BUSINESS CASES WITH DATA SCIENCE

MASTERS' DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

Business Case 1: Hotel Customer Segmentation

Palm & Company

Pedro Santos (M20190420)

Ana Cláudia Alferes (M20190420)

Lennart Dangers (M20190251)

Michael Machatschek (M20190054)

March 2020

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INDEX

1. INTRODUCTION

In this report, we will present you a solution for the problem with our current market segmentation approach. Like our used process model for the development of this solution, the structure of the report is aligned with the CRISP-DM process model. We begin with the business understanding part, where we address the current problems with our customer segmentation model and exemplify the benefits of a more advanced clustering approach. In the next chapter, we will explain the data and the whole data understanding process. Based on that we present to you the model, that we have created, and interpret the results and their implications to our business. Finally, we share theoretical considerations about deployment and maintenance plans and future model improvement.

2. BUSINESS UNDERSTANDING

2.1 BACKGROUND AND CURRENT SITUATION

Our current customer segmentation is a very simple, yet the industry standard, approach based on the origin of the customer. It does not contain information about any demographic characteristics, nor information about the booking behavior of our customers. This oversimplified representation is likely to lead to an erroneous interpretation of our customers. As a result, we experience increased inefficiencies in our marketing campaigns, and we miss opportunities in our CRM and customer acquisition efforts.

As also our new marketing manager made clear, our current customer segmentation approach is not adequate anymore. We therefore propose you a state of the art, data-driven market segmentation solution. Our data resources allow us to create a more sophisticated model.

2.2 BENEFITS OF AN ADVANCED CLUSTERING APPROACH

Our proposed solution has several benefits, which this part of the report will address. We will be able to create more efficient marketing campaigns with a higher conversion rate. By better understanding our existing customers we can not only improve our CRM but also use this information to target new customers with similar characteristics.
Additionally, we may get useful insights about our future product offerings. Finally, as most of our competitors are still using the industry standard market segmentation, we can exploit the maximum potential of a data-driven marketing approach by delivering our customers personalized marketing messages with a higher conversion rate than our competitors.

2.3 OBJECTIVES & SUCCESS CRITERIA

The final objective of our solution is to meet the benefits described above. We want to achieve this by creating new customer clusters, that have a high intra-cluster similarity and a low inter-cluster similarity. The number of clusters should be high enough to make them differentiable, yet low enough to stay manageable.

The success criteria for this project lie in our future marketing efforts. As we mentioned above, we predict a higher conversion rate in our marketing campaigns, a stronger relationship with our existing customers and a more successful customer acquisition process. Measurable KPIs for these goals can be developed with our marketing department.

2.4 PROJECT PLAN

As mentioned above, we followed the CRISP-DM process model in our project.

After discussing the business background and effects, we started to prepare the available data. We began by an exploratory data analysis to get insights for the data preparation and data engineering phase. After selecting the most relevant features, we started the modelling phase. During all these phases we had several interactions between the phases to improve our decisions in previous steps.

3. CUSTOMER SEGMENTATION PROCESS

The customer segmentation process contains three of the CRISP-DM process steps: Data Understanding, Data Preparation, and Modeling. Each process step is connected. Therefore, interactions among each other are possible. For instance, uncertain datatypes in data preparation can cause a step back to get a proper understanding of the data.

3.1 DATA UNDERSTANDING

This chapter contains all parts that explain the data based on the obtained knowledge in business understanding. In general, data understanding explains the variables in a more technical way and show the data source and evaluate the quality of the data. The CRISP-DM process is not unidirectional. For this reason, interaction with the business understanding step might happen if the purpose of specific variables is uncertain.

3.1.1 DATA COLLECTION AND DESCRIPTION REPORT

Data Collection Report

The given dataset includes data about customers of the hotel chain, which has more than four hotels in Portugal and at least one in Lisbon. The customer names and their identity are anonymized by a hash algorithm. All values for time-based columns, such as age and the days since creation, are calculated at the extraction date of the dataset.

Data Description Report

The raw data contains 28 columns, which represent different features. The number of rows is 111.733 and each row stands for one customer, which can be a duplicate. Most of the features are dummy variables (13, all special requests) or numeric (10). Excluding name and ID, there are three categorical variables (nationality, distribution channel and market segment).

