The increase of online cosmetic shop sales using data mining techniques

Mukhamejan Karatayev
Department of Data Science
Nazarbayev University
Nur-Sultan, Kazakhstan
mukhamejan.karatayev@nu.edu.kz

Mukhit Yelemes
Department of Data Science
Nazarbayev University
Nur-Sultan, Kazakhstan
mukhit.yelemes@nu.edu.kz

Meiirgali Mussylmanbay
Department of Data Science
Nazarbayev University
Nur-Sultan, Kazakhstan
meiirgali.mussylmanbay@nu.edu.kz

***Abstract*—The product sales increase is one of the most important objectives of any company, as it is the major source of revenue. The development of modern data mining and big data techniques has helped to reach this goal by properly analyzing their customer data. Therefore, the startup company which sells cosmetic products through a E-commerce web platform decided to initiate a data mining project to encourage customers to buy more of their products. The goal of this research is to conduct the CRISP-DM process framework for this project. First, the optimal clustering algorithm was found, which segments customers into specific groups based on their behavior. Then, the product preferences of each customer were analyzed, and a recommendation system was proposed based on the analysis of the purchased combinations of products. Overall, the results demonstrate that the project justifies its cost by achieving data-mining and business goals.**

***Index Terms*—E-commerce, product sales increase, Cross-Industry Standard Process for Data Mining (CRISP-DM), business understanding phase, data mining**

BUSINESS UNDERSTANDING

I. Determine business goals

A. Business Background

First of all, we have to define the current business situation in our company which motivates this data mining project. We are dealing with a relatively young startup which is on the market for a couple of years. Figure 1 represents the organizational structure of the company, with two core roles - Chief Executive Officer (CEO) or founder and the Chief Technology Officer (CTO). They have two distinct areas of responsibility, the CEO is in charge of the Marketing and Sales departments, Human Resources, and Legal staff management. While CTO is responsible for the UX team, Backend and DevOps engineers, and Data Scientists.

The main objective of the company is to facilitate the purchase of cosmetic products. Also, their special feature is the precise product description, and even video tutorials on the use of some products. Also, they were able to engage a valuable amount of investments from outside that is going to be spent on the development of business objectives.

B. Current business problem

Through the analysis, the marketing and sales team found out that the majority of customers purchase only one cosmetic product. In addition, about 75% of the customers do not buy any product, rather they just view the product page or add it to their carts. Hence, they conducted several meetings where the possible causes and solutions for these problems were discussed. Obviously, it was decided to improve the quality of the web platform and app by improving user experience and fixing bugs, and implementing different marketing techniques. For those purposes, the usage of data mining for marketing purposes was proven to be very effective, so this work focuses on this particular part.

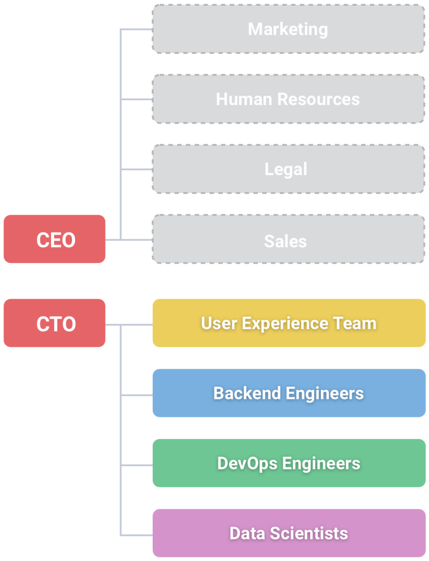


Fig. 1: Startup company organizational structure

C. Business goals and success criteria

To conclude, the overall business goal of this project is to increase the overall sales by 40% for the next year and encourage customers to buy several products at once by providing personalized offers depending on their preferences and behaviors. Apparently, the success criteria are the increase in sales, the increase of average check, also to gain insight about the users of the web platform will be very valuable knowledge. Therefore, the results of this data mining project should be assessed by the Sales and Marketing departments along with the CEO himself.

Personalized service in beauty industry has proven to be very important. For example, it was found that 77% of consumers have chosen, recommended, or paid more for a brand that support personalized services, thus increasing company's revenue [3]. Another example, analytical firm Quantzig in 2019 implemented an association rule learning model on data sources of European food retailers, to maintain a product-bundling recommendation system [1]. These bundling suggestions allowed to increase advertising returns to almost 300% [1].

II. Assess situation

A. Inventory of resources

As we said before, the company attracted investments and allocated $1 million for this project. Also, in terms of human resources, we have a team of data scientists who will be the main executors of the research. Definitely, the participation of those with expert knowledge of the business problem is compulsory. Hence, the DS team will be in tight cooperation with Sales and Marketing team members. Despite, the recent launch, we have already collected a lot of customer data to analyze. Which requires special software solutions and hardware with proper computational power. Another option is to use Amazon Web Services (AWS) which provides all tools for big data analysis. As the system is not extremely complex, open-source Python libraries and software will be enough to obtain successful results.

