Project 1: Predicting Catalog Demand

Jorge Pinzon

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (500 word limit)

Key Decisions:

Answer these questions

1. What decisions needs to be made?

2. What data is needed to inform those decisions?

In this project, the requirement was to provide a predicted value of total profit for a company during a catalog marketing process, based on the results of a previous similar process. A linear regression model was deemed appropriate for this particular situation.

With the model already selected, the attention should be placed on the variables. The first decision is to determine common variables between the training set and the prediction set. Then variables need to be selected for the model (feature selection). This step requires understanding of the type of variable, and the relation between it and the target.

After selecting the variables, it is necessary to determine how the model is going to be run (in my case I used python – multiple options) and analyzed. Finally, a visualization strategy needs to be included incorporating the requirements with the results of the analysis.

For these decisions to be made we need data on previous tests (which we already have) and data on potential targets (new customers).

: Great: ultimately the decision that has to be taken is whether or not to send out the catalogs

: Comment: on top of the two datasets, we are also going to need to know the cost structure of the operation, the gross margin % and the cost per catalog

Step 2: Analysis, Modeling, and Validation

Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

Important: Use the p1-customers.xlsx to train your linear model.

At the minimum, answer these questions:

1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

Since the data was already cleaned, I check for the type of data in the variables, numeric vs categorical. Of the provided variables in the training set, only “Avg_Num_Products_Purchased” is numeric. All other variables are categorical or discreet. This step is important because knowing the type of data dictates the way the variable is analyzed in order to determine if it should be included in the model or not. Numerical contiguous variable can be analyzed against the target in two ways: 1. running a correlation and determine metrics such as correlation coefficient (r^2) and the statistical significance of the relation (p-value) and/or 2. visually assessing the relation on a scattered plot including the abovementioned metrics. For all other variables a correlation model needs to be applied since the visual is less informative.

: Great: that is indeed correct. The scatterplot works well as a tool of exploration for discreet variables, for categorical variables we could plot a box plot to get a first insight into the relation between the variables. However checking the p-value as you did is the right way.

: Great approach to your data here

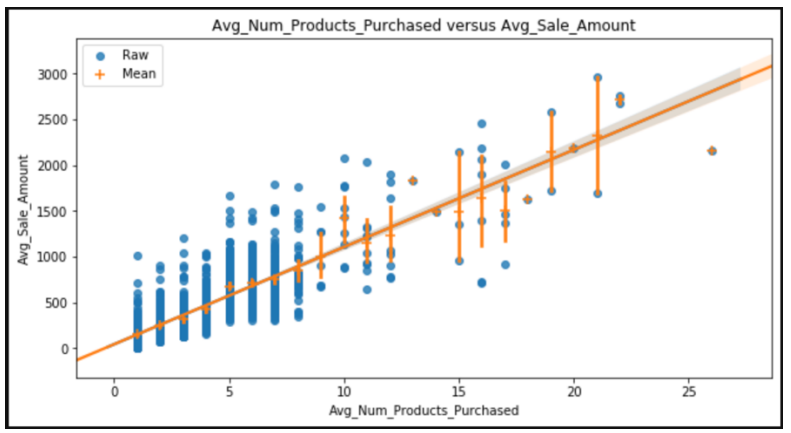
I run the correlation analysis for all variables and this are the results:

	coef	std err	t	P>|t|	[0.025	0.975]
const	107.6301	136.113	0.791	0.429	-159.283	374.543
City	-0.0201	0.565	-0.036	0.972	-1.129	1.089
ZIP	0.0735	0.167	0.441	0.659	-0.253	0.400
Store_Number	0.1117	1.291	0.087	0.931	-2.420	2.643
Avg_Num_Products_Purchased	99.3527	1.582	62.790	0.000	96.250	102.456
Customer_Segment	-27.4969	3.530	-7.788	0.000	-34.420	-20.574
#_Years_as_Customer	-1.0527	1.547	-0.680	0.496	-4.087	1.981

As the table shows, only “Avg_Num_Products_Purchased” and “Customer_Segment” show a significant ($p < 0.05$ ‘P>|t|’) correlation with the target variable (“Avg_Sale_Amount”). All other variable did not show any correlation and I decided not to include them in the analysis.