3.1.2 DATA EXPLORATION AND QUALITY REPORT

Data Exploration Report

At the beginning of the data exploration, it is appropriate to get familiar with the dataset respectively get an overview. For this reason, the following steps in python are a good approach:

Step	Purpose	example method in python
1. examine datatypes and missing values	How is the distribution between numerical and categorical variables? How many missing values (NaN) are existing?	pandas.DataFrame.info()
2. get a statistical overview	What are the main values like the minimum, maximum and mean of each variable?	pandas.DataFrame.describe()
3. examine the distribution with histograms	How is the variable normally distributed?	pandas.DataFrame.hist()
4. detect outliers with boxplot-diagrams	Where are the outliers?	seaborn.boxplot()
5. examine correlation with correlation-matrix	Which variables are highly correlated and could be discarded?	pandas.DataFrame.corr()

Step 1: As already mentioned above, most of the variables are either numerical or dummy variables. According to info-method, there are missing values in Age (4172) and DocIDHash (1001). Missing values in DocIDHash represent customers, who do not want to show their ID at the hotel desk. The technique to deal with missing values is explained in the next chapter.

Step 2: The statistical overview shows duplicates in NameHash (4149) and DocIDHash (8253), which can refer to the same customer. Further occurrences are negative values in Age and AverageLeadTime as well as zero values within BookingsCheckedIn, but values in other variables.

Step 3: The variable Age seems to be normally distributed, whereas other variables are not. The distribution within the special requests (variables that start with “SR”) shows, that most of the special wishes are not requested by a customer.

Step 4: Even though the boxplot diagrams show potential outliers, most of the cases are unusual but correct. We found noticeable values in AverageLeadTime, LodgingRevenue, and OtherRevenue. The outlier treatment is explained in the next chapter.

Step 5: There is no remarkable correlation among all variables. For this reason, no variable can be discarded.

Data Quality Report

All in all, the quality of the dataset is decent. For example, it contains fewer missing values. The ratios of missing values in Age is 3.73 % and 0.90 % in DocIDHash, which is rather low. Most of the data, such as nationality, is reliable because it is registered by the staff at the hotel desk. For applying k-means, we should consider, that the distribution of the categorical value DistributionChannel is not well balanced. Nevertheless, it is vital to deal with the above-mentioned occurrences in data preparation.

3.2 DATA PREPARATION

After passing the steps of business and data understanding, data preparation is the next process step. According to several surveys among data analysts and data scientists, data preparation is the most time-consuming step along with all the steps. This chapter is divided into Data Cleaning, Reformatting and Merging and into Feature Engineering.

3.2.1 DATA CLEANING, REFORMATTING AND MERGING

At first, we dropped customers whose values do not fit the business logic. Such as customer with zero revenue but have checked in more than zero times. In that case, we assume these are room for the staff. If BookingsCheckedIn, BookingsNoShowed and BookingsCancelled have all zero values, these rows also were dropped, because these might be customers from older systems and do not give insights. Through this step, the new number of rows is 77964 (minus 33769).

As mentioned in the previous chapter, we have taken into consideration the missing values in Age and DocIDHash. With the variable Age, we decided to drop missing, negative and values above 100 rather than predict these values. Firstly, the ratio of missing values is only 3.73 % and we assume a solution, that is more accurate. The missing value ratio within DocIDHash is 0.9 %. Since it is not possible to predict a DocIDHash, all rows were dropped. After this step, rows with not unique values in DocIDHash were merged through a group-by method. To be more certain to remove only duplicates, a new variable (IDMerge) was created, which combines NameHash, DocIDHash and the country as a string type. Depending on the variable, the aggregate function was either the mean, the mode (a mode-function for all special requests), the sum or the first value.

