QBUS 6600 Data Analytics for Business Capstone

Big W Group Project

Group 74

520646988

490020122

520389591

510520027

Contents

Executive Briefing

Retail competition is fierce these days, with both strong competitors and changing consumer buying behavior. Big w is a discount department store chain with 179 stores and 20,000 employees worldwide. To maintain a stable market position, companies need to optimize their pricing policies and communication strategies, understand the market situation to develop strategies to cope with sudden changes and combine the current situation with future trends. This report will be based on exploratory data analysis, feature engineering, and model building to achieve the purpose of selling more products to customers without damaging their profitability, forecast the profit fluctuation factors of Big W, and put forward constructive suggestions to help Big W improve its market competitiveness and profits. Based on the following analysis, we suggest that Big W should use more accurate data to design inventory allocation methods based on regional demand and sales patterns, optimize its own inventory level, and use relevant models to improve the forecast of demand. Reduce the number of products in low demand, adjust the sales strategy according to customer demand and seasonal changes in products, and establish a customer loyalty program to stimulate the desire of customers to buy again.

Introduction

Nowadays, strong rivals across a range of industries make retail competitiveness intense. With the variable buying behavior of customers and the emergence of more and more competitors, optimizing pricing and advertising techniques is crucial for firms to sustain a steady market position. Big W is a chain of discount department stores in Australia, and it is a division of the Woolworths Group. The first Big W store opened in 1964, and the store is in the first shopping center that Woolworths developed (*500k Views Spark Big W History Lesson* - *Ragtrader*, 2020). As of today, Big W has 179 stores and employs 20,000 team members around the world, and it has grown in popularity (*500k Views Spark Big W History Lesson* - *Ragtrader*, 2020). Big W is a discount department store that earns a profit by selling high-quality products at a low discount. Companies need to increase sales while reducing expenses to increase revenue to prevent market instability and fluctuations in consumer demand. BIG W needs to put more focus on it, such as learning more about the market situation, knowing the direction of the market, and making strategies for the sudden situation or change. It can let the company formulate the corresponding treatment program in time. Also, high-level managers should combine the current situation with the company's strategies to accurately predict possible future development trends. These can help businesses better prepare for the future and successfully manage sustainable businesses. In this report, we will mainly focus on conducting a related analysis of influential features that can greatly impact the sales units, and subsequently, gross profit.

To achieve this goal, the analysis will involve exploratory data analysis (EDA), which is an analytical approach to identifying general patterns in the data and showing the relationship between results. Outliers and unexpected features of the data are included in these patterns. In any data analysis process, EDA is a crucial initial step (Biswal, 2023). Also, we will do some feature engineering on the data and build a few models. We will use a variety of methods to analyze the data and get the results. This project can help BIG W learn more about itself, including the relationship between sell price, sales units, and profit. It will allow Big W to improve its position in the market and gain more market revenue.

Data Processing

This report involves three datasets, with Big W train and test datasets containing 1.85 million data across four product categories from February 2022 to January 2024. Another competitor dataset covers 77.7k price data for Big W's top five most competitive competitors in the current Australian market from 14

November 2022 to 29 January 2024, with a time lag of 7 months. The following table shows the main variables required for the EDA part.

Response	**Profit-related**	**Sales_Units- related**	**Product-related**
‘gross_rofit’	‘gross profit’	‘sell_price’	‘category’
‘sale_units’	‘scanback’	‘promo_price’	‘brand’
	‘gst_flag’	‘sales_amount’	‘brand type’

As for the BIG W training set and test set, because the time in the competitor data set is only calculated by the weekday, while the time series data in the train data is continuous, the date data is converted to the date-time format for better visualization later. To better illustrate price comparisons with BIG W, price columns for each competitor brand were extracted and combined with training and testing sets to form a more specific data set. Moreover, the amount of data in the competitor dataset is small and exists null values. For the missing value of the promotional price, the average value of the sales price is used to fill the blank value of the promotional price. We also found duplicate columns in the dataset and a negative gross rate, both of which were removed. In addition, since the final profit is composed of gross profit plus scan and tax, the profit column was added to the train set and dummy variables were extracted from the “gst_flag” column to facilitate the judgment of tax status to calculate the total profit value. The above data processing methods are also carried out in the test set, so that all the data processing is completed.

EDA

As can be seen from below figure 1, gross profit fluctuates in an irregular pattern, reaching a peak from November 2022 to January 2023 and then gradually declining. In general, some retail stores and online retailers see a big increase in sales around Christmas until the end of the holiday season. Due to the limited time frame covered by this data set, it is not possible to judge that gross profit will show the same fluctuation trend from year to year. But as seen from the figure, there is no stable profit state in the time range of the observable dataset, so there may be other factors influencing its change.

