1 Introduction

K-Olive is a well-established regional olive producer with over 30 years of successful operations in Victoria's Wimmera and Grampians regions. Despite consistent financial performance over the past two years, K-Olive anticipates shifts in consumer preferences and seeks to strengthen customer relationships while improving demand planning capabilities.

To support these objectives, this report conducts a comprehensive data analysis focused on understanding customer characteristics, repurchase intentions, and production forecasting. The analysis is structured into four key tasks, each designed to provide strategic insights and actionable recommendations.

Section 3 focused on building a model to predict the order quantity.

Section 4 investigates the presence and impact of interaction effects among selected predictors, supported by visual analysis.

Section 5 develops a predictive logistic regression model to examine how quality perception, brand image, and purchase channels influence repurchase probabilities, accompanied by visualizations and managerial implications.

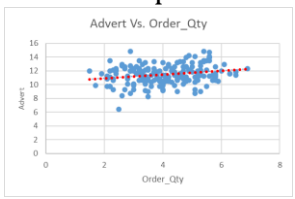
Section 6 applies time-series forecasting methods to predict quarterly production volumes for the next four quarters, aiding operational planning.

This report aims to equip K-Olive with data-driven insights for enhancing customer engagement and optimizing production decisions in response to emerging market dynamics.

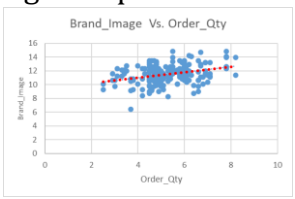
Note: This report leverages a dataset shared by Deakin University and is subject to the university's copyright.

2 Developing Predictive Model

A regression model is used to predict numeric variable[7]. Hence, we are using linear regression to predict order quantity using both numeric and categorical predictors. Below are scatter plots showing each predictor's relationship with the target.



(a) Advert Vs Order



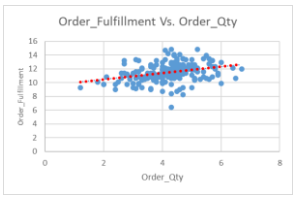
(b) Brand Vs Order

(c) Channel Vs Order

(d) Shipping cost Vs Order

(e) customer type Vs Order

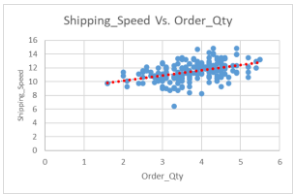
(f) Flexible Price Vs Order



(g) Order Fulfillment Vs Order

(h) Loyalty Vs Order

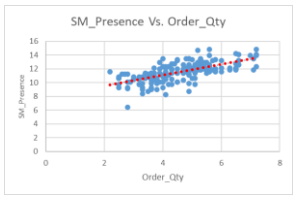
(i) Comp Price Vs Order



(j) Shipping Speed Vs Order

(k) Recommend Vs Order

(l) Region Vs Order



(m) SM Presence Vs Order

(n) Quality Vs Order

(o) Quality Vs Order

Figure 1: Scatter Plot for different attributes

The scatter plots show that quality shows more non linear characteristic Fig 0 compared to linear relationship Fig n. Similarly, order fulfillment Fig g, comp pricing Fig i and shipping speed Fig j show slight non linear relationship but they will further be evaluated through correlation matrix. Remaining attributes are linearly related to the target variable.

To further assess the linear relationship, we can refer to the correlation matrix below:

Figure 2 shows that Shipping Speed and Shipping Cost have a strong correlation, leading to multicollinearity as correlations above 0.8 tend to show multicolinear effects [4]. As a result, Shipping Speed was omitted from the feature selection due to its lower correlation with Order Quantity. Social Media Presence, followed by Shipping Speed,

Shipping Cost, and Recommend, demonstrated a very high correlation with Order Quantity, indicating

