Zappos Insights Report
Venkatasai Varada.

Venkatasai.varada@uconn.edu, Phone:860-994-2825

Contents

Business Problem:

The given dataset related to zappos e-commerce application comes with no specific business problem except that it asks to find a few insights, and perform various techniques like regression analysis, time series forecasting, segmentation.

Executive Summary:

Initial analysis of the dataset is followed by discovering insights about the user trends visiting zappos application through various platforms. The data reference helped me understand variables and actually made the analysis pretty much easy. Various data analysis tools including R, Tableau, SAS 9.4, Excel have been used to perform regression analysis, discover trends and forecasting the gross sales respectively. Excel's pivot tables helped in creating visualizations that make readers better understand the analysis. Linear Regression has been performed to predict the number of orders considering relevant input variables. Details about each method employed can be found in their corresponding sections.

Methodology:

The popular SEMMA (Sample, Explore, Modify, Model, Assess) approach has been used to analyze the dataset. Initial analysis of the dataset which includes checking correlations to see if various variables carry the same information revealed that most of the attributes visits, product page views, search page views are highly correlated. Further information on correlation is discussed in the later part of the report. Linear Regression and hypothesis testing have been performed to see the significance of variables.

Data Analysis:

Checking and eliminating multi collinearity is of primary importance as this may lead to improper and negative estimates for some variables which would have been positive otherwise. For all the numeric variables in the dataset below is a collinearity matrix. **Note**: All this analysis is made considering orders as a target variable.

	visits	distinct_sessions	gross_sales	bounces	add_to_cart	product_page_views	search_page_views
visits		0.995***	0.906***	0.791***	0.952***	0.913***	0.911***
distinct_sessions	0.995***		0.904***	0.772***	0.948***	0.906***	0.902***
gross_sales	0.906***	0.904***		0.590***	0.959***	0.842***	0.831***
bounces	0.791***	0.772***	0.590***		0.655***	0.724***	0.738***
add_to_cart	0.952***	0.948***	0.959***	0.655***		0.885***	0.879***
product_page_views	0.913***	0.906***	0.842***	0.724***	0.885***		0.990***
search_page_views	0.911***	0.902***	0.831***	0.738***	0.879***	0.990***	

Computed correlation used spearman-method with listwise-deletion.

Fig: correlations between all variables.

The power pf exploration is here. With jus initial analysis, it is found out that all the variables together cannot be used in regression as they are highly collinear and may lead to the problem of multi collinearity. Prediction of orders uses the above analysis to build regression models based on visits, product page views and search page views as they have come to be significant variables with hypothesis testing.

Data Cleaning and transformation:

Screening at the Platform variable reveals that there are various platforms through which zappos is accessed. One insignificant platform found in my analysis is Symbian OS. There are only 74 users with Symbian OS and considering the subject matter of operating systems Symbian OS, basically belongs to Nokia phones and Since Nokia has already migrated its operating system to Windows. Putting all these users in the Others category makes no difference and also reduces dimensionality.

A new variable difference has been constructed which is the difference between add to cart and orders.

Significance of this variable is discussed in detail in the insights section.

Talk about outliers and the day variable which represents time series

Time series Data preparation for forecasting: accumulating all the variables by grouping based on day so as to forecast gross sales. Data preparation for time series forecasting is done using SAS enterprise guide. Forecasting is done using SAS 9.4

Insights:

Having a look at the users accessing zappos application through various platforms, there are 18 million windows users which is around 35% of the total users, followed by IOS users. One important thing to notice here is that there are a very few users accessing zappos from android platform.

Recommendation: Technology is getting mobile every day and everyone is trying to use mobile phone than personal computers. There is definitely a need to promote the android application to increase its users. Zappos android app is rated 4.4 by 38k users. Promote android app like redirecting to download app if accessed via web in mobile phones and tablets. It has 5M downloads as of today. The data

belongs to 2013 where the number of users are 2M. There seems to be continual improvement in the usage of app.

