Case Study (PMI)

Pankaj Chandra

Master Degree in Data Science and Process Engineering

Problem Set

Every company wants to succeed and gain an edge on the competition. Achieving the revenue goals translates into maximizing sales. Many companies distribute their goods at physical Point Of Sales (POSs). For all of them the challenge is to devise a strategy that will drive the sales at POSs. Possible solution could be to place the product in the most convenient location for consumers. In this assignment, it is to figure out what surroundings and respective amenities lead to top POS performance.

Proposed Solution : This problem can be taken as Classification Task where target variable can be taken as Total Sales and it can be divided into binary classes of High Sales and Low Sales bracket, furthermore, once the classification model is built then analysis of the dominant surrounding amenities (independent variables) can be studied for their impact in Sales.

Data Retrieval

For performing this analysis, the following are the data sources:

- ‘sales_granular.csv’ - contains information about the sales volumes of a product at particular POS; each POS is uniquely identified by 'store_code'.
- ‘Surroundings.json’ - contains information about 90 different amenities (restaurants, shops, beauty salons etc.) that are in the surroundings of each POS.

Snapshot of the Data

Granular Sales Data

	store_code	8/3/15 9:00	8/3/15 10:00	8/3/15 11:00	8/3/15 12:00	8/3/15 13:00	8/3/15 14:00	8/3/15 15:00	8/3/15 16:00	8/3/15 17:00	8/3/15 18:00	8/3/15 19:00	8/3/15 22:00	8/4/15 8:00	8/4/15 9:00	8/4/15 10:00	8/4/15 11:00	8/4/15 12:00
0	10055	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	10077	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Surrounding Amenities Data

	store_code	subway_station	department_store	embassy	beauty_salon	police	courthouse	cemetery	pharmacy
0	10055	0	0	0	4	0	0	0	3

Data Wrangling (After a combined Data Frame Queried)

- Combination of the two dataset is established taking store codes as the connecting link.
- Two additional features are added : Time of the day and Total Sales aggregated over approx. 23 months for each time stamp. This also helps in augmentation of data since if one row per store code is used then there will be only around 500 observations but using time (hour) of the day as one of the features has augmented the observations to around 13000.
- All negative values are imputed as zero. (could be a mistake in logging)
- Check for null values.
- Check for duplicates.
- Two columns were removed because of zero standard deviation.

