Rossmann Store Sales Prediction

Shreyas Adiyodi, Nidhi Shah, Benjamin Wilson
Darden School of Business | School of Data Science
University of Virginia

***Abstract* —** This research paper highlights the teams use of the Cross-Industry Standard Process for Data Mining (CRISP-DM) approach to address the problem of accurately predicting future sales and customer demand for Rossmann, a company with over 3,000 drug stores in 7 European countries. The analysis assesses the impact of various factors such as promotions, competition, holidays, seasonality, and location on sales, and aims to minimize the Root Mean Square Percentage Error (RMSPE) while predicting store sales for six weeks. The exploratory data analysis highlights correlations, outlier stores, and seasonality, among other factors. Five baseline models are used for modelling, which are combined into a single ensemble model. The paper also includes additional qualitative insights to assist Rossmann in examining promotion and pricing strategies. These insights focus on the three lenses to pricing and psychological factors that influence human nature when shopping while providing suggestions on how to utilize customer value drivers, assess customer demand and propensity to purchase, react to potential competitor pricing models, and understand the level of involvement and visibility customers incorporate into buying behavior.

I. Introduction

Fluctuations in customer demand can greatly impact the financial performance of a business. Therefore, accurately estimating future sales and customer demand is essential for business growth. Sales forecasting involves predicting the sales or demand for a specific product during a certain period. Our paper demonstrates we are able to use modern machine learning principles to predict sales for Rossmann, a company with over 3,000 drug stores in 7 European countries. In our analysis, we assess how sales are impacted by factors such as promotions, competition, holidays, seasonality, and location.

Figure 1 - Rossmann Store Example

Currently, Rossmann store managers are tasked with predicting their daily sales up to six weeks in advance, and this process can lead to inconsistent results due to individual factors. As such, accurate predictions, especially for perishable or time sensitive items, is incredibly important so increasing sales and growing the business. Our objective will be to predict the store sales for six weeks while minimizing the Root Mean Square Percentage Error (RMSPE).

II. Data Description

Within the competition, three core data files were provided. A brief explanation of the files and fields may be found below.

Data Files

- train.csv - historical data including Sales
- test.csv - historical data excluding Sales
- sample_submission.csv - a sample submission file in the correct format
- store.csv - supplemental information about the stores

Data Fields

- Id - an Id that represents a (Store, Date) duple within the test set
- Store - a unique Id for each store
- Sales - the turnover for any given day (this is what you are predicting)
- Customers - the number of customers on a given day
- Open - an indicator for whether the store was open: 0 = closed, 1 = open
- StateHoliday - indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
- SchoolHoliday - indicates if the (Store, Date) was affected by the closure of public schools
- StoreType - differentiates between 4 different store models: a, b, c, d
- Assortment - describes an assortment level: a = basic, b = extra, c = extended
- CompetitionDistance - distance in meters to the nearest competitor store
- CompetitionOpenSince[Month/Year] - gives the approximate year and month of the time the nearest competitor was opened
- Promo - indicates whether a store is running a promo on that day
- Promo2 - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
- Promo2Since[Year/Week] - describes the year and calendar week when the store started participating in Promo2
- PromoInterval - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store

The data may be found on the Kaggle competition page here: https://www.kaggle.com/c/rossmann-store-sales/data?select=store.csv

