Jobathon – Sales Forecasting

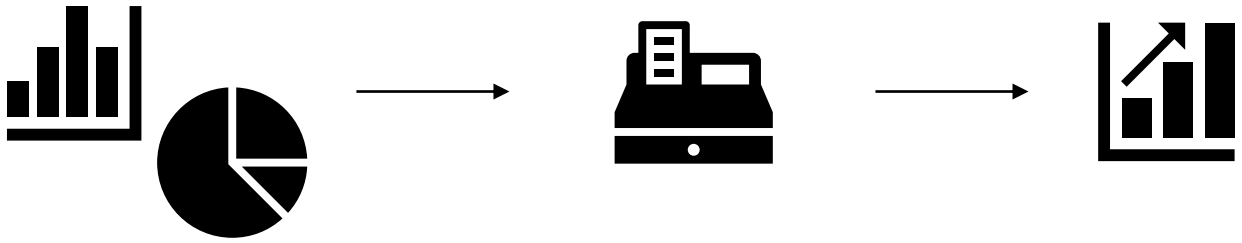
Ramya D

September 2021

Contents

Objective

Predict the store sales for each store in the test set for the next two months.



If we can accurately predict the upcoming cash flows, WOMart can more accurately plan the cashflow at the store level.

Background

Effective insights into upcoming cash flows at a store level is crucial to WOMart's business strategy.

WOMart is a leading nutrition and supplement retail chain that offers a comprehensive range of products for all your wellness and fitness needs.

WOMart follows a multi-channel distribution strategy with 350+ retail stores spread across 100+ cities and effective insights into upcoming cash flows at a store level is crucial to its business strategy.

Brief overview of approach

Two-layer stacking model based on ML regressors on engineered features

- Our modelling approach is to build a regression model rather to estimate the store value whether it is future, current or past with respect to the given data instead of simply building a forecast model at each store level to forecast the future sales data.

- The advantage of our approach is that it restricts the complexity and number of models require to model Sales at a store level

- Also given the additional information available like discount and holiday etc, this approach helps us to account these effects as well unlike the classic time series/statistical models which needs explicit modeling based on exogenous variables

- We have built a two layer stacking model with LR, RF, XGB, GB, BAG regressors in the first level with a Elastic net in second level using engineered features to account seasonality present in the dataset

