MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 2: Market Basket Analysis

Jaime Duarte: 20220675

Shanjida Roman: 20221395

Yousef Ebrahimi: 20221382

Diogo Martins: 20221361

David Martins: 20221006

Group D – Data Vision Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

March, 2023

INDEX

1. EXECUTIVE SUMMARY....2
2. Introduction....3
3. Business understanding....3
   3.1 Business Objetives....3
   3.2 Business Success Criteria....4
   3.3 Situation Assessment....4
   3.4 Data Mining goals....4
4. Data understanding....5
5. Data preparation....7
6. Modeling....10
7. Evaluation....10
8. DEPLOYMENT AND MAINTENANCE PLANS....11
9. CONCLUSIONS....12
10. REFERENCES....13
11. APPENDIX (List of figures)....14

1. EXECUTIVE SUMMARY

Restaurant C is having trouble maintaining its profit margins and steady growth at one of its locations specializing in Asian food. To address this, Data Vision was hired to implement a data mining solution that will utilize the company's sales data to detect differences between dine-in and delivery customers, find patterns in consumption, and assess the adequacy of the current product offering.

The goal was to provide useful insights that will allow for changes in the company's practices, such as creating new menus and products, or promoting cross-selling.

The project focused on creating a market basket analysis using APRIORI algorithm and measures such as Support, Confidence, and Lift to uncover meaningful item relationships that will enable the restaurant to make impactful changes.

After analyzing the data, we found some interesting insights. For example, during certain summer months, the restaurant's revenue during lunchtime was very low, and further analysis revealed that the restaurant was likely losing money during these times. We recommend that the restaurant consider closing on certain days during these months to save on employee expenses. On the contrary, during winter, the restaurant can launch special campaigns, as this is the busiest, and most profitable time for the restaurant.

Additionally, we suggest several cost-saving menu recommendations and set menus for delivery orders. Finally, we provide recommendations for monitoring and maintaining the deployed algorithm, including regularly monitoring model performance, adjusting minimum support and confidence thresholds, and ensuring data quality. It may also be helpful to consider using different algorithms and updating the model as needed.

The project's success will be measured by the restaurant's ability to create new solutions, retaining, and acquiring customers, and increasing profit margins by following our recommendations.

2. INTRODUCTION

The restaurant chain C is a Cypriot company founded more than 20 years ago, with several locations across the country. Due to growing competition and shifting consumer preferences, one of C's first locations, specializing in Asian food, is finding it difficult to maintain its profit margin and steady growth. C has reached out to our team because it desires to make use of its sales data in an AI solution that will help them better understand the consumption and preference patterns of its customers in an effort to reverse this process.

We at Data Vision accepted this challenge and will implement a solution that looks to answer the company's key problems, those being: detecting differences between dine-inn and delivery customers, finding patterns in consumption, and assessing if the current product offering is adequate. The final goal of our analysis will be to provide useful insights that allow for a change in some of the company's choices, such as creating new menus and products, or promoting cross-selling.

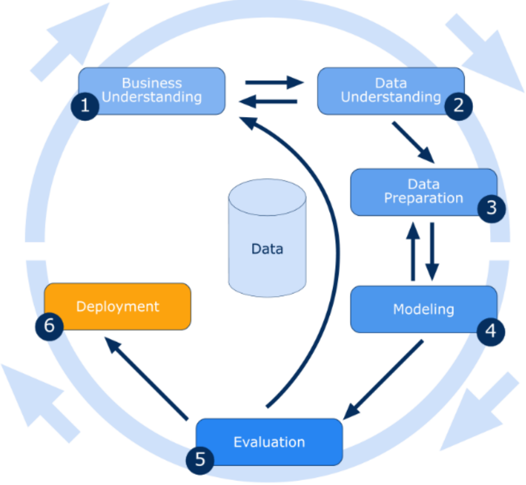


Figure 1: CRISP-DM methodology

We've chosen to use the CRISP-DM methodology to complete this project. In recent years, the most widely used methodology for data science has been the Cross Industry Standard Process for Data Mining (CRISP-DM), which is a process model that has become standardized across industries. To better represent the adopted process model and provide an easy understanding to company C of how exactly we conducted our analysis, we've structured this report to follow this same process.

3. BUSINESS UNDERSTANDING

The restaurant has customers from different locations in the island, mostly around Nicosia. The restaurant has a physical location, where many of its customers choose to eat, but also has a delivery service that distributes its products to addresses near Nicosia. The company keeps a good record of sales data which it is not currently being used optimally.

