BUSINESS CASES WITH DATA SCIENCE

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

Business Case 2: Market Basket Analysis

Palm & Company

Pedro Santos (M20190420)

Ana Cláudia Alferes (M20190932)

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

This report details our several solutions for company C, who is struggling to maintain their growth and profits, through the use of restaurant data collected throughout the year, focusing on behavior and patterns of the customers. The process model for the development of this solution as well as the structure of the report is aligned with the CRISP-DM process model.

The report is composed by a Business Understanding chapter, where we explore the current business approach, its flaws and exemplify the benefits of a market basket analysis (MBA). Afterwards, we will explain the data and the exploratory process, including some key points on customer behavior. Based on our results, we present the MBA model as well as an interpretation of the results and their implications to the business. A prototype of a recommendation system was also created. Finally, we conclude by sharing theoretical considerations about deployment and maintenance plans and future model improvement.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND AND CURRENT SITUATION

One of the first brand restaurant of company C, specialized in Asian food, is currently struggling to maintain their profit margin and their continuous growth due to increasing competition in the market and possible behavior change of the customers. It is increasingly difficult to operate a business without taking full advantage of the data and, therefore, the lack of a good analysis to the customer behavior and needs leads to several problems such as outdated product offerings, that not correspond to the customer's wishes; non-differential approaches between dine-in customers and delivery customers; not being able to identify patterns in the customer consumption that can be used to boost sales and cross-selling.

After analyzing the main problems of the current strategy, we propose the use of a Market Basket Analysis model technique based on the association rules theory, which discovers patterns in regularities between products, i.e., creating rules to assess the frequency of purchasing combinations of products.

2.2. BENEFITS OF A MARKET BASKET ANALYSIS

Our proposed MBA solution has several benefits for the business. We will be able to create a set of menus with items that customers usually buy together. It is also possible to visually place frequently purchased together items close to each other and thereby increasing cross-selling possibilities of products. By having a better understanding of the data, we can help to differentiate the delivery and dine-in customers, as well as create targeted marketing campaigns and products.

Through this analysis, we can take other insights, which could help company C to introduce new products, understand substitute products and create a recommender system, which could boost sales by cross-selling.

2.3. OBJECTIVES & SUCCESS CRITERIA

The final objective is to make use of all the advantages from MBA. In order to look for popular and trending combinations of products, we will start with a deep exploration of data and then begin studying the association rules in the orders from the dine-in customers and delivery customers.

The success criteria for this project lies in finding relevant patterns from customers behaviors, so that company C can make more efficient decisions, increasing sales and improving their customer relationship management.

2.4. PROJECT PLAN

As mentioned above, we will follow the CRISP-DM process which consists of six different phases. After analyzing and understanding the business background, we performed a data exploration in order to get first, relevant insights. Following exploration, there is a preparation phase that consists of data cleaning and feature engineering, which result in a prepared dataset used for the following modelling phase. During all these processes, there were several interactions between the phases, so that we could improve our decisions of previous steps. Finally, after obtaining concrete results, we evaluate them, create a deployment plan und propose future improvements.

3. PREDICTIVE ANALYTICS PROCESS

3.1. DATA UNDERSTANDING

The goal of data understanding is to explore and explain the data based on statistical methods and the obtained knowledge in business understanding. Following the CRISP-DM process, this part contains a data collection and description report as well as a data exploration and quality report.

3.1.1. Data Collection and Description Report

Data Collection Report

The examined dataset encompasses transactional data about customer orders, made in 2018, of a restaurant based in Nicosia (Cyprus), which is owned by the company C. The variables display information about the products, quantities, prices, order dates, and customers. A customer can either be a customer, who eats inside the restaurant or order food for delivery.
Since public holidays can have a significant impact on the foodservice industry, the existing dataset is enriched with the date and the name of public holidays of Cyprus in the year 2018. This is a necessary step to get insights if holidays affect the sales of C.

