Impact of Price Amendments on Sales: A Machine Learning Approach

Nouroz Rahman Amon

Supervisor: Dr. Pascal Killian and Bergfruende GmbH

Abstract

In this paper, we have analyzed the sales dataset of Bergfreunde DE step by step to understand the influence of discounts and other factors on the sales of different products using several machine learning methods. Firstly, we performed extensive exploratory data analysis on the whole dataset to understand underlying features. Then, the dataset was cleaned by reducing unnecessary features and using the univariate outlier detection methods. The seven most important input features for modeling are selected and used for further feature engineering. Finally, multiple ML models were used to predict the future daily sales under the influence of discounts, and performance metrics of Machine learning models are presented, analyzed, and compared. A hyper-parameter-tuned Gradient Boosting model provides the best predictions with a 35% WMAPE error rate. Few key insights from ML models are also discussed in the end.

Introduction

In recent years, tech-based E-commerce businesses have been very popular and one of the main attractions to purchase these products online is discounts. Additionally, COVID lockdown has fueled the need for online retail shopping due to lockdowns worldwide. A large portion of these sales occurs in discounts. As a result, to maintain profitability and maximize customer experience, the study of the relationship between sales and discounts is significantly important for industries. Moreover, the planning teams have been using traditional methods to visualize the relationships between key factors and sometimes lack a proper data-driven decision-making process.

Numerous studies have been developed in the field of market research focused on promotions, their strategic implications, and their impact on sales and other factors [1]. The studies have been conducted previously to compare multiple machine learning models and compare their results in the presence of discounts [2]. Some studies show the effect of discounts on different demographics and compared them statistically [3].

Because of the strong trend and seasonal patterns exhibited by Retail sales [4] the most widespread methods have been Winters exponential smoothing, seasonal autoregressive integrated moving average (ARIMA) model, and multiple regression which have the ability to model trend and seasonal fluctuations presented by aggregate retail sales. However, the linearity in which these algorithms most effectively perform has been compromised. Researchers have proposed a hybrid system based on clustering and classifying the items according to their sales behavior using decision tree classification to forecast sales in a textile retailing environment [5]. Due to economic instability and more fierce competition [6], recent retail sales time-series data show a higher degree of variability and nonlinearity, which decreases the accuracy of these models, hence justifying the use of nonlinear models such as the tree-based models or boosting models which seem more suitable for such complex cases and a large amount of data.

In our research, we aim to use a multivariate approach using many numeric as well as categorical features to build a predictive model that is more explainable and domain-centric.

The objective of this research is to predict the daily sales in the presence of discounts using only the sold-product features of a dataset. To achieve this, we can break the work down into three stages: exploratory data analysis and data cleaning, building and experimenting with several machine learning models, and evaluating multiple key performance metrics of the models for choosing an optimal option. Both linear models and non-linear models have been used for predictions after cleaning the dataset and feature engineering. In the end, we have tried to sketch different possible scenarios for the errors and suggested ways to overcome those too.

Methods

1. Description of data

The dataset consists of sold product data at a product level aggregation from the year 2017. It is data from a sports accessories selling website. This is the first year of available data available up to 2021 end. The data is at the SKU level and can be aggregated in any form. With above 10M item-level data to work on and above 30 usable features 6 of which are categorical. For our case, as we had access to the team maintaining this data set, we were able to proceed with the next step with an in-depth working understanding of the data set, collection procedure, and usage procedure.

2. Exploratory data analysis

In any data science project, the first and most fundamental step is to analyze the dataset, and understand its structure and features. In this step, analysts use different data visualization techniques as well as statistical methods to discern underlying facts about data. This step is an ideal example of descriptive statistics and its usage.

Our first step was to analyze and comprehend the individual features and their trends. For this, we studied the distributions, pairwise correlations, and looked for extreme value tendencies and origins.

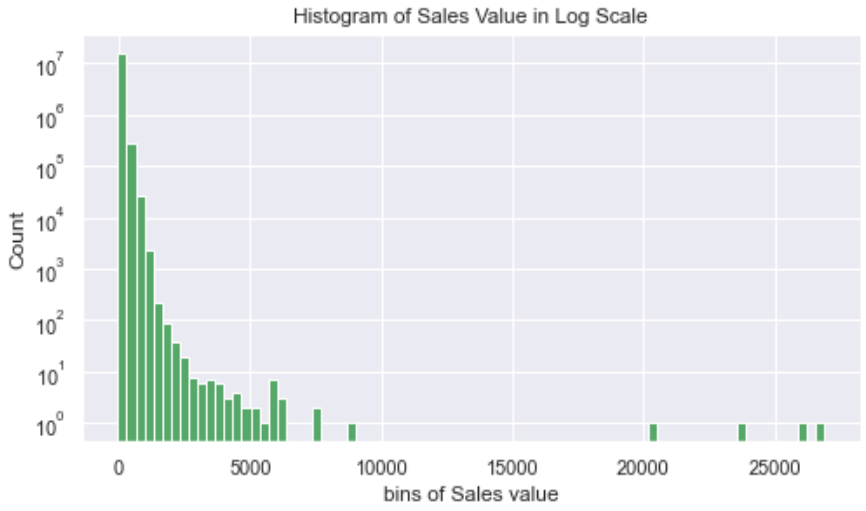


Fig 1: Histogram of sales value

We found several data columns with extreme value tendencies. As seen in Fig 1, The histogram of sales values shows few extreme values (y-axis) at a large distance to majority values (x-axis). Extreme values always have a negative effect on predictive models. That is why we iterated through some data-cleaning methods which are discussed later.

