2023

BIG DATA FOR MARKETING E-COMMERCE DATA

HUNTER'S E-GROCERY'S DATASET FOR PREDICTIVE MARKETING 2023

1.1 ABSTRACT

In today's digital age, data is growing at an unprecedented rate, creating both opportunities and challenges. The accumulation of structured and unstructured information from various sources is known as growing data. The objective of this project is to build predictive models for the supermarket that yield accurate results. These models will help predict whether customers are likely to make a reorder. The purpose of this prediction is to enable the supermarket to implement data-driven marketing and business strategies. The languages utilized in this project include Pyspark and SQL on Databricks. For machine learning, I evaluated Logistic Regression (LR), Random Forest (RF), and Gradient Boosting (GB) algorithms to make predictions, supported by descriptive analyses. Additionally, I conducted a chi-square test to determine the statistical significance of the null hypothesis.

Index terms **– Logistic Regression, Random Forest, Gradient Boosting, Imbalanced Data, Pipelines.**

1.2 INTRODUCTION

Hunter's e-grocery is a well-known French brand based in Brittany that offers a new generation of e-grocery and lifestyle products. With a presence in ten countries, Hunter's e-grocery continually seeks new ways to anticipate and meet customer demands, building strong loyalty among its customers. However, recent black swan events such as Covid-19, the Ukraine crisis, and gas shortages have impacted purchasing behavior, making it necessary to develop a business value proposition to increase customer lifetime value. By leveraging the power of big data analytics, I will explore customer behavior patterns and develop a predictive model to optimize marketing campaigns and reduce expenses. This project will be focusing on understanding customer behavior and estimate their customer retention so that more accurate strategies can be developed. The used data set is extracted from Kaggle. It is provided by the Hunter's to get proposals of business value for informative based decision making. The dataset consists of over 2 million purchase records at a renowned Hunter's supermarket.

Six hypotheses are formulated to investigate the impact of different factors on the probability of customer reorder. These hypotheses explore the influence of the day of the week, hour of the day, days since the previous order, department of the product, order of cart addition, and order number on reorder probability. Null and alternative hypotheses are formulated for each case to analyze their effects.

Big data has transformed marketing by providing marketers with access to vast amounts of data, advanced analytics capabilities, and actionable insights. It enables personalized marketing, real-time analytics, predictive modeling, and data-driven decision making, ultimately leading to more effective and efficient marketing strategies.

According to *Johnson, Williams, and Thompson (2019)*, who explored different predictive modeling techniques used in marketing, including LR, RF, and GB, these models have strengths and limitations (p. 351). The review article provides insights into how these models can be applied to predict customer behavior, enhance marketing campaign effectiveness, and optimize resource allocation.

Zhang et al. (2017) discuss the challenge of imbalanced data in predictive modeling for customer retention. They highlight the issue of disproportionate class distribution, which can lead to biased model performance, a situation that aligns with our study. The authors propose strategies such as oversampling, undersampling, and ensemble techniques to address imbalanced datasets and enhance the accuracy of predictive models, which I will be implementing in our research.

1.3 METHODOLOGY

The dataset used for analyzing supermarket consumer behavior comprises 2,019,501 rows and 12 columns. The columns include the following variables: "order_id" (unique identifier for each order), "user_id" (unique identifier for each user), "order_number" (the order number), "order_dow" (the day of the week the order was made), "order_hour_of_day" (the time of the order), days_since_prior_order" (the number of days since the previous order), "product_id" (the identifier for each product), "add_to_cart_order" (the number of items added to the cart), "department_id" (a unique identifier

for each department), "department" (the name of each department), "product_name" (the name of each product). Additionally, the dataset includes a binary indicator in the "reordered" column, which denotes whether a particular product has been ordered by the user in the past.