Data Description Report

The raw dataset has 12 columns, which represent different features. The number of rows is 84.109 and each row represents one item ordered by a customer. Consequently, several rows can belong to one customer order. Most of the features are categorical variables (8) containing strings or are dummy variables. In general, more information, such as customer since and the city of the customer, are collected only for delivery customers.

3.1.2. Data Exploration and Quality Report

Within the data exploration, most of the analysis requires aggregated data on the order level. Furthermore, a comparison between restaurant and delivery orders is one of the approaches to obtain more insights. For this reason, the new dataset (including holidays) is grouped by the document number, which is one specific order and divided into two datasets: Restaurant and delivery.

Data Exploration Report

The exploratory analysis is separated into the subject areas of time, product and product family, revenue and customer figures, order history and the impact of public holidays.

Time exploration
The time analysis is divided into the hour, day and month regarding both channels, restaurant and delivery.

Within the distribution of the hour, we can see that we have typical a lunch and dinner time orders. In both, we have considerably more orders during dinner time.
In this visualization of the day, it is challenging to see useful patterns for further analysis. In contrast, the distribution of orders in the months of the year shows that in both channels, but especially in delivery, the summer is characterized by fewer orders.

Product families and products
This analysis examines the bestselling products and their related product families, again divided in restaurant and delivery.

In both, products and product family, we can see that starters are the most common product family in restaurant orders and deliveries. Some further similarities between the two channels are the popularity of rice, meat and sizzling. Regarding drinks and spirits, we can see, that these two product families usually are not ordered in deliveries. Furthermore, items of the product family "EXTRAS" are mostly ordered in deliveries. This may indicate that these customers usually want to personalize their meals more often.

Revenue and customer figures
The following figures summarize data regarding the number of persons per visit, how often they ordered, the average spending, the revenue and the number of orders.

Most of the dine-customers come into the restaurant either as a group of two or four. A group of three, five or six is common whereas single customers and groups more than seven are less often. The majority of delivery customers ordered just once (63.6 %). In contrast, only 2.03 % of the customers ordered more than 10 times. Both groups, 2 times or between 3 and 10 times is equally distributed (16.6 % and 17.8 %).

Customers, who eat in the restaurant spent on average more than twice as much as delivery customers (96.75 € to 45.05 €). However, in deliveries, we do not know how many persons ordered, so this may also be the reason for this difference. Like the revenue, the number of orders in the restaurant (56.3 %) is slightly higher than in delivery (43.7 %).

Order History
Looking solely at online customers, it is possible to observe that most of them have only ordered once, with around 63 % of the customers belonging to this bracket. Nevertheless, a respectable amount, approximately 37 % have returned to use the delivering service again. While looking at this returning customers, our goal is to understand if they have favorite products and if they repeat them often or they like to vary their orders and the results are fairly conclusive: at least 83 % of the customers likes to re-order their favorite/most frequent product in, at least, 80 % of their total orders. This could lead to the idea that online customers are not as comfortable trying

new items by themselves, which could be changed by introducing a recommendation system, which we will introduce in another section of this report. It is also possible to avoid customers' attraction by offering discounts if they re-order the same products or if their favorite products are on discount.

The impact of public holidays

During holiday the average total amount per customer is higher in both, delivery and restaurant. When it is a holiday, the average spending per customer is around 31 % in restaurant and 24 % in deliveries above the spending on a normal day.

Usually the customers use the delivery service more often on holidays. The average number of orders decreases by around 17 % in the restaurant, whereas the number of deliveries increases by about 26 %.

Data Quality Report

Overall, the given dataset has a good quality, however, some variables are formatted in the wrong data type. "Customer since" and "InvoiceDateHour" should be in the date format, whereas "TotalAmount" is normally a float data type. In "Locations" one can find some duplicates that are due to typos. The columns regarding the city and the first order ("CustomerSince") of a customer contain null values. This is because there is no information about these two variables for dine-in customers. After the split into delivery and dine-in customers, only the delivery customers encompass 2.106 (6.74 %) missing values, which stand for customers who ordered the first time.

