Store Sales - Time Series Forecasting using Machine Learning

Dhruv Prajapati* (AU1940192), Dhaval Chaudhary* (AU1940180),
Aaryan Mori* (AU1940194), Harshil Doshi* (AU1940279)
School of Engineering and Applied Science, Ahmedabad University
***All Authors have contributed equally**

***Abstract*—Every grocery store keeps the stock of a certain product according to its demand. These stores have thousands of product families as per customer requirements. As they do not have any accurate predictions of sales, many times their can be shortage of products or the theirs is a bunch of products which remained unsold. This affects loss for the store and a lack of customer satisfaction as well. This project will make use of the store sales data published on Kaggle.The data comes from an Ecuador company as known as Corporación Favorita and it is a large grocery retailer.Our main mission in this competition is, predicting sales for each product family and store combinations.**

***Index Terms*—Machine Learning, Time Series, Store Sales, Interpolation of Oil Prices, Exploratory Data Analysis, Feature Engineering.**

I. Introduction

Good forecasts are vital in many areas of scientic, industrial, commercial and economic activity. In time-series forecasting, forecasts are made on the basis of data comprising one or more time series. A time-series is a collection of observations made sequentially through time. We have been provided with time series of gross sales of 54 Favorita grocery stores over four years. The factor that motivated us to choose this project is that two of our group members' father have grocery stores. They have always wanted to predict the sales so that they can save their father from overstocking and waste of food items. After looking into the problem statement, it was the best fit project which our team could have.

The goal of the project is to predict the gross sales for 54 Corporacion Favorita stores stores located in Ecuador. There are 5 data which have been that provided:

1) **The Train Data** contains time series of the stores and the product families combination. The sales column gives the total sales for a product family at a particular store at a given date. There are total 33 unique values which refers to product families. Fractional values are possible since products can be sold in fractional units (1.5 kg of cheese, for instance, as opposed to 1 bag of chips). The "onpromotion" column gives the total number of items in a product family that were being promoted at a store at a given date. As a result, there are 10,48,576 rows to anticipate.
2) **Stores Data** gives some information about stores such as city, state, type, cluster.
3) **Transaction Data** is highly correlated with train's sales column. You can understand the sales patterns of the stores.
4) **Holidays and Events Data** is a meta data. This data is quite valuable to understand past sales, trend and seasonality components. However, it needs to be arranged. You are going to find a comprehensive data manipulation for this data. That part will be one of the most important chapter in this notebook.
5) **Daily Oil Price Data** is another data which will help us. Ecuador is an oil-dependent country and it's economical health is highly vulnerable to shocks in oil prices. That's why, it will help us to understand which product families affected in positive or negative way by oil price.

II. Literature Review

Sales prediction is an important part of modern business intelligence. It can be a complex problem, especially in the case of lack of data, missing data, and the presence of outliers. Sales can be considered as a time series. One important element of sales forecasting is the accuracy of the prediction. Therefore, a lot of efforts have been made to make this process more accurate. In the publications of this field, the accuracy is obtained using some error measurement methods like RMSE, MAPE, after comparing the actual sales with the predicted results.

Classical time-series forecasting methods are not a single best technique to solve time-series forecasting problems (Zhang Kline, 2007). In order to deal with time-series forecasting, each problem might be solved with a different approach. Moving Average (MA) is one of the simplest prediction techniques for making projections about time-series without a noticeable seasonal pattern (Chopra Meindl, 2016). In several papers, a more advanced version of MA which is called Autoregressive Integrated Moving Average (ARIMA) has been used. For instance, Ramos, Santos, and Rebelo (2015) used ARIMA and exponential smoothing method to compare the performances of these methods on forecasting the retail sales of women's footwear, which contain products with repeatable fluctuations in their patterns. The demand for

purchasing boots in winter is an example of these fluctuations. Huber, Gossmann, and Stuckenschmidt (2017) applied multivariate ARIMA successfully to perform demand forecasting on perishable goods. Recently, Yang et al. (2021) applied combined ARIMA and neural network for network traffic forecasting.

