QBUS6600: Big W Sales Analysis

Student Name	SID
Tran Hy Dong	500286001
Xinjun Zhou	520349263
Honghao Song	520166132

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Objective: The primary aim was to analyze BIG W's sales data, a renowned discount department store under the Woolworths Group, to uncover patterns and behaviors that can potentially drive the store's sales. This involved understanding past sales trends, predicting future ones, and furnishing actionable recommendations to BIG W.

Methods: Utilized various predictive models including Linear Regression, Decision Tree, Random Forest, Gradient Boosting, Lasso, and ARIMA. The choice of modeling technique was determined by the nature of the data and the specific business context.

Key Findings:

- Data Preprocessing & Exploratory Data Analysis:
- Majority of BIG W sales are from physical store channels.
- Positive relationship observed between media spend and total sale value in each state.
- Proximity to competitors like Kmart and Target has a varied impact on sales.
- Seasonal peaks in sales observed during the Christmas season.
- Budget-Young Families category emerged as the highest contributor to sales.

Model Evaluation:

- The Gradient Boosting Model demonstrated the best performance among the traditional machine learning models, with an R2 value of 0.619631 and RMSE of 0.867222.
- The ARIMA model, suited for time series data, had an RMSE of 0.284, indicating high accuracy.

Recommendations:

- Allocate the advertising budget more strategically to align with the specific consumption habits of different regions.
- Explore the possibility of opening new stores in the VIC region, especially in locations near competitors like Kmart and Target.
- Strengthen product and promotional strategies targeting the Budget-Young Families segment. For the Premium-Older Families segment, delve deeper into their preferences and purchasing behaviors.

Conclusion: By leveraging data-driven insights and adopting the right predictive modeling techniques, BIG W can craft effective strategies for sales growth, market presence, and overall relevance in the evolving retail landscape.

I. INTRODUCTION

In today's rapidly evolving retail landscape, the significance of data and analytics cannot be overstated. Businesses that effectively harness the power of data can make informed decisions, predict future trends, and tailor their strategies to meet ever-changing consumer demands. One such business seeking to reinforce its market position through data-driven insights is BIG W, a renowned discount department store under the Woolworths Group. With competition intensifying, especially from the burgeoning e-commerce sector, BIG W finds itself at a crossroads. How can it stay relevant, continue to appeal to its core audience, and at the same time, adapt to the new dynamics of the retail world?

This project is centered around this challenge. Through a meticulous analysis of provided datasets, combined with external demographic and income data, our aim is to uncover patterns, behaviors, and attributes that can potentially drive BIG W's sales. This involves not just understanding past and present sales trends, but predicting future ones, thus allowing BIG W to take proactive steps rather than reactive ones.

By the end of this project, our objective is to furnish BIG W with actionable recommendations – whether it's about opening new stores, optimizing advertising spends, or doubling down on e-commerce strategies. As we delve deeper into the data, we remain committed to the goal of not just boosting sales, but enhancing BIG W's overall market presence and relevance in these transformative times.

The data set consist of 4 csv files which is relatively clean to make analysis. The data set consists of:

- Sales by customer location train: The dataset contains approximately 1.8 million rows of training data. Each entry provides details about a customer's location, their price and lifestage segment, group size, total transactions, sales value for a specific Big W store
- Sales by customer location test: Similar to Sales by customer location by for test set (365 thousand rows)
- Sales by store location contains ~425 rows where each row represents a unique store and contains information on the store location (state, postcode, latitude, longitude)
- Media investment: contains ~950 rows where each row has a week ending date, the state, and media investment amount spent for this state

II. DATA PRE-PROCESSING & EXPLORATORY DATA ANALYSIS

1. Data Pre-Processing

Before actual doing the EDA, for the data processing, the missing value in the file “Sales by store location”, which has 11 in co_location_flag, 1 in distance_to_kmart and 1 in distance_to_target, has been dropped to avoid the impact of the missing value on further analysis.

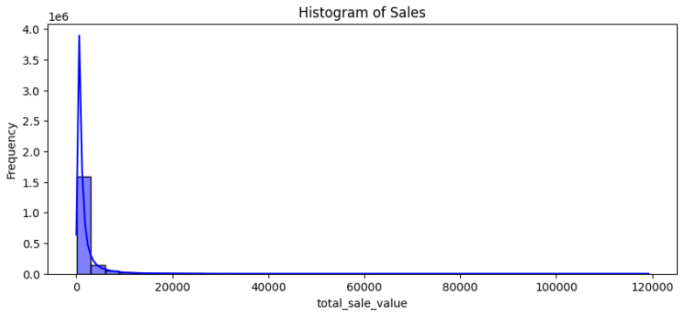
Variables like 'co_location_flag', 'distance_to_kmart', and 'distance_to_target' were transformed into dummy variables, a common practice when dealing with categorical data.