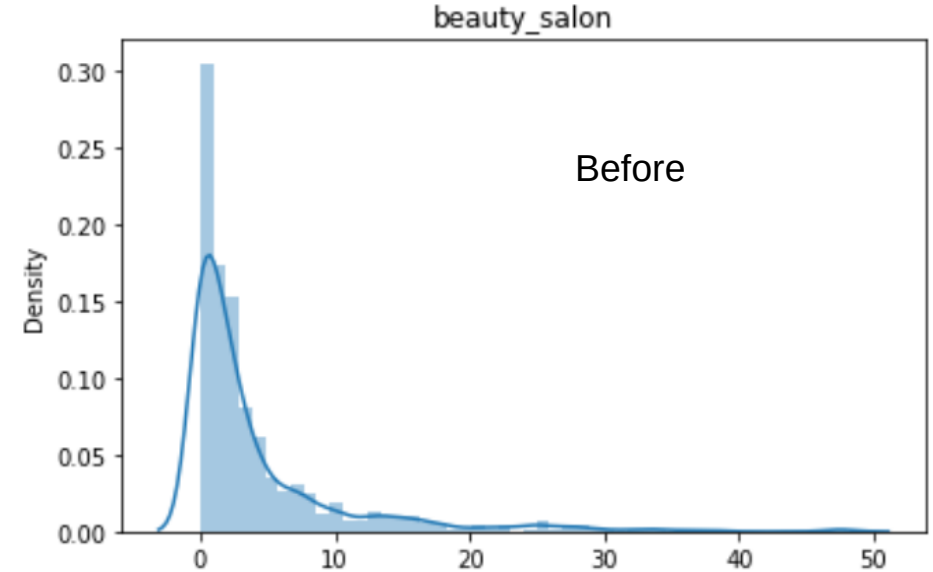
Final DataSet

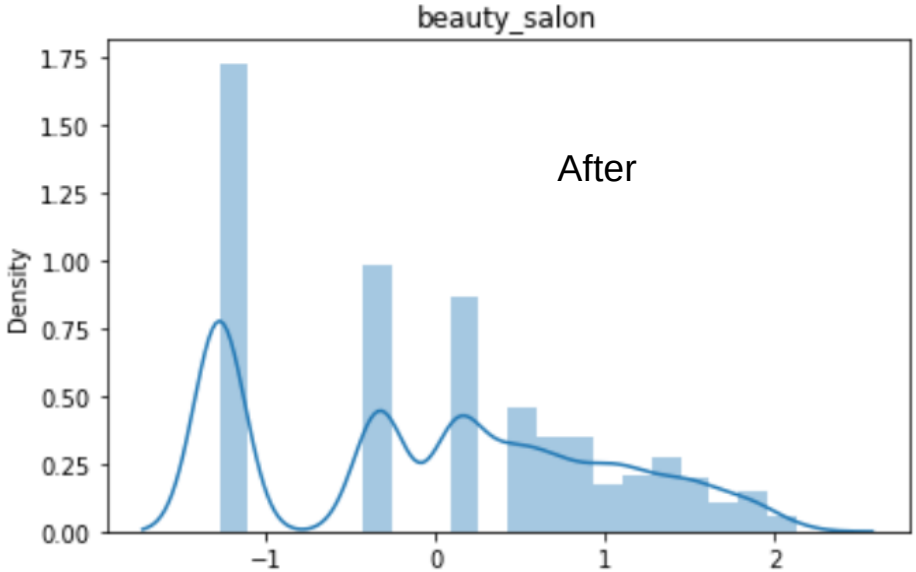
Store Code	Ameneties 1-90	Time of the Day	Total Sales
10599	xxx	9 am	yyy
10599	xxx	10 am	yyy
10899	ccc	9 am	zzz
10899	ccc	10 am	zzz

Target variable is chosen to be **Total Sales** per Time Stamp. This is an obvious choice because aim is to find impact of surrounding amenities on Sales.

Exploratory Data Analysis

Yeo-Johnson Transformation for achieving Gaussian like Distribution for different features