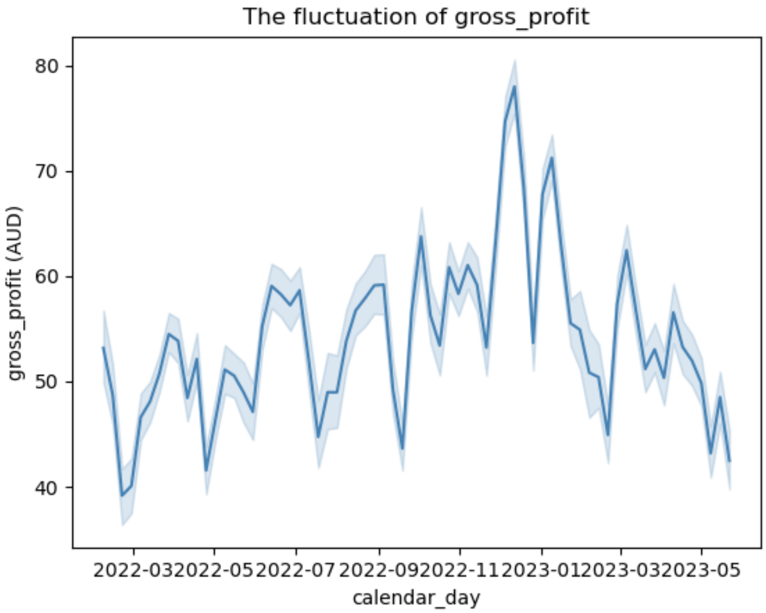


Figure 1: The fluctuation of gross_ profit

In order to achieve the purpose of BIG W selling more products to customers without damaging its profitability, we will explore the factors affecting BIG W's gross profit in this part. Firstly, the correlation map of gross profit is shown in Figure 2. It is found that there were two variables showing strong positive correlations with gross profit are sales amount and sales units, which have a positive correlation of 0.52 and 0.42. It means that as these two variables increase, gross profit will also increase. Sales units and sales amount are two important indicators to measure sales performance (Kokemuller, n.d.). The higher these two variables are, the more popular the company's products or services are and the stronger the market competitiveness.

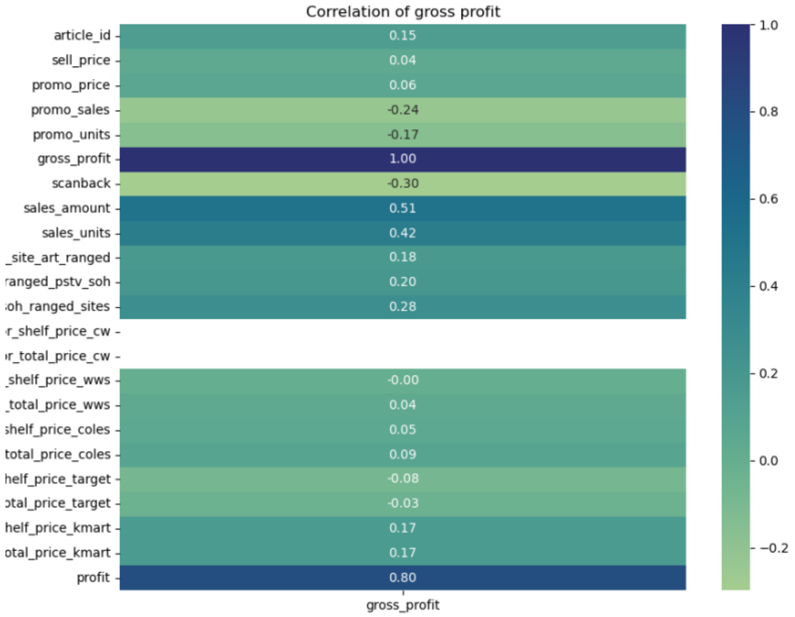


Figure 2: correlation map of gross_profit

Moreover, it can be seen from Figure 3 that the relationship between sales units and sales amount is mutual influence. An increase in sales units usually means that sales amount will naturally increase because more goods are sold; on the other hand, an increase in sales amount may also lead to an increase in sales units, so as to the company can have more money to spend on production as well as marketing campaign, thus attracting more consumers to buy its products (Maverick, 2023). These two variables influence each other, but cannot fully reflect the profitability of the company, still need to consider the cost, pricing strategy and other indicators.

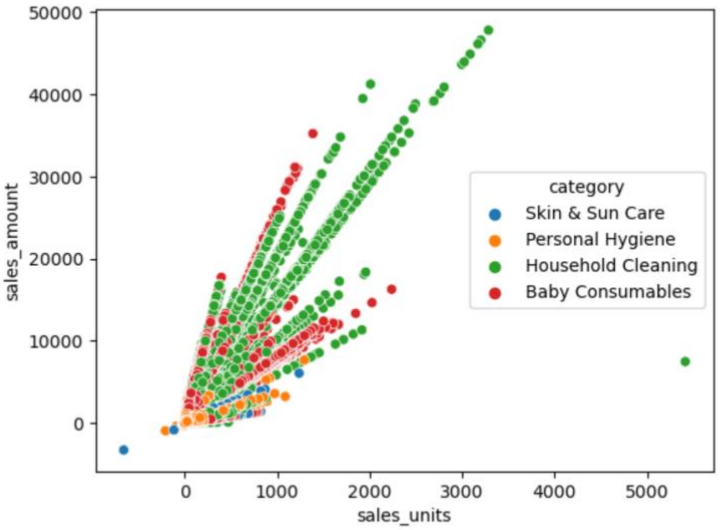
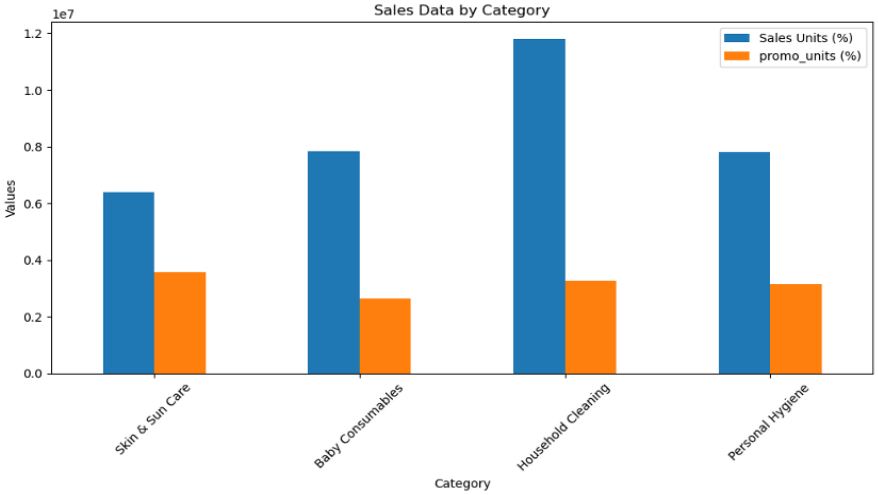