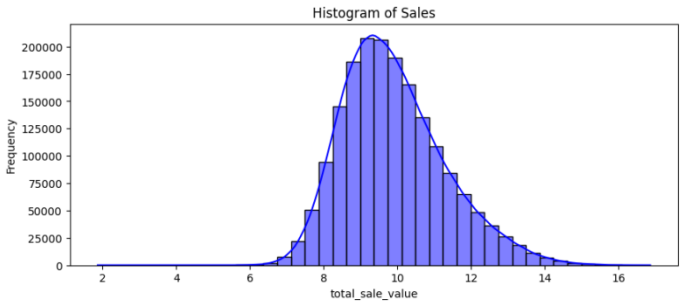
To enhance our data analysis, we have converted the datetime format to facilitate the integration of two distinct data tables. This modification was undertaken to synchronize the data and harness a more

comprehensive insight. The primary objective was to standardize the data into a datetime format, thereby enabling a seamless connection between the two tables.

2. Exploratory Data Analysis

a. Distribution of total_sale_value

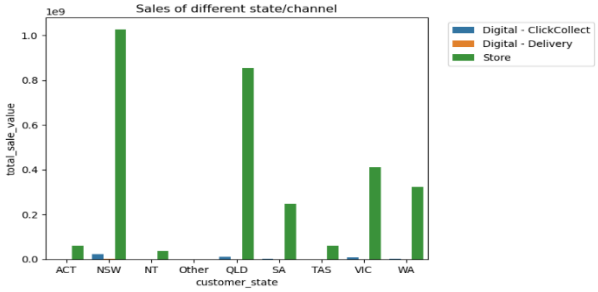




To compare between the Total Sales Value and Log Total Sales Values, the distribution of Total Sales Value is right skewed, but after log-transforming, the skewed value is 0.6, which close to 0 stand for normal distribution. If we want to do the sales-predicting model, log Sales Price should be the variables to be chosen to be predicted.

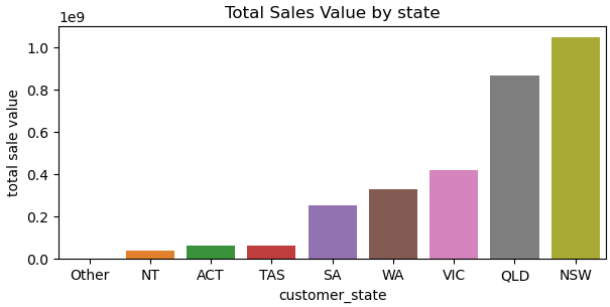
b. Total_sale_value by location

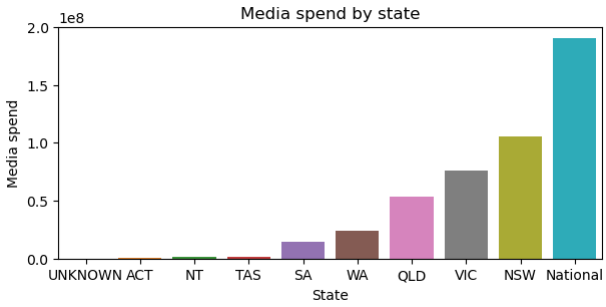
Firstly, in the file "Sales by customer location train" (will be refer as training file instead at followl), it indicates that Big W has three distribution channels to deliver the value to the customers which are store, Digital-ClickCollect and Digital-Delivery. The sales proportion on each channel is shown in the chart:



As we can see, the Store channel dominates the proportion of Big W sales value, where the Digital-ClickCollect channel only contributes a little compared to the Store channel. The sales proportion of the Digital-Delivery channel is so trivial that we can barely see it on the barchart.

Next step is to illustrate the total sales for each state and the funds that are spent on promotion for each state to see if they follow some pattern. The graphs show as follow:

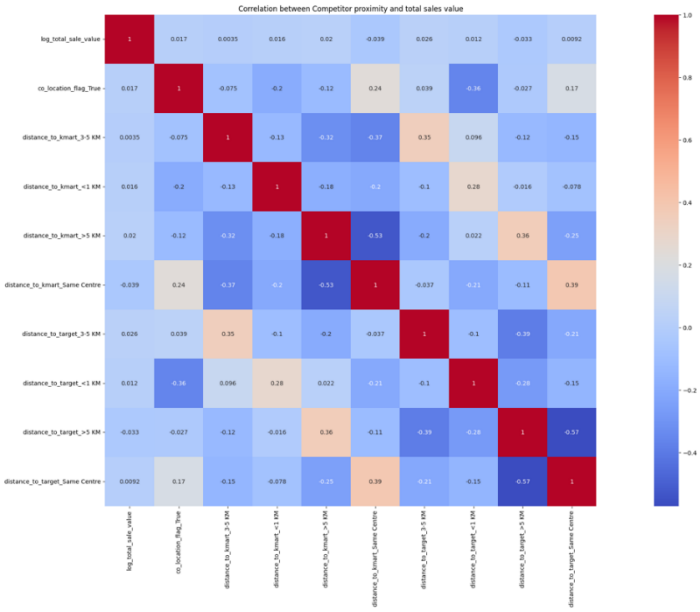




Combine these two graphs together, it roughly shows a positive relationship between the media spend and the total sale value in each state.

c. Impact of Competitor's location on Total Sales Value

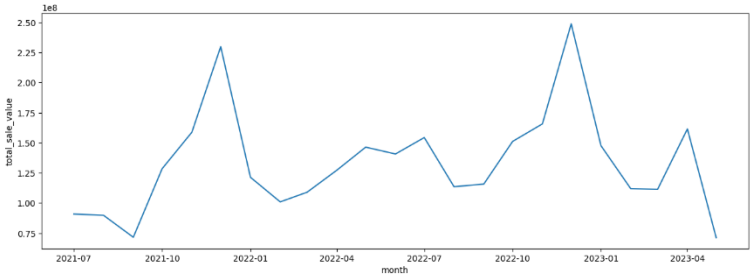
The proximity of competitor stores to Big W locations can significantly influence its total sales value, as nearby competition may divert potential customers and dilute market share (Johnson & Turner, 2022).



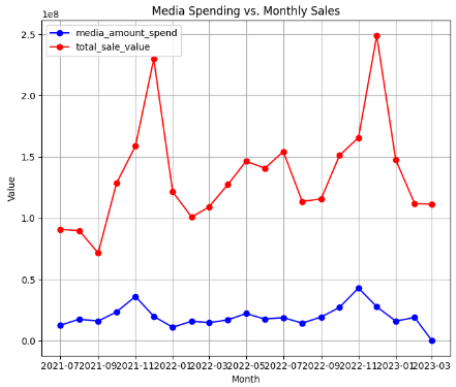
Thus, the columns of "co_location_flag", "distance_to_kmart", and "distance_to_target" should be transformed into dummy variables for further analysis. Combing these dummy variables with logged total sale value, we can have a heatmap which shows the correlation between the logged total sale value and dummy variables:

In this instance, the influence of competitors on Big W's total sales value appears to be relatively minimal. Being situated in the same mall or shopping area can in fact enhance sales, as customers are drawn to locations offering a diverse array of products, and Big W benefits from this arrangement. However, it's crucial to highlight that when Big W co-locates with competitors such as K-mart within the same shopping center, there is a discernible negative correlation, suggesting a potential decrease in total sales value.

d. How seasonality affect total_sale_value

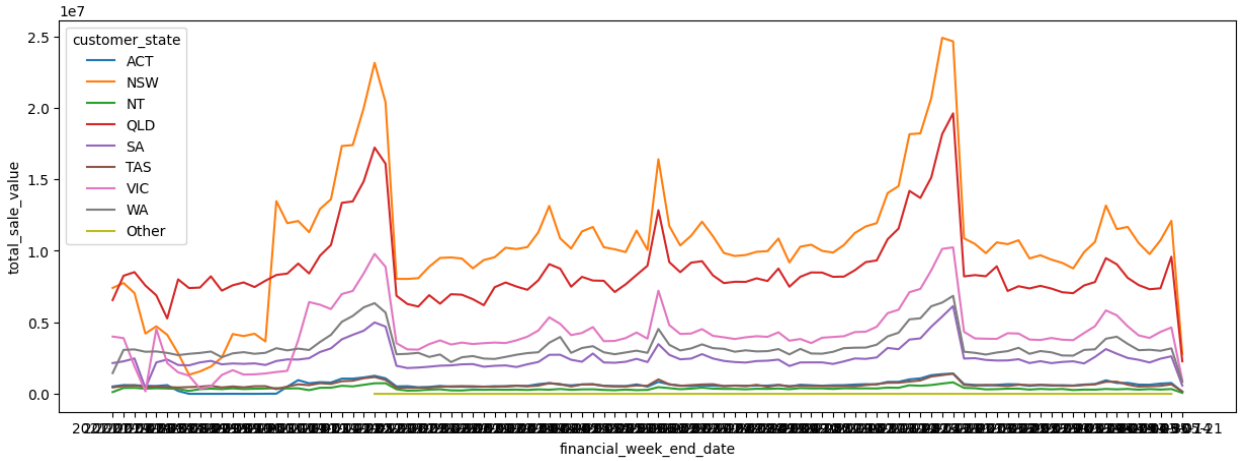


The figure delineates pronounced variations in monthly total sales values spanning from June 2021 to May 2023. Notably, two significant peaks in sales are evident in November 2021 and November 2022, coinciding with the Christmas season. Outside of these peaks, the total sales values tend to remain within a specific range. A substantial decline from April 2023 to May 2023 is discernible, which could pose a significant concern for Big W.