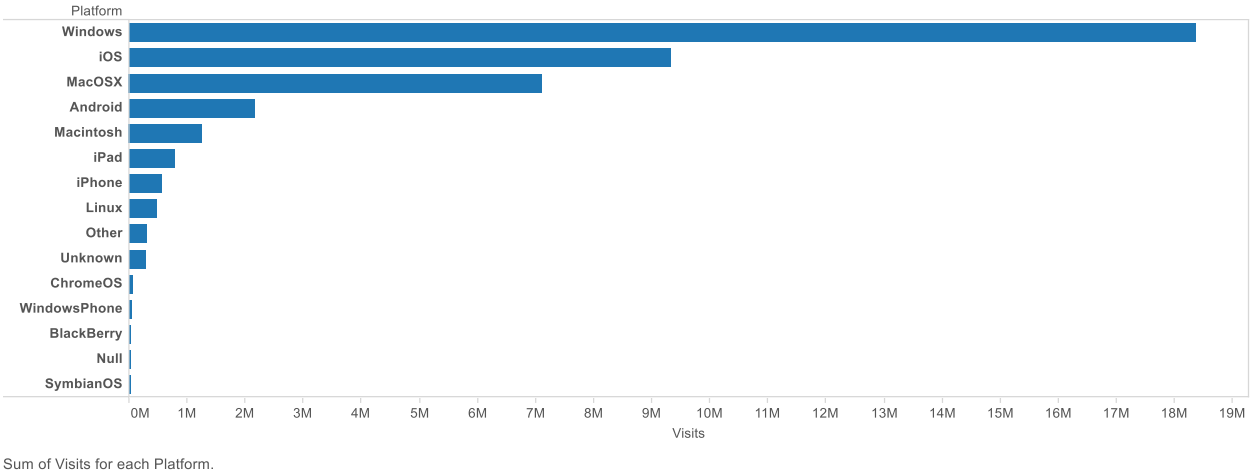


Fig: Tableau screenshot of numbers of users in different platforms.

Orders of each month grouped by new customer or old customer:

Below insight illustrates the number of orders placed by old and new users for a whole month. It is quite obvious that in a month the number of orders placed by old customers will be higher than new customers because of two reasons. First one being, old customers are more in number than new customers. Secondly, every new customer of present month will be an old customer for next month.

The importance of this insight is that the trend of old customers should follow that of new customers. When there is a new customer there are two options 1. Either he will leave zappos for various reasons 2. He will become an old customer for next month. We always want the 2nd option to happen and hence the trend of old customers is determined by the number of orders placed by new customers.

Zappos is positive in this aspect and can continue with the same strategy to attract and retain new customers.

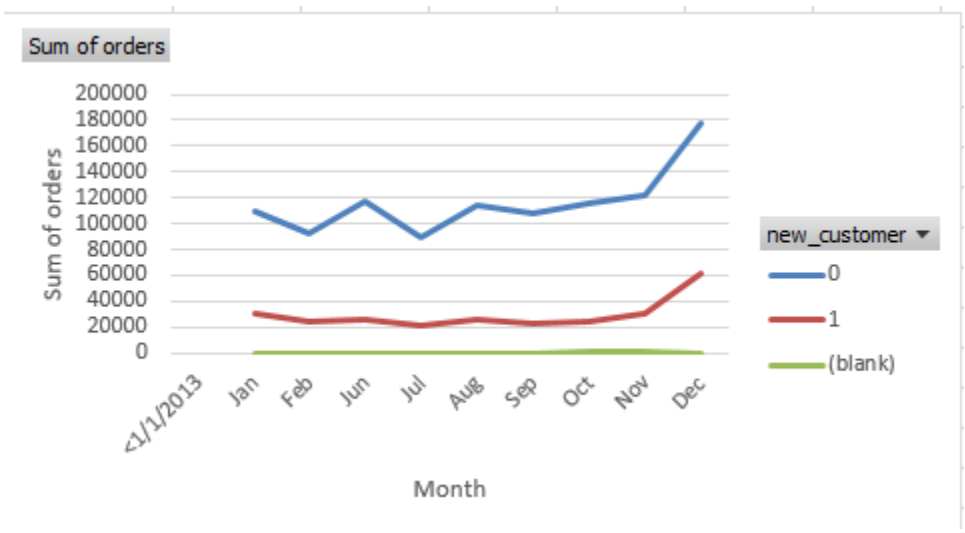


Fig: Monthly orders of new customers and old customers

Difference between add to cart and orders: When a customer adds a product to cart, half your business is done, we need to convert that in to business by forcing the customer to buy by creating interest. The between the number of items placed in the cart and orders placed plays a significant role in determining the business profit. We always have to minimize the difference. When you see at each customers's difference, it is very low. On a monthly basis for all the customers the difference magnifies as shown in the figure below.

Recommendation: By sending alerts to the customers on a periodic basis to remind them that they have placed a product in cart and yet to purchase. By sending promotional offers if there are similar products in cart for different users.

