Retail Data Analysis

Department of Computer Science, Stony Brook University

***Abstract*—The purpose of this project is to analyze the retail data provided by company, Costello's Ace Hardware and provide useful insights which will be beneficial for the company to increase its net sales and improve the implemented customer practices.**

I. Introduction

Retail data demands analysis of all the components incorporated in a business, from sales, inventory, revenue to customer data, transaction details, and coupon codes promotions. Data-driven insights are quite beneficial and give an edge over competitors by tracking the history of customers' purchase patterns to make smarter decisions in the impending time. This is what we leveraged in our models. Our aim is to build models which can derive useful insights and patterns from customer data like the product preference for purchase, time of the day/year when the sales were high, the stores which attract customers from particular zip codes, etc. The whole purpose for this exercise is to assist the stores to grow revenue and business's profitability along with customer-centric experiences.

II. Datasets

A. Costello's Ace datasets

We have been provided with costello_ace_2015-2016 and costello_ace_2017-2018 datasets for analysis. Both the datasets have real transactional data which we have utilized in drawing insights. Few highlights are as follows:

- 39 features depicting the Transaction Date, Receipt Number, Customer Details, Store & Item information, Sales data, etc.
- Customer Number is the unique key for all the different customers
- Transaction date is available no missing values
- After removing redundant data, we have >25L records and all the features are of 'object' type
- 'Return Code' has the maximum number of null values
- All the fields are of object type
- Many numeric fields like Net Sales, Net Sales Units, etc. have 'commas'
- Field Zip Code has alphanumeric values
- Store + Receipt Number is unique
- Item Number and Item Description holds 1:1 mapping.
- Net Sales value is negative if returned value is greater than the purchased value of the transaction
- Zip Code is used to represent the neighboring area of the customer
- Clerk Information is provided as to understand who provides the maximum discounts
- Line Number is the order in which cashier scans the items

B. Online Retail dataset from UCI Machine Learning repository

This dataset was incorporated to understand the segmentation of customers in order to predict the kind of items they will buy in future based on their segmentation.

C. Holiday Dataset from Data.gov

To better analyse the blips in the sales during a period of year, we have accommodated Holiday dataset from Data.gov in our analysis to understand if there is a correlation of off-days/holiday months and the fluctuations in the sales records.

D. Crime Dataset from data.world

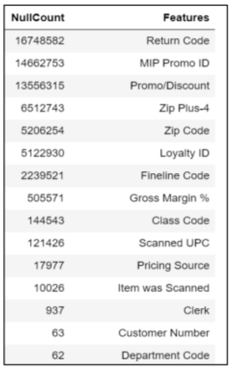
We have introduced crime dataset from data.world to understand why certain products are highly popular in a particular region. Is there some kind of correlation that exists which tend the denizens to buy such items more?

E. Weather Dataset from NOAA

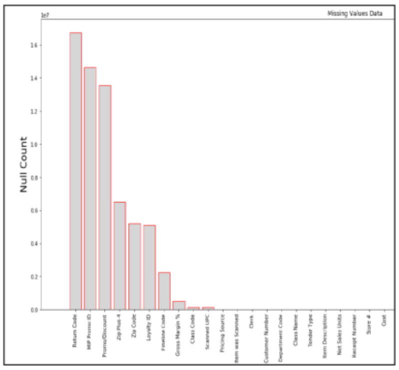
We have leveraged weather dataset from National Centers for Environmental Information to figure out if there is a dependency of sales on the weather of a region. Is the drastic change in weather impacting sales of the store in that region?

III. Preprocessing of the Datasets

For the clean-up task on the Costello's datasets, we have removed the records with blank Customer Number as they can't be assigned any arbitrary number. We have also removed the duplicate entries from the all the datasets. Net Sales and other numerical fields were converted to float datatypes for calculations. As the data involved is huge, we ran garbage collection at frequent intervals to avoid crashing the model.



