Leveraging Data Analytics in FoodMart supermarket for Customer Segmentation and Sales Forecasting

Marcus Romedahl
University of Colorado Boulder
Boulder, United States
maro7104@colorado.edu

Ramy Kassam
University of Colorado Boulder
Boulder, United States
raka0023@colorado.edu

Raleigh Darcy
University of Colorado Boulder
Boulder, United States
rada2150@colorado.edu

Figure 1: Photo by nrd on Unsplash

ABSTRACT

In the age of data-driven decision-making, the Foodmart dataset emerges as a reservoir of comprehensive information encompassing metrics such as workforce size, advertisement budget, gross profit, and store locations. This repository holds immense value for both data analysts and business strategists, offering a rich tapestry for exploration. This research initiative undertakes a systematic exploration to harness the potential of data analytics in unveiling latent patterns and trends within the Foodmart dataset. Employing advanced methodologies in data preprocessing, feature engineering, and machine learning, our objective is to extract actionable insights capable of informing strategic decisions within the retail sector.

The principal aim of this study is to discern high-level features that exert discernible impacts, both positive and negative, on a retail store's gross profit. Through this investigation, we aim to address critical inquiries:

What pivotal expenditures contribute positively or negatively to gross profit? and In what ways does the element of time, encapsulating variables such as store age, manager age and experience, staff number, wages, and opening hours, influence gross profit and sales?

Embarking on this data analytics journey, we anticipate unearthing valuable insights poised to enhance the retail grocery industry. The overarching goal is to steer data-driven decisions that not only optimize profitability but also elevate customer satisfaction in this dynamic and competitive landscape. The principal aim of this study is to discern high-level features that exert discernible impacts, both positive and negative, on a retail store's gross profit. Through this investigation, we aim to address critical inquiries: This research initiative undertakes a systematic exploration to harness the potential of data analytics in unveiling latent patterns and trends within the Foodmart dataset. Employing advanced methodologies in data preprocessing, feature engineering, and machine learning, our objective is to extract actionable insights capable of informing strategic decisions within the retail sector.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Marcus Romedahl, Ramy Kassam, and Raleigh Darcy. 2023. Leveraging Data Analytics in FoodMart supermarket for Customer Segmentation and Sales Forecasting. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROBLEM FORMULATION

In the dynamic landscape of the retail grocery industry, data has emerged as a formidable force, shaping the way businesses operate, strategize, and serve their customers. The Foodmart dataset, an extensive repository of high level information such as number of workers, advertisement budget, gross profit, store locations, etc., offers a unique opportunity to harness the power of data analytics for profound business transformation. The dataset at hand, sourced from Kaggle[1], presents us with such an opportunity, a canvas upon which we shall paint a portrait of knowledge and discovery.

The grocery store of today is no longer a mere physical entity but a multifaceted ecosystem where technology, consumer behavior, and data intertwine. Shoppers, armed with smartphones and wearables, traverse both the physical and digital realms, leaving behind digital footprints that tell stories of their preferences, habits, and aspirations. From the choice of cereal brands to the selection of produce, every decision becomes a data point, every interaction a clue.

The motivation behind this project transcends mere curiosity; it stems from the recognition that data analytics is no longer a peripheral endeavor but a strategic imperative for success in the modern grocery landscape. Retailers and grocers are increasingly realizing that their survival and prosperity depend on their ability to harness the power of data to inform and optimize every facet of their operations.

Our aspiration is to unlock the latent value concealed within the Foodmart dataset. Through the lens of data analytics and machine learning, we aim to extract meaningful patterns, illuminate hidden correlations, and forecast future trends. By doing so, we endeavor to empower retailers and grocers with the insights needed to navigate this evolving landscape successfully.

2 DATA PREPROCESSING

2.1 Missing Values

The first step taken after choosing the data set to work with was to check for any missing values. This is done because few machine learning algorithms handle missing values that well and missing values might introduce bias in the analysis. Any missing values would therefore have to be handled to ensure the performance of the model created. Different data can be handled in different ways, what way the missing data is handled comes down to the context of the data and the objective function. A few ways this could be done is: Deletion, removing the rows or columns that contain missing data; Imputation, filling in the missing values with help of for example mean or median; Prediction, try to predict the missing values with statistical models like k-Nearest Neighbors. The FoodMart Dataset did not contain any missing values so no operations had to be done to handle that.

