Introduction to Data Mining for Business Intelligence

Wisconsin Dells case study

MDS & Perceptual map

Sofia Gervasoni (100448791)

Introduction

The aim of this study is to understand the demographic and geographic characteristics of the customers that visit the Wisconsin Dells (WD), trying to understand if there are any connections between the activities done and the characteristics of the visitors. In addition, we are asked to help the company in determining how to improve its activity of leafleting, by identifying where the company could locate its brochures, based on the customer target.

We are given a dataset that contains the results of a survey on 1698 visitors, containing information about: the number of nights the visitor stayed at WD, the number of adults in the group, the number of children in the group, when they have planned to visit WD, sex, age, education, income and information about the activities done. In particular, there are 33 activities, and for each activity we have “YES” if the visitor did this activity or “NO” if not.

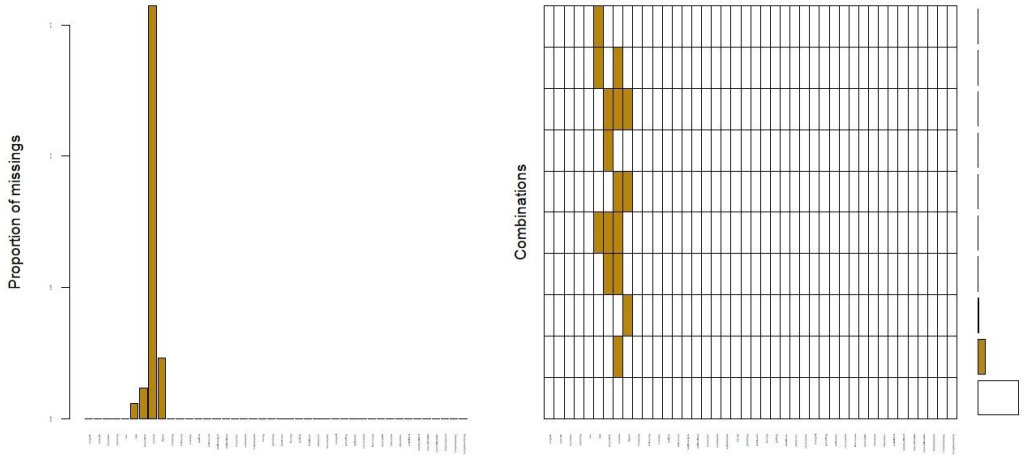
In addition, I have implemented a prediction model (with Machine Learning) that by introducing the characteristics (nnights, nadults, nchildren, age, …) of a group of visitors return a prediction of the number of activities that they will do.

Pre-processing

We start with a dataset with 1698 rows and 42 variables. All the variables imported are considered as “factor”, but there are variables of different nature (such as binary, ordinal, logical and nominal variables), so that I decided to do some transformations to the different types of variables. In particular, I transformed the variables:

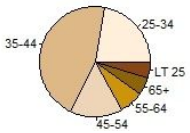
- nnights, nadults, nchildren, planning, age and education in ordinal variables (ord.factor)
- the variables that refers to the different activities in logical (TRUE=”YES” and FALSE=”NO”)

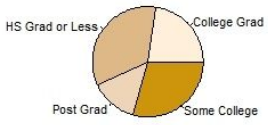
I did not do any transformation for the variables sex (that is a binary variable) and region (that is a character/nominal varible).

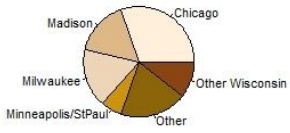


Moreover, I found out that the dataset contained 336 missing values. The missing values are distributed in 4 columns (income, region, education and age) and since they have different impacts for each of these variables, I decided to treat them in different ways.

The variables region, education and age contains respectively 39 (2.3%), 20 (1,18%) and 10 (0,59%) missing values. Since the missing values in these cases have not a great impact I decided to solve this problem by imputing the mode of the different variables (since they are nominal variables, so that I cannot calculate mean and median).