Another part of this step was to look for data inconsistencies and correction needs. This is a necessary and unavoidable step to ensure data sanity for all next steps and can be a time-heavy procedure. Apart from null values rendering several features (e.g.: Size 2, Category 4, Product Season) less impactful & some irregular values, there were no major corrections needed such as string or date manipulations. There were some inconsistencies flagged which are discussed later in the paper.

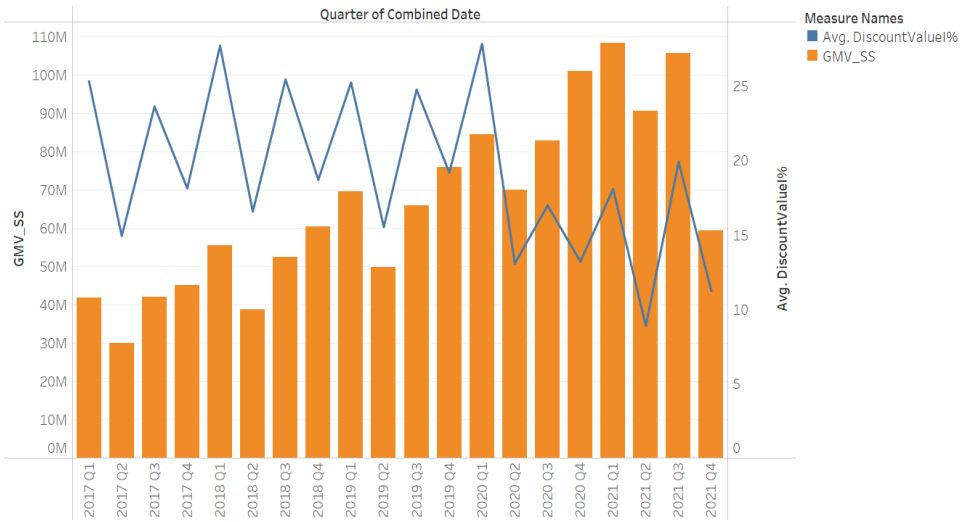


Fig 2: Quarterly sales performance vs. average discount percentage

From this step, in bridge with the discussions from original source team in the primary stage, we concluded that certain data columns such as shipping date, delivery date, and value information were not impactful. Few value fields were also correlated as expected. Data columns such as shop number and SKU details were also skipped as they would not be a factor of influence for results.

Here, we also reached a few other conclusions for the next steps. We included quarter and month as features based on the analysis that showed that discounts had a relational impact based on which month or quarter they were being offered (Fig 2). We also noted a possible prediction concern by noting the inverse relationship of discount versus sales in recent data.

3. Data cleaning

- Extreme value cleaning: In any dataset, observations beyond normal amplitudes are known as outliers. If these data points are not removed in the training process, these observations, can greatly skew the model and reduce the predictive power of the ML model. As a result, we cleaned several features using the Tuckey outlier method [7] which is a well-established univariate outlier detection method in ML projects.

  The above-mentioned method removed 3.5% of the rows from the entire dataset. We ran this method before data aggregation for modeling as well as on aggregated data sets to maximize the number of clean data rows. Since our training dataset was large enough for any ML model, we consider this totally feasible in this step.

- Null Value cleaning: One of the common issues with real-world datasets found in industries is missing entries of rows or columns. Null values can greatly impact optimization steps. One way to handle this is to eliminate those rows with missing entries. Alternatively, extrapolation methods can be used to estimate those data points as well.

  Although our initial EDA showed several null value flags, the features used in the final modeling, discussed later, did not have null values hence we did not truncate the original data set by the presence of any of these. Meaning, the final features were relatively cleaner compared to other features which also indicate the operational priority they hold above other features.

- Improbable Value Cleaning: Improbable values also negatively impacts data models. We cleaned for improbable values in the data cleaning stage as well. Two major improbable values found were zero sales value of a certain product and negative sales value. Both cases can be considered as improbable events given the domain we were working on. Furthermore, we cleaned for improbable positive numerical values for SKUs with a lower average value.

We have also concluded that as the data was received as a for-use data set with end result in mind, from the industry, it was pre-processed to a certain extent before we received it. Due to disclosure agreement concerns, we are not able to expand on this. However, this should be considered when trying to note the processability of such data.

4. Data pre-processing & Features Engineering

Pre-processing and feature engineering are the last key steps of ML modeling. In this step, data is converted into a structure that an ML model can take as an input-output shape and build a mathematical relationship between them.