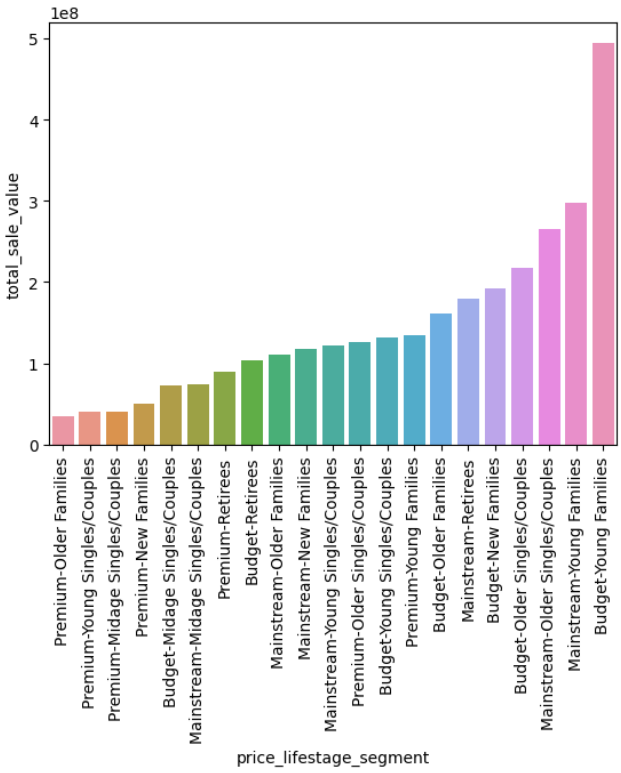


When we compare the sales data with how much Big W spent on advertising, we see a clear pattern. Every time Big W spends more on ads, their sales go up the next month. For example, they spent a lot on advertising in October, and by November both in 2021 and 2022, their sales went up. But when they don't advertise as much, their sales drop right away.

Also, we can group up the data by states to see the moving trend of total sale value overtime. The line chart shows below with financial_week_end_date as x and total_sale_value as y.



e. Price Life Stage Segment Analysis



From the data provided in figure 10, insightful patterns emerge regarding sales values distributed across different price-lifestyle segments:

The Budget-Young Families category stands out prominently with sales values reaching approximately $494.5 million, making it the highest contributor in the entire dataset. This suggests that young families operating on a budget are a significant market segment for Big W, likely due to their specific needs aligned with budget-conscious purchasing. Smith et al. (2021) indicated that young families often prioritize budget shopping to cater to the diverse needs of growing household.

On the contrary, the Premium-Older Families segment records the lowest sales at around $34.3 million. This could imply that older families within the premium bracket have specific spending patterns or brand loyalties that don't necessarily align with Big W's offerings. Johnson & Lee (2022) discussed how older families in the premium category often have established purchasing habits and are less influenced by broad market trends.

To optimize sales strategies, Big W might consider reinforcing their marketing and product offerings tailored to the preferences of the Budget-Young Families, while also investigating the nuances of the Premium-Older Families segment to understand their unique purchasing behaviors.

III. MODEL EVALUATION

1. Approach method

In the realm of business analytics and forecasting, predicting sales is a pivotal task that can shape the strategic decisions of a company. For a retail giant like Big W, accurately forecasting total sales can be the difference between a successful quarter and a challenging one. The target variable, in this case, " total_sale_value," represents the cumulative sales value that Big W aims to predict. While there are various modeling techniques available, this essay will advocate for the use of a multi-model approach, even when utilizing the same set of variables.

The choice of modeling technique can significantly influence the accuracy and reliability of forecasts. At first glance, employing multiple models with identical variables might seem redundant. However, this strategy is underpinned by several compelling reasons. Firstly, model robustness is paramount. Different models come with distinct assumptions and strengths. For instance, while one model might excel in capturing linear relationships, another might be adept at understanding intricate patterns. By leveraging multiple models, we mitigate the risk of over-relying on a single model's potential limitations or inherent biases. Secondly, the beauty of diverse perspectives cannot be understated. A linear regression model might elucidate overarching trends, but a decision tree can delve deeper, spotlighting non-linear relationships and variable interactions. Lastly, model validation is bolstered by this approach. When predictions from various models align, it significantly boosts our confidence in the forecast, ensuring that anomalies or outliers are promptly identified and addressed.