I have checked our dataset for duplicates initially which returned us 0 rows. "days_since_prior_order" is the only column that has 124,342 missing values. These missing values represent new customers. The reordered column is always 0 where the day_since_prior order column is -1. Therefore, it does not provide valuable information for predictive modeling.

The “department” and “product_name” columns have some entries as “missing” and the “department_id” column are always “21” in the corresponding rows of those. Therefore, I have dropped these 4,749 rows since they do not give valuable information.

I checked outliers in ”add_to_cart_order” and “order_number” columns that resulted in 61,823 and 109,659 outliers respectively. I have also discovered that the outliers are populated near each other therefore, I preferred not dropping the outliers since it is quite a large amount of data however I will be diluting them by logarithmic transformation.

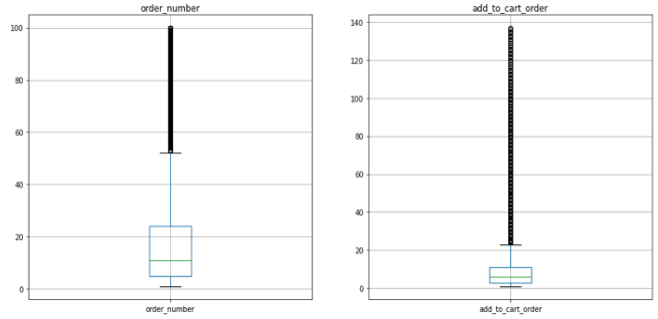

Fig. 1. Outliers in”add_to_cart_order” and “order_number” columns

“order_dow” column has values from 0 to 6 where 6 represents Sunday. I transformed these values into corresponding day names to increase the readability in the graphs. And, I dropped the “user_id” column since I did not have a use for it. Also, I created a new feature named time_slot combining the order_dow and order_hour_of_day columns to use in our predictive models in order to see if it will affect the importance of features in models. Further, I will compare features and get the best features for our models.

For this study, I selected three predictive modeling algorithms: LR, RF, and GB, to predict customer reorder behavior in the supermarket dataset.

To assess the performance of our predictive models, I employed several commonly used performance metrics.

First, the Classification Report provided insights into precision, recall, and F1-score for each class (reorder vs. non-reorder). These metrics helped evaluate the model's ability to correctly identify positive and negative instances and provided a comprehensive assessment of its predictive performance.

Receiver Operating Characteristics(ROC) and Area Under the Curve (AUC) analyses assessed the models' ability to distinguish between reorder and non-reorder instances. The ROC curve visualized the trade-off between true positive rate and false positive rate, and the AUC represented the overall discriminative power of the model.

Cross-Validation was employed to estimate the generalizability of the models. This technique involved splitting the dataset into multiple subsets, training the models on one subset, and evaluating their performance on the remaining subsets. Cross-validation provided an estimate of how well the models would perform on unseen data and helped mitigate overfitting.

1.4 FINDINGS AND DISCUSSION

In the summary statistics, it is observed that the average order number is 17.15. This indicates that there is a significant amount of repeat ordering by users. Orders are spread fairly evenly across the days of the week, with a mean value of 2.74.

The average number of days between orders is 11.39, with a standard deviation of 8.97. This indicates that while there is a significant amount of repeat ordering, the frequency of orders varies widely among users.

To evaluate the distribution of all the columns, I have created subplots and I have discovered that there is a positive skewness in only “order_number” and “add_to_cart_order” columns as shown below:

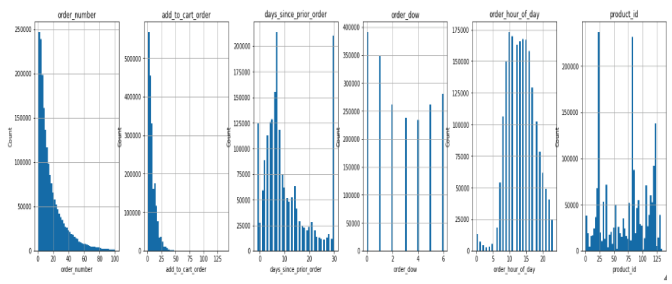

Fig. 2. Distribution of values

Considering that the outliers existing in these columns are not removed, I decided to conduct a logarithmic transformation to reduce the impact of outliers and to normalize the data. Skewness for ”order_number“ is decreased from 1.78 to 0.29. And, skewness for “add_to_cart_order” decreased from 1.89 and -0.318. After the transformation, corresponding columns are named “order_number_log”, ‘add_to_cart_order_log ’.