B. Requirements, assumptions, and constraints

The project has to be completed in six months, and within the allocated budget. However, one of the main dilemmas of such data mining projects is customer privacy, which involves the protection and handling of sensitive personal information of users. As we are going to use this information for customer segmentation and finding insights about their preferences and features. This part will require the Legal department to investigate security obligations, while the technical team will ensure privacy and safety during the mining process.

C. Risks and contingencies

To be prepared for any possible issues and be able to complete successfully the project within time and budget, we have to prepare a contingency plan.

The threat of losing valuable data is a very common situation, which can lead to serious problems, especially in data mining projects. Therefore, we should always have a backup of the data somewhere else. It might increase the overall cost of the project, but definitely worth it.

The hardware or software failure is a significant and popular contingency as well. As even new pieces of hardware have a chance to fail, the software may have bugs, which will increase the cost and shift all deadlines. Hence, all hardware and software should be certified and have a warranty, so they can be replaced in case of failures in a short period.

Despite the efforts, the data mining project may not have desired output. But, we can follow the CRISP-DM framework, and reiterate through the process cycle. It will increase the price and timeline of the project, however, our main goal is to achieve defined business goals.

D. Terminology

Below is the list of specific data mining and business terms that are relevant for our project.

Data mining terms are:

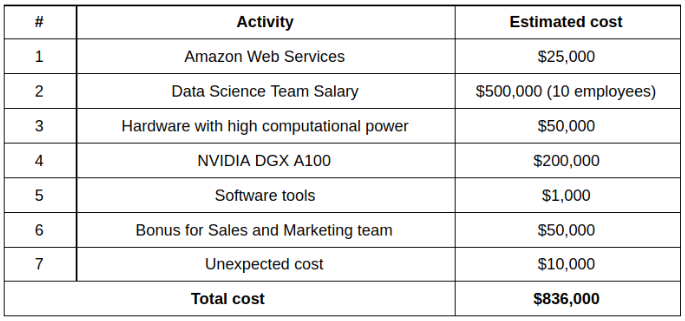
- **Clustering** - type of algorithms which finds group of items that are similar (in our case group of similar customers)
- **Model** - the form of the equation or a set of rules is learned by analyzing data
- **Sampling** - creating representative subset of data from the whole
- **Unsupervised Learning** - Machine learning methods which finds groups in data without defined the use of hand-written class variables

Business terms are:

- **E-commerce** - the means of selling products digitally on the internet
- **Conversion rate** - the amount of users who purchased the product divided by the total amount of users
- **Cookies** - text file sent to a customer's browser about how he interacts with the web platform
- **Email marketing** - engage customers with products and services of the company promoted via emails
- **Upselling** - the process of offering deals on similar products to promote customers to purchase additional products
- **Average check** - the total number of sales divided by the number of customers

E. Costs and benefits

To realize the necessity of the project for the company, we have to derive the possible cost and the benefits before the project is started. Below is the table of required activities and expected prices for each of the:



#	Activity	Estimated cost
1	Amazon Web Services	$25,000
2	Data Science Team Salary	$500,000 (10 employees)
3	Hardware with high computational power	$50,000
4	NVIDIA DGX A100	$200,000
5	Software tools	$1,000
6	Bonus for Sales and Marketing team	$50,000
7	Unexpected cost	$10,000
Total cost		**$836,000**

Fig. 2: Cost of the project

The AWS costs about $19,000-25,000 per terabyte annually, in our case, we will have to pay for 6 months. Additionally, due to the big amount of data, and probably complex model, we need the NVIDIA DGX A100 to train model in a short period of time. The DGX Stations costs about $200,000. The total cost of DS team salaries for 10 members, were estimated by using average data scientists salary ($100,000 per year) in the US. Also, we need 10 high-performance PCs which cost approximately $5,000 each. Also, we have to count the bonus payments for other team members, payments for software tools which are at most $1,000, and some unexpected expenses during the project implementation.

Meanwhile, the benefit can be calculated by the increase of revenues. According to the Sales department, last year, the total value of profit from sales reached $10 million. If we accomplish our business goal by increasing the sales by 40%, it will result in rise of profits by almost $4 million, compared to project cost of $836,000. As we can see, the expected benefit from applying the new marketing technique is valuable, which makes out project quite reasonable for the whole company.

III. Data mining goals

A. Data mining objective

The next important is to convert our business goal into the data mining goal. So to increase the overall sales, we can divide a customer into distinct subgroups based on their specific needs, their preferences, products that they purchased or visited, etc. Then, marketers can tailor personalized messaging and offer special deals based on those insights and the class of the customer. Data mining will make such segmentation meaningful, which eventually should increase the successful purchase rates.

The cluster analysis is the proper technique for these purposes. The Unsupervised methods such as K-means clustering will determine the groups of similar customers, without the labeled data. It means, that we do not have to spend time manually assigning classes to each customer by ourselves. There are many other clustering techniques, one of which is latent class cluster analysis (LCCA), which allows building models using non-numeric data. It was used by a Dallas-area analytical firm to cluster customers and was found to be beneficial [2].

