National College of Ireland

National College of Ireland
Project Submission Sheet – 2020/2021

Student Name:	KARAN VEER SINGH, LAIBA REHMAN, PRIYANKA --
Student ID:	x20146248, x20144032, x20192037
Programme:	MSc in Data Analytics 2021
Year:	
Module:	Domain Application of Predictive Analytics
Lecturer:	VIKAS SAHNI
Submission Due Date:	20/08/2021
Project Title:	Sales Forecasting for Rossmann Stores Using Machine Learning.

Word Count: 2688

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: KARAN VEER SINGH, LAIBA REHMAN, PRIYANKA --

Date: 20/08/2021

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Sales Forecasting for Rossmann Store using Machine Learning

Karan Veer Singh
MSc In Data Analytic
National College of Ireland
Dublin, Ireland
x20146248@student.ncirl.ie

Laiba Rehman
MSc In Data Analytic
National College of Ireland
Dublin, Ireland
x20144032@student.ncirl.ie

Priyanka --
MSc in Data Analytic
National College of Ireland
Dublin, Ireland
x20192037@student.ncirl.ie

***Abstract*— Rossmann is a global company with more than three thousand medical shops in seven European countries. Sales prediction is based on a combination of temporal and economic factors of previous sales data, shop promotions, retail competition, school and state holidays, store location and accessibility, as well as the time of year are all factors to consider. Decision tree will be used for prediction accuracy, with additional information from the data, allowing for the development of a more robust feature set and the strengthening of the sales prediction model. Machine learning solves real-world problem of forecasting retail sales. Store managers in developing efficient staffing plans that boost productivity. For modeling, analysis, prediction, and visualization, we utilized the popular open-source programming language Python.**

***Keywords* — Predictive Analysis, Sales, Visualization, sales forecasting.**

I. INTRODUCTION

Large scale store requires managers to focus on the sales to sustain the market competition. Store managers work on predicting the daily sales in advance to maintain smooth supply and demand every day on every location at every time. Comparative study within the stores of the items and with other stores sales can be accurately predicted using these factors Promotions, competition, school and state holidays, seasonality, and location. Store managers may establish effective staff schedules based on accurate sales with optimization issues such as ideal pricing, discounts, suggestions, and stock levels that may be solved using data analysis methods, retail is one of the most significant business areas for data science and data mining applications. The usual problems handled by information mining applications are Response Modeling, Proposal frameworks, Request prediction, Price Discrimination, Deals Occasion Arranging, and Category Management. Precise estimating of client demand remains a challenge in today's competitive and energetic business environment and minor enhancements in anticipating this demand make a difference expanded retailers lower working costs while improving sales and client satisfaction Predicting the correct request at each retail outlet is crucial for the success of each retailing company since it helps towards stock administration. Demand can depend on an assortment of outside components like competition, climate, regular patterns, etc. and inner actions like advancements, deals occasions, estimating, combination planning, etc., including the complexity of the issue. Consequently, the modeling of request forecast taking under consideration all the variables per retail outlet gets to be fundamental for every retail company. Hence, this paper proposes an approach for request and deals forecast for stores at each outlet. In a perfect world, store supervisors can utilize precise predictions to meet requests whereas minimizing stock footprint and, in this manner, operational costs. Assist, discrete components such as occasions, the opening of competitors, and advancements all have a significant level of request at any given day. We look to analyze the effect of these variables with the help of time series analysis and machine learning techniques. We utilized information from the Kaggle competition Rossmann Store Sales.

