DEMAND FORECASTING FOR A LARGE GROCERY CHAIN IN ECUADOR

Varsha Prabhakar (prabhakv@purdue.edu), Doga Sayiner (dsayiner@purdue.edu), Udita Chakraborty (uchakra@purdue.edu), Thuy Nguyen (nguye441@purdue.edu), Matthew A. Lanham (lanhamm@purdue.edu)
Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907

ABSTRACT

In this project assignment, we are predicting the future sales of a grocery store chain called ‘Corporacion Favorita’. It is one of the leading competitors in the market. Having this much of an importance, Corporacion Favorita proves itself to be a good target for data analytics. Being a dominant and big company creates a bigger risk for stock-out opportunity costs and over-stocking storage and warehouse management costs. Having many stores and customers increases its data availability. This paper explains the learning algorithms applied to predict the unit sales of items sold by Corporacion Favorita. Predictive modeling is the main tool used to solve the analytics problem generated by the business problem. R language is one of the most optimized statistical languages that allow machine learning algorithms to be applied with great efficiency. Especially caret library is used to analyze data and to make predictions.

Keywords: R, caret, predictive modeling, demand forecasting

INTRODUCTION

Corporacion Favorita is an Ecuadorian firm which has more than 50 grocery stores all over the country. Most of its stores are concentrated in Quito, the country's capital. Grocery store chain hold over 4000 items and on average there are 1695 transactions per day per store (almost 100,000 transactions per day across all the stores). In a country with a population of 16 million, mentioned figures show that Corporacion Favorita has an important share in grocery market. Assuming 20% of the country consists of children and teenagers, most of these 100,000 transactions are about the remaining 13 million of the population. This roughly translates to 1 in 100 people choosing to buy from Corporacion. Ecuador is an agricultural country and people generally prefer local merchants to buy goods. Therefore, even though 1% of the population does not seem to be an important figure, it actually means that they can compete against the status quo of the country.

Business problem that Corporacion faces is about having a sound forecast on the unit sales for individual items. Being able to predict the amount sold will enable them to better organize their logistical operations. This will potentially cut both opportunity-costs introduced by stock-outs and warehouse management and storage costs introduced by over-stocking. Forbes contributor Steve Banks claims that "The cost of excess inventory can be huge" and he goes on explaining how much money can be saved while trying to avoid this scenario with better forecasting and other operational means. Having extra material waiting at the warehouses can potentially cost drastic amounts of money. Products just sit on shelves where they are not being utilized. Depending on their type, they can get spoiled which would cause a total loss for the owning company or they can be sold at loss in order to prevent future storage expenses.

Gartner is very involved in making sound and accurate predictions. It is discussed that having the right amount of product at the right time can reduce both opportunity cost and storage costs. However, good predictions/forecasts are not enough on their own. Just-in-time rule first developed by Toyota Manufacturing introduces further reductions in expenses and potential loses. Not having items waiting in production lines increases the line effectiveness and overall profits incurred by it. Moreover, predictions are not only about retail and they hold the power to unlock many more things. Gartner makes a prediction to suggest that the worldwide device shipments will increase 2.1 percent in 2018. Being able to make such a prediction, if the prediction itself considered to be accurate, can lead to many possibilities, especially in terms of investment.

In a similar fashion, Wall Street Journal forecasts that the auto sales will have different effects depending on type of the auto sales, namely the car category or the light truck category. According to Wall Street Journal, United States is experiencing a slowing down in the sales of car whereas light truck sales are soaring. Having this kind of predictive power that would allow investors or even common people planning to buy personal vehicles to better evaluate their options, understand the return of their investments (ROI) and the net present value that these vehicles actually hold today. This creates an environment of consciousness which drives competition between the sellers and creates a driving force for the sake of the economy.

All the points mentioned above suggests that prediction of any kind is very important and is highly regarded in world's one of the most famous news outlets. Therefore working on predicting the sales of grocery market and specific items provided by 'Corporacion Favorita' is an important business that will cut off the wastes (since the products that are mentioned in our case are likely to

be susceptible to being spoiled if not sold in time this waste factor is much more important than other possible waste scenarios) and stimulate growth for both ‘Corporacion’ and the sector that they are a player in, due to expected increase in the competition generated by data science being in play and pushing other companies to adapt and hop on this train of predictive analytics for forecasting unit sales of goods.

In this article, we research machine learning algorithms and methods in general that will allow the prediction of unit sales of products for a big grocery retailer. The importance of this work is emphasized by the news agencies that are mentioned above. They consider forecast to be paramount to achieve higher profit figures and explain how this is a relevant problem by providing examples that are all around us. In the end, we are hoping to provide the tools necessary to achieve high accuracy forecasting results for grocery retail sector that will allow to cut wastes, costs and increase profits along with sustainability.