The data to train the model will consist of all available customer information. Along with their activities and actions related to the services of the company. As it was said before, the privacy and security of personal data will be the main priority during the whole process. Commonly, such datasets will contain a lot of noises, missing and unnecessary data. Thus, the proper data preprocessing will be conducted, which will be described in more detail in the next phase of the CRISP-DM framework in the data understanding and description reports.

B. Data mining success criteria

The success criteria for the clustering methods are not trivial compared to other data mining techniques. In this case, the quality of the predicted clusters is evaluated based on some similarity and dissimilarity standards like the distance between cluster points. Certainly, customers with the same action patterns and characteristics should be in one cluster. The most popular evaluation metrics are Silhouette Coefficient (SC) and Dunn's Index (DI). The first one is bounded between -1(poor clustering) and 1 (dense clustering). While DI represents the minimum inter-cluster distance divided by the max cluster size, a higher score means accurate clustering. Therefore, our goal is to maximize SC and DI scores. However, compared to regression and classification algorithms there is no point to chase high scores. Also, the dominance of one model over the other is not apparent. The true success rate of the customer clustering can be obtained after the analysis of customer groups with the Sales and Marketing departments. As the output of the model will be highly subjective and depend on the company and its audience. For this reason, the model will be tested together with other departments, and learned insights will be discussed and assessed appropriately.

IV. Project plan

The project plan is one of the most significant deliverable of this phase, which will guide the whole process. Below is the list steps of our plan:

1) The first one is the business understanding phase and the deliverable is this report. Also during this step, all required resources and tools should be prepared. We have to meet with Sales and Marketing teams to discuss business objectives and success criteria. This step is the mandatory input for all succeeding steps. **(1 week for completion)**
2) The next is the data understanding step where we collect the customer's data, analyze the quality and find some initial insights, choose important attributes for clustering. The output is the data description report. **(3 weeks for completion)**
3) The output of the previous step will be input for the data preparation part, where we will clean, transform, analyze the data which will be used for modeling. Prepare our final customers dataset. This process might be performed multiple times depending on the model evaluation. **(4-5 weeks for completion could be longer)**
4) Using the prepared data from Step 3, we will test different clustering techniques such as K-means, DBSCAN, LCCA, etc. If necessary step back to the data preparation part to modify it for a particular model. **(4 weeks for completion)**
5) Having multiple customer segmentation models, we have evaluated them based on data mining metrics defined earlier. Also, conduct the meeting with Marketing department to assess whether business all goals are achieved. If results are unsatisfactory, go back to the first step, and repeat the whole process. **(2-6 weeks for completion depending on the results)**
6) Deploy the final model and add it to the web applications with the back-end team. So, the customers will receive personalized offers through emails and see them on their profiles. **(3 weeks for completion)**

A. Initial assessment of tools and techniques

At this point, we do not have the collected data, required hardware, and software, the overall system architecture. As we discussed earlier, we have a budget, so all necessary resources can be obtained later. While there is no shortage of human resources, we have high specialist DS team, and marketing and sales departments. And most important, the data collection part will be conducted in the next phase.

DATA UNDERSTANDING

I. Dataset Features

The dataset contains customer behaviour data for five months (from October 2019 to February 2020) from a medium cosmetic online store. It was collected and published by the Open CDP project from an anonymous company. A customer data platform (CDP) is software that integrates data from different sources to build the dataset of company's customers [4]. The main objective of the project is to provide open-source data from eCommerce projects for data science purposes. In addition, there are other datasets from an electronic store, jewelery store, and multi-category store. For this project, the data from the cosmetic eCommerce store will be used. It was published on the popular web-platform Kaggle in CSV format, and with proper description.

Each row of the dataset represents an event related to the products of the online store and its users. They can be described as the many-to-many relationship between commodities and customers. It contains 19,583,742 unique user event records, with 54,571 unique products, 525 unique categories, and 1,639,358 unique users [4].

Variables Description:

- **Event time** - Time when event happened at (in UTC)
- **Event type** - Type of the event: view (a user viewed a product), cart (a user added a product to shopping cart), remove from cart (a user removed a product from shopping cart), and purchase (a user purchased a product)
- **Product ID** - ID of a product
- **Category ID** - Product's category ID
- **Category code** - Product's category taxonomy (code name) if it was possible to make it. Usually present for meaningful categories and skipped for different kinds of accessories
- **Brand name** - Down-cased string of brand name. Can be missed
- **Price** - Float price of a product
- **User ID** - Permanent user ID
- **User session** - Temporary user's session ID. Same for each user's session. It is changed every time user come back to online store from a long pause.

II. Data Analysis

The general distribution of events by five months is quite similar. There was seen some fluctuation on the value of total events. Overall, there was steady growth by each month, with the exception of dramatic drop from November to December.

