PREDICTING CATALOG DEMAND

OCTOBER 15, 2022

UDACIY

Predicting Nanodegree

TABLE OF CONTENT

ABSTRACT

Objective: The Purpose of this study was to help the Company to predict the profit of selling of products by sending catalogs to new 250 Customers, and recommends whether or not the company should send the Catalogs, assuming the manager agrees to take the approach, as long as profit realized from sale exceeds 10000 amount threshold.

Finding: Revenue generated from the sales from sending catalogs to new customers, based the predicted value of the sale, using the regression model or Continuous model, along with consideration of the probability of NOT – WILLING to buy, and with 50% gross profit, on average, is $**21,987.4, which** exceeds the threshold, set up the managers.

Conclusion: The manager will send the catalogs to new 250 customers.

Step 1: Business and Data Understanding

The company sells high – end Home goods. Last year the company sent its first – printed catalogs, and is preparing to send out this year's catalog in coming month. The company has 250 new customers to send out the catalogs to.

1st Decision: To Send out the catalog, managers want to know the profitability of this approach, or simply want to know How much profit the company would expect from mailing out those catalogs.

2nd Decision: how much would, on average, be gross profit. It is Subjective.
3rd decision. How much of operating Cost would incur for those catalog, this might be predicted too, but it's not done here.

What data need to be collected to inform those decisions?
The past selling amount: Selling amount past revenue realized from sending out this Catalog.
Gross Margin: this could be predicted from historical data.
Customer Segments: Customer variation on behaviors

METHODOLOGY

This Problem – solving framework is used and works its way through methodology map to decide which model is best suit for the particular situation.

as the Analyst wants to predict a variable of interest, then it uses **Predictive MODEL, and** then it wants to find the numerical sales amount from sending those catalogs, therefore, it use the continuous model, but, as data has been already available for analysis or prediction, the data rich. Thus, leading to use the LINEAR REGRESSION MODEL

All Data has already been prepared for the analysis, so, no cleansing, blending, formatting, other crazy stuffs need not to be done.

MODELING

The methodology map used to track which model to use , to arrive the best estimates of the intended outcome. Which are already Discussed in the Methodology section above.

Here given the following data for analysis beforehand,

- Data on previous customers'
- Data on 250 new probable customers, who has less than 1 year relation with the business
- Probability of customer buying the product's or probability of realizing sale also given in data with 250 new customer's data
- Gross margin 50%, already given in the particular situation, by the compnay.
- 6.50 product's distribution cost

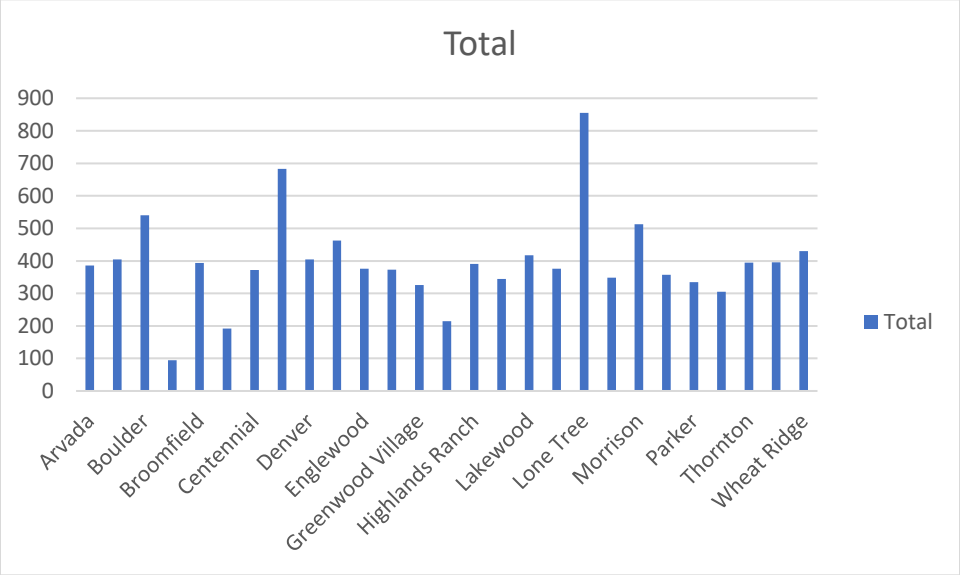
Company wants to know how much profit can be realized from, if it disseminates catalogs to the new 250 customers, in the next month.
So, it's a predictive outcome, it needs predictive model. The outcome is numerical and rich, that's why Regression model is used for this situaiton.