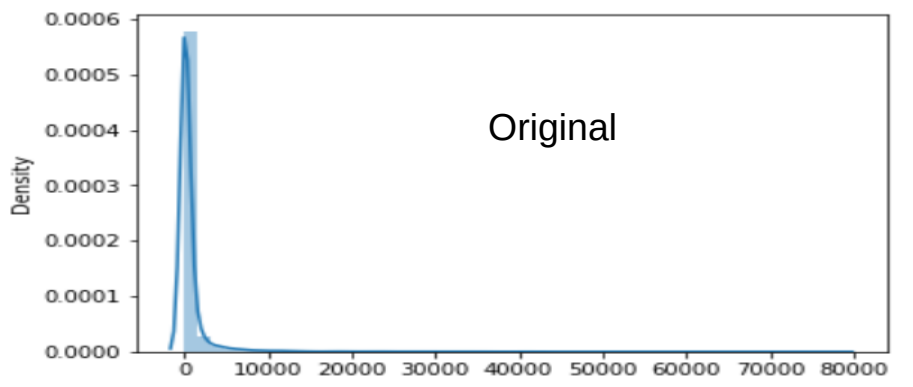


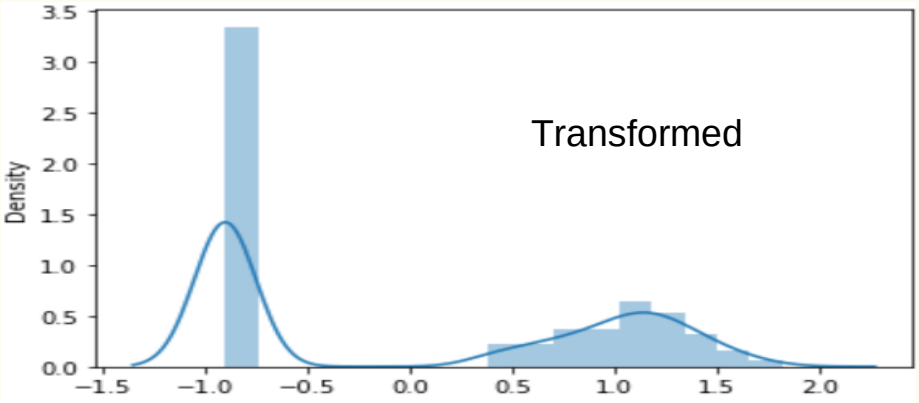


Most Features are Right Skewed in the dataset

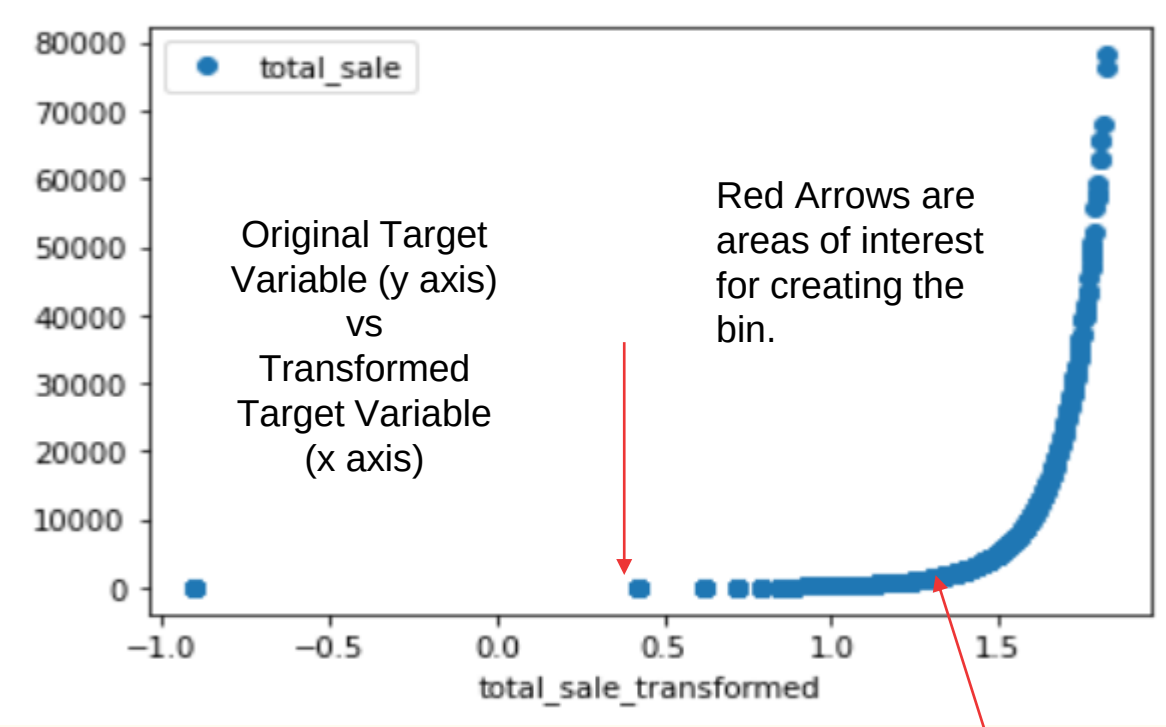
Target Variable is also Right Skewed

	total_sale
count	12960.000000
mean	843.291667
std	3711.704175
min	0.000000
25%	0.000000
50%	0.000000
75%	390.000000
100%	78180.000000
max	78180.000000





