Price Impact Estimation using Generalized Linear Models

Ankur Verma
WalmartLabs
Bangalore, Karnataka
ankur.verma@walmart.com

ABSTRACT

Price Impact Estimator developed by Walmart Labs pricing team provides visibility to the merchants on expected unit volume change of an item (UPC) as well as other items (affined/substitute) due to cross-price effects in response to a price change. Customer transactions sales has been aggregated at item-store-week and then item-week-national level. Forecasting model has been built using regularized regression models to predict the sales units using price, seasonality and other features. Elasticity of an item has been estimated using simulation-based approach to quantify the effect of price alone on the unit sales. Accuracy of approximately 84models built at three different levels of merchandise hierarchy.

KEYWORDS

Machine Learning, Price Elasticity, Simulation

ACM Reference Format:
Ankur Verma. 2021. Price Impact Estimation using Generalized Linear Models. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), London, UK, May 3–7, 2021*, IFAAMAS, 1 page.

1 INTRODUCTION

Pricing is at the core of Walmart shopping experience. An important part of maintaining customer trust in Walmart is to provide customers with consistent prices on items across the box in line with EDLP (Everyday low price) principles.

Merchants are faced with the task of making pricing decisions on a daily basis. One of the major challenges faced by merchants is to identify impact of price changes on unit sales of an item and predicting corresponding demand owing to different pricing scenarios. Currently, merchants do not have a consistent system or process for evaluating the impact of pricing decisions. Merchants often rely on input from suppliers and/or institutional knowledge. To tackle this challenge, machine learning techniques has been used to design a solution to provide visibility to the expected unit volume change of an item and items affected by cross-price elasticity in response to a price change. The solution has been proposed to facilitate data driven decision making on pricing to achieve desired business objectives.

Accuracy of demand equation for each UPC achieved by machine learning techniques has been evaluated using out of time testing with MAPE (mean average percentage error) as an evaluation metric.

Rest of the paper has been organized as follows: Section II describes the data used, formulation of the demand equation and simulation approach for calculating elasticity estimates. Section III includes explanation of features for modelling exercise and modelling techniques. Section IV and V describes model results and evaluation and Section VI provides salient conclusion of the work.

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, London, UK.

2 METHODOLOGY

2.1 Data

Walmart US Stores POS (point of sales) / transaction data exists at the most granular level of sales data of an item at a single transaction. Methodology leverages transaction data which is transformed to a UPC – Week level. Filtering and processing of data includes exclusion of sales occurring under the bracket of clearance and markdown. Rationale for this is to exclude data points having sales at prices which are far off from the base price of a given UPC since the model building for impact of price is accurate when changed in some delta range of the base price which comprises major proportion of the training data set. Unit sales are rolled up at a week level for a given UPC. Price point occurring in majority of stores is assigned as the price of a given UPC for that week. One price point is identified at each week since Walmart pricing landscape is moving towards national level pricing which implies existence of a single price across all store at a given point of time. The solution is constructed keeping in view national pricing landscape.

Cross price effects are an important component of the demand equation to establish causal relationship between unit sales of parent item and price of the halo and substitute UPCs. Top three substitute UPCs within the category of a parent item are identified and the price points at week level is integrated with the base transaction data of the parent UPC at the week level. Top three affined UPCs of a parent UPC outside the category and within the department of parent UPC are identified and the price points at week level is integrated with the base transaction data of the parent UPC at the week level.

As part of base transaction data, month flags are introduced as binary variables indicating whether it belongs to one of the twelve months of the year or not. National holidays are also incorporated as binary flags for the following events - Mother's Day, Thanksgiving Day, Valentine's day, Ash Wednesday, Father's Day, St. Patrick's Day, Good Friday, Independence Day, Easter, New year, Kwanzaa, Labor Day, Memorial Day, Christmas, Halloween, Cyber Monday.

Promotion indicator is created at a week UPC level as a continuous variable which is calculated as percentage of stores on promotion including rollbacks, store tab and competition tab.