Analysis and Modeling

This model used linear regression model. It has one fundamental assumption, all the predictor variables must have linear – relationship with the target variable.

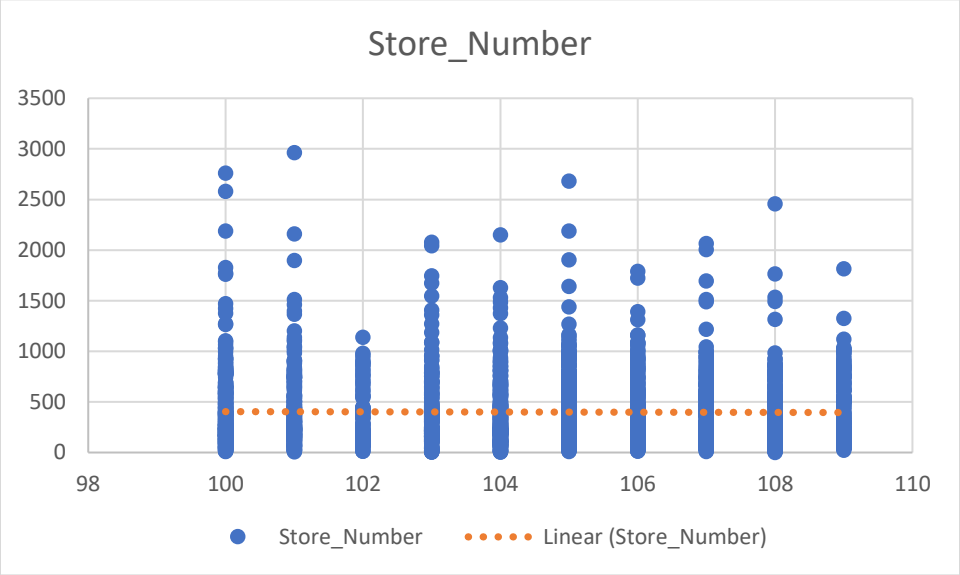
The data, has the following attributes which is not going help the model. The customer name, the CUSTOMER ID, State_Name, City, and years of association of customers.
Why **other variables is not used in the model**, the demonstration is given in the following;

City variable



Where the company earn , on average, its revenue, realized from sending catalog to customer, are in each city of a particular State, is around 350. Only two cities stand out, in short, the city variable would not be much of a help in the regression model.