3.1 Business Objetives

The restaurant is looking to regain its ability to increase its profit margins as well as acquiring and retaining more customers. To achieve this, it's looking to gain a better understanding of the collected data by recognizing patterns in consumption and differences between dine-inn and delivery customers, as well as deciding if the current product offering is adequate. These insights should arm the company with the capacity to achieve its goal of making informed changes to its practices.

3.2 Business Success Criteria

From the business point of view, a successful endeavor would be one in which the company is able to come out with a clear outline to create its new solutions. This means we will look to provide a blueprint for the creation of new menus and products, the substitution of products, possible cross-selling opportunities.

These changes should lead to the betterment of the company's growth path, enhancing its capacity to retain and acquire customers and increase profits margins.

3.3 Situation Assessment

To our knowledge the company is not currently utilizing its data to create any AI based solutions to its problems, so a new model will be created from scratch. However, the company already keeps a good sales record, which will allow for a quality analysis. However, a good data preparation process will be needed in order to extract meaningful insights. Advice regarding future data collection practices will need to be provided, as it will improve future operations. Hardware and software assessments or recommendations will not be conducted, as these fall outside this project's scope.

The only meaningful constraint for the fulfillment of this project is time. We don't believe any restraints at the resource, legal or ethical level will arise. External data regarding Cypriot weather and holidays will be required for our model building, but we believe we'll be able to easily access it.

3.4 Data Mining goals

Our goal is to develop a quality market basket analysis for the restaurant. To do this, we will need to perform a thorough data exploration and preparation. Outcomes will be assessed using measures typically utilized in association rules learning Support, Confidence and Lift.

3.4.1 Support - This will show us how frequently a specific combination of products was ordered by customers. This also shows the percentage of transactions that combine both A and B. Support tells us how frequently an itemset was ordered. To calculate that, we divide the number of transactions that include A by the total number of transactions. Then, we do the same for B.

3.4.2 Confidence -Using confidence, we can be surer about the process of this association rule. Confidence presents us with the probability of the Consequents followed by antecedents. It also tells us how frequently A and B were ordered together for the number of times A were ordered. Hence, confidence is calculated with combined or individual transactions.

3.4.3 Lift - Lift is determined by dividing the observed frequency of two events occurring simultaneously by the expected frequency of those events, assuming that the two events are independent. When the lift value is greater than 1, it means that the two items are more likely than expected to occur together, whereas when it is lower than 1, it means that they are less likely.

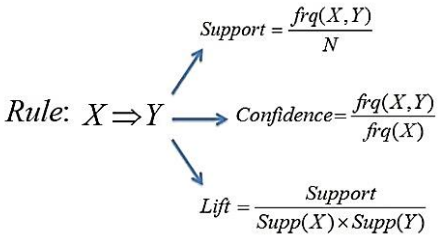

$Support = \frac{frq(X,Y)}{N}$
$Rule:\ X \Rightarrow Y$
$Confidence = \frac{frq(X,Y)}{frq(X)}$
$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$

Figure 2: Success measures

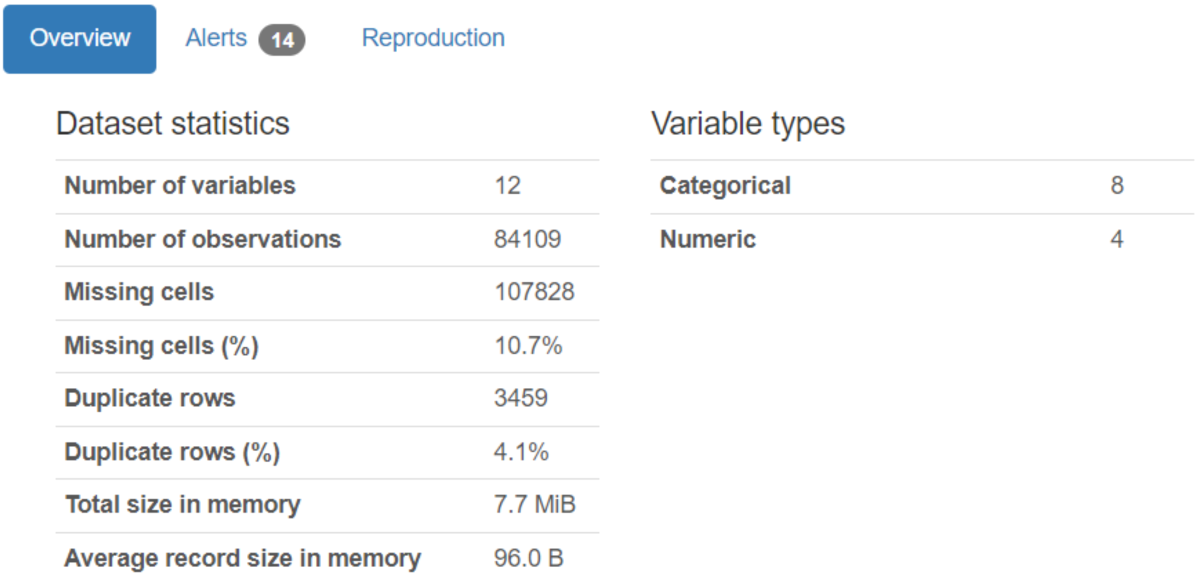
A relevant itemset for any time of business recommendation would need to meet certain thresholds regarding its values for these metrics. We will only take into consideration itemsets with support values over 0.06, confidence over 0.65 and lift over 1.2. This does not mean all itemsets that meet these thresholds will be used for recommendations, they the words case scenario, ideally better values will be achieved.