So that, the modes for *age*, *education* and *region* are respectively “34-44”, “HS Grad or Less" and “Chicago”.

When it comes to the variable *income*, the percentage of missing values is significant (15%), so imputing the mode (as in the previous cases) could be a little bit inaccurate. In this case, since income is an ordinal variable, I used the command *mice.impute.polr()* from the library “mice”. The function *mice.impute.polr()* imputes for ordered categorical response variables by the proportional odds logistic regression (polr) model.

The command mice.impute.polr() requests:

- the vector in which we have to impute the missing values (the column of the variable “income”)
- a logical vector (same length of the previous one) with FALSE where we have missing values and TRUE where we do not have them
- a matrix with numeric values. In this case I took into consideration the variables: nnights, nadults, nchildren, planning, age, education and the total number of activities done (a new variable that I created).

```
89  #install.packages("mice")
90  library(mice)
91  ry=vector()
92  ry[which(is.na(data[,8]))]=FALSE
93  ry[is.na(ry)]=TRUE
94  mice.impute.polr(data[,8], ry, new)
95
96  missing.income=mice.impute.polr(data[,8], ry, new)
97
98  data[which(is.na(data[,8])), 8]=missing.income
99
```

(for the R script about I treat the missing values please take a look at the file “missing.R”)

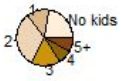
After this pre-processing of our dataset, we can finally start our analysis.

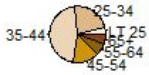
(for the R script about the overall analysis please take a look at the file “official.R”)

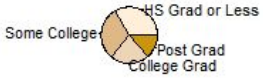
About people that visit Wisconsin Dells

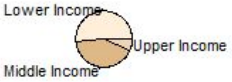
General characteristics

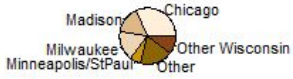
Most of the people that visit WD tend to stay *2 nights or 0 nights*, not many people come to stay just 1 night. Groups are usually made by *2 adults and 2 children or without children* (only 2 adults). People tend to plan the visit at WD with some advance, actually most of the people booked the visit "*one month or more ago*". Most of the visitors that joined the survey were women and they are aged between *35 and 44* years. When it comes to education, most of the visitors have *HS graduation or less* or they attended *some college*. Most of the visitors have a *lower or middle income* and come from *Chicago*.



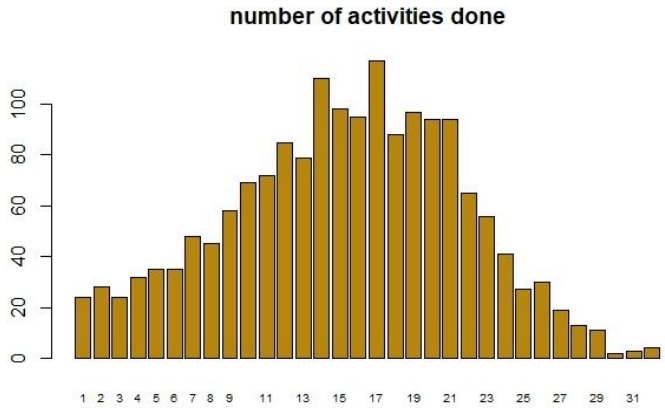








Most of the visitors tend to do between 14 and 19 activities, as we can see from the following barplot (the minimum number of activities done by a visitor is 1, whereas the maximum number of activities is 32).