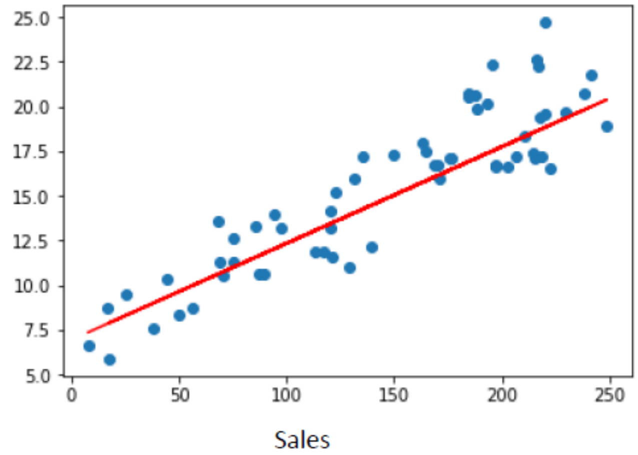
NullCount	Features
16748582	Return Code
14662753	MIP Promo ID
13556315	Promo/Discount
6512743	Zip Plus-4
5206254	Zip Code
5122930	Loyalty ID
2239521	Fineline Code
505571	Gross Margin %
144543	Class Code
121426	Scanned UPC
17977	Pricing Source
10026	Item was Scanned
937	Clerk
63	Customer Number
62	Department Code



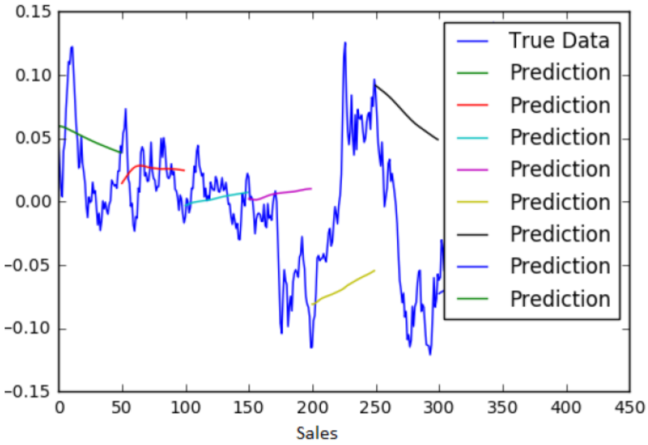
IV. JOURNEY THROUGH THE PROJECT

Project Proposal: We started with analysing the Costello's data for the year 2017-2018 and set certain goals based on the initial study of the data. The aim was to build a number of models which will help Costello's Ace to improve its standing in the market. For this, we desired to predict the sale time and the stock to be maintained for the products of high-desirability. As Costello's Ace values its customers, we wanted to help the company to retain its most frequent customers by segregating the frequent buyers and rewarding them with discounts. Also, we wanted to include external datasets to better understand the retail market and advise Costello's Ace based on the persisting market trends.

Mid-progress: In order to implement our first target, we initiated our analysis by building a model that predicts the net sales of the stores considering the popular time of day/year. Since the model holds linear relationship, we implemented Lasso CV with RMSE 0.218.



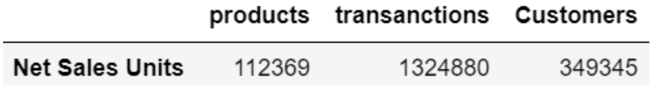
To enrich our model further, we implemented Long Short-Term Memory(LSTM) using TensorFlow library 2 to predict the net sales by reducing overfitting and minimizing the loss.



Final Project: To continue towards the tail of our analysis, we continued our analysis of the original datasets along with the external datasets to have more advanced techniques implemented for customer segmentation and market basket analysis.

V. USEFUL INSIGHTS

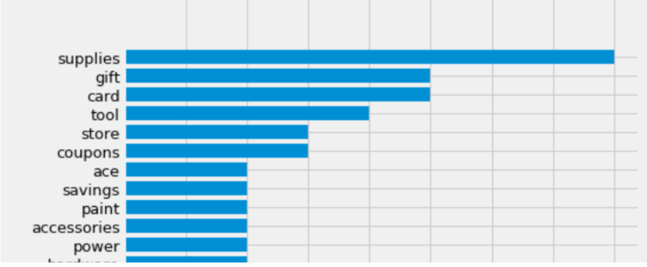
To begin with, we calculated the total number of transactions and found:



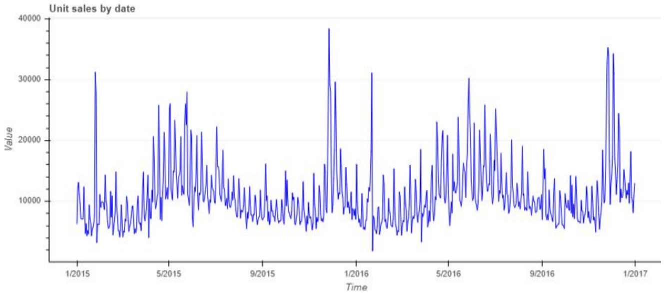
	products	transanctions	Customers
Net Sales Units	112369	1324880	349345

This shows that total of 1324880 transactions were reported but there are only 349345 customers with 112369 products. As we don't have records of cancelled orders, it is possible that multiple items were bought in a single transanction.

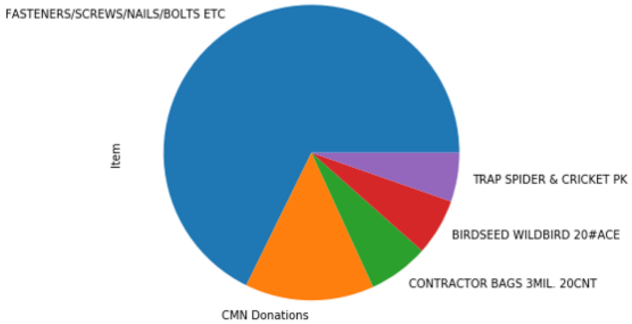
Popular products by occurrances of common words in the names of products. This is helpful in analyzing if same customers have bought similar kinds of products or not.



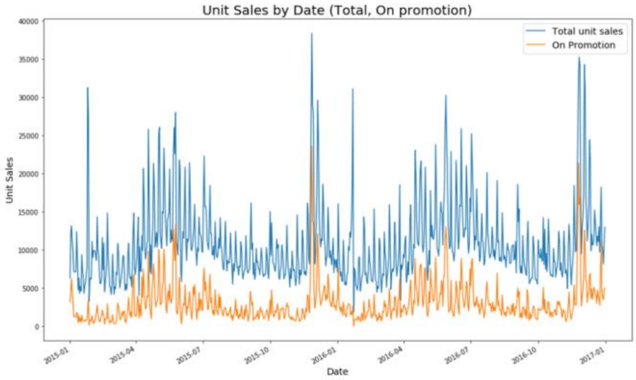
Now, coming to the increase in the sales during December/January period which is predictable due to Christmas/New Year holiday season. There is a small blip during July which may be due to Independence Day celebrations. We have utilized holiday dataset here to understand that due to off-day/holiday, there is a shortage of labor which therefore, is the most probable reason for high sales.



We have leveraged the crime dataset to analyse that Zip Code – 11758 (Massapequa) has the highest sales in items like FASTENERS/SCREWS/NAILS/BOLTS ETC and KEYS. On further investigation, we found that Massapequa has considerable crime rate of 1100/100K.



Investigating on these grounds further, it is observed that Zip Code – 11710 (Bellmore) has lower crime rate but the weather conditions are extreme. Weather dataset was very helpful in displaying the windy nature of the city and therefore, the sales for particular products rise due to requirements.



VI. METHODS AND RESULTS

A. *Lasso Cross-Validation*

We have created a baseline model for predicting net sales of the stores using Lasso CV by implementing the below equation:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Where, tuning parameter, λ controls the strength of L1 penalty which is the absolute value of the magnitude of coefficients. The main goal of using this as the base model was to obtain a subset of predictors that minimizes the prediction error for a quantitative response variable.

B. *Long Short-Term Memory (LSTM)*

We utilized the artificial recurrent neural network (RNN) architecture to classify, process and make predictions based on time series data, since there can be lags of unknown duration between important events in a time series. Below equation is involved in our sales forecasting task:

$$\mathrm{S}_{t+n} = \mathscr{F}_m^n\Big([\mathrm{X}_{t-m}, \mathrm{X}_t], [\mathrm{S}_{t-m}, \mathrm{S}_t], (\mathrm{Z}_t, \mathrm{Z}_{t+n}]\Big)$$

Here, n>0 and m>0 are corresponding to the number of steps from the current time to the predicted future and number of steps that are considered from the history to predict the future, respectively.