Our market basket analysis will hopefully uncover meaningful item relationships, that will enable us to provide data-based insights regarding customer habits and consumption patterns, allowing the restaurant to make impactful changes to their operation.

4. DATA UNDERSTANDING

This phase begins with data collection and proceeds with activities that will enable us to become acquainted with the data. This is done to identify any obvious problems and insights.

4.1 Collect Initial Data – The data we received is in csv file that contains 84109 observations and 12 features in total. Figure 3 shows the overview of collected data. At first sight, it seems to be quality data, although with some missing and duplicate values.



Dataset statistics

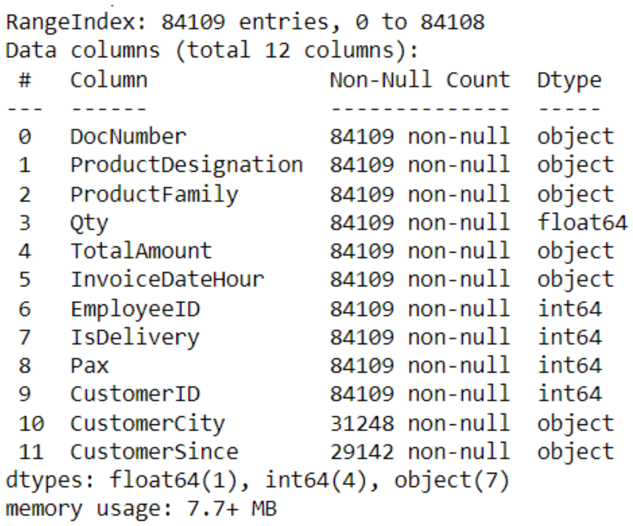
Number of variables	12
Number of observations	84109
Missing cells	107828
Missing cells (%)	10.7%
Duplicate rows	3459
Duplicate rows (%)	4.1%
Total size in memory	7.7 MiB
Average record size in memory	96.0 B

Variable types

Categorical	8
Numeric	4

Figure 3: Data overview

4.2 Describe Data – The dataset has numerical and categorical values with different datatypes such as int64, object, float64. 29 columns describe the nature of data that has been collected from existing customers. The restaurant provided us with meta data which describes each feature and what it means for the business. Only “CustomerSince” and “CustomerCity” have missing values. This tracks with our understanding of the city feature since it is usually only employed in delivery according to the restaurant. These led us to believe that “CustomerSince” is also being employed in delivery, since it has a similar number of missing values. This already tells us that assuming that a good record of customer is being kept, a majority of customers are choosing to dine-in.



```
RangeIndex: 84109 entries, 0 to 84108
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   DocNumber           84109 non-null  object
 1   ProductDesignation  84109 non-null  object
 2   ProductFamily       84109 non-null  object
 3   Qty                 84109 non-null  float64
 4   TotalAmount         84109 non-null  object
 5   InvoiceDateHour     84109 non-null  object
 6   EmployeeID          84109 non-null  int64
 7   IsDelivery          84109 non-null  int64
 8   Pax                 84109 non-null  int64
 9   CustomerID          84109 non-null  int64
 10  CustomerCity        31248 non-null  object
 11  CustomerSince       29142 non-null  object
dtypes: float64(1), int64(4), object(7)
memory usage: 7.7+ MB
```

Figure 4 : Data Information

4.3 Explore Data – We explored the dataset from restaurant C looking at both the raw data and some visualizations. We were able to see that there were 3923 duplicated rows in the dataset and proceeded to eliminate them. Since we have a feature that contains the date and hour of the invoice, we can rest assured that the likelihood of these actually being different orders and not just a record keeping mistake is close to zero.

