Ashley Varma
Final Project Documentation
IMT 572 B
December 14th, 2021

Business Intelligence MLR Analysis

Customer Personality Analysis Dataset

For this project, I decided to use a widely available and popular dataset from Kaggle that records the demographics and spending habits of a store's customers. It tracks spending habits over a 2 year timeline in designated categories of online, catalog, and brick & mortar storefront sales, as well as key demographic information such as age, income, marital status, education, etc. This dataset was created with the intent of tracking which customers were more likely to respond to the business' multiple marketing campaigns. To see which campaigns were most effective in certain demographics, the data demarcates which campaigns were used to make eligible discounted purchases and by which customers. Although the business from which this data is gathered is not explicitly named, the dataset presents a great feat in customer personality analysis for the purposes of business intelligence.

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer
0	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012
1	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014
2	4141	1965	Graduation	Together	71613.0	0	0	21-08-2013
3	6182	1984	Graduation	Together	26646.0	1	0	10-02-2014
4	5324	1981	PhD	Married	58293.0	1	0	19-01-2014

Problem Statement

Utilizing data science while examining customer profiles allows for these wide ranging personalities to be assessed for providing the best experience to customers while bringing more profits for a business based on analytics driven decision processes. The business can now spend marketing funds in a strategic manner towards audiences that will be much more likely to interact with their campaigns in the future. Therefore, the problem statement I will choose to focus on deals with the relationship between spending habits and key demographic information. Specifically, could there be a correlation between spending habits of customers of this business and certain demographic factors (i.e. age, income, categorical item purchases) and how it influences each other.

Hypothesis

'Income', 'Wine Sales', and 'Meat Sales' all positively increase and correlate to a customer's overall 'Spending' habits. E.g. The more a customer spends, the higher the values will be in these 3 categories.

Data Exploration

To begin, I initially began my exploration by familiarizing myself with the columns provided in the dataset. There were 29 categorical and discrete variables (see Figure 1.1) used in the customer personality data, and had approximately 2300 entries and the latest entry date is from 2014. WIth the end goal of being able to perform Multilinear Regression (MLR) and Principal Component analysis (PCA), I used a heatmap visualization to identify the columns with the most correlation to one specific dependent variable (further explained in the data analysis section). Overall, this was a very rich dataset that had many elements that could be further examined for potential multicollinearity with the other independent variables chosen.

Category	Column Name	Value Description
People	ID	Customer's unique identifier
	Year_Birth	Customer's birth year
	Education	Customer's education level
	Marital_Status	Customer's marital status
	Income	Customer's yearly household income
	Kidhome	Number of children in customer's household
	Dt_Customer	Date of customer's enrollment with the company
	Recency	Number of days since customer's last purchase
	Complain	1 if the customer complained in the last 2 years, 0 otherwise

(Figure 1.1: CPA Column Examples)

	ID	Year_Birth	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	...
count	2240.000000	2240.000000	2216.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	...
mean	5592.159821	1968.805804	52247.251354	0.444196	0.506250	49.109375	303.935714	26.302232	166.950000	37.525446	...
std	3246.662198	11.984069	25173.076661	0.538398	0.544538	28.962453	336.597393	39.773434	225.715373	54.628979	...
min	0.000000	1893.000000	1730.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
25%	2828.250000	1959.000000	35303.000000	0.000000	0.000000	24.000000	23.750000	1.000000	16.000000	3.000000	...
50%	5458.500000	1970.000000	51381.500000	0.000000	0.000000	49.000000	173.500000	8.000000	67.000000	12.000000	...
75%	8427.750000	1977.000000	68522.000000	1.000000	1.000000	74.000000	504.250000	33.000000	232.000000	50.000000	...
max	11191.000000	1996.000000	666666.000000	2.000000	2.000000	99.000000	1493.000000	199.000000	1725.000000	259.000000	...

(Figure 1.2: Statistical summary of data values)

Data Preparation

After viewing the data - including later during the analysis stage - it was clear that I could perform some manual data compounding in order to group similar categorical variables and to reduce dimensionality. These variables, when looking from the lens of PCA and MLR data analysis, did not need as much granularity (e.g. 'Kidhome' vs. 'Teenhome'), and oftentimes overlapped in similar domains. Other than these adjustments that came later in the data analysis process, there was no need to restructure or filter the data initially prior to analysis.

Data Analysis (Application of MLR & PCA)

The methods for analysis used were MLR and PCA. For MLR, I first began by identifying my single, continuous dependent variable of interest: 'Spending'. Spending is defined by the total amount of money spent by a customer through all channels of the business (brick & mortar, catalog, and/or online sales). In addition, the three independent variables chosen were 'Wine Sales' (the amount of money spent on wine by a single customer), 'Meat Sales' (the amount of money spent on meat by a single customer), and 'Income' (a customer's yearly household income amount). These feature values were chosen based on the heatmap visualizations below (see Figure 2.1 and 2.2). Figure 2.1 was the first attempt at creating a heatmap, and from the main Jupyter Notebook file, there were a lot of variables to keep track of and it doesn't initially show too many well correlated values (>0.7). This is another reason that the compounding of variables was particularly helpful, when looking at Figure 2.2 you can see much higher correlatory values in the 'Spending' category.