2.2 Demand Equation

First exploratory exercise undertaken was to build most granular level models at the UPC level which leveraged data for each UPC

for constructing a UPC level demand equation. However, in Walmart universe there is a less proportion of UPCs with significant historical price changes. Due to this constraint, only a few UPCs out of the extensive Walmart universe had a model with price as a significant feature. To tackle this issue, multiproduct approach was tested and implemented. Multiproduct approach is proposed for building models which specifies the demand equation having unit sales per store of a UPC as a target variable and explanatory variables mentioned in section III(A). Multiproduct approach refers to building line, fine-line and category level models where data for all UPCs belonging to a line, fine-line and category respectively are clubbed together for modelling. This approach was identified and proposed to overcome challenges of limited data points for UPCs with less transaction history, to identify potential causal price impacts for UPCs which don't have historical price changes and potentially learn from the behavior of UPCs falling within the group (line, fine-line and category) of the UPC in consideration. Modelling techniques used is specified in Section III (B).

2.3 Elasticity Estimates using Simulation

Self-Price elasticity estimate for a UPC is one of the crucial components that defines the characteristic of UPC and enables in computing sales lift and therefore to identify potential UPCs for investment.

Elasticity estimate is arrived at using a simulation procedure. For each UPC, base price is identified and 6 price variants are generated from the base price. They include +5scored at these 7 price points using the multiproduct model's demand equation as mentioned in section II(B). Percentage change in unit sales is calculated between successive price points and is divided by percentage price change to arrive at an elasticity estimate. Average of elasticity estimates at simulated price points is used to arrive at an elasticity figure for a given UPC.

2.4 Model Explanatory Variables

Integrated data described in the earlier sections consists of a set of variables such as self-price, price of halo UPCs, substitute prices, national events flags, monthly flags, and promotion indicator. Different transformations of these variables are used as features. Features used for building line and fine-line level models are stated in Table I.

Most of the features are expressed as relative variables since UPCs with varying magnitude of self, halo and substitute prices are used for model building. This has been incorporated to ensure standardization of variables. An offset is also introduced into the model formulation which has been discussed in detail in section III (B).

Several interactions are also incorporated as a part of the feature set such as interaction of promotion indicator with relative self-price to identify any statistically significant additional impact of price change if it co-occurs with a promotion activity. Binary flags for certain UPCs which contribute to 90group has also been incorporated to identify any additional impact if the UPC in consideration belongs to higher sales contribution UPCs. These high contribution UPC features are also interacted with relative self-price to identify any statistically significant additional impact of price change for a high sales contribution UPC.

Features for category level models have been stated in Table II. Certain week and year related variables have been introduced to capture seasonality effects. Week of transaction for a UPC has been introduced to capture the effects accruing to maturity of transaction cycle for a UPC.

Mean volume of UPC has been introduced for the purpose of target encoding aimed at providing a baseline to the target variable. Volume bucket has been introduced as an interaction between mean price and mean volume of a UPC. Holiday flag has been incorporated to capture any additional impact due to existence of a national holiday period. Promotion indicator has been introduced to capture the effects on unit sales due to varying promotion intensity. Price change percentage variable has been introduced for capturing the impact on target variable due to deviation in price from the base price of a UPC.

2.5 Modelling Approach

Multiproduct models at a line and fine line level have been built using regularized elastic net regression technique with unit sales per store at a weekly being the target variables and the list of features stated in Table I. Regularized regression technique has been used to retain the explainable nature of solution, avoid overfitting and to identify the potential causal impact of changes in the features. Offset for the target variable has been used as rolling average unit sales per store of a UPC. The family of distribution for dependent variable has been set as "Poisson" to enable multiplicative structure and specification of the demand equation. The time period for rolling average has been fixed at 4,13 and 52 weeks and based on the availability of data for maximum lag period, offset has been assigned for each UPC – week level record. R package "glmnet" [1] has been used for this computation.

All self-price, halo and substitute features are constrained to be non-positive. The rationale behind this is to ensure an inverse relationship between self-price and unit sales. Rationale for constraining coefficient of relative substitute price feature is to ensure a direct relationship between substitute prices and unit sales. Similarly halo relative price is constrained to have a negative coefficient to establish an inverse relationship between price of an affined UPC and unit sales of parent UPC. Promotion indicator is constrained to have a non-negative coefficient to establish a direct relationship between unit sales and intensity of promotion activity.

