# 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.

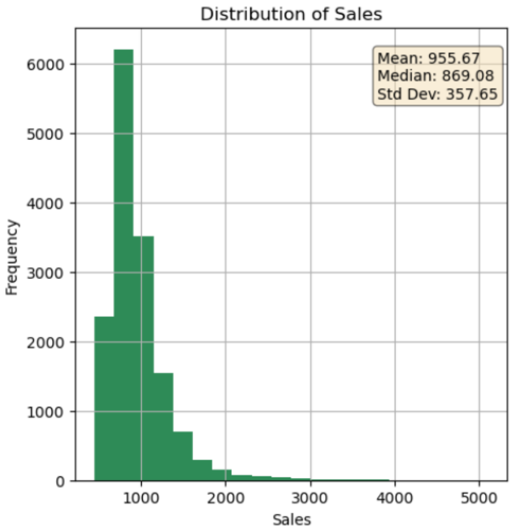
## 2. Exploratory Data Analysis

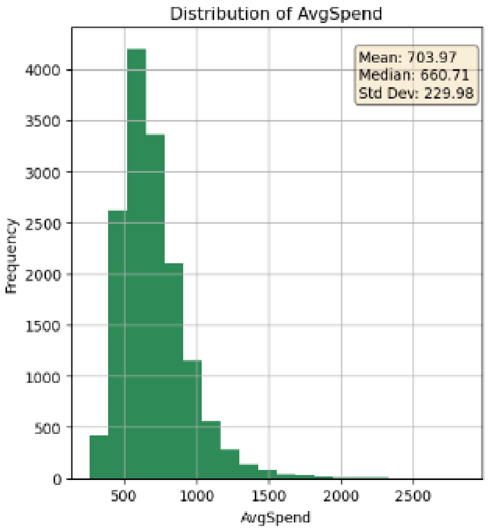
### 2.1 Univariate Analysis:

Here we shortly discuss the distribution of our target variable and strongest predictor.

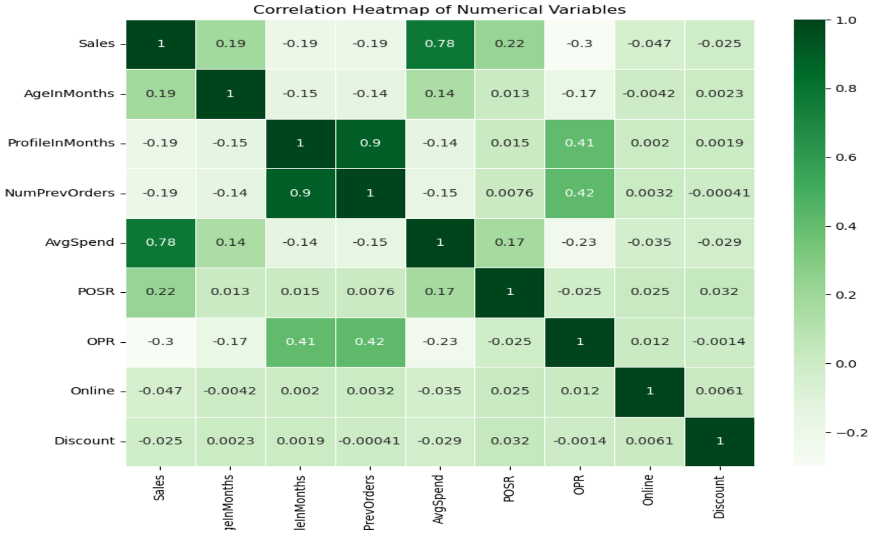
Sales: The distribution is right-skewed. This suggests that most customers spend small amounts, but few spend significantly more, leading to potential outliers.

AvgSpend: This distribution is right-skewed, indicating that a small number of customers have a significantly higher average spend.





