information such as mean, median sales are most important features for sales prediction. These two findings provided perishable food production and distribution companies, or any other entities along the perishable food supply chains with more efficient and better prediction methods to help them avoid potential financial loss, food wastes and lower customer churning rate.

For future work, it is recommended to extent the searching scope of neural network's hyper-parameter with advanced optimization methods, such as Simulated Annealing or Genetic Algorithm; In addition, since only statistical metric (MSE) have been used to evaluate the efficiency of models, it is also recommended to introduce more financial evaluation metrics, such as inventory turnover or ROI to provide different angles of model evaluation for business.

6. Bibliography

Agnew, M.D. and Thornes, J.E. (1995) *The weather sensitivity of the UK food retail and distribution industry*, *Meteorol. Appl*.

Ahmad, M.W., Mourshed, M. and Rezgui, Y. (2018) "Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression," *Energy*, 164, pp. 465–474. Available at: https://doi.org/10.1016/j.energy.2018.08.207.

Akiba, T. *et al.* (2019) "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, pp. 2623–2631. Available at: https://doi.org/10.1145/3292500.3330701.

Arif, M.A.I. *et al.* (2020) "Comparison Study: Product Demand Forecasting with Machine Learning for Shop," *Proceedings of the 2019 8th International Conference on System Modeling and Advancement in Research Trends, SMART 2019*, pp. 171–176. Available at: https://doi.org/10.1109/SMART46866.2019.9117395.

Astudillo-Estevez St Edmund Hall, P.A. (no date) *Towards a Post-Oil Economy: A Complexity Approach to Understanding Natural Resource Dependency and Economic Diversification in Ecuador*.

Aviv, Y. (2007) "On the benefits of collaborative forecasting partnerships between retailers and manufacturers," *Management Science*, 53(5), pp. 777–794. Available at: https://doi.org/10.1287/mnsc.1060.0654.

Blattberg, R.C., Briesch, R. and Fox, E.J. (1995) *How Promotions Work*, *Source: Marketing Science*. Available at: https://about.jstor.org/terms.

Carbonneau, R., Laframboise, K. and Vahidov, R. (2008) "Application of machine learning techniques for supply chain demand forecasting," *European Journal of Operational Research*, 184(3), pp. 1140–1154. Available at: https://doi.org/10.1016/j.ejor.2006.12.004.

Chen, F.L. and Ou, T.Y. (2008) "A neural-network-based forecasting method for ordering perishable food in convenience stores," in *Proceedings - 4th International Conference on Natural Computation, ICNC 2008*, pp. 250–254. Available at: https://doi.org/10.1109/ICNC.2008.275.

Chen, T. and Guestrin, C. (2016) "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, pp. 785–794. Available at: https://doi.org/10.1145/2939672.2939785.

Corsten, D. and Gruen, T. (2003) “Desperately seeking shelf availability: An examination of the extent, the causes, and the efforts to address retail out-of-stocks,” *International Journal of Retail & Distribution Management*, 31(12), pp. 605–617. Available at: https://doi.org/10.1108/09590550310507731.

van Donselaar, K. *et al.* (2006) “Inventory control of perishables in supermarkets,” *International Journal of Production Economics*, 104(2), pp. 462–472. Available at: https://doi.org/10.1016/j.ijpe.2004.10.019.

Dqg, H. *et al.* (no date) “Machine Learning applications in supply chains.”

Ehrenthal, J.C.F. and Stölzle, W. (2013) “An examination of the causes for retail stockouts,” *International Journal of Physical Distribution and Logistics Management*, 43(1), pp. 54–69. Available at: https://doi.org/10.1108/09600031311293255.

Ghalehkhondabi, I. *et al.* (2017) “An overview of energy demand forecasting methods published in 2005–2015,” *Energy Systems*, 8(2), pp. 411–447. Available at: https://doi.org/10.1007/s12667-016-0203-y.

Group, A. and Gijsbrechts, B.E. (2000) *Towards Understanding Consumer Response to Stock-Outs KATIA CAMPO*.

Guo, Z.X., Wong, W.K. and Li, M. (2013) “A multivariate intelligent decision-making model for retail sales forecasting,” *Decision Support Systems*, 55(1), pp. 247–255. Available at: https://doi.org/10.1016/j.dss.2013.01.026.

