# INTRODUCTION

* **Price elasticity** is a measure economists use to capture consumers’ sensitivity to price changes for a particular good or service.
* The price elasticity is defined as the percentage change in quantity demanded, divided by the percentage change in price. Since the quantity demanded generally decreases when the price increases, this ratio is usually expected to be negative. For example, a price elasticity of -0.6 mean a 1% increase in price leads to a 0.6% decrease in quantity demand
* Goods with elasticities less than one in absolute value are commonly referred to as having inelastic or price insensitive demand – the proportional change in quantity demanded will be less than the proportional change in price. In this situation, increasing the price will increase the revenue of the producer of the good, since the revenue lost by the relatively small decrease in quantity is less than the revenue gained from the higher price
* Goods with elasticities greater than one in absolute value are referred to as having elastic or price sensitive demand - the proportional change in quantity demanded will be greater than the proportional change in price. A price increase will result in a revenue decrease to the producer since the revenue lost from the resulting decrease in quantity sold is more than the revenue gained from the price increase
* **Cross Price Elasticity**

The cross price measures the sensitivity of demand for a particular good to changes in the price of another good and is measured as percentage change in quantity demanded of Good A divided by price of Good B. When the elasticity is positive, the two goods are substitutes (e.g. Coca-Cola and Pepsi); when the elasticity is negative the goods are compliments (e.g. coffee and milk)

# BUSINESS QUESTIONS ANWSERED

* How does my elasticity compare to my competitor’s for the same assortment over different time periods?
* Based on competitive elasticity, what is the likely impact of reducing my price?
* Are there opportunities to increase the price in any of my items?
* What is the price recommendation based on competitive prices and my internal constraints?
* How does my competitors’ elasticity compare to mine for similar products?
* Are my pricing policies consistent with the elasticity of my product?
* How does the pricing policy of my competitor affect my sales?
* How does my competitive price elasticity change with changes in assortment?

# Data Types

# Time Series

# A time series is a sequence of data points measured at successive points in time, and spaced at uniform intervals. It captures data on a fixed population over time. Most data used to calculate elasticity is of this time, as elasticity measures price sensitivity over time

# Panel Data

# Panel data refers to multidimensional data, involving multiple measurements over time. It contains observations on multiple phenomenon, observed over multiple time periods for the same individuals/population. Time series and cross sectional data are special cases of panel data in one dimension. It observes a broad section of subjects over time allowing the study of dynamic as well as cross sectional aspects of a problem.

# Types of Models

# Ordinary Least Squares (OLS)

# OLS regression analysis is a method relating the dependent variable, here sales, and other variables that have an intuitive and measured impact on sales, while minimizing the variance (randomness) of the estimates

# It is a method for estimating the unknown parameters in a linear regression model. This method minimizes the sum of squared vertical distances between the observed responses in the dataset and the responses predicted by the linear approximation

# The resulting estimator can be expressed by a simple formula. The regression analysis allows the relationship between demand and price to be isolated and quantified while controlling for other factors that may impact sales, such as discounts, number of substitutes, seasonality, etc

# When calculating elasticity, the model is most often used as a log-log model, such as:

# Ln(sales)= a + bLn(own price) + cLn(competitor price) + d(dummy var) + .. u,

# Where

# sales is the dependent variable;

# price, competitor price, etc are the independent variables,

# and ‘u’ the error

# Dummy variables take on the form of 1 or 0 in any observation and capture any remaining structural reasons for traffic differences in sales. For example, Season is often accounted for using dummy variable, where it is defined such that it equals 1 if during peak season, and equals 0 if off-season.

# The regression analysis estimates the value of the parameters (constant (a), b, c, etc.) on each of the variables, which reflect the relative impact of each of the variables on sales.