## 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.

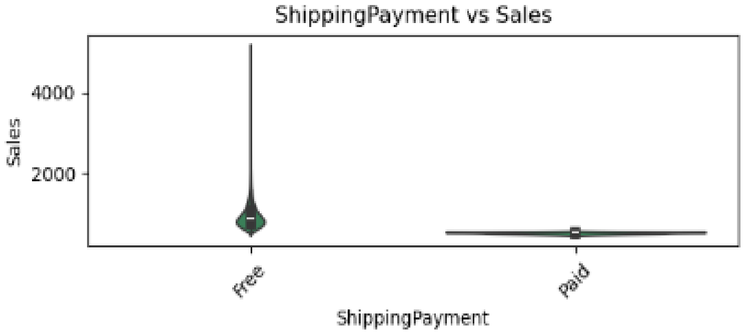
**Multicollinearity:**

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).

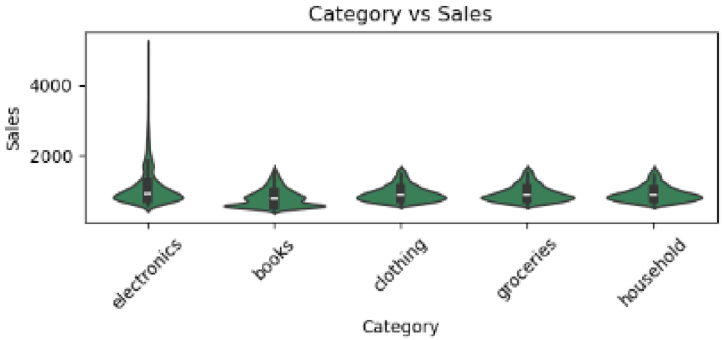
ShippingPayment vs. Sales:

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:

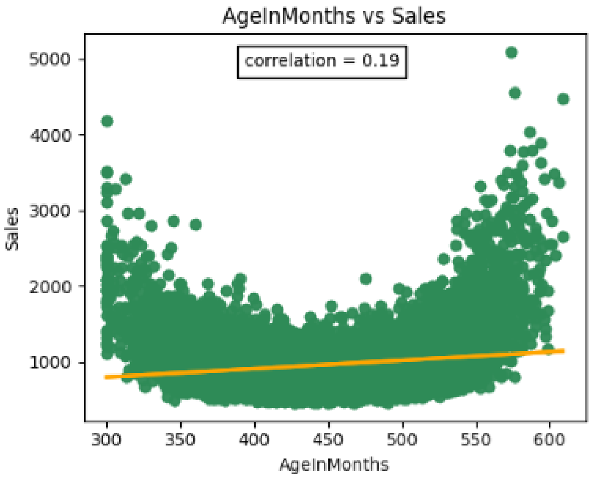
- `Sales`: Outliers in sales likely represent high-value transactions or important customer segments that could provide valuable insights.
- `AvgSpend`: High average spenders could represent premium customers. We believe there is valeuble insights in this data.

### 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.

- **$R^2$**: explains how much of the variance in **Sales** is accounted for by the independent variables. Since our goal is to understand and predict **Sales**, a higher $R^2$ indicates a model that captures key patterns, making it the primary metric for assessing model fit.
- **MSE** : serves as a secondary metric to ensure that our model's predictions are close to the actual sales values. We opted for MSE as our second metric as the value of the MSE is in the same unit as the target variable, giving us a better intuitive feeling for the size of the error.