We noticed some strange values in “ProductDesignation” and “CustomerCity”, with values we believed should represent the same concept, but are being assumed to be different due to small differences in spelling. Different spellings of “KTHMA GEROVASSILIOU WHITE” where corrected and the same was made for several values in “CustomerCity”. Regarding “ProductDesignation” we also saw 344 rows with the value “FOOD”, everyone of these had “SOUP” in the “ProductFamily” feature. This are clearly strange values that we decided to remove. We have no way of knowing which product they actually represent, since they also had different values in “Qty” (quantity ordered) and “TotalAmount” (total value payed by the customer), so we removed these 344 rows. “CustomerID” seems to be an incoherent feature since the vast majority of revenue is being attributed to ID 0. Record keeping regarding this feature is likely flawed.

Boxplots and histograms were created to better visualize the distribution of values for each of the numeric features, these are also important to address future removal of outliers with more accuracy. Where the visual representation pointed to the existence of outliers, we checked the rows. Outliers in “Pax” (with a value over 70) stood out as problematic, with clearly incoherent values, so they were removed. This process was repeated for the categorical variables as well, with several value counts being performed for the different features. We noticed that a majority of deliveries are being made to Egkomi, Strovolos, Niscosia and Lakatameia, with a significant drop off after that. These visualizations also allowed us to understand which products, days of the were the most sought after. Below is an example of one of these visualizations.

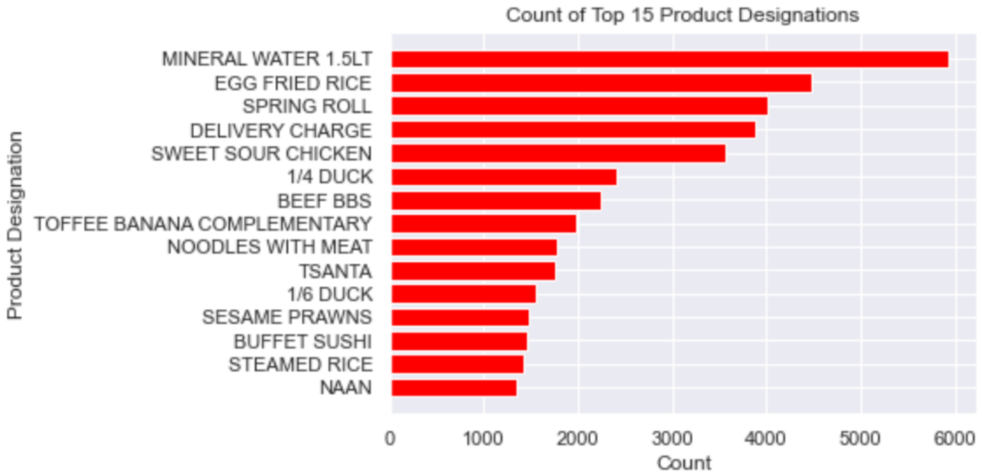


Figure 5: Most ordered products

4.4 Verify Data Quality – Having explored the dataset, we felt that a simple data preparation phase was necessary. The quality of data provided by the restaurant was of good enough quality to perform our analysis.

5. DATA PREPARATION

3.3.1 Clean Data – Data was cleaned before visualizations. This was done by deleting duplicate values, missing values, strange or incoherent values and outliers. To help identify outliers, boxplots were used. This allowed for a manual process of handling outliers. Other methods like using the Inter-Quartile Range were performed, but were not used do to removing more data