2.2 Feature Handling

After concluding that there were no missing values in the data set the next step was to examine the features of the data set. All categorical values exist in two columns, one with text and one with whole numbers representing the category they belong to. This makes half of these columns redundant. The text version was removed since it's easier for models to work with numbers. The data set contained two features representing the basket of food items in each store, one from 2013 and one from 2014, to reduce the dimensionality these two where combined using the mean to a new feature called "Basket" that was used instead. In the data set there were two features called "Wages $m", that displayed the total cost of all wages during the year at that store, and "No. Staff" that displayed the number of full-time staff at the store. These two features were combined to create "WagePerStaff" which was calculated as $WagePerStaff = \frac{Wages\$m}{No.Staff}$. WagePerStaff was created since it's a more comparable feature between stores of different sizes. For the target feature, both "GrossProfit", which contains the stores gross profit for the year, and "Sales $m", which contains the stores total sales revenue for the year, were considered. In the end these two features were combined into "gross_profit_ratio" which is calculated as $gross\_profit\_ratio = \frac{GrossProfit}{Sales\$m}$ and is a common measurement of a business profitability. (Biswas, 2022)

After these feature manipulations had been performed the data set was split into two, namely numerical and categorical data. The numerical data set contained the following features; 'gross_profit_ratio' ,'WagePerStaff', "Adv.$'000", 'Competitors','HrsTrading', 'Mng-Age', 'Mng-Exp','Mng-Train', 'Union%', 'Car Spaces','Basket','Age (Yrs)', and the categorical data set contained the following features; "Loc'n (Num)", 'State (Num)', 'Sundays (Num)','Mng-Sex (Num)", 'HomeDel (Num)'. At this stage this was mostly done as a selection of potentially relevant features for the objective function since these two data sets were instantly merged.

2.3 Outlier Detection

Following the completion of feature manipulation, a critical phase involved the detection and removal of outliers within the data set. Effectively managing outliers is imperative, as they can significantly impact the statistical measures and data distribution, potentially leading to misleading results. To comprehensively identify outliers, a dual approach was employed, utilizing both z-scores and quartiles.

The z-score was applied to the data set with the assumption of a Gaussian distribution of the data. Every data point with a higher z-score than 3 or lower than -3 was considered an outlier and was removed from the data set. Additionally, quartiles were employed to assess the distribution of the data. Any data points located outside the interquartile range (IQR) were also flagged as outliers and systematically eliminated.

By incorporating both z-scores and quartiles, this outlier detection strategy aims to enhance the robustness of the data analysis, ensuring a more accurate representation of the underlying patterns and trends in the dataset.

2.4 Normalization

After the outliers had been taken care of the data was split into two different data sets, one numerical and one catagorical data set, in the same way as in chapter 2.2. The numerical data was then normalized using MinMax scaling. MinMax scaling scales the data according to the following equation:

$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$

The way it was applied in this project it scales the values between [0,1], but it could be used to scale values between [-1,1]. The reason MinMax scaling is used is because some machine learning algorithms are sensitive to the scale of the data. MinMax scaling also keeps the shape of the distribution of the data. This means that the scaling does not reduce the importance of outliers and for MinMax to be useful the data set should not contain any outliers, as they can skew the scaled values of other data points. Therefor it is important to not skip the outlier detection step that is presented in chapter 2.3.