First, we truncated the data set to a particular location as one location contributed to above 85% of total sales volume (Fig 3). This highly simplifies the model space and reduces extended features after encoding without sacrificing a major usability factor for the end-user. Our priority was on making accurate predictions in most cases.

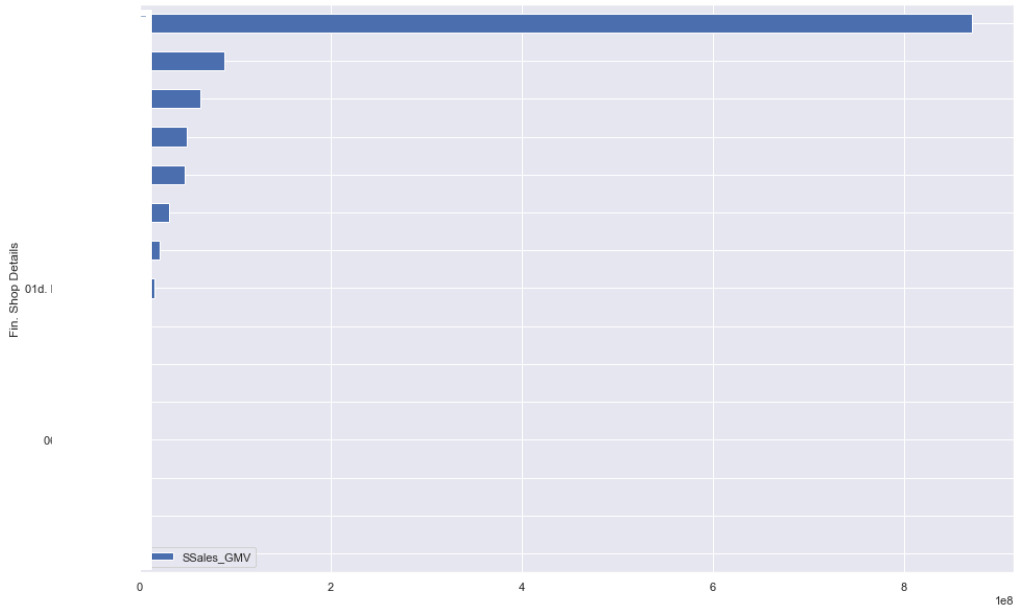


Fig 3: Histogram of location-wise sales value

Additionally, we did not include all available feature points for this modeling. We removed columns that did not impact sales such as color-coding of the product in inventory as well as columns that did not impact discounting decisions by the organization such as size.

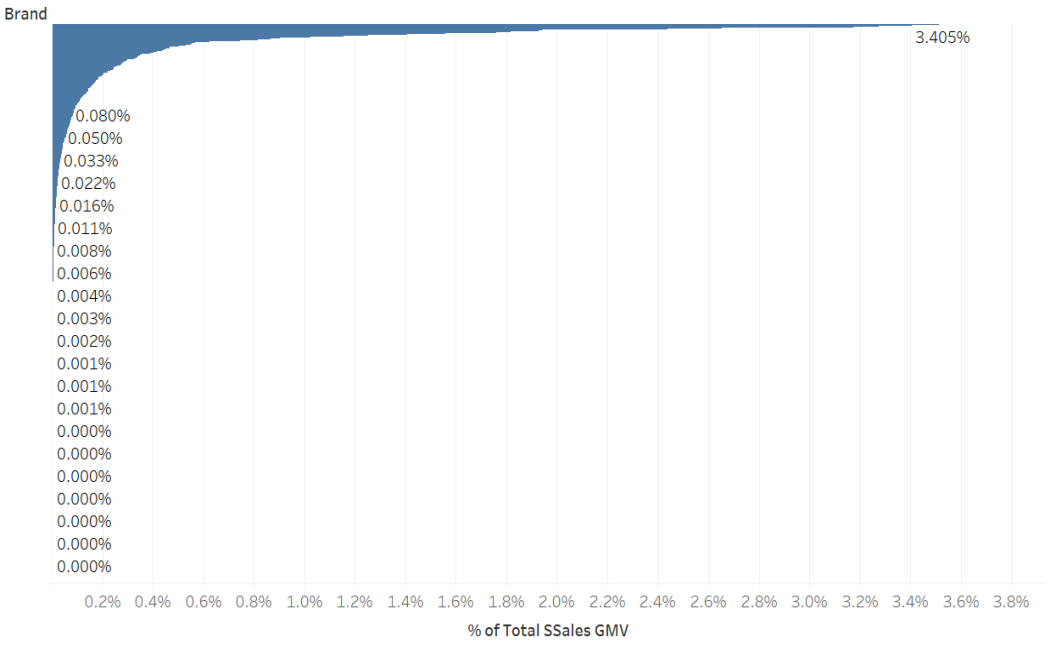


Fig 4: Histogram of sales contribution percentage by Brand

We only included brands that contribute to the top 80% of the sales by sales volume for simplicity (Fig 4, Brands names are truncated due to non-disclosure agreement). This ensures accurate predictability for most sale-volume contribution brands as well correct detections of elasticity for customers of this brand.