Lastly, the handling of outliers is done with respect to the visual (boxplots-diagrams) and the detailed (verify all features of the potential outlier) exploration of each numerical variable. As a result, outliers were removed at a defined threshold. The table below shows the development of the number of rows after each step in data cleaning.

rows	Step	Delta to previous step
111733	Raw dataset	
77964	After dropping customers with zeros in all: BookingsCheckedIn, BookingsNoShowed and BookingsCancelled + Customers, who have CheckedIn above zero, but no revenue	minus 33769
75648	After merging DocIDHash (merge duplicates)	minus 2316
72922	After removing nans in ages	minus 2726
72889	After dropping outliers	minus 33

3.2.2 FEATURE ENGINEERING

It can be rather difficult to train a model without having relevant features. Most likely, raw data does not show all the relevant information which can be extracted from the dataset. Therefore, it is vital to extract and create new features out of the raw dataset. This process is called feature engineering.

In the given dataset we created the following new features:

Feature Name	Computation	Purpose
PricePerNight	LodgingRevenue / RoomNights	to drop outliers by defining a minimum threshold per night (20 Euro per night)
OtherRevenuePerPersonNight	OtherRevenue / PersonsNights	to drop outliers by defining a maximum threshold per person and night (10000 Euro per person and night)
PersonPerRoom	PersonsNights / RoomNights	to distinguish between for example singles and couples, or customer with children
GroupSize	PersonsNights / BookingsCheckedIn	to distinguish between for example singles and couples, or customers that booked for a group
AvgNights	RoomNights / BookingsCheckedIn	to distinguish between short and long stays

SRLocation	Sum of all variables related to the location	to combine all variables, which are related to the location of the room within the hotel
SREquipment	Sum of all variables related to the equipment of a room	to combine all variables, which are related to the equipment of a room
Nationality2		to group the most common to obtain a clearer view regarding nationality
DistributionChannelNew		to reduce categories to get a better distribution

To obtain a more balanced distribution, we binned some of the variables. “Binning methods smooth a sorted data value by consulting its “neighborhood,” that is, the values around it. The sorted values are distributed into a number of “buckets,” or bins.” (Han, Kamber, Pei,2012). The binned variables are Age, AvgNights, GroupSize, AverageLeadTime, and DaysSinceCreation. For the bin size of age, we decided to stick to the standard of the hotel industry: 0-24, 25-44, 45-64, 65+. All other binned variables are following neither the equal-width nor equal-frequency approach. Instead, we created bins, which improve the distribution and follow a business and cluster perspective.

The last steps before modeling are to convert all categorical variables into numerical. This step is called encoding and splits one categorical variable into the number of given categories. Whether one category occurs, it will count 1. All other columns will have 0. (Provost, Fawcett, 2013).

3.2.3 FEATURE SELECTION

To select the features we would use, we first ran the model with all features in order to take quick insights. The goal is to remove features that i) are not diverse throughout the clusters, ii) are not considered important during the PCA or iii) do not add enough value/information for our clustering goal. Through several tries and adjustments, we reached a good model with desirable results.

The end model contains 9 features (divided in bins, therefore, totaling 31 features): Age, AvgNights, GroupSize, AverageLeadTime, Nationality2, PricePerRoom, DistributionChannelNew, SREquipment and SRLocation.

3.3 MODELING

Our final model uses two widely known Data Mining processes: Principal Component Analysis (PCA) followed by K-Means.

Due to the high number of features, it was necessary to use a method to reduce dimensionality in order to summarize and exclude redundant features. The PCA technique is used with this goal in mind: we selected the first 23 components.

Afterwards, we applied K-Means on top of the 23 components to find similarities. For this technique, we aimed to minimize Sum of Square Errors while keeping a manageable number of clusters: 6. We will talk in more detail about this choice later.

4. RESULTS EVALUATION AND PROFILING

Our cluster solution is one that both minimizes the sum of square distances but also maximizes business understanding, therefore, we found that 6 clusters represent a good combination between both.

Cluster 0: 'Locals' and Corporate travelers:

- Contains 11.123 customers (15%)
- Mostly middle aged and younger adults
- Mostly Iberian (21% Portuguese, 9% Spanish)
- 57% prefer short stays (1-2 days)
- 47% travel in pairs, whereas 38% travel alone
- Very spontaneous group (57% book up to 30 days in advance)
- Moderate to high spending
- Channel: Other, Includes majority of Corporate customers
- Rarely request

Cluster 1: Young European adults who travel in pairs:

- Contains 10.481 customers (14%)
- Younger adults (25-44)
- Mostly Europeans (except Iberia)
- Medium stays only (3-4 Days)
- Prefer to travel in pairs
- Early bookers
- Moderate spenders
- Channel: Travel Agency
- Often request equipment related

Cluster 2: Solo travelers and late bookers with high spending:

- Contains 9.093 customers (12%)
- Mostly middle aged and younger adults
- Evenly distributed nationalities
- Short stays only (1-2 days)
- Prefer to travel alone
- Tend to be late bookers (ie, more spontaneous)
- Moderate to high spenders
- Channel: Travel Agency

Cluster 3: Solo travelers and early bookers with low spending:

- Contains 13.375 customers (18%)
- Diverse age group
- Rest of the word (22%) followed by Germany (19%)
- Short stays only (1-2 days)
- Always travel in pairs
- Tend to be earlier bookers (ie, longer lead times)
- Moderate to low spenders
- Channel: Travel Agency
- Often request equipment related

Cluster 4: Early booking Europeans who enjoy longer stays and lower spending:

- Contains 9.896 customers (14%)
- Mostly middle aged and younger adults
- Mostly Europeans (except Iberia)
- Long stays (5+ days)
- Prefer to travel in pairs or bigger groups
- Very early bookers (37% book within 121-365 days)
- Low spenders
- Channel: Travel Agency

- Often request equipment related

Cluster 5: Older Europeans who enjoy medium stays:

- Contains 18.950 customers (26%)
- Older customers 45+, including 24% 65+
- Mostly Europeans (except Iberia)
- Medium Stays only (3-4 days)
- Travel mostly in pairs
- Very early bookers (34% book within 121-365 days)
- Moderate to low spenders
- Channel: Travel Agency
- Often request equipment related

5. DEPLOYMENT AND MAINTENANCE PLAN

The new customer segments can be used from our marketing department in a manual way. There is no need to automate any process here. Additionally, it would be advisable to cyclically rerun the segmentation analysis with the updated data. We suggest doing it after each main season.

After doing the segmentation analysis for several seasons, we can analyse the changes over time. Additionally, we can also develop a model for Demand Forecasting.

As this model should be renewed every few seasons, we do not need any further maintenance of the model. Of course, the database and the data collection processes should be maintained. Improvement regarding these processes will be discussed in the next chapter.

6. IMPLICATIONS FOR BUSINESS

Hotel customers were divided into 6 clusters as this represents an optimal business solution, where we avoid redundancies and avoid creating too many marketing profiles. With this in mind:

- Demographically, some clusters tend to show some similarities, such as clusters 0 and 2, and are usually populated by individuals between 24-65 years old, where cluster 5 contains older customers and cluster 1 contains younger adults;
- Behavior wise, there seems to be a relation between spending fewer nights with smaller lead times, as well as higher prices, as shown by clusters 0, 2 and 3. The opposite is also verified, shown by clusters 1, 4 and 5. Cluster 2 has a preference on traveling alone whereas most clusters show that it is more common to travel in pair.

Targeted Marketing Initiatives

For marketing initiatives, there are some strategies to both acquire new customers and encourage older customers to return.

- Discount based vouchers for in-hotel activities for groups to stay longer (such as cluster 1,4 and 5), in order to increase spending;
- Discount based vouchers for longer staying groups (cluster 4) for extra nights, ie, percentage discount for each night they stay above a certain threshold;

- Discount based ‘Last Minute Offers’ for late bookers, such as clusters 0 and 2, to increase occupancy rate during low seasons;
- Offer tourist itineraries for core city attractions for short staying customers, such as cluster 0,2 and 3, to create deeper interest in the city and encouraging a return;
- Loyalty based rewards for returning customers, which could appeal to younger customers who stay for medium and long periods, such as clusters 1 and 4;
- Partnership and cross-sell of paid city activities (for example, zoo) for lower prices to medium and long staying customers (cluster 1,4,5);
- Event based discount, targeting Corporate customer (cluster 0), for example, web summit.

7. CONSIDERATIONS FOR MODEL IMPROVEMENT

To improve the customer segmentation analysis, we have two suggestions.

Firstly, we would suggest standardizing the data collection processes among our hotels and combine the results in a single data warehouse.