Figure 3: sales_units Vs sales_amount

Sales Units Vs Promotion

As can be seen from the correlation map, the scanback, promotional sales and promotional units are negatively correlated with gross profit. For retail stores, promotional activities can increase the sales unit of the product, but will reduce the price and profit margin of the product. Such promotions may appeal to some price-sensitive consumers; however, these consumers may only buy the products for the promotion, and not really need the products (Smith, 2019). In this way, for enterprises, promotional activities bring short-term sales growth, rather than long-term customer loyalty and profit growth. In a price-sensitive market, whether the length of a price cut or promotion leads to an increase in product sales depends on a number of factors. In the formulation of price reduction strategy, enterprises should consider market competition, demand elasticity and other factors to formulate a suitable price strategy.

Businesses often adapt their promotional and marketing strategies to seasonal trends. They might offer some seasonal discounts or host events with a seasonal theme. It can increase the company's visibility as well as attractiveness, and the most important is to raise the sales units. Promotion will only affect the sales units in short-term, it will not have long-term impact.



From the bar chat, we can see that the sales units for Household cleaning are higher than the other three sales units: Baby consumables, Personal hygiene, and Skin & Sun care. Also, the sales units of Baby consumables, Personal hygiene, Skin & Sun care are very close. For the promotions, only the number of promo_units of Skin & Sun care account for almost half of the total sales units. When comparing the total percentage of sales units with the percentage of promo_units, Household Cleaning is the least; it only has 0.3% of the 1.18%. Overall, when a company has promotions, the number of promo_units only has a little difference between Household cleaning, Baby consumables, Personal hygiene and Skin & Sun care.

Promotion will also increase price sensitivity; when the price goes down, people are more willing to make a purchase (Nielsen, 2017). Some of them might purchase more units than they originally planned.

Category	**Household cleaning**	**Baby consumables**	**Personal hygiene**	**Skin & Sun care**
Price sensitivity	Moderate	High	Low	Moderate

The table shows that Baby consumables have high price sensitivity and Personal hygiene has low price sensitivity. The market for baby consumables is typically large, and there are various brands and retailers. As a consumable, baby products are in high demand, and people need to purchase and store large amounts of them. When the price of Baby consumables from any brand decreases, it will attract large numbers of customers to buy it. Competition between different brands is intense, and it causes a rise in price sensitivity. Personal hygiene is a necessary item that people basically use every day. Thus, it makes them less sensitive to price.

Sales Units Vs Scanback

In addition, scanback itself is funded by the vendor during the promotional period and reduces the purchase cost for the merchant. The higher the scanback, the more the purchase cost is reduced, usually prompting the company to lower the price of the product in order to attract more consumers and increase sales (SIMISTER, 2013). It can also be seen from Figure 4 below that the higher the scanback, the greater the profit margin of product discount, which leads to an increase in product sales. This is very effective for clearing inventory or driving seasonal sales. However, if the scanback provided by the manufacturer is lower or the retailer's cost is higher, scanback may lead to a decrease in gross profit.

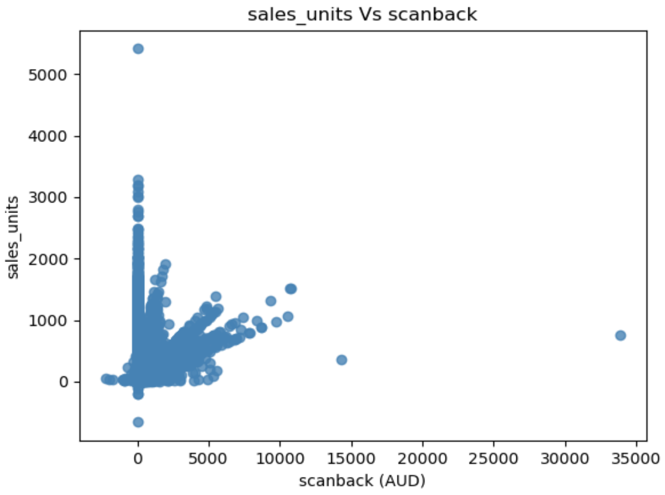


Figure 4: sales_units Vs scanback

Although promotional activities can stimulate the consumption desire of consumers to a certain extent, and effectively convey the information of products or services to consumers. However, it cannot bring long-term profits for merchants, and the sales units are affected by scanback, resulting in changes in profits. To solve this problem, merchants can flexibly adjust sales strategies according to customer needs and seasonal changes of products, which is conducive to establishing long-term relationships with customers. It is also possible to combine different promotional tools and methods to consider the influencing factors to achieve the best sales results.

Sales Units Vs Seasonality

As stated before, there are some connections between sales trends and seasonality, and we can look at how they vary across product areas. For the relationship between sale units and seasonality is how the number of units sold changes from season to season or period to period. Seasonality tends to influence consumer behavior due to various factors, such as different climates and holidays. For example, swimsuits sell more in the summer than in other seasons, supermarket sales are higher during the holidays, and some cultural products sell more during the corresponding festivals. However, not all products are affected by seasonality; some products sell well in all seasons. To observe and understand consumer behavior is essential for companies to effectively build their inventory management, operations, and marketing strategies.

