Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

The company manufactures and sells high-end home goods. As part of their sales strategy the company circulated their first print catalog last year. They now have an additional 250 customers on their mailing list. The purpose of this report is to examine the expected profitability of printing and circulating these additional 250 catalogs. The company do not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000. The aim of this report is to determine if the company should send the new customers the catalog. To answer this question, historical sales data and customer characteristics are required to forecast the expected revenue from these additional 250 catalog clients.

Step 2: Analysis, Modeling, and Validation

Data from 2375 customers was utilized in this report to build a predictive model to forecast the expected revenue from the proposed additional 250 catalog clients. All data analysis was conducted in Alteryx Designer 2018 (Boulder, Colorado). The modeling process included three steps.

Step 1 Data exploration and visualization.

From initial data investigations, it was decided to drop unique data identifiers (e.g. ‘Name’, ‘Address’), features with no variation (‘State’) and variables not present in the dataset to be tested (‘Responded_to_Last_Catalog’). In addition, after visualizing the distribution of the ZIP code (fig 1), this variable was transformed to three categories (< 80100, > 80200 and 80200 < 80100 named A, C and B respectively). Checking the assumptions of the linear regression model, the continuous predictor variables are not correlated (fig 2), however examining the

variance inflation factor (fig 3) suggests the presence of multicollinearity for both ‘City’ and ‘ZipGroup’ (Pallant 2010). The continuous variable ‘Average Number of Products Purchased’ has a linear relationship with ‘Average sale Amount’, however years as a customer does not appear to have a linear relationship and should be considered for removal (fig 4). Finally, a base model containing six variables (Customer Segment, City, Store Number, Average Number of Products Purchased, Years as Customer, and Zip Group) was generated with a R^2 of 0.839 and a RMSE of 136.4. Examination of the beta coefficients and p values suggests that store number contributes little to the predictive model and was subsequently removed from further analysis.

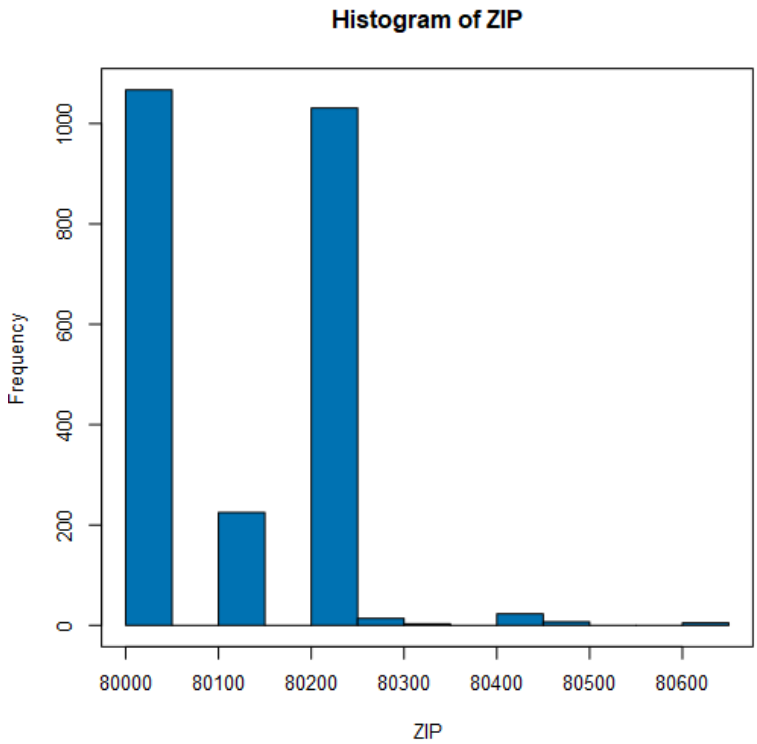


Fig 1. Histogram of ZIP

Table 1. Correlation Matrix of continuous variables.

	Avg_Sale_Amount	Avg_Num_Products_Purchased	Years_as_Customer
Avg_Sale_Amount	1.000000	0.855754	0.029782
Avg_Num_Products_Purchased	0.855754	1.000000	0.043346
Years_as_Customer	0.029782	0.043346	1.000000

Table 2. Variance Inflation Factor table

	(G)VIF	DF	Standardized (G)VIF
Customer_Segment	2.2431	3	1.1441
City	588.6318	26	1.1305
Store_Number	1.3338	1	1.1549
Avg_Num_Products_Purchased	2.1972	1	1.4823
Years_as_Customer	1.0183	1	1.0091
ZipGroup	456.0650	2	4.6212