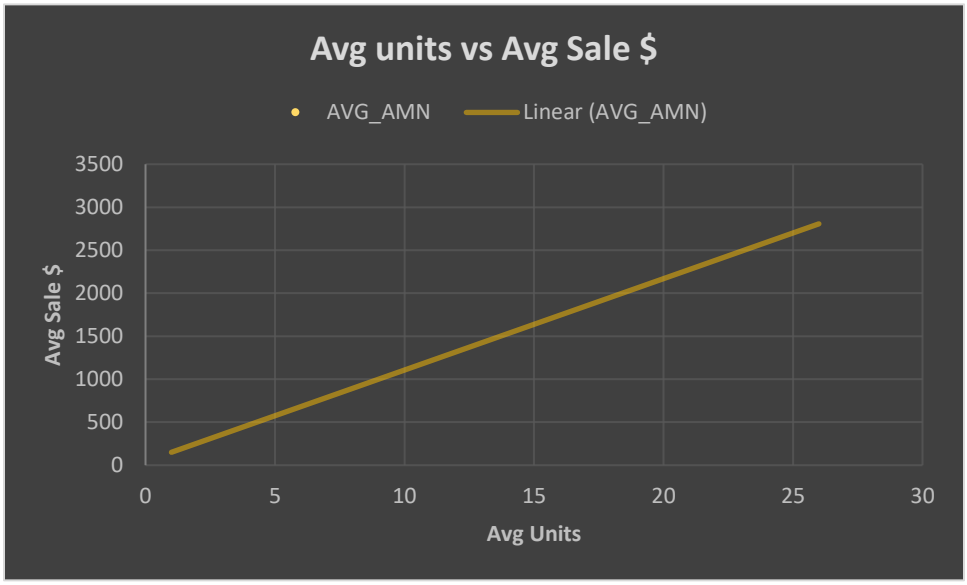
STORE NUMBER or ID

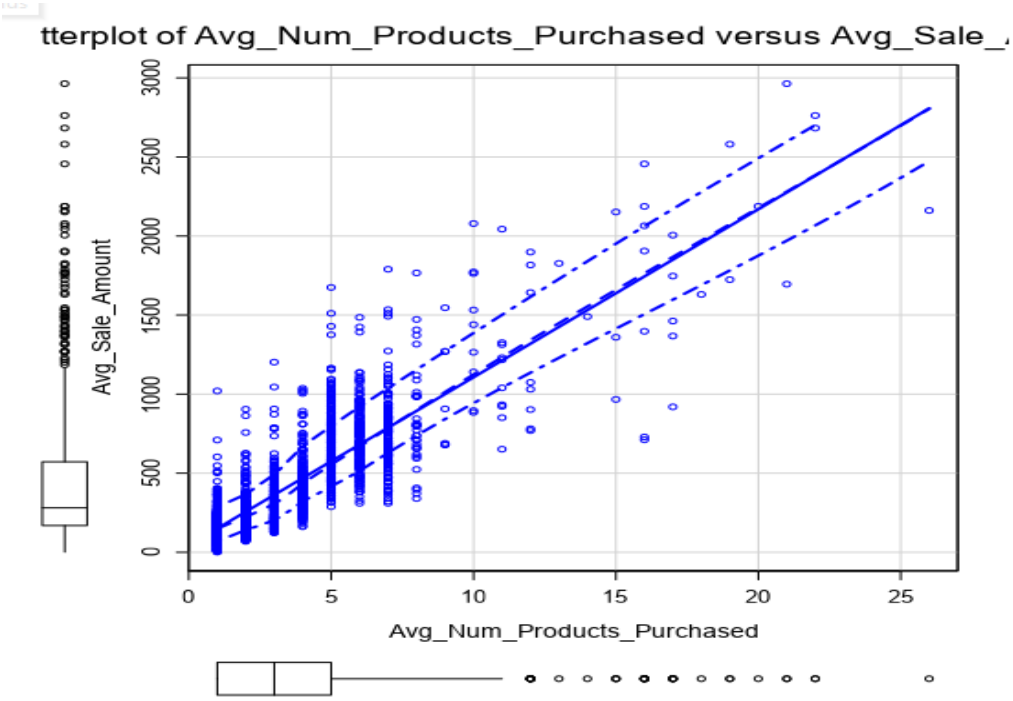


Store ID have no relation with the target variable

Relation Does Exits

The units of sale, realized from sending out catalogs: This is numerical predictor variable, which does have **POSITIVE** r





Avg num of product sale's relation with the Target Variable, the sale amount

Avg sales units has positive strong relation with the avg sale amounts in DOLLER. Therefore. A linear relation exists between the continuous predictor variable and the target variable.