	Loyalty	Cust_ Type	Region	Dist_Channel	Quality	SM_Presence	Advert	Brand_ Image	Comp_ Pricing	Order_Fulfillment	Flex_ Price	Shipping_ Speed	Shipping_ Cost	Recommend	Order_ Qty
Loyalty	1.00														
Cust_Type	-0.05	1.00													
Region	0.07	-0.03	1.00												
Dist_Channel	0.22	-0.14	-0.26	1.00											
Quality	0.08	-0.05	-0.54	0.38	1.00										
SM_Presence	0.31	0.07	0.02	0.41	0.24	1.00									
Advert	0.26	0.01	0.22	0.17	-0.05	0.40	1.00								
Brand_Image	0.26	0.01	0.38	0.29	-0.12	0.67	0.63	1.00							
Comp_Pricing	0.08	0.15	0.59	-0.33	-0.45	0.00	0.10	0.20	1.00						
Order_Fulfillment	0.14	-0.01	0.00	0.26	0.08	0.36	0.23	0.28	-0.06	1.00					
Flex_Price	0.06	0.04	0.58	-0.22	-0.49	0.15	0.26	0.27	0.47	0.42	1.00				
Shipping_Speed	0.20	0.02	0.01	0.25	0.07	0.41	0.32	0.30	-0.06	0.77	0.51	1.00			
Shipping_Cost	0.17	0.02	0.00	0.24	0.14	0.42	0.25	0.30	-0.09	0.70	0.36	0.84	1.00		
Recommend	0.18	-0.02	-0.11	0.42	0.35	0.45	0.18	0.30	-0.13	0.31	0.07	0.39	0.37	1.00	
Order_Qty	0.41	-0.05	-0.12	0.40	0.43	0.67	0.24	0.34	-0.22	0.31	0.00	0.43	0.50	0.50	1.00

Figure 2: Correlation Matrix

their strong influence on the target variable. Loyalty, Distribution Channel, Quality, and Brand Image showed moderate influence, contributing to the Order Quantity, but with less impact compared to the other factors.

After several iterations, the first model included all shortlisted attributes. Below are the results.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.81
R Square	0.66
Adjusted R Square	0.64
Standard Error	0.81
Observations	200.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	13.00	235.97	18.15	27.84	0.00
Residual	186.00	121.27	0.65		
Total	199.00	357.24			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	7.31	0.73	10.07	0.00	5.88	8.74	5.88	8.74
Loyalty	0.07	0.01	4.48	0.00	0.04	0.10	0.04	0.10
Cust_Type	-0.12	0.12	-1.04	0.30	-0.36	0.11	-0.36	0.11
Region	0.29	0.18	1.59	0.11	-0.07	0.66	-0.07	0.66
Dist_Channel	-0.04	0.15	-0.30	0.77	-0.34	0.25	-0.34	0.25
Quality	0.18	0.06	2.90	0.00	0.06	0.29	0.06	0.29
SM_Presence	0.61	0.08	7.53	0.00	0.45	0.77	0.45	0.77
Advert	0.00	0.07	-0.01	0.99	-0.13	0.13	-0.13	0.13
Brand_Image	-0.17	0.10	-1.77	0.08	-0.36	0.02	-0.36	0.02
Comp_Pricing	-0.11	0.05	-2.32	0.02	-0.21	-0.02	-0.21	-0.02
Order_Fulfillment	-0.19	0.10	-1.97	0.05	-0.38	0.00	-0.38	0.00
Flex_Price	-0.03	0.08	-0.40	0.69	-0.19	0.13	-0.19	0.13
Shipping_Cost	0.34	0.07	4.77	0.00	0.20	0.49	0.20	0.49
Recommend	0.37	0.14	2.60	0.01	0.09	0.65	0.09	0.65

Figure 3: First Regression Summary Output

The model is statistically significant (F-score ¡ 0.05) with strong explanatory power (R = 0.81). Insignificant variables (e.g., Advert) will be removed iteratively to refine the model.

After several iterations, and removing insignificant attributes, we come to the following model as shown in figure 4 below.

The model is statistically significant with improved F-statistic (from 27.84 to 50.56), but minimal gains in R-squared and error metrics suggest limited improvement in accuracy. Residual analysis is needed to evaluate the adequacy of the fitted model.

Residual plots confirm linearity for key predictors[5], but show skewness in Quality, heteroscedasticity in Competitive Pricing, and non-normal residuals—indicating assumption violations.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.81
R Square	0.65
Adjusted R Square	0.64
Standard Error	0.81
Observations	200.00