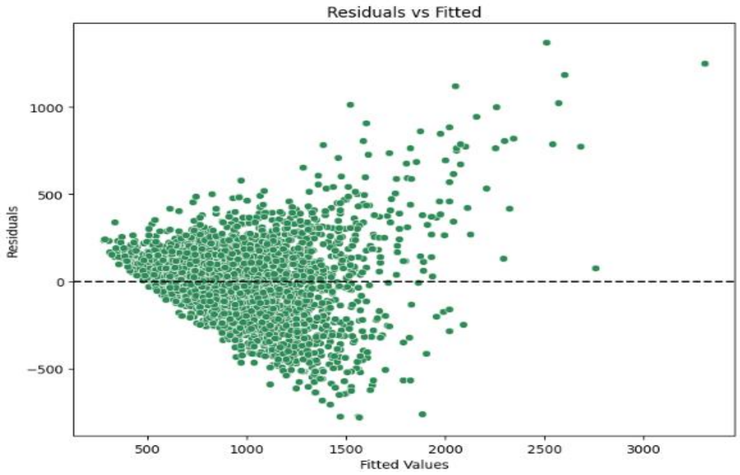
## 5. Linear Regression Analysis

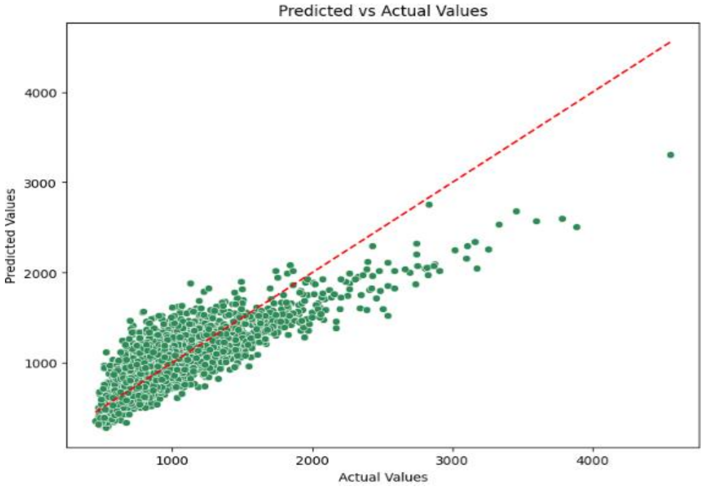
All models' effectiveness in predicting Sales where measured by R-squared in combination with their Mean Squared error.

### 5.1 Initial linear model

The initial linear model considers all predictors, about 42 after dummy variables inflated that number significantly. An Ordinary least squares model was fitted, and yielded average results. We then decided to implement some transformations and interaction terms in our preprocessing stage, as well as changing our approach to dummy variable encoding. This brought us to 47 predictors with marginally better results shown below.

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 as sales gets larger.

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

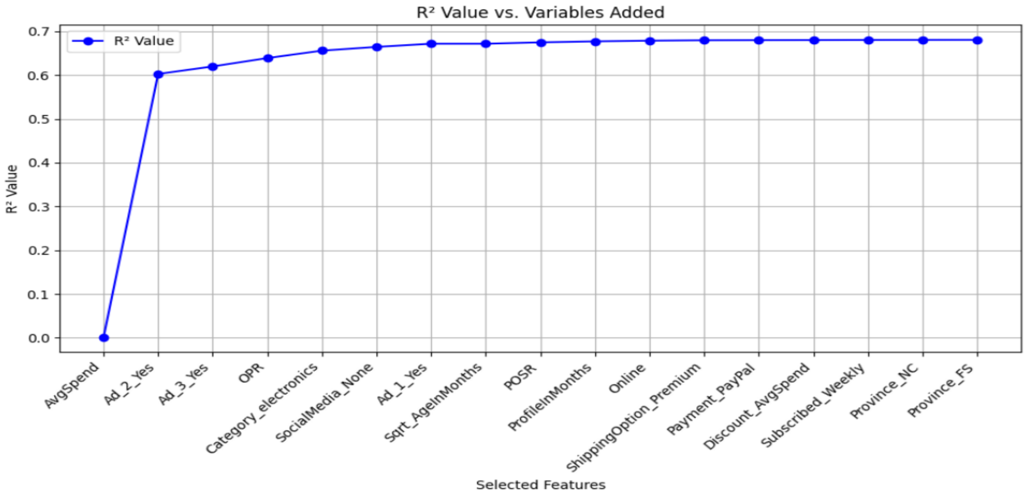
### 5.2 Reduced linear model

We then decided to build a linear model with only the most significant predictors. 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 multicollinearity.