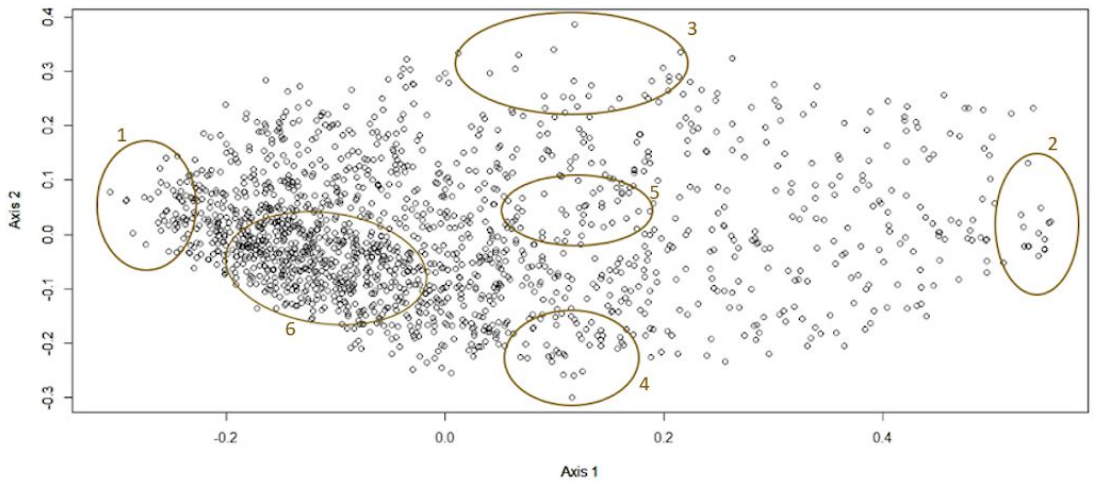


Customer segmentation and type of activities (MDS)

Thanks to the Multidimensional Scaling analysis I was able to understand the common (or not) characteristics of the different visitors. Actually, the MDS analysis returns information about how similar or dissimilar are the different observations. As closer they are, as similar they are (viceversa, as far they are, as dissimilar they are).

In order to apply this analysis it is necessary to obtain the distance matrix. Since in this case our data frame has no numeric variables, we cannot use the command *dist()*, but we can use *daisy()* from the library "cluster". Moreover, since our dataset is made up by a variable type mix and we do not have numeric variables, we cannot use the euclidean distance, but we can use the Gower distance.

So, thanks to the commend *daisy()* and the Gower distance, I obtained the distance matrix on which I applied the command *cmdscale()* and then I plotted it, obtaining the following graph.



The main difference between the observations on the left side of the graph and the ones in the right side lies in the number of activities done. Actually, the visitors in the right side of the graph are the ones that tried the lower number of activities, whereas the ones in the left side are the ones that tried most of the activities (some of them tried all of them).

But this isn't the only characteristic that makes possible a distinguishment between the individuals. In order to better understand if it is possible to segment customers, I established 6 main groups of individuals, located in different positions of our graph, and I found out some interesting results.

I started my analysis with the ***first group*** of observations *on the left* (indicated in the graph with number 1). As said, these individuals tried most of the activities (generally speaking they tried more than 29 activities) so it is not possible to discern between the different

activities, even if the activity less done is *bungee jumping*. Generally, they stay at WD more than 4 nights and they mostly come from Chicago. The groups are usually made up of 2 adults and at least 1 child. They usually plan to come to WD with a large advance (One Month or More Ago) and they are aged between 25 and 44.

The ***second group*** analysed is the one in the *right side* (indicated with 2). The individuals in this group tried very few activities (like 1 or 2), but the most interesting thing is that they mainly come for the *Water Park*. They mostly come from Madison and they do not stay more than one day (zero nights). The groups are usually made by 1 or 2 adults with not many children (an average of 2). They usually plan the visit "last minute" (this week) and they are usually aged between 25 and 64 years. They usually have a lower/middle income.

The ***third group*** is the one *up in the middle* (group 3) is characterized by having an average number of activities done (between 7 and 16). There is no fixed number of nights of staying for this group but an average of 2. This type of group of visitors is made up of adults only, with an age higher than 55. They do not usually plan the activity with much advance (this month or week) and they usually have a lower or middle income. They come to WD for shopping, scenery, museum, gambling, boat tours and ride ducks, and they eat fine. They avoid activities like outdoor pool, boat swim, amusement park, go kart, water park and bungee jumping and they do not eat fast food.

