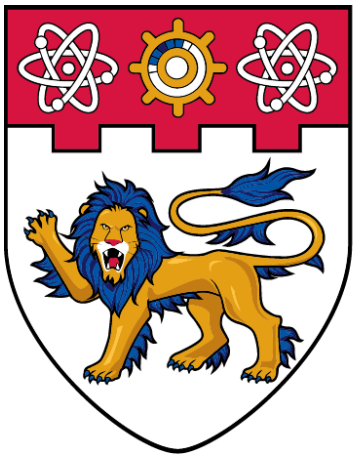


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

CZ4041 - MACHINE LEARNING

Project Report
Kaggle Challenge: Store Item Demand Forecasting Challenge

GROUP 11

TEKWANI PUNEET (U1722127D)
NIKHITA MENON (U1722287L)
FAZILI NUMAIR (U1822056F)
ARORA MANAV (U1822077D)
RAJURAVI VISHAL RAJ (U1822268B)

Table of Contents

1. Problem Statement

Background Information

The Kaggle competition chosen for the project allows us to explore different time series techniques on store-item sales data. The competition provides us with store-items sales data for the past 5 years.

Problem Statement

Given the 5 years of store-item sales data, the Kaggle competition requires us to forecast the sales for the year 2018 for 50 different items at 10 different stores.

Sales forecasting is a crucial process adopted by several companies today, to help them predict their upcoming sales and make informed business decisions accordingly.

Evaluation

Evaluation of our model will be based on the Symmetric Mean Absolute Percentage Error (SMAPE or sMAPE) accuracy metric, which is based on the percentage errors or relative errors. SMAPE can be given by :

$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

where A $\square$ is the actual value and F $\square$ is the forecast value.

SMAPE = 0 when the actual and predicted values are both 0.

2. Understanding the Dataset

We have been provided with 3 files of datasets; namely, *train.csv, text.csv* and *sample_submission.csv*. The structure size of the fields and records of each of these three datasets have been detailed below.

Training Dataset

This is the main training dataset we have been provided with, which contains 5 years of store-item sales data, at 10 different stores.

Data Fields:

1. date: The transaction date of the specified item at the specified store, ranging from 1/1/2013 to 31/12/2017.
2. store: A unique numeral store identifier (1-10).
3. item: A unique numeral item identifier (1-50).
4. sales: The turnover of the specified item at the specified store on the specified date.

Number of Records: 913,000

Test Dataset

This is the test dataset we have been provided with, which will be used for testing and evaluating the model.

Data Fields:

1. id: A unique numeric identifier for a specific date, store and item.
2. date: The transaction data of the specified item at the specified store, ranging from 1/1/2018 to 31/03/2018.
3. store: A unique numeral store identifier (1-10).
4. item: A unique numeral item identifier (1-50).

Number of Records: 45,000

Sample Submission

This is a sample submission file that provides the format of the prediction for the competition.

Data Fields:

1. id: A unique numeric identifier for a specific date, store and item.
2. sales: The predicted turnover of the specified item at the specified store on the specified date.

3. Data Preprocessing

Data Wrangling

The data we have been provided with for this project does not contain any null, duplicate or undefined values.

Feature Engineering

Since our project is based on a time-series problem, we decomposed the ‘date’ data field into the following components. These derived features will allow us to effectively capture the seasonality patterns.

1. Day
2. Day of the week
3. Day of the year
4. Week of the year
5. Month
6. Year

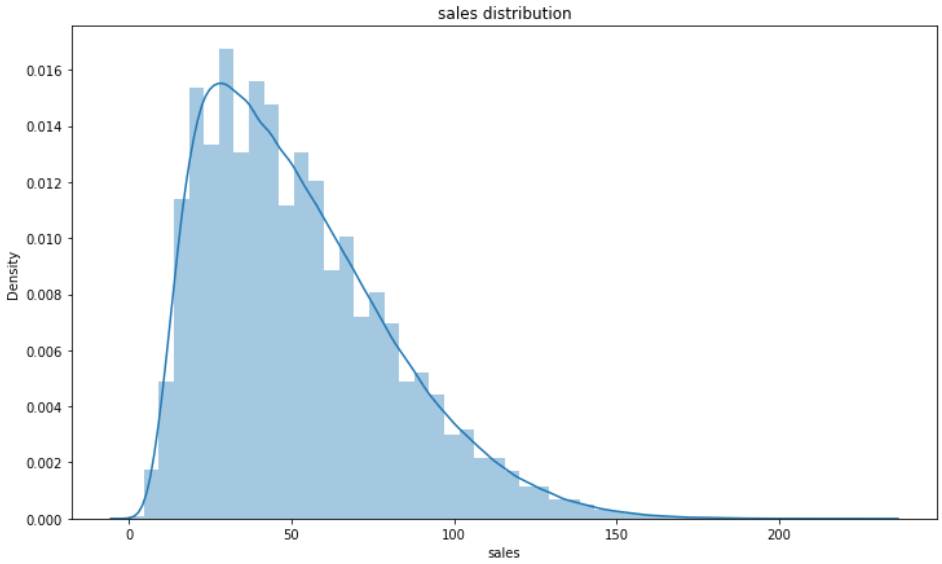
Field Name	Min	Max	Mean	Standard Deviation
date	2013-01-01	2017-31-12	-	-
store	1	10	-	-
item	1	50	-	-
sales	0	231	52.3	28.8

Since the ‘Store’ and ‘Item’ attributes have discrete values in range [1,10] and [1,50], we classify these as categorical data types.