Category	Household cleaning	Baby consumables	Personal hygiene	Skin & Sun care
Seasonal Trend	Increasing	Stable	Stable	Decreasing

From the above, it shows that seasons have a negative correlation with Skin & Sun care, although a positive correlation exists between household cleaning supplies and the changing of the seasons. In many traditional cultures, when spring arrives, it's time to do a thorough cleaning of the home to get rid of the dust and dirt that have built up over the winter. This might lead to an increase in the market for household cleaning products. In the winter, people will reduce the cleaning time because it is too cold. Thus, the demand for household cleaning products might decrease (Schoenbauer, n.d.). During the winter months, people may spend more time indoors than outdoors. When they engage in outdoor activities, as they wear more clothes to expose less skin, the demand for sunscreen products may drop at this time, and the sales of sunscreen products may decrease as well.

Sales Units Vs Competitor

BIG W has 5 competitors Woolworths, Coles, Chemist Warehouse, Kmart and Target. When there are more competitors in the market, it leads to an increase in customer demand for the product. Customers have more choices, which increases the probability that they will make a purchase. Also, competitors will do promotions to lower their prices to stay competitive, and they might place large quantities of items. When large quantities of items exist, their price will decrease. Moreover, some competitors might design new products to attract customers. These new products have more unique features or benefits than others. These behaviors will motivate people to buy more, which will lead to an increase in sales units.

Feature Engineering

Feature engineering was conducted to refine the dataset for more effective and predictive modeling.
A new column named `is_only_supplier` was added to indicate whether an SKU (product available at the store) is exclusively available at Big W or also at competitors, 1 means exclusive to Big W and vice versa. This helps in identifying how exclusivity impacts sales. Another column, `competition_level`, was introduced to count the number of brands within each subcategory, revealing the level of competition in different subcategories. A column `num_skus_on_sale` was created to track the number of distinct SKUs on promotion each week per subcategory, providing insights into the intensity of promotion competition. Furthermore, the `promo_power` was calculated to quantify the discount depth by comparing the regular and promotional prices each week. To better represent the influence from categorical features and avoid overly enlarged dataset, we choose the feature sub-category to be used for the dummy variable.

At last, we decided to represent the time effect on sales units by creating a new feature called days since the last promotion. When a product has never been promoted in the dataset at the given moment, it would be marked as -1. For the remaining entries, we would calculate the number of days since the last promotion for a product on the given day via the difference between the current date and the latest date since the price change.

Model Fitting

Based on the native characteristics of the dataset and the discovery captured in the EDA phrase, we have selected three distinct models: Elastic Net, Random Forest, and XGBoost. It is expected that the chosen models with unique strengths and focuses with respect to the input data will sufficiently provide contribution to the prediction of unit sales from Big W with constrained conditions.

- Elastic net is an advanced linear regularization that attempts to combine the advantages of ridge regression and LASSO, namely shrinkage and sparsity together (Naidu Narisetty, 2020). In other words, Elastic Net is a compromise between Ridge and LASSO. For this model, we expect that it can effectively utilize key features and reveal linearity within the data.
- As an ensemble of multiple decision trees and an extension of the bagging method, the random forest algorithm utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. (IBM, 2023). This model will help us to detect nonlinearity that is unable to unveil from plotting and linear models within data and provide more precise prediction.
- XGBoost is an ensemble of multiple learning methods in the field of machine learning. In short, Gradient Boosting is about taking a model that by itself is a weak predictive model and combining that model with other models of the same type to produce a more accurate model (West, 2022). It is also the top-pick choice for every Kaggle competition due to its premier characteristics in modeling.

In the modeling phase, we will exclude some variables, including promo_sales, promo_units, gross_profit, and sales_amount, and scan back (funding from vendor), to avoid data leakage. The reason is that these variables are either highly correlated with our target variable, unable to be obtained in real life, or directly include information about the unit sold.

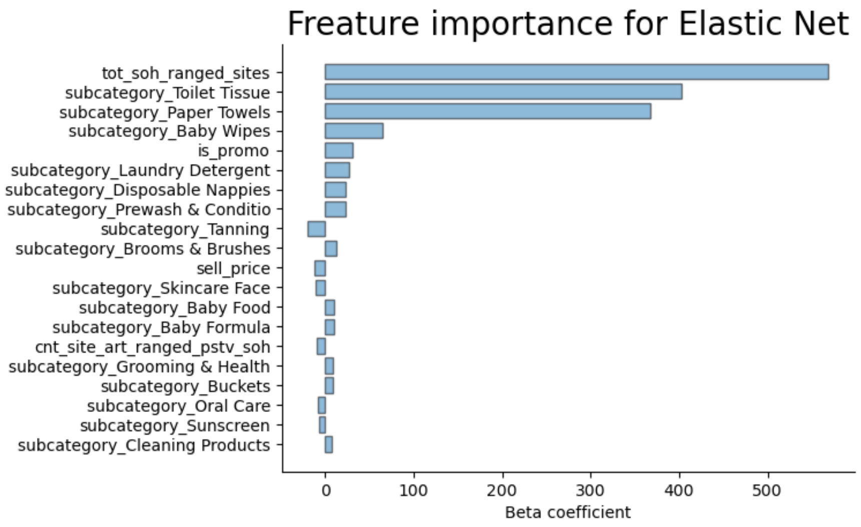
To measure the performance of each selected model, we will use root mean square error (RMSE) on the test dataset to justify which model is better than others. RMSE can be calculated by the square root of the sum of the difference between the predicted response variable and the true response variable, divided by the size of the test dataset.