There are different types of exponential smoothing methods. Kalekar (2004) explained the difference between these methods and used the triple exponential smoothing method or Holt-Winters (HW) technique to deal with the seasonality in dataset. Exponential smoothing has been a powerful forecasting tool for prediction.

III. Implementation

A. Exploratory Data Analysis

The data we have been working has been provided to us from an Ecuador company as known as Corporación Favorita and it is a large grocery retailer. Also, the company operates in other countries in South America. There are 54 stores and 33 prodcut families in the data. The time serie starts from 2013-01-01 and finishes in 2017-08-31. However, Kaggle gives splitted two data as train and test. The dates in the test data are for the 15 days after the last date in the training data. Date range in the test data will be very important to us while we are defining a cross-validation strategy and creating new features.

- The training data, comprising time series of features store_nbr, family, and onpromotion as well as the target sales.

  store_nbr identifies the store at which the products are sold.

  family identifies the type of product sold.

  sales gives the total sales for a product family at a particular store at a given date. Fractional values are possible since products can be sold in fractional units (1.5 kg of cheese, for instance, as opposed to 1 bag of chips).

  onpromotion gives the total number of items in a product family that were being promoted at a store at a given date.

- Stores data has **Store metadata**, including **city**, **state**, **type**, and **cluster**. cluster is a grouping of similar stores.

- Oil Price data includes **Daily oil prices**. Includes values during both the train and test data timeframes.

- Holidays data includes**Holidays and Events**, with metadata.

  Transferred column: A holiday that is transferred officially falls on that calendar day, but was moved to another date by the government. A transferred day is more like a normal day than a holiday. To find the day that it was actually celebrated, look for the corresponding row where type is Transfer. These are frequently made up by the type Work Day which is a day not normally scheduled for work (e.g., Saturday) that is meant to payback the Bridge.

  Additional holidays are days added a regular calendar holiday, for example, as typically happens around Christmas (making Christmas Eve a holiday).

Additional Notes:

- Wages in the public sector are paid every two weeks on the 15th and on the last day of the month. Supermarket sales could be affected by this.
- A magnitude 7.8 earthquake struck Ecuador on April 16, 2016. People rallied in relief efforts donating water and other first need products which greatly affected supermarket sales for several weeks after the earthquake.

B. Light GBM

LightGBM is a decision tree-based gradient boosting framework that improves model efficiency while reducing memory utilization. It employs two innovative techniques: Gradientbased One Side Sampling and Exclusive Feature Bundling (EFB), which address the drawbacks of the histogram-based approach used in most GBDT (Gradient Boosting Decision Tree) frameworks. The properties of the LightGBM Algorithm are formed by the two methodologies of GOSS and EFB explained below. They work together to make the model run smoothly and give it an advantage over competing for GBDT frameworks.

LightGBM does not grow a tree row by row, unlike most other implementations. Instead, it grows trees leaf-by-leaf. It selects the leaf that it believes will result in the most significant loss reduction. Furthermore, unlike XGBoost and other implementations, LightGBM does not use the sortedbased decision tree learning algorithm, which searches for the optimum split point on sorted feature values. Instead, LightGBM uses a highly optimized histogram-based decision tree learning algorithm, which provides significant performance and memory savings. Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) are two unique techniques used in the LightGBM algorithm to run faster while maintaining excellent accuracy.

C. Ridge + Random Forest

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated.(Hilt, Donald E.) It has been used in many fields including econometrics, chemistry, and engineering.Ridge regression was developed as a possible solution to the imprecision of least square estimators when linear regression models have some multicollinear (highly correlated) independent variables—by creating a ridge regression estimator (RR). This provides a more precise ridge parameters estimate, as its variance and mean square estimator are

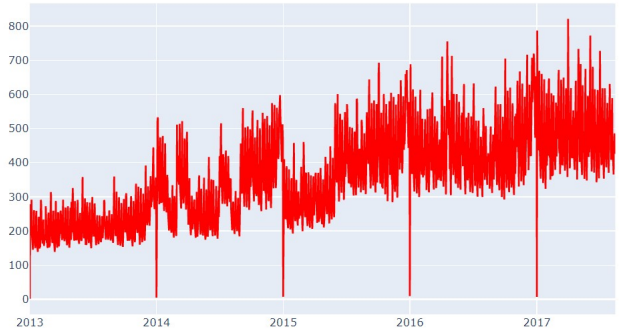
often smaller than the least square estimators previously derived.