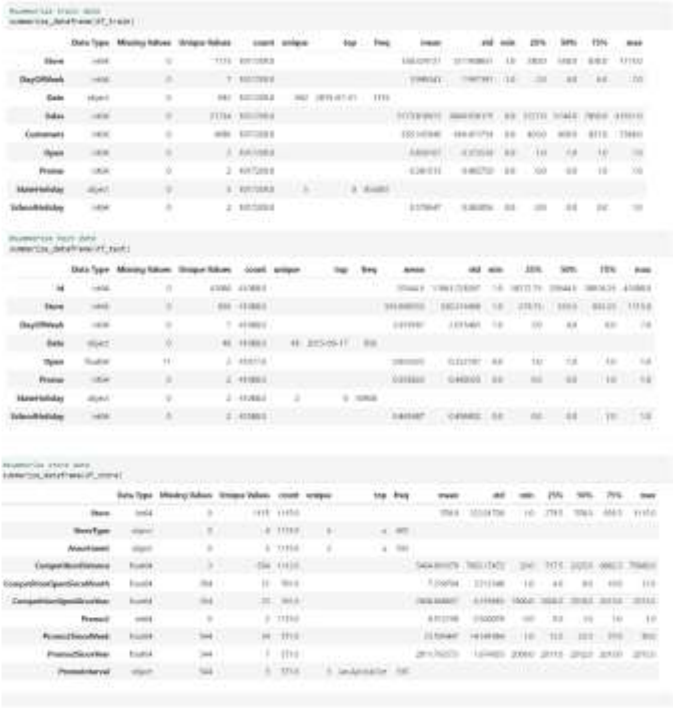


Figure 3 - Summary of Rossmann Data

III. Methodology

In this research paper, we present the application of the Cross-Industry Standard Process for Data Mining (CRISP-DM), which has been widely adopted in the data science community. The CRISP-DM framework was leveraged to guide the modeling approach throughout the project, ensuring a systematic and rigorous process. The framework consists of five stages: Define & Understand Business Problem, Collect & Understand Data, Explore & Analyze Data, Build & Validate Model(s), and Integrate Qualitative Insights.

1. Define & Understand Business Problem - Understanding the business objectives and processes, translating business problems into an analytical approach, and outputting a problem definition.

2. Collect & Understand Data - Understanding data availability, quality, and insights and forming hypotheses of expectations, culminating in data feasibility.

3. Explore & Analyze Data - Exploring the data to identify associations, anomalies, trends, patterns, and relationships, ultimately resulting in the selection and cleansing of data.

4. Build & Validate Models - Building and iterating the model to refine performance and conducting sensitivities to evaluate the validity of results, culminating in analytical results.

5. Integrate Qualitative Insights - Translating analytical results into business terms, integrating qualitative insights for well-rounded insights, and ultimately resulting in actionable business insights.

Overall, the application of the CRISP-DM framework ensured a systematic and rigorous modeling approach throughout the Rossmann capstone project, resulting in valuable insights and actionable business recommendations.

IV. Exploratory Analysis

Figure 3 illustrates the performance of four different store types. Despite having fewer stores, Store type B outperforms the others consistently with an average of $4k higher sales and 1k+ greater customers. The analysis also identifies over 20 outlier stores with above-average sales and customers, which significantly contribute to sales predictions. In contrast, Store type D consistently performs the lowest compared to the other three types.

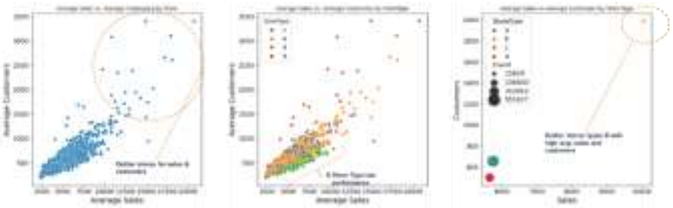


Figure 2 - Store Comparison by Sales & Customers

Similar to peers in the retail industry, Figure 4 demonstrates that Rossmann stores are affected equally by seasonality in their sales cycles. Multiple cycles in sales can be observed throughout the figure, with a -40% decrease in sales in November, followed by a ~10% growth in December, indicating that the holiday season materially influences sales. Additionally, a ~15% increase in sales can be seen in April, which should be considered in further analyses to understand the cause for future predictions.

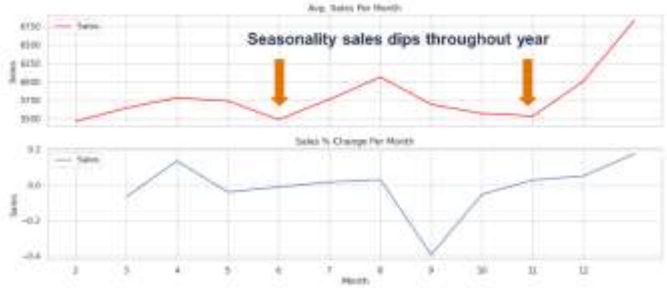


Figure 4 - Sales Seasonality

Our analysis reveals also reveals that Sunday store closures have a significant influence on the weekly sales cycle, leading to Monday sales spikes. There is a gradual decrease in store sales throughout the week, with a minor sales increase on Fridays (day 5 on the Figure 4), indicating end-of-week purchases by customers.