ANOVA

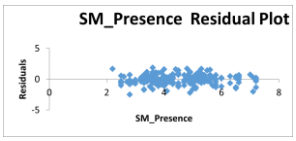
	df	*SS*	*MS*	*F*	*Significance F*
Regression	7.00	231.61	33.09	50.56	0.00
Residual	192.00	125.64	0.65		
Total	199.00	357.24			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6.82	0.64	10.60	0.00	5.55	8.09	5.55	8.09
Loyalty	0.07	0.01	4.60	0.00	0.04	0.10	0.04	0.10
Quality	0.18	0.05	3.65	0.00	0.08	0.28	0.08	0.28
SM_Presence	0.50	0.06	8.16	0.00	0.38	0.62	0.38	0.62
Comp_Pricing	-0.10	0.04	-2.35	0.02	-0.18	-0.02	-0.18	-0.02
Order_Fulfillment	-0.21	0.09	-2.40	0.02	-0.39	-0.04	-0.39	-0.04
Shipping_Cost	0.34	0.07	4.90	0.00	0.21	0.48	0.21	0.48
Recommend	0.33	0.14	2.41	0.02	0.06	0.60	0.06	0.60

Figure 4: Seventh Regression Summary Output

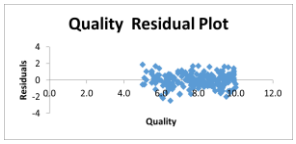
(a) Order fulfillment

(b) Loyalty

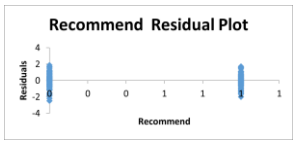


(c) Social Media Presence

(d) Competitive Price

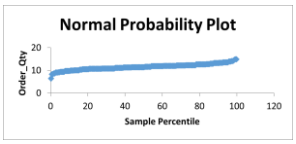


(e) Quality



(f) Recommended

(g) Shipping Cost



(h) Normal Probability Plot

Figure 5: Residual and Normal Probability Plots

Figure 6 shows the final regression predictive model. The regression model explains approximately 65% of the variation in order quantity, with the remaining 35% attributed to factors not included in the model. The overall model is statistically significant, as indicated by an F-statistic p-value less than 0.05, confirming that at least one predictor contributes meaningfully. Additionally, all included variables have p-values below 0.05, establishing their individual significance. The model shows improvement, with the Fstatistic increasing [1] from 27.84 to 50.85. The residual plot shows improvement in quality attribute as shown in figure below:

Intercept (7.48): This is the baseline value of order quantity when all predictor variables are set to zero.

Loyalty (0.07): For every 1 unit increase in Loyalty, the order quantity is expected to increase by 0.07, holding all other variables constant.

Quality² (0.01): Since Quality is squared, this indicates that the relationship between Quality and order quantity is nonlinear. A 1 unit increase in Quality² will increase order quantity by 0.01.

SM Presence (0.49): For each 1 unit increase in Social Media Presence, order quantity is expected to increase by 0.49, holding all other factors constant.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.81
R Square	0.65
Adjusted R Squa	0.64
Standard Error	0.81
Observations	200.00

ANOVA

	df	SS	MS	F	Significance F
Regression	7.00	232.07	33.15	50.85	0.00
Residual	192.00	125.17	0.65		
Total	199.00	357.24			

	Coefficients	Standard Erro	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	7.48	0.51	14.56	0.00	6.47	8.49	6.47	8.49
Loyalty	0.07	0.01	4.68	0.00	0.04	0.10	0.04	0.10
Quality^2	0.01	0.00	3.75	0.00	0.01	0.02	0.01	0.02
SM_Presence	0.49	0.06	8.00	0.00	0.37	0.61	0.37	0.61
Comp_Pricing	-0.09	0.04	-2.25	0.03	-0.18	-0.01	-0.18	-0.01
Order_Fulfillment	-0.21	0.09	-2.41	0.02	-0.39	-0.04	-0.39	-0.04
Shipping_Cost	0.34	0.07	4.87	0.00	0.20	0.48	0.20	0.48
Recommend	0.33	0.14	2.38	0.02	0.06	0.59	0.06	0.59

