ElecKart Market Mix Modeling

Pratibha Deshmukh

Naveen Masali

Amit Shyamsukha

Nikhil Tiwari

18/03/2018

Agenda

- **Business & Data Understanding**
- Data preparation & Exploratory Data Analysis
- Model Building
- Recommendation
- Challenges Faced

Business Understanding & Objective – ElectKart Market Mix Modelling

Background :

ElecKart is an e-commerce firm specialising in electronic products. To enhance their revenues they have done significant investment in their marketing efforts, like promotions over last one year. They are about to create a marketing budget for the next year which includes spending on commercials, online campaigns, and pricing & promotion strategies. They want to reallocate their budget optimally across different marketing levers to improve the revenue response using Market Mix modelling. CRISP-DM framework will be used for modelling purpose

Objective :

To develop a market mix model for 3 product sub-categories - Camera accessory, Gaming accessory and Home Audio to observe the actual impact of different marketing levers over sale of last one year (July 2015 -June 2016) and recommend the optimal budget allocation for different marketing levers for the next year.

Data Understanding : Business has provided below 5 dataset for analysis.

Media Investment Detail: This file contains different marketing spend at monthly level.

Column Name	Significance
Year	Year
Month	Month
Total Investment	Monthly Total ad spend in CR
TV	Monthly Total TV ad spend in CR
Digital	Monthly Digital ad spend in CR
Sponsorship	Monthly Sponsorship spend in CR
Content Marketing	Monthly Content marketing spend in CR
Online marketing	Monthly Offline marketing spend in CR
Affiliates	Monthly Affiliates spend in CR
SEM	Monthly SEM spend in CR
Radio	Monthly Radio spend in CR
Other	Monthly Other spend in CR

Yearly Promotional calendar:

Column	Significance
Sale details	Promotion Name along with date of sale (e.g. Eid & Rathayatra sale (18-19th July)

Monthly customer satisfaction score: : It contains month wise customer satisfaction score in percentage.

Column	Significance
Month	Monthly customer satisfaction score

Product Details:

Column Name	Significance
Product category	Category name
Frequency	Frequency of the products sold
Percent	Percentage w.r.t to total sales

Data Understanding :

Order Details : It contains daily level order details for a year (July 2015 and June 2016). 1648824 number of rows in order data file

Column Name	Significance
fsn_id	The unique identification of each SKU
order_date	Date on which the order was placed
Year	Year of the order
Month	Month of order
order_id	The unique identification number of each order
order_item_id	Suppose you order 2 different products under the same order, it generates 2 different order Item IDs under the same order ID; orders are tracked by the Order Item ID.
GMV	Gross Merchandise Value or Revenue
Units	Number of units of the specific product sold
Deliverybdays	Dispatch delay from Wearhouse
Deliverycdays	Delivery delay to customer
s1_fact.order_payment_type	How the order was paid – prepaid or cash on delivery
Sla	Number of days it typically takes to deliver the product
cust_id	Unique identification of a customer
Pincode	Zip code
product_analytic_sub_category	Product sub category
product_mrp	Maximum retail price of the product
product_procurement_sla	Time typically taken to procure the product

- Business & Data Understanding
- **Data preparation & Exploratory Data Analysis**
- Model Building
- Recommendation
- Challenges Faced

Data Preparation : Data issues

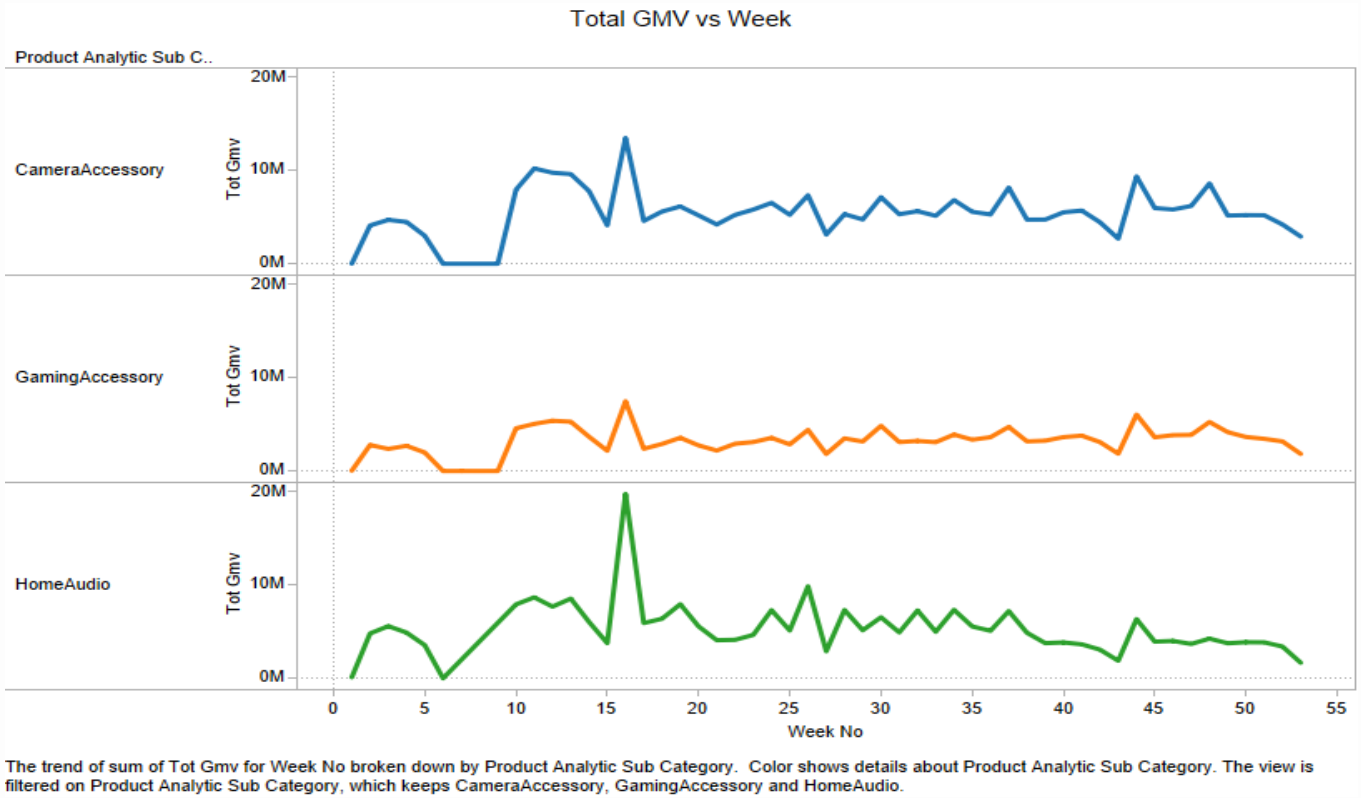
Missing data and data issues:

- Missing values in GMV, customer id and pin code column. There were total 4904 missing values in entire dataset across all product categories.
- Orders having GMV value of 0
- Few records where GMV value was higher than MRP value.
- Negative values in deliverybdays and deliverycdays column.
- Negative values in customer id and pin code columns.
- Negative values in product_procurement_sla column.
- Outlier test is checked for all the relevant variables and outliers are detected.

