Principal Component analysis Applied to Market Change Analysis
Belami E-Commerce VAM Department 2018
By: Carlos Monsivais
August 31, 2018

Table of Contents

Case Study Summary

This is a proof of concept showing exactly how Machine learning Techniques, specifically the technique called Principal Component Analysis has the capability of looking at data from a completely different perspective figuring out patterns or anomalies that would be practically impossible for a human to look at. For example, the process and output below given by the code is something that could be used given more data to look at differences in the market for a very specific manufacturer. This is a skeleton and a process by which we can get a more in-depth analytical point of view regarding how markets shift with an in-depth perception to even the smallest changes.

In 2017 Belami E-Commerce made $624,340.5 worth of sales for Galtech between the months of April to July however in 2018 we only had $437,392.8 worth of sales. We are looking at the months of April to July because those are the seasonal months when patio umbrellas which is a bulk of Gaeltacht's business have the highest sales. Because of the plummet in sales I was tasked to take a quantitative approach as to why this sudden shift occurred. I began by using correlation values to see if there were any massive explanatory changes regarding the correlation values of how the variables interacted with each other however there was no significant outcomes. As a result, I used Principal Component Analysis to get a more in depth look at what exactly happened between the variables and what may have shifted from year to year.

In Conclusion, I used Principal Component Analysis because by using a simpler method of looking at the interaction between the variables by looking at the Pearson correlation I couldn't see any significant changes between the variables. As a result, I used Principal Component Analysis to get a more in depth look at these values and analyze what about these variables changed causing a probable shift in the market or caused sales reduction by so much from year to year.

Limitations

The limitations of this analysis and model is that we are assuming the only variables that make up the whole market are Sum of Sales, Impressions, Cost and Conversions. This is way too simplistic to describe the Galtech Market as a whole. Therefore, we are only able to describe the Belami E-Commerce Galtech Markets for 2017 and 2018. Therefore, there are a lot of limitations, for a more in depth analysis a lot more data would have to be inputted in the model especially regarding variables that are representative of the market because at the end of the day, what you put into machine learning regarding the quality of data is what you should expect to get out in regards to the quality of the data.

With the given limitations we are assuming that the data is representative of the Belami Galtech data for the year for simplicity. However, remember the disclaimer *We are assuming the whole Belami E-Commerce Galtech Market is made up of the 4 variables including Sum of Sales, Impressions, Cost and Conversions however this is too simplistic of a view to analyze the market changes therefore this is a proof of concept as to how this could be done. An overview of what the markets were like, in 2017 the Belami E-Commerce Galtech market was considered successful in how there was high growth and high sales from April to July but in 2018 it was considered a failure with negative growth and low sales from April to July 2018.

Conclusively, this model is very limited as it was built to show a proof of concept of the types of results that can be obtained by using this machine learning technique and what can be obtained by the use of this code because we would just have to add more variables to the R code and then the interpretation of the output is the same however regarding the changes the new variables had on the analysis.

Variables Used

Below is a list of the variables used for this analysis and a description of what they represent in the data:

- Sum of Sales: The summation of sales for that given day.
- Clicks: The summation of the number of Google Clicks received on an advertisement.
- Cost: The summation of marketing costs for Google on any given day.
- Conversions: The number of times a customer purchases from the Site ID 1 website through Google.

These variables were used however the thing to keep in mind is that I started out with 12 variables however due to the variables being extremely highly correlated with Pearson Correlation values of 0.99 I had to remove them and ended up with 4 variables.

Even though Principal Component Analysis assigns numerical values to the data such that the Pearson Correlation value is 0 between all the variables I still saw that by keeping the 12 variables it affected the data therefore I played it safe and removed the highly correlated variables form the beginning to not have results that are affected by multicollinearity.