Subsequently, we will also focus on the gross profit based on our best prediction result. It is intuitive that high unit sales do not always imply high gross profit. We will discuss more in the later section.

Model 1: Elastic Net (Linear model)

The first model we used was Elastic Net. As the baseline model, it is reasonable to use a linear model at first to settle a basis for our analysis. As we illustrated before, Elastic net is “a compromise between Ridge regression and LASSO regression”. To achieve this goal, some tuning is needed for the crucial parameters called L1 ratio and L2 ratio. These parameters range from 0 to 1 with the sum of 1 regulating the penalty contributed by Ridge and LASSO. Thanks to the built-in cross-validation function for the elastic net, we could search for the best L1 ratio from multiple ideal choices. To prevent the violation of the assumption of the linear model, we rescaled the whole independent variable using Min-Max scaling. In our case, the best L1 ratio is 1. From this number, we could conclude that the L1 penalty (coming from LASSO) has dominated the contribution of the regularization. As a result, our model will exhibit characteristics similar to LASSO regression, focusing more on shrinking coefficients to zero, rather than shrinking their magnitude. Since LASSO will shrink the coefficients of some features to zero, it would automatically perform feature selection for us.

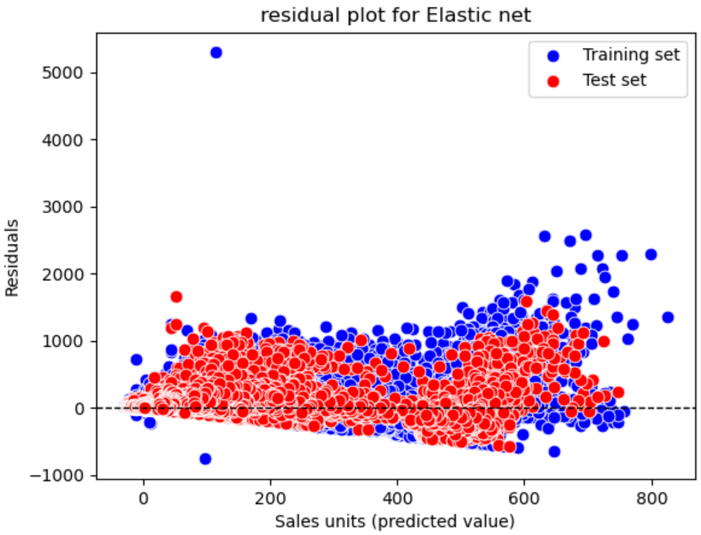
When we apply the optimal L1 ratio to our model and fit it with training data, the estimated RMSE between the predicted sales units from the test dataset and true sales units is 47.64. For the comparison, the RMSE on the training dataset is 45.70, which reveals that the performance of Elastic Net on our model is wonderful. To dive deeper into the concept of feature importance, the following plot is drawn:



In this plot, we notice that there are three most important variables in the model. At the top, the most influential feature is the Total stock on hand across all stores. For products with a large inventory on hand, promotion would always lead to more units sold. Meanwhile more stock on hand will lead to greater sales deals, which uplift the sales. The second and third feature that played a crucial influence on sales units is whether the products are categorized under the subcategory of Toilet Tissue or Paper Towels. It made sense because as a local store focusing more on selling daily goods, towels and Tissue are the most welcome product for the majority of the place. Along with the effectiveness of each feature, we could discover that the promotion did play a crucial role in the increment of units sold. Speaking of features with a negative relationship to the sales units, the most influential one is the subcategory of tanning. This could tell us that tanning products are unwelcome in Australia because of the intensive UV rays.

The resulting gross profits derived from the predicted sales unit also achieved ideal accuracy on the test dataset, with an RMSE of 395.78 between real gross profit and predicted one.

To check the linearity of our result, we plotted residual value against predicted sales units:



In this plot, we could observe the spreading of the data point is fan-shaped, indicating some increment in variance among the predicted value and implying that a minor heteroscedasticity (non-constant variance) is present. This symptom may introduce inconsistency to our predicted result, or potentially invalidate our model. We could also find some outliers at the top and bottom of the plot, revealing that there are some extreme values that generated significant deviations from the residual.

Some other limitations can be exhibited in our model. Firstly, the imbalanced L1 and L2 ratio might lead to insufficient utilization from ridge regression, leading to a high correlation problem between used features. Secondly, the complex nature of Elastic Net can increase the difficulty of model interpretation. Thirdly, the adjusted R square score for Elastic Net is 0.4571, this means that approximately 45.71% of the variance in the target variable is explained by the model. For simplicity, this means that there is still a significant amount of variability that cannot be explained by our model, which prompts us to use other models.