The grouping was done based on the years 2020-2021. Despite it being the part of the data set containing sales that are pandemic-influenced, it is the most recent data and therefore a better representation of the currently most impactful brands in this business. This is important to consider for end-user usability. Additionally, since the company had some sales volume in recent years, recent data provides a better view of brands.

Building on the understanding of the data and process of use from the previous steps we have the date, available stock, categorization, size, brand, color, gender, and location information available corresponding to sold products along with discount, time of sales, and sales value information for our final features list. The features we ultimately finalized for use in our model also have extracted time features.

The following tables show all the data columns with a brief description:

Table 1: Brief description of columns

Name	**Type**	**Definition**
Combined date	timestamp	Final product processing date
Category-1	Categorical	First level product categorization
Category-2	Categorical	Second level product categorization; under Category 1
Brand	Categorical	Brand og origin as the website is an aggregator
Gender	Categorical	Target customer gender assigned to the product and does not indicate final customer gender
Discount %	Numeric	Discount as a percentage of GMV for the SKU
Retail/OB/Closeout	Categorical	If a source of product is categorized as Retail or Own Brand or Closeout
Sales value (Y var.)	Numeric	Gross value accounting for discounted value as well

As the data set mostly consisted of categorical data, we dummy coded all our categorical variables and extracted our time variable into extracted time features of the year, month, and quarter and discount percentage to be modeled to predict the sales value. The data was aggregated to date level as we will be predicting average sales per day for the condition set by the features. The final aggregated data set then contained 200 features for model processing and above 1M entries of rows.

5. Data Modeling

The target variable “sales value” is a numeric variable and in our work, we predict the daily aggregate value of sales using previous data of sales at the SKU level. This approach makes it a regression type of problem in machine learning.

To achieve our goals, we fed the input variable into a model and train it using optimization methods.

For initial model exploration, we applied the models several times on several subsets of the data, grouped by date. This was primarily because of the time range used, 50% of the time range was a global pandemic that naturally affected buying patterns. This method would ensure that we detect anomalies in performance due to time training. We proceeded with models that performed equally well on all subsets.

We ran all of the following while including the Brand feature and while not including the Brand feature. This indicated a significant difference in data weight as well as prediction ability for the models besides the useability for the organization. All models were trained with an 80%-20% train-test split.

I. Linear Regression: We have used a linear regression model without any regularization first. The error of the linear model was very high and provided a low R-squared value between the predicted value and the original value on the test set. The R-squared value was 0.25 and the WMAPE error was 60%. We understand that as the dataset did not have a linear pattern, a linear model was not appropriate in this case. Additionally, results did not improve with regularization modelings as well. However, they are not recorded below as they had nearly equivalent results.

II. Support Vector Regression: Building on the performance of the linear regression model, we pursued the SVR model as such models fit a hyperplane. This would improve the probability of an accurate fit. This algorithm converged on earlier timeline subsets (2017-2018) but did not converge on all subsets and neither on consolidated data set despite several hyperparameter tuning iterations. We discuss relevant future work in the Discussions section below.

III. Gradient Boosting Regression: Based on the performance of the previous supervised algorithm we explored Gradient Boosting as it performs considerably accurately for a wide variety of constraint situations. Additionally, since it performs by reduction of error based on previous iteration errors, it has a higher probability to converge. Gradient boosting was able to perform the best of all algorithms with 35% WMAPE. We have graphically discussed these results in the results section.

IV. Random Forest Regression: Based on the performance of the previous ensemble algorithm, we explored a hyperparameter tuned random forest regression. It performed on par but not better than Gradient Boosting.

V. Artificial Neural Network: We also explored supervised algorithms, namely ANNs. With several iterations of hyperparameter tuning. With up to 8 layers and 100 nodes, we did not reach convergence. We did not pursue this further because the objective of this work is to reach a viable model for daily use and we reached acceptable results with the previous stage. Furthermore, with ANNs, it would have been difficult to provide explanations of how features interact with each other and feature importance of the model.

A tabular comparison of all the above results is in Table 2 below in the Results section.

Results

A lot of evaluation metrics have been used by researchers to check model performances of machine learning-based models [8]. We have used score, RMSE, MAE, WMAPE, and MdMAE as evaluation metrics to articulate our results and the reason for the errors in different segments of product-level data.

We discuss below the results of the hyperparameter-tuned Gradient Boosting algorithm as we chose this model for final predictions and inferences. The results are shown below to analyze the validity of proceeding with this model.

Gradient Boosting (Without Brand Feature)

We see from Fig. 5 and 6 below that predicted values are mostly very closely predicted to real values. Any considerable deviation starts at extreme values such as 6000 and above. The graph of residual values also has a normal distribution indicating sanity of the model prediction and assumptions.