Data Preparation : Data clean up

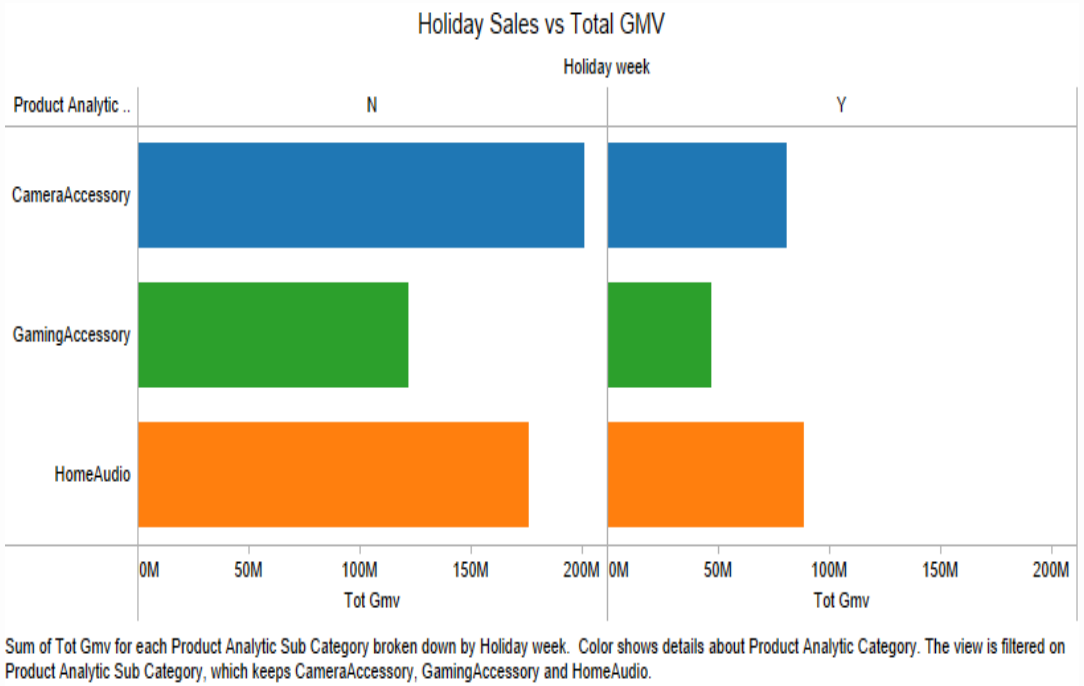
- Rows(orders) with missing values for GMV, Customer ID, PIN Code are deleted (Less than 0.5% of entire dataset)
- Rows(orders) having GMV value greater than MRP, value of MRP is replaced by their GMV price.
- Rows (orders) with 0(MRP) values are deleted, since it is not possible to have product with MRP value as 0.
- Negative deliverybdays and deliverycdays are replaced by 0.
- Negative product_procurement_values are replaced by 0.
- Since there were lots of outliers, they were replaced at appropriate cut off value.
- Daily order level data has been aggregated at weekly level for duration between June 2015 to July 2016 for 3 product sub categories - CameraAccessory, Home audio and GamingAccessory.
- Monthly level ad spend has been converted into weekly ad spend.
- Promotional data has been transformed to weekly level which signifies whether that particular week had any promotions, this is derived from Promotion dates given.
- Monthly level NPS score has been considered for each week of the month.
- All different datasets are merged together to form a single master file for carrying out modelling.
- Filtering has been done to create 3 different datasets for CameraAccessory, Home audio and GamingAccessory.

Data Understanding : EDA Analysis



The trend of sum of Tot Gmv for Week No broken down by Product Analytic Sub Category. Color shows details about Product Analytic Sub Category. The view is filtered on Product Analytic Sub Category, which keeps CameraAccessory, GamingAccessory and HomeAudio.

- We can observe that there is sharp spike during the weeks between 10 to 15, in the events of Dussehra and Diwali.
- There are also subtle spikes during the promotional periods across the year.



Sum of Tot Gmv for each Product Analytic Sub Category broken down by Holiday week. Color shows details about Product Analytic Category. The view is filtered on Product Analytic Sub Category, which keeps CameraAccessory, GamingAccessory and HomeAudio.

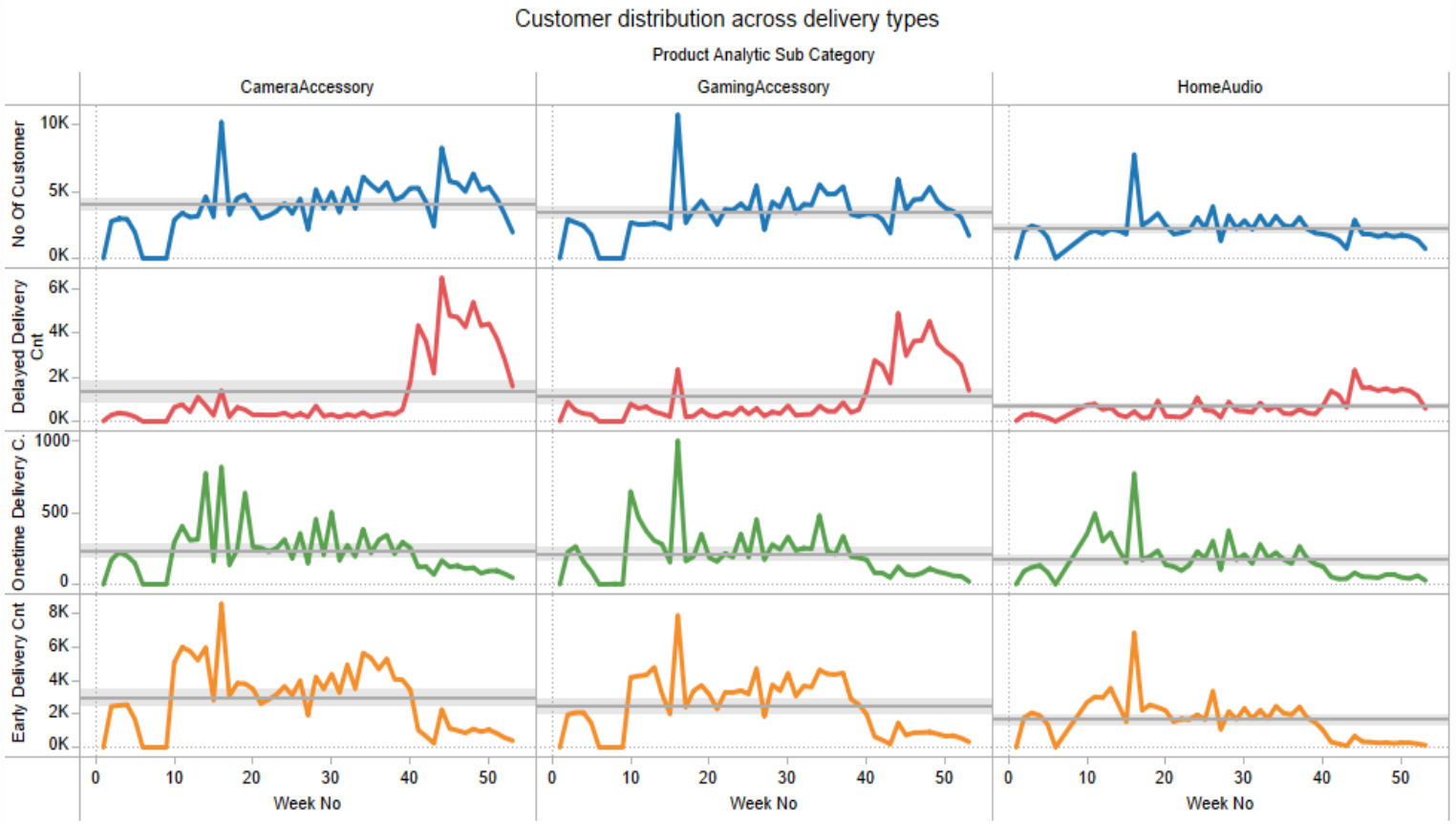
- The ratio of holiday weeks to Non holiday weeks is 1:5.
- Hence, the amount of sales during the Holiday week is far more than the normal week. This suggests holiday weeks are more profitable.