The Categorical Variable

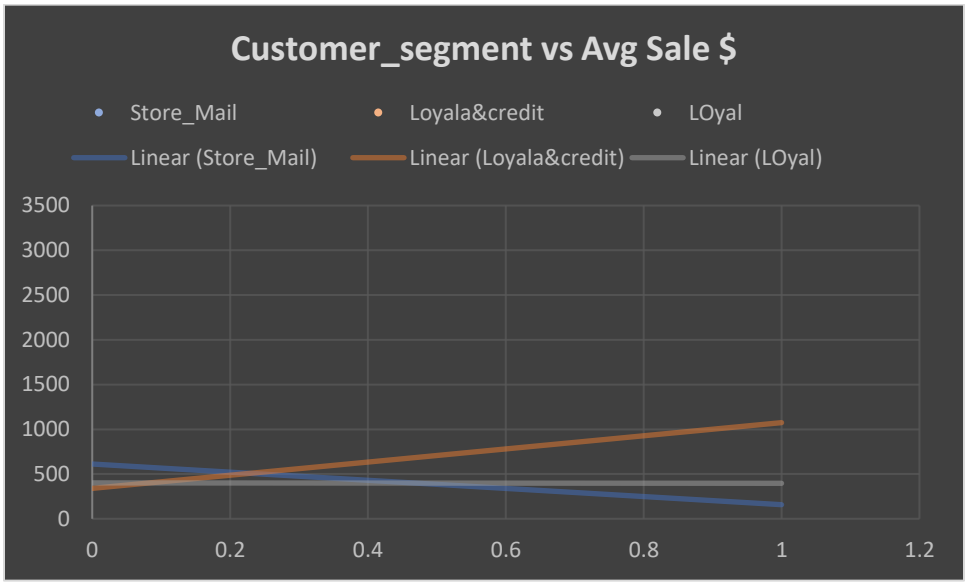
The only Categorical variable which can be used for this model is CUSTOMER SEGMENTS, others are not qualified for being the independent variable for this analysis, for example, the city has 29 distinct cities, each city might have some relation with the target variable but it just doesn't justify the inclusion of 29 variable in the regression model.

Customer segments: Each Customer segments has some different degree of relation with the target variable, which is demonstrated as follows;

There are the following customer segments, given in the data

- Loyalty Club only
- Loyalty and Credit Card
- Credit Card
- Store Mail

Only Credit Card is used as the base for including the categories in the model, therefore, ti does not appear in the model, because all the other categories are compared to the Credit Card's data.



The Charts explains the relations between the variables, such as, the Store mail has some negative relation with the target variable, while the Loyal and Credit Customer has positive relation with the target variable, and the only loyal customer has moderate relation with the target variable. The relationships of categorical variable with the target variable has been explained in the model, with the explanation of Co – efficient, and p- value.

Record Report

1

Report for Linear Model Linear_Regression_6

2 *Basic Summary*

3 Call:

lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = the.data)

4 Residuals:

5

Min	1Q	Median	3Q	Max
-663.8	-67.3	-1.9	70.7	971.7

6 Coefficients:

7

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	303.46	10.576	28.69	< 2.2e-16 ***
Customer_SegmentLoyalty Club Only	-149.36	8.973	-16.65	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	281.84	11.910	23.66	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-245.42	9.768	-25.13	< 2.2e-16 ***
Avg_Num_Products_Purchased	66.98	1.515	44.21	< 2.2e-16 ***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

8 Residual standard error: 137.48 on 2370 degrees of freedom
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366
F-statistic: 3040 on 4 and 2370 degrees of freedom (DF), p-value < 2.2e-16

9 *Type II ANOVA Analysis*

10 Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	28715078.96	3	506.4	< 2.2e-16 ***
Avg_Num_Products_Purchased	36939582.5	1	1954.31	< 2.2e-16 ***
Residuals	44796869.07	2370		

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The Model Validity:

The Following is the Regression model, resulting from Altryex tool and Excel , two predictive variable were considered to anticipate the target variable, among the predictive variable, one is categorical type which require dummy variable for K – 1 categories, but in Altryex tool that won't be needed, it automatically do that for us. Three statistics of this regression model is necessary to validate the model;

R – Squared: how many target variable is explained by the predictive variable, the portion is **.** 0.8369, which indicate a robust portion of target variable is explained by predictive variable.