In terms of the required time to create the prediction models and performance measures that are obtained by using the models to make predictions, the best method to be applied is determined to be “Gradient Boosting”.

The remainder of this paper is organized as follows: A review on the literature on various criteria and methods used for retail forecasting is presented in the next section. In Section 3, the proposed methodology to approach the problem at hand is presented thoroughly, and the criteria formulation is discussed. In Section 4, various models are formulated and tested. Section 5 outlines the performance of our models. Section 6 concludes the paper with a discussion of the implications of this study, future research directions, and concluding remarks.

LITERATURE REVIEW

Prediction of sales is an important field in the grocery and food industry and due to new technologies, it has recently gained a lot of attention in order to improve business operations and profitability. However, historically the industry has relied on traditional statistical models but in recent years, more advanced machine learning methods has gained attraction.

Therefore, with the topic of sales forecasting, various studies are analyzed to understand the current methodology among different organizations in the industry and some of the best practices are found, including SVM, neural networks and weighted moving averages, which may be considered for the purpose of this paper.

Paper I "Demand Forecasting in Retail Grocery Stores in The Czech Republic"

Research Problem	Algorithm/Methodology	Result	Further analysis
Grocery store retailers in Czech Republic rely in demand forecasting on their own intuition and experience in retailing, and therefore qualitative methods of forecasting are applied most frequently)	A quantitative research was conducted in 75 selected retail stores in the Czech Republic. The sample included only retail stores where groceries predominate in the range of goods	Judgmental method (40%), (Moving) Average (21%), Naive method (19%), Customer expectations (9%), Unknown methods implemented in software (5%), Analogy method (4%), Simple regression (3%), Time series decomposition (1%), Exponential smoothing (0%), ARIMA models (0%) Advanced forecasting models (0%)	Further research should therefore be focused on identifying the causes of the current level of demand forecasting in the retail business, including specifying the possibility of removing barriers to the implementation of more suitable approaches to demand forecasting in the surveyed retailers.

Paper II_"Optimization of The Sales Forecast Algorithm for a Supermarket Supply Chain"

Research Problem	Algorithm/Methodology	Result	Further analysis
This paper represents the results of the study of different forecasting models applied to sales data of products sold by Auchan Portugal, with the objective of improving/optimizing its main storehouse stock management.	The paper focuses on a study of three forecasting models: moving averages, weighted moving averages, and moving averages with exponential smoothing. The study was performed using weekly sales data of the same product, in order to compare the results obtained.	The sales forecasts that achieve the results closest to the effective sales are the ones that make use of the weight moving averages, which give a greater weight to the most recent sales data. The simple moving averages model also follows the evolutionary trend of effective sales but presents greater deviations.	Even though the this is a good model in itself, it forecasts based on its own trends and does not account for external events and characteristics that can severely impact sales. Hence there is a tremendous need for statistical models that can describe dependencies, predict sales and support inventory management.

Paper III "Predictive Analysis of Big Data in Retail Industry"

Research Problem	Algorithm/Methodology	Result	Further analysis
The paper focuses on big data in retail industry, providing a summary of the state-	**Data mining** refers to techniques are used to extract patterns from data, such as rule learning, cluster analysis, classification and regression,	**Customer targeting:** Customer's individual behaviors can be segmented by using big data analytics and collect customer behavior at each touch point. By	This paper show how big data can help to improve the retail business and can be applied in the sector and help improve margin.

of-the-art research on big data analytics.	which can be used to determining for example the characteristics of successful employees or even determine customer purchase behavior **Optimization methods** are the numerical techniques using to redesign a system or process. Optimization methods can be applied to improve performance according to a certain measure **Neural networks** refer to computational models based on biological neural networks and used for detecting patterns in data which can be used for pattern recognition, image analysis, optimization and adaptive control. Machine Learning: Machine learning is an artificial intelligence technique which allows computers to adapt behavior based on empirical data in order to making intelligent decisions based on information **Predictive modelling** uses a set of models in order to predict the probability of an event occurring, which can be applied for example in order to predict the potential that make a customer can be cross sold another product **Cluster analysis** uses techniques allow to turn a diverse group into a smaller with similar characteristics, it can be used for segmenting consumers into groups to perform better marketing campaigns and projects.	analyzing customers, we can personalize product recommendations to increase customer satisfaction. **Inventory management:** Big data analytic tools can help improve inventory management. For example, combining data of sales histories and seasonal sales can improve stock forecasting and predict changes in demands. Also, a retailer can automate stock replenishment by analyzing data such bar code systems which can reduce stock delay. **Price optimization:** The granularity of data of sales and pricing can be used to analyze market demands on price or product changing, which can then be derived to get an optimal pricing decision. **In-store behavior and customer sentiment analysis:** Retailers can collect information on customer's in-store behavior such as footpath and time spent in different parts of the store, and the data collected can be analyzed to improve store layout, shelf positioning and product mix. Also, data on social media about customers' reactions can help decision makers to monitor marketing campaigns.	However, there are some barriers to using big data analytics such as the privacy of information and scalability of analytic algorithms. In order to help analyze big data, retailers can use analytic techniques and technologies to help analyze big data in order to help with supporting decision making.