UpGrad

Data Understanding : EDA Analysis

- The number of customers visiting the website goes up especially during Dussehra and Diwali month.
- There have been on time delivery in these periods.
- The deliveries are delayed during the last 3 months (i.e. April to June) , which need to be addressed.
- Week 1 is considered as 1st week of July 2015 and week 53 as last week of June 2016.

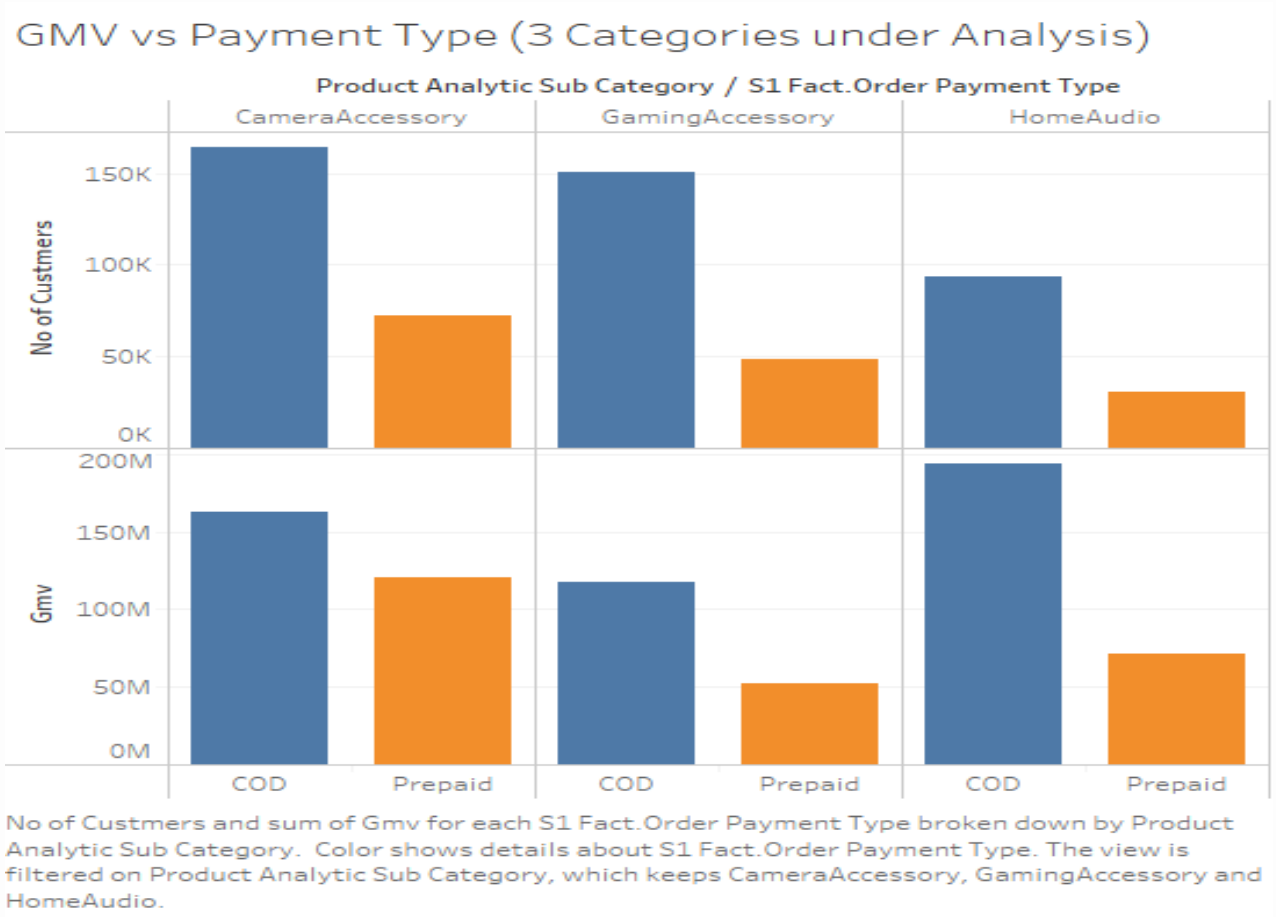
Customer distribution across different delivery types