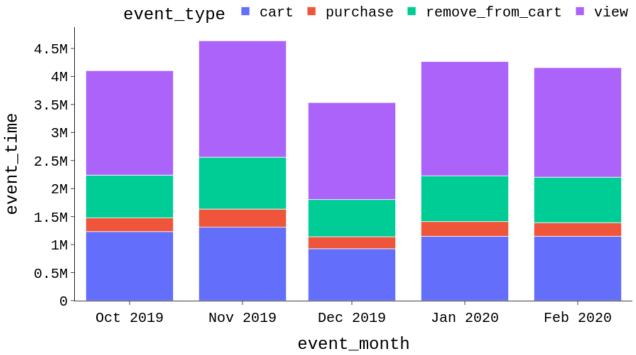


Fig. 3: Events distribution by month

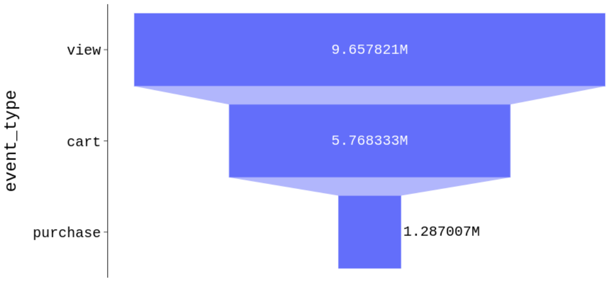


Fig. 4: Customer Funnel for Purchase Journey

The first business problem of our project was that very small proportion of clients actually proceed to the purchase process. Similar case could be observed in the Figure 4. The funnel of client's path from viewing the product to adding it to cart and then purchasing it narrows extremely. If at the beginning, there were close to 10 million views on products, at the end there happened only 1.3 million purchases. So, the filter ratio was the half and then one-fourth at each of the stages.

Our second business problem is low value of ticket size, which indicates that clients tend to purchase products of lower price category and/or few products at once. From the Figure 5, there could be observed the distribution of number of purchases and ticket sizes for the customers in the dataset. For the number of purchases, the median value is 6, mean value is 11.6. As regard to the ticket size, the median value is around 33 dollars and mean is 57 dollars.

Problems:

The outlier analysis revealed nominal negative price irregularities in the "Price" attribute, this is assumed to be caused due to human error. "Price" attribute is also assumed to be the selling price of the product. Looking into the structure of the data, revealed some irregularities with "Category code" and "Brand". Since more than 50% of "Brand", 98% of "Category code" and less than 1% of "User sessions" observations were

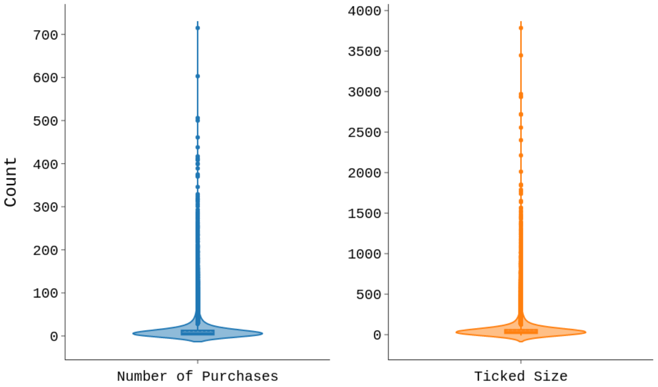


Fig. 5: Summary statistics for number of purchases and ticket size of purchases

Attributes	Missing values
event_time	0
event_type	0
product_id	0
category_id	0
category_code	20339246
brand	8757117
price	0
user_id	0
user_session	4598
event_month	0

Fig. 6: The number of missing values in the dataset

not labelled. The number of missing values shown in the Figure 6. These attributes will not be used in the analysis.

DATA PREPARATION

The first step of the data preparation was merging data sets for five months together and adding the month as the new attribute. With that, there was 20,692,840 rows in the data set overall. Secondly, there was done removal of duplicate rows from data set. Next, the entries with ”purchase” event type were selected, as only this type of action is necessary for future models.

As there was going to be two group of models based on the data types, at this stage the data preparation went in two directions. For the users model, ’event_time’, ’event_type’, ’price’ and ’month’ columns and for the products model, ’user_id’, ’user_session’ and ’product_id’ columns were left.

For the final stage, with each of the approaches, outlier removal and normalization of features between 0 and 1 were performed.

MODELING

As we said before, our data mining goal is to segment customers based on their event history. For this reason, different clustering algorithms such as K-means, Density-based spatial clustering of applications with noise (DBSCAN), and BIRCH (balanced iterative reducing and clustering using hierarchies) were implemented and compared.

After the data preparation phase, the dataset contained 908,776 rows and 5 columns (’event_time’, ’event_type’, ’price’, ’user_id’, ’month’). Regarding the quality of the data, as a result of the data cleaning step, there were no missing data or duplicates. As we can see, to work with this amount of data, a PC with large enough RAM and a high-performance CPU was required. Because of the sufficient budget, all required computational resources were provided in time.