Paper IV "Machine Learning Methods for Demand Estimation"

Research Problem	Algorithm/Methodology	Result	Further analysis
The paper compares methods of modeling consumer behavior to	Eight different models are tested if suitable for estimating demand; linear regression, the	Machine learning algorithms bridge the gap between parametric models with user	There is a concern about the relative paucity of econometric theory for

standard econometric models that are used by practitioners to study demand.	conditional logit and six machine learning methods; namely stepwise regression, forward stage-wise regression, LASSO, SVM(Support Vector Machines), bagging, and random forests. A method proposed by Bates and Granger which dates back to 1969 is also discussed to be useful to implement. The models created are treated as regressors and together they form a combined model by regressing the dependent variable, which is the output that is trying to be predicted, on to the prediction of each component model.	selected covariates and completely non-parametric methods. Further it is mentioned that linear regression can improve the predictions with very little extra work.	machine learning models. In related work (Bajari et al., 2014), asymptotic theory results for rates of convergence of the underlying machine learning models were presented. While several of the machine learning models have non-standard asymptotic, with slower-than-parametric rates of convergence, the model formed by combining estimates retain standard asymptotic properties. This simplifies the construction of standard errors for both parameters and predictions, making the methods surveyed even more accessible for the applied practitioner.

Paper V_"A Multivariate Intelligent Decision-Making Model for Retail Sales Forecasting"

Research Problem	Algorithm/Methodology	Result	Further analysis
This paper addresses a multivariate sales forecasting problem which forecasts the overall sales of a retail product based on its early sales volume, which is helpful for related retail enterprises to make scientific and reliable replenishment forecasting and thus improve the performance and efficiency of their retail supply chains.	An effective multivariate intelligent decision-making (MID) model is developed to provide forecasts by integrating a data preparation and preprocessing module, a harmony search-wrapper-based variable selection (HWVS) module and a multivariate intelligent forecaster (MIF) module. The HWVS module selects out the optimal input variable subset from given candidate inputs as the inputs of MIF. The MIF is established to model the relationship between the selected input variables and the sales volumes of retail products, and then utilized to forecast the sales volumes of retail products.	The MID model can tackle the investigated multivariate sales forecasting problem effectively and it is statistically significant that the proposed model could provide much superior performance over the IELM model and the generalized linear model. It is also showed that the HWVS module can effectively find the appropriate variable input by eliminating reluctant and irrelevant inputs whichever candidate input variables are used, which results in less model parameters and higher forecasting accuracy. The proposed model is effective and widely applicable to multivariate sales forecasting problems.	Future research will focus on utilizing the proposed MID model to handle other multivariate forecasting problems, such as multivariate time series forecasting problem, and compare the performance of proposed model with multivariate time series forecasting model. Besides, the effects of different variable selection methods on the forecasting performance of MID model should be compared for further improvement.

Paper VI "An Intelligent Fast Sales Forecasting Model for Fashion Products"

Research Problem	Algorithm/Methodology	Result	Further analysis
The paper forecasts sales in the fast-moving industry of fashion by using single-hidden-layer feedforward neural networks (SLFN) called extreme learning machine (ELM).	In this paper, a method employing extreme learning machine (ELM), with the combination of statistical methods is used for forecasting. ELM not only learns much faster with a higher generalization performance than the traditional gradient-based learning algorithms but also reduces learning time of ANN dramatically and it even makes it possible to apply ELM in real-time applications such as real-time control. The forecasting accuracy and time cost of the ELM model are both explored.	Considering both the computation time and the stability of ELM, an appropriate value of parameter P must be found. Referring to (Sun et al., 2007), P = 100 is a reasonable parameter for ELME.	Future research will be conducted to explore further on how an intelligent fast forecasting model can be developed for time-series forecasting with different targets, such as consumer products and financial indices.

Paper VII "Prediction of Consumer Purchasing in A Grocery Store Using Machine Learning Techniques"

Research Problem	Algorithm/Methodology	Result	Further analysis
Analysis based on linear models are insufficient to satisfy the requirement of academics and practitioners for prediction of customer's purchase behavior, among the development of machine learning techniques.	This paper employs two representative machine learning methods: Bayes classifier and support vector machine (SVM) and investigates the performance of them with data from a mid-size supermarket in Japan.	SVM is used to apply to forecast purchase behavior, which was independent of the distribution and relationship of variables even though they were linear or nonlinear aspect of variables. In the numerical example, SVM demonstrated better forecasting performance related to linear discriminant analysis, logistic regression analysis and even Bayes classifier.	Based on this work, better method needs to be researched for the highest accuracy level possible for consumer behavior extraction.