The ***fourth group*** is the one *low in the middle* (group 4) is characterized by having an average number of activities done (usually 10). They stay on average 2 nights and they are more likely to be families. Actually, the group of visitors is usually made up of 2 adults and 2 children and the interviewed person has between 35-44 years. They tend to plan the visit with some advance (usually One Month or More Ago) and they usually have a low/middle income. They usually visit WD for the indoor and outdoor pool, the amusement park and the water park and they usually eat fast food. They avoid activities like antiquing, hiking, gambling, fishing, golfing, circo, tbski, helicopter, horse, stand rock, theater, bar/pub and bungee jumping and they do not eat fine.

The ***fifth group*** is the one located *in the middle* of the graph and it is characterized by having an average number of activities done (usually between 8 and 12). Even in this case the number of nights is not defined, but we could say an average of 2. These groups of visitors are mostly made up by adults (always more adults than kids). There are not relevant common characteristics for age, planning and region but they usually have a lower income. They mostly come for shopping or shopping broadway, but for the other activities we do not have relevant common trends.

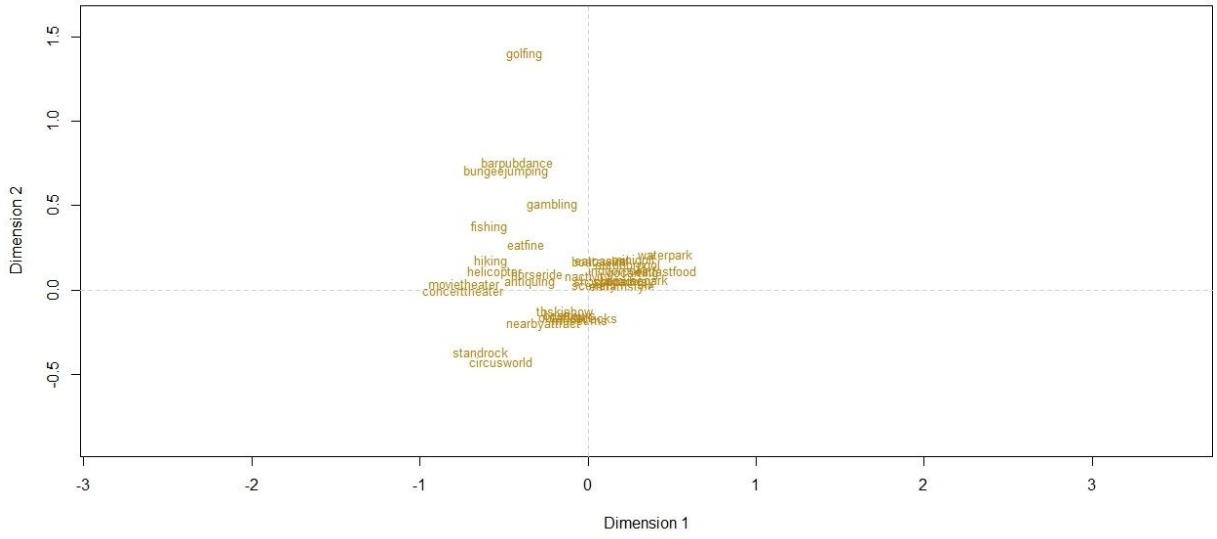
I also decided to analyse this ***sixth group*** (the one indicated with 6), because here are located most of the individuals, so that in general they represent what people mostly do when they come to WD. The number of activities tried in this case range between 19 and 22. They mostly stay in WD 2 or 3 nights, they are more likely to be families (usually adults with 2/3 children) and they tend to plan the visit with some advance (This month or One

month or more ago). The age of the people interviewed in this case mainly range between 35 and 44 years. They mostly have a middle income and they mostly come from Chicago. They tend to do activities like shopping, scenery, outdoor pool, amusement park, minigolf, go kart, waterpark and shopping broadway, and they usually eat casual or family style. The activities less done are antiquing and bungee jumping.

Patterns in visitor activities (perceptual map)

I continued my analysis identifying any possible relationship between the existing activities. In order to do so, I created a perceptual map using the command *anacor()* from the library anacor (this library does not accept character variables, so I did my analysis on the variables that refers to the different activities and the total number of activities done).