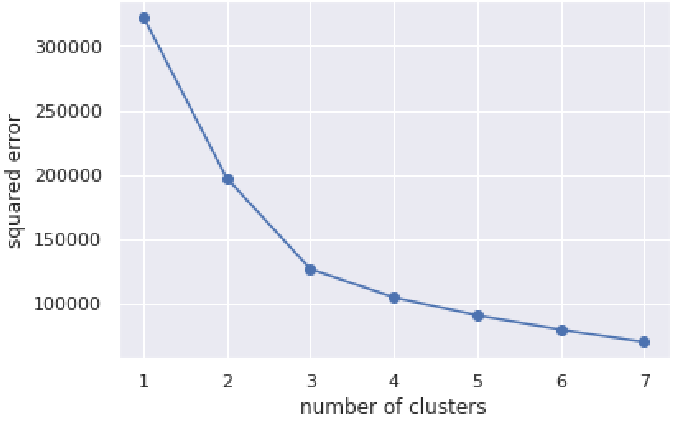


Fig. 7: Graph of the Elbow method

First of all, each customer’s behavior should be described by some quantitative metric. One of the most simple but very effective approaches is the RFM analysis. RFM stands for recency, frequency, and monetary value, each corresponding to some key customer trait. The recency value represents the freshness of the customer activity (e.g. the time since the last user’s activity with the product). This dataset has a value from 0 for February 2020 (latest records) to 4 for October 2019 (earliest records). While the frequency represents how often the customer performs those activities, in our case the number of purchase events for each user. Lastly, the monetary value illustrates the purchasing power of the customer, for our data it is the sum of product prices for each individual. These metrics can effectively describe the behavior of the customer and derive the customer’s lifetime value. After the calculation of RFM scores for each person, the outliers were removed. Finally, to obtain proper clusters, the user id column was dropped and data was standardized with the special function from sklearn library.

The combination of the RFM model with the K-means clustering algorithm was successfully used for customer segmentation problems [5]. K-means clustering is one of the most popular unsupervised machine learning algorithms. Unsupervised methods make inferences from data utilizing only input vectors without using labeled outcomes, as in our problem. After initializing the desired number of clusters or number of centroids representing the center of the cluster, K-means algorithm iteratively identifies the best possible position of

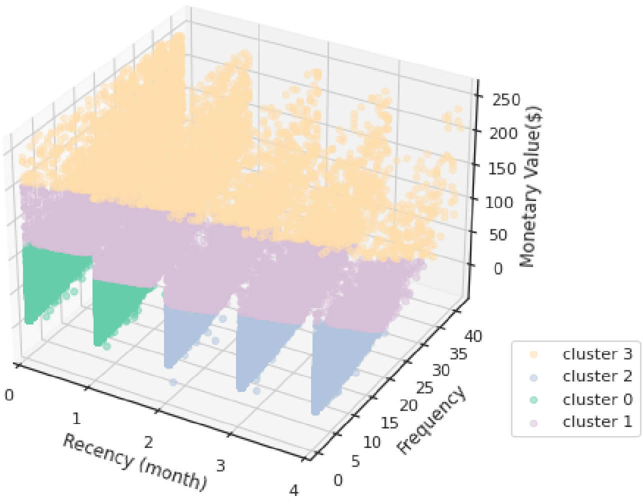


Fig. 8: 3D scatter plot of K-means clustering algorithm

the centroids. While allocating every data point to the nearest cluster center and keeping the centroids as small as possible.

Therefore, using the elbow method we have to identify an optimal number of clusters 7. We computed the Error Sum of Squares (SSE) which shows how well data is segmented, for different numbers of clusters. According to Figure 7, the most optimal k-value is equal to four. Then, the data with RFM scores was segmented into four clusters with the K-means algorithm. To visualize the output of the model, we derived the 3D graph with the following axises Recency (month), Frequency, and Monetary Value. As we can see from Figure 8, four customer groups are well separated and we can assign each cluster to some defined group. After the analysis of this result with the marketing and sales team, we decided to assign these clusters to the following classes:

- **New customers** - customers that started shopping very recently and as a result, they didn't make purchases often nor spend much money
- **Potential Loyalist** - the group that shops quite often and spent a reasonable amount of money, but not as much as loyal customers
- **At-Risk** - customers that have high recency (last purchase was a long time ago), low frequency, and monetary values
- **Loyal Customers** - the group of customers that purchase regularly and spent the highest amount of money, although they might have different recency ranging from October 2019 to February 2020

According to Figure 9 the biggest group is the at-risk customers, while the smallest is the loyal customers. Also, it was found that potential loyalist customers spend on average $72, which is three times higher than new and at-risk customers with an average monetary value of $24. The loyal customers spend on average $146 and have median recency of only one month.