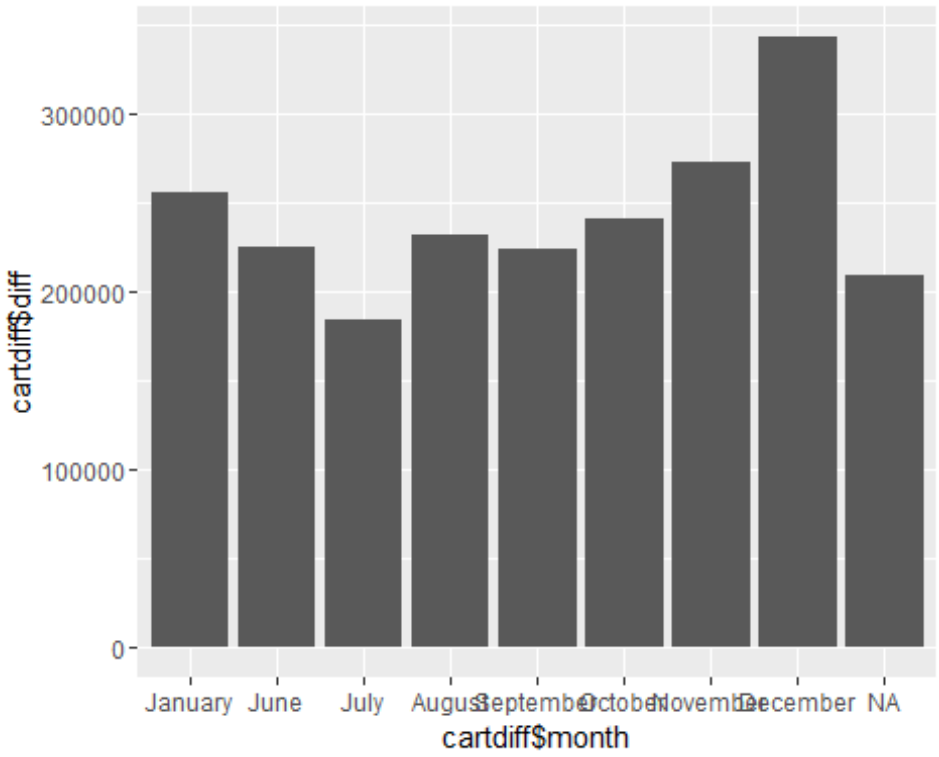


Fig: Difference between cart and orders month wise

Predicting orders:

Predicting orders for a whole month is not a good idea as this would not lead to any good recommendation to improve business. Had the data been customers and orders data, we could predict the orders a customer would make based on his/her time spent on the website, search page views etc., Then, we can focus on the customers who would order less in number as per the predicted model and then send promotional offers to those users to bring them back to business, Instead the data given here is web information. Based on the scope of the given data set the orders of each month is predicted so that we could compare them with the previous months to see if we are actually improving as a business. Three Regression models have been built based on visits, search page views and product page views individually as these variables determine the interest of a customer to place an order in general. But as analysts we need statistical proof. hypothesis testing revealed that these attributes are significant to predict orders but again, there is a problem of multi collinearity with the data. Hence, Individual models have been built in r by taking a 60, 40 train and test set.

Forecasting gross sales:

Before forecasting, let us have a look at the trend of sales of zappos by day.Gross sales trend from January through December reveals that there is no information for Mar, April, May months. Hence sub setting the data so that there is continuous time series to forecast sales.