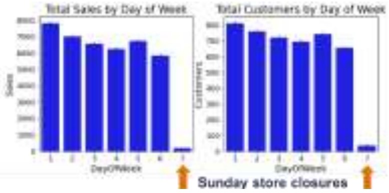


Figure 5 - Daily Average Sales

Through analysis of the holiday impact to store sales, we identified that store closures for state holidays, but not for school holidays, create an effective prediction point in assessing the number of store days open for sales week. Similarly, like Sunday store closures, few stores choose to open on state holidays, causing significant sales increases on days before such closures occur. Predicting store sales days prior to store closures for holiday purposes could create a strong prediction variable. However, school holidays show little to no difference in sales between stores as seen in Figure 6, indicating that it is a less powerful attribute to perform predictions with throughout modeling process. Correlation tests indicate a similar variable importance.

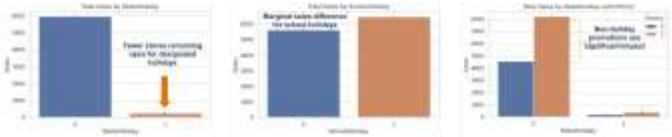


Figure 6 - Sales by Holiday & Promotion

Furthermore, we can see that non-holiday days with promotions create a positive interaction effect. Promotions are at initial glance an effective driver of sales and become a strong predictor of future store sales. However, few promotions are run on or during holidays given that most stores choose to close.

The impact of the two different promotion types on store sales can be seen in Figure 7 where promotion 1 leads to an average sales increase of approximately $4k, while promotion 2 (consecutive promotions) results in an average sales decrease of around $750. Roughly 27 more stores engage in consecutive store promotions with no similar decrease in customer count (i.e. the same number of customers come during promotion 2, yet spend less money). In contrast, participating in a single promotion appears to be successful in driving sales growth. Therefore, the consecutive promotion appears to be unsuccessful in driving sales.

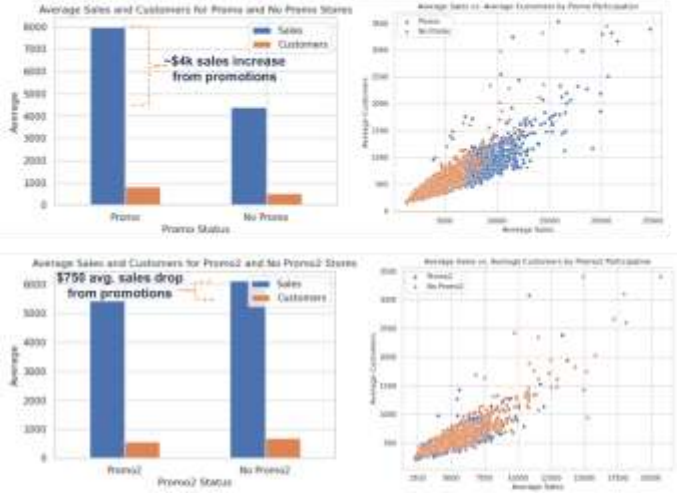


Figure 7 - Sales Promotion Effectiveness

The promotion 1 effectiveness may not be as effective as thought though, given customers buying behavior as seen in Figure 8. Clear trends appear pre and post promotion periods, depicting clear customer buying patterns. An apparent pre-promotion sales dip indicates customers predict the promotion will occur and choose not to buy until the promotion takes effect. Similarly, a post-promotion sales dip occurs from customers buying items in bulk to hold over until the next promotion occurs. The consistent sales dip prior to promotion periods combined with consistent promotion periods by Rossmann allows us to understand that customers are successfully predicting promotions and buying items in bulk to hold themselves over until the next promotion occurs. This behavior is quite common in retail when these two factors exist.

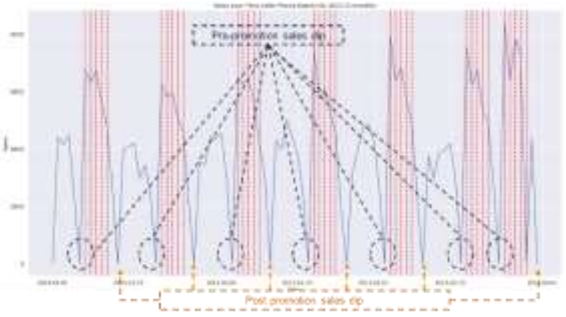


Figure 8 - Promotion Prediction Pattern

Maintaining effectiveness with promotions though is complicated with the increasingly competitive landscape for Rossmann stores. Figure 9 illustrates the depicts this competitive landscape as seen in the 15-year period leading up to 2015, the peak entry of competitors between 2007 and 2013. As competition has increased, they have chosen geographically close locations to Rossmann stores, creating a larger competitive environment and leading to a more aggressive landscape.