# As log formulations approximate percentage changes in impacts, the parameters of the logged independent variables can be directly interpreted as elasticity

# INPUT VARIABLES

|  |  |  |
| --- | --- | --- |
| **Price Variables** | | |
| Own list price | Price listed on retailer's site | **-** |
| Own discount percent | % discount offered | **+** |
| Own final price after discount | Final price after discount | **-** |
| Own Shipping price | Shipping price mentioned | **-** |
| Competitor list price | Competitor's list price | **+** |
| Competitor discount percent | Discount given by competitor | **-** |
| Competitor final price after discount | Competitor's final price | **+** |
| Competitor shipping price | Competitor's shipping price | **+** |
| Number of variants | Total number of variants available | **-** |
| Number of substitutes within own retailer | List of buying options customer has for the given SKU within own retailer | **-** |
| Number of substitutes across competitors | List of buying options customer has for the given SKU across retailer | **-** |
| Length of copy for product description | Total number of words in main copy | **+** |
| Availability | Explains whether product is in stock or out of stock | **+** |
| Price difference across retailers(relative price) | Price difference across retailers for same product. For example, price of product p1 is $12 on client’s site and $15 at competitor’s site then price difference variable will have value as -3 | **-** |
| Price ratio between retailer | Ratio calculated by considering own price(numerator) and competitor price(denominator) |  |
| Recency in week | Determines how recent the product is in terms of week. It is the difference between first date at which product was available at PiceTrac and today’s date. The difference is calculated in weeks | **+** |
| Recency in month | Similar difference as above is calculated in months | **+** |
| **Review Variables** | | |
| **Note –**   * **Review and image data is aggregated on main product id level** * **Same set of variables(review and image) have been created for competitors in which case they will have opposite impact on sales(for example, reviews for competitor will have negative impact on client’s sales)** | | |
| Review count | Total number of reviews present | **+** |
| Average rating | Average rating mentioned on the site | **+** |
| Average helpful votes(v1) | Average of helpful votes mentioned on the site | **+** |
| Average helpful votes(v2) | Total review helpful votes divided by review total votes | **+** |
| Rating1 | Count of "only 1 Star" rating for the product | **+** |
| Rating2 | Count of "only 2 Stars" rating for the product | **+** |
| Rating3 | Count of "only 3 Stars" rating for the product | **+** |
| Rating4 | Count of "only 4 Stars" rating for the product | **+** |
| Rating5 | Count of "only 5 Stars" rating for the product | **+** |
| Rating4 and rating5 | Variable is created by adding rating4 and rating5. Additional variable is created by dividing given summation with reviews count | **+** |
| Rating1 and rating2 | Variables are created using similar approach as rating4 and rating5 | **+** |
| Rating and review count | Individual ratings are divided by reviews count | **+** |
| Average sentiment score | Sentiment score is calculated using text review present on the site. The review was broken into words and each word was compared against positive and negative list of keywords. And thus we derive at the score | **+** |
| Total sentiment score | summation of sentiment score | **+** |
| Average positive sentiment score | Average score of only positive reviews | **+** |
| Total positive sentiment score | Addition of score for only positive reviews | **+** |
| Average negative sentiment score | Average score of only negative reviews | **-** |
| Total negative sentiment score | Summation of score for only negative reviews | **-** |
| Average weighted score | Average helpful votes(v2) multiplied by average sentiment score | **+** |
| Total weighted score | Average helpful votes(v2 )multiplied by total sentiment score | **+** |
| Total reviews in last month | Number of reviews between first day of the week and month back | **+** |
| Total reviews in last two months | Number of reviews between first day of the week and two months back | **+** |
| Total reviews in last three months | Number of reviews between first day of the week and three months back | **+** |
| Total reviews in last six months | Number of reviews between first day of the week and six months back | **+** |
| Saving dollar | How much customer saves in case of promotions or markdown | **+** |
| Promo flag | Dummy variable for promotions. It keeps into consideration free super saving shipping , free returns | **+** |
| **Image Variables** | | |
| Number of images | Total number of images present on the retailer's site | **+** |
| User generated images | Out of the available images, how many are uploaded by user | **+** |
| Not user generated images | Out of the available images, how many are not uploaded by user | **+** |
| **Additional variables** | | |
| Seasonality | Specific time period when there is sudden change in sales |  |