Huber, J., Gossmann, A. and Stuckenschmidt, H. (2017) “Cluster-based hierarchical demand forecasting for perishable goods,” *Expert Systems with Applications*, 76, pp. 140–151. Available at: https://doi.org/10.1016/j.eswa.2017.01.022.

Huber, J. and Stuckenschmidt, H. (2020) “Daily retail demand forecasting using machine learning with emphasis on calendric special days,” *International Journal of Forecasting*, 36(4), pp. 1420–1438. Available at: https://doi.org/10.1016/j.ijforecast.2020.02.005.

J. Gustavsson *et al.* (2011) *Global Food Losses and Food Waste- Extent, Causes and Prevention.*

Ju, Y. *et al.* (2019) “A model combining convolutional neural network and lightgbm algorithm for ultra-short-term wind power forecasting,” *IEEE Access*, 7, pp. 28309–28318. Available at: https://doi.org/10.1109/ACCESS.2019.2901920.

Kalousis, A., Prados, J. and Hilario, M. (2007) “Stability of feature selection algorithms: A study on high-dimensional spaces,” *Knowledge and Information Systems*, 12(1), pp. 95–116. Available at: https://doi.org/10.1007/s10115-006-0040-8.

Kandananond, K. (2012) “A comparison of various forecasting methods for autocorrelated time series,” *International Journal of Engineering Business Management*, 4(1), pp. 1–6. Available at: https://doi.org/10.5772/51088.

Khoshgoftaar, T.M. and Allen, E.B. (2001) *Controlling Overfitting in Classification-Tree Models of Software Quality*, *Empirical Software Engineering*.

Lee, H.L., Padmanabhan, V. and Whang, S. (1997) “Information distortion in a supply chain: The bullwhip effect,” *Management Science*, pp. 546–546. Available at: https://doi.org/10.1287/mnsc.1040.0266.

Lee, H.L., Padmanabhan, V. and Whang, S. (2004) “Information distortion in a supply chain: The bullwhip effect,” *Management Science*, 50(12 SUPPL.), pp. 1875–1886. Available at: https://doi.org/10.1287/mnsc.1040.0266.

Lee, K. *et al.* (2020) “Comparison of Artificial Intelligence Methods for Prediction of Mechanical Properties,” in *IOP Conference Series: Materials Science and Engineering*. IOP Publishing Ltd. Available at: https://doi.org/10.1088/1757-899X/967/1/012031.

Lee, W.I. *et al.* (2012) “A comparative study on the forecast of fresh food sales using logistic regression, moving average and bpnn methods,” *Journal of Marine Science and Technology*, 20(2), pp. 142–152. Available at: https://doi.org/10.51400/2709-6998.1832.

Li, J. *et al.* (2017) “Feature selection: A data perspective,” *ACM Computing Surveys*. Association for Computing Machinery. Available at: https://doi.org/10.1145/3136625.

Liang, W. *et al.* (2020) “Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms,” *Mathematics*, 8(5). Available at: https://doi.org/10.3390/MATH8050765.

Marcos Roberto Machado, Salma Karray and Ivaldo Tributino de Sousa (2019) *LightGBM: an Effective Decision Tree GradientBoosting Method to Predict Customer Loyalty inthe Finance Industry*.

Melgani, F. and Bruzzone, L. (2004) “Classification of hyperspectral remote sensing images with support vector machines,” *IEEE Transactions on Geoscience and Remote Sensing*, 42(8), pp. 1778–1790. Available at: https://doi.org/10.1109/TGRS.2004.831865.

Minner, S. and Transchel, S. (2017) “Order variability in perishable product supply chains,” *European Journal of Operational Research*, 260(1), pp. 93–107. Available at: https://doi.org/10.1016/j.ejor.2016.12.016.

Morgan, J. (2003) *SAMPLE SIZE AND MODELING ACCURACY OF DECISION TREE BASED DATA MINING TOOLS*, *Academy of Information and Management Sciences Journal*.

Ning, A. *et al.* (2009) “Fulfillment of retailer demand by using the mdl-optimal neural network prediction and decision policy,” *IEEE Transactions on Industrial Informatics*, 5(4), pp. 495–506. Available at: https://doi.org/10.1109/TII.2009.2031433.