Next, we used the same data as an input for another popular unsupervised clustering method called DBSCAN. It is the density-based algorithm that works by identifying dense regions in space separated by less dense regions. First of all, we have to define two hyperparameters: epsilon (maximum distance between two samples for one to be considered as in the neighborhood of the other) and minimum number of samples (minimum number of points present in the neighborhood required to form a cluster). Unlike the K-means algorithm, DBSCAN does not require the initialization of the number of clusters, rather it is found by the model automatically. After multiple experiments, it was found that eps = 0.5 and min_samples = 6000 are the most optimal values. With these parameters, the DBSCAN algorithm segments customers into six groups. We plotted the same 3D graph that was shown above for the K-means algorithm. According to Figure 10a, the classes are separated mostly by the recency value ('cluster 0', 'cluster 1', 'cluster 2', 'cluster 3', 'cluster 4'). The rest of the data is classified as noise ('cluster -1'). As you can see, it is hard to define these clusters in terms of business view. Therefore, the DBSCAN algorithm cannot be used for the purposes of this project.

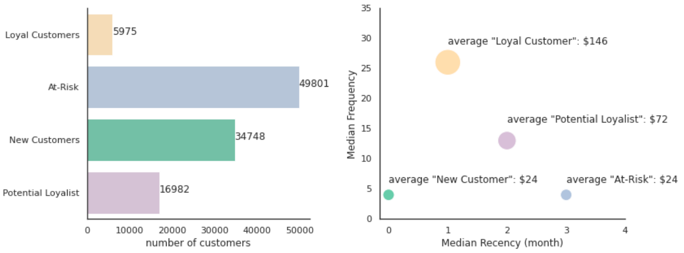
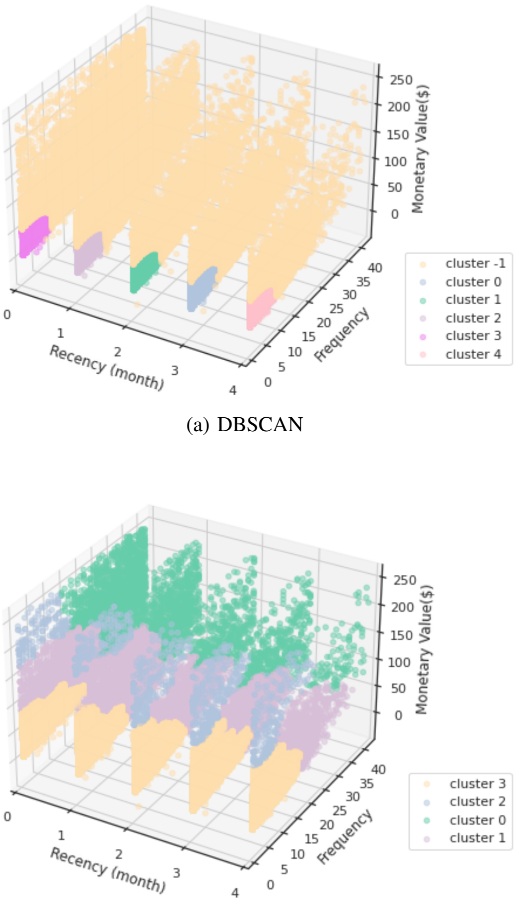


Fig. 9: Groups description of K-means clustering algorithm

The third method that was tested is the BIRCH clustering algorithm. It is a memory-efficient, online-learning algorithm invented as an alternative to the MiniBatchKMeans approach. BIRCH constructs a tree data structure with the cluster centroids being read off the leaf. As well as the K-means method, it requires the number of clusters to be initially defined. Therefore, we used the same number of customer groups, as it was proven to be the most effective and explainable. The same data was segmented by BIRCH into four clusters. Figure 10b illustrates that customers are divided accurately. Also, we can see that groups are not split by only one attribute, like with the DBSCAN algorithm. Therefore, we can characterize obtained clusters in the same way as in the K-means case. However, the distribution of customers by class shows that they are not representative. According to Figure 11, new customers and at-risk customers have the same recency median. All four groups contain customers with all possible recency values, which does converge with our descriptions of the clusters.

In order to compare and evaluate these clustering algorithms, we calculated the Silhouette Coefficient and Davies Bouldin score. The Silhouette Coefficient illustrates how similar a data point is to its own cluster (cohesion) compared to other clusters (separation). The best value is 1 and the