Figure 6: Final Regression Summary Output

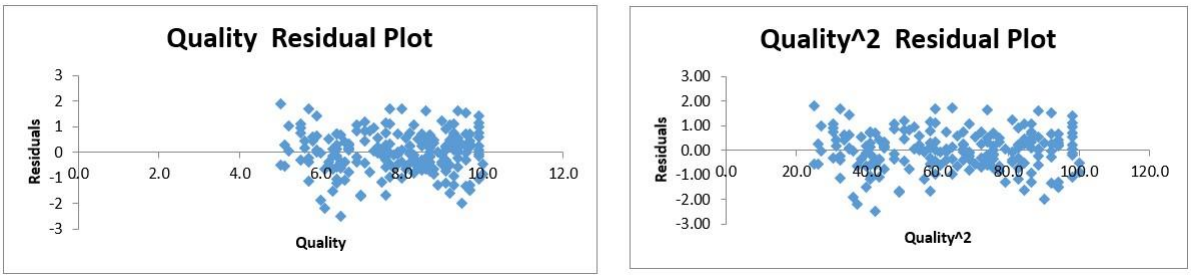


Figure 7: Residual plots of Quality before and after correction

Comp Pricing (-0.09): A 1 unit increase in Competitive Pricing will decrease order quantity by 0.09, suggesting that higher competitive pricing negatively impacts the order quantity.

Order Fulfillment (-0.21): A 1 unit increase in Order Fulfillment will decrease order quantity by 0.21, indicating a negative effect on the dependent variable.

Shipping Cost (0.34): A 1 unit increase in Shipping Cost will result in an increase of 0.34 in order quantity.

Recommend (0.33): A 1 unit increase in the Recommend variable will increase order quantity by 0.33.

Hence, the values can be predicted by the equation as below:

$$\text{Order Quantity} = 7.48 + 0.07 \times \text{Loyalty} + 0.01 \times \text{Quality}^2 + 0.49 \times \text{SM Presence} - 0.09 \times \text{Comp Pricing} - 0.21 \times \text{Order Fulfillment} + 0.34 \times \text{Shipping Cost} + 0.33 \times \text{Recommend}$$

Limitations of the prediction model:

1. **Heteroscedasticity:** Present in Comp Pricing despite transformations, affecting coefficient reliability as shown below:
2. **Unexplained Variance:** Model explains 65% of the variation; remaining variance may stem from unobserved factors like seasonality or promotions.

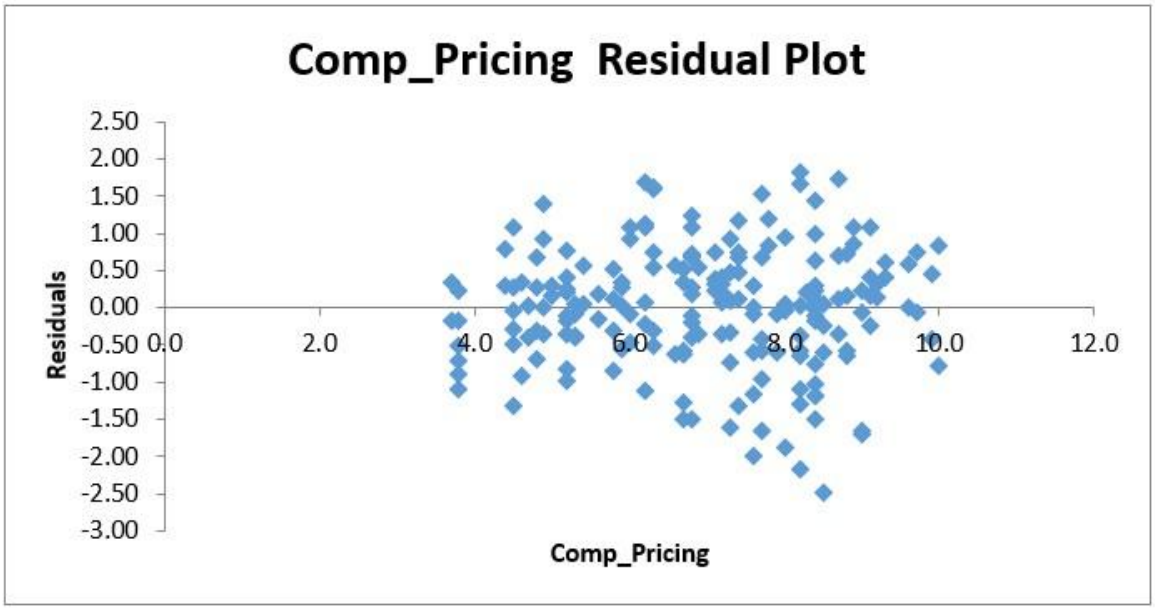


Figure 8: Heteroscedasticity as observed in Comp Pricing

3. **Linearity Assumption:** May not hold consistently across all predictors.
4. **Multicollinearity:** Addressed by removing Shipping Speed, which may have excluded relevant information.
5. **Sample Size:** Larger and more diverse samples could improve accuracy and generalizability.