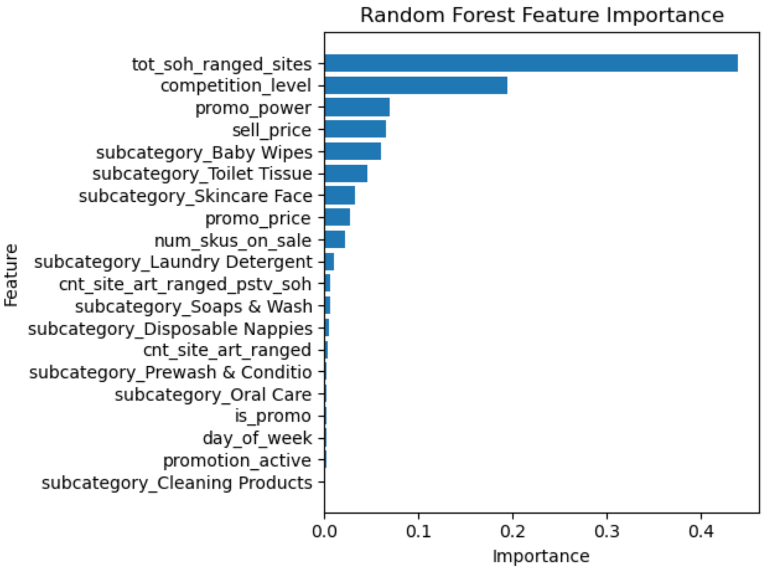
Model 2: Random Forest (Non-linear model)

Since the linear model did not perform well, we expect to get better results from the non-linear model. Random forest is one of the advanced tree-based models in machine learning. With some powerful techniques like bagging and ensemble, random forest successfully gathered attractive advantages from the basic decision tree method and comparatively fewer assumptions on input features compared to linear models. Due to its inherent randomness, the result will be different in every trial. To avoid it, we should set a seed (random state) to ensure our predicted result is consistent. Besides this, several other hyper-parameters need to be tuned. With the optimal tuning, we expect to get a more accurate prediction than the elastic net.

To save up time, we used randomized search for hyper-parameter tuning. The choices of hyper-parameter being tuned would also create an impact on our result. In our case, the hyper-parameters we tuned are the number of trees in the forest (n_estimators), the minimum number of samples required to be at a leaf node (min_samples_leaf), and the maximum number of features used for splitting (max_fatures). In the optimal case, the value of these hyper-parameters is 110 for n_estimators, 971 for min_samples_leaf, and 34 for max_features.

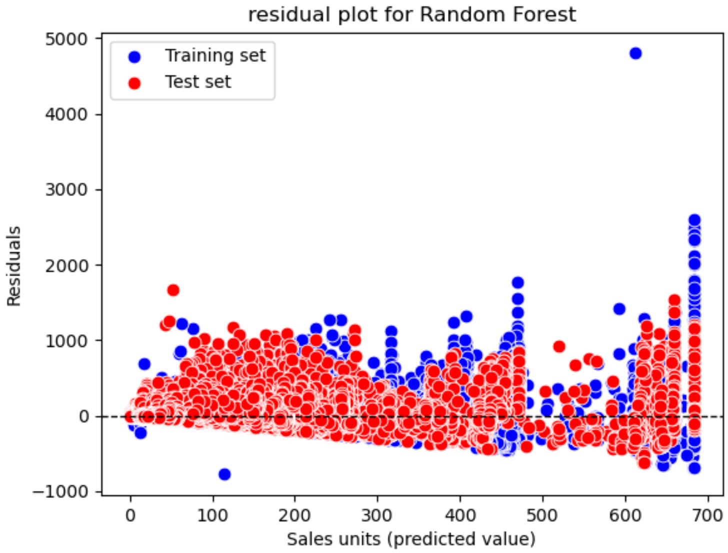
After fitting the training data and predicting the sales units by features from test data, the RMSE on the test dataset is 42.75. For the comparison, the training RMSE of random forest is 39.11. Compared to Elastic Net, we achieved a small improvement in random forest and didn’t overfit or underfit the training dataset. For the resulting gross profit calculated using predicted sales units, the test RMSE is 378.35, which is also better than the result from our previous model.

We are also interested in the features that played an important role in prediction.



Similar to the plot from Elastic Net, total stock on hand ranged across all the stores is still dominating in all used features. The second most important feature is the competition level within each subcategory. When more distinct brands are presented under the same subcategory, vendors tend to settle for lower selling prices or higher discount percentages, resulting in better sales units. Promotional power took second place in features of importance for Random Forest. Sell price placed a leading position in the list of important features. Since the consumers of Big W are price-sensitive, less price would always lead to better popularity. Among the categorical features, baby wipes became the most welcome product among all of the available subcategories. Combined with the discovery in the EDA part, we could conclude that families tend to hoard baby wipes periodically. Besides the features that also appeared to be important to Elastic Net, like toilet tissue, the product under the Skincare face subcategory became one of the top ten important features in Random Forest. The rising of this feature can explain why tanning is unwelcome from another perspective.

In random forest, the R squared score is 0.5641, showing that approximately 56.41% of the variations in our dataset are explained by random forest. This is a great improvement from Elastic Net. The rise in the R-squared score implies that the non-linear factors beneath our dataset are dominant over linear factors.



The residual plot may tell us more about the performance of random forests. At first glance, we could observe a fan shape spreading on the plot, suggesting that our model might not fully capture some aspects of the underlying data. On the other hand, the similar distribution for training and test datasets implies that the model is reasonably decent and didn't overfit the training data.

Some limitations exist in random forests. Firstly, training can be computationally intensive for random forests, especially for high-dimensional datasets. Secondly, the increasing complexity leads to higher difficulty in interpretation, compared to linear models. Thirdly, Random Forest may capture noise in data, causing bad performance at the end. This could be proven by the large variation for large numbers in the residual plot.