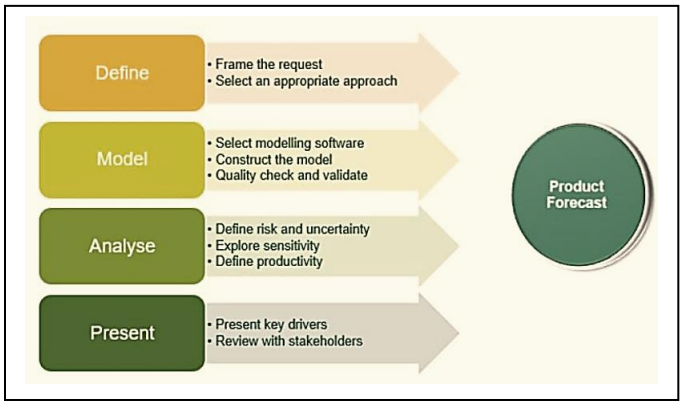

Fig. 1. Flow chart for Product forecast

Upon further examination, we realized that typically since Sales variations were generally driven by these discrete occasions, whereas the time-series patterns of regular or inter-year patterns were minimal. We accept that this finding is generalized to numerous forecasting problems, where more granular day-to-day predictions are required. For selling medical equipment and medications, retail pharmaceutical is a significant pillar in the healthcare business. It has several channels that include product production, retailers, wholesalers, branding, and licensing. Unlike a traditional pharmacy, a retail pharmacy deals with non-prescription medications.

II. APPLICABLE TECHNIQUES AND KEY FEATURES

A lot of time and effort has gone into figuring out how to anticipate sales. The Decision Tree Model has been used for sales forecasting in various scenarios due to its promising performance. The secret to winning in supply chain management is accurate demand forecasting. In retail sales, there are several techniques that can be applied for forecasting demand. The emergence of data mining techniques has resulted in the application of business intelligence in a variety of commerce industries. This study attempts to collect customer classification information using a decision tree in retail sales for demand forecasting. The study suggests a methodology that has been utilized in retail sales to enhance demand forecasting and inventory performance in whole supply chain management. The combination of the decision tree model with the inventory replenishment system leads to an enhancement in customer service and a decrease in inventory [1]. In this article, it is stated that, for modern retail organizations with a large network of stores, precise sales forecasting is critical in driving the company's corporate development growth, as well as its success or failure. Sales forecasting facilitates businesses to better manage resources, such as production and cash flow, and to make more informed business decisions. They have proposed a machine learning-based sales forecasting model that is both accurate and efficient. Feature engineering is used to extract features from past sales data at first. Furthermore, they employed eXtreme Gradient Boosting (XGBoost) to estimate future sales amounts using these characteristics. The findings of their experiment using a publicly reachable Walmart retail products dataset provided by the Kaggle competition enabled their model to perform very well for sales prediction while using less memory and computational time [2]. This article describes the implementation of a retail product recommendation system and sales forecast for a chain of retail outlets. The relative significance of consumer demographic variables is calculated and applied in the model for properly estimating the sales of each client type. When confronted to a single aggregate model developed for the whole dataset, modeling data at a higher degree of detail by clustering across client kinds and demographics delivers better results. The system implementation is explained in detail, as well as the practical difficulties that emerge in such real-world applications. Initial reports from test stores over a one-year span show that the system raised sales and improved efficiency considerably [3]. The rising level of retail engagement in today's consumer goods industry is illustrated by the growing number of alternatives for obtaining pharmaceuticals, consumers with varying tastes and high expectations are driving the market, combined with technological advancements that make retail services and product delivery more efficient [4]. Retailers and Manufacturers utilize Decision Trees for the graphical depiction to study how consumers form decisions. They assist in simplifying the consumer experience, for category segmentation, and developing overall category sales by providing specified product hierarchies [5]. Deep, prominent promotions on high-pull items have a huge net unit effect, but they have a significant negative impact on promotional margin, resulting in lower net profit. As a result, when it comes to promotion options, the retailer must undertake some hard trade-offs between profit goals and revenue [6].

III. EMPLOYED TECHNIQUE

In this research, the machine learning technique employed is the Decision Tree Model.

A. Exploration of Business Features using Visualization

Competition distance: As shown in figure 1, The stores that are the furthest away have the highest average sales and number of consumers. This doesn't necessarily indicate that the far competition distance is better, but it does highlight the fact that when there isn't much competition nearby, retailers tend to sell more and attract more consumers because they have a near-monopoly in this area.
As observed, stores with more competition distance have the highest average sales and number of consumers.