However, some values have poor predictions which skewed the mean of the predictive model to 34%. It is evident that the model can predict values better when original daily sales of a certain category are between 1000-4000 units. When sales volume is more than 6000, the predictive power starts to decrease. We account this to discounts on some popular brands or a lack of promotional activities from other players in the market.

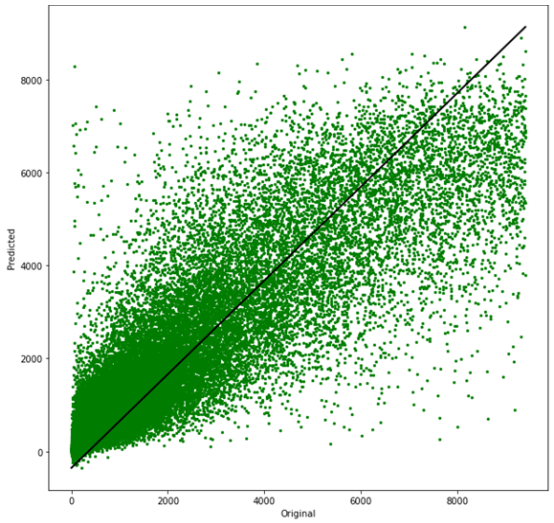


Fig 5: Scatter plot of Original vs Predicted Sales Values

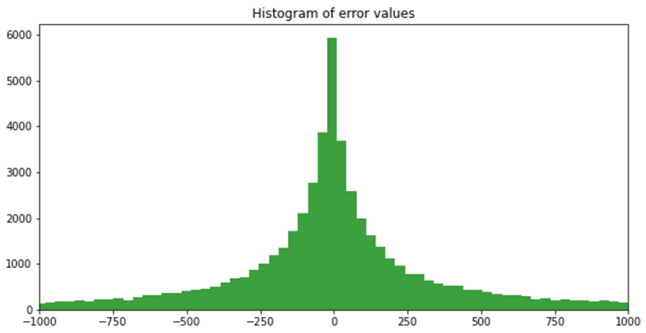


Fig 6: Histogram of error values

We see from Fig. 7 below that the Mean Absolute Percentage Error peaks near to 0 and is at a lower level for extreme values which also indicates the sanity of predicted values. This also indicates that there are a considerable amount of times when the model accuracy is better.

These are the points where the model is fairly accurate. This could be one of the possible use cases of this work in the retail industry.

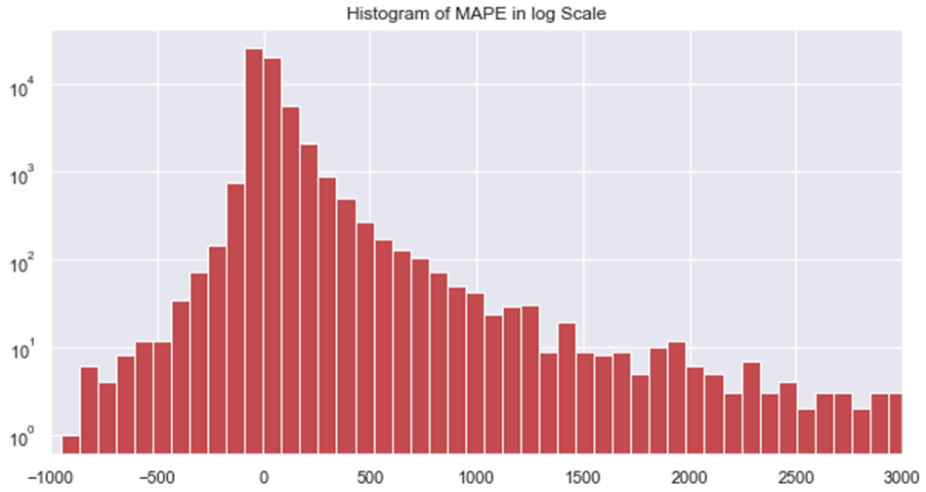


Fig 7: Histogram of MAPE in log scale

Fig 8: Sales vs Discount Percentage Scatter Plot - Retail and Own Business product types

We see above in Fig. 8 that there is also an inverse relationship between sales value and the discount percentage that we also found in the EDA stage. This automatically indicates a difficult inaccurate prediction for the model. This could also be attributed to the pandemic stage too. Such a non-linear relationship between discount and sales value makes it difficult for any linear regression model to predict these values accurately. However, we have shown that a non-linear model like gradient boosting can provide decent predictions in most cases.

In Fig 9. Below the predictions metrics for Retail: R-Sq: 0.78, WMAPE error is: 35%, prediction metrics for OB: R-Sq: 0.75, WMAPE error is: 42%. This, in relation to Fig 8 shows that despite this inverse relationship, the model predicts relatively well. The deviation starts at 6000 for retail and 3000 for OB. This we conclude to be an effect of a lower proportion of OB sales in the original data set.