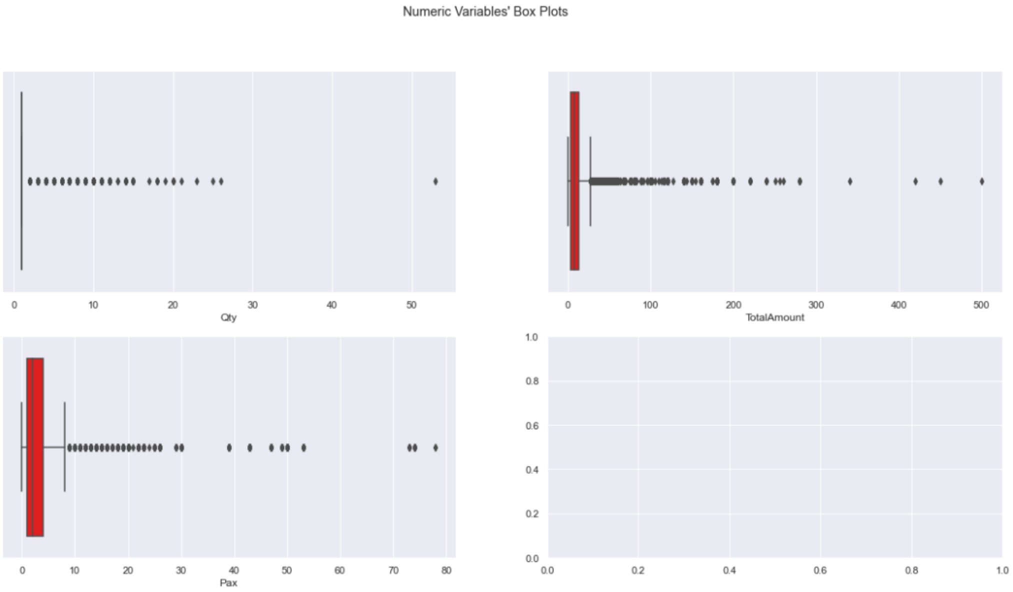


Figure 6: Metric data boxplots

3.3.2 Integrate Data – We felt we could make our analysis more trustworthy if we integrated some external data. We chose to load data regarding weather patterns in Cyprus in the year of 2018, as well as national holiday's information. We believe these can alter consumer behavior and incorporating them in our model will likely yield more reliable results. With this data we created some variables.

3.3.3 Format data/Feature engineering – Several features were created, with some only being used for intermediate steps. Using data from the external sources and the invoice feature, we created the following relevant features:

Features	**Description**
Date	Date of the invoice in a yyyy-mm-dd format
DayName	Day of the week in which invoice was created
MonthName	Month when invoice was created
Season	Season when invoice was created
Weather_HeavyRain (Boolean)	If there was a holiday that day (1 = yes)
Weather_Storm (Boolean)	If there was a holiday that day (1 = yes)
Holiday (Boolean)	If there was a holiday that day (1 = yes)
Weather_Dust (Boolean)	If there was widespread dust that day (1 = yes)
MealTime	Was the order at lunch, dinner, or other time
YearsAsCustomer	Years since creation of customer profile

Table 1: Created Features

These new features allowed for a better understanding of consumer behavior, so several visualizations we performed to aid our understanding and provide easily digestible information to the restaurant. Using "YearsAsCustomer" we can see that the biggest groups are customers with a profile less than a year old, followed by customers that began ordering over 5 years ago. Lunch service consistently brings in less revenue than dinner across all months. The restaurant served more customers in the Winter and Autumn, with Summer being the least popular. We were also able to see that, unexpectedly, the rate of orders by delivery barely increases in days with any of the adverse weather conditions, with thunderstorms being responsible for the biggest (yet still small) increase between the three. Holidays also do not provide the restaurant with a higher average revenue. We were also able to visualize which months were bringing in more revenue, as you can see in the graph bellow.