C. *Customer Segmentation*

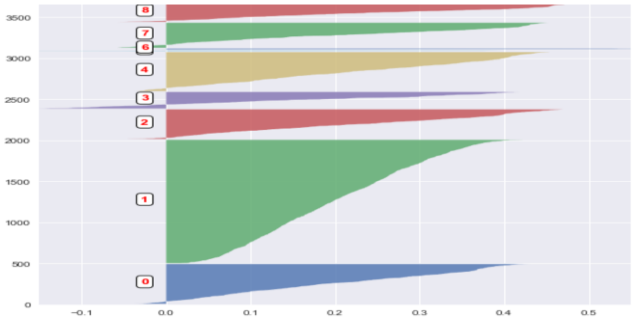
We wanted to perform customer segmentation on the given dataset to categorize the customers in a particular segment based on their purchasing patterns. Our aim was to build a model which can possibly predict the kind of items which are highly likely to be purchased by a particular segment of customers. Our simplest approach to this problem statement was to Natural Language Toolkit (NLTK) Bag-of-words (BoW). Through this model, we extracted common product names from the datasets and then converted them into matrix of occurrence of words. We followed below algorithm for our model:

Step1: Tokenize product name: We started by removing stopwords from the names. Stopwords are words which do not contain enough significance to be used in our algorithm. Tokenization is an act of breaking up a sequence of strings into pieces, such as words, keywords, phrases, or even whole sentences. Punctuation marks are discarded in this process.

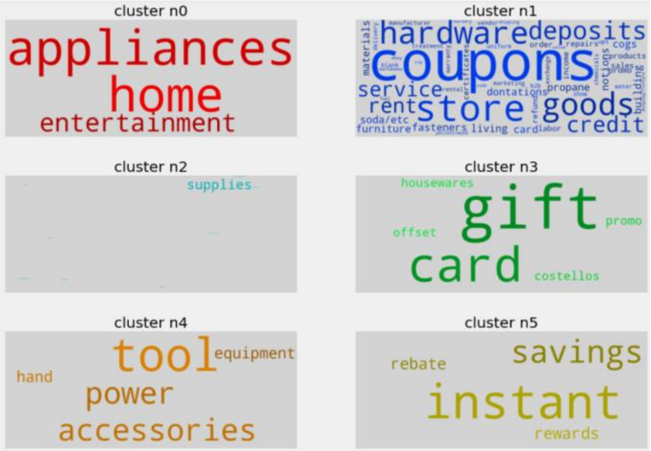
Step2: Apply tokenization to all the product and department names.

Step3: Build vocabulary and generate matrix of word of occurrences.

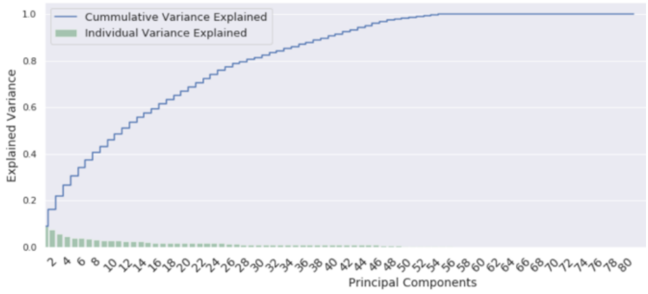
We used only preserved words, like Binary Bag of Words and then, converted this into a product matrix with different products as rows and different words as columns. A cell contains a ‘1’ if a particular product has that word in its description, else it contains ‘0’. Then, we used this matrix to categorize the products. Thereafter, we used K-means clustering to categorize similar products under one cluster and calculated the intra-cluster silhouette distance:



Analyzing the word cloud clusters demonstrate that all the related items are clustered together which will help facilitate customer segmentation on the basis of their purchase trend. For ex., gift card, promotional coupons, discounts related items are clustered in Cluster n1.