(a) DBSCAN

(b) BIRCH

Fig. 10: 3D scatter plot of DBSCAN and BIRCH clustering algorithm

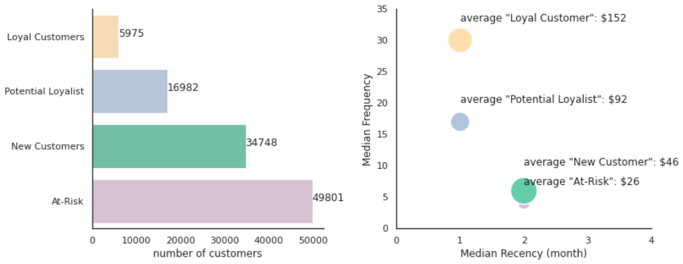
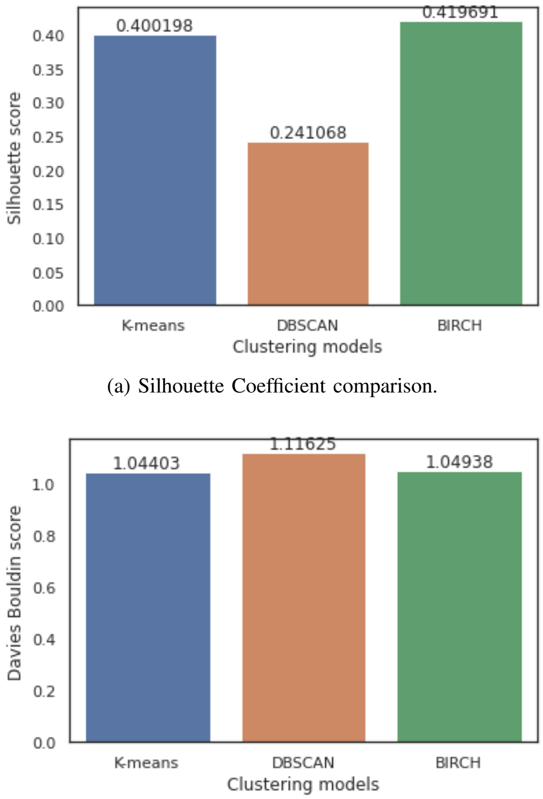


Fig. 11: Groups description of BIRCH clustering algorithm

worst value is -1, a score near 0 expresses overlapping clusters. The Davies-Bouldin score is defined as the average similarity measure of each cluster with its most similar cluster. The minimum score is zero, with lower values indicating better clustering. The bar chart in Figure 12a shows the Silhouette Coefficient scores for all three clustering algorithms. The results demonstrate that K-means and BIRCH have similar performance, while the DBSCAN has a significantly lower score that indicates poor customer segmentation. However, according to the comparison of the Davies-Bouldin scores in Figure 12b, the difference in performance is not noticeable. Nevertheless, the DBSCAN model has the worst score. Besides these evaluation metrics, to meet all business goals it is important to obtain customer groups based on their actions and behavior. Our marketing and sales department should be able to understand the results of the model in order to successfully apply this valuable information for finding the proper advertisement approach for each customer group. Which will in turn increase the sales of the products and overall profit of the company. Analyzing the K-means and BIRCH with marketing and sales teams, we decided to use the K-means algorithm. This model segments customers into logically understandable and reasonable groups.



(a) Silhouette Coefficient comparison.

(b) Davies Bouldin score comparison.

Fig. 12: Clustering algorithms performance comparison.

Furthermore, in order to increase product sales, we have to understand the preferences of each customer. Therefore, the list of unique products that were purchased during the whole period was derived for each customer. Our data contains user id, user session, and product id, so grouping by first two columns we obtained the list of unique products, for each user. From this information, we can understand preferred combinations of products. In combination with the customer segmentation algorithm, we can use a specific marketing strategy for each customer group and even determine the individual approach for each customer in this group. Therefore, it may significantly increase the customer's interest in our offers and deals.

We tried to find even deeper insights from our dataset, by identifying which product combinations are purchased more frequently. For example, if the customer adds the product to the cart, our algorithm will recommend a list of products that have been regularly purchased with this particular product. From the previous analysis, we obtained the set of purchased products during the unique user session. Then we count the number of occurrences of each unique combination of products. Dropping single products and combinations that occurred only once, the probability of each set of products was calculated. The obtained data contained 1,281 unique product combinations. With this information, we can recommend the list of products with the highest purchase probability for a chosen commodity. A simple python script was prepared to demonstrate this method in action. It takes the data source and product id as the input, and outputs the ten most popular recommendations. All these three algorithms will help the company to increase overall sales.

EVALUATION

As noted in the previous section, the models will allow us to determine the categories of consumers, with which our marketing workers will work to increase the number of consumers and sales, respectively. Thus, we can declare that the business goal set by us will certainly be achieved. According to model evaluation results, our clustering algorithm meets our data mining goals, obtaining well-divided customer groups and, demonstrating acceptable Silhouette Coefficient and Davies Bouldin scores. In addition, analyzing the customer's preferences and establishing an efficient recommendation system provided the essential insights for a successful marketing campaign. As we have all required computational and human resources, sufficient data management infrastructure, the deployment of these models will not cause any problems.