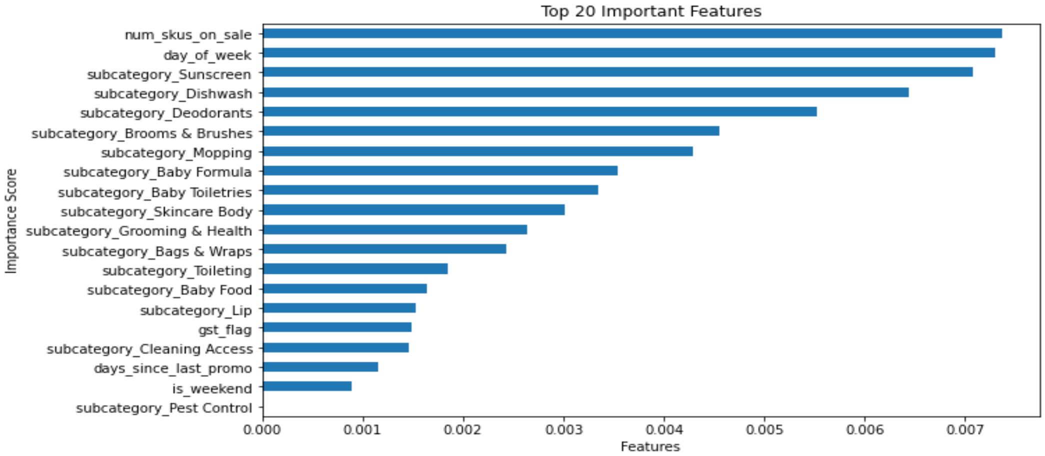
Model 3: XGBoosting

The XGBoost is known for its efficiency and effectiveness, particularly with large datasets which is exactly the situation we need to deal with. The model configuration (such as Max Depth and Learning Rate) was optimized using Randomized Search since the dataset is large and we have only limited computing power. This method tests various combinations of these parameters to identify the most effective setup efficiently.

Interpretation of the Estimated Model

The optimal configuration determined through Randomized Search was found to have a maximum depth of 4 (depth of each decision tree model), a learning rate of 0.05(how fast the model converges), and 500 trees (the overall size of the model). Furthermore, to prevent overfitting—a scenario where the model memorizes data patterns without generalizing well—methods such as cross-validation were implemented. The model was evaluated using the Mean Squared Error (MSE) metric, which measures the average of the squares of the differences between observed and predicted values. A lower MSE value indicating a model that predicts more accurately.

Fitting with test data, the best model configuration resulted in an MSE of 1969.33 for sales units and 785 for gross profit, which showed room for improvement. In comparison, the Random Forest has an MSE of 39, suggesting a more accurate prediction. Despite the relatively high MSE values, the model achieved an adjusted R^2 of 0.66, demonstrating substantial explanatory power regarding the data. This justifies further interpretation of the feature importance matrix.



Some key takeaways of the Feature Importance plot are:

1. num_skus_on_sale: Indicates how many SKUs are on promotion within the same subcategory at the same week. A high importance score for this feature suggests that the level of competition among promoted items in a category significantly impacts sales performance.
2. day of week/is_weekend: Sales performance varies by day, indicating potential consumer purchasing patterns that vary throughout the week.
3. Product Subcategories: Certain subcategories such as Sunscreen, Dishwash, and Deodorants show significant sales, pointing towards their critical role in the overall sales/cross promotion strategy.
4. days_since_last_promo : this feature is crucial for understanding the relationship between promotion frequency and consumer purchasing habits. A shorter interval may suggest that frequent promotions are necessary to drive sales and vice versa.

Limitation of the Model

XGBoost builds complex models that are difficult to interpret. Furthermore, the model is sensitive to outliers (commonly seen in our dataset) which led to skewed predictions.

Conclusions and recommendations

In conclusion, in the previous EDA parts we find that sales units and sales are two major factors that can cause gross profit fluctuation and they influence each other, but they cannot fully reflect the profitability of the company. Since the promotion may attract some price-sensitive consumers, although the promotion is able to increase the sales units of the product, it may reduce the price and profit margin of the product. Promotions lead to short-term sales growth and do not sustain long-term customer loyalty and profit growth. In addition, competition between different brands can lead to increased price sensitivity. When product selectivity increases, it stimulates the purchasing power of the masses, thereby increasing sales and gross profit. On the other hand, the sales unit will also be affected by the scanning code; The higher the scanback, the lower the purchase cost, prompting companies to lower the price of products to increase sales, especially when clearing inventory or boosting seasonal sales. We used exploratory data analysis to find out some key findings that emerged, and used them to build the following models.

The random forest model uses 3 hyperparameters for random search, and its performance is better than Elastic Net. The R-square score is higher than that of the elastic network, indicating that the nonlinear factors are more dominant than the linear factors in the data set. Given that the model is acceptable and that there is no overfitting of the training set. We used the first two models to identify several features that were most beneficial to sales units and finally determined that the key features affecting sales included the number of SKUs promoted, the timing of the promotion, and the change in sales by day and product subcategories. However, the baseline model Elastic Net indicates that the model is more likely to be LASSO, and the proposed XGBoost model is proven to need improvement. Although the XGBoost model has a high R2 score, it is sensitive to outliers which leads to skewed predictions. By comparing the three models, we can choose different models according to the different scenarios in business. Random forest is appropriate for those with high error requirements, while the XGBoost model is suitable for those with high dataset fit requirements.

Based on the above analysis, we suggest that Big W should optimize its own inventory level based on the data of day_of_week and is_weekend, so that the inventory on hand is sufficient rather than overstocked. Use more accurate data to design inventory allocation methods based on regional demand and sales patterns, and use relevant models to improve demand forecasting. In terms of products, we should reduce the release of unpopular products, such as tanning products. Establish a customer loyalty program to encourage repeat purchases of popular products. Use data on num_sku_on_sale to inform customers about pricing strategies and provide targeted promotional posters for hot items. It should also flexibly adjust sales strategies according to customer needs and seasonal changes in products, establish customer loyalty programs, and encourage customers to repeat purchases of popular products.