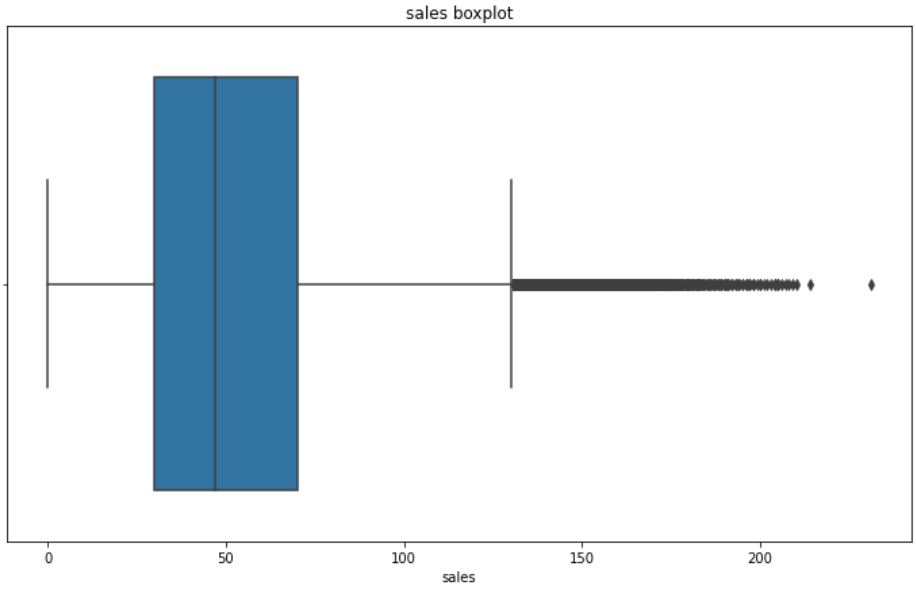
4. Exploratory Data Analysis and Outlier Detection

Sales Distribution

The histogram of sales indicates a right skewed normal distribution with mean, median and mode of 52.25, 47 and 30 respectively.

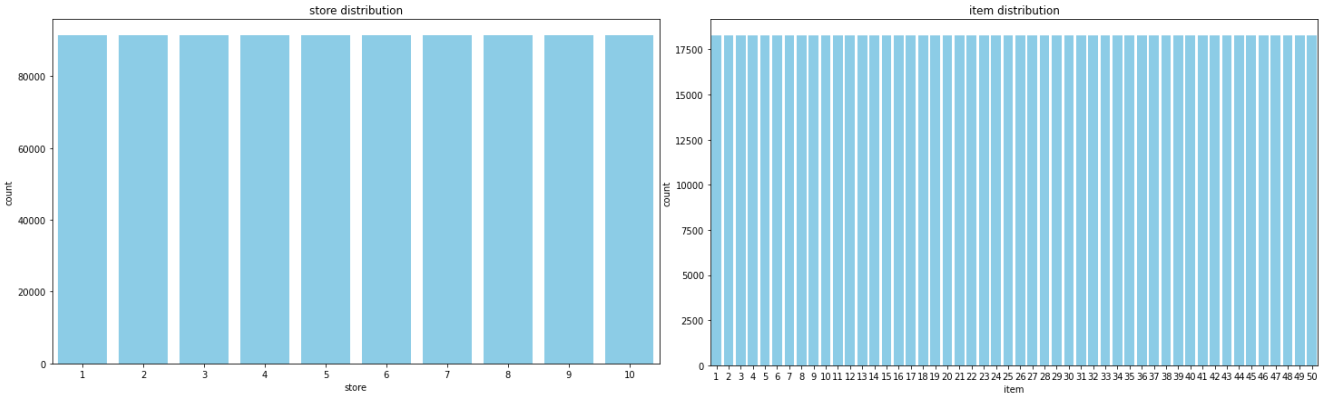


This indicates that overall sales are dominated by extremely high values and this is confirmed from the box plot below.



While values beyond the max whisker (Q3 + 1.5 IQR) are acceptable for a right skewed normal distribution, the values greater than 200 seem extremely high and on further analysis, we see that these values primarily correspond to store 2 sales and the total percentage of these values is only 0.001% of the total sales of store 2. Therefore, we can classify these values as outliers and eliminate them from the dataset.

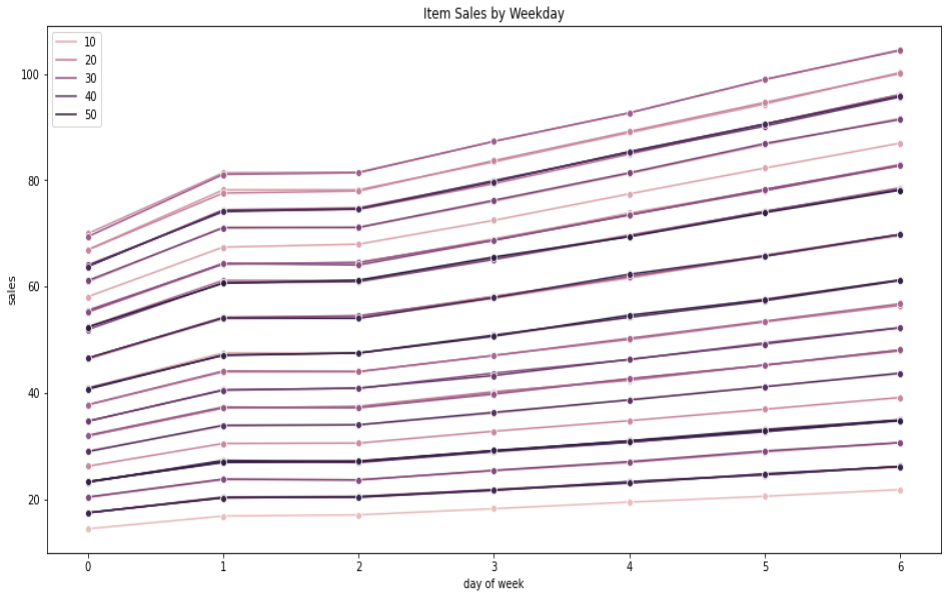
Store and Item Distribution



The distribution of both store and item attributes follows a uniform distribution. This demonstrates that all sales are equally distributed across different stores and items and there is no class imbalance.

