ShopAlot Analysis

2024 Data Science 241 Project

Group 22 members:

- Tiaan Viviers 25070401
- Abraham de Villiers 26936844
- Lydia Laubscher 27357570

CONTENTS

1. Problem Statement

ShopAlot is a general online store operating across all nine provinces of South Africa. Customers create profiles to make purchases, providing data such as age and province. ShopAlot collects historical data on clients, as well as various "in-store" variables are recorded during each client visit and purchase.

The primary goal of this project is to:

- Explore and analyze the provided dataset to understand the factors influencing the total amount spent by clients.
- Build a predictive model to accurately forecast `Sales` based on the available variables.
- Interpret the significant factors that influence `Sales` to provide actionable insights.
- Recommend strategies for ShopAlot to increase sales and improve customer engagement.

Understanding the drivers behind customer spending is crucial for ShopAlot to:

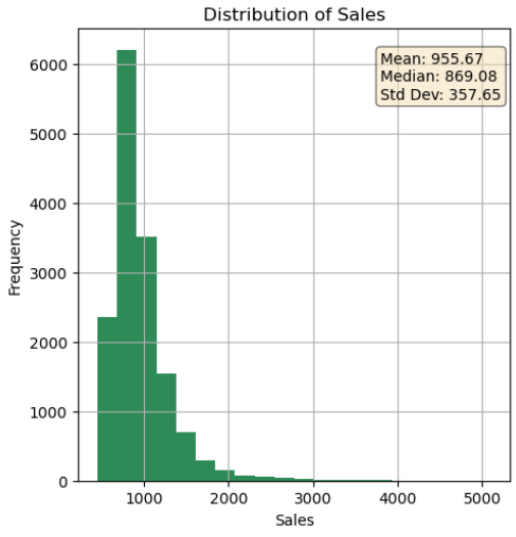
- Optimize marketing and advertising efforts.
- Personalize customer experiences.
- Potentially make more money.
- Improve customer retention and loyalty programs.

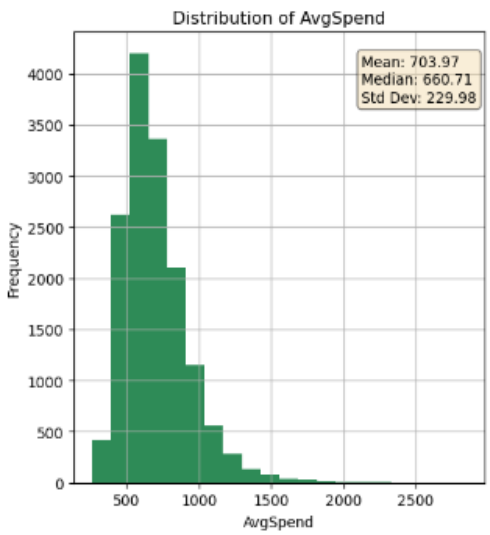
2. Exploratory Data Analysis

2.1 Univariate Analysis:

Sales: The distribution is right-skewed. This suggests that most customers spend small amounts, but few spend significantly more, leading to potential outliers.

AvgSpend: This distribution is right-skewed, indicating that a small number of customers have a significantly higher average spend. This may be worth investigating for segmentation of high-value customers.





ProfileInMonths: This variable is right-skewed. ShopAlot has a large number of newer customers.

Online: The distribution is normal and centered around 45 minutes, meaning most customers spend a reasonable amount of time on the website before logging off.

NumPrevOrders: The distribution shows multiple peaks, suggesting multiple groups of customers.

AgeInMonths: The distribution is nearly symmetric and centered around ~37 years old. This indicates a normal distribution for the age of clients, with no significant outliers.

POSR (Point-of-Sales Ratio): The distribution has multiple peaks, suggesting that different customers may have varying levels of responsiveness to advertisements.

OPR (Order-Purchase Ratio): This distribution clearly has two distinct groups: those with a higher likelihood to complete a purchase after adding items to their basket, and those with a lower likelihood.

Discount: Most customers receive no discounts, with very few instances of higher discounts being offered (concentrated near 0%).