Plotting the results of the command a*nacor()*, I obtained the following graph.



From this graph, is it possible to understand that there are some relationships between the different activities offered by WD.

The activities in the *first quadrant* contain the activities that are less done by the visitors (even if there are any other reasons for which these activities are so related). For example, bungee jumping and bar/pub/dance are activities usually made by young people, that could be the reason why they are so close. But in general we could say that these activities are in this quadrant for 2 main reasons. The first reason is that most of these activities are expensive and most of the visitors have a low/middle income, so not many of them can afford these activities, because they are expensive (e.g. golfing, eating fine, helicopter …). The second reason is that most of the visitors came with the family (so with kids) and are aged between 35-44, so that they tend to avoid this type of activities because they maybe prefer doing activities that each member of the family can do (for example kids cannot gamble or doing bungee jumping and probably they prefer amusement / water park

instead of antiquing or hiking). For the same reasons, *eating fine* belongs to this group of activities, actually families prefer to eat casual, fast food or family style.

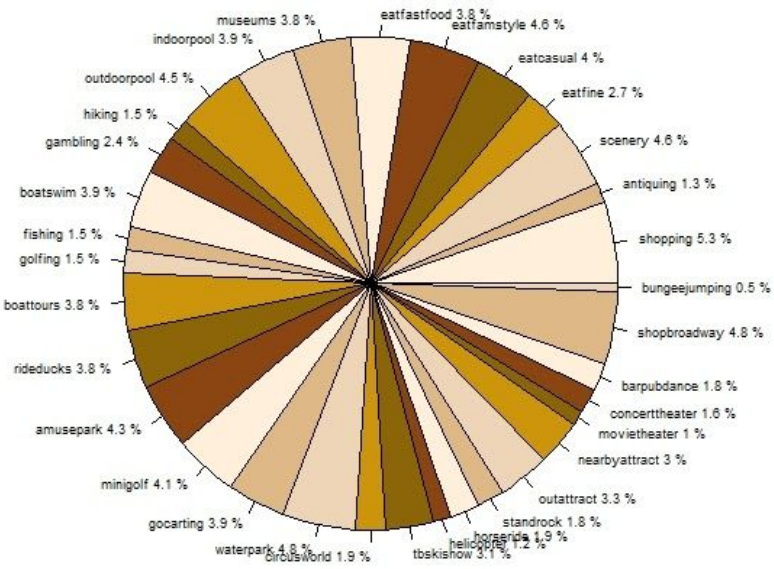
The activities that belong to the *second quadrant* are the most done by the visitors, that are for example the water / amusement park, shopping /broadway shopping, indoor and outdoor pools and so on. To confirm what was said before, since the main visitors are families, these are the activities that could be done by each member of the family and are also affordable activities, that is why they are so common. Moreover, visitors that do these activities tend to eat fast food, casual or family style (and not *fine*).

From the graph it is also possible to see the relation between some activities like movie and concert theater, stand rock and circus world, tb sky show and nearby attractions. A common trend for the third quadrant could be the fact that there are more or less all activities in which there are shows to watch or some interest point (like stand rock) just to visit.

Advertising strategy

The company should locate most of the brochures in the most common attraction, in doing so they can reach a wider segment of customers.

Thanks to the following pie chart, it is possible to understand which are the activities that most of the visitors do when they come to WD.



Shopping (5.3%) and Shopping broadway (4.8%) are the activities most done by the visitors, so the company should focus its advertising strategy in these two attractions. Even if there are some activities like water park (4.8%), amusement park (4.3%), minigolf (4.1%), scenery (4.6%) and outdoor pool (4.5%) that are also so common.When it comes to the resurats, the most common is the “eat family style” (4.6%).