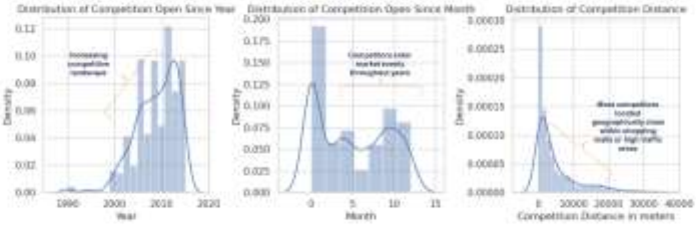


Figure 9 - Competitive Landscape

The exploratory analysis is summarized successfully in Figure 9 where we can see unsurprisingly an almost perfect correlation between customers and sales. It is worth noting, although obvious, that customers is naturally not a factor which can be used for predicting the future store sales. The promotion 1 and promotion 2 remain strong predictors of sales given their effectiveness in both increasing and decreasing store sales numbers. As seen as well, the state holiday's are an effective measure as well, given that stores simply will close down for those days, negatively impacting sales. It should be highlighted that, although the competitive environment has increased considerably for Rossmann, those attributes associated with competition have less correlation with sales than perhaps expected. No further data is provided in terms of competitor prices or promotions as this may cause instances of stronger correlation in predicting Rossmann store sales.

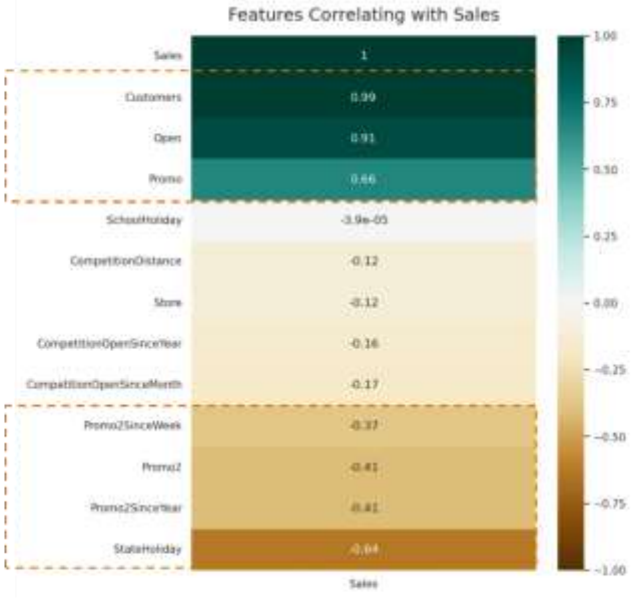


Figure 10 - Feature Correlation with Sales

V. Modelling Approach

In order to improve the future predictive accuracy of our model if Rossmann were to implement this over the long-term, we chose to is to integrate multiple models into an ensemble. Ensembles leverage the strengths of individual models and mitigate their weaknesses, resulting in a more robust and accurate prediction tool.

In developing each model, we employed cross-validation techniques to evaluate the performance. A 75/25 train/test split was chosen to evaluate the model's generalizability to new data. Additionally, we separated the training data further into smaller training and validation datasets for individual model training purposes.

To combine the individual models' results, we used a linear regression weighting approach, assigning weights to each model based on their individual performance, which ensured that the optimal models contribute more to the ensemble's final prediction. In order to evaluate the ensemble's predictive performance, we utilized several evaluation metrics, including root mean squared error (RMSE), mean absolute error (MAE), R-squared (R2), and root mean squared logarithmic error (RMSLE). These metrics were chosen because they provide a comprehensive evaluation of the ensemble's accuracy, precision, and generalizability.