Paper VIII "Predicting Sales In A Food Store Department Using Machine Learning"

Research Problem	Algorithm/Methodology	Result	Further analysis
Sales in a food store department	This study aims to compare three machine learning methods for sales prediction in the food industry: Multilayer Perceptron (MLP), Support Vector Machine (SVM) and Radial Basis Function Network (RBFN). The methods were compared in	There is a statistically significant difference between the SVM, MLP and RFBN when predicting the sales in a food store department. The SVM performed lower error measures than the other two methods.	Since this study used limited data, thus, one could hardly draw the conclusion that the SVM is always the most accurate method to use for sales prediction in a food store department. However, the result of this

	terms of their prediction accuracy on daily sales in a food store department. The performances of the models were determined using the performance measures: Mean Average Percentage Error (MAPE) and Root Mean Squared Error (RMSE).		study can indicate what methods to look at when implementing machine learning methods to predict sales in the food industry.

Paper IX "Predicting Sales Revenue by Using Artificial Neural Network in Grocery Retailing Industry: A Case Study in Turkey"

Research Problem	Algorithm/Methodology	Result	Further analysis
The paper aims to forecast the sales revenue of grocery retailing industry in Turkey with the help of grocery retailers marketing costs, gross profit, and its competitors' gross profit by using artificial neural network.	Artificial neural networks are models which are used for forecasting because of their capabilities of pattern recognition and machine learning. ANN method is used to forecast the sales revenue of upcoming period.	According to results there are high similarities between forecasted and actual data. Forecasted results of this study are bigger or smaller than the actual data for only 10%. Because of this high accuracy, companies at grocery retailing industry in Turkey can use ANN as a forecasting tool.	Other factors that could affect sales revenue can be also put into the mix for further research.

Paper X "The study of a forecasting sales model for fresh food"

Research Problem	Algorithm/Methodology	Result	Further analysis
As fresh food products have time effectiveness, the purpose of this research is to discuss and develop a mechanism for controlling the order and managing the stock for CVSs, a fresh food manufacturer and retailer in Taiwan.	The ''Ordinary day and holiday moving average method" and ''back-propagation neural network" were proposed and tested based on the operating characteristics of business circle and sale forecasting.	In the process of sales, products may not show stable sales amounts (stationary series) due to their PLC. When other factors are involved, some products' sales amounts will continuously increase (ascending series) or decrease (descending series). In addition, with more days accumulated, BPNN will obtain smaller errors and MSEs, which means higher precision.	More methods could be researched to get higher accuracy.

DATA

We are using a Kaggle dataset that includes sales information for a grocery chain based in Ecuador. Brick-and-mortar grocery stores are never too sure about purchasing and sales forecasting. Retailers are often over or understocked as a result of over or under-prediction. The problem becomes more complex as retailers add new locations with unique needs, new products, ever transitioning seasonal tastes, and unpredictable product marketing. The retailer in focus, Corporación Favorita operates hundreds of supermarkets, with over 200,000 different products on their shelves.

Table1: Data used in study

Variable	Type	Level	Description
id	Integer		Record ID
date	Factor	4	Date
store_nbr	Integer		Store Number
item_nbr	Integer		Item Number
unit_sales	Numeric		No. of units sold
onpromotion	Logical		On promotion or not
holiday_type	Factor	6	Transferred holiday or extra days that are added to a holiday
holiday_locale	Factor	3	Local or regional
holiday_locale_name	Factor	24	Name of location
holiday_description	Factor	103	-
holiday_transferred	Factor	2	A holiday that is transferred officially falls on that calendar day but was moved to another date by the government.
item_nbr	Integer		Item Number
item_family	Factor	33	Product Family
item_class	Integer		Product Class
item_perishable	Factor	2	Yes / No
dcoilwtico	Numeric		Daily oil price
store_nbr	Integer		Store Number
store_city	Factor	22	City name
store_state	Factor	16	State name
store_type	Factor	5	Type name
store_cluster	Integer		Cluster for store
date	Date		Date
transactions	Integer		Number of transactions