3 Interaction effect between Quality and Brand Image

To test Cindy's hypothesis—that the effect of quality on order quantity is stronger for customers with favorable brand perceptions—we modeled an interaction between Quality and Brand Image as below:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.60
R Square	0.35
Adjusted R Square	0.35
Standard Error	1.08
Observations	200.00

ANOVA

	df	SS	MS	F	Significance F
Regression	3.00	126.80	42.27	35.95	0.00
Residual	196.00	230.45	1.18		
Total	199.00	357.24			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.75	2.31	0.33	0.74	-3.79	5.30	-3.79	5.30
Quality	1.04	0.28	3.69	0.00	0.48	1.59	0.48	1.59
Brand_Image	1.30	0.40	3.21	0.00	0.50	2.09	0.50	2.09
Quality*Brand_Image	-0.10	0.05	-2.08	0.04	-0.20	-0.01	-0.20	-0.01

Figure 9: Regression output for interaction analysis

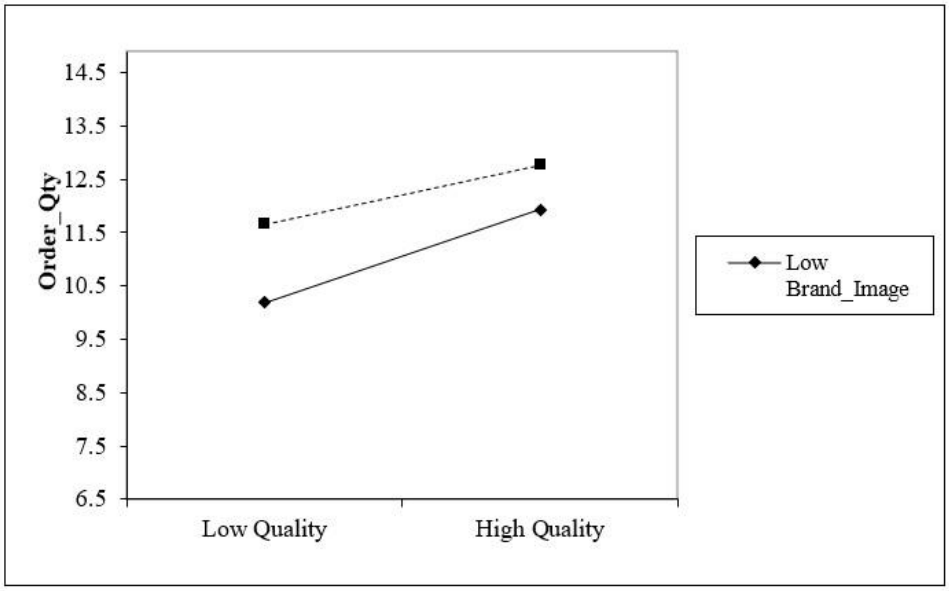


Figure 10: Interaction between Quality and Brand Image

The regression model in figure 9 reveals that Quality and Brand Image both significantly predict Order Quantity (p ¡ 0.05). However, the interaction term (Quality × Brand Image) is also significant but negative (p = 0.04), suggesting that as brand image improves, the marginal effect of quality on quantity ordered slightly decreases. This may reflect a ceiling effect[2], where high brand image already signals high quality, reducing the added influence of perceived quality.

Figure 10 illustrates an interaction between product quality and brand image in influencing order quantity. As product quality improves from low to high, the order quantity increases for both low and high brand image groups. However, the increase is more pronounced when the brand image is low, suggesting that product quality plays a more significant role in driving orders when the brand image is weaker. In contrast, when the brand image is strong, order quantities are relatively high regardless of quality, indicating that a strong brand can partially offset the effects of lower product quality. This interaction shows that the effect of product quality on order quantity depends on the level of brand image.

4 Developing a model to predict the likelihood of recommending K-Olive to others

A logistic regression model was developed (Figure 11 to predict the likelihood of customers recommending K-Olive based on three predictors: Distribution Channel, Quality, and Brand Image. A cutoff score of 50% was taken as the classes are balanced. All variables in the model were found to be statistically significant (p ¡ 0.05), indicating that each contributes meaningfully to the likelihood of recommendation.

The Distribution Channel has a positive and strong effect (B= 1.06) with an odds ratio of 2.88 suggesting that customers reached through more effective channels are nearly three times more likely to recommend the brand.