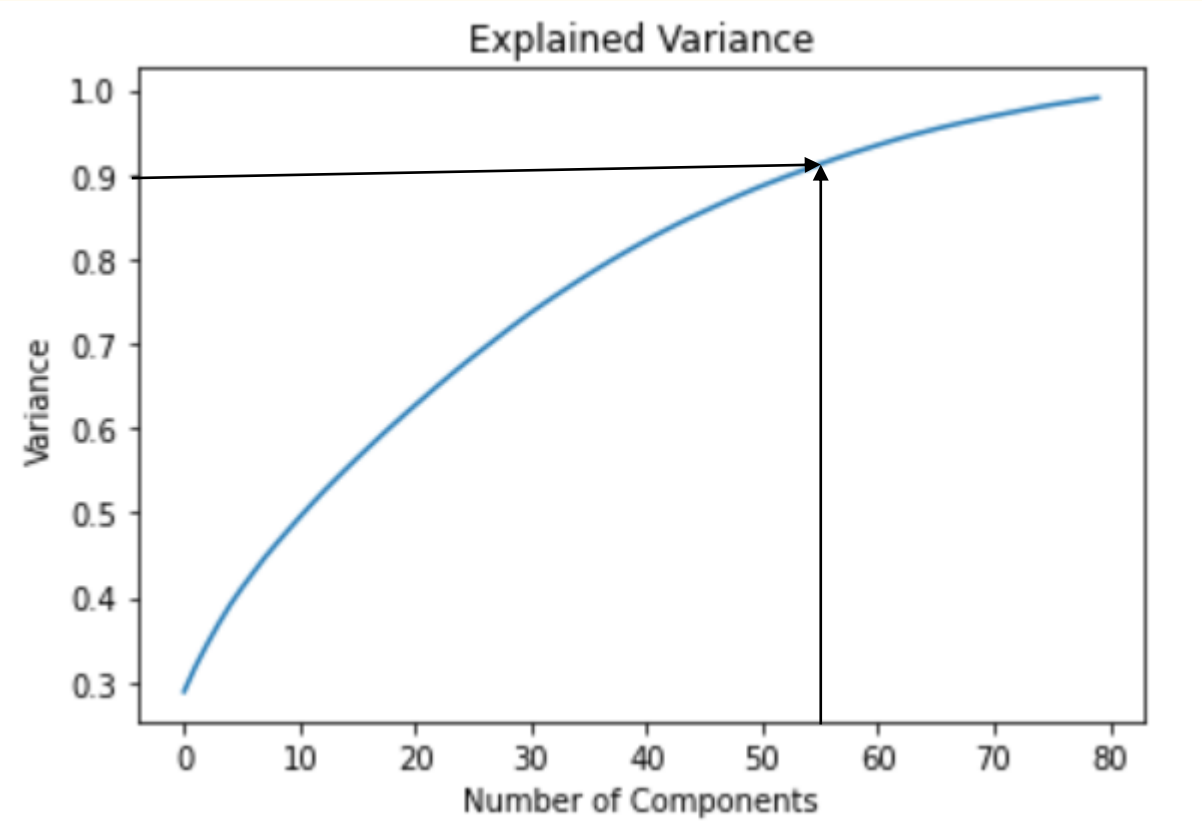
Creation of Bins on Target Variable



These two points of interest are analyzed for binning the target variable into two class :

- Low Sales Volume Cases
- High Sales Volume Cases

PCA and VIF for the reduction of dimensions



Principal Component Analysis(PCA) suggests that 90 percent of the variance in the dataset can be explained by 55 components/features.

Variance Inflation Factor (VIF) Study also showed that 23 features were redundant. Therefore, there were removed from the Study. (Pl.see github link)

Model Building (70 % Training and 30 % Testing)

The classification algorithm used is XGBoost i.e. a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. Following are the prediction made by the model on the test dataset.