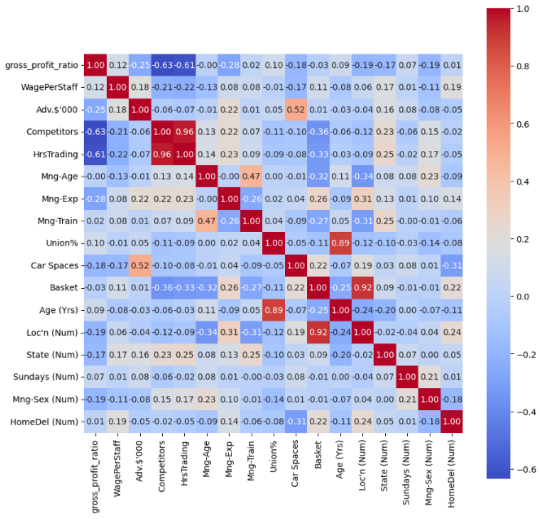

Figure 2: Heatmap of the correlation between features in the data set

2.5 Feature Selection

After the numerical values had been normalized the two data sets were merged into one data set again. The merged data set was used to find correlations and multicollinearity between features. By removing highly correlated features (that is not correlated with the target feature) and features with high multicollinearity we reduce the risk of overfitting the model. To find the correlations between features a correlation matrix was created, from which a heatmap as seen in figure 2 was generated.

Not only was a correlation analysis done, but also a collinearity analysis using VIF. VIF stand for the Variance Inflation Factor and provides a measure of multicollinearity in a set of multiple regression variables. The VIF analysis is done on all features except the target feature and returns a VIF score. If the VIF score is high it means that the feature is highly correlated with other features and can be predicted by the other features. These features can be removed since they do not provide unique or independent information to the model. Highly correlated features introduce noise to the model and risk overfitting.

From the first round of correlation and collinearity analysis the features HrsTrading, Basket and Mng-Age was removed from the data set. HrsTrading was removed because it was highly correlated with Competitors. Both features could be argued to be relevant for the objective function. HrsTrading was removed because it had the highest VIF of about 55. Basket and Loc'n was also highly correlated. Taking the objective function into consideration it could be argued that location is more important, since it's easier to affect the location of a store than the basket. This is the reason why Loc'n was not removed even though it had a higher VIF. In contrast to the two earlier features, Mng-Age was removed because it showed now correlation what so ever with the target feature.

The feature selection step is a recursive process since the removal of one feature will effect the multicollinearity of all features in the data set. Because of this the process was redone a few more times and by the end Union% and WagePerStaff had

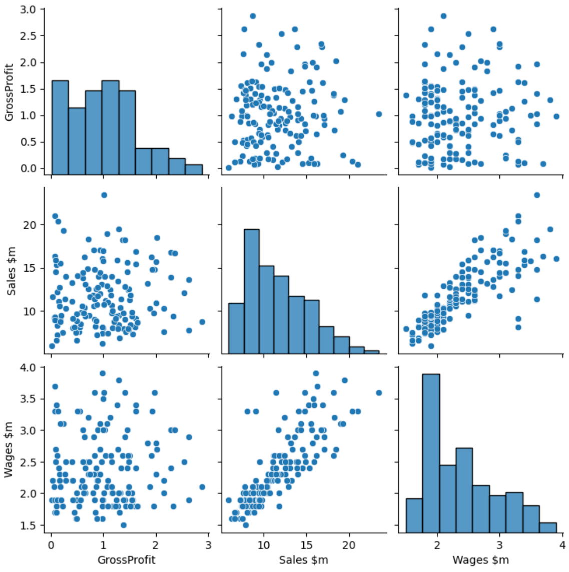

Figure 3: Multivariate linear regression model

also been removed. Union% was removed because Union% and Age (yrs) had a high correlation and Union% had a much higher VIF. WagePerStaff, Location and Advertisment all had high VIFs, through testing it was concluded that dropping WagePerStaff had the biggest effect on reduction of multicollinearity for the other features so WagePerStaff was dropped. Some features still had what could be considered a high VIF of around $7-8$, but these features was kept to not reduce the features in the data set to a dimension where the model becomes unuseable in real life.

Moreover, for the identified locations with attributes like wages, staff numbers, sales, and delivery option were scrutinized to understand their potential significance. This qualitative analysis complemented the quantitative approaches of correlation analysis and VIF, providing a more holistic view of the data set.

This study investigates four key aspects across different stores to discern relationships and make comparisons. Store 34, positioned in a country, is examined for having the highest wages at 3.9. Store 65, also in a country, stands out for achieving the highest sales at 23.5. Store 44, situated in a strip, is scrutinized for managing the maximum number of staff at 117. Finally, Store 52, also located in a strip, is analyzed for achieving the maximum gross profit at 2.872. By examining these distinct aspects—highest wages, highest sales, maximum staff, and maximum gross profit—across diverse store locations, this study aims to identify patterns, correlations, or notable differences that contribute to a comprehensive understanding of store performance. Such insights could potentially inform strategic decision-making and optimization strategies for retail operations.