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

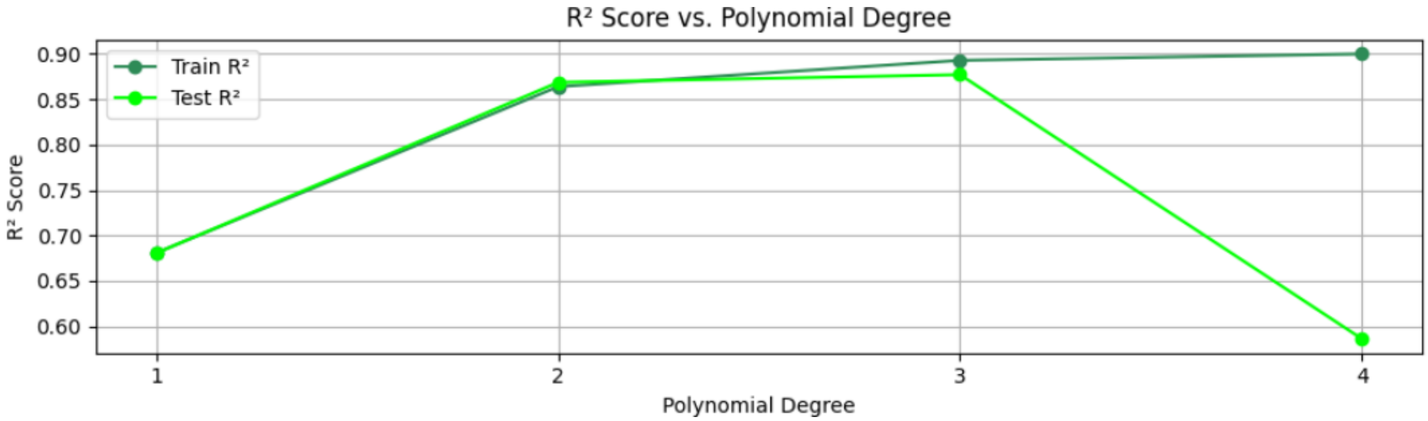
The variables deemed relevant by the forward stepwise selection method in order of significance are:
AvgSpend, Ad_2_Yes, Ad_3_Yes, OPR, Category_electronics, Ad_1_Yes, Sqrt_AgeInMonths, POSR, ProfileInMonths, Online, ShippingOption_Premium, Payment_PayPal, Discount, AvgSpend, Subscribed, Weekly, Province_NC, Province_FS.
SocialMedia_None was removed as it had perfect multicollinearity with Ad_1_Yes.



## 5.3 Polynomial transformation model

After examining various plots we saw that a polynomial function might be a good fit. We proceeded to experiment. A polynomial regression model was created using the significant variables mentioned above, in degrees 1 to 4. The polynomial transformation instantly yielded results.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 further. Results for 3rd degree polynomial:

Training Data - $R^2$: 0.8929,   MSE: 13677.3   |   Testing Data - $R^2$: 0.8771,   MSE: 15796.5



## 5.4 Model Coefficients

Since our best performing model was the polynomial regression, this makes interpretation of coefficients very complicated. We will shortly discuss the most significant predictors' coefficients, using it as an inference model to gain business insights.

- **AvgSpend = 0.98**: For every unit increase in average spending, sales increase by nearly 1 unit, highlighting the strong relationship between customer spending and sales.
- **Ad_2_Yes = 145.91**: Running Ad_2 has a significant positive impact on sales, increasing them by 145 units, making it a highly effective advertising strategy.

- **Ad_3_Yes** = **122.60**: Ad_3 also boosts sales significantly, by around 122 units, suggesting it is another successful campaign.
- **OPR = -175.65**: A higher OPR is associated with a decrease in sales, showing that issues in order processing reduce sales performance.
- **Category_electronics = 70.34**: Selling electronics increases sales by around 70 units, showing it's a key product category for growth.
- **SocialMedia_None = -53.99**: Customers who do not interact with social media platforms result in lower sales, indicating the importance of social media engagement.
- **Sqrt_AgeInMonths = 13.42**: Older customer profiles are associated with slightly higher sales, as represented by this positive and significant coefficient.