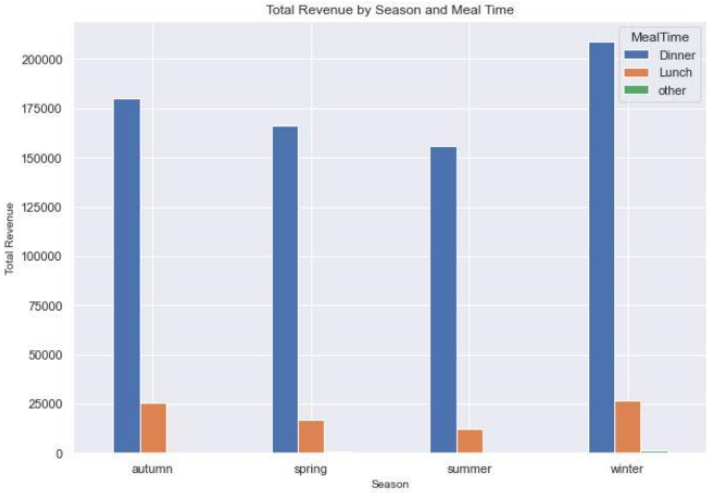


Figure 7: Total Revenue by season and mealtime

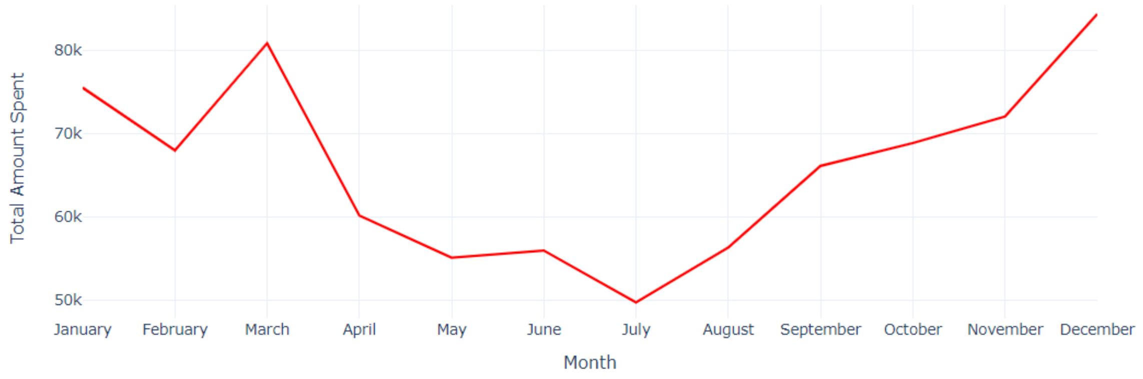


Figure 8: Total Revenue by Month

The low values we saw regarding revenue in some of the summer months (June, July, and August) led us to a deeper dive. Knowing "TotalAmount" values for Lunch were considerably lower for every month, as we've previously mentioned, we questioned how much revenue exactly was being brought in during lunch services on these months. A very obvious issue arises from these observations, which we will discuss further down in the deployment recommendations.

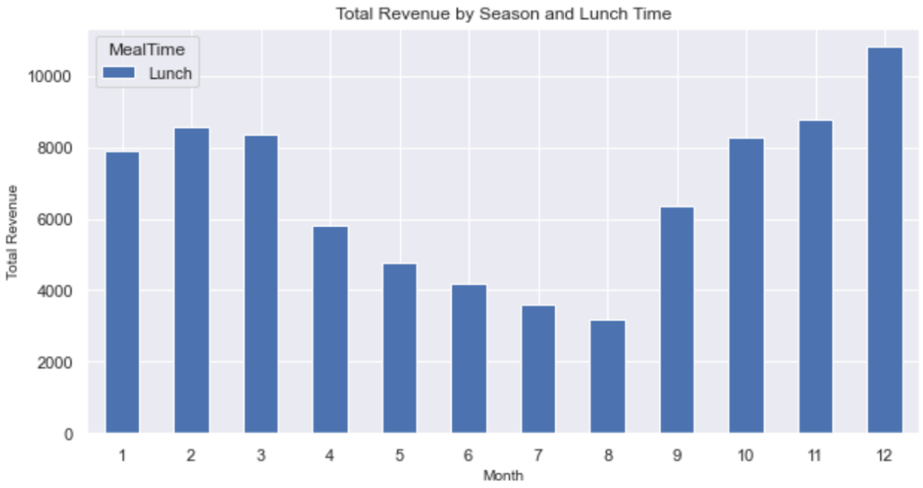


Figure 9: Total Revenue by Season (Lunch Time)

6. MODELING

6.1 Select Modeling Techniques – For this task we chose to use the Apriori algorithm. The Apriori algorithm is a popular algorithm used for association rule mining, which is a data mining technique that aims to discover relationships and patterns in large datasets, exactly what we are looking to achieve in this project. It is used to find frequent itemsets in a dataset, (a set of items that appear together in a transaction). The algorithm generates a set of candidate itemsets, prunes those that do not meet a minimum threshold and repeats the process until no more frequent itemsets can be found. It's efficient and can handle large datasets with many items and transactions. It works by generating candidate itemsets in a breadth-first manner, which avoids the need to generate all possible itemsets. It is also scalable, being able to analyze market basket data from a variety of sources. It produces results in the form of association rules, which are easy to understand and interpret.