This exploration was aligned with the recursive nature of the feature selection process. As features were removed, we ensured that the model retained essential attributes influencing gross profit and sales. The trade off between reducing multicollinearity and preserving key features was carefully considered, striking a balance that ensures model usability in real-life scenarios.

Store	Loc.	Sales	Max Staff	Wages	Profit
34	Country	16.1	98	3.9	0.974
65	Country	23.5	89	3.6	1.018
44	Strip	15.3	117	3.4	0.104
52	Strip	8.7	53	2.1	2.872

Table 1: Store Information

In other words, our feature selection process went beyond statistical analyses, incorporating a qualitative exploration of different attributes to uncover their relationship with gross profit and sales, thus contributing to a more robust model.

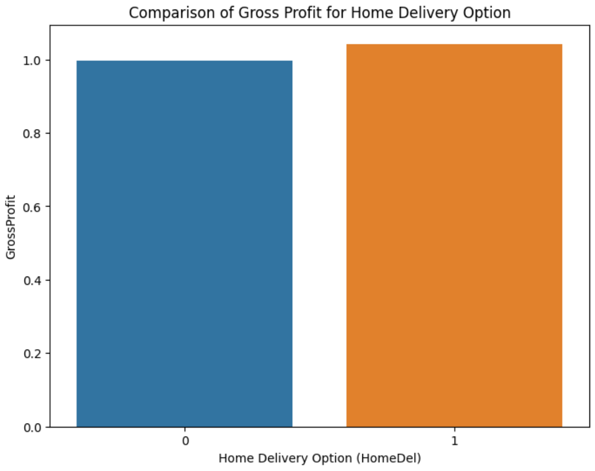

Figure 4: Delivery Analysis

3 METHODOLOGY

When the preprossecing mentioned in chapter 2 was completed a model had to be selected to answer the objective function. The data set was split into a training and test set. The training set made up 80% of the total data set and the test set made up 20% of the total data set. To find the most fitting model a k-fold crossvalidation was made on the training set, where $k = 5$. The modals linear regression, decision tree regressor, random forest regressor, XGB regressor and SVR was chosen because all of them can analyse feature importance in some way, which is important to answer the objective function.

Linear regression returns a coefficient of each feature that can be interpreted as the change the target variable will receive for each unit change in the feature analysed. This modal assumes that the relationship between each feature and the target variable is linear, which it might not be. It also does not account for interactions between features. Decision trees assign an importance score to each feature based on the total reduction of the criterion brought by that feature. In contrast to the linear regression this method does not show how the change in a particular feature will effect the target variable, only what features are most important for the target features value. Random forest is an ensemble of decision trees. The feature importance scores are averaged across all trees to give more robust estimates. XGBoost uses a similar method as random forests but also considers the number of times a feature appears in the trees. SVR doesn't directly provide feature importance. However, in the linear case, the weight coefficients can give an idea of feature importance.

Thanks to the k-fold crossvalidation we can figure out what model fits our data set the best. This is done by calculating the mean squared error (MSE), mean absolute error (MAE) and R-squared values of all the model performances. The result rounded to five decimal places is presented in table 2.

The MSE is the average of the squared differences between the predicted and actual values. This results in higher errors having a bigger impact on the result. MAE is the average of the absolute differences between the predicted and actual values and is therefore less sensitive to outliers compared to MSE. Both of these measure error and should therefore be as low as possible. R^2 measure how well the model's prediction fit the data. R^2 ranges from 0 to 1, where 1 indicates a perfect fit. Therefore R^2 should be as high as possible, however if the R^2 value is to high it might indicate that the model has been overfitted to the data.

After examining the MSE, MAE and R^2 values of all models the conclusion was drawn that random forest is the best choice for the data and objective function. This conclusion was drawn since random forest has the lowest MSE, second lowest MAE and highest R^2 value out of all the tested models.