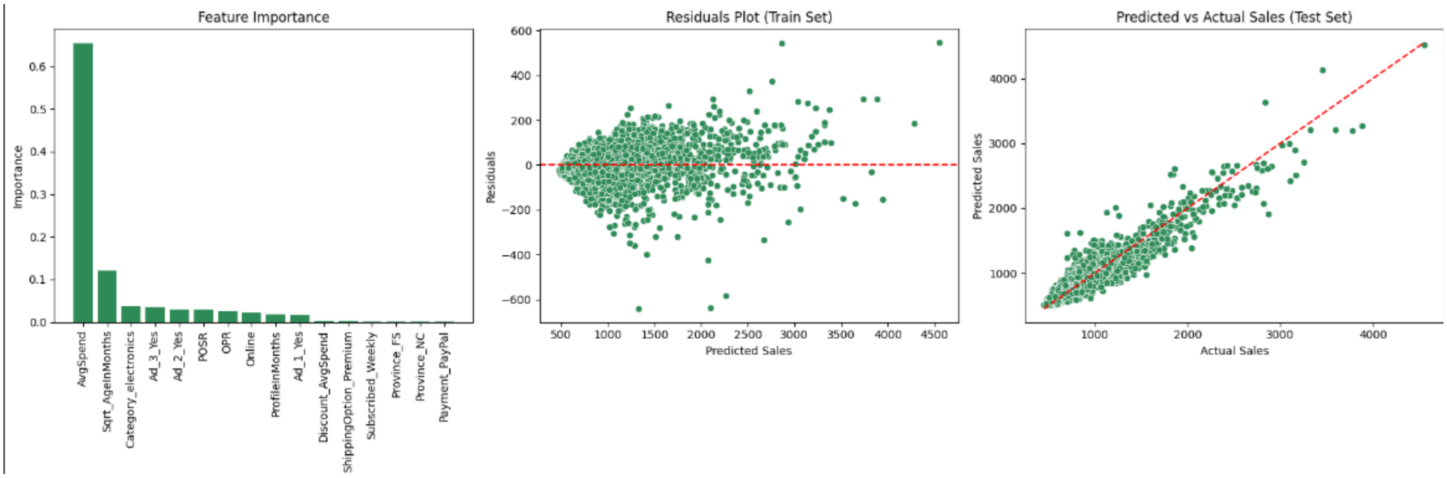
### 5.5 Conclusion

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.