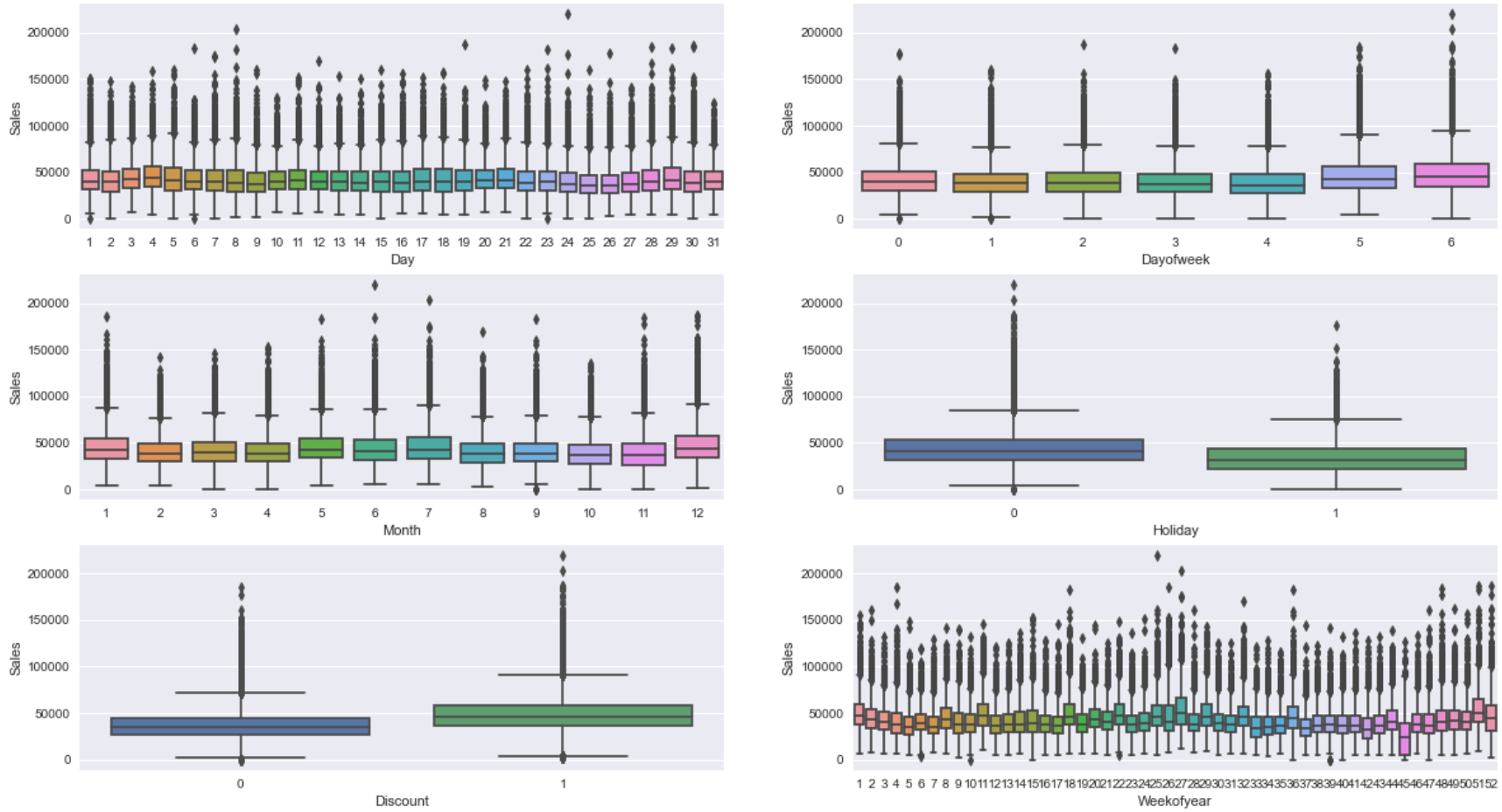
Pre-processing/EDA/Feature Engineering

Standard data checks on train dataset

Sl.No	Check	Presence in Data	Approach
1	Incorrect rows	None	
2	Summary rows	None	
3	Dulicate/Extra/Unnecessary rows	None	
4	Missing Column Names	#Orders not present in Test	Excluded from modelling
5	Inconsistent column names	None	
6	Unnecessary columns	Row ID	Excluded from modelling
7	Columns containing Multiple data values	The data is a collection of time series data of 365 stores	
8	No Unique Identifier	None	
9	Misaligned columns	None	
10	Disguised Missing values	None	
11	Significant number of Missing values in a row/column	None	
12	Non-standard units	No mention of currency	Ignored
13	Abnormally high and low values	None	
14	Non standards in text variables	None	
15	Incorrect data types	Store id is given as an integer	Treated it as a categorical variable and relevant encoding performed
16	Correct values beyond range	Sales and orders are non-negative	
17	Validate internal rules	Sales is zero when orders is zero and vice versa/ Each store has exactly 517 observation of time series	

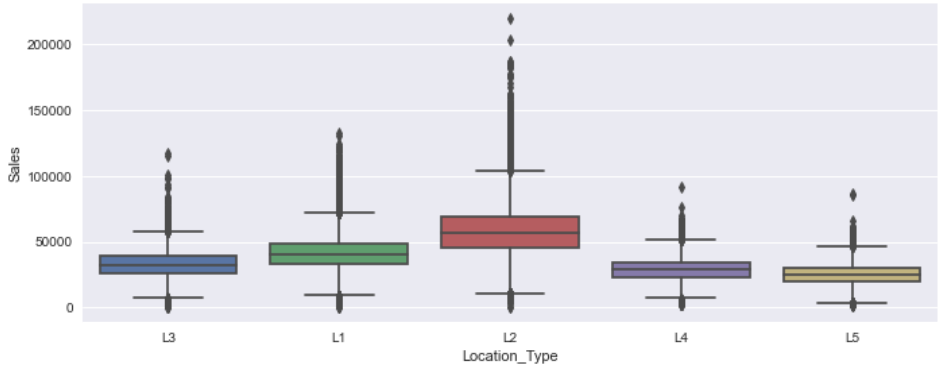
Pre-processing/EDA/Feature Engineering (contd)