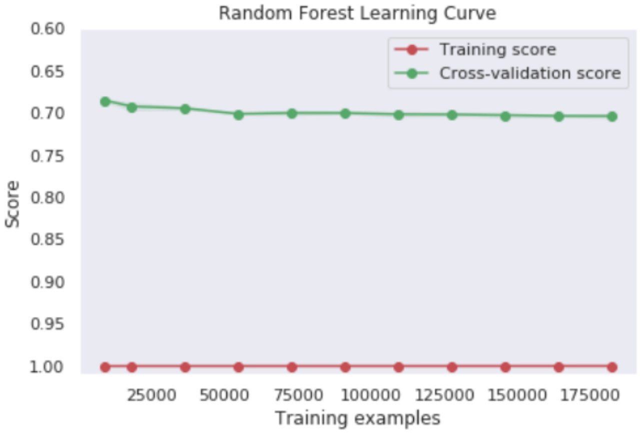
For dimensionality reduction, we performed Principal Component Analysis (PCA) and explained the amount of variance in the model:



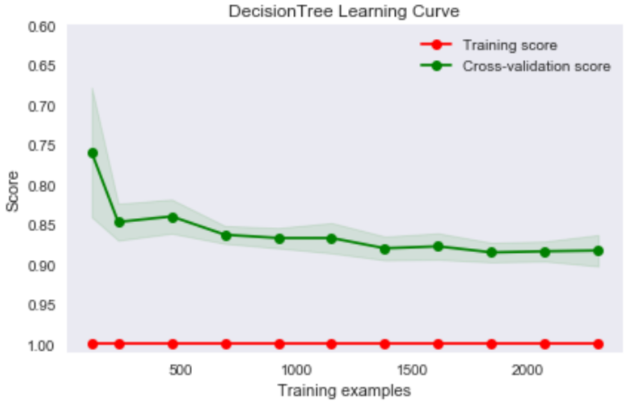
We, thereafter, used the already generated product categories and created few new features which tell us to which category the product belongs to. Time-based splitting is performed which will give us information about every customer on how much do they purchase, total number of orders, etc. We now proceed to build customer segments and focus on customers with more than one orders. Since this shows loyalty to the store, and the store would like to retain them by giving extra discounts.

We took the dataset, scaled the same, splitted it into training and test datasets and applied Random Forest Classifier, Decision Tree Classifier, and K-Nearest Neighbor Classifier to achieve the maximum accuracy.

Learning Curve of Random Forest Classifier (Acc: 51.85%)



Learning Curve of Decision Tree Classifier (Acc: 70.62%)



Now, we used Ensemble Modeling – Voting Classifier to collate the best performing models and achieved 73.13% accuracy.

D. Apriori Algorithm

After much investigation, we analyzed and observed the fact that if the products are sold in pairs, i.e. products that are closely related to each other or are frequently bought, may result in the increase of sales. This will benefit both the company as well as people residing nearby.

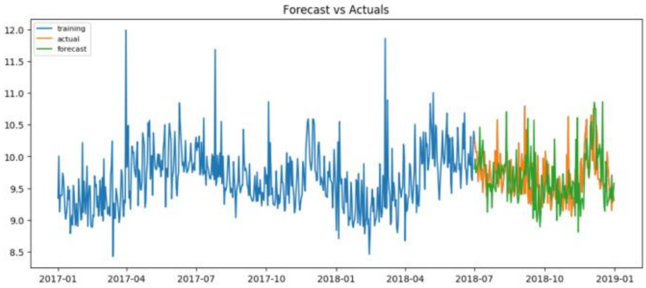
Market Basket Analysis is the process of extracting purchasing trends from the company's database records. This analysis takes into consideration the products that are bought by customers in a single transaction. According to the Apriori algorithm, all the sub-clusters of an uncommon itemset can't be frequent, or in other words, if an itemset doesn't provide the minimum threshold value, the itemset's parent groups do not provide the minimum threshold value, either.