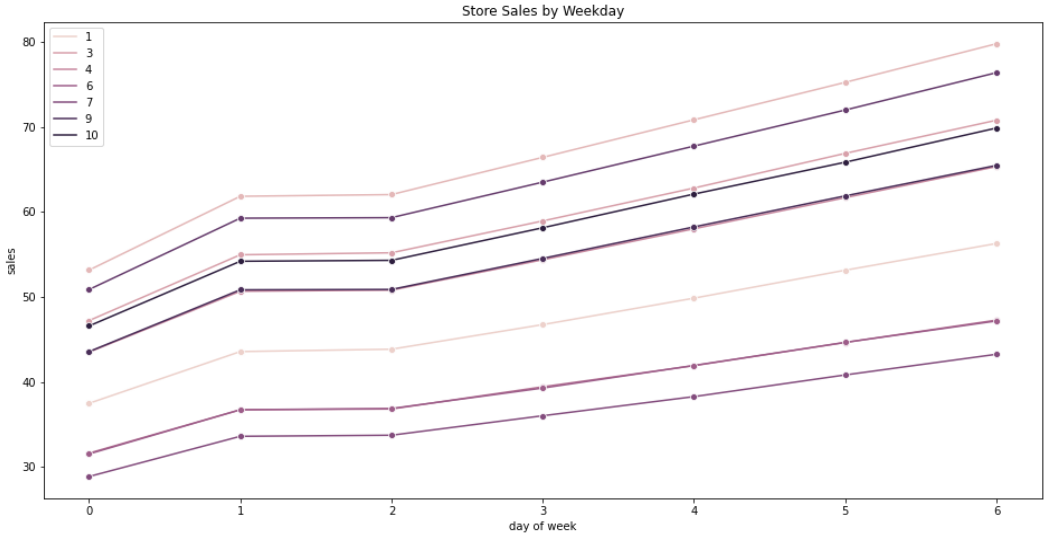
Daily Trends

The graphs in this section show the daily trend in sales across the days of the week, both for all 10 stores combined and by store.



The first graph above shows the daily sales trend for all stores combined. We can observe that with the exception of the stagnancy in sales volume between Tuesday and Wednesday, there is an upward trend of sales across the days of the week, from Monday to Sunday. Sales peak on Sunday, which can explain the typical weekend rush and the large volume of purchases that consumers tend to make on Sundays, in preparation for the week ahead.

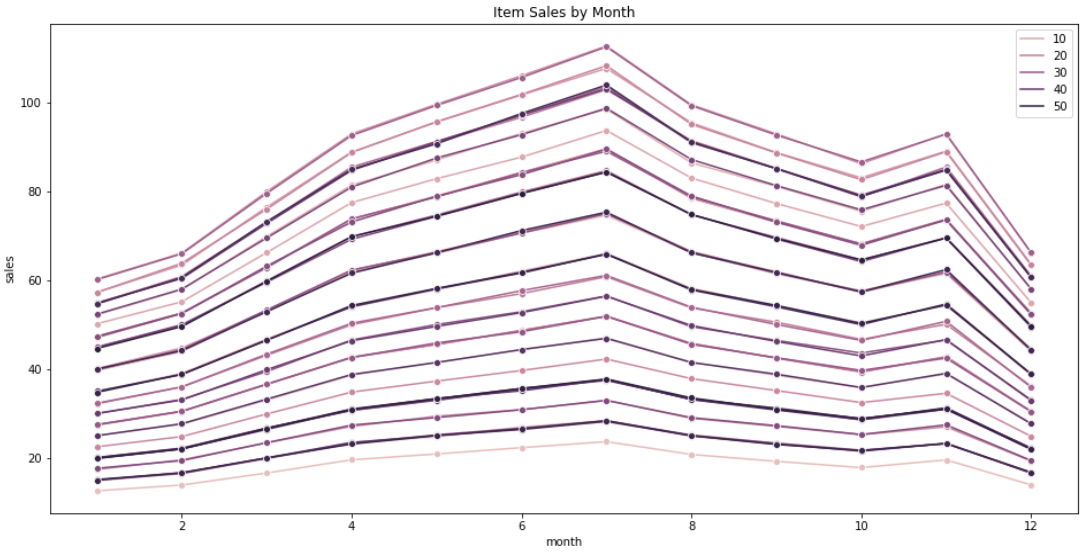
The hue in the graph above represents the different items being sold, and we can observe their respective sales across the week. Each category of items is spread out evenly along the y-axis across the sales figures, occupying both ends and the mid-range of the spectrum.