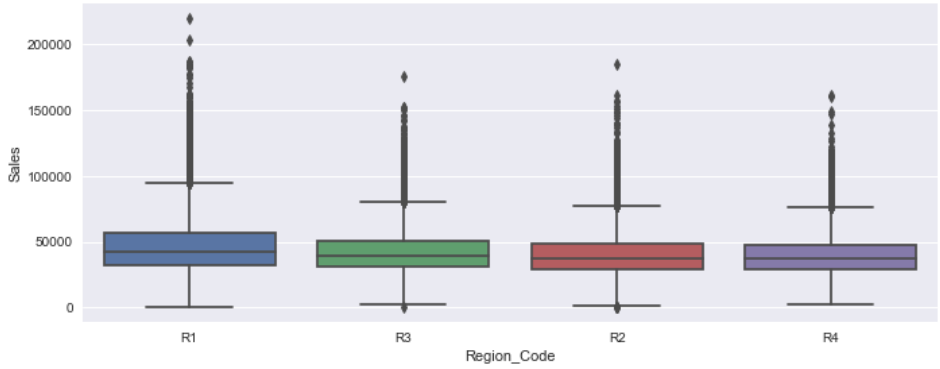
Sales is a highly seasonable variable – Discount and Holiday has a great effect on it

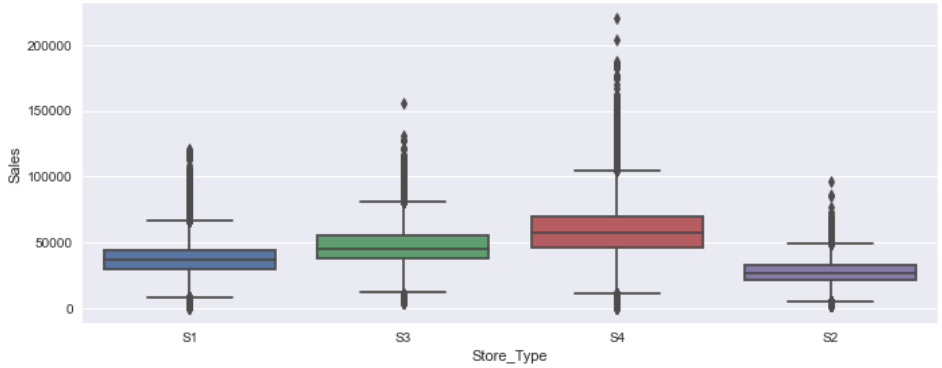


Pre-processing/EDA/Feature Engineering (contd)

L2 location and S3 store type generate higher sales on an average







Pre-processing/EDA/Feature Engineering (contd)

Engineering various features from the Date column

- Extracted following variables from Date column
  - Month
  - Day
  - Day of year
  - Week of year
  - Day of week
  - Is month start/end, Is quarter start/end

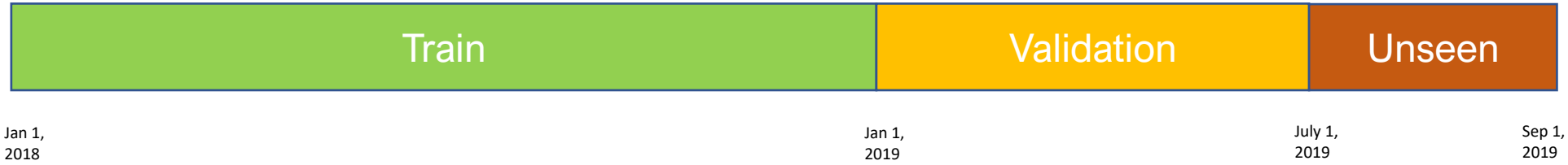
- Transformed seasonable variables Month and Day with sine and cosine transformations to have a period of 12 and 7 respectively.

- Target encoding for store_id variable based on median of sales.
- Transformed Sales column based on Yeo-Johnson transformer to convert into a variable that follow normal distribution

Final Model

Engineering various features from the Date column

- Split the dataset into Train (data till 31 December 2018) and Validation (remaining portion starting from 1st January 2019



- Built the below models on Train dataset with transformed Sales as target. All the Hyperparameters are tuned using a 10 - fold Grid search CV
  - Linear Regression, Random forest Regressor, Gradient boost Regressor, XGB Regressor, Elastic net is used as Meta model in level 2

Final Model

Engineering various features from the Date column

Sl.No	Model	R2 score	RMSE
1	Linear Regressor	58.58	12529
2	Random forest Regressor	57.18	12739
3	XGB Regressor	63.33	11788
4	Gradient Boosting Regressor	64.04	11673