Below is the algorithm followed by our model:

```
Apriori$(T, \epsilon)$
    $L_1 \leftarrow \{\text{large } 1 - \text{itemsets}\}$
    $k \leftarrow 2$
    while $L_{k-1} \neq \emptyset$
        $C_k \leftarrow \{c = a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a, \{s \subseteq c \mid |s| = k-1\} \subseteq L_{k-1}\}$
        for transactions $t \in T$
            $D_t \leftarrow \{c \in C_k \mid c \subseteq t\}$
            for candidates $c \in D_t$
                $count[c] \leftarrow count[c] + 1$
        $L_k \leftarrow \{c \in C_k \mid count[c] \geq \epsilon\}$
        $k \leftarrow k + 1$
    return $\bigcup_k L_k$
```

Where k-itemset represents the itemset containing k number of elements, Lk refers to the frequent itemsets with k elements and Ck signifies the frequent candidate itemsets with k elements.

In this model, we have not considered the items that are associated with discount coupons. We have implemented the above algorithm to understand two frequent products that are bought together. Based on our analysis, we observed that 'PAINTBRSH FOAM 3" JEN', 'PAINTBRSH FOAM 2" have been purchased frequently with confidence = 0.9963 and 'BIRDSEED WILDBIRD 20#ACE', 'PEANUT CRUNCH SUET' with confidence = 0.8876.



E. Long Short-Term Memory(Improved)

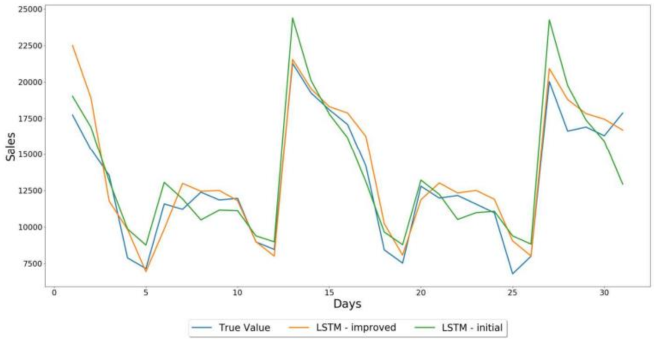
For this model, we conducted our analysis as a multivariate forecasting task, and therefore, we employed several other features apart from using the historical daily sales values. For the initial stage, we wanted to study how historical data can be employed to forecast the number of sales analogous to traditional time series forecasting tasks. The original dataset includes Date, State Holiday, Promotion Availability, Store Open/Close information and the Number of Customers.

To proceed with our model building, we decomposed the composite attribute Date into three separate features – day, month and year. Through empirical analysis, we identified that day of the week, promotion availability information and school holiday details are the best combinations of historical features that maximize the forecasting accuracy. For the implementation model, we used these three features to feed our model along with daily sales values. Now, we've extended our initial model employing the abovementioned information that described the future which is known to us ahead of time.

We've divided the dataset into training and test subsets store-wise and scaled them using min-max scaling. Scaling is imperative here as each feature was operating in significantly different intervals.

Now, in order to build an improved model of LSTM, we have incorporated new dataset and removed unnecessary and insignificant attributes. We realized that the LSTM model requires tuning too many hyperparameters and manually tuning each hyperparameter for the enormous search space is not a feasible task. The evaluation included 29 stores and needed tuning 13 hyperparameters for two different and needed tuning 13 hyperparameters for two different LSTM, forcing us to tune 13×29×2 hyperparameters if we can run each experiment exactly once. Therefore, the need to automate the hyperparameter optimization process became mandatory.

To automate the hyperparameter optimization process, we employed a Bayesian optimization based on the Gaussian Process (GP). Bayesian optimization finds a posterior distribution as the function to be optimized during the parameter optimization, then uses an acquisition function to sample from that posterior to find the next set of parameters to be explored. Since Bayesian optimization decides the next point based on more systematic approach considering the available data it is expected to yield achieve better configurations faster compared to the exhaustive parameter optimization techniques such as Grid Search and Random Search. Therefore, Bayesian optimization is more time and resource efficient compared to those exhaustive parameter optimization techniques, especially when we are required to optimize 13 parameters including 3 parameters with a continual search space.



The better performance of LSTM is due to its superior ability to model time-series features. Machine learning algorithms have no notion of the different time steps of data or any kind of time series specific information, they merely perform a regression task on the given data, whereas the LSTM understands the concept of times steps and are strong tools used extensively in time-series forecasting. LSTMs are capable of modelling long-range dependencies. The LSTM architecture contains a cell state in addition to a hidden state, that enables the LSTM to propagate the network error for much longer sequences while capturing their long-term temporal dependencies. LSTMs can also fit a wider range of data patterns compared to the traditional models. These factors have enabled the LSTM to produce more accurate forecasts compared to two conventional machine learning models. The reduction in error is significant (20%-21%) when considered the new features for sales forecasting.