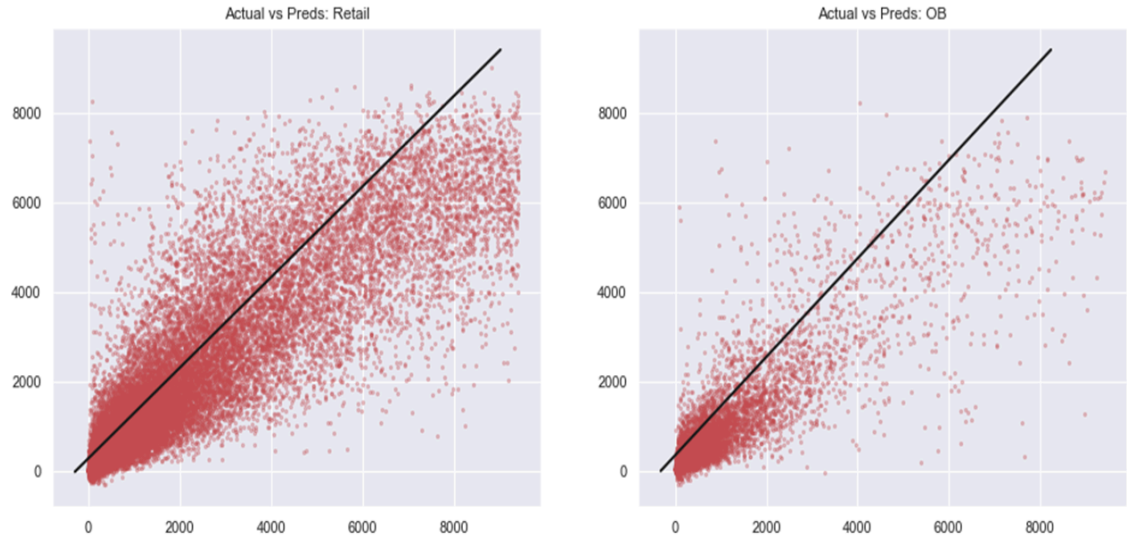


Fig 9: Actual vs Predicted value Scatter Plot - Retail and Own Business product types

In Fig. 10 and 11. Below, we see the same pattern (inverse relationship between sales value and the discount percentage that we also found in the EDA stage) for the Outdoor Bekleidung Category 1 type. The majority of products fall under this category by count for the website. Again, a non-linear model like gradient boosting can provide decent predictions in most cases. In the graph below, correlation score: 0.79, WMAPE: 35%. We analyzed similar results for other features as well.

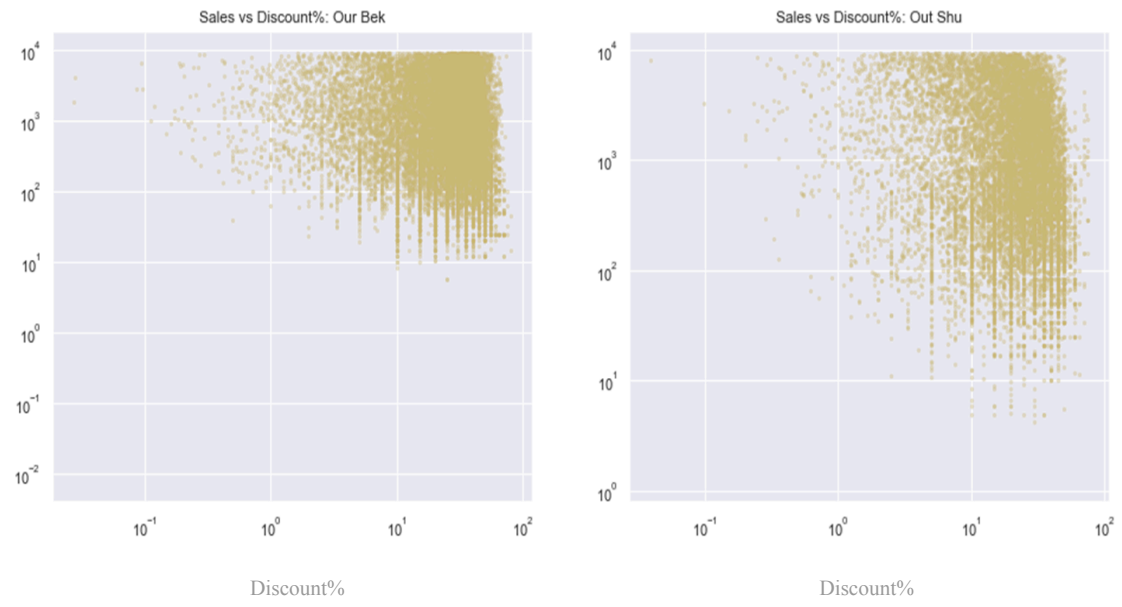


Fig 10: Sales vs Discount Percentage Scatter Plot - Outdoor Bekleidung Category 1 product type