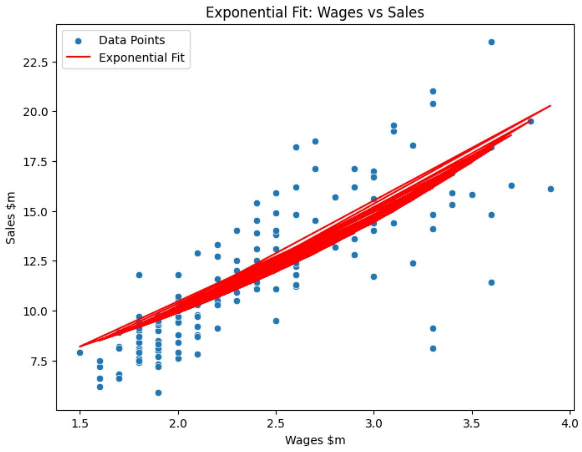

Figure 5: Exponential curve

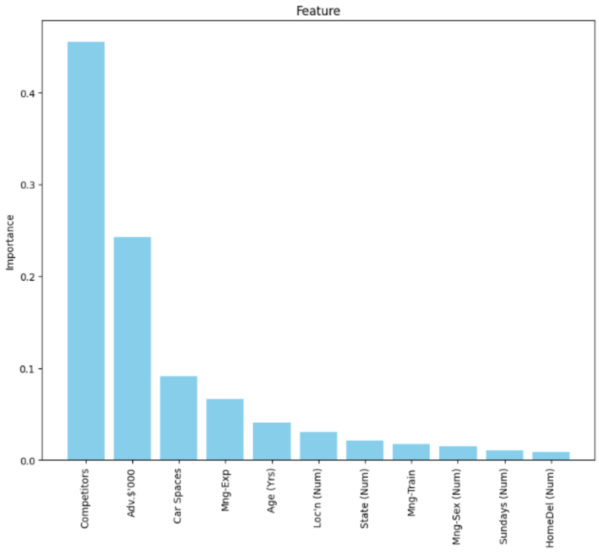

Figure 6: Feature importance

4 RESULTS AND ANALYSIS

In our data analysis, we adopt an approach to explore specific relationships and ascertain the impact of independent variables on both profits and sales. Notably, the examination of gross profit in isolation may be influenced by unaccounted factors or undisclosed expenses not present in the dataset. Instances in which sales and profits do not exhibit proportionality suggest that heightened sales do not necessarily guarantee elevated profits. However, a discernible correlation emerges between wages and sales, indicating that a higher wage structure is likely to correlate with increased sales figures.

This preliminary analysis underscores the importance of meticulously identifying and evaluating each attribute, serving as a foundational step in elucidating these intricate relationships. Subsequent to this initial exploration, a more comprehensive analysis involving diverse calculations and ratios will be conducted to discern optimal courses of action for retail operations, ultimately contributing to informed decision-making and strategic planning.

After concluding that random forest was the best choice for the chosen data set and objective function it was used to predict the gross profit ratio (GPR) of the test set. The predictions can be compared in table 3.

The feature importance is presented in figure 6. In the figure we can see that the number of competitors in the area is the feature that has the greatest effect on the GPR. Competitors is almost twice as important as the second most important feature, advertisement budget.

To know what value to choose for what feature to get the highest possible GPR we need to know how different choices of the values affect GPR. To understand how the choice of numerical values affect the GPR a correlation analysis was made on the data set. The coefficients and what that means for the choice of the feature value is presented in table 4. To understand what choice to make when it comes to the categorical data the categorical values were grouped and the average GPR was produced for each group. Most of the values importance pale in comparison to the two most important features. How the choice of these values affect the GPR will still be presented, but there is really no need to conider them. The calculation for location can be seen in table 5, from this it can be concluded that it's slightly better to be at a Strip or a Mall then in the Country. The calculation for which state is the best choice can be seen in table 7, from this it can be concluded that Queensland is the most profitable state, followed by Victoria and Western Australia as good choices as well. The worst state by far is Australian Capital Territory. This feature is most useful if you want to locate your store in Australia, but it suggests that state or region might have some implication on the GPR even in different countries. However from the feature importance it becomes clear that what state the store is located in does not matter much at all. The same is true for all categorical features. The rest of them is presented in the tables 8 to 10 but no further analysis will be made for them because of their insignificance.