To evaluate both LSTM and machine learning models, we used Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) as error metrics. We have employed RMSE for the hyperparameter optimization task of both LSTM and machine learning models. Considering y and ȳ respectively as the true sales and predicted sales, shown in Equations 1 and 2 are the respective equations for RMSE and MAE.

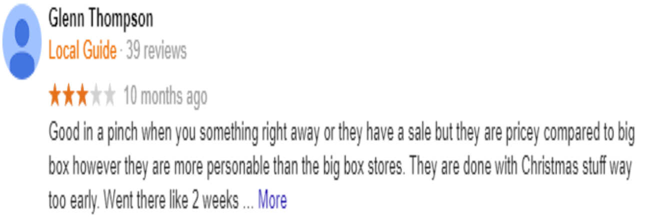
$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{n}}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)}{n}$$

Final accuracy achieved: 84.43%.

VII. Market competition and trends

We have studied and researched the company's market standing and the competitors challenging the Costello's Ace current revenues. For the motive of analysis, we scrapped data from Yelp and Google reviews and observed that the reviews provided by customers are highly satisfactory, however the prices offered by the store is higher than the neighboring competitive stores in the zone area.



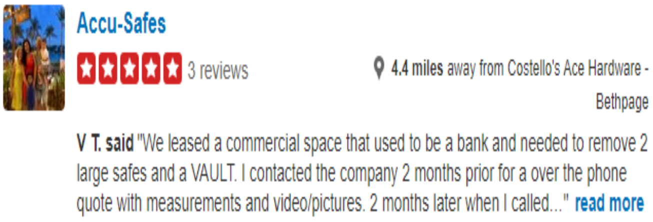
Glenn Thompson

Local Guide · 39 reviews

★★★☆☆ 10 months ago

Good in a pinch when you something right away or they have a sale but they are pricey compared to big box however they are more personable than the big box stores. They are done with Christmas stuff way too early. Went there like 2 weeks ... More

We have found customers with only this single concern since the same products are available at 'Big Boss Stores' at competitive rates.



Accu-Safes

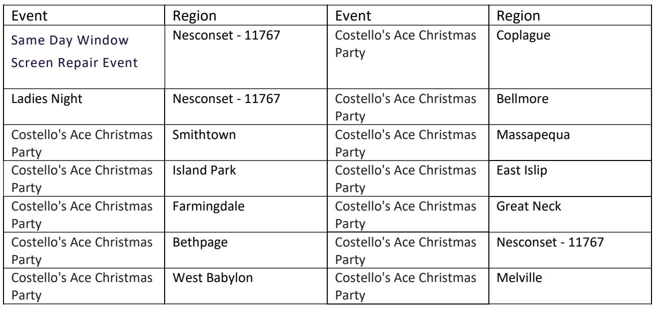
★★★★★ 3 reviews

4.4 miles away from Costello's Ace Hardware - Bethpage

V T. said "We leased a commercial space that used to be a bank and needed to remove 2 large safes and a VAULT. I contacted the company 2 months prior for a over the phone quote with measurements and video/pictures. 2 months later when I called..." read more

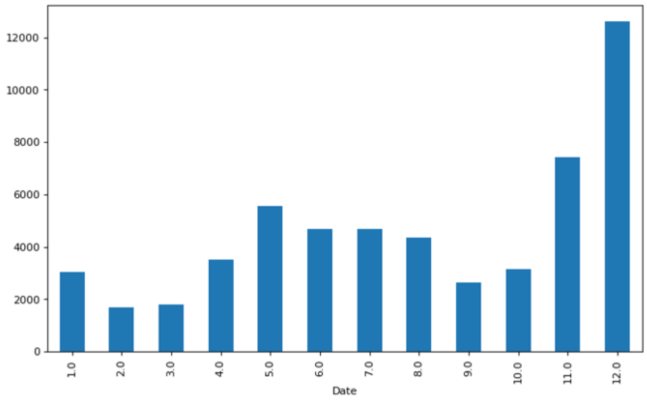
We performed this analysis for 2 areas with Zip Codes: 11751 and 11703 which is in Suffolk county since the retail sales for these areas is fairly less (<5k). Since there are a lot of frequent customers who buy different products, it is advisable that the company must either reduce the price of frequently bought items or sell products at a reasonable price by collaborating frequently bought items.