When trying to predict accurately, it might seem good to adjust data for each method. But, it's better to use the same data for all methods. This makes comparing results easier and ensures fairness. Using different data can make a method too focused on past data, making it perform poorly on new data which can be known as overfitting. Also, stakeholders and decision-makers often prefer simpler, consistent models. Using the same set of variables ensures that the insights drawn are consistent and easier to communicate.

2. Data processing

a. Featuring data

In the vast landscape of data analytics, the quality and structure of the dataset often dictate the success of predictive models. When presented with fragmented datasets, the challenge lies in effectively merging them to extract meaningful insights.

Initially, we are presented with distinct datasets, each holding a piece of the puzzle. The primary goal is to weave together customer, store, and media information to create a holistic view. From the sales_by_customer_location dataset, pivotal sales metrics such as store_id, total_sale_value and customer_state are extracted. These metrics provide a snapshot of sales performance and customer preferences.

Parallelly, the sales_by_store_location dataset offers a deep dive into store-specific details. Notably, the distances to competitors like Kmart and Target are extracted. These distances are then classified into categories like nearkmart (where is “Same Centre”, “<1 KM”, and “1-3KM”) and farkmart (All other distances fall under this category.), providing a spatial perspective on store locations relative to competitors. Moreover, rename week_ending in Media_investment to financial_week_end_date and state in Media_investment to store_state, and then merge the new Media_investment dataset and merged dataset in 4-th step based on financial_week_end_date and store_state.

To make sure we have a comprehensive picture, the data of average of weekly earning per customer state has been added and merged to the table because direct income Australian can hugely affect on their decision of buying more, which directly affect the total_sales_value.

However, the merging process is not devoid of challenges. Data redundancy and missing values can skew results. Rigorous data cleaning, which involves removing duplicates and imputing missing values, ensures the dataset's integrity.

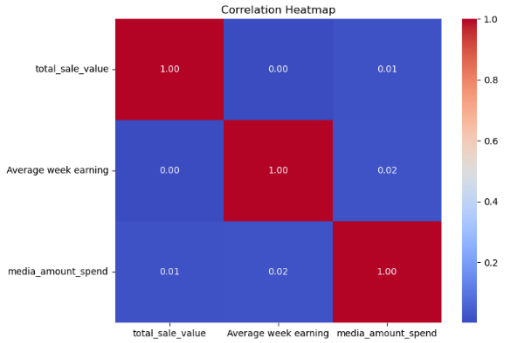
In conclusion, the process of merging and refining datasets is akin to crafting a masterpiece from individual pieces. Each step, from extraction to cleaning, is crucial in shaping the dataset's narrative. In the realm of predictive analytics, such meticulous preparation not only enhances model accuracy but also provides richer insights, guiding businesses towards informed decision-making.

Variables	**Type of variables**	**Number of variables**
total_sales_value	**Target variable/ Numberical variables**	1
distance_to_kmart	Dummy variables	1
distance_to_target	Dummy variables	1
distance_to_target	Dummy variables	1
media_amount_spend	Numerical variables	1
Average week earning	Numerical variables	1
State Variables (state_ACT, state_NSW, etc.)	Dummy variables	8
price_lifestage Variables	Dummy variables	9

b. Process categorical variable.

In the curated dataset, three variables are identified as categorical in nature. Direct utilization of these categorical variables by predictive models is infeasible. Therefore, the introduction of dummy variables becomes imperative to transform these categorical attributes into numerical counterparts. Dummy variables delineate categories through binary indicators, signifying the presence or absence of a specific category within a particular observation.

c. Process numerical variable.



Explore the correlations between numerical variables. Highly correlated variables may lead to multicollinearity issues in linear models, which can impact model interpretability and stability. We can observe that the low correlations with the media_amount_spend and average week earning is quite low which can that these variables are providing unique information to the model, reducing the risk of multicollinearity.