References

500k views spark Big W history lesson - Ragtrader. (2020, November 24). Www.ragtrader.com.au. https://www.ragtrader.com.au/news/500k-views-spark-big-w-history-lesson

Biswal, A. (2023, February 17). *What is exploratory data analysis? steps and market analysis: Simplilearn*.Simplilearn.com.

IBM (2023). *What is Random Forest? | IBM*. [online] www.ibm.com. Available at: https://www.ibm.com/topics/random-forest.

Kokemuller, N. (n.d.). *What Are the Elements of the Gross Profit Ratio?* Small Business - Chron.com. https://smallbusiness.chron.com/elements-gross-profit-ratio-55271.html

Maverick, J. B. (2023, February 20). *Is It More Important for a Company to Lower Costs or Increase Revenue?* Investopedia. https://www.investopedia.com/ask/answers/122214/company-it-more-important-lower-costs-or-increase-revenue.asp#:~:text=Profit%20margins%2C%20which%20are%20computed

Naidu Narisetty, N. (2020). *Principles and Methods for Data Science*. Elsevier.

Nielsen, L. (2017, November 21). *The effect of sales promotion on Sales Volume*. Small Business - Chron.com https://smallbusiness.chron.com/effect-sales-promotion-sales-volume-5051.html

Schoenbauer, A. (n.d.). *Season of Change for Spring Cleaning*. Numerator. Retrieved May 4, 2024, from https://www.numerator.com/resources/blog/season-change-spring-cleaning/

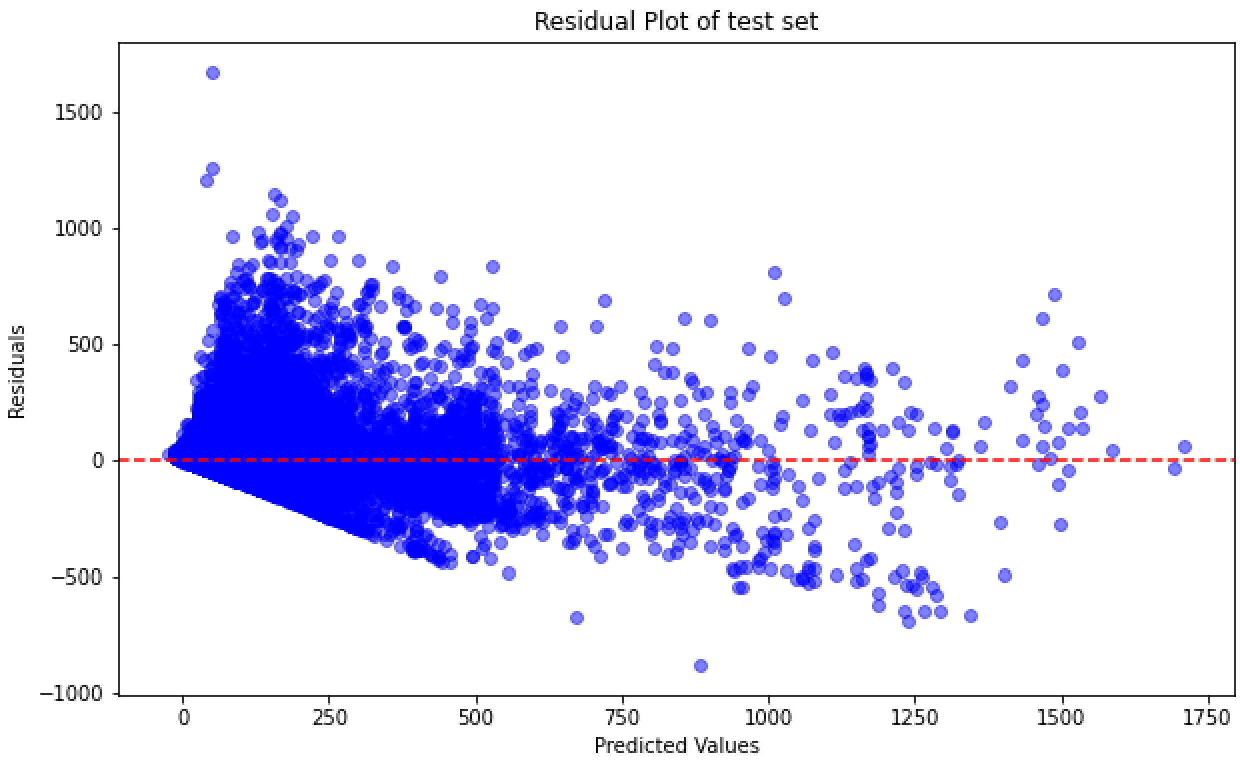
SIMISTER, P. (2013, April 30). *The Cost Volume Profit Relationship*. Businessdevelopmentadvice.com. http://businessdevelopmentadvice.com/blog/the-cost-volume-profit-relationship/

Smith, K. (2019, January 19). *How to Calculate the Profit from Your Promotions*. Www.mighty roar.com. https://www.mightyroar.com/blog/promotion-profit#:~:text=However%2C%20the%20bad%20thing%20about

West, M. (2022). *What is the difference between the R gbm (gradient boosting machine) and xgboost (extreme gradient boosting)?* [online] Quora. Available at: https://www.quora.com/What-is-the-difference-between-the-R-gbm-gradient-boosting-machine-and-xgboost-extreme-gradient-boosting.

Appendix

- Residual Plot of model 3


Residual Plot of test set
Residuals
Predicted Values