Parfitt, J., Barthel, M. and MacNaughton, S. (2010) “Food waste within food supply chains: Quantification and potential for change to 2050,” *Philosophical Transactions of the Royal Society B: Biological Sciences*. Royal Society, pp. 3065–3081. Available at: https://doi.org/10.1098/rstb.2010.0126.

Peters, J. (2012) *Improving the promotional forecasting accuracy for perishable items at Sligro Food Group B.V.*

Qu, T. *et al.* (2017) “Demand prediction and price optimization for semi-luxury supermarket segment,” *Computers and Industrial Engineering*, 113, pp. 91–102. Available at: https://doi.org/10.1016/j.cie.2017.09.004.

Sari, K. (2007) “Exploring the benefits of vendor managed inventory,” *International Journal of Physical Distribution and Logistics Management*, 37(7), pp. 529–545. Available at: https://doi.org/10.1108/09600030710776464.

Sari, K. (2008) “On the benefits of CPFR and VMI: A comparative simulation study,” *International Journal of Production Economics*, 113(2), pp. 575–586. Available at: https://doi.org/10.1016/j.ijpe.2007.10.021.

Shang, G. *et al.* (2020) “Using transactions data to improve consumer returns forecasting,” *Journal of Operations Management*, 66(3), pp. 326–348. Available at: https://doi.org/10.1002/joom.1071.

Shannon, C.E. and Weaver, W. (1949) *THE MATHEMATICAL THEORY OF COMMUNICATION*.

Song, L. *et al.* (2016) “Architecture of demand forecast for online retailers in China based on big data,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Verlag, pp. 759–764. Available at: https://doi.org/10.1007/978-3-319-31854-7_75.

Srinivasan, S. *et al.* (2004) “Do promotions benefit manufacturers, retailers, or both?,” *Management Science*, 50(5), pp. 617–629. Available at: https://doi.org/10.1287/mnsc.1040.0225.

Stamell, J. *et al.* (no date) “Strengths and weaknesses of three Machine Learning methods for pCO 2 interpolation.” Available at: https://doi.org/10.5194/gmd-2020-311.

Sterman, J.D. (1989) *Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Experiment*, *Science*. Available at: https://about.jstor.org/terms.

Suban, D.T. and Bogataj, M. (2015) “An optimal ordering cycle at interactions of fuzzy parameters and high disposal fees of food or drug supply systems,” in *IFAC-PapersOnLine*, pp. 2374–2379. Available at: https://doi.org/10.1016/j.ifacol.2015.06.443.

Tarallo, E. *et al.* (2019) “Machine learning in predicting demand for fast-moving consumer goods: Exploratory research,” in *IFAC-PapersOnLine*. Elsevier B.V., pp. 737–742. Available at: https://doi.org/10.1016/j.ifacol.2019.11.203.

Taylor, D.H. and Fearne, A. (2009) “Demand management in fresh food value chains: A framework for analysis and improvement,” *Supply Chain Management*, 14(5), pp. 379–392. Available at: https://doi.org/10.1108/13598540910980297.

Thron, T., Nagy, G. and Wassan, N. (2007) “Evaluating alternative supply chain structures for perishable products,” *The International Journal of Logistics Management*, 18(3), pp. 364–384. Available at: https://doi.org/10.1108/09574090710835110.

Tsoumakas, G. (2019) “A survey of machine learning techniques for food sales prediction,” *Artificial Intelligence Review*, 52(1), pp. 441–447. Available at: https://doi.org/10.1007/s10462-018-9637-z.

Wang, W. (2011) “Analysis of bullwhip effects in perishable product supply chain - Based on system dynamics model,” in *Proceedings - 4th International Conference on Intelligent Computation Technology and Automation, ICICTA 2011*, pp. 1018–1021. Available at: https://doi.org/10.1109/ICICTA.2011.255.

Xing, E.P., Jordan, M.I. and Karp, R.M. (no date) *Feature Selection for High-Dimensional Genomic Microarray Data*.

Zhu, S. *et al.* (2010) “Feature selection for gene expression using model-based entropy,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(1), pp. 25–36. Available at: https://doi.org/10.1109/TCBB.2008.35.

Zinn, W. and Liu, P.C. (2008) “A COMPARISON OF ACTUAL AND INTENDED CONSUMER BEHAVIOR IN RESPONSE TO RETAIL STOCKOUTS,” *JOuRNAL OF BuSINESS LOGISTICS*, 29(2).