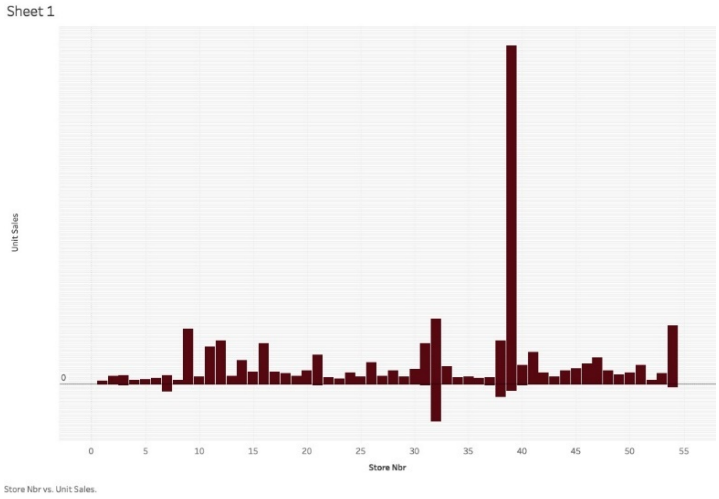
METHODOLOGY

Accumulated retail data for unit sales was available at state, city, store and item level for each day between 2013 and 2017. Items are further classified by family, class and perishability. Each store was classified by its client defined cluster, type and location. Furthermore, holidays relevant by location and oil prices over the span of observation time were also included.

Our main focus was to predict the sales of Bread/Bakery products as they are perishable in nature, making it important to understand the demand accurately to avoid shortage as well as wastage. For our study, we extracted data for one year, spanning across August 2016 to August 2017 as sales trends are best predicted by recent trends.

The data available needed some preprocessing steps for it to be compatible with the different modelling techniques.

- Pre-processing:
    - Pulled out observations relevant to the selected time frame (Aug-16 to Aug-17) and family of products (Bread / Bakery)
    - Identified outliers in the data through exploratory data analysis and dropped observations where sales is less than 0 (indicating returns) or greater than 300



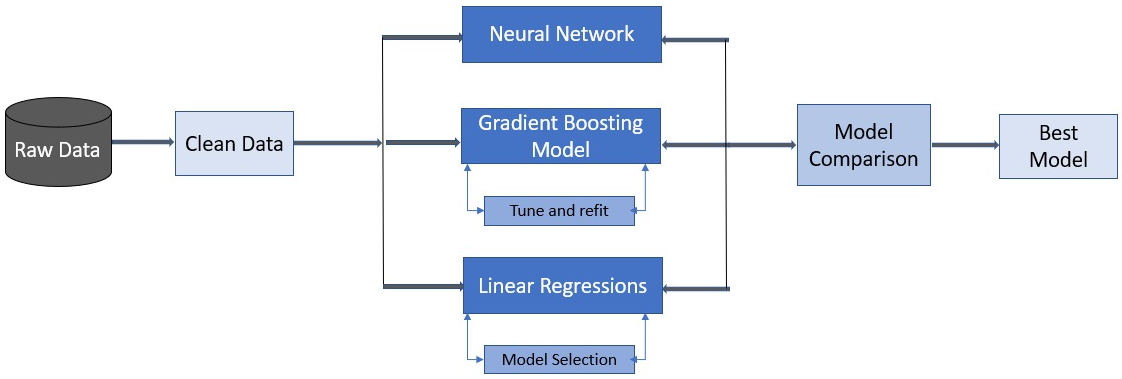
From our data analysis, we see that Q3 + 1.5*IQR = 20.7, but we also observe some trends in the classes of items with high sales as they could indicate specific trends.

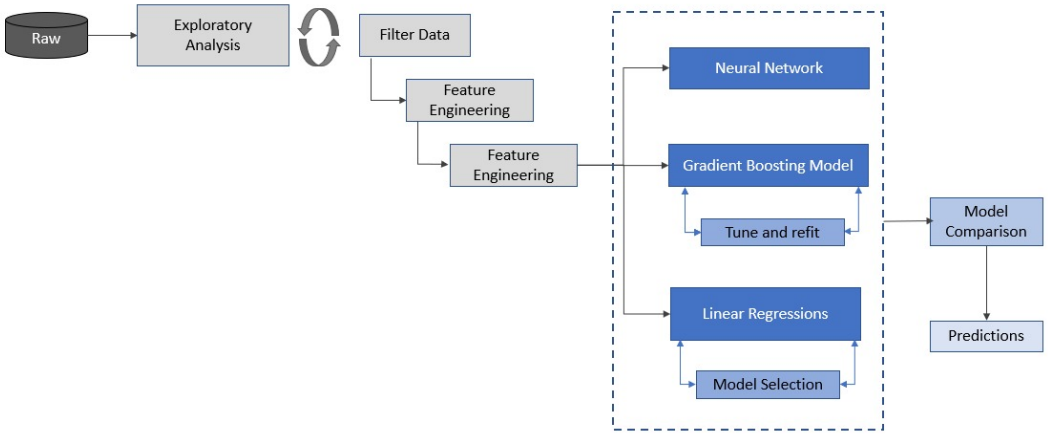
    - Developed new features like flags to identify local and national holidays, month etc, to determine their impact on sales
    - Dropped irrelevant variables like product class and perishability since they are common to all observations for the selected business case
    - One-hot encoded all categorical variables and dropped columns based on correlation greater than 0.85, near zero variance and presence of linear combinations between them

- We train the models using different learning techniques and use a 6 – fold cross validation approach to get a good balance between bias and variance. K = 6 allows for a fold size that does not take too long to run while also leading to good performance measures
- We partitioned the data using the createPartition() function from the caret package as it creates sets representative of the data distribution.