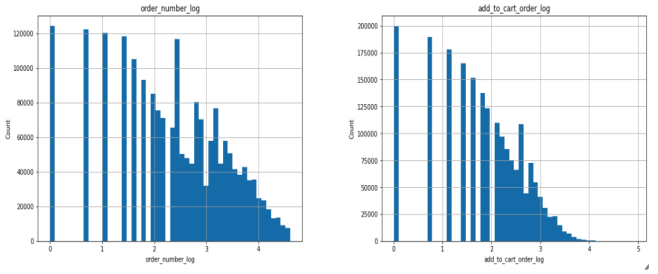

Fig. 3. order_number and add_to_cart_order after the log transformation

Looking at the distribution of the number of purchases by day and hour below, Monday and Sunday were the days that had the most purchases and from 10 a.m. to 4 p.m. hourly whereas the least purchases were made between 2-5 a.m. on Wednesday and Thursday.

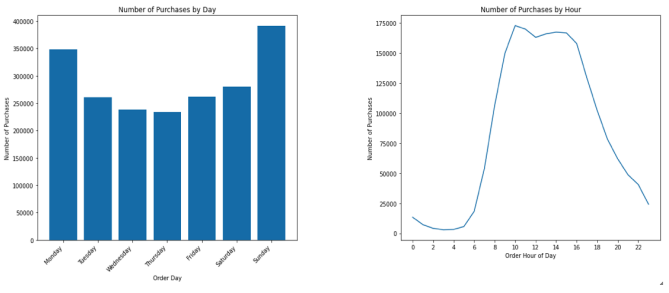

Fig. 4. Distribution of purchases by day and hour

The most purchased products are fresh fruits, fresh vegetables and packaged vegetables fruits whereas the least ones were bread, soy lactose free and chips pretzels. The department that had most purchases is produce, dairy eggs and snacks and the least is dry goods pasta, deli and canned goods.

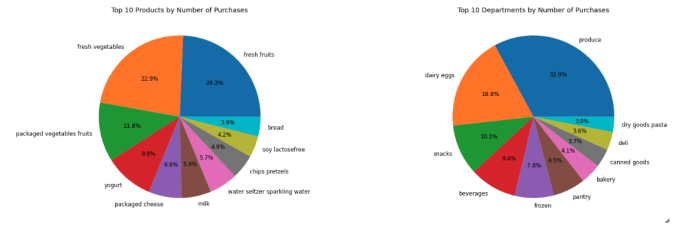

Fig. 5. Top 10 products and departments by number of purchases

In association rules, when sorted by support level, it was observed that the combination of "packaged vegetables fruits" and "fresh vegetables" appeared most frequently, indicating that these items are commonly purchased together in approximately 18.658% of the transactions. This association exhibited a confidence level of 69.1% and a lift value of 1.556, indicating a positive relationship between the items.

On the other hand, when sorted by confidence level, the combination of "condiments", "canned jarred vegetables", "fresh herbs", "packaged vegetables fruits" and "fresh fruits" demonstrated a confidence level of 97.5%, indicating a strong likelihood of customers also purchasing "fresh vegetables" alongside these items.

Frequent items provide insights into popular choices, while association rules highlight specific item associations that can be leveraged for targeted marketing, cross-selling, or store layout optimization.