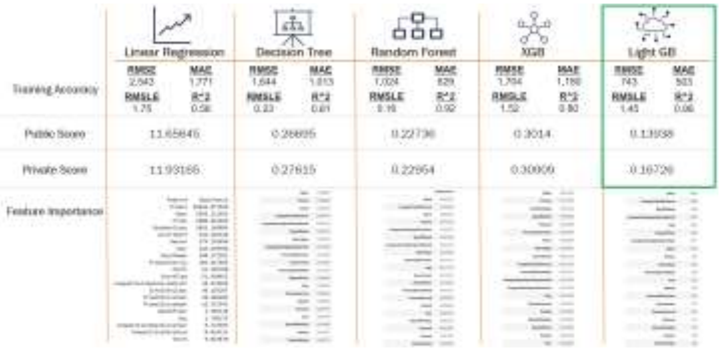
To develop an effective ensemble, we utilized a combination of five models: linear regression, decision tree, random forest, XGB, and Light GB. Each model has a unique strength that contributes to enhancing the ensemble's overall long-term effectiveness. We assessed each model's individual performance and feature importance to determine optimal features for the ensemble.

Linear regression was chosen as it is a simple and interpretable model that identifies linear relationships between features. This allowed our team to initially test whether such relationships exist in the greater model beyond those seen in the correlations during exploration. Next, tree based models were chosen to test interaction effects of the data. Specifically, the decision tree model was chosen to capture the complex interactions between the various features, especially those of a categorical nature. Additionally, random forest was chosen as an ensemble itself, but of prediction trees, in order to overcome decision tree limits and handle outliers and noise that exist from the ~20 stores with larger sales an customer numbers. An XGB model was then added as it takes advantage of regularization to prevent overfitting and assess feature importance as the tree based models (especially decision trees) may have overfit during the training process. Further, light GB model was finally used as it tends to be optimal for datasets with numerous features, making them well-suited for semi-high-dimensional data.

VI. Model Results

As can be seen from Figure 11, the prediction scores greatly enhanced as more advanced models were trained, providing a top R^2 score of 96% from the Light GB model. The more simple linear regression model's poor performance can be attributed to the lack of linear correlation that exists between sales and the numerous variables. Although the tree based models performance were superior to that linear regression, its inability to lower their public and private scores below 0.22 can be attributed to lack of tree depth and a lack of interaction effects in the data.

The top performance of the light GB model can be attributed to multiple factors, top of which is the built in regularization. Light GBM models use L1 and L2 regularization techniques to prevent overfitting and remain flexible, an attribute that allowed it to be far superior to other models in this instance.



	Linear Regression		Decision Tree		Random Forest		XGB		Light GB	
Training Accuracy	RMSE 2,543	MAE 1,771	RMSE 1,644	MAE 1,013	RMSE 1,024	MAE 829	RMSE 1,704	MAE 1,180	RMSE 743	MAE 503
	RMSLE 1.75	R^2 0.58	RMSLE 0.23	R^2 0.81	RMSLE 0.16	R^2 0.92	RMSLE 1.52	R^2 0.80	RMSLE 1.45	R^2 0.96
Public Score	11.65846		0.26895		0.22736		0.3014		0.13938	
Private Score	11.93166		0.27615		0.22654		0.30009		0.16726	
Feature Importance	[illegible]		[illegible]		[illegible]		[illegible]		[illegible]	

Figure 11 - Individual Model Performance

For the Kaggle competition, it is worth noting that the Light GB model would be advantageous to submit as it has the optimal performance. For Rossmann though, it should be recognized that an ensemble model has its benefits for long-term use for enhanced generalization and to reduce overfitting by a single model. To do so, our team performed a linear regression between sales and the predicted sales by each of the five models to determine the coefficients or weights that should be assigned to each model prediction. This results in each model prediction being multiplied by its assigned linear coefficient and summed together with the linear intercept to provide a single ensemble model predicted score. As assumed given that it is a stacking of various models, the ensemble model does not outperform the light GB model for this competition, as can be seen in Figure 12.

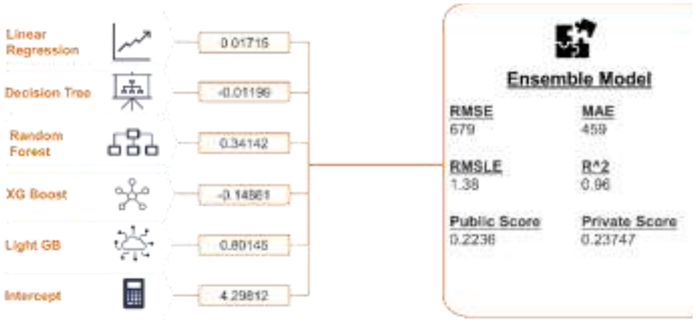


Figure 12 - Ensemble Model Input Weighting & Performance

VII. Qualitative Insights

Qualitative insights to customers propensity to buy is critical in predicting, and otherwise increasing, future sales given the nearly perfect correlation that exists between customers and store sales. The optimal pricing and promotion behavior required for Rossmann stores is emphasizing by the balance required between the three lenses to pricing - economics, customers, and competitors – and the importance of understanding the underlying psychological responses that influence customer behavior – level of involvement and human nature.