Now that there are 3 independent variables, I can use them to predict a dependent variable by fitting a best linear relationship. With these variables, using MLR I computed the intercept, and multiple slope values (or coefficients) concerning the 'Spending' feature. The results of this show that the 'Wine Sales' and 'Meat Sales' are closely linear to the dependent variable, while 'Income' tends to have a more logarithmic relationship. This can also visually be seen through Figure 2.3, which is a scatterplot of the independent variables and the dependent variable. For the purposes of splitting the data, 80% of the data was used for training while the other 20% was used for testing and then implemented the linear model.

In addition, PCA was also used to see which of the chosen variables were potentially more closely related to the dependent variable. PCA was indirectly used to reduce dimensionality after the first reduction of variables, and in the preliminary stages of data exploration and analysis I did not find there to be a significant correlation between eigenvalues shown and the variance between principal components in an elbow plot beyond the three chosen independent variables.

Heatmap Visualizations

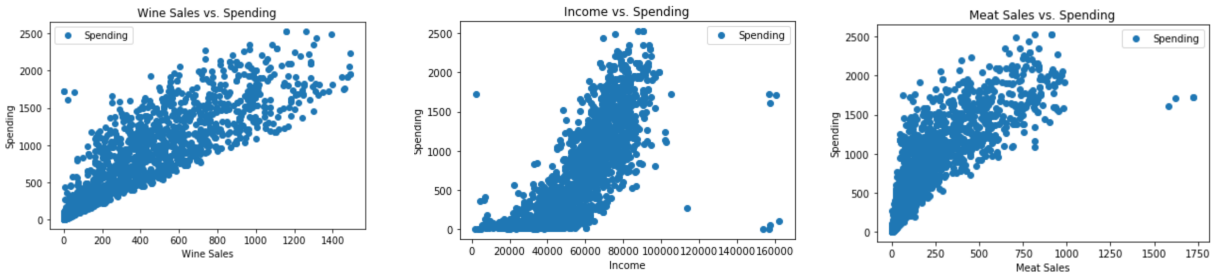
	ID	Year_Birth	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts
ID	1	2.76245e-05	0.0130955	0.00240561	-0.00257989	-0.0465241	-0.0228783	0.00460006	-0.00443722
Year_Birth	2.76245e-05	1	-0.161791	0.230176	-0.352111	-0.0198711	-0.157773	-0.0179172	-0.0308724
Income	0.0130955	-0.161791	1	-0.428669	0.0191334	-0.00396976	0.57865	0.430842	0.584633
Kidhome	0.00240561	0.230176	-0.428669	1	-0.0361331	0.00882673	-0.496297	-0.372581	-0.437129
Teenhome	-0.00257989	-0.352111	0.0191334	-0.0361331	1	0.0161977	0.00484641	-0.176764	-0.26116
Recency	-0.0465241	-0.0198711	-0.00396976	0.00882673	0.0161977	1	0.0160639	-0.00430564	0.0230561
MntWines	-0.0228783	-0.157773	0.57865	-0.496297	0.00484641	0.0160639	1	0.389637	0.562667
MntFruits	0.00460006	-0.0179172	0.430842	-0.372581	-0.176764	-0.00430564	0.389637	1	0.543105
MntMeatProducts	-0.00443722	-0.0308724	0.584633	-0.437129	-0.26116	0.0230561	0.562667	0.543105	1
MntFishProducts	-0.0244749	-0.0416254	0.438871	-0.387644	-0.204187	0.00107897	0.399753	0.594804	0.568402
MntSweetProducts	-0.00764156	-0.0181326	0.440744	-0.370673	-0.162475	0.0226696	0.386581	0.567164	0.523846
MntGoldProds	-0.0134378	-0.0618182	0.325916	-0.349595	-0.0217253	0.0166933	0.387516	0.392995	0.350609
NumDealsPurchases	-0.0371658	-0.0608456	-0.0831009	0.221798	0.387741	-0.00109837	0.0109399	-0.132114	-0.122415
NumWebPurchases	-0.0189239	-0.14504	0.387878	-0.361647	0.1555	-0.0107263	0.542265	0.296735	0.293761
NumCatalogPurchases	-0.00344014	-0.121275	0.589162	-0.502237	-0.110769	0.0251105	0.635226	0.487917	0.723827
NumStorePurchases	-0.0149269	-0.128272	0.529362	-0.499683	0.0506952	0.000798836	0.6421	0.461758	0.479659
NumWebVisitsMonth	-0.00744618	0.121139	-0.553088	0.447846	0.134884	-0.0214447	-0.320653	-0.418383	-0.53947
AcceptedCmp3	-0.03604	0.0617745	-0.0161744	0.0146744	-0.0426769	-0.0329906	0.0622018	0.0147269	0.0182718
AcceptedCmp4	-0.0253867	-0.0605096	0.1844	-0.1616	0.0388864	0.0188256	0.373286	0.0101522	0.102912
AcceptedCmp5	-0.00751702	0.00712254	0.335943	-0.205634	-0.19105	0.000128909	0.472613	0.215833	0.373769
AcceptedCmp1	-0.0216142	-0.00593032	0.27682	-0.172339	-0.14009	-0.0192828	0.354133	0.194748	0.309761