# DATA PREPRATION

# STAGE 1 – Key variables

## PriceTrac Data

1. PriceTrac data is available on weekly basis. Identify the time period of analysis from sales data. For given time period merge PriceTrac weekly crawled files across retailer for given category
2. Calculate discount percent and week number as follows –
   1. Discount percent - (regular price-final price)/regular price
   2. Week number - ( extraction date – first date from where analysis starts)/7)+1
3. Split the merged file per retailer. There has to be one client identification number per product id for given week. Hence take first product id if there are multiple product ids present for a given week
4. Prepare the data in such a way that, competitor information (product name, price) is populated in columns for own retailer. Own retailer’s product name, product id, prices, client identification number and week must be unique in each row
5. Calculate following additional variables –
   1. Price difference - price difference between client and each and every competitor. Hence if there are 3 competitors then we will have 3 variables
   2. Product description length - how long the product description is. Here we identify total length of the product description
   3. Availability flag - 1 if availability is “Yes” otherwise 0
   4. Promo flag - If additional info column contains "Free Super Saving Shipping & Free Returns" then 1 else 0
   5. Saving dollar – Dollars that customer saves. This can be identified by searching for $ in the Price promo column and extract the number after $
   6. Recency – First identify first and latest extraction date for a particular product id and then determine the difference between today’s date and first date. Find out recency in terms of week and month
   7. No of variants **–** Take the count of unique client identification id per product id and week
6. Append all these new variables to the dataset calculated in step 4

## Sales Data

1. Import the sales data on SAS
2. Determine gross sales by subtracting return units from net sales
3. Aggregate sales data on product id level
4. Now transpose the data in such a way that, sales quantity will be a single variable(instead of variable for each week) for unique combination of week number and product id
5. Extract week number from variable name(for example for Week22\_gross\_sales, week number will be 22)
6. Repeat 4 and 5 for the required variables and merge all of them

## Content Data

1. Content data has 3 sections –
   1. Product file - Basic information about product
   2. Review file - Contains reviews for a particular product. Product id will be repeating if there are multiple review for a given product
   3. Image file - URL of the image, also contains information about whether the image is uploaded by the user. In this case, there can be multiple instances of the product id depending on the number of images product has
2. Following variables are additionally calculated for modeling from Review file
   1. Calculate week number using review creation date. This will be used when we want to map review data to PriceTrac
   2. Using the week number, identify first and last day of each week. And then compute one, two, three and six months date
   3. When we are aggregating review variables , we need to consider all the reviews/ratings those are present before each week ending. While calculating velocity variables, we will be considering duration which is between week ending and specific month prior
   4. Sentiment score
      1. Sentiment analysis using reviews helps to evaluate how the product been perceived, whether the reviews are positive or negative
      2. Each word in the review is compared against list of positive and negative key words. Sum of positive keywords – Sum of negative keywords will give final sentiment score
      3. Thus the score can range from most negative to most positive number
      4. This score is generated in R and output file is imported in SAS
      5. Data is then rolled up on main product id level per week
      6. Few variations such as average and total sentiment score, average positive score, average negative score is calculated
   5. Helpful votes
      1. Week number and main product id wise average of Review helpful votes variable
      2. Another variation is dividing review helpful votes by review total votes
   6. Average weighted score
      1. Multiply average sentiment score with second variation of helpful votes
   7. Velocity of the reviews
      1. It includes preparing variables such as number of reviews in last month from week ending(Sunday) of the week for particular product id. Thus we will be calculating dynamically number of reviews per week for given product id
      2. Similarly various other variables are prepared such as
         1. Number of reviews in last two months
            1. Number of reviews between two months back from week ending day of the week
         2. Number of reviews in last three months
            1. Number of reviews between three months back from week ending of the week
         3. Number of reviews in last six months
            1. Number of reviews between six months back from week ending
   8. Rating variables
      1. Creating rating variables using rating score such as rating1, rating2 and so on
      2. Rating count on product id level on weekly basis
      3. Average rating on product id level on weekly basis
3. Below variables are created using image data –
   1. Number of images - Main product wise count of url
   2. Not user generated image - If UGC = “No” then 1
   3. User generated image -Number of images - Not user generated image
4. Finally merge all three files on main product id level for week number > 0