Data Understanding : EDA Analysis

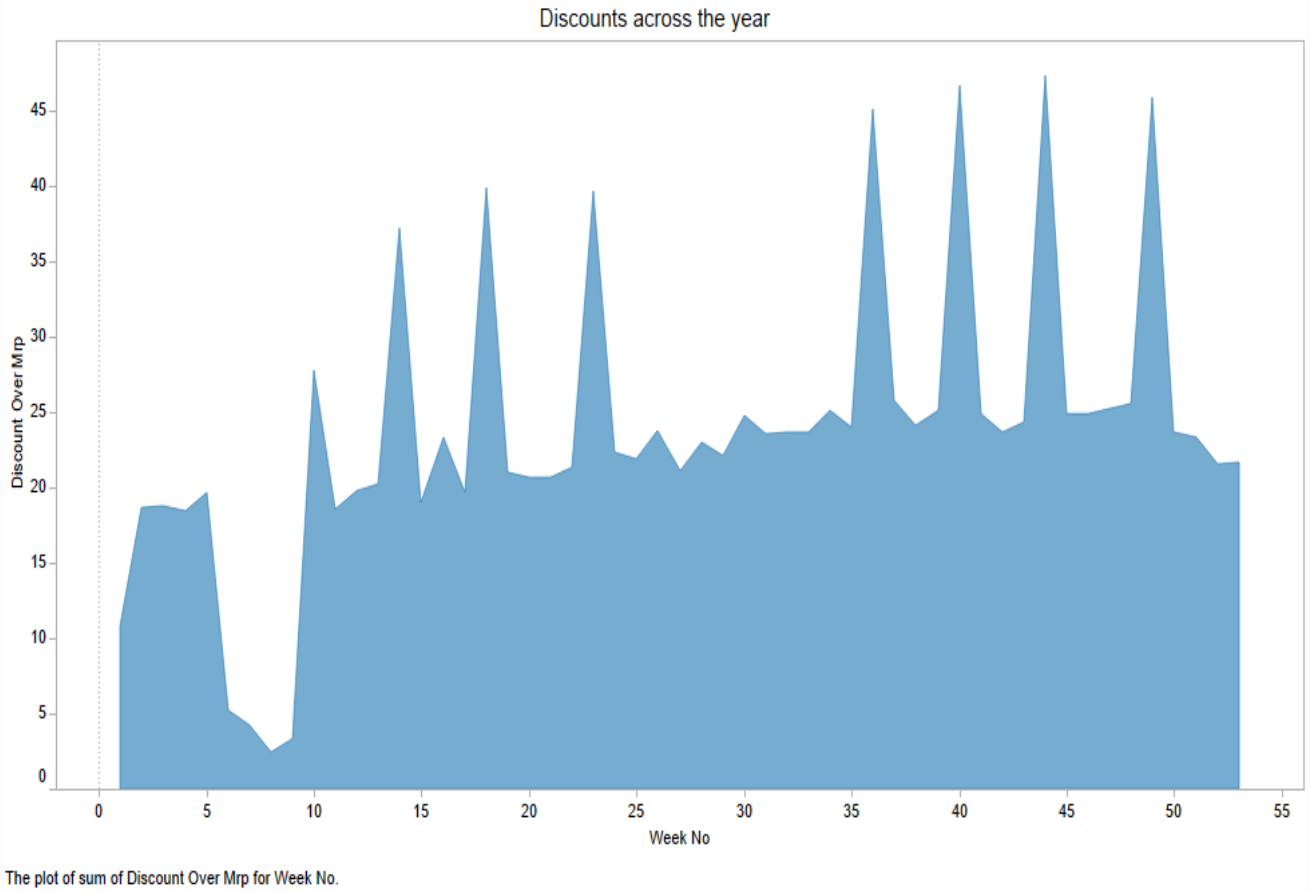
- Across all the categories, the **Cash on Delivery** payment type contributes largely to Revenue.
- The Cash on Delivery mode should be made more convenient to enhance sales.
- Also introspect on the Prepaid mode and plan for better and smooth processes to increase sales from this as it is beneficial to the company from both operational and financial point of view.

GMV/No. of Customers across payment types

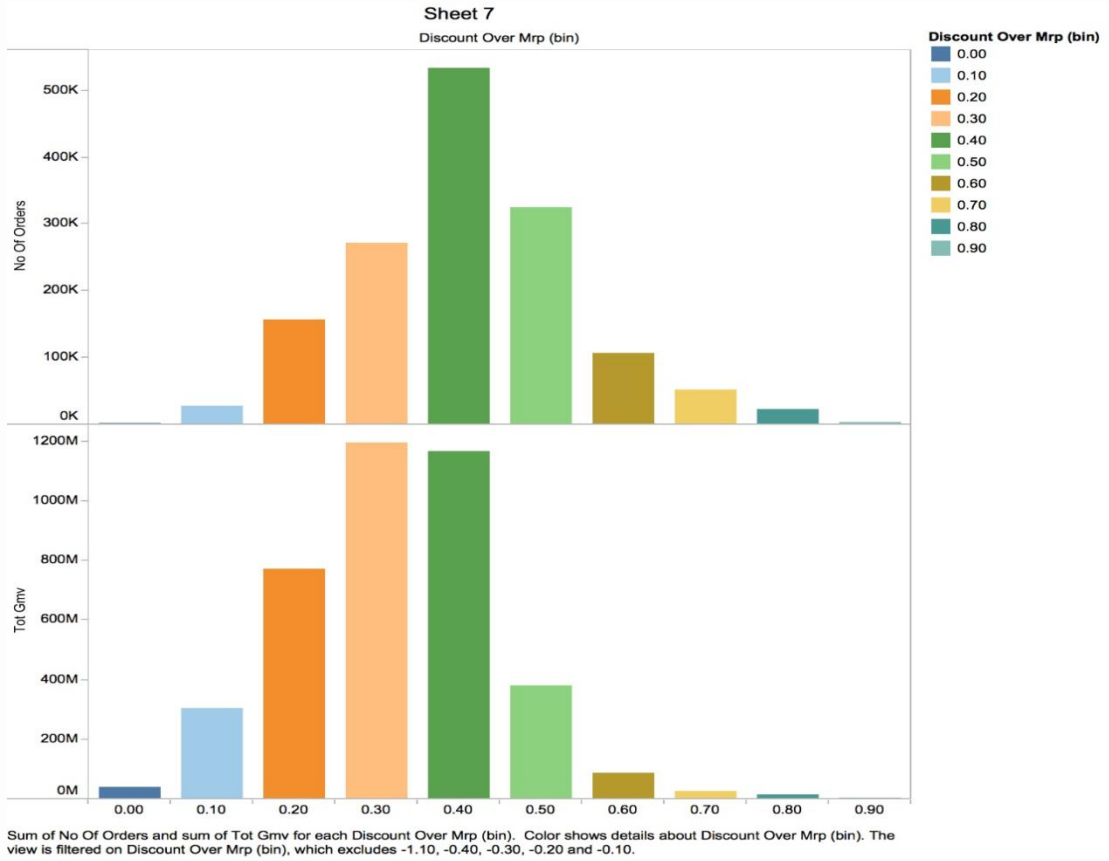


UpGrad

Data Understanding : EDA Analysis



The plot of sum of Discount Over Mrp for Week No.