The second graph just above shows the daily sales trend on a store-by-store basis. In addition to the observations made previously, we can also closely observe the comparison of the daily sales of specific stores as compared to the others. Stores 7 and 3 have the lowest and highest sales respectively across days of the week. Stores 7, 6 and 1 fall within the lower half in terms of their relative sales volume for the store while the remaining Stores 3, 4, 9 and 10 fall within the higher half in terms of their relative sales volume for the store. The reason for the differences in sales across stores could be due to the location of the store (in a prime location closer to residential areas or further away from them), the quality of service at a store, the promotions provided by specific stores and several other factors.

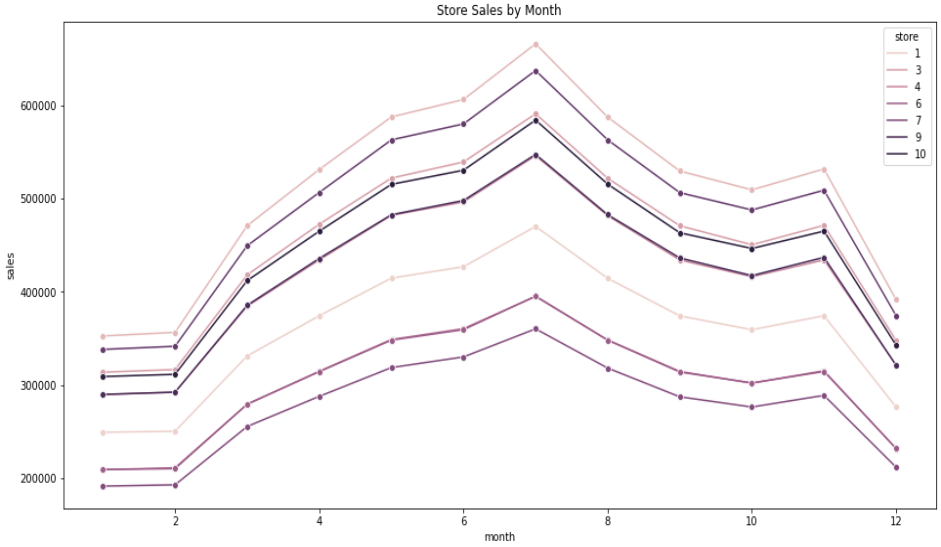
Monthly Trends

The graphs in this section show the monthly trend in sales across the 12 months of the year, for all 10 stores combined and by store.



The first graph above shows the monthly sales trend for all stores combined. We can observe that there is a consistent rise in item sales from January to July, after which it drops until October, rises again in November and drops again sharply in December. The reason for the peak of sales in July could be due to the increase in purchases over the summer holiday period, while the sharp dip in sales in December could be due to the decrease in purchases caused by the higher number of consumers traveling and being away from the country during that period. Thus, the sales depend on and fluctuate according to the seasonal consumer behavior patterns.

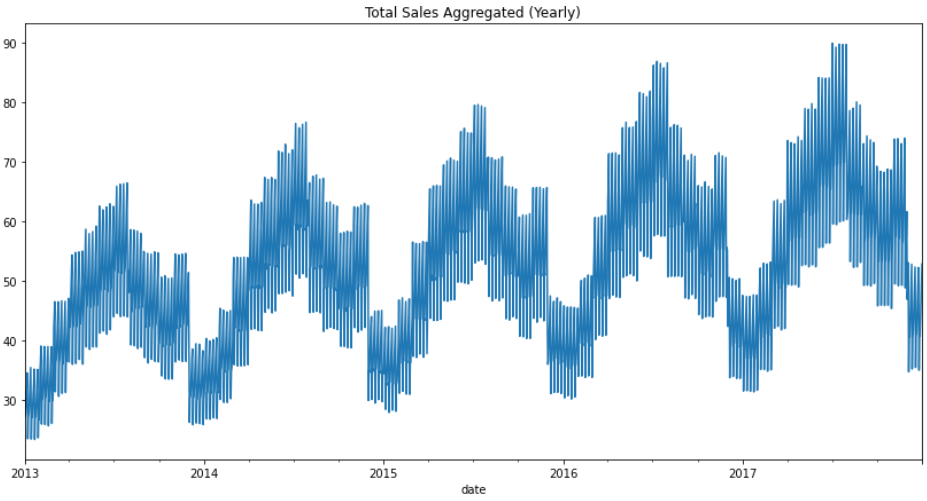
The hue represents the different items being sold, and we can observe their respective sales across the year.



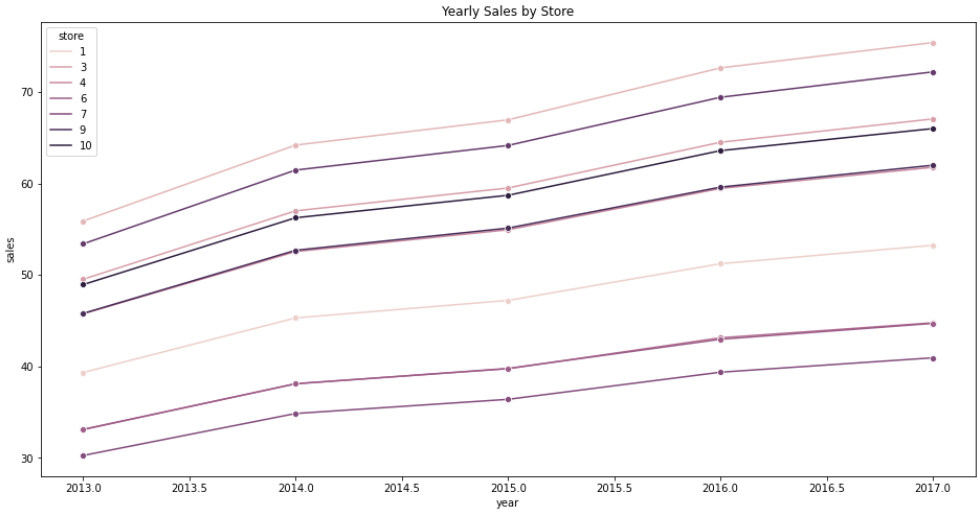
The second graph just above shows the daily sales trend on a store-by-store basis. In addition to the observations made previously, we can also closely observe the comparison of the monthly sales of specific stores as compared to the others. Just as we observed previously under the daily sales trends, Stores 7 and 3 have the lowest and highest sales respectively across the months in the year. Stores 7, 6 and 1 fall within the lower half in terms of their relative sales volume for the store while the remaining Stores 3, 4, 9 and 10 fall within the higher half in terms of their relative sales volume for the store. The reason for the differences in sales across stores could be due to the location of the store (in a prime location closer to residential areas or further away from them), the quality of service at a store, the promotions provided by specific stores and several other factors.