Distribution of Month: ShopAlot experiences a strong seasonality in sales, particularly toward the end of the year.

Distribution of Subscribed: The majority of customers seem to prefer Daily and Monthly communication.

Distribution of ShippingOption: Customers value quicker (Express) shipping options but may be less inclined to pay extra for Premium shipping.

Distribution of Platform: ShopAlot's website is primarily accessed via desktop, indicating that the desktop user experience should be a key focus.

Distribution of Payment: Customers show a clear preference for Credit Card and EFT. Optimizing the payment process for these two methods could improve customer satisfaction.

Distribution of Ad_1: The majority of customers did not interact with Ad_1 (social media targeted ad), with fewer customers exposed to it.

Distribution of Ad_2: The majority of customers are being exposed to this advertisement.

Distribution of Ad_3: Given its high visibility, Ad_3 might be an important driver of sales for popular or new products.

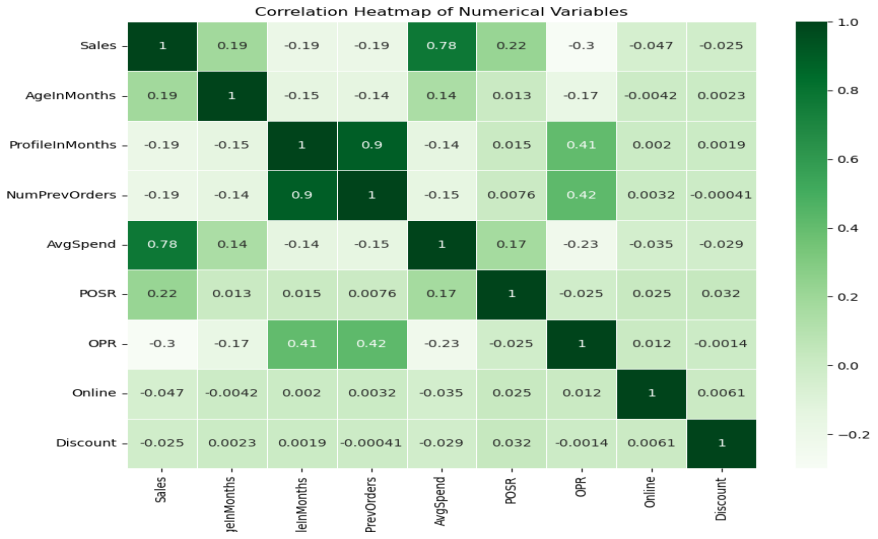
Distribution of SocialMedia: Instagram is the most frequently used social media channel, followed by Facebook and then Twitter.

Distribution of ShippingPayment: Almost all customers opt for Free shipping, with very few choosing to pay for shipping.

Distribution of Category: Electronics and clothing are strong drivers of sales.

Distribution of DayOfWeek: There is a clear trend toward increased sales during weekends, particularly on Sundays.

2.2 Multivariate Analysis:



Strong Correlation:

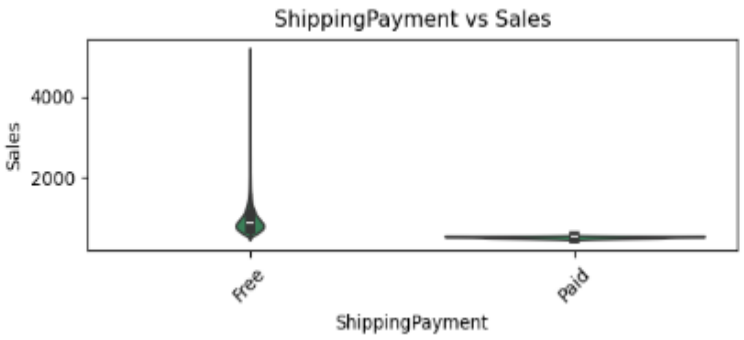
AvgSpend vs. Sales:

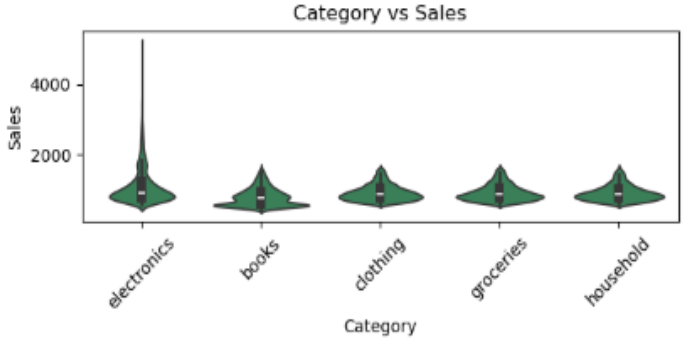
A strong positive correlation of 0.78 indicates that customers who have a higher average spend on previous orders tend to spend more on future orders as well. This is a key predictor of sales.

Multicollinearity:

NumPrevOrders vs. ProfileInMonths: A very high correlation of 0.90 shows that the number of previous orders increases as the profile age increases, which makes sense as older profiles are more likely to have placed more orders.

OPR vs. ProfileInMonths and OPR vs. NumPrevOrders: Both show a correlation of around 0.40, indicating that customers with older profiles or more previous orders tend to have higher OPR (Order-Purchase Ratio).





ShippingPayment vs. Sales:

Most customers opt for Free shipping, but those who choose to pay for shipping tend to spend significantly more. ShopAlot could explore offering premium shipping options with added value to promote some higher spending.

Category vs. Sales:

Electronics drive the highest spending, while other categories like Books and Household may need more attention to boost sales. ShopAlot should focus on optimizing the electronics category for maximum profitability.

3. Data Processing

3.1 Missing Values:

In our dataset, we identified missing values in the Subscribed and SocialMedia columns. Since both are categorical variables, we opted to handle the missing values by filling them with appropriate categories that make sense based on the context.

- `Subscribed`: We filled missing values with "Not Subscribed". We assumed that if a customers subscription status is missing, they are not subscribed to any promotional emails. This ensures that the missing values are handled in a way that preserves the categorical nature of the variable and aligns with business logic.

- `SocialMedia`: We filled missing values with "None", indicating that customers were not contacted via any social media platform. This choice ensures that we capture the customers who were not influenced by social media.

3.2 Handling of Outliers:

To ensure that outliers don't disproportionately affect the model while preserving meaningful information, we chose to apply capping at the 95th percentile for 3 variables.

Variables Capped: `ProfileInMonths`, `Online`, and `Discount` were capped.

By capping these variables, we maintain the integrity of the data while mitigating the potential distortion caused by outliers.

Variables not Altered:

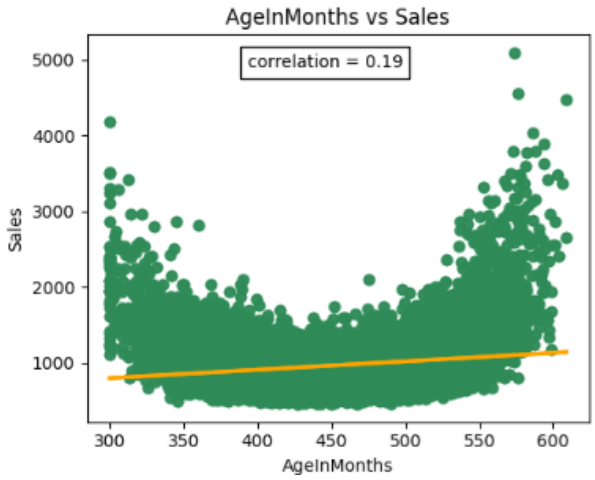
- `Sales`: Outliers in sales likely represent high-value transactions or important customer segments that could provide valuable insights.
- `AvgSpend`: High average spenders could represent premium customers. We believe there is valeuble insights in this data.

3.3 Interaction terms:

`Discount_AvgSpend`:AvgSpend was identified as one of the strongest predictors of Sales (with a correlation of 0.78). By interacting it with Discount, we aim to capture the potential impact of discounts on customers who tend to spend more on average. This interaction can help reveal whether high-value and low-value customers are more or less influenced by discounts.