Quality (B= 0.56) increases the odds of recommendation by 75% underscoring its influence on customer satisfaction.

Brand Image (B= 0.67) also shows a strong positive relationship with an odds ratio of 1.95 meaning customers with a favorable view of the brand are nearly twice as likely to recommend it.

Logistic Regression

		# Iter	20.00		Alpha	0.05	
	coeff	s.e.	Wald	p-value	exp(b)	lower	upper
intercept	-8.34	1.64	25.80	0.00	0.00		
Dist_Channel	1.06	0.36	8.77	0.00	2.88	1.43	5.80
Quality	0.56	0.14	15.75	0.00	1.75	1.33	2.30
Brand_Image	0.67	0.18	14.30	0.00	1.95	1.38	2.75

Figure 11: Logistic Regression Model

8.34 + 1.06*Dist Channel + 0.56*Quality + 0.67*Brand Image

The logistic regression model shows a significant improvement over the baseline model, with a log-likelihood (LL) of -107.20 compared to the baseline model's LL0 of -138.63. The Chi-square statistic of 62.87 (with 3 degrees of freedom) yields a p-value of 0.00, confirming the statistical significance of the model. This indicates that the model with Distribution Channel, Quality, and Brand Image provides a strong fit.

LL	-107.20
LL0	-138.63
Chi-sq	62.87
df	3.00
p-value	0.00
alpha	0.05
sign	Yes

Overall Fit

R-sq (L)	0.23
R-sq (CS)	0.27
R-sq (N)	0.36

R square

	Coeff (B_i)	EXP (B_i)
Intercept	-8.340	0.000
Dist_Char	1.060	2.886
Quality	0.560	1.751
Brand_Im	0.670	1.954

Coefficients

Figure 12: Logistic Regression Model Performance Metrices

The model explains 27%–36% of the variation (Cox & Snell R^2 = 0.27, Nagelkerke R^2 = 0.36), indicating acceptable explanatory power.

Distribution Channel has a significant positive effect on the likelihood of recommending K-Olive. A switch from a direct to a network channel increases the odds of recommendation by 188.6% (EXP(B) = 2.886). Quality also positively impacts the recommendation likelihood, with each 1-point increase in the quality score increasing the odds of recommending by 75.1% (EXP(B) = 1.751). Brand Image has the strongest effect, with each 1-point increase in brand image increasing the odds of recommendation by 95.4% (EXP(B) = 1.954).

	Intercept	Dist_Channel	Quality	Brand_Image
Intercept	2.69	0.14	-0.2	-0.23
Dist_Channel	0.14	0.13	-0.02	-0.01
Quality	-0.2	-0.02	0.02	0.01
Brand_Image	-0.23	-0.01	0.01	0.03

Figure 13: Correlation Matrix

The covariance matrix shows low correlation values between the predictors, suggesting that multicollinearity is not a concern. The predictors appear to be independent, supporting the robustness of the model.