The holiday dataset includes the dates of the clock change, which cannot be considered as a public holiday.

3.2. DATA PREPARATION

Data Cleaning

In general, keeping the main task of a market basket into consideration, this dataset does not require much data preparation. Most of the preparation is done during the exploratory analysis to obtain the right insights and visuals. For instance, to examine the time of the orders, it is necessary to change the invoice date into the date format and separate the hour, day and month.

In total, the above-mentioned points such as typos in location and the change of the data type are the main steps in data cleaning. The missing values in "CustomerSince" were not replaced because this variable is not meaningful for the approach of the market basket analysis.

Feature engineering

For the application of the Apriori algorithm, we decided to do it without further features. Since the goal was to do a market basket analysis on the existing products, it does not require any new features.

For the recommendation system, which is described later, three variables were binned to obtain less categorical values to get better conclusions. The binned variables are the time, holiday and person. The purpose of binning the time is to distinguish whether it is a dinner or lunch. The categorical variable holiday displays whether it is a public holiday (=1) or not (=0). The number of persons eating in the restaurant is grouped into 1-2, 3-5 and more than 6 to interpret the behaviors of different group sizes.

3.3. MODELING

3.3.1. Introduction of new products and menus

After running the *apriori* algorithm and analyzing the results, we decided that in order to increase customer satisfaction, some adjustments on the current restaurant menu are required. These adjustments include implementation of new products and promotion of cross-selling products through a creation of a set of pre-defined menus. We believe that with this approach, company C will gain competitive advantages over there competitors and keep up to date with their customer needs.

The two possible channels of consumption in the restaurant, there is dine-in and delivery, have their own characteristics and specific type of customers. So, to have a more detailed view, we analyze them separately.

In order to be able to create our menu suggestions, we used three specific rules: **support**, or the proportion of transactions of a specific product/product combination in the database, **confidence**, or the proportion of transactions that contain the product/product combination X and also contain the product Y, and the **lift**, or the ratio of the confidence of the rule and the expected confidence of the rule, meaning that if a transaction has a lift lower than 1, the products are rarely bought together and represent substitute goods, while if the ratio is bigger than 1, they represent complementary goods.

After the first analysis we found Mineral Water 1.5LT to be a product that has a big influence in our rules because it's a very generic product which can be combined with everything and it is requested very often. In order to have a more insightful view on the rules, we decided to remove this product from the analysis.

In a general overview, for both channels, after combining thresholds for two association rules, Confidence >= 0.5 and Lift >= 4, we reached the conclusion that the order **NO MEAT -> NOODLE WITH MEAT,** with Confidence =1 and Lift rounding 5, means that the restaurant does not have any vegetarian meal. Our suggestion for this problem is to introduce a new product that will be just **PLANE NOODLES,** so customers can make their own combinations with other products from the restaurant.

For dine-in meals with threshold for Confidence (>= 0.5) and Support (>= 0.2), we were able to find the order **SPRING ROLL -> EGG FRIED RICE** with Confidence=62% and Support=22%.

Another rule, that we found interesting to investigate, is **SWEET SOUR CHIKEN -> EGG FRIED RICE,** with Confidence=71% and Support=21%. Both dine-in rule orders do not present high values for complementary aspects (Lift = 1.5, approximately) but we believe that with the creation of a menu, the restaurant will increase the probability of selling these items together.

Our suggestion is to make a menu with **SPRING ROLL, EGG FRIED RICE** and **SWEET SOUR CHIKEN,** considering an option to order a Soft Drink for a minimal price increase.

In the delivery orders with threshold for Confidence (>= 0.5) and Lift (>= 4)

we could see the order **JIRA PULAU -> NAAN** with Confidence=50% and Lift =4.5, which has complementary effects. They are an Indian side dish and a starter, respectively, so just with these two items, it is not possible to consider them as a main dish. However, since these two items are the only Indian items present in our rules, we suggest the restaurant to invest in an Indian Menu with **JIRA PULAU** and **NAAN**.