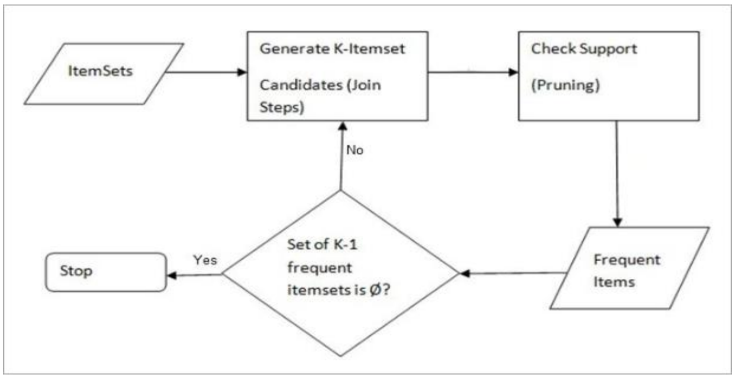


Figure 10: Flow chart of Apriori-algorithm

6.2 Assess Model – Our model seems to be fast and reliable. We analyzed its Support, Confidence and Lift scores, as well as some least common measures: Leverage and Conviction. We also performed a subjective assessment of how useful the itemset could be for eventual changes to the business operation. The itemset analysis was carried out for contexts, meaning we used several values of different features that we believe can make for relevant changes in consumer behavior (a specific season, weather condition, day of the week, etc.). Quality itemsets were produced, with several meeting our initial requirements of support values over 0.06, confidence over 0.65 and lift over 1.2.

7. EVALUATION

7.1 Evaluate Results - Evaluation allows us to check the original business objectives and all other findings. The result shows that the objectives of the business meet, and some good, targeted marketing strategies developed. Meaningful values for Support, Confidence and Lift were achieved. We believe these outcomes allowed us to provide the business with meaningful recommendations that will.

7.2 Next Steps & Considerations for model improvement - An effort should be made to store data with less inaccuracies. Same products should have the same names, this will allow for an easier upkeep of the data. From our understanding, customer ID is not being correctly reported, since a majority of income is coming from one customer ID. If a proper record is created it will allow, for example, to perform a quality customer segmentation process, if desired. The hotel can standardize the format of their data to ensure that it is consistent across all sources. This can help to reduce errors in the data and make it easier to analyze. By implementing these suggestions, we believe restaurant C will gain a more comprehensive view of their customers. This can help improve the accuracy of their analysis and inform future marketing strategies.

8. DEPLOYMENT AND MAINTENANCE PLANS

8.1 Plan Deployment – Deployment should be accompanied by development of business strategies that take advantage of our findings.

Seasonal recommendations:

1. As we've mentioned, the values for "TotalAmount" during Lunch time in some of the summer months (June, July, and August) is very low. By checking the average revenue brought in on each day of the week during these months, we find an even more disturbing practice. If we consider 2018's Cyprus minimum wage (just under 6€/h) and assume that 2 workers (cook and waiter) will work for 3 hours (very conservative estimations) in order to provide these lunch services, we can easily reach the conclusion that the restaurant is not only failing to profit, but very likely losing money during these days. As you can see in the table below, we have at least 4 services that fail to reach the 36€ those 2 employees would require. Knowing this, we recommend that the restaurant considers the following:

- **Strongly Recommend:** Closing on Monday in July; Tuesday in June and July; and Thursday in August

- **Moderately Recommend:** Closing on Monday in July; Tuesday in August; Wednesday in July and August; Thursday in June; and Friday in June

	June	July	August
Monday	N/A	25.9500	72.375
Tuesday	25.900	19.4250	47.175
Wednesday	78.500	47.2250	57.275
Thursday	50.475	139.7000	12.475
Friday	54.800	141.5250	170.825
Saturday	195.700	87.9125	81.725
Sunday	642.575	436.0500	355.850

Table 2: Average Revenue by Day of the week

Assuming the company follows our recommendations and taking into consideration the same conservative estimations regarding expenses with employees, the restaurant would be saving 30h of employee pay per week (around 700€ per month). Obviously, some profit is being gained with the limited amount of customers currently being saved, but the actual saving amount will never deviate much from our estimation.

2. During winter, the restaurant can launch a special campaign - when a customer orders any dish, they can be also offered the egg fried rice.

Cost Saving Menu Recommendations:

1. Take out meat from Noodle with meat.
2. Noodle with meat can be replaced with egg fried rice.

3. Noodle with meat can be replaced with noodle with spring roll without meat.

Overall recommendations:

1. Have a special menu for delivery orders.
2. For dine-in, increase the price of these dishes below and serve it with water:
   - TOFFEE BANANA COMPLEMENTARY, 1/4 DUCK, EGG FR...
   - SPRING ROLL, EGG FRIED RICE, BEEF BBS, SWEET ...
   - SPRING ROLL, 1/4 DUCK, SWEET SOUR CHICKEN
3. For deliveries: combine these dishes to create new set menus:
   - 1/4 DUCK, SWEET SOUR CHICKEN, EGG FRIED RICE
   - SWEET SOUR CHICKEN, EGG FRIED RICE
4. For delivery dinner time:
   - Offer the EGG FRIED RICE beside SWEET SOUR CHICKEN, 1/4 DUCK
5. For delivery lunch time:
   - Increase the price for the egg fried rice.