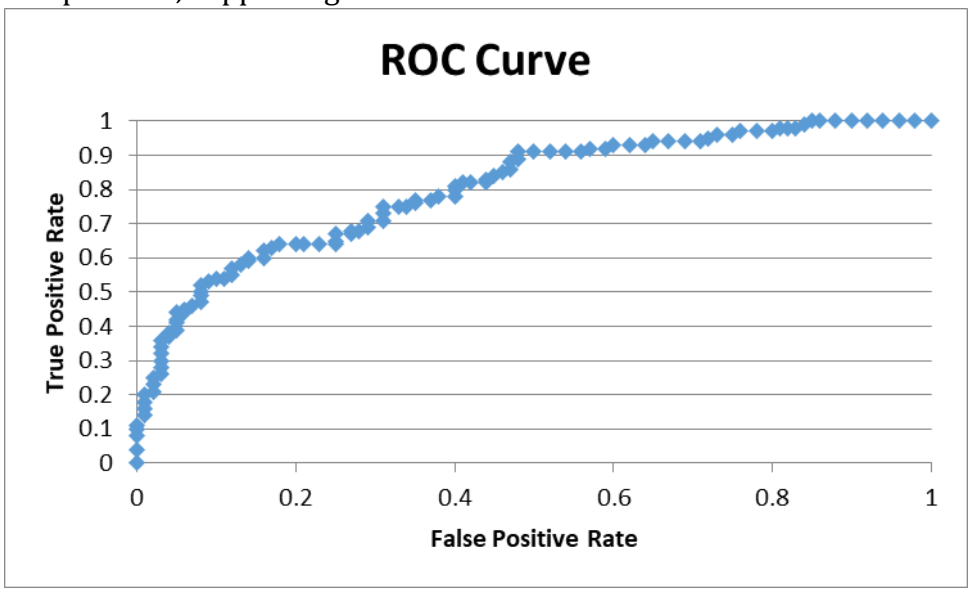


Figure 14: ROC Curve

The AUC is near 1.0, indicating excellent predictive power, with the ROC curve well above the baseline, showing strong classification accuracy.

To assess the practical significance of the logistic regression model, confusion matrix and accuracy rate are suitable [3]. The overall classification accuracy was 71.0%, meaning 71% of product recommendations were accurately classified. While the model performed well, the remaining misclassification rate could be reduced by adding more relevant independent variables.

Out of the 95 recommended products, 68 were correctly classified, resulting in a 71.6% accuracy for recommended items. For the 105 non-recommended products, only 25.7% were correctly classified, highlighting areas for improvement.

CLASSIFICATION TABLE

	Success-Observed	Fail-Observed	Total
Success-Predicted	68	27	95
Fail-Predicted	32	73	105
Total	100	100	200
Accuracy	0.680	0.730	0.710
Cutoff			0.5

Figure 15: Confusion Matrix

Comparing the accuracy rate to the hit ratio shows the model's practical significance. The model's 71.0% accuracy exceeds both the PCC hit ratio (0.5) and the standard hit ratio (0.644), confirming that the logistic regression model is significantly better than random chance at classifying observations.

To understand how three factors—perceived quality (scored 1–10), a strong brand image (score of 10), and distribution channel (direct vs. distributor)—affect the likelihood that customers will recommend K-Olive products, following curve is plotted:

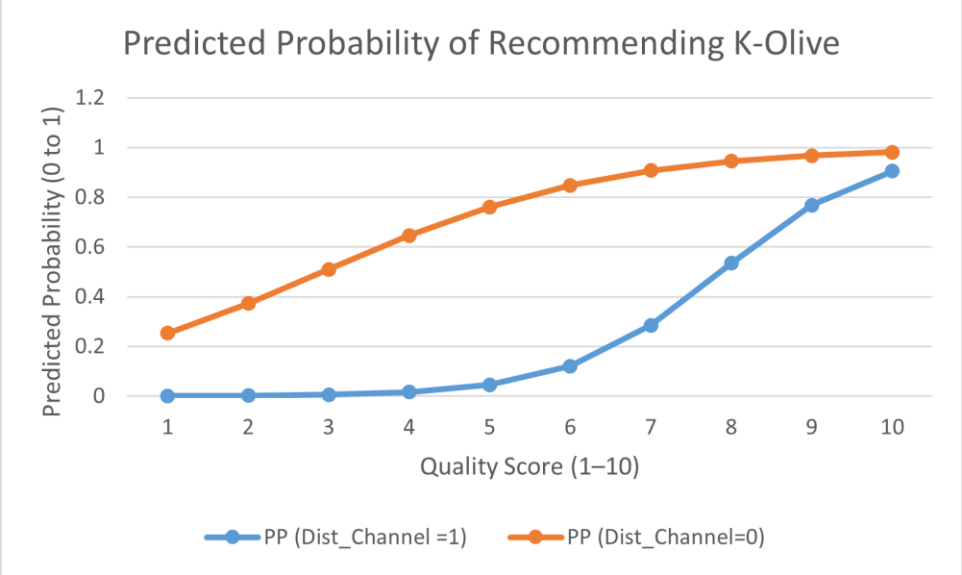


Figure 16: Predicted Probability of Recommending K-Olive

The graph shows that as quality scores increase, the predicted probability of recommending K-Olive also increases for both distribution channels. However, customers who purchase directly (Dist_Channel = 0) consistently show higher recommendation probabilities than those buying through a distributor (Dist_Channel = 1). The gap narrows as quality reaches the highest levels, indicating that excellent quality can overcome distribution-related differences in recommendation likelihood.