Pearson Correlation Coefficients

Pearson Correlation Coefficients							
Variables	**Sum of Sales**		**Impressions**		**Cost**		**Conversions**
Sum of Sales	1						
Impressions	**2017**	**2018**	1				
	-0.05	0.18					
Cost	**2017**	**2018**	**2017**	**2018**	1		
	0.12	0.32	0.52	0.58			
Conversions	**2017**	**2018**	**2017**	**2018**	**2017**	**2018**	1
	0.24	0.48	0.28	0.3	0.44	0.5	

Pearson Correlation Coefficient Formulas

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

The standard formula used to calculate the correlation values in the table.

From the graph we can see that there were either non-significant correlations such as Conversions to Impressions or that the correlation values barely changed form 2017to 2018 such as Cost and Conversions. As a result, by looking at this data from a very simplistic point of view there was essentially no changes from 2017 to 2018 and therefore I needed to use a tool to give me a more in depth look at what was going in within the data. Therefore, simple statistical analysis would not suffice in terms of explaining the behavior of the market.

How Principal Component Analysis Works

Principal Component Analysis (also known as PCA) essentially reduces the dimensionality of your data set. For example, if you have an Excel file each column is one dimension of data. If we have a data set with 20 Excel columns, we would in theory have a 20th dimensional data set. As a result, anything that is over 3 dimensions is very difficult to imagine since as human beings we live in a 3-dimensional world. We need some method that will reduce the dimensionality of this data set into something that we can comprehend and visualize to perform statistical procedures on it. Before the analysis of PCA, the data must be scaled so that it can create an orthogonal set of numerical values, they must all have the same mean and be scaled within a certain distance.

As a result, PCA is a statistical procedure that converts a set of variables that are possibly correlated and converts them into a set of variables that are linearly uncorrelated. After the conversion into linearly uncorrelated variables, these new values are called principal components. The principal components are converted so that the first principal component known as PC1 has the largest variance which means that it accounts for as much of the variability in the data as possible. Afterwards, PC2 has the second highest variance which is interpreted as the second highest principal component that accounts for the second highest variability in the data.

To perform PCA, we do the following:

1. Gather the n samples of m dimensional data $\overrightarrow{x_1}, \dots \dots \overrightarrow{x_n}$ in your data set $|R^m$ where we compute the following:
   - Sample Average: $\vec{u} = \frac{1}{n}(\overrightarrow{x_1} + \cdots \dots + \overrightarrow{x_n})$
   - Build the Matrix B: $B = [|\overrightarrow{x_1} - \vec{u}|, \dots \dots \dots, |\overrightarrow{x_n} - \vec{u}|]$
   - Compute matrix S (Covariance Matrix): $S = \frac{1}{n-1} \times B \times B^T$

2. Find the eigenvalues $\lambda_1, \dots \dots \dots, \lambda_m$ of S arranged in decreasing order as well as an orthogonal set of eigenvectors $\vec{u}_1, \dots \dots \dots, \vec{u}_m$

3. Interpret the results as following:
   - Are a small number of the λ_i much bigger than the others?
   - If so, this indicates a dimension reduction is possible.
   - Which of then variables are most important in the first, second, third, etc... principal components?
   - Which factors appear with the same or opposite signs as the others?

Advantages of PCA	Disadvantages of PCA
Useful for dimension reduction for high dimensional data analysis	Only numerical values can be read in
Helps reduce the number of predictor items using principal components	Prediction models are usually less interpretable
Helps to make predictor items independent to avoid multicollinearity problems	
Allows you to interpret many variables in a 2- dimensional plot	
Can be used to develop predictive models	

Principal Component Analysis: Step 1 Eigenvalues and Variation

After running the Principal Component Analysis function in R Studio, below are the Eigenvalues I got for the years 2017 and 2018:

2017 Eigenvalues	
PC1	1.8674404
PC2	1.0785001
PC3	0.6176599
PC4	0.4363996

2018 Eigenvalues	
PC1	2.1884039
PC2	0.9320306
PC3	0.5112613
PC4	0.3683042

*The rule of thumb according to the Kaiser Rule of Thumb we want to only look at principal components with an eigenvalue greater than 1. Therefore in 2017 we will use PC1 and PC2 and in 2018 we will use PC1 and PC2(Even though PC2 is at 0.93 I considered it to be adequate because of its very close proximity to 1 and to compare PC1 and PC2 in 2017 and 2018).