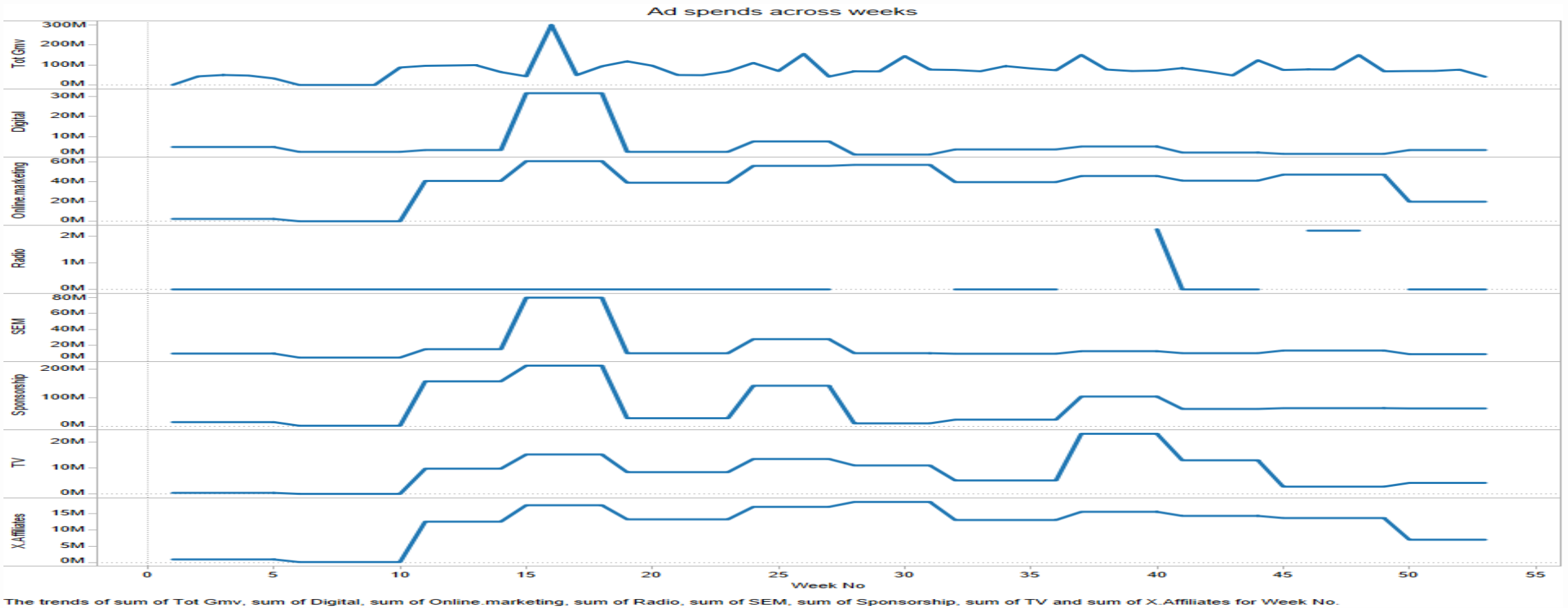


Sum of No Of Orders and sum of Tot Gmv for each Discount Over Mrp (bin). Color shows details about Discount Over Mrp (bin). The view is filtered on Discount Over Mrp (bin), which excludes -1.10, -0.40, -0.30, -0.20 and -0.10.

- Company is providing heavy discount mostly during promotional weeks. But discount over 50% is not very effective.