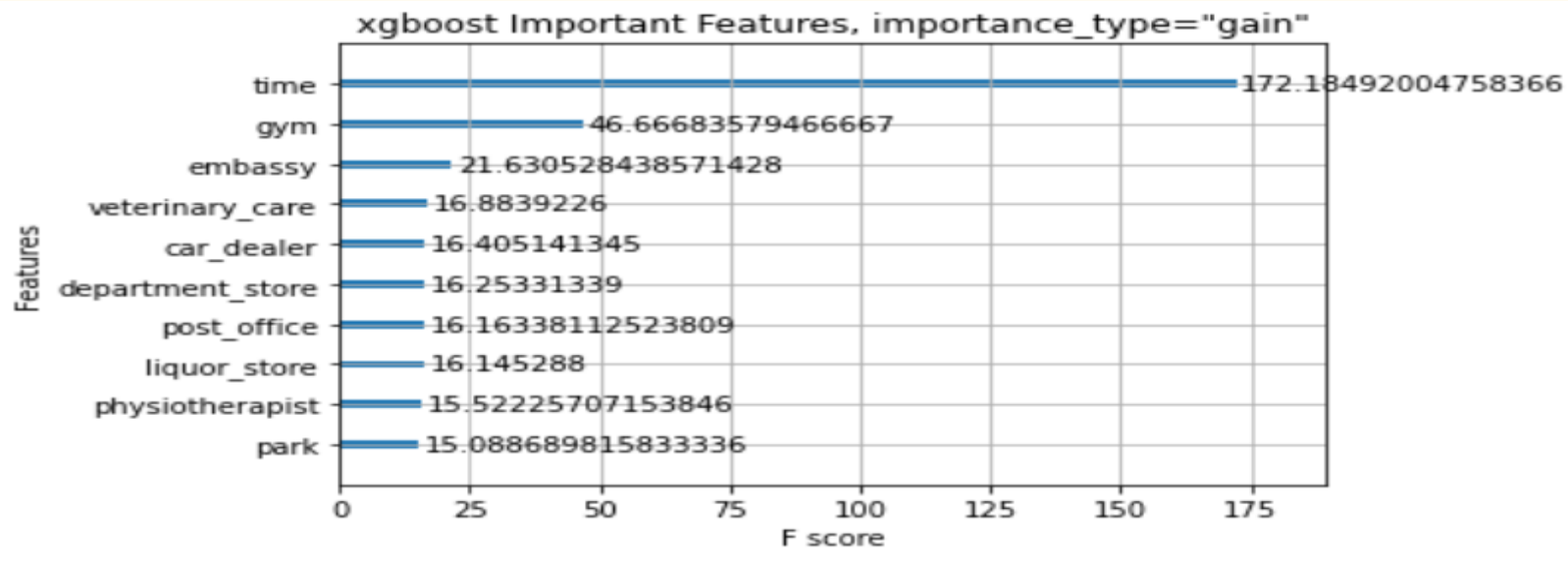
Confusion Matrix	Predicted Low Sales	Predicted High Sales
Actual Low Sales	1807	305
Actual High Sales	161	1615

Accuracy Score 0.8801440329218106

	precision	recall	f1-score	support
0	0.92	0.86	0.89	2112
1	0.84	0.91	0.87	1776
accuracy			0.88	3888
macro avg	0.88	0.88	0.88	3888
weighted avg	0.88	0.88	0.88	3888