Random forests ca be defined as an grouped learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of over-fitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

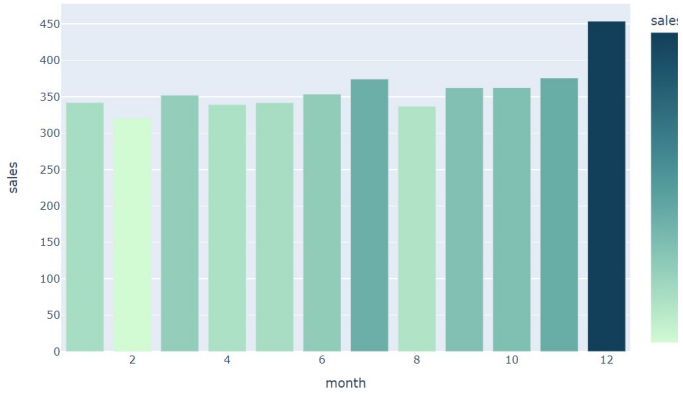
IV. Results

A. *Exploratory Data Analysis*

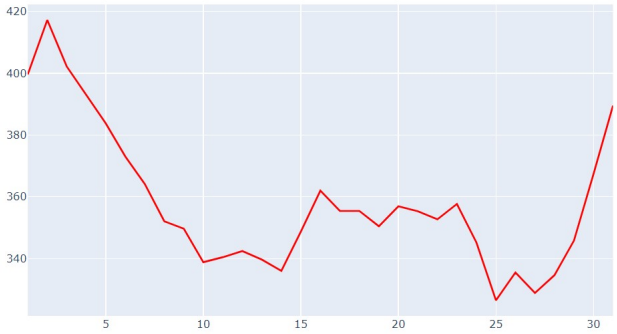
We explored the data and the additional factors which affects the data. Our primary step was to understand the data, as there were millions of rows and so many parameters spread across the files. To tackle the problem, it was a necessary job of getting a clear understanding of the data. We implemented exploratory Data Analysis. We began with plotting graphs with respect to various time feature to get an understanding of overall sales trend. We got the following graphs of the sales trend shown below:



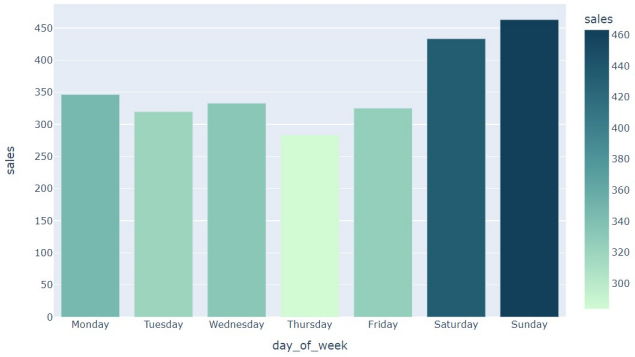
The above graph shows the overall sales trend over the years from 2013 to 2017. We can derive that the sales is growing till 2014 but after a deep low in 2015, the sales has been constant.



The above graph shows the overall sales trend over the months. We can derive that the sales is rising at the end of the year. Which can be because of large number of holidays in December.



The above graph shows the overall sales trend over the days of months. We can derive that the sales is rising at middle of the month and end of the month. Which can be because the public wages are paid during the same period.large number of holidays in December.