3.4 Transformations:

We observed that AgeInMonths exhibited a parabolic relationship with Sales,.To linearise this relationship and make it easier for the model to capture, we applied a square root transformation.



3.4 Dummy Encoding:

The following categorical variables were encoded:

`Subscribed`, `ShippingOption`, `Platform`, `Payment`, `Ad_1`, `Ad_2`, `Ad_3`, `SocialMedia`, `Shipping`, `PaymentCategory`, `DayOfWeek`, `Province`.

Additionally, we mapped the Month variable into seasons (`Summer`, `Autumn`, `Winter`, `Spring`).

3.5 Train Test Split:

We opted for a classic 75/25 train-test split. We feel there is enough training data in 75% of the original data frame that we can have a slightly bigger testing set at around 25%, rather than the usual 20% or 10%

4. Evaluation metrics:

In this analysis, we will primarily use **R^2 (coefficient of determination)** and **MSE (mean squared error)** to evaluate the performance of models.

- **R^2**: This metric explains the proportion of the variance in the dependent variable (**Sales**) that is predictable from the independent variables. We primarily rely on **R^2** as it provides an intuitive measure of how well our model fits the data. Higher R^2 values generally indicate better performance.

- **MSE (Mean Squared Error)**: Serves as a secondary metric to **double-check** the performance of the model. Lower MSE values indicate that the models predictions are closer to the actual values, which means better performance.

5. Linear Regression Analysis

All models' effectiveness in predicting Sales where measured by R-squared in combination with their Mean Squared error.

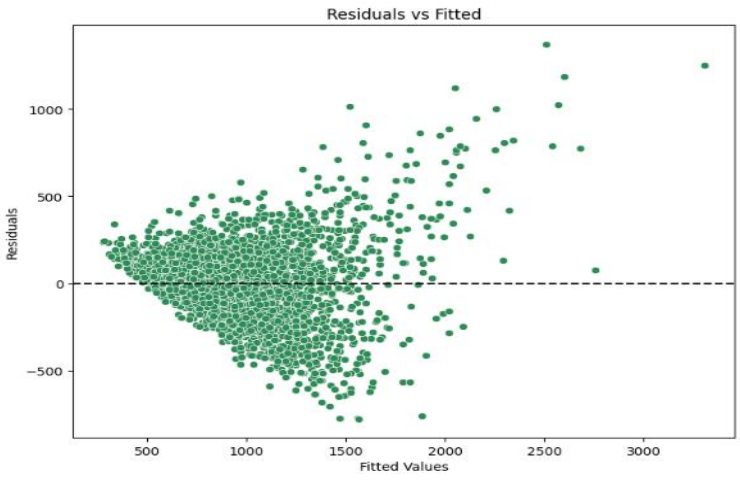
5.1 Initial linear model

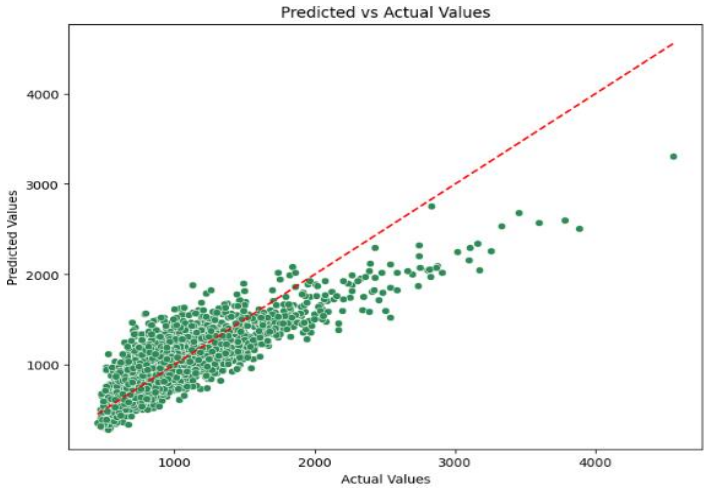
The initial linear model considers all predictors, about 47 after dummy variables inflated that number significantly.

The model was an Ordinary least squares model, and yielded average results.