Yearly Trends

The graphs in this section show the yearly trend from 2013 to 2017, for all 10 stores combined and by store.



The first graph above shows the yearly sales trend for all stores combined, on a year-by-year basis. We can observe fluctuating sales over each year, starting off with a rise in sales volume, followed by a decline. The fluctuation in sales seems to follow a standard pattern every year. We can also observe that the sales volume across all the stores increased every year, from 2013 to 2017. This rise could be due to a rise in demand for the items, inflation, increase in standard of living and several other factors.



The second graph just above shows the daily sales trend on a store-by-store basis. In addition to the observations made previously, we can also closely observe the comparison of the yearly sales of specific stores as compared to the others. Just as we observed previously under the daily and monthly sales trends, Stores 7 and 3 have the lowest and highest sales respectively across days of the week. Stores 7, 6 and 1 fall within the lower half in terms of their relative sales volume for the store while the remaining Stores 3, 4, 9 and 10 fall within the higher half in terms of their relative sales volume for the store.

The reason for the differences in sales across stores could be due to the location of the store (in a prime location closer to residential areas or further away from them), the quality of service at a store, the promotions provided by specific stores and several other factors.

5. Results and Evaluations from EDA

After performing a detailed exploratory data analysis, our review suggests that the data is synthetic and we base our conclusion on the following assessments

- Seasonality patterns are consistent without any aberrations.
- The weekly, monthly and yearly seasonal patterns are very stable for every store-item combination.
- Seasonal sales for all item/store combinations follow the same pattern and thus prediction of one combination allows us to predict the rest.

6. Models

The goal of this project is to predict the sales for 2018 (from January to March) and we implemented two approaches.

LightGBM (Benchmark Model)

Light GBM is a gradient boosting framework that uses a tree based learning algorithm. It is unique because it grows trees vertically while other algorithms grow trees horizontally meaning that Light GBM grows tree leaf-wise while other algorithms grow level-wise. It will choose the leaf with max delta loss to grow.
When growing the same leaf, the Leaf-wise algorithm can reduce more loss than a level-wise algorithm which makes this model enticing for our use case. This property has led to boosting algorithms being some of the best performing models in Kaggle competitions recently.

Motivation and Rationale

Light GBM was chosen because of the following reasons:

- Light GBM performs well for categorical features and thus is a great fit for our dataset which contains discrete variables within a specific range.
- Light GBM can handle the large size of data and takes lower memory to run.
- It is mainly used for ranking, classification while the development focus is on the performance and scalability.
- Its main principle is to boost the set of weak learners to strong learners.
- Moreover, it also gives importance to instances that are misclassified.

All of these reasons made LGBM a good fit for the dataset at hand and thus LGBM was used as a baseline model.

Preprocessing

The following actions were undertaken to preprocess the data for this approach :

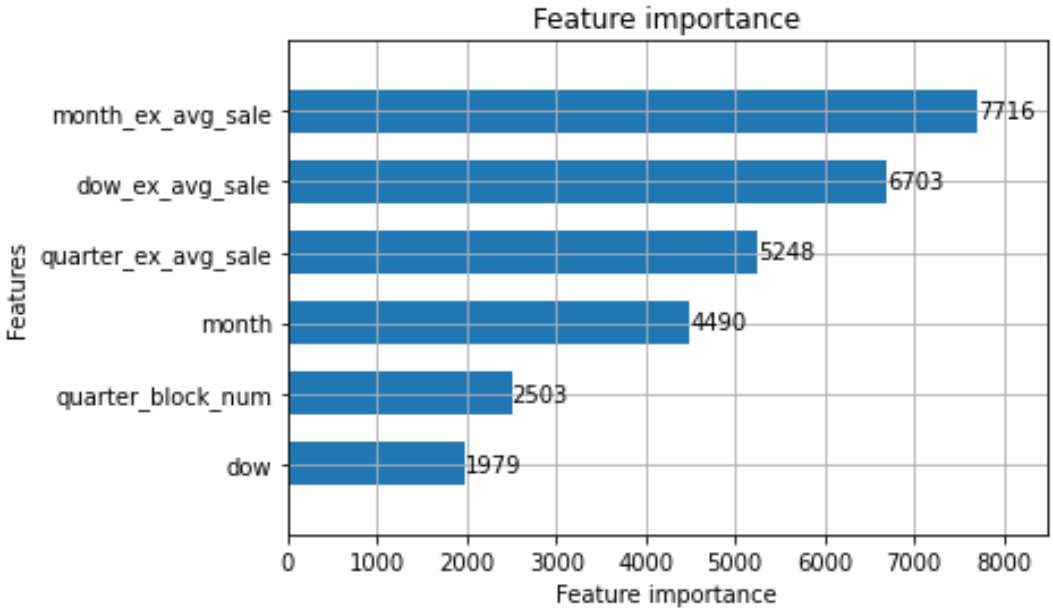
- Finding the slope and intercept of a linear fit for sale values grouped by *store*, *item* and *day of week,* followed by fitting a linear model to the sale values grouped by *store, item* and *day of week* to use this as a trend.