5 Demand Forecasting for K-Olive

A forecast has been developed for next four quarters. The time series data shows the quantity sold each quarter across different years. Time-series analysis was deemed appropriate for this task due to its ability to capture the temporal dependencies in the dataset, where past values provide valuable information for predicting future outcomes. This approach allows for the identification of trends, seasonality, and other cyclic behaviors within the data, making it ideal for predicting demand, sales, or order quantities over time. The 4-centered moving average method is useful for smoothing these fluctuations[6] to forecast future sales because it helps eliminate seasonality and short-term variations by averaging over a period. For example, to forecast Q3 2017, the 4-CMA will average the sales of Q2 2017, Q3 2017, and the two adjacent quarters (Q1 and Q4 of 2017), thereby providing a balanced forecast based on nearby data points.

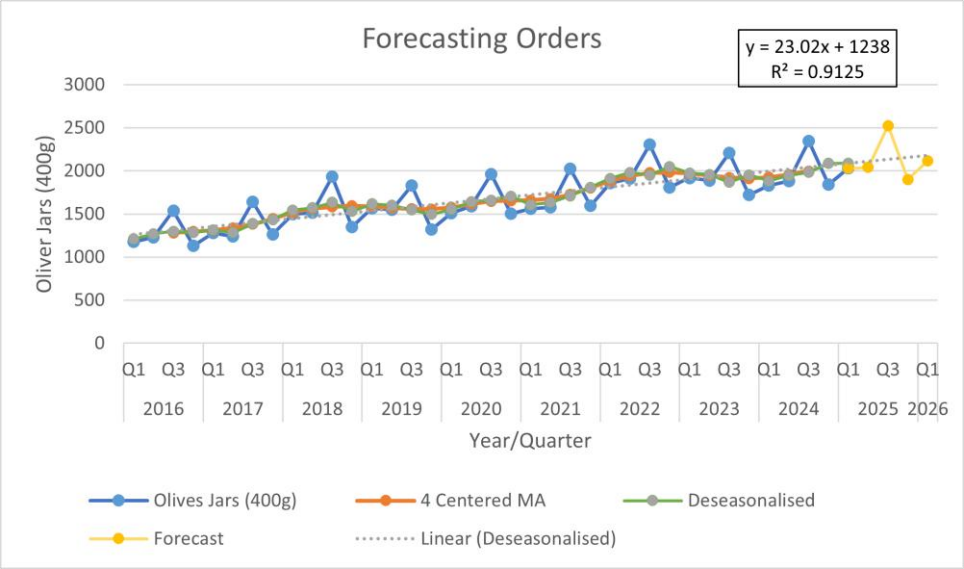


Figure 17: Forecasted value of next four quarters

A MAPE (Mean Absolute Percentage Error) of 3.68% in the forecasting model indicates that, on average, the predictions made by the model deviate from the actual values by approximately 3.68%. This is a relatively low error percentage, suggesting that the model is performing well in predicting the target variable.

6 Conclusion and Recommendation

The analysis identifies Loyalty, Social Media Presence, and Shipping Cost as key business drivers influencing order quantity. These insights suggest that businesses can increase demand by strengthening customer loyalty programs, investing in targeted social media campaigns, and ensuring timely and cost-effective delivery. While the model shows strong predictive accuracy, issues like pricing inconsistencies and unexplained variance indicate opportunities for refinement. Addressing these through pricing strategy adjustments and improved data collection can further enhance forecasting accuracy. Overall, aligning marketing, pricing, and distribution strategies with these insights can drive stronger sales and customer engagement.

References

[1] Model building. *Wiley series in probability and statistics*, pages 23–49, 09 2020.

[2] Peter C Austin and Lawrence J Brunner. Type i error inflation in the presence of a ceiling effect. *The American Statistician*, 57:97–104, 05 2003.

[3] Mahmoud Fahmy Amin. Confusion matrix in binary classification problems: A stepby-step tutorial. *Journal of Engineering Research*, 6:0–0, 12 2022.

[4] Jong Hae Kim. Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*, 72:558–569, 07 2019.

[5] Julia Martin, David Daffos Ruiz de Adana, and Agustin G. Asuero. Fitting models to data: Residual analysis, a primer. *Uncertainty Quantification and Model Calibration*, 07 2017.

[6] Raveendran Vadakkoot, Mitul Devendra Shah, and Suyashi Shrivastava. Enhanced moving average computation. 01 2009.

[7] Shiyu Zhou and Yong Chen. Linear model for numerical and categorical response variables. pages 81–108, 01 2022.