For this study, we partitioned using a 75:25 ratio. The data set is large enough to ensure good representation in both sets in with this ratio and it also does no lead to exclusion of certain observation types

- We used RMSE and R-square as measures of estimation of model performance. We choose these because we are predicting a continuous variable. RMSE gives a measure of error of prediction in terms of how far each observed point lies from the expected value on an average. R square indicates how much of the variation is explained by the resulting model or how well the model fits the data
- Provide a diagram created in PowerPoint or Visio that shows the steps you took (e.g. pulled data from DB, created new features, pre-processed data (how?), partitioned data, build model, evaluated models, etc.) This should make it crystal clear to the reader what your entire workflow does and help you explain to others in a PowerPoint or poster presentation later on. Detail here is a good thing. Below are two examples from previous projects.





MODEL(S)

Linear Regression

In linear regression, we fit a model based on the relationship between the dependent variable and the set of independent variables. The model identifies the best model by minimizing the mean of squared errors between the observed and predicted values. Since this method assumes the relationship to be linear, it tends to have a high bias, which we counter by training the model on 6 folds of data. Further we used backward as well as forward selection methods to identify the most important variables and choose the model with best performance measures. We differentiated the least squares formula and equate it to zero to get the model coefficients, leading to a closed form given by:

$$b_1 = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)2}$$

$$b_0 = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{n\sum x_i^2 - (\sum x_i)2}$$

Gradient Boosting Method

In this technique, predictors are chosen using decision trees which essentially divide the data set into smaller data sets based on the descriptive features until you reach a small enough set that contains data points that fall under one label with characteristic properties.

- **Boosting** is an ensemble technique in which the predictors are not made independently, but sequentially.
- **Gradient Boosting** is an example of boosting algorithm which employs the logic in which the subsequent predictors learn from the mistakes of the previous predictors. So, the intuition behind gradient boosting algorithm is to repetitively leverage the patterns in residuals and strengthen a model with weak predictions and make it better.

The objective function of the algorithm is to minimize the residual error between the predicted value and the true observed value.

- Tuning Parameters:
  - **n.trees** – Number of trees
  - **interaction.depth (Maximum nodes per tree)** - number of splits it has to perform on a tree
  - **Shrinkage (Learning Rate) –** It is considered as a learning rate.
  - **n.minobsinnode -** the minimum number of observations in trees' terminal nodes
  - **bag.fraction (Subsampling fraction) -** the fraction of the training set observations randomly selected to propose the next tree in the expansion.
  - **train.fraction -** The first train.fraction * nrows(data) observations are used to fit the gbm and the remainder are used for computing out-of-sample estimates of the loss function

Model Approach:

1. Initialize model with a constant value for n = number of observations:

$$F_0(x) = arg_\gamma \min \sum_{i=1}^{n} L(y_i, \gamma)$$

2. For m = M ; M: number of iterations (for i = 1 to n)
   - Compute pseudo residuals

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

   - Fit a base learner (e.g. Tree) $h_m(x)$ to pseudo residuals (train using training set) $\{(x_i, r_{im})\}_{i=1}^{n}$
   - Compute multiplier γ_mr by solving the following one-dimensional optimization problem:

$$\gamma_m = arg_\gamma \min \sum_{i=1}^{n} L(y_i, F_{m-1}(x) + \gamma h_m(x))$$

   - Update the model

$$F_m(x) = F_{m-1}(x) + \gamma h_m(x)$$

3. Output $F_m(x)$

Neural Networks

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A neural network can learn how to do tasks based on the data given for training and it can create its own organization or representation of the information it receives during learning time. Our model implements one hidden layer that trains through responsibility assignment factor and we specify are tuning parameters through a tuning grid as follows:

- Size = [1,2,3,4,5,6,7,8,9,10]
- Decay = [0.1,0.2,0.3,0.4,0.5]

Our business problem deals with predicting the volume of sales of bakery products per day, which essentially is a continuous variable. All of these methods perform well on regression type of problems GBM and neural network are non-parametric techniques that learn well in our problem. We can see from the results later that train and test models have similar performance, indicating no over-fitting, which is often a concern in non-parametric models, thus proving to be candidate models.