(Figure 2.1: Heatmap of some values from original dataset)

	Age	Income	Spending	Wines	Fruits	Meat	Fish	Sweets	Gold
Age	1	0.198064	0.113241	0.159225	0.0176531	0.0334988	0.0402651	0.020013	0.0640298
Income	0.198064	1	0.79265	0.688269	0.507959	0.692465	0.520351	0.523746	0.389204
Spending	0.113241	0.79265	1	0.893098	0.613249	0.845853	0.642311	0.60697	0.528599
Wines	0.159225	0.688269	0.893098	1	0.386977	0.568752	0.397602	0.390178	0.392588
Fruits	0.0176531	0.507959	0.613249	0.386977	1	0.547796	0.593407	0.571581	0.396443
Meat	0.0334988	0.692465	0.845853	0.568752	0.547796	1	0.573507	0.535048	0.359328
Fish	0.0402651	0.520351	0.642311	0.397602	0.593407	0.573507	1	0.583804	0.427056
Sweets	0.020013	0.523746	0.60697	0.390178	0.571581	0.535048	0.583804	1	0.357336
Gold	0.0640298	0.389204	0.528599	0.392588	0.396443	0.359328	0.427056	0.357336	1

(Figure 2.2: New correlation heatmap of compounded variables)

Correlation Analysis



(Figure 2.3: Strongest correlatory feature values scatterplots)

Inference + Description of Results

P-values Observed

To clarify, a p-value simply quantifies the strength of evidence in support of a null hypothesis. Thus, if the p-value is less than the significance level, I can reject the null hypothesis. Because the p-values calculated for each of the independent variables in this analysis were approximately 0, I can reject the null hypothesis and have more confidence in the plausibility of the initial problem statement and hypothesis presented. Therefore, these values are also statistically significant, as seen in Figure 3.1.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-30.4596	7.784	-3.913	0.000	-45.727	-15.192
Wines	1.0219	0.011	95.702	0.000	1.001	1.043
Income	0.0023	0.000	11.798	0.000	0.002	0.003
Meat	1.2458	0.016	79.227	0.000	1.215	1.277

(Figure 3.1: Ordinary Least Squares [OLS] regression results w/ p-values shown)

Accuracy Interpretation

Expanding on the scatterplots shown in the Correlation Analysis section, the results of the MLR model created were tested for its accuracy in predicting the positive correlation to ‘Spending’. From these results, I’ve found that the R Squared value, the measure of good fit which indicates the variance explained by predictor features, of 96.83% means that the model explains

almost all of the variation in the dependent variable around its mean. This is because higher R-squared values represent smaller differences between the observed data and the fitted values. In contrast, the adjusted R Squared value, which considers the number of variables and only increases when the variable improves fit more than chance alone, was much lower than the originally reported R Squared. This could mean that during the PCA phase, the variables chosen were not improving the fit of the model.

Additionally, the Mean Absolute Error represents the average of the absolute difference between the actual and predicted values in the dataset. In general, the smaller the MAE is, the better. The mean values in my dataset for 'Spending' were very widely ranging from the low 10s to the 100s and 1000s, so this value seems reasonable. Moreover, the standard deviation of residuals reported (also known as the Root Mean Square Error) was 110.6, which is a large value due to - again - the wide range of values in the 'Spending' category. In contrast, the same does not apply to the Mean Square Error value. This value was extremely high, which means that the amount of error is also large as well. The large prediction error may mean that the data was overfit, which explains the high R Squared value as well.

Conclusion

All in all, although the beginning of the MLR and PCA phases seemed promising with the prediction outputs, ultimately the statistical accuracy checks have shown that the model produced may not be the best for the multicollinearity based prediction of the spending habits of this business' customers. Although the individual predictor variables themselves have strong correlations with the response variable, I would not recommend moving forward with using this model to help with the prediction and interactions with future marketing campaigns for this business.

It was worth noting that, from a business intelligence perspective, this type of investigation is very much worth the time and resources to investigate underlying causes behind the successes of such campaigns. Those individual variables with large correlations are still worth looking into as well from this deep dive. For all of today's modern businesses, a solid grounding in the fundamentals of data science has much more far-reaching strategic implications.