We will compare the following values:

- 2017: PC1 and PC2

VS

- 2018: PC1 and PC2

Since we will only focus on PC1 and PC2 for 2017 and 2018 below is the amount of variation in the data that they represent:
*We are omitting PC3 and PC4 because they don't explain lot of variation in the data and because of the Kaiser Rule of Thumb

2017 Explained Variation		
	PC1	PC2
Proportion of Variance	0.4669	0.2696
Cumulative Proportion	0.4669	0.7365
2018 Explained Variation		
	PC1	PC2
Proportion of Variance	0.5471	0.2330
Cumulative Proportion	0.5471	0.7801

From the tables above, we can see the following:

- 2017: PC1 and PC2 account for 73.65% of variation in the data.
- 2018: PC1 and PC2 account for 78.01% of variation in the data.

This is important because now we can compare apple to apples because we are going to compare Principal Components that explain approximately the same amount of variation in the data and can therefore compare them approximately evenly.

Principal Component Analysis: Step 2 Principal Component Values

*We are omitting PC3 and PC4 because they don't explain lot of variation in the data and because of the Kaiser Rule of Thumb

2017 Principal Component Values		
Variables	PC1	PC 2
Sum of Sales	0.2061700	0.8476706
Impressions	0.5306945	-0.4437460
Cost	0.6199796	-0.1289767
Conversions	0.5398912	0.2605937

2018 Principal Component Values		
Variables	PC1	PC 2
Sum of Sales	0.4324174	0.6462527
Impressions	0.4685220	-0.6148829
Cost	0.5632444	-0.3000682
Conversions	0.5256025	0.3379875

*Another rule of thumb is to only look at principal component loading values that have an absolute value of 0.30 or greater because anything under that value is not vey important.

Interpretations:

- PC1 2017 (46.69% of Explained Variation):Impressions, Cost, and Conversions have the largest positive values; therefore, they have a positive correlation with PC1. As these values increase, more can we explain in PC1 for 2017.

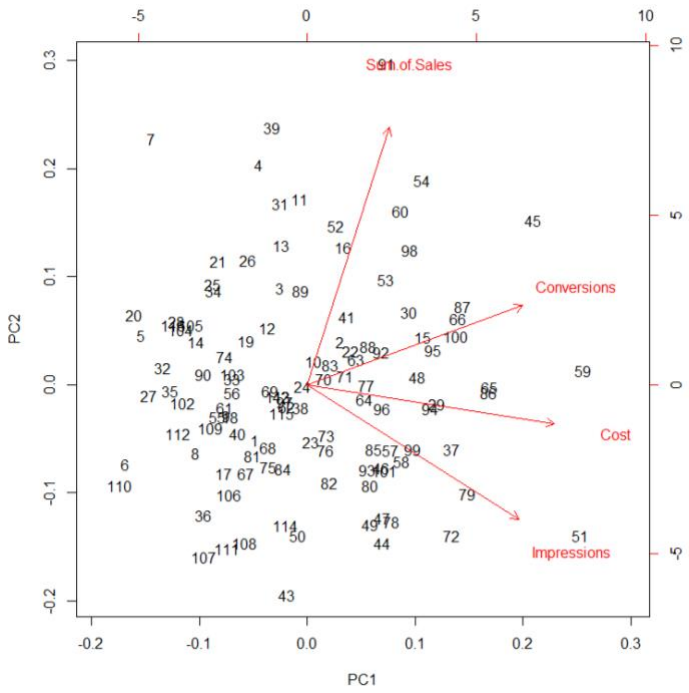
- PC2 2017 (26.96% of Explained Variation): Sum of Sales has the largest positive value; therefore, it has a positive correlation with PC2. As this value increases, more can be explained in PC2 for 2017. On the other hand, Impressions has a relatively high negative value; therefore, it has a negative correlation with PC2. As this value increases, more can we explain in PC2 for 2017.