Modelling data is divided into two parts comprising of approximately 104 weeks of data for each UPC as the training set and around 13 weeks of data for each UPC as testing data for out of time testing using MAPE (mean absolute percent error) as an error metric. Model coefficients are arrived at using 10-fold cross validation on the training set.

Category level models have been built using linear boosting techniques [2] based on parallel coordinate descent. Boosting technique has been implemented to leverage larger pool of available data points. Target variable is unit sales per store of a UPC. The set of features have been mentioned in Table II and rationale for incorporating the features has been explained in detail in section III(B). The model is specified to have an additive structure. Hyperparameter tuning has been implemented to arrive at a learning rate of 0.5. Number of iterations has been specified as 1500. MAPE

(mean average percentage error) has been used as the evaluation metric for training the model.

3 EXPERIMENTS AND RESULTS

3.1 Loss Function

The models have been judged on various criteria for their Mean Absolute Percent Error. As discussed in section III(B), 13 weeks of data for each UPC has been kept as test data. We calculate MAPE for each of these UPCs by doing prediction on test data using the trained model. It is an indicator of how accurate the models are. A model has been considered a legitimate model only if the MAPE is less than 30

3.2 Model Stability

The products available in Walmart have dynamic nature, their nature and characteristics keep on changing over time. In this regard, it becomes very important to ensure that models which have been built are robust and successfully capture the dynamic nature of the products.

To do this, a model for an UPC has been trained on datasets having same features but are of different time periods. This has been done on five different datasets for an UPC and elasticity values have been generated using the simulation technique as described in section II(C). The average change in coefficient values of price variables as well as elasticity values for all the UPCs was less than 10getting overfit on the train data.

3.3 Volume Lift Accuracy

As discussed in previous sections, most of the products in the Walmart eco-system do not go through significant number of price changes, in order to estimate the performance of the models during a price change, volume lift accuracy has been calculated.

This has been done by considering all those weeks where price change has occurred in train and test data. In all those weeks, next eight weeks of actual as well as predicted volume have been taken and MAPE, discussed above, has been calculated. The rationale behind taking next eight weeks is that the effect of a price change such as jump in demand has always been observed at least a week after the price change. One of the reasons behind this lag is customers take time to internalize the price change and react to it.

3.4 Discussion

Walmart landscape.

Table III discusses the overall coverage of our models for 40 odd categories available in Walmart eco-system. To get coverage of a model, we only consider UPCs which have models with MAPE less than 30only the UPCs which have elasticity in range [-4,0) have been considered. The baseline for the coverage calculation includes only those items which have been sold in more than 2000 stores in last 1 year to include only national items not regional items.

Total number of UPCs column in Table III gives the baseline for the categories and other columns depict the coverage from three different models. The last column is the overall coverage which has been found by combining the results of three different models.

Consider the category number 1867 in Table III, this category has 729 UPCs which have been sold in more than 2000 stores in 1 year, out of which ,561 UPCs have been successfully captured by these three models i.e. 561 UPCs are more than 70accurate and have elasticity in [-4,0). Similarly, category 1483 has exceptionally good coverage of 97is not satisfactory for all the categories for e.g., category 8539(Television) has overall coverage quite low. One of the reasons might be highly elastic nature of electronic appliances. This shows that other approaches also need to be tried out to capture these nuances.

Table IV summarizes the coverage from three models. Overall, the line, fineline and category models are capturing 70of national items for these 40 odd categories.

4 CONCLUSION

Walmart eco-system goes through very small number of price changes, building a machine learning model on these few price changes is a challenging task. The multiproduct approach, regularized regression technique and linear boosting technique have been proved to be quite successful to get demand equations as well as elasticity values of products sold in Walmart. However, new modelling techniques need to be developed to capture dynamic nature of the products. An approach such as Bayesian Structural Time Series could be a good start.