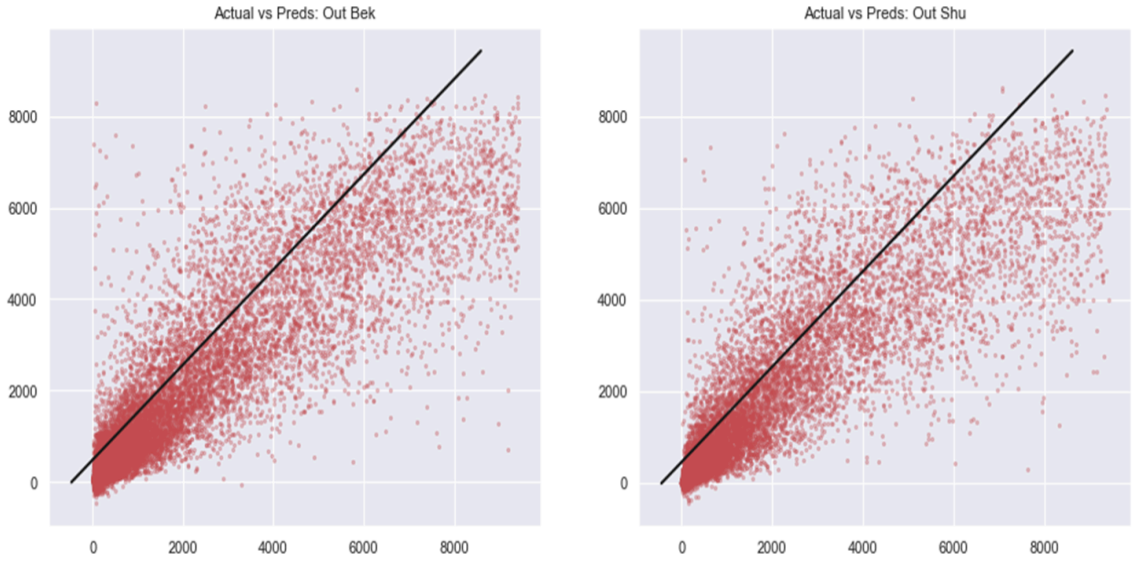
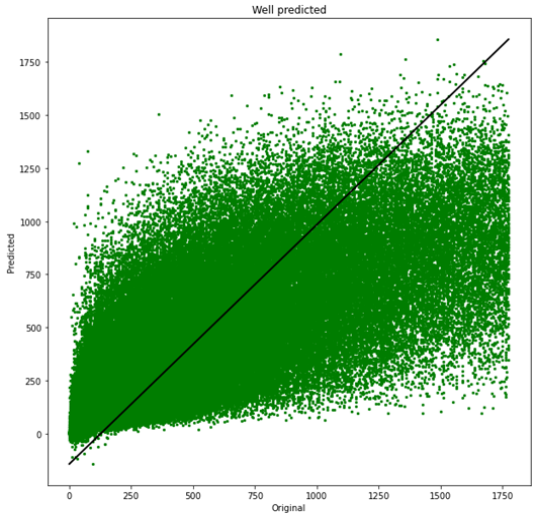


Fig 11: Actual vs Predicted value Scatter Plot - Outdoor Bekleidung Category 1 product type

Gradient Boosting (Without Brand Feature)

We see from Fig 12 and 13 below that predicted values are mostly very closely predicted to real values. Any considerable deviation starts at values such as 1000 and above. This is lower

performance than a model that did not include the brand feature however, it is still an acceptable performance considering the useability it offers. The graph of error values also has a normal distribution indicating the sanity of the model prediction and assumptions.



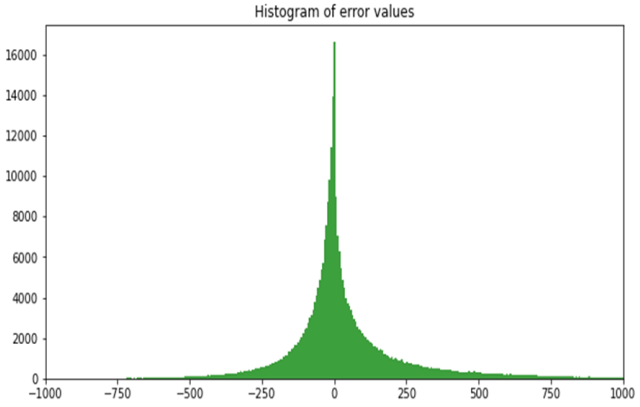


Fig 12: Scatter plot of Original vs Predicted Sales Values

Fig 13: Histogram of error values

We have summarised Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Weighted Mean Absolute Percentage Error (WMAPE), Median Absolute Percentage Error (MdAPE) in a table below. We have used multiple evaluation metrics to depict the fact that although RMSE is a normal distribution in all the cases, the WMAPE error is not. There are several categories of products where the error percentage is considerably high and that skews the results. Alternatively, we have not used MAPE error because the actual value of some product sales might be very low and predictions could be high which could have resulted in a significant error percentage further increasing the mean value.