More specifically, when it comes to different ages, visitors tend to prefer different attractions. For example, people that are older than 55 that usually come in groups of anly adults, usually do activities like scenery, museum, gambling, boat tours and ride ducks, and

they eat fine, so that in this attraction the company should locate advertising for this target of people. When it comes to younger people that usually come with children, tend to do activities like indoor and outdoor pool, the amusement park and the water park and they usually eat fast food, so in these attractions WD should locate advertising for this other type of visitor.

Studying the correlation between the different activities, I also found out a relationship between some of them. The stronger relationship is between *boat tours* and *ride ducks,* that means that most of the people that do one of these activities do also the other one, so that the company should locate advertising of boat tours in ride ducks and vice versa. The same happens for other activities like *indoor and outdoor pool*, *shopping and shopping broadway* and *amusement and water park*. There also some other relationships (even if a little bit weaker) between: *eating family style and amusement park*, *outdoor pool and amusement park/go karting*, *hiking and fishing*, *tbs sky shows and boat tours/rideducks*, *amusement park and go karting/minigolf*, *minigolf and go karting* and *movie and concert theater*. By taking into consideration these relationships the company could locate the advertising in a strategic way, because the visitor that likes one of these activities likes also the other one, so it is likely that will take into consideration the opportunity to also try the other activity. In this way we create like a virtuous circle.

Additional analysis: prediction model (Machine Learning)

(for the R script please take a look at the file "wisconsin_machinelearning.R")

I tried working in the machine learning space with the aim of finding out a prediction model, which by imputing the characteristics of the individual (nnights, nadults, nchildren, age, sex, education, region, panning and income), return the number of activities that the goup of visitor will try. We are in a regression problem, where the number of activities is the response variable.

I started working with an *rpart* regression model and I also scaled the data. I decided to impute the mode when it comes to NAs. I also added drop constants and dummy encode, but they do not seem to affect our overall error.

Since we are working with a machine learning process, we are trying to train a new model based on the existing data, so that it can make the best possible decision. Since we can train it on the existing data, it makes perfect sense to divide all the rows into a training part and a test part, because if we were to train our model with all of the rows (all the data available), we would not have a correct measure of its ability (it would be completely aware of all the situations). Having certain test indexes, makes sure that we can evaluate it in a correct way and I also chose to do so using a root mean square error measure. There are different ways in which we can split up the data (into training and test), and I decided to use the *holdout method* (but also cross validation could have been valid).

It is imperative to remember that if we want to reproduce and repeat the calculations it is necessary to use *set.seed()* .

The error founded seems to be acceptable, but in order to understand if it was truly acceptable, I decided to test the model and the data using a k-means regression method. I worked with it in the same way I did before and I found out that the errors were similar but rpart had a slightly lower one, so I decided to keep on using this method.

Something else that can be done when it come to machine learning is to adjust the hyper parameters of the methods and when it comes to rpart, it is very easy to find that the main parameters are called *"maxdepth"* and *"minsplit"* (online are available the lower and the upper values that these parameters can at most take). So that, at this point I tried to test the best possible values for both parameters (and not the default values as before), creating a new division of the data (with a control grid method of evaluation), finding out that the ideal value of the hyperparameters were: minsplit=40 and maxdepth=30 (in order to have the lowest error possible which is rmse.test.rmse=5.7860489). So I forced the learner to take on the values that I found to be ideal (keeping all the other conditions same as before), finding out the new error.

Finally I was able to create the final model, training the learner found after setting the hyper parameters. So I found the prediction that the model would make and the RMSE of this final model to understand how precise it would be.

With saveRDS() I saved the model to make future predictions so that the company can use it any time new data gets entered into its system.

I also create some “fake” data to see potentially what each group of customers would imply for the company, and I get the following results:

	nnights	nadults	nchildren	planning	sex	age	education	income	region
1	2	3	0	This Week	Female	35-44	Post Grad	Middle Income	Chicago
2	1	2	3	This Month	Male	LT 25	Some College	Lower Income	Other

```
> pred
Prediction: 2 observations
predict.type: response
threshold:
time: 0.03
  response
1 16.36634
2 13.47668
```

The first individual is most likely to do 16 activities, whereas the second one is more likely to do 13 activities.