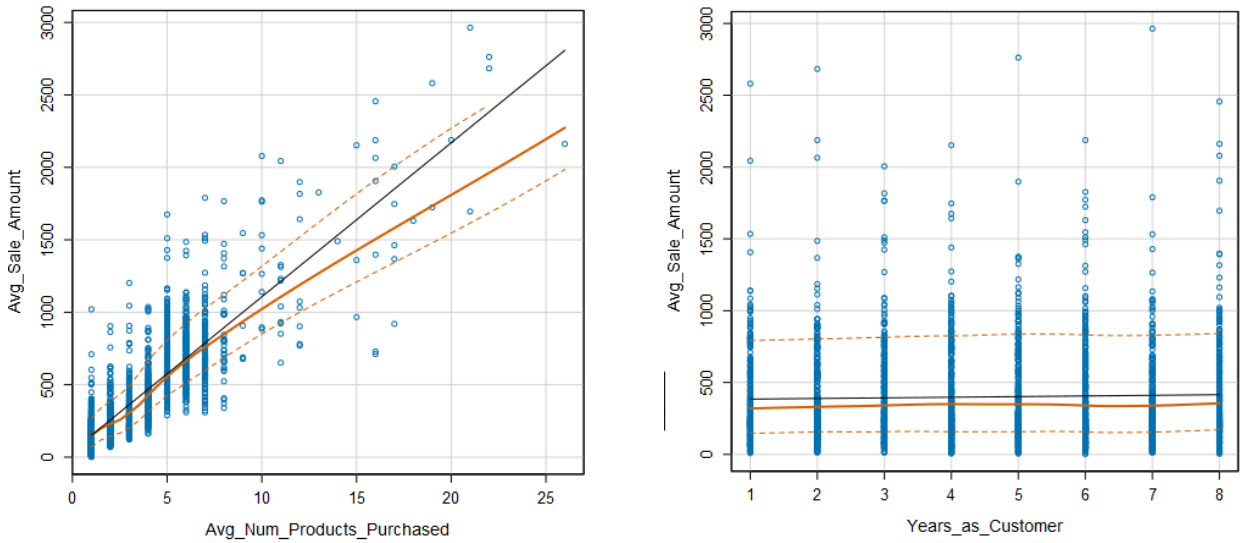


Fig 2. Scatter plots depicting the bivariate relationship between the continuous predictor variables and the outcome measure.

Step 2 Model Testing

A total of six possible models were trained using a random 70% subset of the data from the 2375 customers. To have confidence in the generalizability of the generated models, these were

then tested on remaining 30% of the data. Both 'Average Number of Products Purchased' and 'Customer Segment' were retained in each model tested given their high bivariate correlation and high beta coefficients respectively. In addition to these two features, a test all subsets approach was taken with three further features (Years as Customer, City and Zip Group).

Step 2 Model Selection and application

The results from this analysis demonstrates little difference in the performance of the six models (table 3). In addition, both 'Zip group' and 'Years as customer' provided a non-significant contribution to the generated model once customer segment and average number of products were included.

Table 3. Comparison of the six models tested

Model	Correlation	RMSE	MAE	MPE	MAPE
1. Customer Segment, Average Number of Products Purchased	0.9212	141.473	95.631	-46.021	67.891
2. Customer Segment, Average Number of Products Purchased, City	0.9196	142.838	96.411	-48.292	70.165
3. Customer Segment, Average Number of Products Purchased, Years as Customer	0.9214	141.340	95.524	-45.225	67.095
4. Customer Segment, Average Number of Products Purchased, Zip Group	0.9214	141.361	95.598	-45.873	67.795
5. Customer Segment, Average Number of Products Purchased, Years as Customer, City	0.9197	142.677	96.326	-47.457	69.329
6. Customer Segment, Average Number of Products Purchased, Years as Customer, Zip Group	0.9215	141.231	95.473	-45.052	66.954

Given that a sparse model is generally preferred for prediction (Field 2009), model 1 with two predictor variables (Customer Segment, Average Number of Products Purchased) was chosen

and demonstrated generally good predictive ability (particularly for the lower sale amounts) (fig 3). The following regression equation was generated:

*Y = 303.46 + 66.98 * 'Average Number of Products Purchased' - 149.36 (if type Loyalty Club Only) + 281.84 (if type Loyalty Club and Credit Card) - 245.42 (if type Store Mailing List) + 0 (if type Credit card only)*

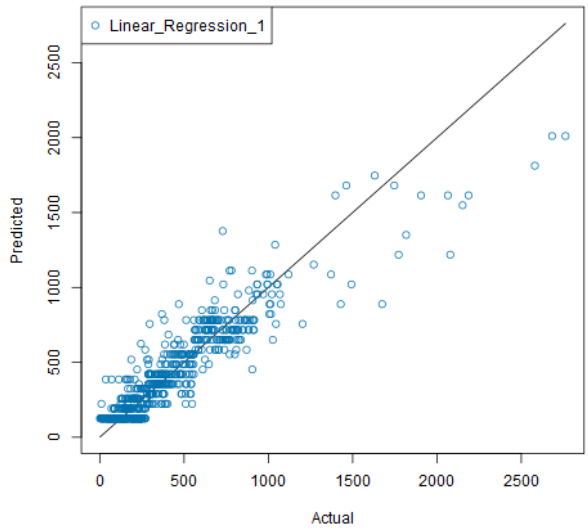


Fig 3. Scatter plot of chosen model's predicted vs actual sale amount

Step 3: Presentation/Visualization

The model generated within this report could explain 84% of the variance in average sale amount and demonstrated good generatability in out of sample testing ($R^{2:}$ 0.85, RMSE 141.5). Applying this mode to the prospective 250 clients estimates a predicted sale amount of $138292.13. The gross margin is 50% on products sold through the catalog and the cost of printing and distributing is $6.50 per catalog. This results in an estimated profit of $67521.06.