1.5 RESULTS

I evaluated the performance of three machine learning models by making predictions on a balanced testing dataset using pipelines and calculated the area under the ROC curve and AUC as a metric with visualizations. The AUC values obtained were 0.706 for LR, 0.705 for RF, and 0.718 for GB. Among the results I will take the two best models LR and GB, and I will perform hyperparameter tuning using cross-validation and grid search.

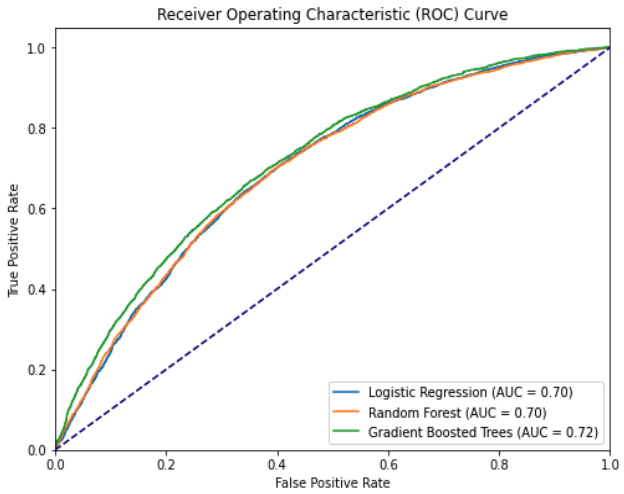

Fig. 6. Comparison of models LR, RF and GB

With the best parameters, the results are as below:

LR: Elapsed time: 1445.459469795227 seconds and Best LR AUC: 0.7026129570756284

GB: Elapsed time: 9834.941793680191 second and Best GB AUC: 0.7215015036282618

In the classification report, the best performing model was GB, which achieved an accuracy of 0.6601. It outperformed the LR model, which had an accuracy of 0.6458. Both models showed similar precision, with GB achieving slightly higher recall and F1-Score values, indicating its ability to better balance between identifying positive cases and minimizing false negatives.

Features and Coefficients Importance Extraction:

In both LR and GB, the most important feature is "order_number_log", indicating a strong positive impact. Other important features in LR include "order_hour_of_day", "days_since_prior_order", and "department_id", which also contribute positively. The "order_dow" feature has a slightly negative impact.

In GB, the importance of features aligns with LR, with "order_number_log", "department_id", and "add_to_cart_order_log" being the most important. The "days_since_prior_order" and "order_hour_of_day" features have relatively lower importance, and "order_dow" has the least impact. Time slot was excluded from features since it doesn't have significant impact among the features and the results.

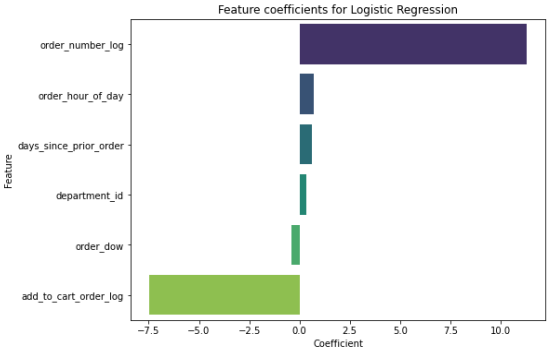

Fig. 7. Feature importances for LR

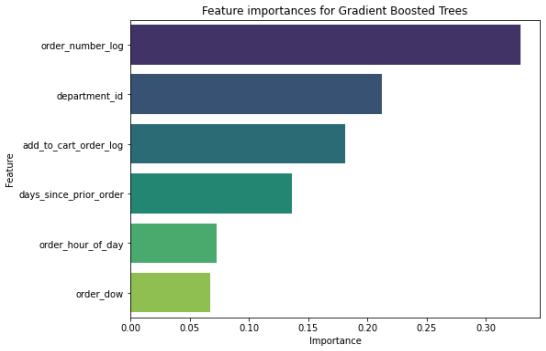

Fig. 8. Feature importances for GB

Feature Selection Experiment:

I have a list of different numbers of features to select, defined by the variable num_features_to_select. This process allows us to compare the performance of LR and GB models with different numbers of selected features and determine the impact of feature selection on the models' predictive accuracy. The printed AUC scores provide insights into the models' performance using various feature subsets, aiding in the analysis and decision-making process. The experiment suggests us to input all the features for best results out of the dataset.

```
AUC using 2 selected features for Logistic Regression: 0.620
AUC using 2 selected features for Gradient Boosted Trees: 0.656

AUC using 3 selected features for Logistic Regression: 0.692
AUC using 3 selected features for Gradient Boosted Trees: 0.712

AUC using 4 selected features for Logistic Regression: 0.696
AUC using 4 selected features for Gradient Boosted Trees: 0.714

AUC using 5 selected features for Logistic Regression: 0.708
AUC using 5 selected features for Gradient Boosted Trees: 0.717
```

Fig. 9. Comparison of models with selected features

Hypothesis Testing Results:

```
Summary of test:
+-------+----------------+-------------------+
|pValues|degreesOfFreedom|statistics         |
+-------+----------------+-------------------+
|[0.0]  |[98]            |[200117.5381761258]|
+-------+----------------+-------------------+
```

Fig. 10. Hypothesis testing results for order_number

```
Summary of test:
+-------+----------------+-------------------+
|pValues|degreesOfFreedom|statistics         |
+-------+----------------+-------------------+
|[0.0]  |[136]           |[45465.51184043615]|
+-------+----------------+-------------------+
```

Fig. 11. Hypothesis testing results for add_to_cart_order

Since the p-values are less than the significance level (typically 0.05), I can reject the null hypotheses for all the variables. This means that there is evidence to suggest that each of these variables has a statistically significant association with the probability of a customer making a reorder.

Therefore, I can conclude the following based on the results:

department_id, order_number, add_to_cart_order order_hour_of_day, day_since_prior_order and, order_day_of_week have a significant impact on the probability of a customer making a reorder, accepting the alternative hypotheses for all the independent variables as a result of the chi-square analysis.

1.6 CONCLUSION

This study analyzed a dataset containing over 2 million orders, investigating patterns and associations in customers' ordering behavior. Machine learning models were employed to predict reorder probabilities, with GB demonstrating superior performance compared to LR and RF. Hyperparameter tuning further slightly improved the models' accuracy. Features such as order_number, add_to_cart_order, and department_id had notable impacts on reorder probabilities. Additionally, hypothesis testing confirmed the statistical significance of all the independent variables selected.

The results provide valuable insights for the business, enabling targeted marketing, cross-selling strategies, and store layout optimization. Several marketing strategies can be suggested to leverage the insights gained from the analysis of customer ordering behavior:

Personalized Recommendations: Provide personalized product recommendations based on common product combinations to enhance cross-selling opportunities.

Promotions and Discounts: Focus promotions on popular products like fresh fruits and vegetables, and offer discounts during off-peak hours to stimulate sales.

Enhance Store Layout: Optimize store layout by allocating more space to high-demand departments and arranging related products together for cross-category buying.

Seamless Reordering Experience: Simplify the reordering process with features like one-click reorder and order reminders to encourage frequent purchases.

Social Media Engagement: Use social media platforms to engage customers, share recipe ideas, and encourage user-generated content to generate brand awareness and word-of-mouth marketing.

Future research can focus on exploring additional factors influencing customer ordering behavior to further enhance predictive models.

REFERENCES

Johnson, A., Williams, B., & Thompson, S. (2019). Predictive modeling in marketing: A review. Journal of Marketing Analytics, 7(4), 349-373.

Zhang, Y., Zhang, Q., Zhu, H., & Yang, J. (2017). Imbalanced data in customer retention predictive modeling. Expert Systems with Applications, 73, 221-230.