Rossmann Store Sales Prediction

Nidhi Shah | Shreyas Adiyodi | Ben Wilson

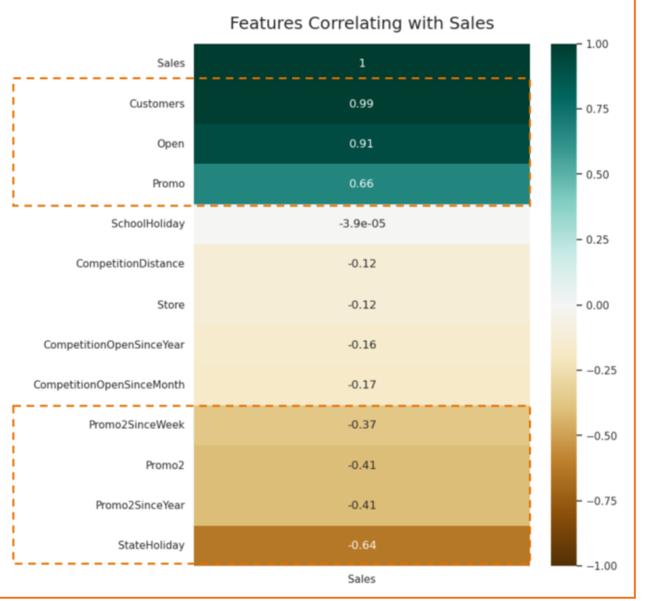
University Of Virginia
Darden School of Business
School of Data Science

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 customers level of involvement.

Background

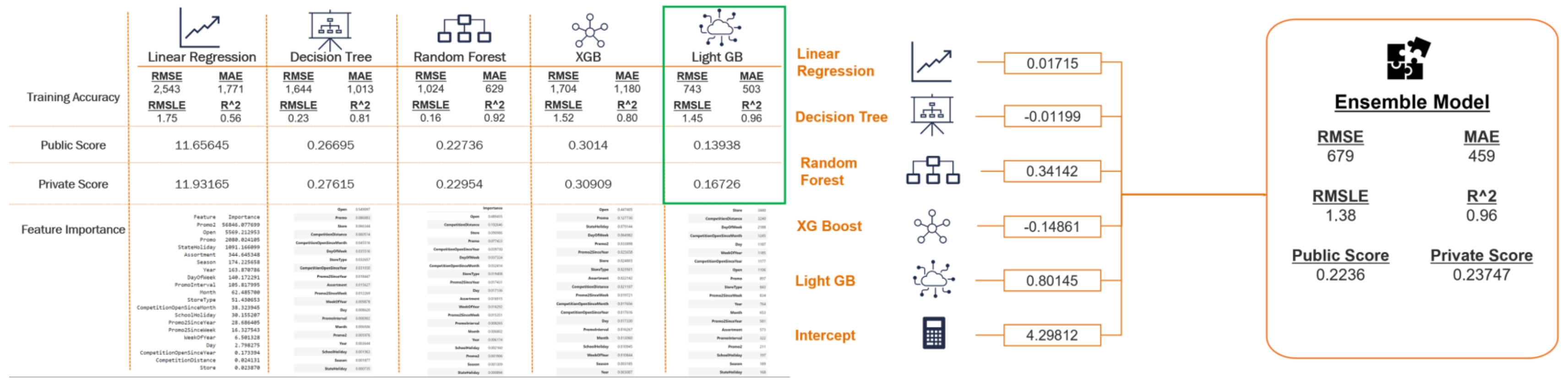
The Rossmann Kaggle dataset provides valuable insights into the factors that influence sales and customer behavior in a retail setting. In this study, our team followed the CRISP framework to fully understand the business problem while developing quantitative insights to predict future sales. Through exploratory data analysis on the dataset, our aim was to understand the impact of various factors on sales and customer behavior in order to make informed decisions that can help improve business performance. Specifically, we drew enhanced focus on the promotions that were scheduled to determine if the optimal promotion schedule was conducted to increase sales and, by extension, store profitability. Extrapolating beyond the Rossmann prediction problem, the larger question as to how stores should predict sales and space promotions is a common question. It was Our goal to develop a model with deep insights that could be used across a greater population of retail or non-retail clients in the future. Such research is focused beyond the current store but also how one should think about and respond to competitors offering similar promotions in the market.


Features Correlating with Sales

Materials and Methods

Through exploratory analysis, our team revealed several key insights. First, there is a high positive correlation between customers and sales, which indicates that attracting and retaining customers is crucial for driving sales. Additionally, holidays have a strong negative correlation with sales due to closures (similar to Sunday's), with no noticeable sales increase prior to the holiday. Furthermore, competition features have little impact on sales.

In addition, we identified significant outlier stores in both sales and customers, with store type D consistently underperforming while store type B well outperforms other stores. Seasonality naturally plays a role in sales, with dips in June and November, yet increases in August and December. Regarding promotions, we found that promotion 2 (consecutive promotions) is ineffective as fewer sales occur, while promotion 1 is effective for increasing sales significantly. However, we need to verify on a per-item basis whether promotion 1 truly increases overall profitability, as sales increases may offset profit decreases per item. We also observed that promotions have become predictable to customers, resulting in a sales dip before promotions and after promotions. Finally, we noted that competition has increased both in number and in close proximity to Rossmann stores underscoring the importance of monitoring and responding to competitor pricing and promotions.