# STAGE 2 – Merging of files

## Merging PriceTrac and sales data

1. PriceTrac and sales data can be merged using week number and product id

## Merging PriceTrac and content data

1. Client’s content data
   1. In order to map own content data, product id from PriceTrac and main product id from content can be used
2. Competitor content data
   1. Competitor’s product id and product id from PriceTrac

# STAGE 3 – Preparing additional variables

***Note – Following variables will be prepared only after merging the datasets***

## Merging PriceTrac and sales data

* After merging above two files, extract unique list of product names
* If required, list of product names should be divided into segments(for example, briefcase, carry-on, backpacks in luggage category)
* Identify key features from the product description(for example, laptop facility, wheeled, expandable etc)
* Prepare separate column in excel for each feature
* If the feature is available in the product name then mark 1 else 0
* Import the file in SAS

1. Number of substitute within own retailer
   1. Take a count of product names having similar feature
   2. If a product is having more than one feature(for example, briefcase which is wheeled and having laptop facility) then products which are matching with any one of the feature should be counted
   3. From the total count subtract one as total count is accounting for own name as well
   4. For example, consider product Rockland Giraffe Wheeled Duffel Bag. Its features are specially designed and wheeled facility. That is, it has value 1 for specially designed and wheeled column. Now go to next row and check whether either of specially designed and wheeled columns have 1 for that particular product name. If yes then that product will be counted as substitute. Continue this process till you reach last product name
2. Number of substitute across competitor
   1. Consider only those product for which final price is greater than zero
   2. Now for these product names, repeat same steps creating number of substitute for own retailer

# STAGE 4 – Preparing data for model

1. Drop unnecessary variables from the datasets
2. Take log of sales and prices
3. Data is now ready to be used for modeling

# MODEL BUILDING

1. Once data for modeling is prepared, we need to split the data in 70-30 format as follows –

* Model build on 70% population
* 30% kept as holdout for model validation

1. Model building is an iterative process. We need to check for following errors and correct them in order to get the desired model

|  |  |  |
| --- | --- | --- |
| **Error** | **Definition** | **How it skews the Model** |
| Multicollinearity | Occurs when 2 or more explanatory variables are highly correlated | Estimates of the variable impact of individual regressors tend to be inaccurate |
| Heteroscedasticity | Unequal variance of the error terms | Biased standard error, leading to wrong inference of the model |
| Autocorrelation | Correlation between errors in time series data | Biased Standard error |

1. Once the model is built for 70% of the data, run the same model on 30% of the data. If both the models give similar visualizations/trends then model is accurate

# To do for Model Building

1. Sales vs Price; log(Sales) vs log(prices)
2. Check for multi-collinearity : Correlation among all variables
3. Keep on iterating by removing variables having VIF > 5 with refer to correlation values.
4. Repeat till variables having VIF < 5.
5. List those significant variables for model building.
6. Sales vs Reviews
7. Principal component analysis for variable selection: PCA
8. Methods choosing significant variables: Chose PC having cumulative proportion at-least by 90% (or/and)
9. Calculate differences between eigen values wherein consider variables till there is sharp drop from one eigen value to other.
10. Choose variables having high correlation on either direction which is actually explaining larger variations among variables.
11. Sales vs Price+Review+Images : Model building