Jackie Gan

CIS 5526

Dr. Vucetic

Predicting Stores Sales in Ecaudor

**Introduction:**

Over the years, retail stores are having a hard time justifying which items to restock, how much to restock, when to restock, and where. If the store overstocks, they will be stuck with overstocked and perished goods, implies lost profit and revenue. If the store understocks, items sold out too soon would upset customers and worsen future growth. In this project, we will explore a Kaggle competition, Store Sales – Time Series Forecasting. Our goal in this project is to predict the last 2 weeks sales from mid-August 2017 to end of August 2017.

**Approach:**

Data Analysis:

In this project, it had contained training.csv (training set), transaction.csv (log of daily transaction across all stores), stores (information about each store), holiday\_events.csv (Holidays in the country), oil.csv (daily oil prices), and test.csv (testing set) on the supermarkets Favorita stores located in Ecuador. Some important notes from the dataset, Ecaudaor is a very oil dependent oil, the oil price from the oil dataset can help forecast the health of the economy. Next, employees in the public sector are paid every 2 weeks, once on the 15th and the other at the end of the month. Another important note, Ecuador was struck with a 7.8 earthquake on April 16, 2016. I expect sales will drop within the month of April. All of the data was read through pandas data frame and plotted using seaborn and matplotlib.

What does each dataset mean and why are they important?

* The training dataset consist of a couple columns, date, store\_nbr, family, sales, and onpromotion. I used this dataset to examine what type of product families were selling well and what were the trend. What does sale look like from year to year? What sales look like within a month, workdays, weekends, which store and what cities had most of the sales? What is the relation between sales and promotions. Through our exploratory Data Analysis, we found sales has been increasing from 2013 to 2017. We found almost all family categories had some sort of turning point in mid early 2016, this is highly likely due to the earthquake in 2016. However, almost product family had recovered and stabilized in sales except a few families of product such as liquor, frozen foods, and wine. With further examining, we found the top 5 family of product sold are grocery, beverages, produce, cleaning, and diary (Figure 7 and 8). At the same time, I found the same product that are being promoted also had a higher in sales and they were also the same top 5 families of product.
* In our oil dataset, we found that the oil prices have dropped significantly from early 2013 at 110 dollars a barrel to as low as 26 dollars a barrel in early 2016 and it stabilized around 40 – 50 dollars around 2016 to late 2017 (Figure 4). As oil prices getting cheaper, we also saw an increase of sales within those year (figure 5). The correlation between oil prices and sales is around 0.5, which is relatively positive and strong (figure 6). Which implies oil is a significant feature that will determine out future sales.
* Public sector pay is also another factor in our data, as we examine this feature. We found that sales on each day is very different. People are spending most of their money during the end of the week especially Saturday and Sunday (figure 9). Then when we analyze the sales average by the day of the month, we found day 15, day 16, day 17 has an increase of sales. Followed by the end of the month day 30th, 31st, 1st and 2nd we still see a high sale. This shows us, that people are more willing to spend during weekends and near their paycheck.
* Holiday is big in Ecuador, especially during Christmas. We explored the data in sales by month, we found December yield the highest sales compared to the rest of the other months (figure 10). This might be due to the reason of Christmas and New Year’s. In this holiday, dataset we also found that additional day, and bridge day yield very high in sales. Since bridge days and additional days are specifically plays in the favor of extending a holiday to the weekend and adding additional days to holidays. However, we should not consider transfer days, because in reality it is just another workday.
* Store dataset is also a big factor. We found that store 44 yield the highest store sales compared to all other stores. Store 44 is also in a very specific type of store, which is type A which grossed more sales than all other types. We found a city; Quito has the most stores and the most sales. In the same, store 44 is in fact also in Quito, type A and have the most sales.

Data Cleaning, Data Preparation and Machine Learning Approach:

In modern dataset, a lot of datasets will have missing data due to mishandling or failure to record. Luckily, this only happened in our oil dataset. First, in our oil dataset, we found a lot of missing data that are ‘NAN’. To ensure my model runs smoothly, I decided to fill in the gaps of the data with linear interpolation. Linear interpolation will estimate the missing points. However, this might not be the best approach because linear interpolation is not always accurate. But this might not be a problem since more than 95% of the data has already existed. Next, I realized the first row, first day of 2013 is missing, I had to use row 2 value to fill in the gap of row 1.