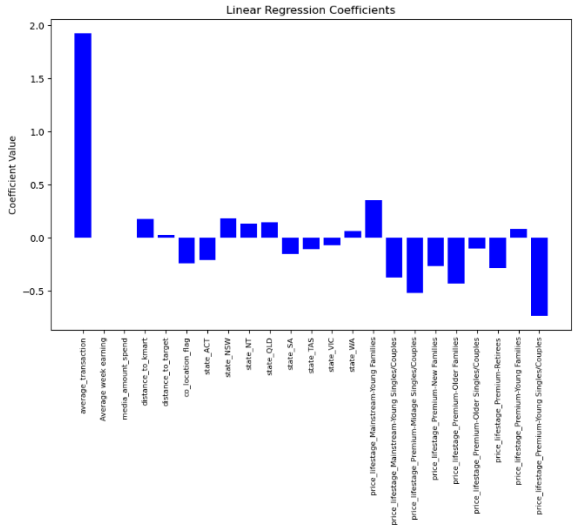
d. Split Training Set and Validation Set

Before we start training our models, we need to split our initial dataset into two parts: one for training the models and the other for checking how well they work on new, unseen data. This split helps us figure out how good our models are and allows us to choose the best one and make any needed adjustments.

In this task, we randomly pick 75% of our original data for training, and the remaining 25% is set aside for validation. This way, we can make sure our models learn from a good chunk of data while also testing them on data they haven't seen before. It's like having a practice run before the real thing, ensuring our models are as accurate as possible when they encounter new information.

3. Model Evaluation

a. Linear Regression Model



Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The simplest form of linear regression involves one dependent and one independent variable and is referred to as simple linear regression. Linear regression is simple model which find a best fit data linear equation to minimize the sum of the squared differences between the predicted and actual value.

In our analysis, the average_transaction predictor, with a coefficient of 1.9258, emerged as paramount, indicating that even slight changes in transaction values can significantly influence the outcome. Similarly, the price_lifestage_Mainstream-Young Families segment, with a coefficient of 0.3569, showcased its distinct influence, suggesting that this demographic has unique behaviors affecting the predicted outcome. Conversely, the almost negligible coefficient of media_amount_spend (5.997278e-09) hints at its limited impact, prompting a reconsideration of media spending strategies. In essence, while all predictors play roles in shaping outcomes, some, like average_transaction, wield more influence, guiding data-driven decision-making.

	RMSE	**R-squared**
Linear Regression Model	1.314797	0.125698

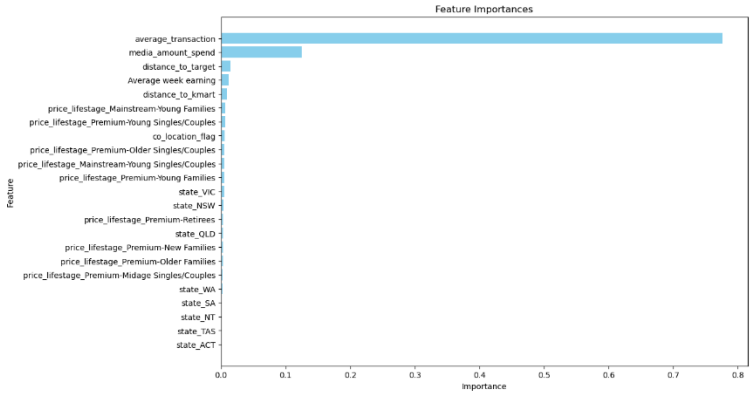
The Linear Regression Model's performance, as indicated by the RMSE of 1.314797, reveals that the model's predictions deviate from the actual values by an average of approximately 1.314797 units.

Furthermore, the R-squared value of 0.125698, or 12.57%, indicates a relatively low proportion of the variance in the dependent variable being explained by the model. This suggests that a significant amount of variability in the data remains unaccounted for by the model's predictors.

In essence, based on these metrics, the model might not be considered highly predictive.

b. Decision Tree Model

The Decision Tree model is a popular and versatile machine learning algorithm that belongs to the family of supervised learning methods. It works by recursively splitting the dataset into subsets based on the most significant attribute(s) at each level. Therefore, Decision Trees offer an intuitive and visual approach to decision-making, making them a favorite choice for tasks that require human interpretability.



Besides the average transaction (0.77), media_amount_spend (0.125031): The second most influential feature, though considerably less than the first. It indicates that the amount spent on media also plays a notable role in the model's decisions, but not as prominently as the average transaction.

distance_to_target (0.014614) and other features with lower importance scores: These features contribute to the model's decisions but have a relatively minor influence compared to the top features. Their impact on the model's predictions is more nuanced and might be more evident in specific contexts or data subsets.

	RMSE	R-squared
Decision Tree Model	1.001441	0.492781

The Decision Tree Model has an RMSE of 1.001441, indicating predictions are off by about 1 unit on average. With an R-squared of 0.492781, the model explains roughly 49% of the data's variance, suggesting a moderate fit with potential for improvement.

c. Random Forest Model

Random forest is an ensemble machine learning algorithm that combines the prediction of multiple decision tree to improve the accuracy and robustness of model. Based on random selection of trained samples and feature, it can handle complex data and reduce overfitting.