REFERENCES

Biswas, S. (2022) 'Gross Profit Ratio: Meaning, Calculation, Formula and Significance', ClearTax. Available at: https://cleartax.in/s/gross-profit-ratio (Accessed: 'Date').

Table 2: Result of k-fold crossvalidation

	MSE	MAE	R^2
Linear Regression	0.02490(+/−0.00343)	0.11877(+/−0.00565)	0.48424(+/−0.08649)
Decision Tree	0.02792(+/−0.00408)	0.12830(+/−0.00564)	0.41550(+/−0.12649)
Random Forest	0.02322(+/−0.00418)	0.11959(+/−0.00857)	0.5260(+/−0.03761)
XGBoost	0.02926(+/−0.01007)	0.13575(+/−0.01987)	0.41338(+/−0.12058)
Support Vector Machines	0.02674(+/−0.00813)	0.12804(+/−0.01712)	0.45530(+/−0.14057)

Table 3: Prediction vs Reality

Store No.	Predicted GPR	Actual GPR	Diff (% rounded to one decimal)
50	0.35851824	0.426353	84.1%
87	0.52701836	0.547539	96.3%
113	0.31821062	0.250600	127.0%
15	0.2886206	0.204526	141.1%
63	0.21328386	0.148600	143.5%
81	0.27709348	0.228907	121.1%
98	0.04388074	0.043524	100.8%
55	0.3628831	0.366706	99.0%
138	0.33641409	0.117861	285.4%
137	0.27829742	0.652288	42.7%
107	0.57201804	0.570425	100.3%
141	0.24168037	0.407476	59.3%
61	0.30454308	0.317193	96.0%
131	0.23628084	0.064068	368.8%
111	0.35978257	0.100193	359.1%
34	0.16063038	0.046972	342.0%
72	0.27653349	0.208649	132.5%
65	0.42087533	0.335942	125.3%
53	0.19907962	0.673177	29.6%
70	0.14925133	0.031035	480.9%
106	0.13945877	0.019049	732.1%
7	0.3488638	0.158005	220.8%
74	0.36630816	0.251486	145.7%
33	0.63022261	0.460434	136.9%
103	0.49360623	0.220734	223.6%
67	0.2966841	0.230993	128.4%
39	0.22132102	0.285941	77.4%
95	0.20541434	0.234826	87.5%
100	0.42292478	0.361392	117.0%

Table 4: Choice of numerical features

Feature	Coefficient	What this means when opening a new store
Competitors	−0.633270	Place the store in an area with as few competitors as possible.
Mng-Exp	−0.277368	Choose a manager with low experience.
Adv.$'000	−0.254429	Have a low advertisement budget.
Car Spaces	−0.181642	Have less car spaces for your customers.
Mng-Train	0.020115	You get a slightly better result if you send your manager to training.

Table 5: Location

Category	Average GPR
Strip	0.346020
Mall	0.339481
Country	0.233047

Table 6: State

Category	Average GPR
New South Wales (NSW)	0.289773
Victoria (Vic)	0.355573
Queensland (Qld)	0.404097
South Australia (SA)	0.288813
Western Australia (WA)	0.352077
Tasmania (Tas)	0.206885
Northern Territory (NT)	0.127443
Australian Capital Territory (ACT)	0.059133

Table 7: State

Category	Average GPR
New South Wales (NSW)	0.289773
Victoria (Vic)	0.355573
Queensland (Qld)	0.404097
South Australia (SA)	0.288813
Western Australia (WA)	0.352077
Tasmania (Tas)	0.206885
Northern Territory (NT)	0.127443
Australian Capital Territory (ACT)	0.059133

Table 8: The managers sex

Category	Average GPR
Male store manager	0.332188
Female store manager	0.224365

Table 9: Open on sundays?

Category	Average GPR
Yes	0.325352
No	0.296598

Table 10: Do the store have home delivery?

Category	Average GPR
Yes	0.316627
No	0.312964