Secondly, we suggest collecting the following information from our guests:

- type of room
- group size
- average persons per room
- travels with children
- purpose of stay (holiday/business)
- uses room service
- voucher used
- uses free extra service (gym, pool)
- season

With these additional features we have a more detailed view of our customers.

8. CONCLUSIONS

To conclude it can be said that the advantages of the new customer segmentation outweigh the old approach. More features were taken into consideration, which provides a more meaningful understanding of each customer cluster.

Even though a minority of clusters could be slightly more meaningful, further filtering enables the marketing department to reach more specific customers. Since a customer segmentation is a static analysis, it is vital to renew this analysis constantly. However, it is not necessary to maintain the model itself.

As a final recommendation, the collection of more data, such as the type of the room or the purpose of stay would help to improve the model to obtain improved insights about the customers.

9. REFERENCES

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Data Mining: Concepts and Techniques. https://doi.org/10.1016/C2009-0-61819-5, page 89

Provost, F., & Fawcett, T. (2013). Data Science - What You Need to Know About Data Mining and Data-Analytic Thinking. Journal of Chemical Information and Modeling. https://doi.org/10.1017/CBO9781107415324.004, page 158

10. APPENDIX

Data Dictionary: A final data dicionary after feature engineering.

Columns	Type	Length	Description
ID	int	8	Costumer ID
Nationality	str	4	Nationality of the customer
Age	int	4	Age of the customer
DaysSinceCreation	int	4	Number of days since the customer was created
NameHash	varbinary	32	Hash of the customer name
DocIDHash	varbinary	32	Hash of the customer personal document identification
AverageLeadTime	int	4	Average number of days before arrival data the customer makes bookings
LodgingRevenue	float	8	Total amount of lodging revenue pais by the customer so far
OtherRevenue	float	8	Total amount of other revenue paid by the customer so far
BookingsCanceled	int	4	Number of bookings the customer made but subsequently canceled
BookingsNoShowed	int	4	Number of bookings the customer made but subsequently made a "no-show"
BookingsCheckedin	int	4	Number of bookings the customer made, which actually ended up staying
PersonNights	int	4	Total person/nights the customer has stayed at the hotel so far
RoomNights	int	4	Total of room/nights the customer has stayed at the hotel so far
DistributionChannel	str	30	Distribution channel normally used by the customer to make bookings at the hotel
MarketSegment	str	30	Current market segment of the customer
SRHighFloor	dummy	1	Indication if the customer usually asks for a room in a higher floor
SRLowFloor	dummy	1	Indication if the customer usually asks for a room in a lower floor
SRAccessibleRoom	dummy	1	Indication if the customer usually asks for an accessible room
SRMediumFloor	dummy	1	Indication if the customer usually asks for a room in a middle floor
SRBathtub	dummy	1	Indication if the customer usually asks for a room with a bathtub
SRShower	dummy	1	Indication if the customer usually asks for a room with a shower
SRCrib	dummy	1	Indication if the customer usually asks for a crib
SRKingSizeBed	dummy	1	Indication if the customer usually asks for a room with a king size bed
SRTwinBed	dummy	1	Indication if the customer usually asks for a room with a twin bed
SRNearElevator	dummy	1	Indication if the customer usually asks for a room near the elevator
SRAwayFromElevator	dummy	1	Indication if the customer usually asks for a room away from the elevator
SRNoAlcoholInMiniBar	dummy	1	Indication if the customer usually asks for a room with no alcohol in the mini bar
SRQuietRoom	dummy	1	Indication if the customer usually asks for a room away from the noise
PricePerNight	float	4	Price the customer spend per night for lodging
OtherRevenuePerPersonNight	float	8	Price the customer spend per night for other services (besides lodging)
PersonPerRoom	int	4	Number of people per room
GroupSize	int	4	Number of guests registered in the customer's name
AvgNights	int	4	Average number of nights spendind on the hotel by the customer
SRLocation	dummy	1	Indication if the customer usually asks for specific location on the hotel
SREquipment	dummy	1	Indication if the customer usually asks for specific equipment on the hotel
Nationality2	str	4	The most frequent nationality were grouped common to obtain a clearer view regarding nationality
DistributionChannelNew	str	30	The most frequent channel were grouped in order to reduce categories to get a better distribution