Adj – R Squared: the efficiency of inserting any variable in the model is justified by the adj – r squared, as the portion is .8366, this indicate that taking up categorical variable does not cause deficiency in the model, rather it seems robust, as it is more than 70%.

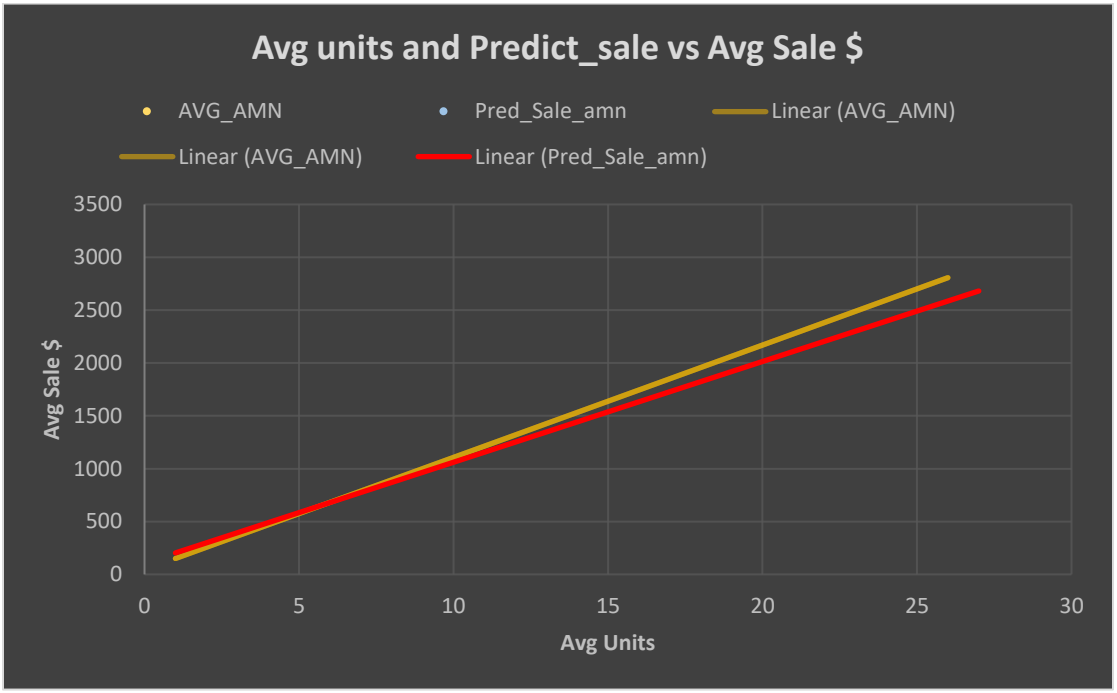
P – Value: P – Value is the measurement of **likelihood of having relation with the target variable by fluke.** The Normal threshold is 0.05, it gets lower than .05, then the probability is not significant. Or simply the probability of having relation of each predictive variable by fluke with the target variable is near to zero, therefore, the co – efficient is justified by the P – value, Each independent variable having p – value is < 2.2e – 16, which is near .0000000000000022, which is really tiny. Therefore, the predictive variable, does indeed have a significant relation with the target variable.

The Equation:

The linear equaiton can be made from this regression model is as follows;

Avg amn of sale = ***303.46 + 66.98 ****** *Avg units of products purchased* ***+ 281.84 ****** *Customer segment Loyalty and credit card customer segment* ***- 149.36 ****** *Loyalty Club only* ***- 245.42 ****** *Mailing list Customer segment*

Presentation/Visualization



As we can see, the model trendline of predicted sales is pretty similar to the original past trendline of sale.

Recommendation: The company should take the initiative to send out the catalogs to new Customers

The Expected Profit, realized from sending out the catalog would be **21987.44**.
The profit was calculated, by multiplying the probability of sale to a particular customer, given in the data, multiplying the gross margin 50%, with the predicted sale amount, and deducting the cost of distribution of printing or other operating cost related to the catalogs. Then sums predicted profit, realized from up all the 250 new customers.