As a result of reviewing the process of the inital data mining project, our ”e-shop” developed a greater appreciation of the interrelations between steps in the process. Going through all the stages of CRISP-DM, we see that the cyclical nature of the process increases its power. The process review has also led our team to understand that:

- When something unusual appears in another phase of the CRISP-DM process, a return to the exploration process is always warranted.
- Data preparation requires experience and skill, as it is necessary to choose the right attributes for data analysis.
- It is critical to stay focused on the business problem at hand, because once the data are ready for analysis, it's all too easy to start constructing models without regard to the bigger picture.
- Once the modeling phase is over, business understanding is even more important in deciding how to implement results and determine what further studies are warranted.

Probably, one of the key activities that were missed is the testing of our models on real customers. We could select a small target group from our customers, then apply our clustering model, and by identifying their preferences try to suggest them special offers. After this process, we would assess the effectiveness of our models in a real-world situation. Other than that, the data mining process went smoothly without any serious issues. In the beginning, we thought that the clustering model will be enough to meet all business goals, however, during the development process, we found that we need more tools. Therefore, customer preferences analysis and recommendation system were implemented as well.

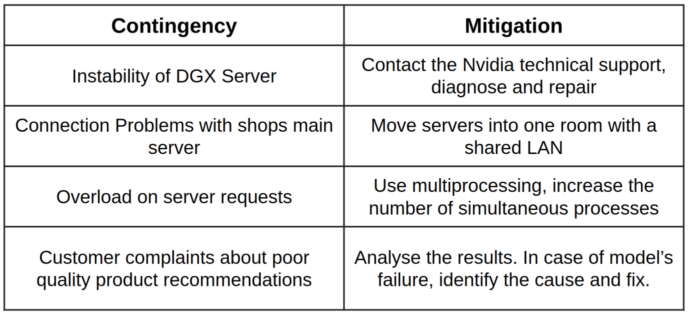
All over, our team reasonably confident in the accuracy and relevance of the project results, so we are proceeding to the deployment phase.

DEPLOYMENT

As a result of the modeling phase, we have a K-means clustering model, and functions that determine a customer's product preference list, and recommend a list of products that can be purchased with a given product. Next, the deployment plan for our models is the following:

1) The first step is to think over the online cosmetic shop's business logic with addition of new models to it. The data workflow will be between the website's server and Nvidia DGX server where models will be set. Current services will be updated based on that new workflow.
2) There have to be made the updates to the website's interface, in order to properly display the new feature introduced by this project, that is a recommendation of products based on the taste of client. It should be executed by UI/UX specialists.
3) The server session must be set on DGX workstation, that will handle the incoming requests, apply models to compute the output and execute the response in real time.

As regard to the possible problems and complications, there was prepared a contingency plan that will help to solve the issues and mitigate the negative consequences. The detailed contingency plan is provided below:



Contingency	Mitigation
Instability of DGX Server	Contact the Nvidia technical support, diagnose and repair
Connection Problems with shops main server	Move servers into one room with a shared LAN
Overload on server requests	Use multiprocessing, increase the number of simultaneous processes
Customer complaints about poor quality product recommendations	Analyse the results. In case of model's failure, identify the cause and fix.

Fig. 13: Contingency plan

In order to provide the maintenance for new system, there have to be made several additional actions. First of all, the new service usage statistics will be collected. This includes both the internal data, like the system load, performance, and external data, like the client's feedback on new features, the quality assessment of recommendations. All of them will be useful, as it will allow the company to keep the situation under control and immediately react in case of any indicators falling. Furthermore, due to the accumulation of store's sale history, every 6 month the model will be tuned with addition of new data entries. Because of this, the model's predictions will be kept up-to-date with possible changes in the trends.

To sum up, the original business problem was low average check and small purchase-to-view rates on online cosmetic shop. This problems were found by analysis of store sales information. There did not happened the deviations from the original project plan and the final project cost met the expectation of $836,000. The evaluation phase showed that the modelling reached the expected data mining goals, and both clients segmentation and product recommendation models are suitable for their application tasks. As regard to deployment, the overall process in general is to set the server with resultant models on the DGX workstation and completely integrate the system to online store's workflow. For the future project in similar directions, there could be recommended to get acquainted with the methods used in this project and possibly search for even wider datasets in terms of scale, as it may help to achieve even more general and universal models as a result.

References

[1] Specialty Foods Retailer Achieves 3.5x Increase in ROAS with The Help of Market Basket Analysis. (2019, December). Retrieved October 25, 2021 from https://www.quantzig.com/

[2] S. Gossett. (2021, September). Use Data Mining to Predict If Your Product Will Crash and Burn. Retrieved October 25, 2021 from https://builtin.com/big-data/data-mining-marketing

[3] S. Broyd. (2020, November). Transforming the Beauty Industry with Data and Analytics. Retrieved October 25, 2021 from https://clarkstonconsulting.com/insights/beauty-industry-transformation/

[4] M. Kechinov. (2019, November). eCommerce Events History in Cosmetics Shop, Version 6. Retrieved November 5, 2021 from https://www.kaggle.com/mkechinov/ecommerce-events-history-in-cosmetics-shop/metadata.

[5] C. Daqing. (2012, November). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Database Marketing & Customer Strategy Management 19, 197-208.