To have confidence in our estimate, the estimated profit was multiplied by the probability that a customer will in fact purchase from the company:

Expected profit = [(predicted sale amount * probability of purchase) * gross margin)] - costs

Using this above formula, the calculated expected profit was $21987.44. Some caution is required, since the positively skewed nature of both 'Average Number of Products Purchased' and 'Average Sale Amount', suggests that the model may be inaccurate at large values of these variables (Fig 4). Indeed, this was observed in the out of sample test (fig 3).

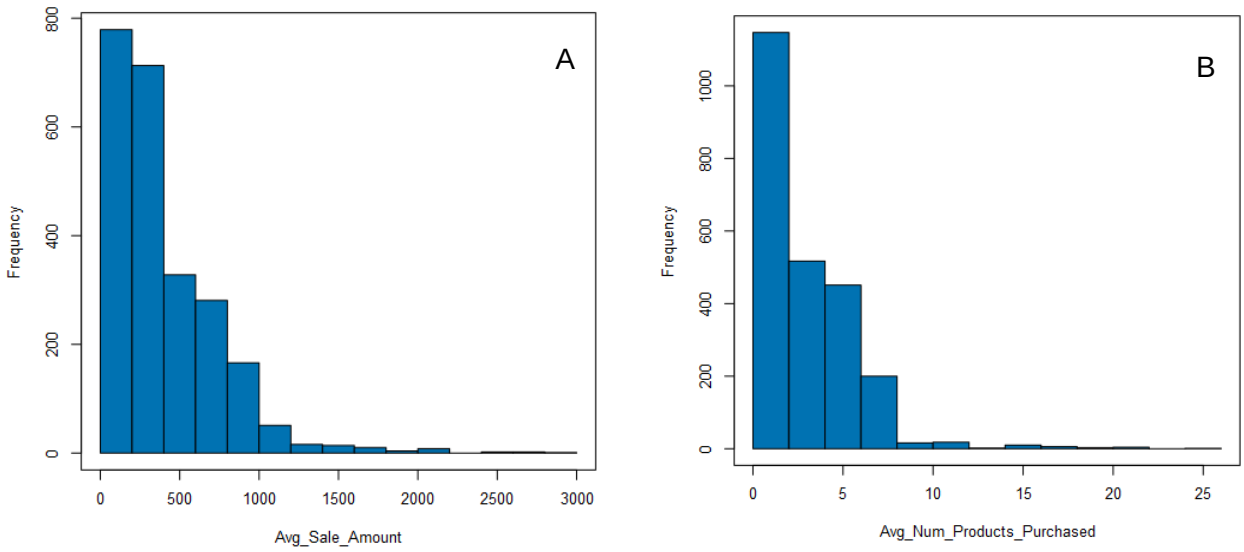
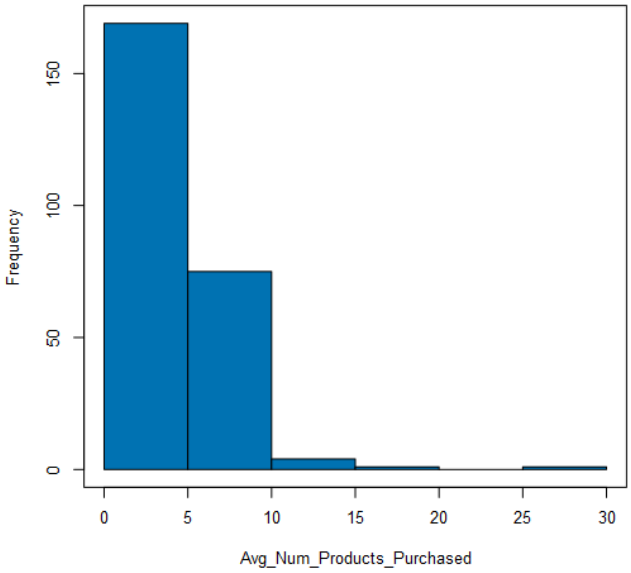


Fig 4. Histograms depicting the distribution of average sale amount (A) and average number of products purchased (B)

However, given the positively skewed nature of the 'Average Number of Products Purchased' in the purchase history of the prospective 250 print catalog clients (fig 5) and the fact that the calculated expected profit was 2.2 times greater than the $10,000 threshold set by the

company, it is recommended that the company circulate their print catalog to the additional 250 clients.



.

Fig 5. Histograms depicting the distribution of average number of products purchased for the 250 clients being assessed.

Future data collection should be conducted on clients with high purchase orders to further enhance the accuracy of future predictive models.

References:

Field, A. (2009). *Discovering statistics using SPSS*. Sage publications

Pallant, J. (2010). *SPSS survival manual: A step by step guide to data analysis using SPSS*. McGraw-Hill International.