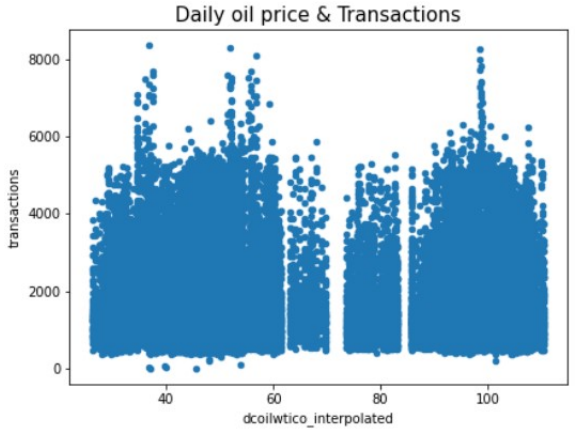
The above graph shows the overall sales trend over the days during a week. We can derive that the sales is rising at the weekend. Which can be because of more transaction due to holidays.

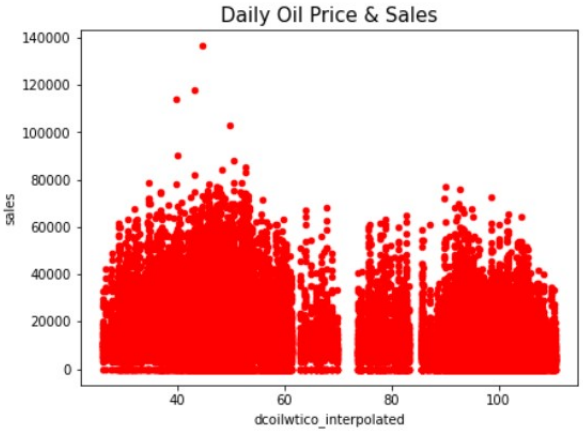
We have also been provided with additional data of holidays and events. After analysing all the days, we plotted the gross sales with respect to holidays, work days, events, transfer etc. The graph for the sale is plotted below:

The economy is one of the biggest problem for the governments and people. It affects all of things in a good

or bad way. In our case, Ecuador is an oil-dependent country. Changing oil prices in Ecuador will cause a variance in the model. We can see the trend and predict missing data points, when we look at a time series plot of oil price. Below graph shows the oil prices trends over the years.

There are some missing data points in the daily oil data. We can treat the data by using various imputation methods. However, we chose a simple solution for that. Linear Interpolation is suitable for this time series. First of all, let's look at the correlations for sales and transactions.





The correlation values are not strong but the sign of sales is negative. Maybe, we can catch a clue. Logically, if daily oil price is high, we expect that the Ecuador's economy is bad and it means the price of product increases and sales decreases. There is a negative relationship here.

After completing the Exploratory Data Analysis, we implemented feature engineering. Feature engineering refers to select and transform the most relevant variables while creating a predictive model. We have transformed 3 features in our data with respect to holidays, date and oil price trend:

- Only keeping national holiday for simplicity
- Setting date as index
- Dropping duplicated holiday
- last 7 day avg oil price
- leave the holiday with higher avg sales, which means more important holiday

Then we merged our entire data and divided it into three parts for modelling:

- Train Data: 01/04/2017 to 31/07/2017
- Test Data: 16/8/2017 to 31/8/2017
- Validation Data: 1/8/2017 to 15/8/2017

B. Modelling

We ran a total of two models, one is Light GBM and the other is Ridge Regression as well as Random Forest. While it is pretty evident Ridge Regression performed better than Light GBM. Light GBM failed to work even after we specifically tuned the parameters for its model. It will be an interesting thing to look on.

We have implemented the Fourier features to preserve the correlation between neighboring data points. However, the temporal structures are lost in the frequency domain. This makes it possible that different signals produce very similar magnitude representations under Fourier mappings.

We have used Optuna for tuning hyper-parameters. To control the behaviour of our machine learning model we have tuned the hyper-parameter. If we don't correctly tune our hyper-parameters, our estimated model parameters produce sub-optimal results, as they don't minimize the loss function. This means our model makes more errors.

