Corporación Favorita Grocery Sales Forecasting

Arizona State University, W.P. Carey School of Business

2021 Spring, Cohort A, Team 8

Team Members: Jinhang Jiang, Bhavana Patil, Dhruv Tyagi, Jinghuan Li

Executive Summary

This project was launched by Corporacion Favorita on kaggle.com three years ago. In this project, we were challenged to predict the unit sales for thousands of items sold at different Favorita stores located in Ecuador for the following 16 days starting on August 16, 2017. We also aimed to study which store had the most impact on unit sales and what other factors that influenced the unit sales of the Favorita stores in Ecuador.

Corporacion Favorita gave us one training and one testing file for analysis. The training data includes dates, store IDs and item IDs, whether that item was being promoted, as well as the unit sales from January 1, 2013 to August 15, 2017. We were also given six additional files that include supplementary information, such as daily oil price, individual transaction information, store and item information.

We used RStudio and Power BI Desktop as our analytic tools to perform the descriptive analysis, visualization, and predictive analysis on the sales data. During the analysis process, we encountered disk memory exhaust issues several times, and we also found that the sales data has significant changes through the years either due to natural disasters or business expansions. To improve the efficiency and relevancy of the analysis, we decided to chunk the data into two versions.

After studying the visualizations and the statistics, we learned that in the year of 2017, unit sales from Quito which is the capital of Ecuador accounted for 51% of the total unit sales in Ecuador. Item #502331 was the most popular item that was carried by the most stores in Ecuador while item #2011451 and #2015898 were tied for the least popular items.

We applied the LightGBM model to predict the unit sales with a testing dataset including 3,370,464 observations and 65 features. The final submission score was 0.529 and ranked the 710th place out of 1671 teams.

Background Introduction

Sales forecasting is always a complicated problem for the brick-and-mortar grocery stores. If you overestimate, grocers will be left with overstocked, perishable products. If you guess a little short, popular products can easily sell out, leaving you with money on the table and angry customers.

As retailers introduce new locations with specific demands, new items, ever-changing seasonal preferences, and volatile product marketing, the challenge becomes more complicated. Corporación Favorita, a big Ecuadorian supermarket chain, is aware of this as well. They own and run hundreds of supermarkets that store over 200,000 different products. Currently, they are performing arbitrary forecasting approaches with no evidence to back them up and little technology to carry out plans. Therefore, they want to study how machine learning will help them better serve consumers and build up a well-functioning Just-In-Time inventory system.

Data Description

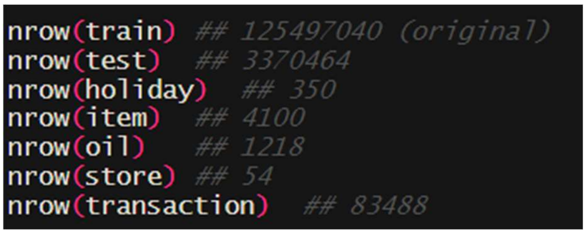
Originally we had seven data files, including holidays_events.csv, items.csv, oil.csv, stores.csv, test.csv, train.csv, and transactions.csv. The training data includes the target unit_sales by date, store_nbr, and item_nbr and a unique id to label rows. The “onpromotion” column tells whether that item_nbr was on promotion for a specified date and store_nbr. The testing data including the date, store_nbr, item_nbr combinations that are to be predicted, along with the “onpromotion” information. The full description of data may be found here or in “Appendix B.”

Because the original data file size was 4.7G which was beyond the computing power of our computer, and the data was dated all the way back to 2013 which highly likely has become irrelevant to our analysis. For example, Ecuador was struck by a magnitude 7.8 earthquake on

April 16, 2016. Citizens came together to help with relief efforts, donating water and other essential needs, which had a huge impact on grocery sales for several weeks after the earthquake. The data fell in this time frame is very likely to become outliers for predicting. Thus, we decided to chunk the data into a smaller and more relevant set. And we got two versions of the subset of the data for visualization purpose and modeling purpose respectively. First version that was created for visualizations is dated from August 16, 2016 to August 15, 2017. The second version that was created for building the predictive model contains data only after January 1, 2017. The number of the sales data entries decreased to 45,497,040 and 23,808,261 respectively from 125,497,040.