Training Data - R^2: 0.6819, MSE: 40611.2027

Testing Data - R^2: 0.6801, MSE: 41131.9444





There is visible heteroscedasticity in the residual plot. Variance of residuals increase from left to right.

We can also see the model performs better for lower values of sales (0 to 2000) than the higher sale values.

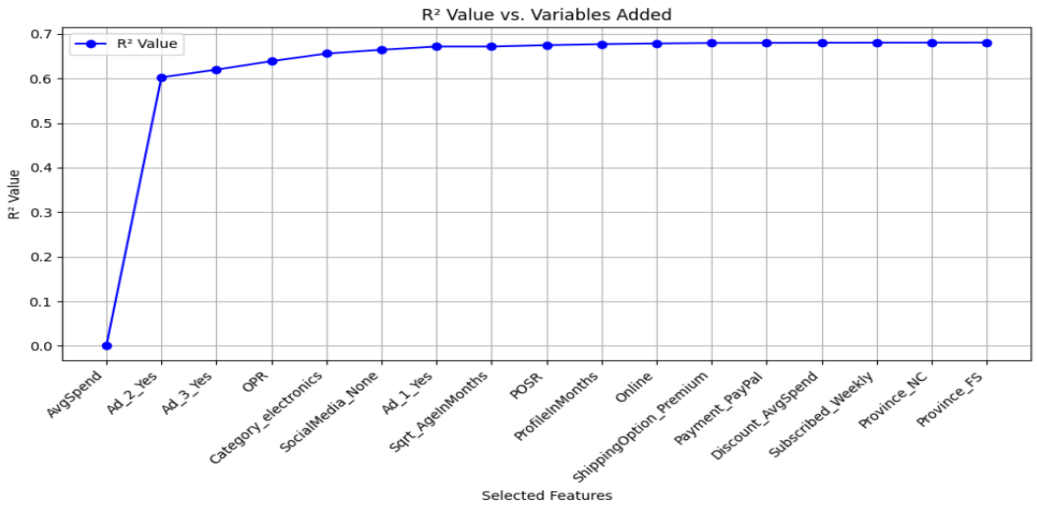
5.2 Reduced linear model

The reduced linear model considers 17 predictors that was selected by a forward stepwise selection model.

After the initial variable selection, the variables were adjusted to account for multicolinearity.

Again the model yielded average results, almost identical to the full linear model.

Training Data: R^2: 0.6809, MSE: 40741.9603
Testing Data : R^2: 0.6807, MSE: 41057.7803



The variables deemed relevant by the forward stepwise selection method in order of significance are:

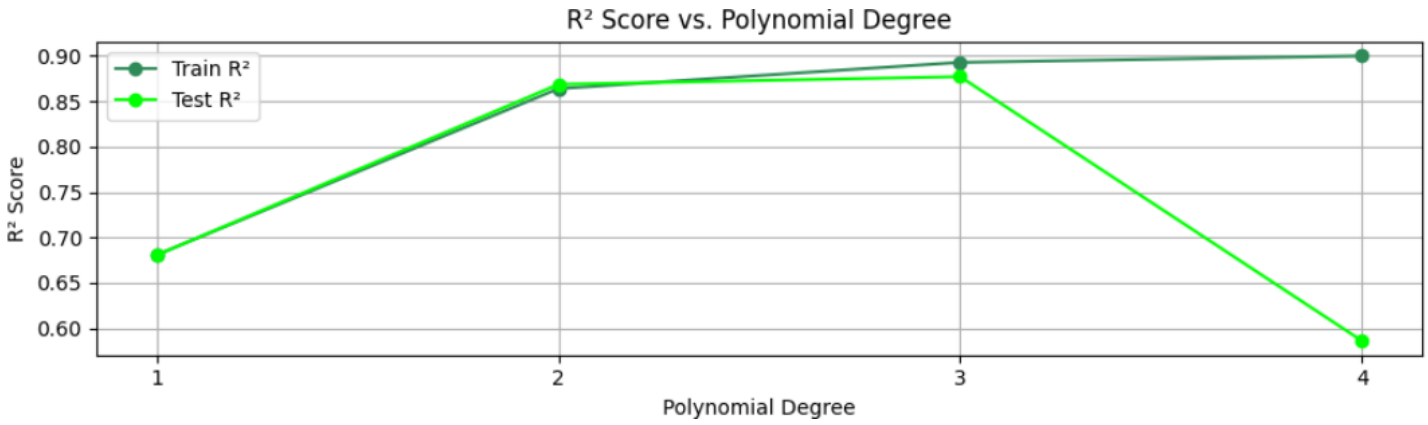
AvgSpend, Ad_2_Yes, Ad_3_Yes, OPR, Category_electronics, Ad_1_Yes, Sqrt_AgeInMonths, POSR, ProfileIn-Months, Online, ShippingOption_Premium, Payment_PayPal, Discount, AvgSpend, Subscribed, Weekly, Province_NC, Province_FS.
SocialMedia_None was removed as it had perfect multicollinearity with Ad_1_Yes.