Changing the threshold and metrics for Confidence (>= 0.5) and Support (>= 0.2), we come across one more cross-selling possibility, namely a **SWEET SOUR CHIKEN -> EGG FRIED RICE** menu**,** with Confidence=62% and Support=22%.

3.3.2. Recommendation System

As an alternative or even supplementary way of creating value from Company C's customer data and consequently improving the operations of the restaurant, we propose a real-time recommendation system for dine-in and delivery customers. The tool is aimed to recommend items and meals based on customer characteristics and their ordered items. From a business point of view this tool should increase customer satisfaction and simultaneously generating more revenue.

As Company C have two distinct distribution channels, with their very own specificities, we exemplify the use cases of the recommendation system for each of those separately. We start by explaining the main functionalities of the prototype to give an idea of how the system can be used. Based on that, we suggest a possible deployment strategy and future improvements.

In the restaurant context, the recommendation system needs the number of persons and the ordered items as an input. It automatically gets information about the time and possible holidays. Based on this information, the tool returns items and meals that have the highest lift with the ordered items.

The prototype for the delivery app considers the location, the time and a possible holiday as features to filter the database. Based on the items in the shopping cart, the user gets the items with the highest lift as a recommendation.

Please note that with the current prototype version, it is not assured that there are suitable recommendations for all possible feature combinations. This is a tradeoff that we must make, if the recommendations should have a certain level of validity (minimum support, confidence and lift) and additionally be targeted on specific customer segments. Once we implement our suggested deployment strategy and future improvements, this problem will continuously decrease with a higher amount of data.

4. DEPLOYMENT AND MAINTENANCE PLANS

Generally, we suggest starting with a specific use case and extend the system with further use cases in the future. This enables us a low-cost and quick way to create a proof of concept. One of these specific use cases in the restaurant context is to recommend a dessert based on starters and/or main dish.

From an operations point of view the system should be deployed as follows:

In the restaurant, the tool should be integrated with the POS-system and inform the waiter in real-time about possible recommendations for the customer. In the use case mentioned above the waiter would only need to enter the number of people into the system, in addition to the ordered items. Before he asks the guests for a dessert, he can consult the tool for recommended dessert for these guests.

For the delivery orders, the recommendation system should be integrated in the delivery app. The goal here is that the user gets recommendations right before the checkout. For the app we also would suggest recommending a dessert based on the other items in the shopping cart.

From a technical perspective the following steps are required:

It would be advisable to develop a data mart for each of the channels. Initially this data mart should be built with historical data from the last years. Based on that we could create a first production-ready model for each of the channel.

In the first year the data marts should be updated every month with the data from the POS-system and the app. In this period the model should be refined and updated in the same frequency.

After this period, the usefulness of the tool can be reassessed and, if beneficial, further use cases can be added to the system and the improvements described in the next section can be implemented.

5. FUTURE IMPROVEMENTS

Generally, we see a lot of potential in a more detailed customer segmentation. For this purpose, it would be advisable to collect more data about the customers. Features that can be collected from the waiter are e.g. number of children or gender of customers.

From a technical perspective the recommendation model can be supplemented with supervised learning techniques by collecting data about the performance of the system i.e. the acceptance of the recommendations. Furthermore, we could improve the model by including the order history of customers.

6. CONCLUSIONS

To conclude, we believe that an MBA approach joint with a recommender system is a good solution for Company C, which could help tackle the surging competition. The introduction of new menus and vegetarian options will surely attract new customers while the recommender system is able to help increase customers' spending. We also believe that this strategy will allow delivery customers to order more frequently with the possible introduction of tailor-made discounts and the menus and recommender systems mentioned above.

As for the future, the collection of more data will allow the recommender system to be more accurate and we will be able to observe different patterns of consumption.