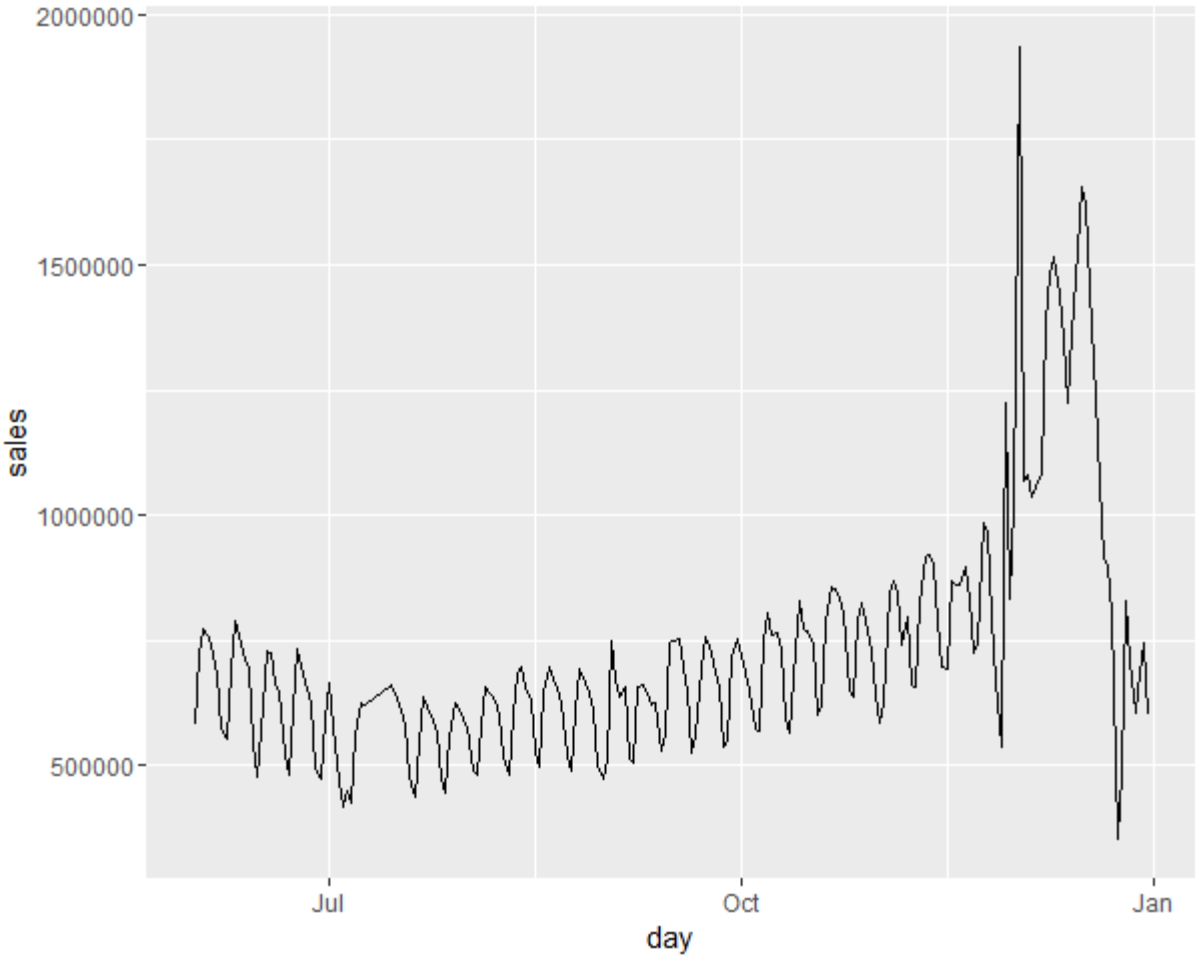
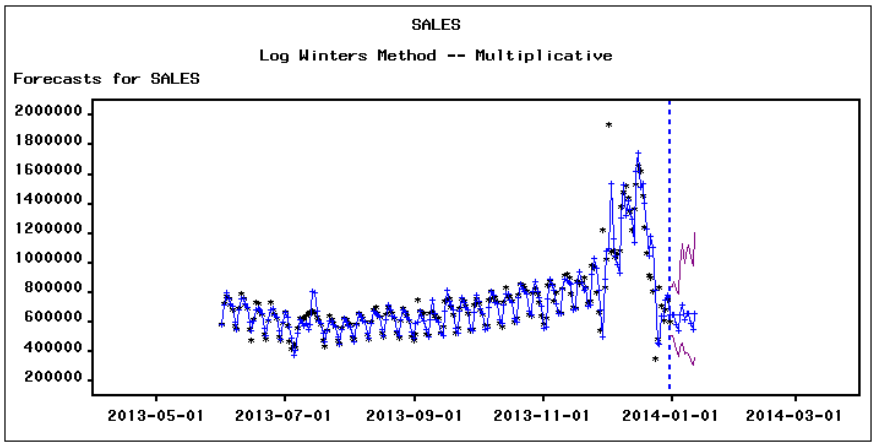


Fig: Gross sales trend from Jun to December

Above graph reveals that the sales have a general increasing trend over the given period and has a very high sale in November and December. It is quite obvious from the trend that the high sale in December is due to Christmas. Let us use this data to try and find out how sales would be in January. Had there been one more year's data about gross sales, the accuracy of forecasting would have been much better.



Sales Forecasting using Log winters multiplicative method.

Projected sales of Zappos in January every day can be seen in the above graph. However, the confidence interval is a little higher and I strongly feel that it can be reduced if there is more data of previous year's.

Forecast Data Set

SALES

Log Winters Method -- Multiplicative

DAY	ACTUAL	PREDICT	U95	L95	ERROR	NERROR	R
2014-01-02	.	647368	873983	479511	.	.	.
2014-01-03	.	586392	824764	416914	.	.	.
2014-01-04	.	540513	788385	370573	.	.	.
2014-01-05	.	648456	982444	428009	.	.	.
2014-01-06	.	719511	1127717	459066	.	.	.
2014-01-07	.	618624	994628	384763	.	.	.
2014-01-08	.	657761	1098807	393745	.	.	.
2014-01-09	.	654262	1121565	381662	.	.	.
2014-01-10	.	592590	1038050	338291	.	.	.
2014-01-11	.	546190	977207	305282	.	.	.
2014-01-12	.	655360	1208087	355518	.	.	.

Fig: Forecasted sales for the next 12 days in January 2014.