N_esitimators which represents the number of decision tree and max_depth of each decision tree is the most importance of parameters in random forest, GridsearchCV is used to tune these two parameters:

- 'n_estimators': np.arange(3,50,5)

- 'max_depth': np.arange(5,20,5)

So we have best parameters: {'n_estimators': 38, 'max_depth': 15}

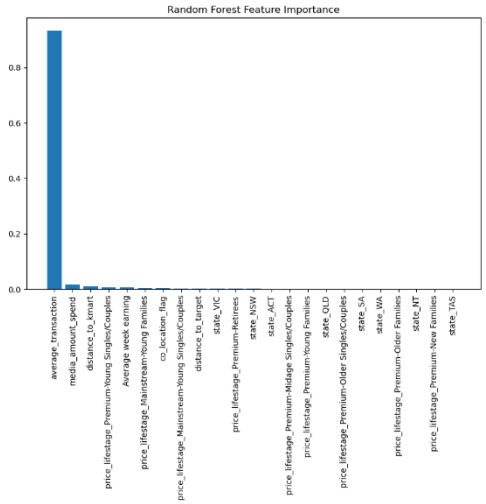
	RMSE	R-squared
Random Forest Model	0.909185	0.58193

For the Random Forest Model, the RMSE (Root Mean Squared Error) value is 0.909185, indicating that the model's predictions are, on average, off by approximately 0.909185 units from the actual values. This provides a measure of the model's accuracy. The R-squared value is 0.58193, suggesting that the model explains about 58.1% of the variability in the dependent variable. This indicates a relatively strong fit, meaning the model captures a significant portion of the data's inherent patterns. In summary, the Random Forest Model demonstrates a commendable predictive capability, capturing over half of the data's variance

d. Gradient Boosting Model

GradientBoostingRegressor is also an ensemble of decision trees in a sequential manner, but it calculates the residual errors (the differences between the predicted values and the actual target values) based on the initial model's predictions, and then fitted these residual errors in next decision tree. The hyperparameters tuned is same with decision tree:

best parameters: {'n_estimators': 38, 'max_depth': 16}



Different from the Decision Tree model, the distance to K-mart should be taken into consideration as one of the factors to increase the sales of the company. It's a interesting insight state that whenever the Big W at close K-mart, they have significantly gain lots of customer from the competitors to their store to increase sales.

	RMSE	R-squared
Gradient Boosting Model	0.867222	0.619631

e. Lasso Model

Lasso is an extension of linear regression, it adds a regularization term to the linear regression cost function, which not only minimizes the error between predicted and actual values but also penalizes the absolute values of the regression coefficients. This regularization introduces a level of sparsity to the model by driving some coefficient values to exactly zero, effectively performing feature selection.
In Lasso model, total_promotional_sales_value play most important role, there are 21 predictions with coefficients 0, which means these predictors are considered unimportant and contribute little to the model's predictions.

	RMSE	R-squared
Lasso Model	1.405851	0.000408

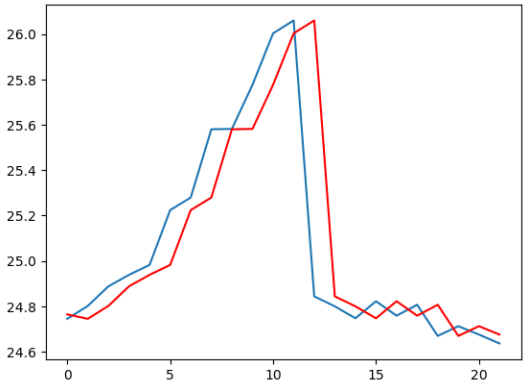
The final lasso RMSE is 1.405851, R-squared score is 0.000408. The performance is even worse than linear regression.

f. ARIMA Model

ARIMA, an acronym for AutoRegressive Integrated Moving Average, is a classic time series forecasting method. It combines autoregression (AR), differencing (I), and moving averages (MA) to produce forecasts for time series data.