In the training dataset, my approach was to convert all the data into the dates. To be able to analyze stores sales, the dates are essentials and specially to transform into smaller chucks such as year, month, day, and day of the week. Next, I made a date\_str column to merge the data for oil and holidays, since the days as already been converted to date objects in the beginning. Using the date\_str will help merge the data without conflicting date objects from the beginning. Now, we merged the store dataset by the store\_nbr (store number), since date is not a column in the dataset. We don’t need to convert anything since we never converted store\_nbr to a different object.

To better understand the data, I have used matlibplot to generate histograms, scatter plots, and bar charts for each numerical feature. We used the groupby() and mean() function to group all the data together by dates to analyze the each specific part of the data. Using these methods, I was able to extract and analyze the daily, monthly, and yearly average. I was also using the corr() function to analyze oil price, sales, and transaction dataset.

The categorical features were mapped to all different types of values. In the beginning of the training dataset, we converted all the data to year, month, and day. Time series is crucial to be part of our feature since it will be used to forecast the future of our sales. We were also able to use them as part of our features that is being added to our training set. Later, we used one-hot encoding to transform whether it’s a workday, weekend, regional holiday, local holiday, and national holiday. We also eliminated transfer days, since Kaggle told us it is not celebrated as a holiday. However, there are still data that are words such as city, state, family of products, making them one hot encoding would not be ideal since it will make training a lot longer with a lot of unique attributes. We were able to use label encoder from sklearn that will label them as numerical values.

Linear Regression is widely used to solve forecasting problems. The Linear regression algorithm will learn how to make a make a weighted sum from its features. This also means, we need to incorporate lag features. Lag feature are features that is derived from previous values of a time series. This will help predict and learn about future from temporal trend and dependencies. When I was determining lags, I was looking through the data from what days correlated the best sales. There were also resulted when employee getting, so I added lag for the end of the month and middle of the month. Namely, 16, 17, 18, 28 and 31, due to month structure. This is not the best way to add lags. There are better ways to add them such as suing autocorrelations using diff functions. Lag features also works in decision trees, and XGboost.

**Result:**

We first split the data into 80% training and 20% testing using the sklearn train\_test\_split(). Then I trained 3 different models, linear regression, decision tree, and xgboost. We will be using MSLE (mean square log error) and R^2 (R squared) as metrics for these models. In the first model, using linear regression with lag and time series features we were able to use previous data to build a trend for predict futuristic data. After using cross validation to check the data with 10 folds. We found the model performed relatively well with a 0.04 MSLE and 0.96 R2 (figure 12).

In our decision tree model, we found it performance worse than linear regression and XGBoost. Part of the reason is that the cross-validation score higher than the MSE score, this implies the model is overfitting. Even though decision tree can see nonlinear relationships, in this dataset we have too many features to deal with. This would result the model to perform worse and trying too hard in the model. However, the performance is not the worst, we got a 0.07 MSLE (figure 14).

In our last model we used, we used XGboost. XGboost is a gradient boosted tree that would have advantages in dealing with large feature set, this dataset had 28 columns of features. After cross validation, the model is shown that it is also overfitting when the score is higher than testing set. Next, I used an early stopping technique prevent overfitting in our model. This technique did improve our model but not enough to consider it is overfitting. We have resulted a R2 score of 0.97 and 0.0327 (figure 13). XGBoost also have a very interesting feature plot that we can generate. We were able to get F score about the importance of each feature (figure 15).

**Conclusion:**

In this project, I first generated a report on exploratory analysis to find the important features in our dataset. After examining through the analysis. I was able to extract the features and data clean of the data. Next, I built three models different regression models to predict the sales of Ecaudor store. Our model preformed relatively well compared to Kaggle competitions. However, there are some things I would do different and want to explore more next time. I would want to explore more on lag features by building autocorrections plots to find which lag features are more significant with plotting results. I would also want to try exploring into the hyperparameters of each model and tune them better. Especially in xgboost and decision, this might help us eliminate some of the overfitting problems that we had in our results.

PS: I have spent around 28 hours on this project

**Resources:**

https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data <https://www.kaggle.com/competitions/store-sales-time-series-forecasting/overview>

<https://www.kaggle.com/code/berkayalan/linear-regression-with-time-series#Introduction>

<https://www.kaggle.com/code/carlolepelaars/understanding-the-metric-rmsle/notebook>

<https://www.kaggle.com/code/ryanholbrook/time-series-as-features>

<https://www.kaggle.com/code/prashant111/xgboost-k-fold-cv-feature-importance>