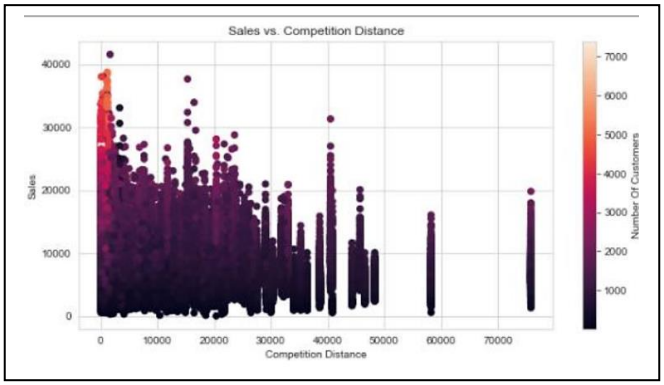

Figure 1 Sales versus Competition distance

School Holiday: When we look at the influence of school holidays on sales, Fig.3 we can observe that the impact isn't that significant.
Hypothetically, during school holidays sales should have the highest average sales and number of consumers but as observed during school holidays sales has no impact.

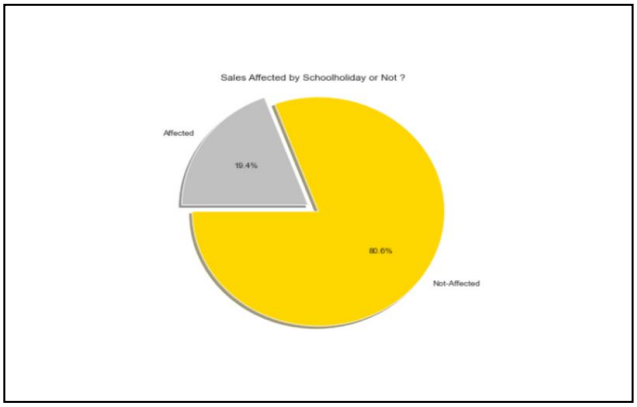

Figure 3 Sales during school holidays

Sales increase during weekends: The stores sales increased on weekend as most of the organisations and offices are closed during that period. Parents wants to take the kids for regular medical check up, also for themselves purchases the required dietary supplements and medical equipment's which increases the overall sales.

As observed, sales are not impacted significantly on school holidays. When we look at the influence of weekends on sales, we can observe that sales on weekends decreased.

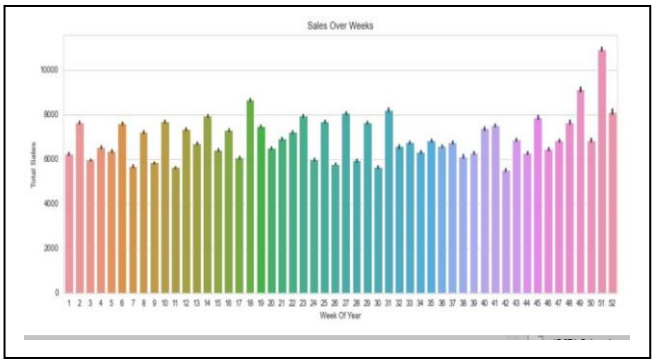

Figure 4 Sales during weeks

Sales increase during the second half of the year: When we look at the influence of year timeline on sales, we can observe that sales in the second half of year are less as compared to the whole year.

As observed, Sales decrease during the second half of year.

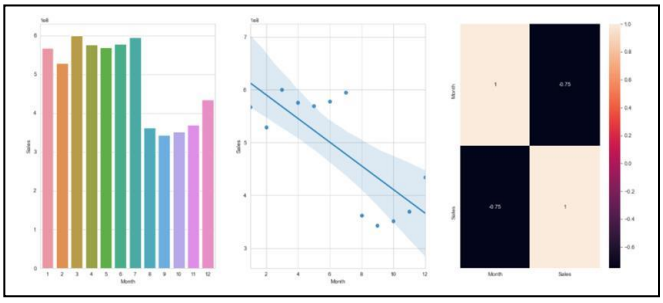

Figure 5. Sales during second half of the year

Sales over a time: The figure depicts the average of sales over a period of time. We can observe that sales are at their greatest around the time of the New Year, i.e., around Christmas.