For the economic aspect of pricing, the customer segments' demand flexibility, the stores local market share, and the general size of the local channel are critical to setting prices. This magnifies the need for Rossmann to determine the segment and elasticity of each segment of customers, enhance its understanding of each location's market share compared to competitors, and use the smaller channels' elasticity in determining how to run future promotions.

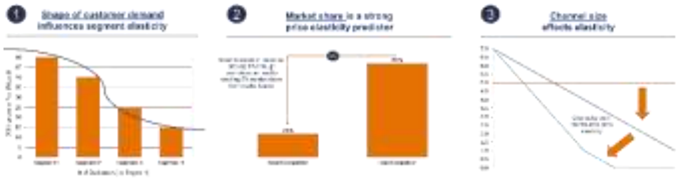
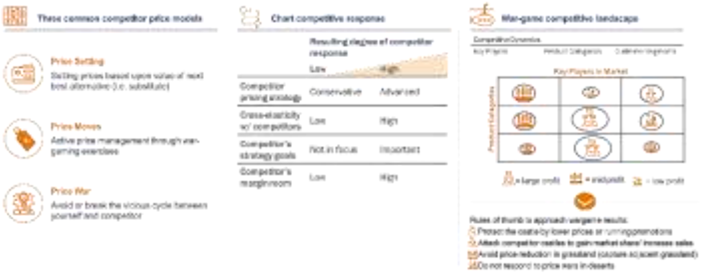


Figure 13 - Economic Lens to Pricing

From the customer lends, the complexity of retail customers requires Rossmann to balance the three value drivers per customer segment with their propensity to buy. Understanding the drivers which can be flexed with each segment allows Rossmann to better assess each customers propensity to purchase and otherwise enhance store sales predictions.

The competition landscape is also a crucial factor as has been seen from the growing level of competition in a short distance to Rossmann stores. The three common competitor price models - price setting, price moves, and price wars – can greatly shape the effectiveness of store promotions as well as have a greater effect on overall profitability. Understanding the competitive landscape not simply in-term of competitor pricing models but in competitive response to Rossmann price setting and promotion schedules will aide Rossmann balance leading and reacting to price promotions effectively. This should be measured through war-gaming the competitive landscape, identifying by market and category which sectors it should compete for profitability versus allow competition to take.


These common competitor price models
Price Setting
Setting prices based upon value of next best alternative (i.e. substitute)
Price Moves
Active price management through war-gaming exercises
Price War
Avoid or break the vicious cycle between yourself and competitor
Chart competitive response
War-game competitive landscape
Key Players in Market
Product Categories
= large profit
= mid profit
= low profit

	Resulting degree of competitor response	
	Low	High
Competitor pricing strategy	Conservative	Advanced
Cross-elasticity w/ competitors	Low	High
Competitor's strategy goals	Not in focus	Important
Competitor's margin room	Low	High

Rules of thumb to approach wargame results:
Protect the castle by lower prices or running promotions
Attack competitor castles to gain market share/ increase sales
Avoid price reduction in grassland (capture adjacent grassland)
Do not respond to price wars in deserts

Figure 14 - Competition Lens to Pricing

From a psychological perspective, Ross should understand that of the four types of purchase decisions customers make based on external visibility and internal mental involvement, the average item purchased from their stores is found in the low visibility, low involvement quadrant. This results in customers being highly elastic to price thresholds and price points as they make purchasing decisions. As such, Rossmann should offer smaller, varied promotions for minor utility gains as customers will seek deals to influence future buying decisions while integrating price raises at once to limit negative utility losses to the end customer.