8.2 Plan Monitoring and maintenance

These are some of the recommended practices for a quality monitoring and maintenance of the deployed algorithm:

- Regularly monitor model performance: Monitoring the performance of the model on a regular basis ensures that it is still providing accurate results. This can be accomplished by contrasting the predictions of the model with actual sales data.
- Monitor changes in customer behavior: Keep an eye out for changes in customer behavior, such as new menu items, promotions, or seasonal changes, and update the model accordingly. This will help ensure that the model continues to provide relevant recommendations.
- Adjust minimum support and confidence thresholds: Depending on changes in customer behavior, you may need to adjust the minimum support and confidence thresholds used in the Apriori algorithm. This will help ensure that the model is still identifying relevant associations between items.
- Consider using different algorithms: While the Apriori algorithm is a popular choice for market basket analysis, there may be other algorithms that are better suited for your particular dataset. Consider exploring other algorithms and comparing their performance to the Apriori algorithm.
- Ensure data quality: Make sure that the data used to train the model is clean and accurate. This can be done by regularly auditing the data and correcting any errors or inconsistencies.
- Update the model as needed: Finally, be prepared to update the model as needed. As customer behavior changes and new data becomes available, you may need to retrain the model or adjust its parameters to ensure that it continues to provide accurate recommendations.

9. CONCLUSION

We embraced the challenge of trying to aid restaurant C in changing their trajectory. With our understanding of the data provided, and an intricate market basket analysis, we were able to find ways of providing insightful recommendations that will help the restaurant apply changes to their operations. We also provided strategies for monitoring and maintenance of our (and eventually other) algorithm. We believe the task was completed successfully and that the restaurant will find our work helpful. We look forward to having another opportunity of working with the company.

10.REFERENCES

[1] CRISP-DM help overview. (n.d.). Retrieved March 28, 2023, from https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

[2] Holidays and observances in cyprus in 2018. (n.d.). Retrieved March 28, 2023, from https://www.timeanddate.com/holidays/cyprus/2018

[3] Hotz, N. (2023, January 19). What is CRISP DM? Data Science Process Alliance. Retrieved March 28, 2023, from https://www.datascience-pm.com/crisp-dm-2/

[4] Kadlaskar, A. (2023, March 13). Market basket analysis: A comprehensive guide for businesses. Analytics Vidhya. Retrieved March 28, 2023, from https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/

[5] Li, S. (2017, September 27). A gentle introduction on Market Basket Analysis - Association rules. Medium. Retrieved March 28, 2023, from https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce

[6] Simplilearn. (2023, February 28). What is market basket analysis? overview, uses, types, and examples: Simplilearn. Simplilearn.com. Retrieved March 28, 2023, from https://www.simplilearn.com/what-is-market-basket-analysis-article

[7] Tymvou, Nicosia weather historystar_ratehome. Weather Underground. (n.d.). Retrieved March 28, 2023, from https://www.wunderground.com/history/monthly/cy/%CF%84%CF%8D%CE%BC%CE%B2%CE%BF%CF%85/LCEN

[8] YouTube. (2017). YouTube. Retrieved March 28, 2023, from https://www.youtube.com/watch?v=WGlMlS_Yydk.

[9] YouTube. (2019). YouTube. Retrieved March 28, 2023, from https://www.youtube.com/watch?v=guVvtZ7ZClw&t=313s.

[10] YouTube. (2020). YouTube. Retrieved March 28, 2023, from https://www.youtube.com/watch?v=43CMKRHdH30.

[11] YouTube. (2022). YouTube. Retrieved March 28, 2023, from https://www.youtube.com/watch?v=aqsa-gO_aq4.

[12] Cyprus - minimum wages 2023 (2023) countryeconomy.com. Retrieved March 29, 2023, from https://countryeconomy.com/national-minimum-wage/cyprus#:~:text=In%202023%2C%20the%20national%20minimum,account%2012%20payments%20per%20year.

11.APPENDIX (LIST OF FIGURES)

Figure 1: CRISP-DM methodology

Figure 2: Success measures

Figure 3: Data overview

Figure 4 : Data Information

Figure 5: Most ordered products

Figure 6: Metric data boxplots

Figure 7: Total Revenue by season and mealtime

Figure 8: Total Revenue by Month

Figure 9: Total Revenue by Season (Lunch Time)

Figure 10: Flow chart of Apriori-algorithm

Table 1: Created Features

Table 2: Average Revenue by Day of the week