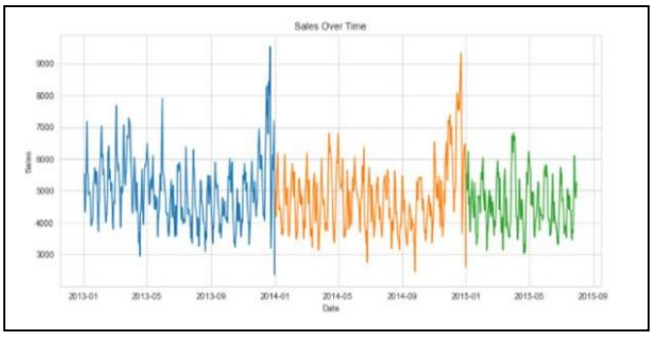

Figure 6. Sales over Time

Sales over assortment: The assortment is used to increase the product size as it focuses on uplifting the low-sale value product with the high sale value. most retails as well focus on this strategy that just changes the replacement of the products over the counter or add an offer to its combination. From figure 8 it can be observed average sale for assortment type B is more. It indicates assortment B is attracting more customers. Assortment A and C have equal no. of average sales.

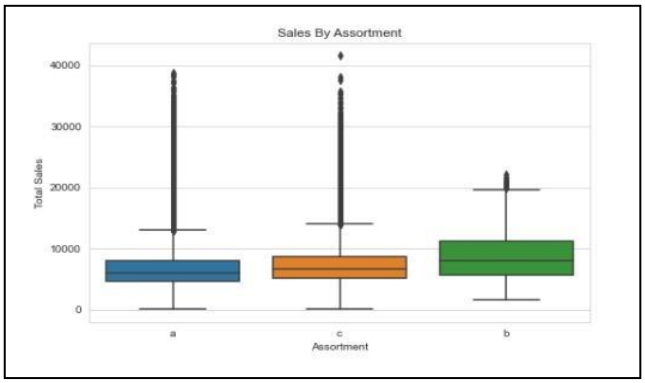

Figure 7. Sales by Assortment

Store Type: shows the average sales concerning the store type. It is observed from the plot for store type 'B' average sales are more and for the rest store type average sales are almost constant. The difference between sales of 'B' store type is almost 4k. It means Store type B is the best performing store.

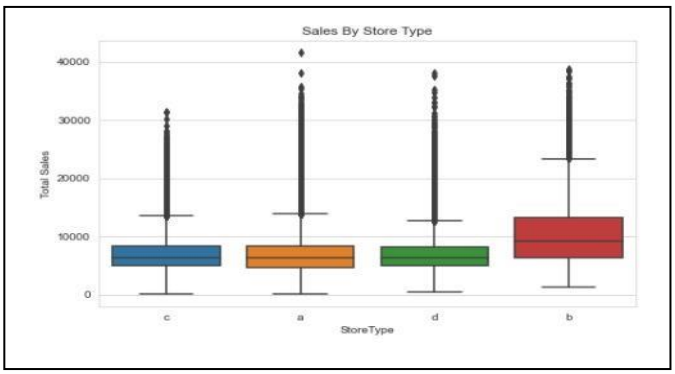

Figure 8. Sales by store Type

Sales During days of week: The combination of bar plot which represents the average no. of customers on the day of the week and line graph which shows the average sales is shown in fig 3. For the day of the week axis 1 denotes Monday, 2 denotes Tuesday, and so on till 7 which denotes Sunday. It can be observed from the plot that the customer footfall is more on Monday, Tuesday, and Friday in comparison to Wednesday, Thursday, Saturday, and Sunday, because most stores are not open on Sunday, so the sales volume is so less. The average sales are directly correlated to the customers, as the average no. of customers shrinks in Sunday sales gets affected as well.