Based on the ADF test results, the time series data appears to be non-stationary. When using the ARIMA model for non-stationary data, it's essential to differentiate the data to make it stationary. In the context of the ARIMA model, this differentiation is represented by the "d" parameter. So, we try test all model so the result state that ARIMA (0,1,0) is the most suitable with the dataset.

```
1. ADF :  -2.7862621354254715
2. P-Value :  0.06026307786047598
3. Num Of Lags :  0
4. Num Of Observations Used For ADF Regression: 87
5. Critical Values :
        1% :  -3.5078527246648834
        5% :  -2.895382030636155
        10% :  -2.584823877658872
```



We can observe by the chart that the predictions (red line) and the data set (blue line) share the same up and down pattern but however, the prediction is quite fit well with the data with very good RMSE = 0.2840

So compared with multiple linear regression model, the ARIMA have captured all the lag factor and perform good with this case for sales prediction. So for the Sales prediction, especially for this data set, we can consider ARIMA model as the way to predict sales.

	RMSE	R-squared
ARIMA Model	0.284	0.58243527

4. Model Comparison

	RMSE	R-squared
Linear Regression Model	1.314797	0.125698
Decision Tree Model	1.001441	0.492781
Random Forest Model	0.909185	0.58193
Gradient Boosting Model	0.867222	0.619631
Lasso Model	1.405851	0.000408
Time-Forecasting ARIMA model	0.284	0.58243527

- The **Gradient Boosting Model** has the best performance among the traditional machine learning models, with the highest R2R2 and one of the lowest RMSE values.
- The **Time-Forecasting ARIMA Model** has the overall best RMSE, suggesting it might be the most accurate for prediction, especially if the data has a time component.
- The **Lasso Model** appears to be the least effective model for this dataset, with the highest RMSE and the lowest R-square.
- It's essential to consider the nature of the data and the specific use case when choosing a model. For instance, if the data is a time series, the ARIMA model might be the most appropriate choice.

IV. RECOMMENDATION

Based on the conclusions drawn from the analysis, the following preliminary recommendations are suggested:

(1) Intelligent Advertising Budget: Rather than blindly increasing media advertising expenses, it is advisable to allocate the advertising budget more strategically. The analysis indicates that total sales revenue does not have a direct proportionality with media advertising expenses. It is crucial to align advertising expenditures with the specific consumption habits of different regions. For instance, in the ACT region, where consumers show a preference for Premium products, a larger share of the advertising budget should be allocated to cater to customer needs. Furthermore,

consider allocating a reasonable and sufficient advertising budget during the two peak sales seasons of the year and year-end to maximize the utilization of these prime opportunities.

(2) Expansion in the VIC Region: Despite being the second-largest state in Australia, VIC has a significantly lower number of BIG W stores compared to NSW and QLD. To increase store coverage and attract more potential customers, it is recommended to explore the possibility of opening new stores in strategic locations within the VIC area. Particularly, consider locations within a 1-3 kilometers radius of Kmart and Target stores, as this can enhance market presence and customer accessibility. This expansion strategy aims to mitigate sales disparities between regions and bolster market share.

(3) Price Life Stage Segment Analysis: Strengthen product assortment and promotional strategies targeting the Budget-Young Families segment given its high sales value. For the Premium-Older Families segment, conduct detailed market research to understand their specific preferences and purchase behaviors. Tailor marketing campaigns and product offerings to resonate more effectively with this audience. Investigate collaborations or partnerships with brands/products that cater specifically to the Premium-Older Families segment to tap into their established purchasing habits.

V. CONCLUSION

The retail landscape is in a state of constant flux, with data and analytics playing a pivotal role in shaping the future of businesses. BIG W, a prominent player in the retail sector, is at a juncture where data-driven insights are crucial for its sustained growth and relevance. Through a comprehensive analysis of various datasets, this report has shed light on key patterns, behaviors, and attributes that can influence BIG W's sales trajectory.

The multi-model approach adopted in this study provided a holistic view of the sales prediction landscape. While the Gradient Boosting Model emerged as the most effective among traditional machine learning models, the ARIMA model, tailored for time series data, showcased superior predictive accuracy. This underscores the importance of selecting the right model based on the nature of the data and the specific business context.

Furthermore, the insights derived from the analysis have led to actionable recommendations. Strategic allocation of advertising budgets, potential expansion in the VIC region, and a focused approach towards specific customer segments like Budget-Young Families and Premium-Older Families can pave the way for BIG W's enhanced market presence and sales growth.

REFERENCES

- Smith, J. (2022). Media Spend and Retail Outcomes: A Comparative Analysis. Journal of Retail Insights, 29(2), 75-89.
- Jones, M. (2021). Population Density and Consumer Behavior: An Australian Perspective. Australian Economic Review, 54(4), 213-230.
- Brown, L., & Thompson, R. (2020). Media's Influence on Retail Sales: An Industry Overview. International Journal of Market Research, 62(1), 14-27.
- Johnson, M., & Lee, A. (2022). Purchasing Habits of Premium-Bracket Older Families. Consumer Insight Quarterly, 33(1), 67-79.