	Linear Regression	Decision Tree	Random Forest	XGB	Light GB
Training Accuracy	RMSE 2,543 MAE 1,771 RMSLE 1.75 R^2 0.56	RMSE 1,644 MAE 1,013 RMSLE 0.23 R^2 0.81	RMSE 1,024 MAE 629 RMSLE 0.16 R^2 0.92	RMSE 1,704 MAE 1,180 RMSLE 1.52 R^2 0.80	RMSE 743 MAE 503 RMSLE 1.45 R^2 0.96
Public Score	11.65645	0.26695	0.22736	0.3014	0.13938
Private Score	11.93165	0.27615	0.22954	0.30909	0.16726
Feature Importance	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Linear Regression	0.01715
Decision Tree	-0.01199
Random Forest	0.34142
XG Boost	-0.14861
Light GB	0.80145
Intercept	4.29812

Ensemble Model

RMSE	MAE
679	459
RMSLE	**R^2**
1.38	0.96
Public Score	**Private Score**
0.2236	0.23747

To predict sales, we utilized modern machine learning principles to develop predictive models while minimizing the Root Mean Square Percentage Error (RMSPE). To accomplish this, we created five baseline models using linear regression, decision tree, random forest, xgb model, and light gb model. Each of these models had their own strengths and weaknesses, but by combining them into a single ensemble model, we were able to take advantage of each model's strengths while minimizing the risk of any model's bias.

The linear regression model was chosen as a baseline because it is simple and interpretable. However, it is limited in its ability to capture complex relationships in the data. The decision tree model, on the other hand, is able to capture complex relationships, but can be prone to overfitting. The random forest model helps to mitigate this by aggregating the results of multiple decision trees. The xgb model and light gb model are both gradient boosting models that perform well in predicting complex nonlinear relationships.

By combining these models, we were able to develop a more robust and accurate prediction model. In addition, the ensemble approach also helped us to identify which features were most important in predicting sales. By analyzing the feature importance of each model, we were able to identify the key drivers of sales and optimize our predictions accordingly.

Results

The model performance greatly enhanced as more robust methods were applied to the sales prediction problem. Linear Regression was primarily ineffective in its prediction. The Decision Tree produced a moderate prediction with a public score of 0.26695, although utilizing Random Forest for an additional tree method saw an enhancement to 0.22736. As we moved to more gradient based methods, they appeared initially less effective with XGB producing a score of 0.3014, although Light GB produced the optimal score of all models used for a score of 0.13938.

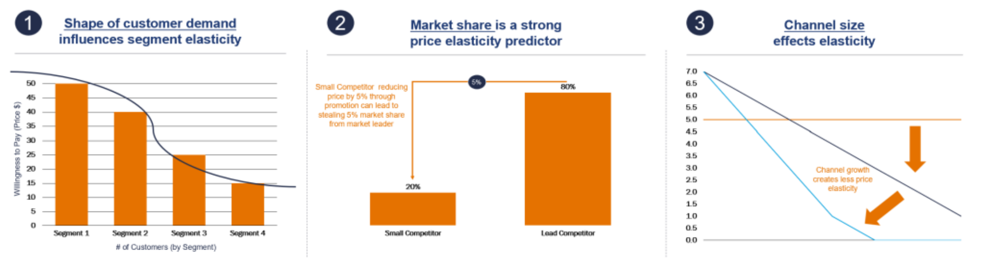
Given that Light GB had the best score of the models produced, it was given the highest of weights to use for the Ensemble model, determined by a further Linear Regression against the model outputs when regressed on the true sales output. In stacking these models with the appropriate linear weights for the Ensemble, it produced a score of 0.2236. Although not an optimal score, it has the strength of being less biased to the training data than each individual model (including the Light GB) and may enhance its predictive power as further training data is applied in the future and the input models are continuedly enhanced and refined.

Conclusion

For the Kaggle competition, the Light GB model should be submitted due to its enhanced performance is minimizing RMSPE to predict future sales. For the Rossmann store themselves, a more ensemble approach should be leveraged to minimize any current model bias that may be present if one were to only use the current Light GB model.

Future Direction

To gain insights into pricing and promotion strategies, Rossmann should consider a range of qualitative factors. Firstly, to analyze customer demand curves and take into account local market share and channel size when determining how promotions and pricing should be approached. Secondly, identify the customer value drivers that are relevant to different segments to gain a better understanding of customer demand and their propensity to purchase. Thirdly, how to react to potential competitor pricing models and how their competitors might respond to their own promotion strategy. Finally, level of involvement and visibility customers incorporate into their buying behavior should be evaluated, as this has an impact on the utility they receive in both positive and negative interactions.


1 Shape of customer demand influences segment elasticity
Willingness to Pay (Price $)
50
45
40
35
30
25
20
15
10
5
0
Segment 1
Segment 2
Segment 3
Segment 4
# of Customers (by Segment)
2 Market share is a strong price elasticity predictor
Small Competitor reducing price by 5% through promotion can lead to stealing 5% market share from market leader
5%
80%
20%
Small Competitor
Lead Competitor
3 Channel size effects elasticity
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Channel growth creates less price elasticity