Table 2: A table showing results of different models and designs:

'Brand' Feature	**Model**	**Score**	**RMSE**	**MAE**	**WMAPE**	**MdAPE**
Included	Linear Regression	0.23	312	221	71%	63%
	Random Forest Regression	0.64	212	130	42%	39%

	Gradient Boosting Regression	**0.62**	**216**	**128**	**41%**	**38%**
Excluded	Linear Regression	0.41	1483	1060	78%	76%
	Random Forest Regression	0.76	945	510	37%	38%
	Gradient Boosting Regression	**0.79**	**877**	**473**	**35%**	**36%**

We notice above that when brand is included in the RMSE and MAE decrease. This is due to the increase in data dimension and this does not indicate better performance of the model. Hence, looking at percentage error metrics would make more sense to the reader. However, WMAPE also does not increase a significant amount on the inclusion of Brand.

Discussion and Future Work

In an E-commerce business, understanding customer behavior and prediction of sales are crucial for efficient planning, logistics, business growth, and profitability. Although it seems like a simple time series problem, accurate prediction is often very difficult. Hence, it requires a lot of research as well as experiments.

In our above literature, it is evident that the ML model can not give exact predictions for all the cases. This is mostly due to some external factors as well as the quality of data. However, our research can be used as a baseline to build predictive tools for the retail industry or planning teams to support their day-to-day operations to support the traditional ways of forecasting and planning. Additionally, the exploratory visualizations created in the process can be used as dashboards to make more informed data-driven decisions by managers of online retail industries.

One of the major perturbations that we have observed in our work is in the data generated during COVID-19 lockdown time. There were fewer people buying from direct stores due to the lockdown and as a result so companies experienced unprecedented organic growth in customer acquisition and new orders. This phenomenon has greatly impacted the trends of companies in the online retail industry. Secondly, since we did not have any data on competitors, a counter discount offer from other companies might have disrupted the market to change the sales pattern of one company. Thirdly, some brands have less elasticity among users due to their popularity but we have not used any such information for building our models. This undermines the weight of their possible sales volume at a lower discount and all brands are equally treated as a feature after encoding. In future research, if we can overcome these ideas, we believe higher accuracy through the Machine Learning model is plausible. It is evident that sold product features need to be used along with customer attributes to gain an accurate prediction as customer elasticity is easier to map from such features.

We also observe that WMAPE also does not increase a significant amount on the inclusion of Brands. This indicates that we can include other features and possibly swap existing features to further improve the quality of the predictions. Another point of view that can be added here is the inclusion of brands further decreases the size of each segment. This indicates that machine learning models could not capture the patterns of such categories of products that

have a lower number of training examples. In future research, balancing each category will a good amount of training examples that can greatly improve predictive results.

We also suggest that with data sets that do not have major irregularities such as a pandemic influence (which here introduced an inverse pattern), support vector regression may be able to perform well as well.

In our work, we have shown that a hyperparameter-tuned gradient boosting model can predict daily sales at 35% WMAPE error. However, if we consider MdAPE error which is more accurate in this case due to the skewness of some products' sales volume being so high, that is also 36%. This is close to some previous research benchmarks of 16% and 24% - which were also done for the retail industry but on a monthly basis [5-6] and on a significantly less amount of features [2-3]. Furthermore, there are some product categories in our dataset where the absolute error percentage was less than 5% which is considerably better than any work in our research done on the online retail industry.

Our conclusion is that our currently chosen model is a reliably useable model for our end user but we have been able to identify scopes for furher work as mentioned above.

Acknowledgment

The authors acknowledge and thank the Methods Center at The University of Tuebingen and Bergfuende DE for guidance, constant support, and help throughout the research work.

References

1. Levy, M., Grewal, D., Kopalle, P. K., & Hess, J. D. (2004). *Emerging trends in retail pricing practice: implications for research. Journal of Retailing*, 80(3), xiii-xxi.

2. Carreira, A. N. D. A. R. (2017). *Retail forecasting under the influence of promotional discounts* (Doctoral dissertation).

3. Alon, I., Qi, M., & Sadowski, R. J. (2001). *Forecasting aggregate retail sales:: a comparison of artificial neural networks and traditional methods*. Journal of retailing and consumer services, 8(3), 147-156.

4. Osman, S., Fah, B. C. Y., & Foon, Y. S. (2011). *Simulation of sales promotions towards buying behavior among university students*. International Journal of Marketing Studies, 3(3), 78-88.

5. Thomassey, S., & Fiordaliso, A. (2006). *A hybrid sales forecasting system based on clustering and decision trees*. Decision Support Systems, 42(1), 408-421.

6. Ali, Ö. G., Sayın, S., Van Woensel, T., & Fransoo, J. (2009). *SKU demand forecasting in the presence of promotions*. Expert Systems with Applications, 36(10), 12340-12348.

7. Seo, S. (2006). *A review and comparison of methods for detecting outliers in univariate data sets* (Doctoral dissertation, University of Pittsburgh).

8. Hyndman, R. J., & Koehler, A. B. (2006). *Another look at measures of forecast accuracy*. International journal of forecasting, 22(4), 679-688.