RESULTS

In our analysis, we tried 4 different models. Comparing all the models based on the RMSE, gradient boosting performs the best. Here, the performance for the training and test models are similar, which make them candidate models.

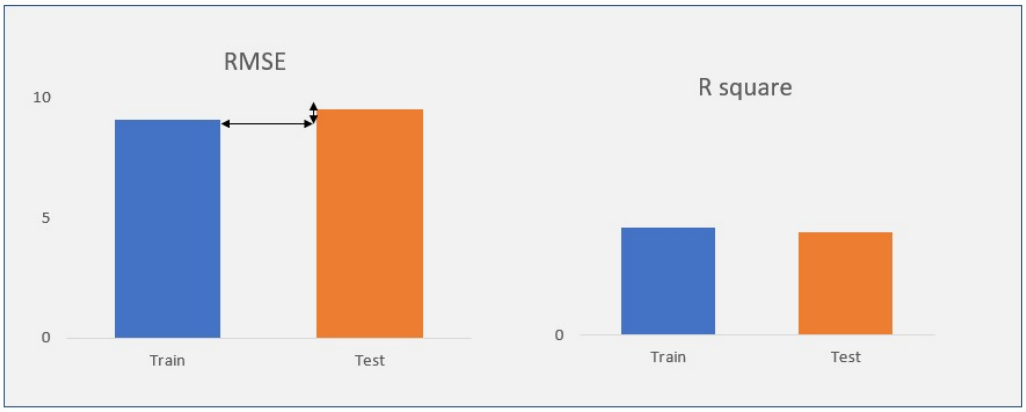
Results from different models

Model	RMSE_Train	RMSE_Train
Linear Regression	11.47	11.97
Gradient Boosting	9.07	9.5
Neural Network	9.28	9.82
Support Vector Machine	Still running(from 5 hours)	Still running(from 5 hours)

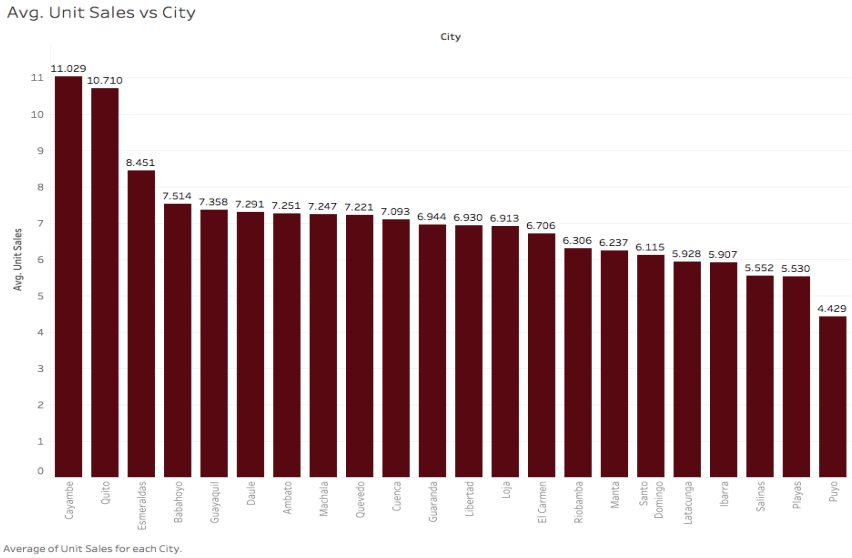
While SVM is promising to give good results for our business problem, it takes a long time to train the model for a dataset of this size. We intend to track the performance for the once the model is ready. Time take for different models to run:

Model	Time taken for the model to train
Linear Regression	20 minutes
Gradient Boosting	40 minutes
Neural Network	90 minutes
Support Vector Machine	Still running(from 5 hours)

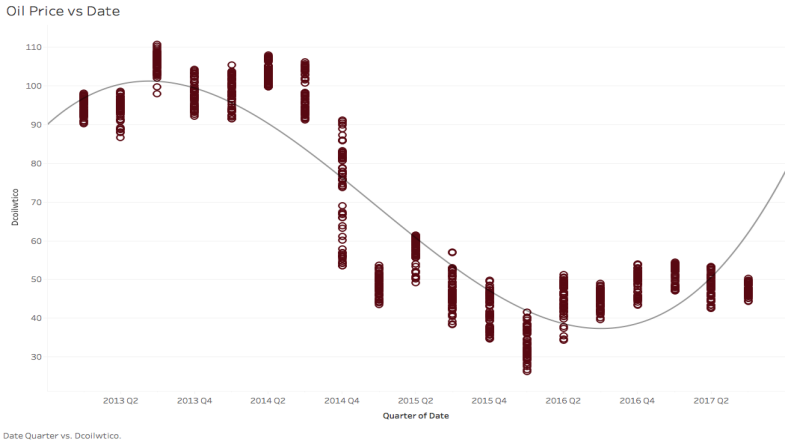
Results from Gradient Boosting Model:

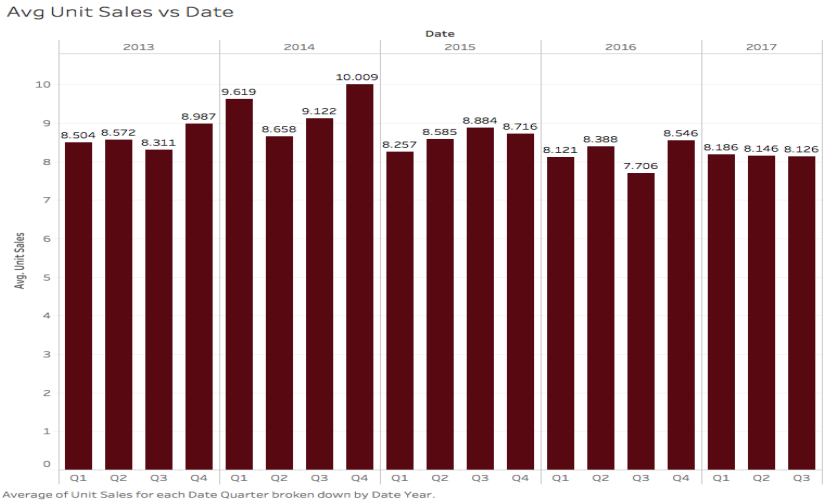


	Set	RMSE	Rsquared	MAE
1	Train	9.071796	0.5078484	4.841416
2	Test	9.498994	0.4857778	4.853988



Plots from Exploratory Data Analysis:





CONCLUSIONS

Predicting sales especially in grocery stores is of paramount importance to a store owner. Looking at the future predictions for a particular item, we can determine inventory to prevent overstocking or stockouts. Also, revenue of stores can be improved by looking at items which can perform better and promoting those items.

In the current situation, owing to the data set size and computational difficulties associated with it, we have chosen gradient boosting as our optimum model. However, given more time we can optimize our neural network model.

Further, this prediction problem can be considered as a time series forecast to utilize the relations between the various time series parameters to improve our prediction. We can use models like ARIMA, HoltWinters.

REFERENCES

1. Bajari, Patrick, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang (2014). “Demand Estimation with Machine Learning and Model Combination.” University of Texas at Austin.
2. Bajari, Patrick, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang (2015). "Machine Learning Methods for Demand Estimation." American Economic Review, 105 (5): 481-85.
3. Chen-Yuan Chen, Wan-I Lee, Hui-Ming Kuo, Cheng-Wu Chen, Kung-Hsing Chen (2010). “The study of a forecasting sales model for fresh food”. Expert Systems with Applications, Volume 37, Issue 12, Pages 7696-7702, ISSN 0957-4174.
4. Dilek Penpece and Orhan Emre Elma, "Predicting Sales Revenue by Using Artificial Neural Network in Grocery Retailing Industry: A Case Study in Turkey," International Journal of Trade, Economics and Finance vol.5, no.5, pp. 435-440, 2014.
5. Hamza Belarbi, Hamid Bennis, Abdelali TAJMOUATI, El Haj Tirari Mohammed (2016). “Predictive Analysis of Big Data in Retail Industry”. 1st International Conference on Computing Wireless and Communication Systems (ICCWCS-2016)
6. Paták, Michal & Branská, Lenka & Pecinova, Zuzana. (2015). "Demand Forecasting in Retail Grocery Stores in The Czech Republic". 10.5593/SGEMSOCIAL2015/B22/S7.089.
7. Patrícia Oliveira; Fátima Rosa; Miguel Casquilho (2012). "Optimization of the sales forecast algorithm for a supermarket supply chain". Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa (Portugal)
8. Siwerz, R., & Dahlén, C. (2017). Predicting sales in a food store department using machine learning (Dissertation). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-208888.
9. Yu, Yong & Choi, Tsan-Ming & Hui, Patrick (2011). “An intelligent fast sales forecasting model for fashion products”. Expert Syst. Appl.. 38. 7373-7379. 10.1016/j.eswa.2010.12.089.
10. Y. Zuo, K. Yada and A. B. M. S. Ali (2016). "Prediction of Consumer Purchasing in a Grocery Store Using Machine Learning Techniques," 2016 3rd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), Nadi, 2016, pp. 18-25.
11. Z.X. Guo, W.K. Wong, Min Li (2013). “A multivariate intelligent decision-making model for retail sales forecasting”. Decision Support Systems, Volume 55, Issue 1, Pages 247-255, ISSN 0167-9236.