- PC2 2018 (54.71% of Explained Variation): Cost has the largest positive value; therefore, it has a positive correlation with PC1. As this value increases, more can we explain in PC1 for 2018.

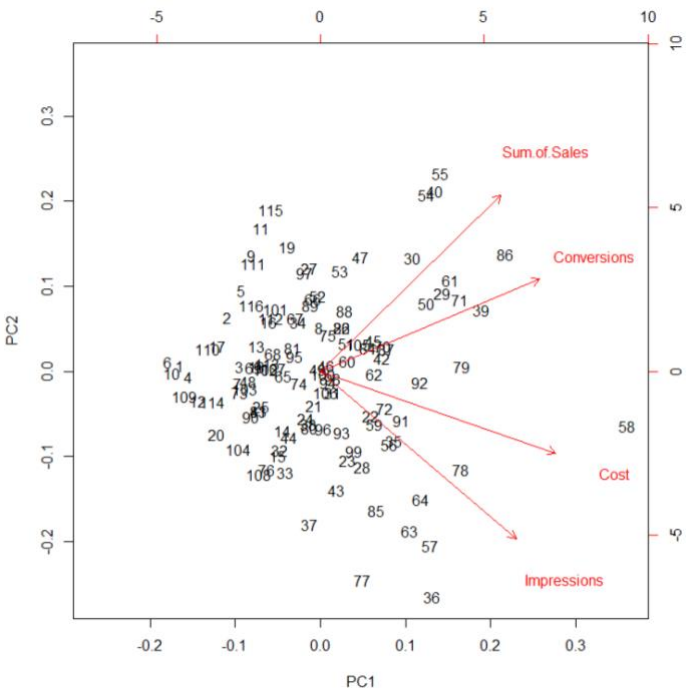
- PC2 2018 (23.30% of Explained Variation): Sum of Sales has the largest positive value; therefore, it has a positive correlation with PC2. As this value increases, more can we explain in PC2 for 2018.

Principal Component Analysis: Step 3 Bi Plots

2017



2018



Above are the bi-plots regarding the Principal Component data however what we will look at is the relationship between Sum of Sales and the Conversions variables because of the movement of their vectors from 2017 to 2018:

2017
Here we can see that the Sum of Sales and Conversions variable are relatively close together and correlated however there is still a good amount of space between them meaning in 2017 that if we did not have high Conversions, then it did not affect our Sum of Sales as much.

2018
Here we can see that the Sum of Sales and Conversions vector are a lot closer to each other than in 2017. Since they are very close to each other we can say that Sum of Sales and Conversions are now extremely correlated with each other in how if we did not have high conversions then we also did not have very high Sum of Sales.

Principal Component Analysis: Step 4 Varimax Loadings

Here we are using the Varimax Loadings which essentially maximizes the sum of variances of the squared loadings. Therefore, we are taking the loadings produced by the principal component analysis and then applying this function which maximizes the sum of variances.

You only want to apply the varimax function to the loadings matrix with p rows and k<p columns. For example, in this case we have 4 rows because we have 4 variables and so we would need to have less k<p columns. Since we are only looking at PC1 and PC2 because of their high variability explanation we are going to look at the 4 rows (because of the 4 variables) and at 2 columns representing PC1 and PC2.

Essentially, this function tells us the most important variable in every single principle component and we get the following output:

2017 Varimax Values		
Variables	PC1	PC 2
Sum of Sales	-0.151	0.859
Impressions	0.664	-0.194
Cost	0.620	0.131
Conversions	0.390	0.455

2018 Varimax Values		
Variables	PC1	PC 2
Sum of Sales	-0.124	0.768
Impressions	0.762	-0.131
Cost	0.617	0.164
Conversions	0.154	0.606

What we are looking for is the variable with the largest absolute value because that variable will tell us it's the most important one in the principle components. Here we can compare the following:

- 2017 PC1: Impressions = 0.664
- 2017 PC2: Sum of Sales = 0.859

- 2018 PC1: Impressions = 0.762
- 2018 PC2: Sum of Sales = 0.768

We can see that the most important variables in each component stayed the same with relatively similar values therefore the importance of each variable in each year did not really change.