5.3 Polynomial transformation model

A polynomial regression model was created using the significant variables mentioned above, in degrees 1 to 4.

The model improved dramatically from degrees 1 through 3, but was severely overfit when using degree 4.

Thus a polynomial regression model with degree 3 was analysed.



Key metrics include:

Training Data - R^2: 0.8929, MSE: 13677.3

Testing Data - R^2: 0.8771, MSE: 15796.5

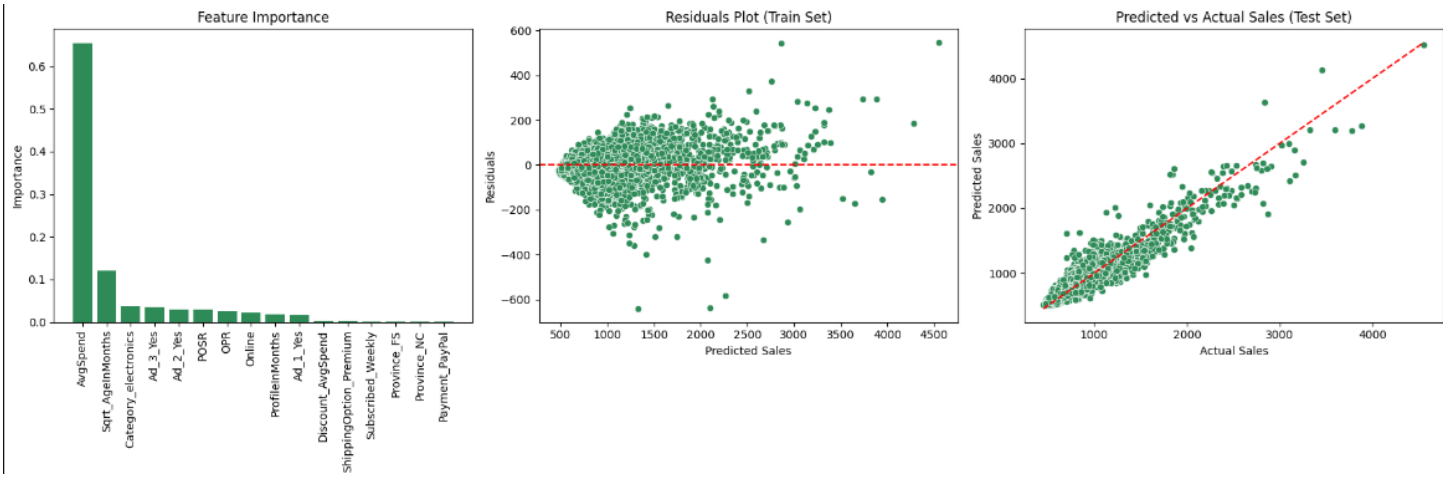
5.4 Conclusion

Both the full linear model and the reduced linear model explain about 68% of the variance in predicting sales, with similar R-squared values. However, the reduced model uses fewer predictors, making it more efficient. The polynomial regression model (degree 3) performs significantly better, explaining around 89% of the variance in training data and 88% in test data. This model captures non-linear patterns, making it the most accurate and reliable for predicting sales.