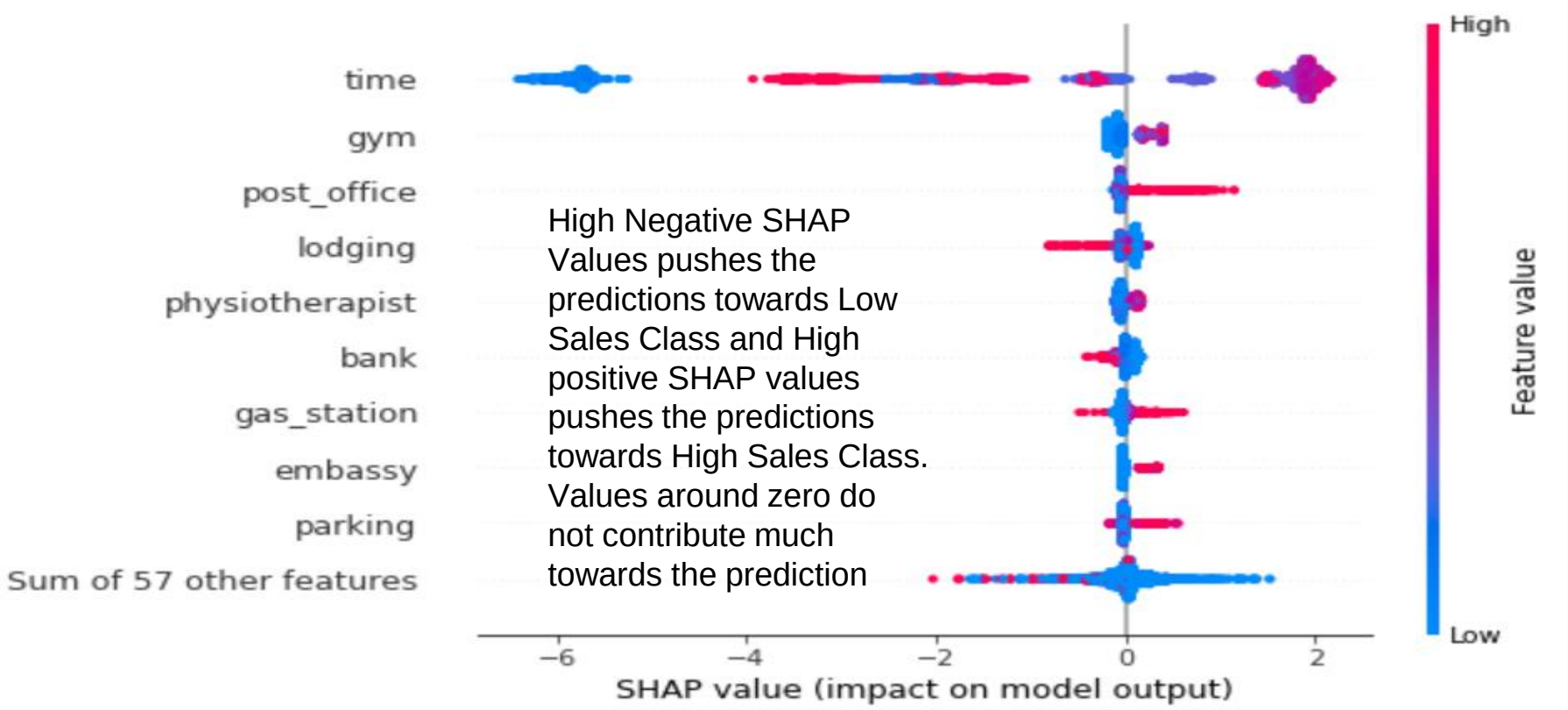
A good f1 score on the both the classes suggests the model is well equipped for the classification task.

Feature Importance (Top 10 out of 60+ features)