We came across another interesting observation while exploring the data on the web (Costello's Ace Events). Since the primary objective of the company was to increase the sales, we identified that several events are hosted by the company which accounted for an increase in its sales. The events are organized in only 11 regions and hence, the sales in these regions are the most.



Event	Region	Event	Region
Same Day Window Screen Repair Event	Nesconset - 11767	Costello's Ace Christmas Party	Coplague
Ladies Night	Nesconset - 11767	Costello's Ace Christmas Party	Bellmore
Costello's Ace Christmas Party	Smithtown	Costello's Ace Christmas Party	Massapequa
Costello's Ace Christmas Party	Island Park	Costello's Ace Christmas Party	East Islip
Costello's Ace Christmas Party	Farmingdale	Costello's Ace Christmas Party	Great Neck
Costello's Ace Christmas Party	Bethpage	Costello's Ace Christmas Party	Nesconset - 11767
Costello's Ace Christmas Party	West Babylon	Costello's Ace Christmas Party	Melville

As we can see from the graph, the events mostly took place during the months of November and December, and at the same time, an increase in sales for the region Nesconset – 11767 is experienced.



We have also seen that the sales had a peak during that period and therefore, if the events will reach on a broader scope, it will certainly impact the sales to a greater extent. If they can reach out to other regions, there is a high possibility of increase in the sales. This is just an assumption based on a small dataset scraped from Costello's website.

VIII. Future work

1. In the future, we are planning to explore the ability to incorporate multiple stores with a single LSTM to extract cross-series information to improve forecasting accuracy. We expect such features to improve time-series forecasting by comprehending the interdependencies between the stores such as competition, partnerships, market distribution etc. Moreover, it is interesting to investigate the importance of incorporating information that describes the future beyond the day being predicted. For instance, the customer buying behavior for a day can significantly affect the fact whether the store is going to be closed on the following day. Yet, the time-series models may not be able to anticipate such relationships without explicitly providing information that represents the future even beyond the day that is being forecast. Therefore, we will be exploring such extensions with our technique in the future.
2. As mentioned, there were not much analysis could be done based on promotional events since there was not much data available for exploration. In future, we can keep track of all the events conducted and create a dataset and how it impacted on the sales.

IX. Conclusion

In this report, we have successfully analyzed different aspects of the dataset. We have grouped several different segments of the data provided and created multifarious models for prediction and analysis.

Also, we advised the company to take several casual steps at their end to improve their sales in a few regions where are getting seriously challenged by other stakeholders.

As retail data analysis is very vast topic, it always demands more and more exploration. So, we will be continuing our research and investigation for the abovementioned future work.

X. References

i. Online Retail dataset from UCI Machine Learning repository
ii. Holiday data from Data.gov https://catalog.data.gov/dataset?tags=holiday
iii. Crime dataset from data.world https://data.world/datasets/crime
iv. Weather dataset from NOAA https://www.ncdc.noaa.gov/cdo-web/search
v. Google Reviews
vi. Yelp https://www.yelp.com/biz/costellos-ae-hardware-bellmore
vii. Events data https://allevents.in/ and https://www.costellosace.com/events/
viii. Wikipedia for Model and algorithm understanding https://en.wikipedia.org/wiki/
ix. https://pdfs.semanticscholar.org/a8d0/f4372c36b7b0472aaf2b8d120fb7040a75bd.pdf?_ga=2.215142224.694753852.1574980912-1436127251.1574980912
x. https://www-ai.cs.tu-dortmund.de/LEHRE/SEMINARE/SS09/AKTARBEITENDESDM/FOLIEN/Frequent_Itemset_Mining_Methods.pdf
xi. https://arxiv.org/pdf/1804.01182.pdf
xii. http://cs229.stanford.edu/proj2015/219_report.pdf