- Removal of increasing trend and yearly seasonality from the sale values. Followed by normalisation of the stationary sale values and identification of the outliers using the normalized stationary sale values.

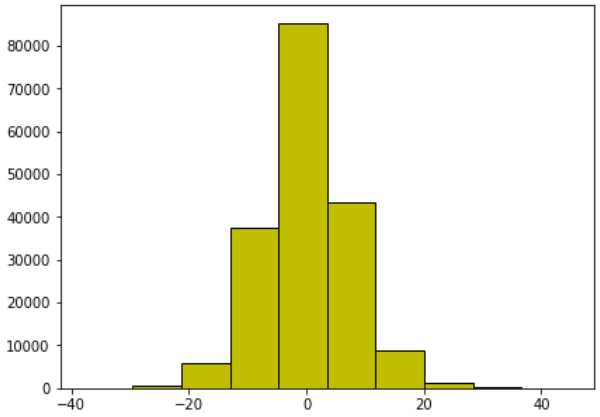
- Handling the outliers using interpolation to get corrected sales and building the expanding mean sale values which are grouped by *store, item, (day of week, month,* and *quarter*).

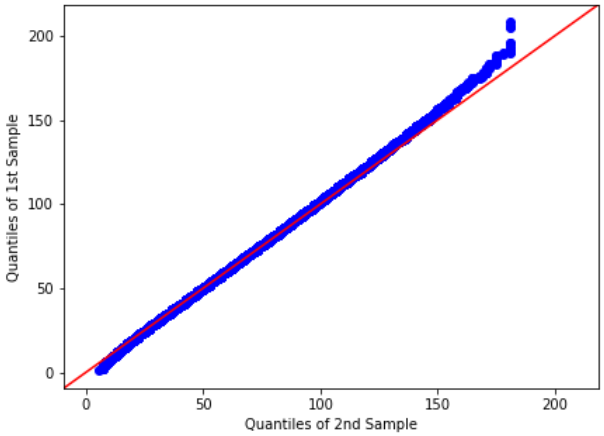
- Finding *stores* and *items* whose mean sales value is below the 50% percentile. Thereby, in the prediction for year 2018 of these stores, these items will be multiplied by a factor smaller than one.

Feature Importance



Error Distribution and Analysis





LGBM for 2018 Prediction

The prediction is carried out by initially defining the parameters and hyper-parameters. For better results, the hyper-parameters should be tuned carefully. Moreover, more critical store items, the prediction is multiplied by a factor slightly smaller than 1.

parameter	**value**
Task	Train
Boosting type	GBDT
objective	Regression
Num leaves	10
Max depth	3
Metric	SMAPE
Learning rate	0.1
Boosting rounds	10000

Final Prediction Score

Public Leaderboard	**Private Leaderboard**	**Validation Score**

13.94421	**12.67741**	**12.59647**

Analytical Solution (Final model)

Motivation and Rationale

Our conclusion from the EDA suggested that the dataset is synthetic and thus an analytical solution would be the most appropriate for our problem.

We created a baseline model using item sales combinations and subsequently experimented with different configurations to find the best approach. We immediately surpassed our previous scores using the LGBM with only our rudimentary baseline models which indicated that the analytical solution is in fact the best approach.

Solution

The baseline model is a linear combination of the predictors with no hyperparameters. This approach is different from a regression model as we are not explicitly training the model or learning weights (regression is only used to find the annual growth). Furthermore, the predictors are only transformed using aggregations or normalized using the mean to compute their associated factors.

In the next few iterations, we experimented with various approaches to find the best possible solution and thus our final analytical solution is detailed as follows

$$predicted\ sales = sales\ (store, item) * salesFactor\ (dayoftheweek) * salesFactor\ (month) * growth(year)$$

$$where\ salesFactor\ is\ defined\ as\ \frac{aggregated\ sales\ (on\ any\ column)}{average\ sales\ (total\ cummulative\ sales)}$$

We subsequently used alternative configurations to find the best model and using the following approaches

$$predicted\ sales = sales\ (store) * salesFactor\ (dayoftheweek) * salesFactor\ (month) * growth(year)$$