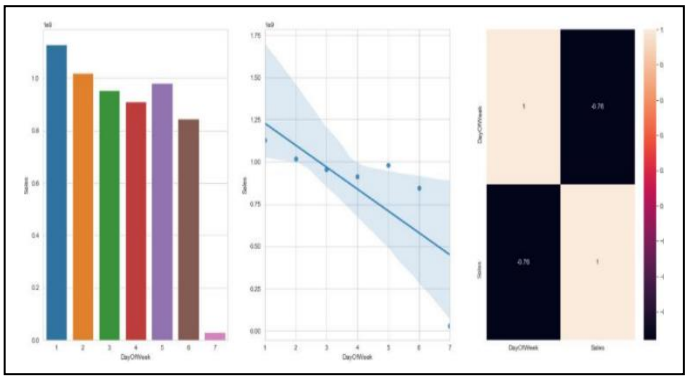

Figure 9. Sales during days of week

B. Data Pre-Processing

The information presented for the training classified the store into four groups. Sales data, "target factors" for each business, and numerous features such as distance from nearest competition, promotions, and customer count are all included. The data is a collection of both continuous and discrete variables. All retailers receive training to create a predictive model for each location. This enables us to drop characteristics like competition distance and store categorization, which are constant across all stores. The day of the week, the month, the promotion, and the school holidays are the four characteristics of those chosen for training.

C. Model Design for Predictive Analysis

We can recognize that the correlation between Customers and Sales is 0.82, implying that they are positively associated, as mentioned before in the analysis. It's intriguing to see that Sales per Customer is 0.28 and Promo is 0.28. Both have a positive correlation, as running a promotion improves that number. Sales per Customer also has a positive relationship with Competition Distance is 0.21), as I previously stated.

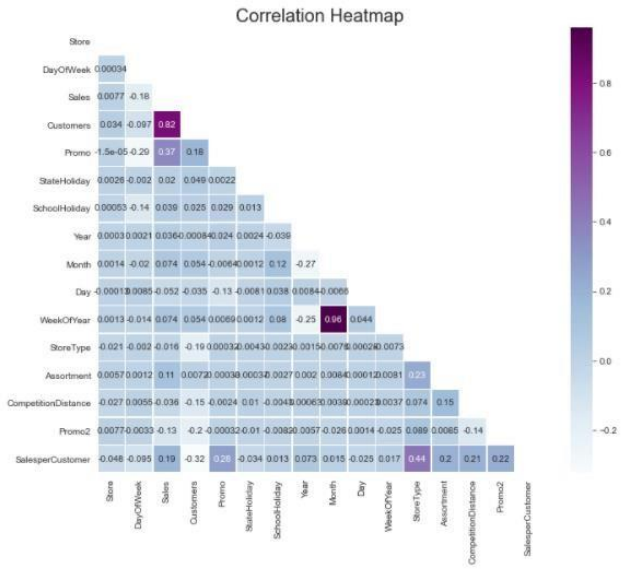

Figure 10 Predictive Analysis

The larger the competition distance, the more sales per customer we obtain, which makes sense because the more monopolization Rossman can achieve in the region. Furthermore, as we previously said (0.22), the effect of promo2 on Sales per Customer did cause a change in the buying pattern and enhanced it when continuous promotions were used. Finally, we can see that StoreType has a significant impact on Sales per Customer (0.44). This is most likely due to my encoding of the store type variable, which suggests that the high categories, such as d, which equals 4, have more weight, but if not, it makes sense that the last categories, such as d, would explain the increase in Sales per Customer.

D. Feature Importance by the Model

The below image shows that the features like 'average sales' , 'Promo' ,'weekofyear' , 'CompetitionDistance ' , 'AvgCustomer' , 'Month' are more important then the other features like 'SchoolHoliday' ,'Asortment_b' , 'StoreType_b' and other feature in model.