Data Exploration & Visualizations

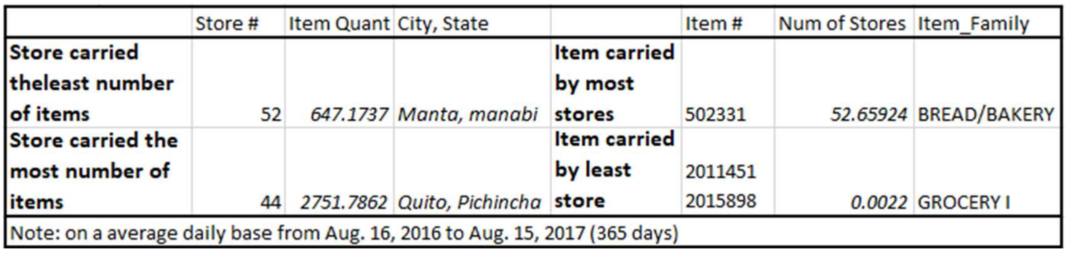
Figure 1 shows the number of the original observations of each data file we were given. The original number of observations for training data is more than 125 million. That information was collected for 54 stores and 4100 items throughout a six-and-half year time frame across 22 cities in Ecuador.



```
nrow(train) ## 125497040 (original)
nrow(test)  ## 3370464
nrow(holiday)  ## 350
nrow(item)  ## 4100
nrow(oil)   ## 1218
nrow(store) ## 54
nrow(transaction)   ## 83488
```

Figure 1. The Number of Observations for Each File

We used a frequency analysis to investigate the effect of individual stores and products. Figure 2 lists the country's biggest and smallest stores based on the amount of items they sell on a regular basis. It also indicates which products were the most and least popular, depending on the amount of stores that sold them.



	Store #	Item Quant	City, State		Item #	Num of Stores	Item_Family
Store carried theleast number of items	52	*647.1737*	*Manta, manabi*	**Item carried by most stores**	502331	*52.65924*	BREAD/BAKERY
Store carried the most number of items	44	*2751.7862*	*Quito, Pichincha*	**Item carried by least store**	2011451 2015898	*0.0022*	GROCERY I
Note: on a average daily base from Aug. 16, 2016 to Aug. 15, 2017 (365 days)							

Figure 2. The Frequency Analysis for Stores and Items

From Figure 2, we learned that while the store #44 which is located in Quito was the store that carried the largest number of items, 2751.78 items on an average daily basis, in the country; store #52 located in Manta carried the least amount of items, 647.17, on an average daily basis. Item #502331 was the most popular item, which belongs to the BREAD/BAKERY item family, and it was carried by about 52 out of 54 stores in Ecuador. On the other hand, item #2011451 and #2015898, both belonging to the GROCERY I item family, were carried by the least number of stores (less than 0.0022 stores on an average daily basis), tied for the least popular items across the country.

Corporacion Favorita had 54 stores located in 22 cities. By looking at the data on a city basis, Figure 3 under "Appendix A" shows that city Manta, Quito and Guayaquil, among others, most affected the sales. The percentage of records were 2.7%, 40.1%, and 13.3% respectively.

By looking at the data on a monthly basis, we found an interesting fact that while store #52 was the smallest store in the country, it had the highest unit sales increase (0.42 million) from April to May. This information may be found in Figure 4 under "Appendix A."

Figure 5 under "Appendix A" is an interactive dashboard we created to study the unit sales. You may find the information regarding the unit sales on a Quarter basis, Month basis, weather on promotion, location basis, and the top stores, etc.

Since "Ecuador is an oil-dependent country and it's economical health is highly vulnerable to shocks in oil prices," Corporacion Favorita believed that incorporating the oil data may help better understand the unit sales. However, as Figure 6 illustrates, there is no obvious relationship between oil price and unit sales in the year of 2017.

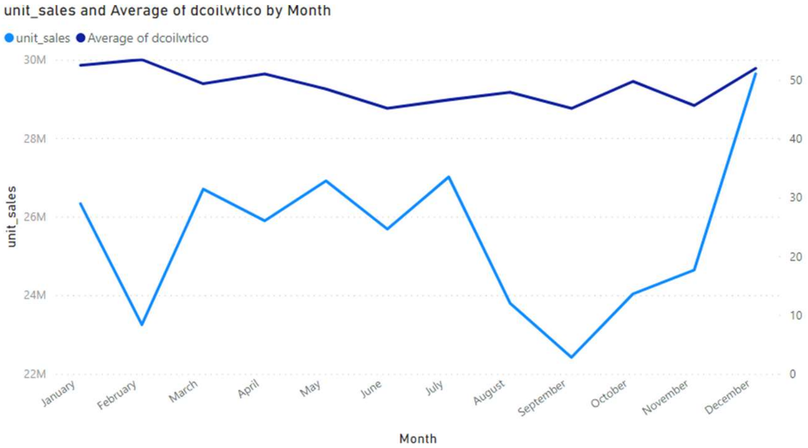
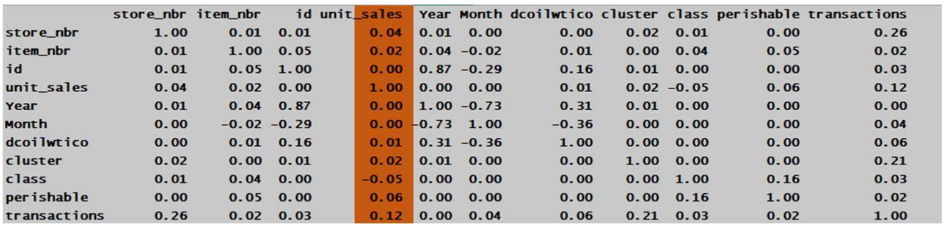


Figure 6. Oil Price vs. Unit Sales

Feature Engineering and Modeling

To optimize the modeling process, we did some necessary feature engineering, and created the following columns: "Year" and "Month" from the "date" column. And then, we did a left join on the training and testing data from all the supplementary files except "holiday.csv". To convert the text data into numeric format, we generated dummy variables for both "item_family" column and "city" column. The final dataset contained 65 features.