$$predicted\ sales = sales\ (store, item, month, dayoftheweek) * Growth(year)$$

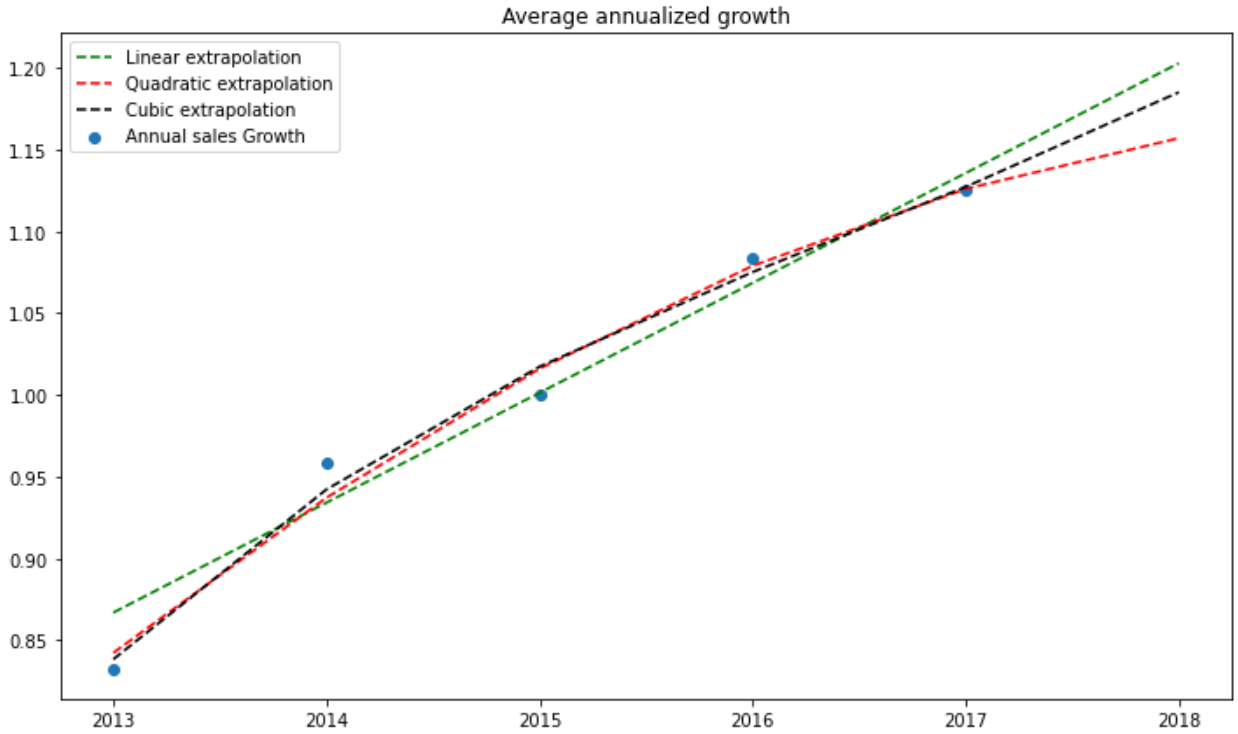
$$predicted\ sales = sales\ (store) * salesFactor\ (dayoftheweek) * salesFactor\ (month) * growth(year)$$

$$predicted\ sales = average\ sales * salesFactor\ (dayoftheweek, store, item) * salesFactor\ (month) * growth(year)$$

$$predicted\ sales = average\ sales * salesFactor\ (dayoftheweek, item) * salesFactor\ (month, item) * salesFactor\ (store, item) * growth(year)$$

The sales (store,item) are computed using the aggregated values of the original dataset using the two attributes. For the salesFactor, the aggregated data frame was normalized for a particular column using the mean and this provided an approximate growth pattern.

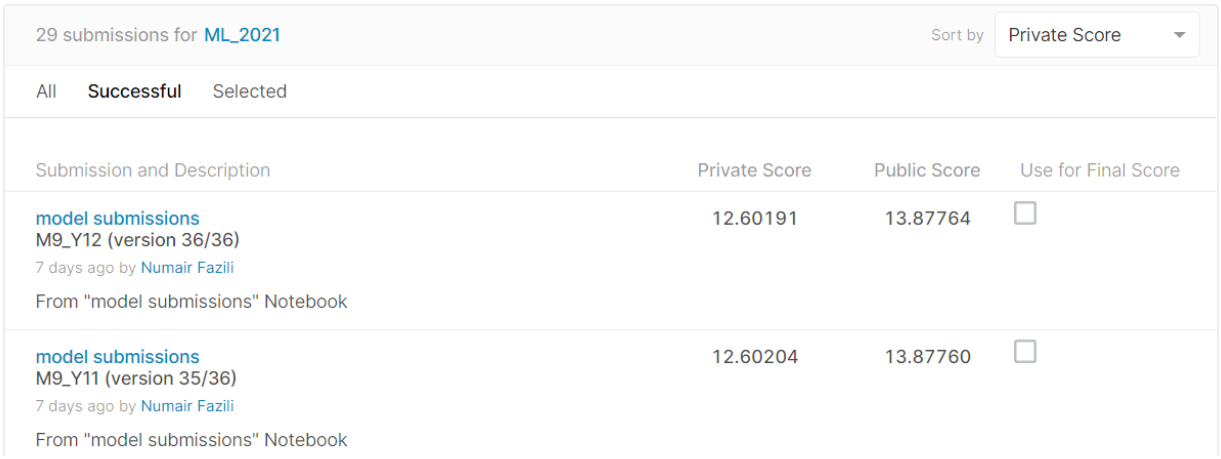
To compute the annual growth, we performed cross-validation on yearly growth with polynomial fits of linear, quadratic, and cubic degrees using the numpy library.