Principal Component Analysis: Step 5 Results Summary

Step 1 Eigenvalues and Variation

- 2017: We should only look at PC1 and PC2 because of the Kaiser Rule of Thumb

- 2018: We should only look at PC1 and PC2 because of the Kaiser Rule of Thumb

- We can compare 2017 and 2018 Principal Components because they explain very similar amounts of variability in the data

Step 2 Principal Component Values

- 2017 Highest Correlated Variables to PC1: Cost (Positive)
- 2017 Highest Correlated Variables to PC2: Sum of Sales (Positive), Impressions(Negative)

- 2018 Highest Correlated Variables to PC1: Cost (Positive)
- 2018 Highest Correlated Variables to PC2: Sum of Sales (Positive), Impressions(Negative)

Step 3 Bi Plots

- 2017: Sum of Sales and Conversion are correlated however if Conversions are not as high it won't affect Sum of Sales as much, there are other variables not consider that affect Sum of Sales a lot more. (This year the average Conversions was 4.5 Conversions per day)

- 2018: Sum of Sales and Conversion are extremely highly correlated in how if Conversions suffer, then Sum of Sales will suffer. (This year the average Conversions was 4.1 Conversions per day)

Step 4 Varimax Loadings

The most important variable sin each component was the following:

- 2017 PC1: Impressions
- 2017 PC2: Sum of Sales

- 2018 PC1: Impressions
- 2018 PC2: Sum of Sales

Conclusion

In conclusion the biggest change from 2017 to 2018 in this analysis where we reduced the dimensionality of the data using Principal Component Analysis is that in 2017 we were less dependent on Conversions affecting our Sum of Sales. In this case there were other variables not measure in this case study that affected Sum of Sales. However, in 2018 Sum of Sales were highly correlated with Conversions and therefore, Sum of Sales was very dependent on Conversions. . For example, if Conversions increased then Sum of Sales increased and if Conversions decreased then Sum of Sales decreased. In 2018 we had an average of 4.12 Conversions per day which was during a year where Sum of Sales really depended on Conversions, however, in 2017 we had an average of 4.51 Conversions per day during a year where Sum of Sales did not depend on Conversions anywhere near as much.

Another Conclusion from the Varimax loadings is that in 2017 Impressions had more of a weight in 2017 than they did in 2018. For example, in 2017 they had a correlation with PC1 of 66.4% meaning that if Impressions increased than all other variables would increase, including Sum of Sales with a high correlation. However, in 2018 Impressions had almost a 10% increase by having a correlation with PC1 of 76.2% meaning that if Impressions increase then all other variables including Sum of Sales would increase with an even higher correlation. This tells me that in 2017 Impressions were not as important when it came down to the amount of Sales however in 2018 Impressions were important maybe meaning that the quality of advertisements has gone up and Belami has not adjusted to them.

Below is a quick list of takeaways:

Conclusions Summary	
2017 Conclusions	2018 Conclusions
Conversions did not affect Sum of Sales as much	Conversions really affected Sum of Sales
Impressions affected Sum of Sales (Correlation of 66.4% in PC1)	Impressions really affected the Sum of Sales (Correlation of 76.2% in PC1)

*We only look at PC1 Correlation Values from the Varimax Loadings because PC1 explains the most variability of the data for example, PC1 in 2017 explains 46.69% of variability and PC1 in 2018 explains 54.71% of the variability. Therefore, we get the most explanation of variability from these two Components.

To improve this analysis, I believe that more data should be fed into the model because at the end of the day as stated before I don't believe it's accurate to say that only these four variables represent the whole Belami E-Commerce Galtech Market, I believe the market is a lot more complicated with a lot more moving parts that we did not measure. However again, this is only a proof of concept showing what kind of results can be obtained by using machine learning techniques and I believe using this process to detect changes or anomalies would be a very suitable process to use.