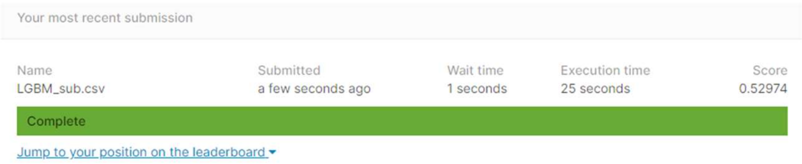


	store_nbr	item_nbr	id	unit_sales	Year	Month	dcoilwtico	cluster	class	perishable	transactions
store_nbr	1.00	0.01	0.01	0.04	0.01	0.00	0.00	0.02	0.01	0.00	0.26
item_nbr	0.01	1.00	0.05	0.02	0.04	-0.02	0.01	0.00	0.04	0.05	0.02
id	0.01	0.05	1.00	0.00	0.87	-0.29	0.16	0.01	0.00	0.00	0.03
unit_sales	0.04	0.02	0.00	1.00	0.00	0.00	0.01	0.02	-0.05	0.06	0.12
Year	0.01	0.04	0.87	0.00	1.00	-0.73	0.31	0.01	0.00	0.00	0.00
Month	0.00	-0.02	-0.29	0.00	-0.73	1.00	-0.36	0.00	0.00	0.00	0.04
dcoilwtico	0.00	0.01	0.16	0.01	0.31	-0.36	1.00	0.00	0.00	0.00	0.06
cluster	0.02	0.00	0.01	0.02	0.01	0.00	0.00	1.00	0.00	0.00	0.21
class	0.01	0.04	0.00	-0.05	0.00	0.00	0.00	0.00	1.00	0.16	0.03
perishable	0.00	0.05	0.00	0.06	0.00	0.00	0.00	0.00	0.16	1.00	0.02
transactions	0.26	0.02	0.03	0.12	0.00	0.04	0.06	0.21	0.03	0.02	1.00

Figure 7. Correlation Matrix

According to the correlation analysis in Figure 7, we can tell that there is no clear linear relationship between the unit_sales and other variables. So we moved to a more advanced regression model: LightGBM regressor.

The NWRMSLE of our final submission was 0.529, which stands for Normalized Weighted Root Mean Squared Logarithmic Error, ranked as 710th place out of 1671 teams by the time we submitted. Figure 8 shows the result.



Your most recent submission

Name	Submitted	Wait time	Execution time	Score
LGBM_sub.csv	a few seconds ago	1 seconds	25 seconds	0.52974

Complete

Jump to your position on the leaderboard ▾

Figure 8. Kaggle Submission

Figure 9 under “Appendix A” shows the top 10 most important features generated by the model. They are transaction volume, item number, item class (such as Meats, Poultry, Beverage, etc.), promotion events, store clusters, and store numbers.

Conclusion

In this project, we found the top 3 cities that had the most impact on the unit sales were Manta, Quito, and Guayaquil. Store #44 and #52 are the largest and smallest stores in the country based on the amount of items they carried. However, the smallest store had the most unit sales increase from April to May, so the store size cannot solo determine the sales. We also learned that even though Ecuador's economical health is heavily influenced by oil prices, we could not find evidence to back up that the company’s sales have a relationship with the oil price yet.

In the modeling process, we generated the most important features. Most of them are directly related to items themselves. By looking back to the item class in detail, the company may find more useful insights to help them make better decisions.

We could not find a way to incorporate the holiday information into the analysis as the company expected. But theoretically, we do believe that adding the holiday information to the modeling will certainly help optimize the result.

APPENDIX A – SUPPLEMENTARY MATERIALS

You may check the partial codes and relevant tables on my GitHub

or

https://github.com/jinhangjiang/CIS591_SUPPLEMENTARY_MATERIALS

1. R code for the analysis
2. Figure 3: Sales on City Basis
3. Figure 4: Sales on Monthly Basis
4. Figure 5: Interactive Dashboard for Unit Sales
5. Figure 9: Top 10 Most Important Features

APPENDIX B – REFERENCE

Full Data Description: https://www.kaggle.com/c/favorita-grocery-sales-forecasting/data

LightGBM R-Package: https://lightgbm.readthedocs.io/en/latest/R/index.html

Andrewmvd. “LightGBM in R.” Kaggle, Larxel, 26 Nov. 2017,

www.kaggle.com/andrewmvd/lightgbm-in-r