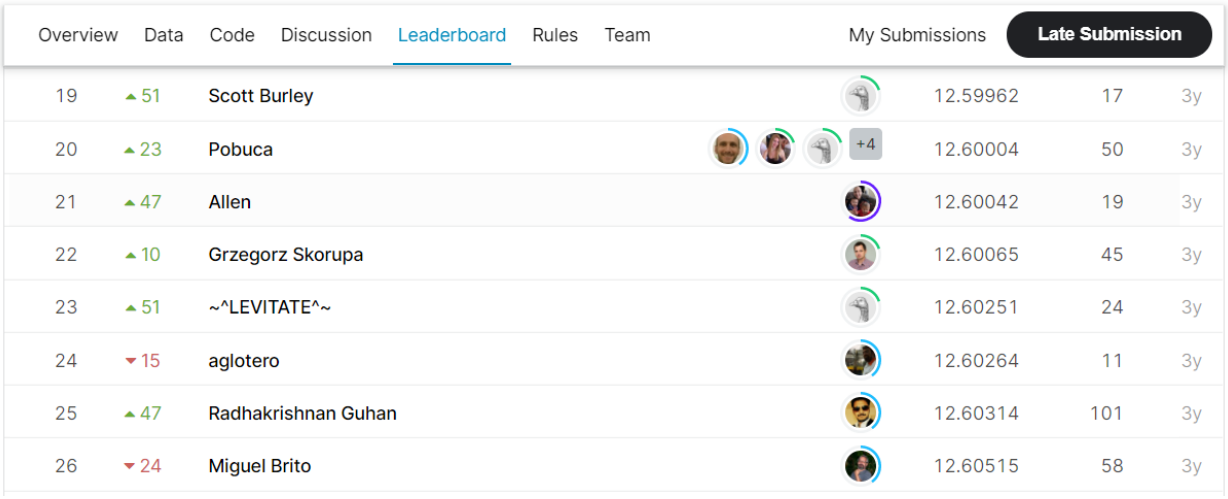


The model for yearly growth uses a polynomial fit of quadratic degree which has the optimal tradeoff between bias and variance. For the polynomial regression, the weights decay exponentially increases over years. This enables us to prioritize recent sales (sales in 2017 have more bearing on 2018 sales in comparison to sales in 2013).

Results

Final Score	Leaderboard	Position
12.60191	**Private**	**23/459 (Top 5%)**





7. Summary

In this project, we analyzed the sales data as part of a Kaggle competition. We started by preprocessing the data and then implementing exploratory data analysis to study the variables and their corresponding relationships in detail. This process allowed us to understand the seasonality trends, sales, and volatility patterns. We also concluded that the dataset provided to us was synthetic and thus prepared our subsequently modeling approach accordingly.

The dataset provided contained almost a million records and thus for the predictive model, we required an algorithm that was fast and accurate. To this end, we used LGBM as a benchmark model due to its fast and high-performance gradient boosting approach. While the results obtained from this approach were decent, the scores on the Kaggle leaderboard were still too far ahead and even after tuning the hyperparameter, we were unable to make advances in the leaderboard rankings.

To this end, we moved towards developing analytical solutions to solve this problem. The intuition behind the analytical solution was that since the sales patterns are consistent with respect to the seasonalities (day of the week, month, year), leveraging this information should provide a decent estimation for future predictions. Therefore, we experimented with various configurations in pursuit of the best model. The final model implemented allowed us to break into the top 5 percent of the Leaderboard ranking with a SMAPE difference of only 0.0217 to the best submission.

8. Challenges

We faced numerous challenges during the course of this project especially since none of our team members had any prior experience with time series modeling or Kaggle competitions.

Computational Bottlenecks

The training dataset had almost a million records, this made training difficult due to resource constraints on our local machines. Therefore, we used google collaboratory for conducting EDA and training our models. The resources provided by Google collab helped us reduce the training time by over 500%.

Exploratory Data Analysis

The scale of the dataset made it difficult to analyze and draw conclusions using simple plots. Therefore, we had to apply some group by transformations (for example, on the ‘item’ data field) to be able to analyze the seasonality sales patterns more effectively.

Tuning LGBM Model

Model tuning requires lots of trial and error as some parameters may prove not to be useful in improving the overall score. Furthermore, though tuning the parameters may result in a better score, over-tuning the parameters may be detrimental to the training model as too many iterations may result in overfitting of the model.

Tuning Analytical Model

Finding the best configuration for the analytical model was the biggest challenge we faced. The slightest of the modifications caused the score to fluctuate dramatically and thus required a significant number of trial and error approaches towards the final solution.

9. Learnings and Conclusion

This project provided us with an exceptional opportunity to work together as a team in solving the machine learning challenge on Kaggle. Moreover, the competitive nature of the problem made it difficult to break into the top rankings of the leaderboard. Nevertheless, all the effort and hard work paid off as we were able to enter the top 5%.

Frameworks

We learned how to decompose machine learning tasks effectively such that there is minimum overlapping between tasks while ensuring all tasks are completed as per schedule. Moreover, we maintained strict quality standards by cross-verifying each other's components.

Communication

Communication between members was an important aspect of our project, and the work by the EDA team and their recommendations allowed the modeling team to effectively capture the intrinsic details of the dataset and plan the execution strategy accordingly. The process was interactive and this constant communication cycle between the two teams increased the quality of work while reducing the overall time and effort by individual members.

Python Programming

While conducting the EDA, our queries were inefficient due to which rendering times were too long. We addressed this issue by using group by aggregations provided by pandas. These are powerful aggregation queries that significantly improved our code performance.

10. Work Distribution

Task	**Assignee**
Exploratory Data Analysis	Puneet Tekwani, Nikhita Menon
Models	Numair Fazili, Manav Arora
Documentation (Report and Presentation Slides)	Nikhita Menon, Numair Fazili, Vishal Raj, Manav Arora
Presentation Video	Nikhita Menon, Vishal Raj, Numair Fazili, Puneet Tekwani