UpGrad

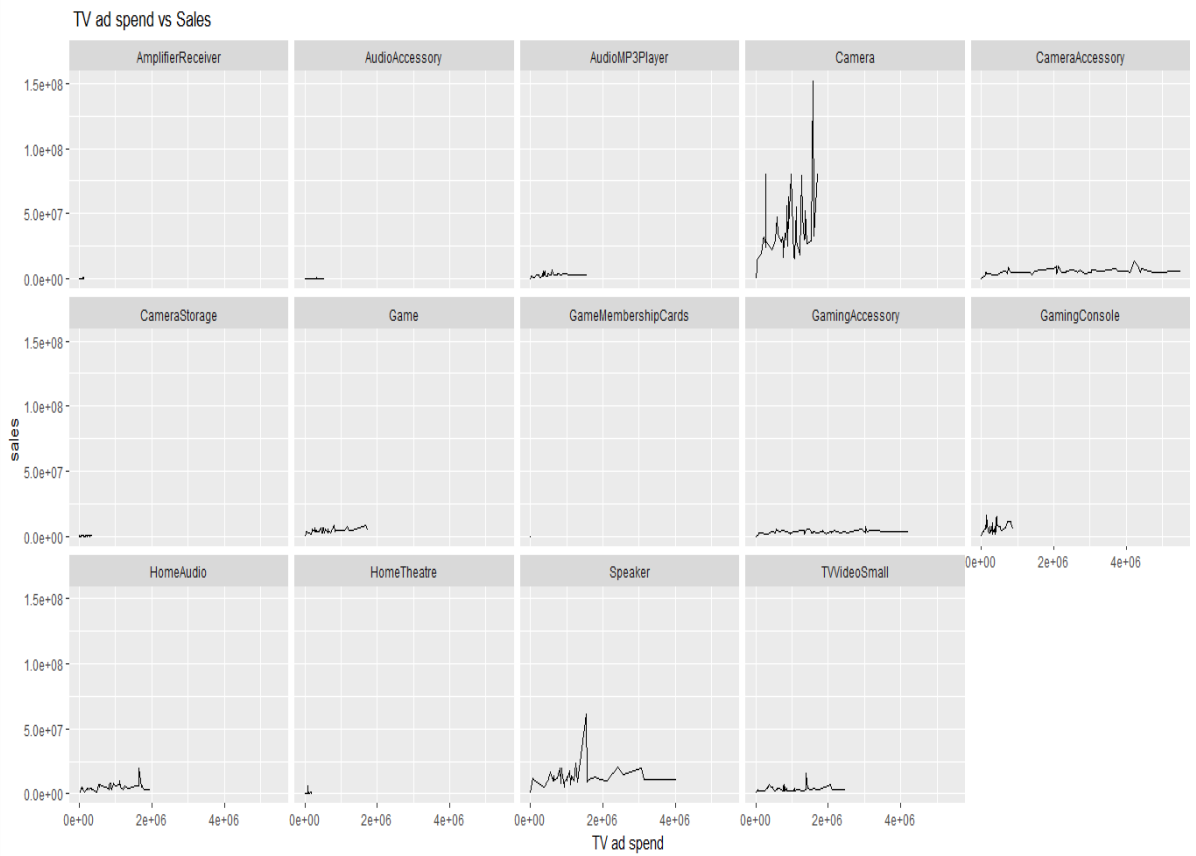
Data Understanding : EDA Analysis

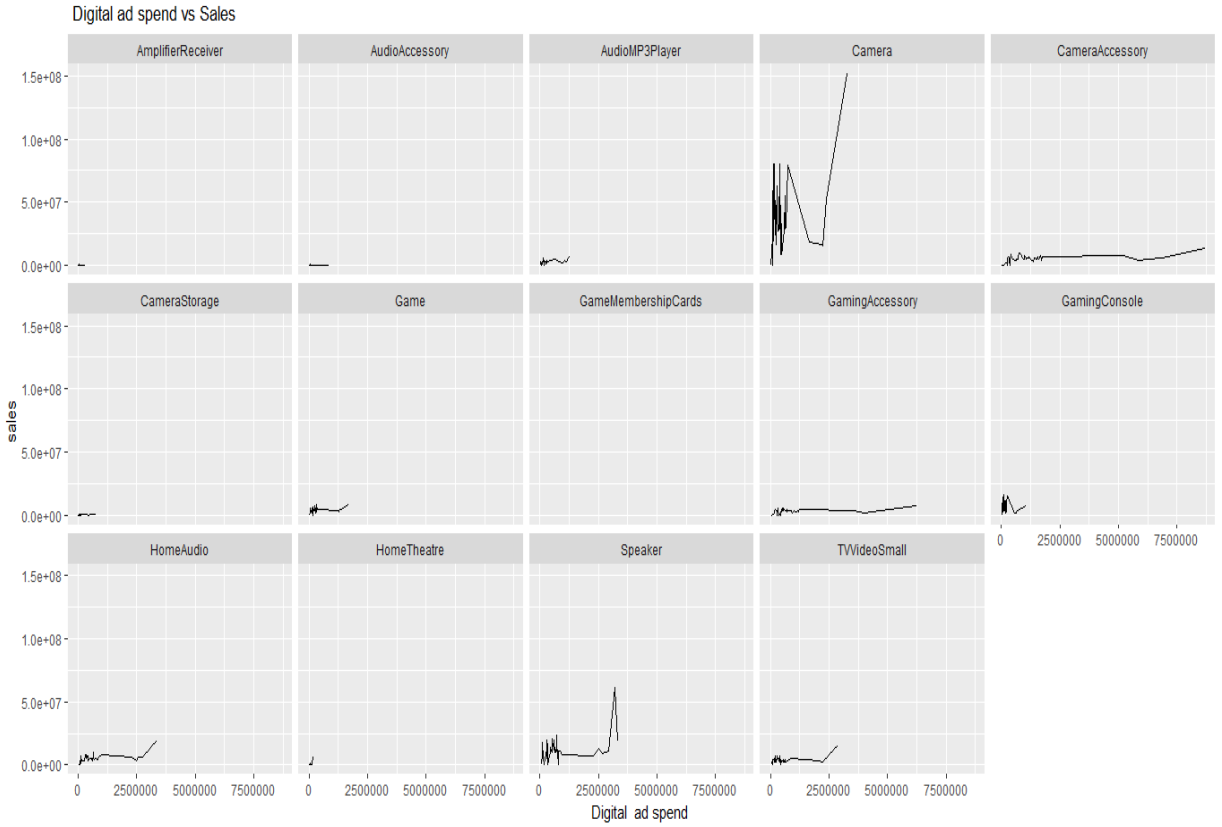


Ad spends are varying through out the year and higher during the promotional events

UpGrad

Data Understanding : EDA Analysis





Mostly there are two types of behaviour with ad spend. For few media categories (e.g. TV and Sponsorship) revenue is getting maximized at specific ad spend. Beyond that, with increasing of ad spend, sale is increasing at much lower rate. For few media categories, initially revenue is increasing with ad-spend, then it is almost constant but after a certain ad spend revenue started increasing at much higher rate (e.g. Digital , Content)

UpGrad

Data Preparation – Derived KPIs

- **Following variables have been derived by the insights from EDA**

1. Average GMV
2. Average MRP
3. List price
4. Lag GMV values for last 3 weeks
5. Lag price values for last 3 weeks
6. Discount over MRP
7. Average number of orders
8. Promotion type
9. Ad stock value
10. Holiday week
11. Value per visit
12. Delivery statuses(Early, Ontime, Delayed)

- Business & Data Understanding
- Data Preparation & EDA
- **Model Building**
- Recommendation
- Challenges Faced

Camera Accessories – Outcome of 5 models

Model	Significant Variables	Adjusted R Square on Test data set	Cross validation(MSE)
Simple Linear Model	Lens + Daussera.sale + Affiliates + Sponsorship	0.84	0.24
Multiplicative Model	NPS + list_price + Camera_Tripod+ Flash + Dussera.sale+ Lens + Camera_Battery + Affiliates	0.79	1.14
Koyck Model	Lens + Dussera.sale + Eid + Rathyatra Sale + Affiliates + Digital	0.73	0.23
Distributed Lag Model	Lens + Dussera Sale+ Affiliates + Discount Over MRP+NPS+ Sponsership	0.80	0.32
Distributed lag + Multiplicative Model	Lens +List_Price + Affiliates_ad_stock + GMC Change from week3 + GMV Change from week2 + GMV Change week1 + Sponsership	0.87	0.67

- The MSE figures are based on the 10 fold cross validation.

Simple linear model is chosen as best model out of 5 models based on high R square and low MSE value.

Game Accessories – Outcome of 5 models

Model	Significant Variables	Adj R Square on Test data set	Cross validation(MSE)
Simple Linear Model	Game Pad + Gaming Headset+ Gaming Mouse + Daussera.sale + Gaming keyboard + Sponsorship	0.54	0.29
Multiplicative Model	TV + List_Price + Game Pad + Gaming Mouse + Gaming Headset	0.69	1.24
Koyck Model	Game Pad + Gaming Headset + Gaming Mouse + Dussera Sale + Gaming Keyboard + BSD5 + Affiliates + Sponsorship	0.67	0.32
Distributed Lag Model	GamePad + GamingHeadset + GamingKeyboard + GamingMouse + Sponsorship + affiliate_ad_stock + Daussera.sale	0.68	0.32
Distributed lag + Multiplicative Model	GamePad + Gaming Headset + Gaming Mouse+ Dussera Sale	0.76	0.91

- The MSE figures are based on the 10 fold cross validation

Distributed Lag model is chosen as best model out of 5 models based on high R square and low MSE value.

Home Audio – Outcome of 5 models

Model	Significant Variables	Adj R Square	Cross validation(MSE)
Simple Linear Model	AudioSpeaker + FMRadio + Eid.RathYatra.sale + Big Diwali sale + Content Marketing + Sponsorship	0.81	0.16
Multiplicative Model	List_Price + Sponsership+FM Radio+Home Audio Speaker+Dussera Sale	0.68	0.86
Koyck Model	Home Audio Speaker+FM Radio+Dussera Sale	0.89	0.41
Distributed Lag Model	Home Audio Speaker + Dussera Sale+Boom Box+ FM radio+Discount over MRP+Affiliate Ad stock	0.74	0.42
Distributed lag + Multiplicative Model	X.Affiliate + gmv_change_w3 + HomeAudioSpeaker + Dock +Boombox + DockingStation	0.78	0.89