: Great: the linear relation in the graph is clear and the variables that you picked as predictors are correct

The findings on “Avg_Num_Products_Purchased” can be corroborated with a visual assessment of the relation:



The r^2 value for this relation is 0.73 with a p-value < 0.05 indicating a positive correlation with “Avg_Sale_Amount”.

2. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. For each variable you selected,

please justify how each variable is a good fit for your model by using the p-values and R-squared values that your model produced.

The linear regression model trained with ***p1-customers.xlsx*** data showed significant correlation (p= 0.0000) and a coefficient (r^2 = 0.74). These metrics indicate that the model show a direct relation between the selected feature and the target variable and in most cases the prediction is accurate. However, having an r^2 of 0.74 also means that there is room to improve the model or exploring other models. Develop new features, collect more data, generate dummy features or engineer some of the ones already in the dataset, and/or normalize the data are among the options to improve the model. I would argue that the model as is, is usable and provides the required information.

3. What is the best linear regression equation based on the available data? Each coefficient should have no more than 2 digits after the decimal (ex: 1.28)

Important: The regression equation should be in the form:

*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

For example: Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

Note that we **must** include the 0 coefficient for the type Cash.

Note: For students using software other than Alteryx, if you decide to use Customer Segment as one of your predictor variables, please set the base case to Credit Card Only.

ANSWER

My best linear regression equation is:

Predicted_amount = 117.06 + -27.52 * Customer_Segment + 99.31 * "Avg_Num_Products_Purchased

: Required change: the multiple and adjusted R^2 should be slightly above 0.8. The predictor variables you selected are correct, but I believe you did not convert the categorical variable customer segment into dummies.
Alteryx does that by itself but in python you need to use pandas.get_dummies here the documentation htt

ps://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html Remember also check drop first as true, that is going to be your baseline categorical variable with ß equal zero.

: Comment: When doing multivariate regression (linear regression using more than one predictor variable), we should always use the adjusted R-squared rather than the multiple R-squared. This is because the latter always increase when you add new variables into the model, even if the variables are bad predictors. On the other hand, adjusted R-squared, as by its name, is adjusted to increase only when the new variable actually improves the model.

: Required change: in the regression equation each category should be included with its coefficient. Like in the example above, where a categorical variable has 3 categories, Credit card, Mortgage and Cash. Cash being the baseline.

Step 3: Presentation/Visualization

Use your model results to provide a recommendation. (500 word limit)

At the minimum, answer these questions:

1. What is your recommendation? Should the company send the catalog to these 250 customers?

My recommendation is to proceed and send the catalog to the 250 new customers.

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)

While the model has room for improvement, it is already strong and sound with high correlation metrics (p-value < 0.05 and r2 = 0.74), indicating that decisions made on its results should be accurate. After running the model to predict the Sale Amount for each customer, I calculated the total profit across customers. To get the total profit, I adjusted the predicted value by willingness of the customer to respond to the catalogue ("Score_Yes") and subtracted the cost of the catalogue (US $6.50), for each customer after that, I added all values for all customers. The predicted total profit (US $22,141.93) is 2.2 times the minimum expected amount (US $10,000.00) set by management, therefore the campaign should be successful in bringing new sales to the company.

: Required change: do not forget to mention that you also accounted for the gross margin % by multiplying by 0.5 the expected revenue

3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?

According to the linear regression model the new 250 customers are expected to generate a total net profit of US $22,141.93.

: Required change: the profit is off by a bit, it should be slightly below $22,000. Fix the linear equation and you will have the right outcome

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.