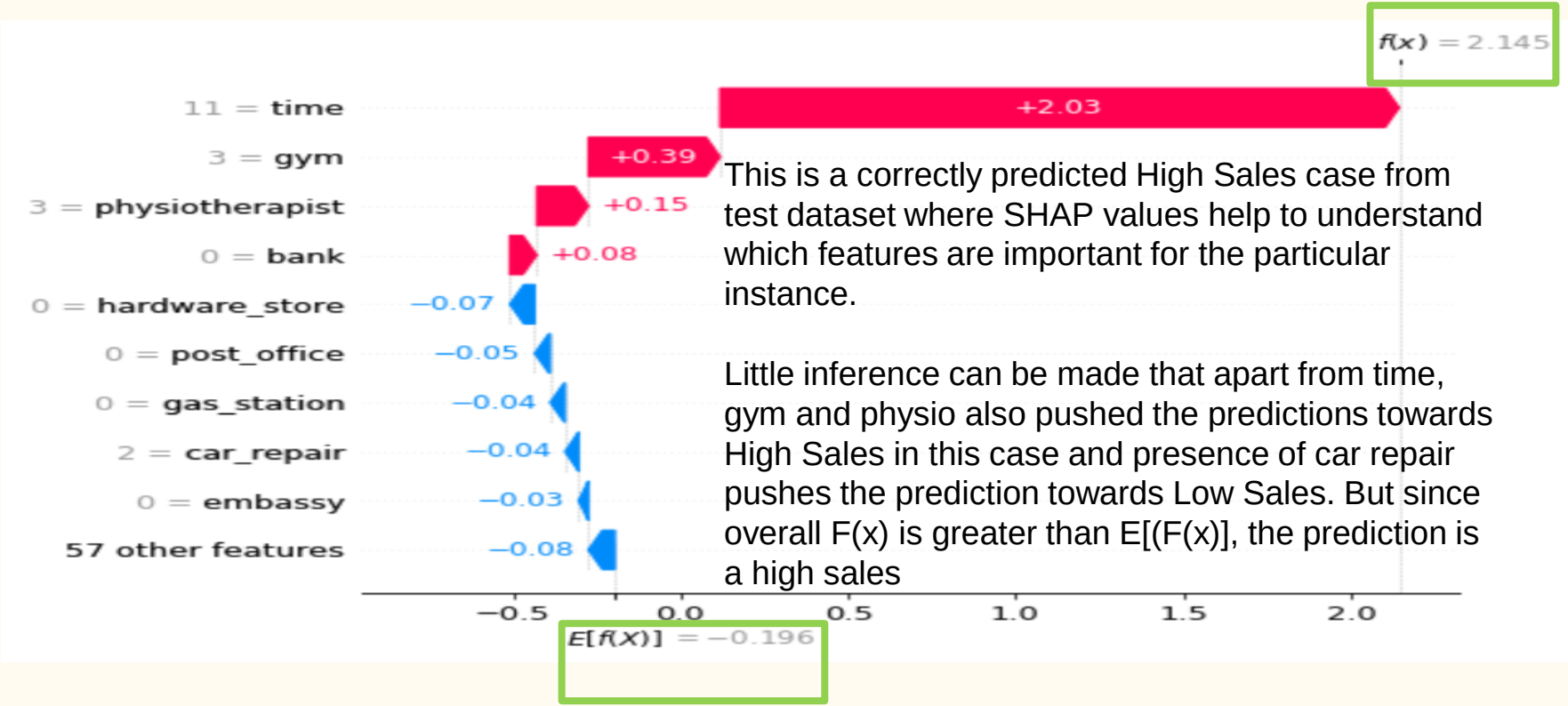


The Gain implies the relative contribution of the corresponding feature to the model calculated by taking each feature's contribution for each tree in the model. A higher value of this metric when compared to another feature implies it is more important for generating a prediction.

Global Agnostics using Shapley Values



Local Agnostics using Shapley Values



This is a correctly predicted High Sales case from test dataset where SHAP values help to understand which features are important for the particular instance.

Little inference can be made that apart from time, gym and physio also pushed the predictions towards High Sales in this case and presence of car repair pushes the prediction towards Low Sales. But since overall F(x) is greater than E[(F(x)], the prediction is a high sales

Conclusion on Important Features (i.e., the goal)

- The goal of the assignment is to find the most important amenities that impact sales. Due to the lack of time, initially surrounding features were chosen as predictors and Total sales per Store for the 23 months of data as the target variable, but the model I built could not perform the classification task between high sales data and low sales data.
- Once the inclusion of time factor is made, then model starts to perform much better but sad part is time is not the feature corresponding to the amenities (dominant ones) which one needs to focus here. The suggestion here for the current state of assignment is to utilize all other features except time to conclude the assignment for now. Therefore, following features can be considered as highly important : **Gym, Embassy, Post Office, Gas Station and Bank.**

How this Analysis can be Improved ?

- Retrieving features like store locations (coordinates) and then mapping the surroundings with high and low sales to understand if the stores with similar surroundings fall into similar sales bracket or not?
- Studying the interaction of features because a group of few features can influence the model to find better patterns for it's task. Also, this will help to visualize how different interaction impact the sales.
- Obviously, more feature engineering on new features like monthly or quarterly to visualize seasonality in the dataset can help the cause.
- Clustering method could be used to analyze if there are some distinct clusters different ranges of sales data. High dimensional data visualization tools can also be employed such as t-sne; to understand if there are genuine clusters or not after feature engineering.
- Cross Validation and Hyper-Parameter tuning will surely help in achieving a more robust model but for now the aim should be to find a concrete base solution via utilizing extensive feature engineering.