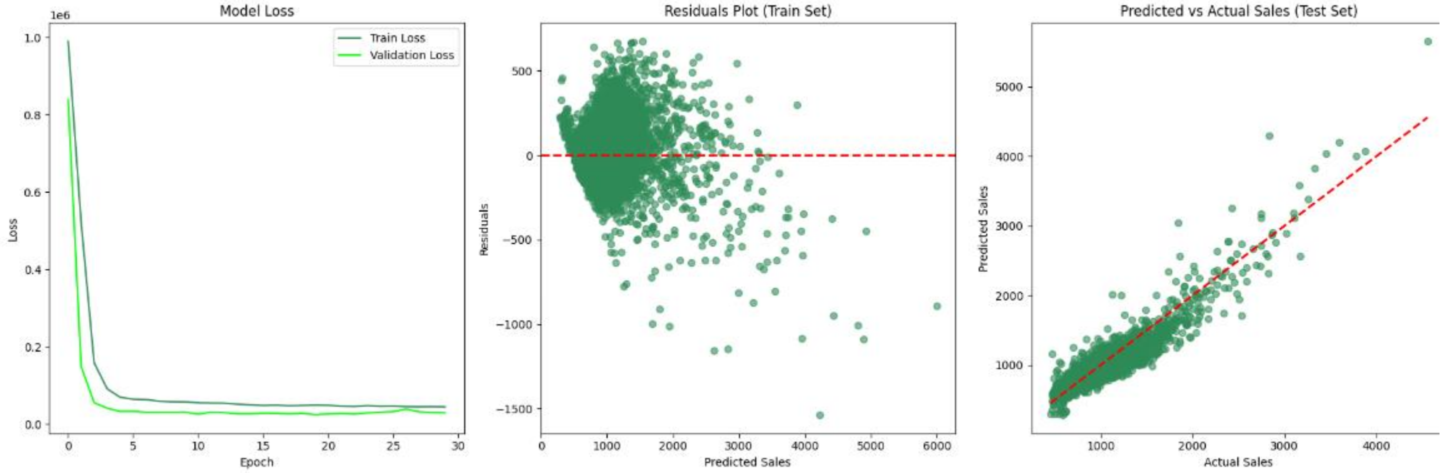


This model achieved a $R^2$ of 0.9785 in training and 0.8594 in testing. The **Random Forest model** performs well for lower and moderate sales values. The **AvgSpend** feature dominates the model's predictions, while other variables have a smaller impact. The residuals are centered around zero, which is a good indication of a well-fitted model, however there is slight heteroscedasticity. The predictions align well with the actual values, particularly for lower and moderate sales values.

## 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.

This model achieved a $R^2$ of 0.9046 in training and 0.8702 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.



<u>Model Loss (Train & Validation Loss) Over Epochs:</u> The first plot shows the training and validation loss over 20 epochs. Both curves decrease and flatten, which is a positive sign indicating that the model is learning and converging.

<u>Residuals Plot (Train Set):</u> There is some pattern and spread of residuals, indicating that while the model fits well in general, it struggles with certain predictions.

<u>Predicted vs. Actual Sales (Test Set):</u> Most of the points are close to the line, suggesting that the model is making reasonable predictions.

# 7. Concluding Remarks

**Model Comparisons**: *(Table listed below)*

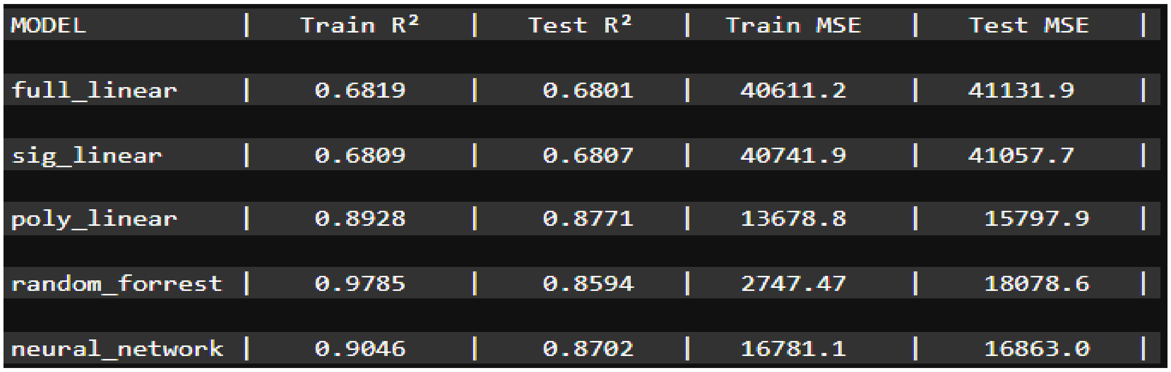
From this table we can see predictive accuracy across different approaches. The **full_linear** and **sig_linear** (significant predictor linear model) models perform similarly, with moderate $R^2$ values and relatively high MSE, suggesting that these models, while interpretable, may not capture the complexity of the data well.

The **poly_linear** model shows a significant improvement in both $R^2$ and MSE, indicating its ability to capture non-linear relationships more effectively, while still maintaining reasonable generalization as seen in the test performance. This was our strongest model.

The **random_forest** model delivers the highest $R^2$ on the training set but shows a notable drop in test performance, signaling potential overfitting. Despite this, it offers strong predictive accuracy overall.

The **neural_network** achieves a high $R^2$ score but a weaker MSE score, with both train and test $R^2$ values being high and MSE being higher than the Polynomial model. With further tuning, this model could be the best predictive model.

In conclusion, while the **Polynomail model** provides the best predictive power, and is a robust option with high interpretability and performance.



```
MODEL          |  Train R²  |  Test R²  |  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 |   0.9046   |   0.8702  |   16781.1   |   16863.0  |
```

**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.