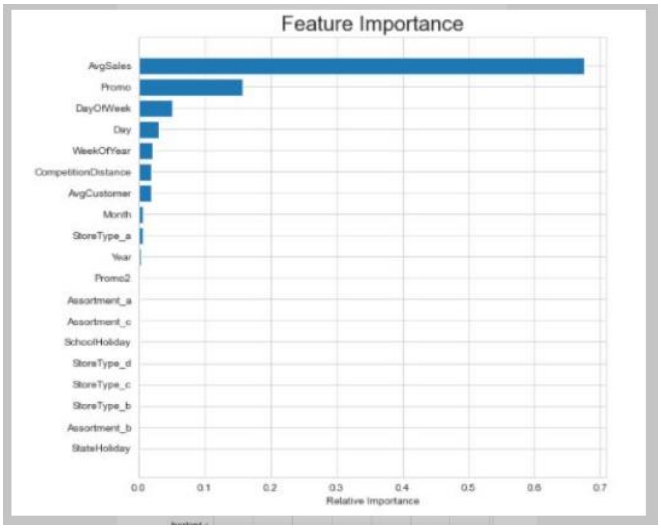

Figure 11. Feature Importance

IV. EVALUATION OF MODEL

1. **Qualitative Analysis:** Our approach will assist local shops to assist them in deciding marketing methods and also in budgeting and establishing financial rules. It is possible to estimate everything on average with effective sales forecasts, allowing average labor and production capacity to be completely utilized over the entire period. As a result, forecasting can help to overcome seasonal variations. Contribute to inventory management and avoid the dangers of overstocking and understocking. We may use projections to determine which products generate the most profit and which products should be phased out. To handle the foreseeable challenges, we believe that all organizations will contemplate projecting their sales at some time in the future.

2. **Quantitative Analysis:**

```
Desion tree Regression:

RSS: 326932956426.5358
Mean absolute error: 726.6503320023668
Mean squared error: 1205229.4697618384
Root Mean Squared Error: 1097.8294356419117
Accuracy: 87.4704500425944%
R Squared: 0.8747045004259439
Adjusted R Squared: 0.8746961856713058
Mean Absolute Percentage Error(MAPE): 11.0 2
```

Figure 12 Applied Machine Learning Model

Our decision tree model gathered the best and most consistent results with accuracy of 87.47%, Mean absolute error 726.650, mean squared error and R squared as 1205229.46 and 0.8747 in the validation, respectively.

With all the graph details, visualizations and after applying the machine learning models, we were able to gather very valuable information including that it's clear to see how the number of days after a competition opened its doors has an adverse effect on sales. There will usually be a lot of discounts a few days before a new competitor launches. The former store's sales will suffer a severe decline. However, the effect will fade over time. The dataset's original information is the date and the store. The sales trends in different stores will be different. The fact that the day of the month had such an impact on the forecast astonished us the most. We couldn't come up with a plausible explanation. We expect that sales fluctuate according to the seasons. The characteristics we extracted in the analysis were StoreMonthCustomers and StoreDayAverage. The average monthly sales are used to reflect the store's monthly changes. Sales will, on the surface, differ on different days of the week. However, they may follow the same pattern on the same day of the week. Weekends, for example, are more likely to have higher sales than weekdays.

CONCLUSION

We developed a model that allows Rossman shop managers to successfully forecast sales. Based on this projection, the company will be able to create an efficient work schedule for its staff. Our next step could be to design a visual interface for predicting sales. In any case, we currently have a model that can be used to create successful sales and programs using the Rossman system.

REFERENCES

[1] P. K. Bala, "Decision tree-based demand forecasts for improving inventory performance," *2010 IEEE International Conference on Industrial Engineering and Engineering Management*, 2010, pp. 1926-1930, doi: 10.1109/IEEM.2010.5674628.

[2] X. dairu and Z. Shilong, "Machine Learning Model for Sales Forecasting by Using XGBoost," *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, 2021, pp. 480-483, doi: 10.1109/ICCECE51280.2021.9342304.

[3] M. Giering, "Retail sales prediction and item recommendations using customer demographics at store level,"ACM SIGKDD ExplorationsNewsletter, vol. 10, no. 2, pp. 84–89, 2008.

[4] E. J. Fox and R. Sethuraman, "Retail competition," inRetailing inthe 21st Century. Springer, 2010, pp. 239–254.

[5]R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] **URL:** https://eyesee-research.com/decision-tree-root-of-behavior/

[7] K. L. Ailawadi, B. A. Harlam, J. Cesar, and D. Trounce, "Promotionprofitability for a retailer: the role of promotion, brand, category, andstore characteristics,"Journal of Marketing Research, vol. 43, no. 4,pp. 518–535, 2006.