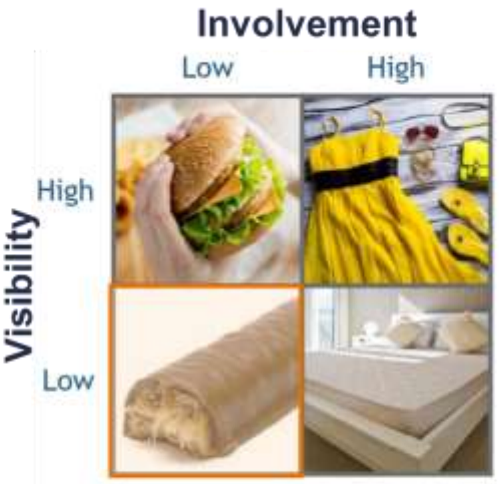


Figure 15 - Visibility vs. Involvement of Customer Decisions

Finally, the importance of considering the customer's utility curve, which emphasizes negative experiences more than equal positive experiences, influences Rossmann’s future promotion and price raising strategy. Negative impacts to customers are felt more than equivalent positive impacts due to how we value negative versus positive utility in human nature. Therefore, promotions should be small, spread out, and varied for customers to receive consistently small utility increases, while price raises should come all at once toward the end of the negative utility function, where customers experience diminishing rates.

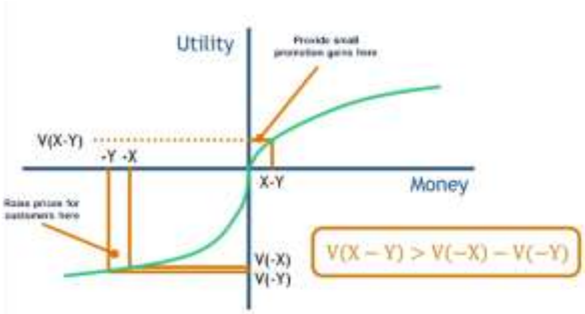


Figure 16 - Customer Utility Curve

VIII. Conclusion & Recommendations

Based on the analysis and findings, it is recommended that Rossmann consider the following strategies to enhance its pricing and promotion strategies:

1. **Enhance Store Promotion Effectiveness**: Rossmann should remove consecutive promotions that have been shown to be ineffective in increasing sales or profit. Future promotions should be randomized to limit customers ability to successfully predict promotion schedules and stockpiling goods. Frequent, smaller promotions should be run to incentivize recurring customer shopping patterns while spreading out actions that provide customers positive utility feedback. Bundling all price increases for one time will limit the negative utility impact on customers as well. This strategy recommendation is highly realistic to implement with a high expected impact on sales and thus, should be prioritized first.

2. **Utilize Customer Elasticity of Demand**: Rossmann should segment its customer base and evaluate each segment's level of demand. Market share and channel size of stores across markets should be calculated to review pricing power. Inputs should be utilized to analyze the elasticity of demand for customers, and promotions should be set based on the results. This strategy is moderately realistic to implement with a moderate expected impact on sales.

3. **Understand Competitive Mindset and Tactics**: Rossmann should assess the competitions pricing and promotion strategy in terms of its impact on Rossmann's future store sales. The company should determine the competitive response to historical Rossmann store promotions and war-game the pricing and promotion landscape to optimize the strategy for future responses. This strategy is moderately realistic to implement with a moderate expected impact on sales.

4. **Target Customer Value Drivers**: Rossmann should weight the three customer value drivers to determine customer preferences in the form of utility gained. The store experience and promotions should align with the weighted drivers to increase customer surplus. Doing so would allow Rossmann to enhance customer propensity to buy and, in turn, increase customers average basket size. This strategy is more difficult to implement and has a more limited impact on sales, therefore, it should be prioritized last.

Overall, the recommended strategies will help Rossmann optimize its pricing and promotion strategies, increase customer satisfaction, and boost sales. It is suggested that the company prioritize the implementation of the first two strategies and gradually move towards the others. Further research and analysis can be conducted to assess the impact of these strategies and make necessary adjustments to optimize their effectiveness.

Acknowledgment

Our team would like to express our sincere gratitude to several individuals who have been instrumental in the success of our capstone project. Firstly, a thank you to Doug Thomas and Alex Cowen from the Darden School of Business for their timely assistance in finding a suitable client for our project, despite the challenges we faced. Their support was invaluable.

Secondly, we would like to express a sincere thank you to Ron Wilcox from the Darden School of Business for his invaluable expertise and guidance in advising us on price and promotion trends. His insights and advice were crucial in shaping our project and achieving our objectives.

Finally, we would like to express our appreciation to Asifur Mehedi from the Darden School of Business for his flexibility and understanding throughout the capstone project. His willingness to adapt and adjust to changing circumstances was greatly appreciated, and his contributions were instrumental in the successful completion of our project.