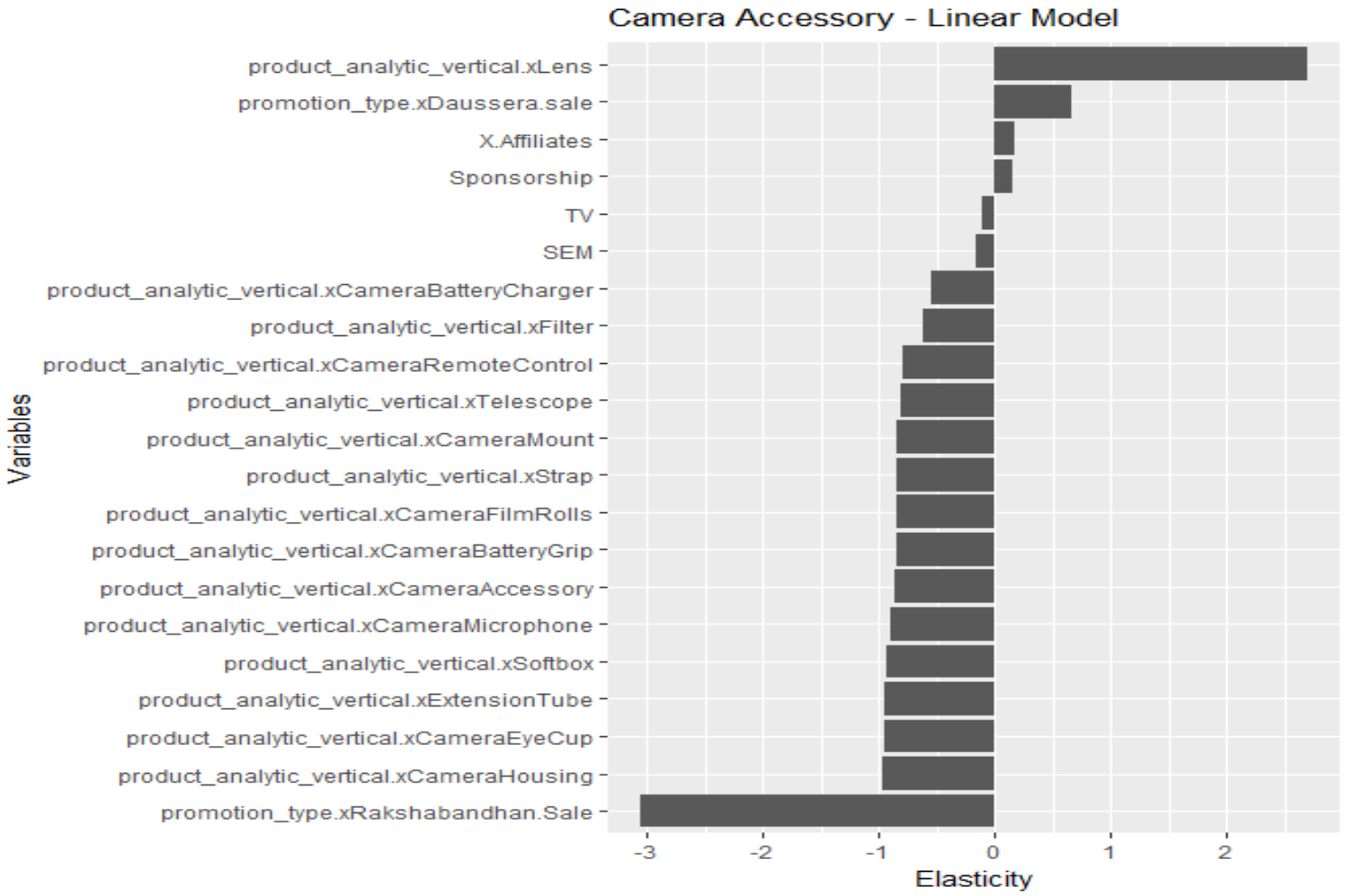
- The MSE figures are based on the 10 fold cross validation

Simple Linear model is chosen as best model out of 5 models based on high R square and low MSE value.

- Business & Data Understanding
- Data Preparation & EDA
- Model Building
- **Recommendation**
- Challenges Faced

Camera Accessories – Recommendations Based on Elasticity of KPIs

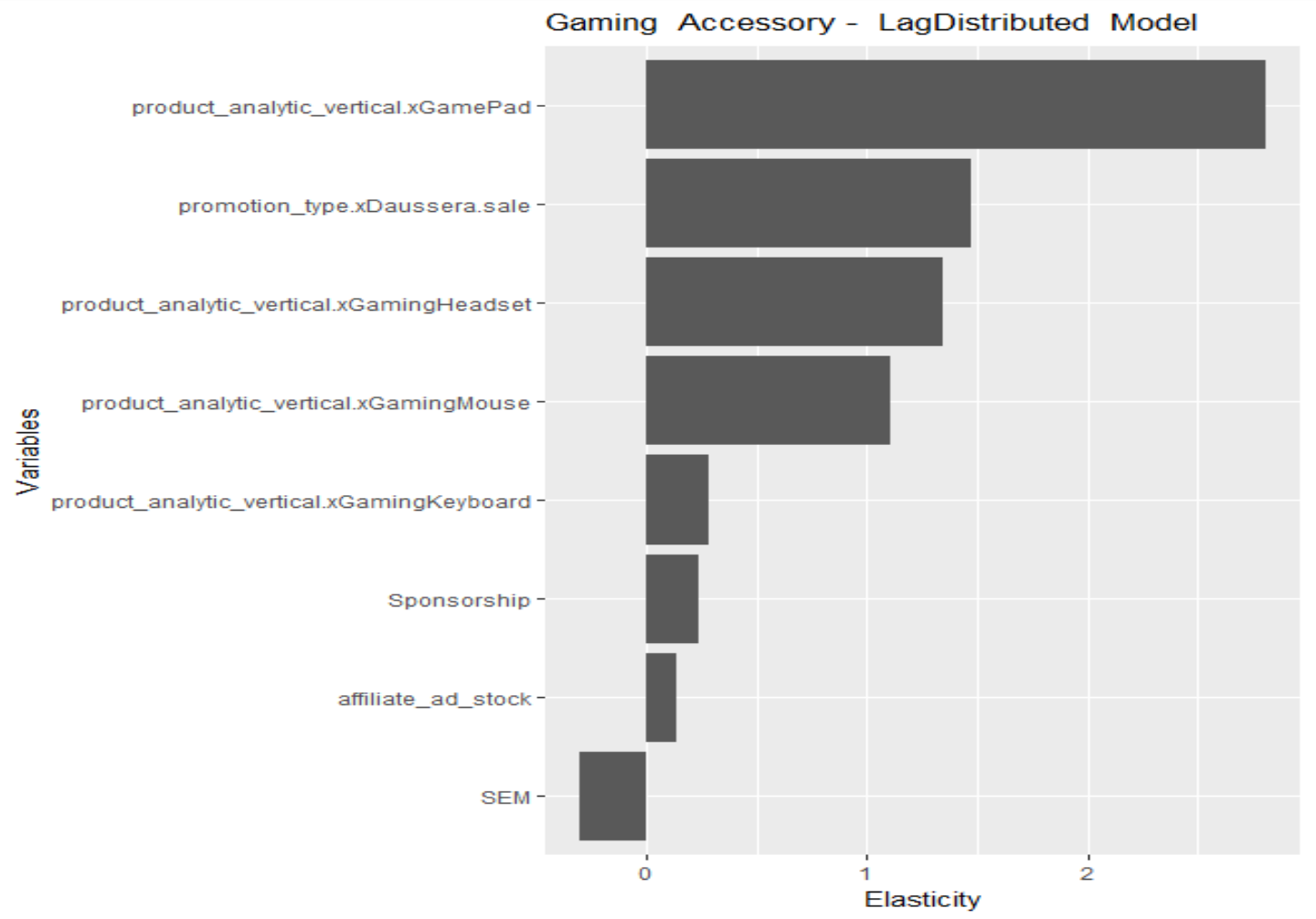
- The adjoining figure represent the elasticity of different variables w.r.t the overall sales figure.
- Positive elasticity means, increasing the value of the KPI would lead to increase in the sales.
- **Affiliate** spend has positive impact on sales. One unit of this ad spend will increase the sale by 0.17 units.
- **Sponsorship** spend has positive impact on sale and one unit of this ad spend will increase sale by 0.16 units.
- **Dussehra** sale has very positive impact. It will increase the sale by .65 unit.
- Company should promote **Lens** product as it has very positive effect on Revenue.