After tuning hyper-parameters, we implemented our both models. We applied Root Mean Squared Log Error to minimize the loss and maximize the accuracy.

Root Mean Squared Log Error:

$$RMSLE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2}$$

After implementing both models and applying RMSLE to both we got the following results:

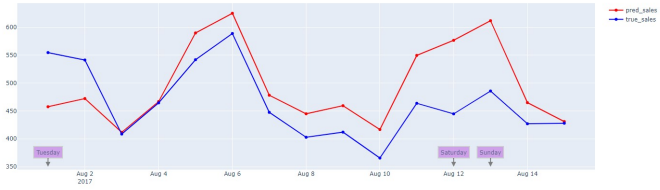
Model	RMSLE
Light GBM	0.3232
Ridge + Random Forest	0.2395

We can observe that $\text{RMSLE}_{LightGBM} = 0.3232$ is much greater than $RMSLE_{Ridge} = 0.2395$.

The important feature is that School and Office supply sales performs exceptional both the models. It's RMSLE in LightGBM is 4.04. We were able to decrease its RMSLE to 1.77 by applying Random Forest model. Therefore, we can derive that Ridge+RF model has better accuracy than Light GBM model.

C. *Error Analysis*

Error analysis is a process to isolate, observe and repair erroneous Machine Learning predictions therefore helping understand highs and lows in performance of the model. We have annotated the days on which our model has predicted bad. The Final scatter plot is shown below:



V. Conclusion

Time-series forecasting in simply refers to forecast or to predict the future value over a period of time. Initially, we didn't have any knowledge of our data and the outliers present in our data. After implementing, Exploratory Data Analysis, we had a brief information available for our data. After finding the missing data, we added some of the features for better predictive modelling. Since, Ecuador's economy is oil-dependant. We have find a correlation of oil prices with transactions and sales. The correlation values are not strong but the sign of sales is negative. Which derives that There is a negative relationship here. Logically, if daily oil price is high, we expect that the Ecuador's economy is bad and it means the price of product increases and sales decreases. We have implemented two models. We obtained the lowest RMSLE value of 0.2395 through our Ridge regression model which significantly better than Light GBM.

References

1) Hilt, Donald E.; Seegrist, Donald W. (1977). Ridge, a computer program for calculating ridge regression estimates. doi:10.5962/bhl.title.68934

2) Papacharalampous G, Tyralis H., Koutsoyiannis D, ”Univariate time series forecasting of temperature and precipitation with a focus on machine learning algorithms: A multiple-case study from Greece” Water Resour. Manag. 2018, 32, 5207–5239. [Google Scholar]

3) B. Pavlyshenko, “Machine-Learning Models for Sales Time Series Forecasting,” Data, vol. 4, no. 1, p. 15, Jan. 2019, doi: 10.3390/data4010015.

4) S. Chopra, P. Meindl Supply chain management: Strategy, planning, and operation Community.tableau.com (2016) [online] Available at https://community.tableau.com/docs/DOC-1236

5) J. Huber, A. Gossmann, H. Stuckenschmidt Cluster-based hierarchical demand forecasting for perishable goods Expert Systems with Applications, 76 (2017), pp. 140-151

6) H. Yang, X. Li, W. Qiang, Y. Zhao, W. Zhang, C. Tang A network traffic forecasting method based on SA optimized ARIMA–BP neural network Computer Networks, 193 (2021), Article 108102

7) Kaggle.com. 2022. Store Sales - Time Series Forecasting Use machine learning to predict grocery sales. [online] Available at: https://www.kaggle.com/competitions/store-sales-time-series-forecasting/overview.

8) P.S. Kalekar Time series forecasting using holt-winters exponential smoothing Kanwal Rekhi School of Information Technology, 4329008 (13) (2004)

9) G.P. Zhang, D.M. Kline Quarterly time-series forecasting with neural networks IEEE Transactions on Neural Networks, 18 (6) (2007), pp. 1800-1814