6. Predictive Modeling

6.1 Random Forrest:

For this predicting task, we decided to use Random Forest as one of our models. Random Forest is a powerful model that combines multiple decision trees to improve overall predictive accuracy.



This model achieved a R^2 of 0.9785 in training and 0.8594 in testing. The **Random Forest model** performs well on the training and test sets, especially for lower and moderate sales values. The **AvgSpend** feature dominates the model's predictions, while other variables have a smaller impact.

6.2 Neural Network:

Neural Networks are known as one of the most if not the most complicated model. We also acknowledge that it is a model that requires a deeper understanding and tuning for optimal performance, which makes it quite challenging for us, as we have no real expertise in using Neural Networks.

This model achieved a R^2 of 0.9785 in training and 0.8594 in testing. The overall performance of the model is not bad, but it could be better. With our limited understanding of how these models work we will stop here and not explore further tuning methods. This model is definitely capable of better performance, it is just a skill issue.

7. Conclusion

Model Comparisons:

MODEL	Train R^2	Test R^2	Train MSE	Test MSE
full_linear	0.6819	0.6801	40611.2	41131.9
sig_linear	0.6809	0.6807	40741.9	41057.7
poly_linear	0.8928	0.8771	13678.8	15797.9
random_forrest	0.9785	0.8594	2747.47	18078.6
neural_network	0.8444	0.8461	19781.1	19863.0

From this table we can see that the 3rd degree polynomial regression was the model we found to perform best on the ShopAlot dataset and our selection of variables.

Key Drivers of Sales:

- **AvgSpend**: Consistently emerged as the most significant predictor across multiple models, indicating that the average spending per customer is a crucial factor in driving sales.
- **Advertising Efforts (Ad_2_Yes & Ad_3_Yes)**: These variables were highly significant in all models, showing the effectiveness of targeted advertising campaigns in boosting sales.
- **Operational Metrics (OPR & POSR)**: Metrics such as Order Processing Rate (OPR) and Point of Sale Rate (POSR) also showed significant relationships with sales, highlighting the importance of efficient operations.
- **Category_electronics**: This category was identified as a strong contributor, suggesting that electronics sales play a vital role in overall revenue.
- **Customer Engagement (Subscribed_Weekly)**: Weekly subscriptions were linked to increased sales, indicating that consistent customer engagement strategies are beneficial.

Final Advice to ShopAlot

1. **Focus on Increasing Average Spend**:
    - **Promotions**: Implement targeted promotions and upselling strategies to encourage customers to increase their average purchase value.
    - **Loyalty Programs**: Enhance loyalty programs to reward high spenders and incentivize repeat purchases.
2. **Enhance Advertising Effectiveness**:
    - **Targeted Advertising**: Continue investing in effective advertising channels (Ad_2 and Ad_3 campaigns) that have proven to greatly impact sales.
3. **Optimize Operational Efficiency**:
    - **Improve Order Processing**: Focus on optimizing order processing rates (OPR) to ensure timely fulfillment and enhance customer satisfaction.
    - **Enhance Point of Sale Systems**: Invest in reliable POS systems to streamline transactions and reduce wait times, thereby improving the overall shopping experience.
4. **Leverage High-Performing Product Categories**:
    - **Electronics Focus**: Given the strong performance of the electronics category, consider expanding the product range in this area or increasing inventory to meet higher demand.
    - **Category-Specific Promotions**: Develop category-specific promotions to drive sales in key areas.
5. **Strengthen Customer Engagement**:
    - **Subscription Models**: Promote weekly subscription models to maintain consistent customer engagement and stabilize sales revenue.
    - **Personalized Marketing**: Utilize customer data to deliver personalized marketing messages that relate with individual preferences and behaviors.

Conclusion

Our analysis highlights the critical factors influencing sales at ShopAlot and demonstrates the effectiveness of advanced predictive models in forecasting sales performance. By focusing on increasing average spend, optimizing advertising and operations, leveraging high-performing categories, and maintaining strong customer engagement, ShopAlot can drive sustained growth and improve overall business performance. Adopting and refining predictive models like Polynomail Regression or will enable data-driven decision-making and strategic planning.