UpGrad

Game Accessories – Recommendations Based on Elasticity of KPIs

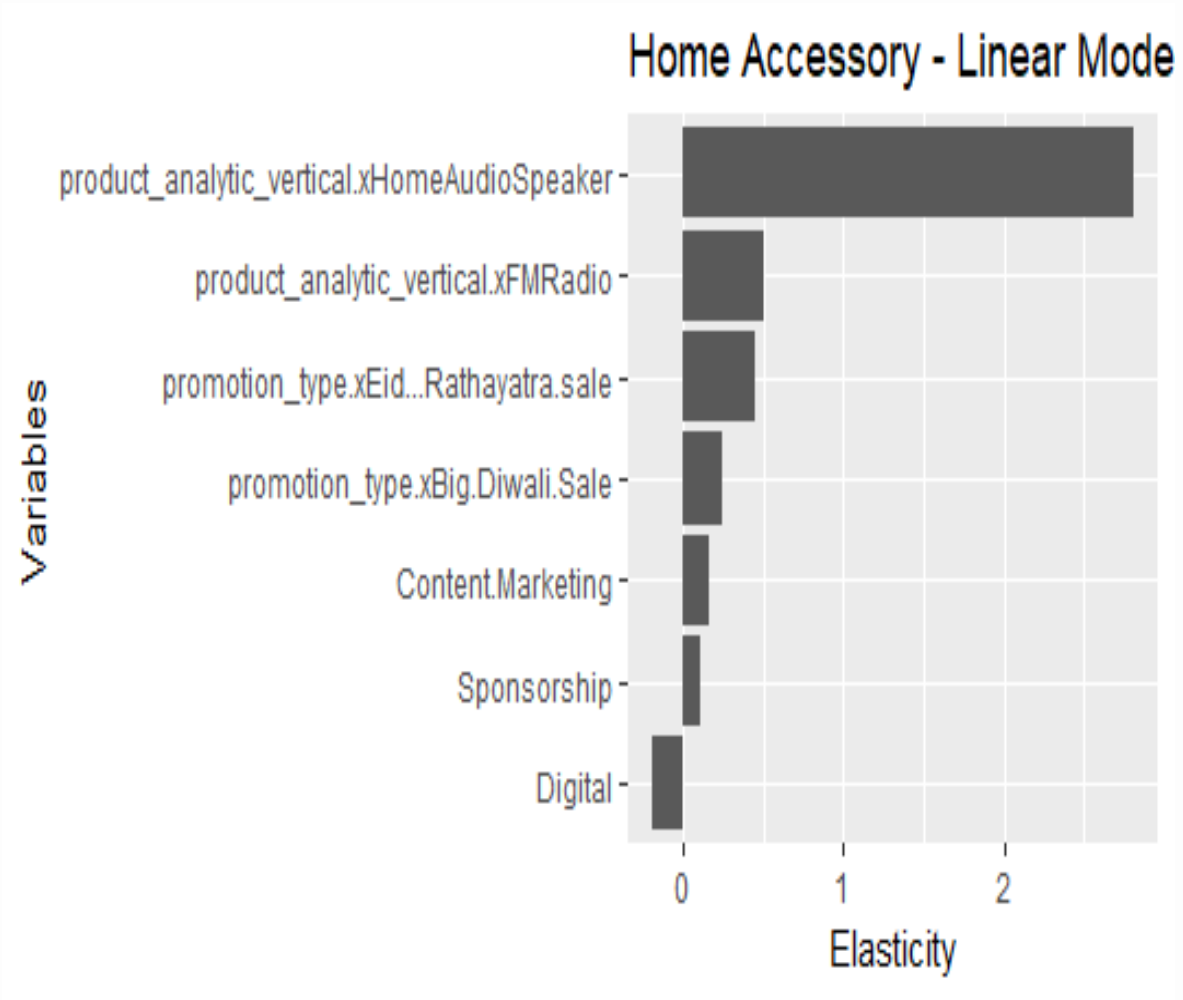
- The adjoining figure represent the elasticity of different variables w.r.t the overall sales.
- Positive elasticity means increasing the value of the KPI would lead to increase in the sales.
- **Affiliate** spend has positive impact on sales. One unit of this ad spend will increase the sale by 0.13 unit.
- **Sponsorship** spend has positive impact on sale and one unit of this ad spend will increase sale by 0.23 unit.
- **Dussehra** sale has very positive impact. It will increase the sale by 1.4 unit.
- Company should promote **Gamepad , Gaming Headset , Gaming Keyboard** and **Gaming Mouse** products as these have positive effect on revenue.



UpGrad

Home Audio – Recommendations Based on Elasticity of KPIs

- The adjoining figure represent the elasticity of different variables w.r.t the overall sales.
- Positive elasticity means increasing the value of the KPI would lead to increase in the sale.
- **Content Marketing** spend has positive impact on sales. One unit of this ad spend will increase the sale by 0.16 unit.
- **Sponsorship** spend has positive impact on sale and one unit of this ad spend will increase sale by 0.10 unit.
- **Diwali** sale has positive impact. It will increase the sale by 0.24 unit. Same for Eid,Rathayatra sale. It will increase the sale by 0.44 unit.
- Company should promote **FM Radio** and **Home audio speaker** as these have positive effect on revenue.



UpGrad

- Business & Data Understanding
- Exploratory Data Analysis
- Model Building
- Recommendation
- **Challenges Faced**

Challenges Faced during Model building

- Deciding on number of derived KPI needed which are important as well as logical.
- Arriving at a base dataset because many iterations were performed as the model results were leading to perfectly fitting model.
- To solve this problem, correlation concept was used to remove correlated variables.
- Selection of significant variables which are finally decided by correlation matrix and non zero variance.
- Identifying the functions/ packages to perform ad stock and creating lag variables.
- Limited availability of the sample implementation of Kyock, Lag + Multiplicative models in R on internet.
- Decision on removal of variables from analysis.