**NATIONAL RESEARCH UNIVERSITY**

**HIGHER SCHOOL OF ECONOMICS**

Faculty of Computer Science

Bachelor’s Programme “HSE and University of London Double Degree Programme in Data Science and Business Analytics”

**Research project report**

on the topic: "Optimization of client lists for communication by means of mathematical modeling."

(the first stage)

**Student:**

group БПАД182 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Sign Name, Surname, Patronymic

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date

**Supervisor**:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name, Surname, Patronymic

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Position

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Company

Date \_\_\_\_\_\_\_\_\_2020 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Grade Sign

**Moscow 2020**

**Content:**

**1. Introduction**

1.1 Abstract;

1.2 The role of members of the research project;

1.3 Instruments;

1.4 Main result;

1.5 Next steps;

**2. Review and comparative analysis**

2.1 Review of abstracts which you used for project and review of analogues;

**3. Selection of methods, algorithms, and models for project implementation**

* Information about filtering of data;
* Creating pivot table;
* Analyze pivot tables;
* Creating an analytical base table

**4. List of sources**

**Introduction:**

* 1. **Abstract:**

Nowadays one of the most popular areas of predictive analytics is the search for an optimal offer for a client or Next Best Offer. Our project is dedicated to customer analytics by unfiltered cutting data received from kid’s online store given by SAS INSTITUTE. In order to form the most suitable proposal, it is necessary to do a number of works: building analytical clustering (lifestyle - segmentation) using machine learning, identifying customer profiles based on purchases, the formation of a strategy to maintain the existing customer base through effectively built communications with customers for each segment, the choice of the best personal offer for each profile (Next Best Offer). Within the project it is necessary to allocate segments of clients which with the greatest probability will make the order in a category using results of the constructed mathematical models. Also, the project considers modern methods for detecting groups of goods characterized by mutual influence on each other in terms of demand volume (Market Basket Analysis), using Machine Learning methods.

* 1. **The role of members of the research project:**

Another participant of the project besides me is Demchenko Karina. Both students from the program of Faculty of computer science, Bachelor’s Programme ‘HSE and University of London Double Degree Programme in Data Science and Business Analytics’, 2 course. The work of Karina was dedicated to filter (‘clean’) and analyze data that she get from our database and sort the data so that only the orders that were bought out are left in the table. While I was filtering and processing all the orders placed, cleaning up the wrong positions and analyzing the placed purchases. Then we divided the work on building pivot tables according to the obtained data. We needed to get pivot tables by purchased and placed orders. My job was to build tables by categories and subcategories of goods and time. Karina did the work on the other tables - days of the week, months, regions and others. Next, Karina combined the resulting table and made a third type of table called " Buyout". Then, my work was to make visualization (histograms, charts) on the resulting tables. The next step was to analyze the resulting summary tables and visualizations and to propose marketing hypotheses for products. We divided this work: Karina carried out the analysis for "Regions", "Categories of goods (Group 2 and Group 3)" summary tables. I analyzed data on the tables for "Days of the week", "Hours of shopping", "Months". Next, we began to form the analytical base table. Karina was collecting data for the purchased goods, and I was collecting data for the placed goods. Then we combined the two received tables and got the analytical base table. Then, Karina did Analytical Base table and gave it to me& Thus, I could try to implement the algorithm of k-means.

* 1. **Instruments:**

1. The unfiltered data for our project was given to us by our mentor <https://drive.google.com/open?id=1-OTO8kN5qdbMVQuNF_eRlLQ9RezcrOdc>
2. Python, Anaconda, jupyter notebook
3. Python libraries: numpy, pandas, matplotlib.pyplot, sklearn.cluster, sklearn.decomposition
4. Microsoft Excel for Mac, version 16.33
5. Google documents
6. Messages with mentor by Skype

**1.4 Main result:**

The main result of the project is an attempt to find a solution to an actual applied data analysis task. This project proposes to develop an approach using machine learning methods based on real data. We should develop a full system for segmentation of client base and definition of clients with the greatest probability of purchase in a category, create its description for business. Another purpose is selection of the best personal offer for each profile (Next Best Offer).

* 1. **Next steps:**

- Segmentation of the customer base and development of a strategy for segment development

- Building a joint purchase model based on customer transaction data (Market Basket Analysis)

- Identification of rules for general purchases by customer product groups.

- Evaluation of the effect of the proposed approach

- Constructing life-style segmentation

- Calculation of financial and economic substantiation of the project

**2. Review and comparative analysis**

**2.1 Review of abstracts which you used for project and review of analogues:**

The project is dedicated to learning how to filter datasets and implement got data in the next purposes. There is an analogue to cope with the task. It is SAS instruments that our mentor uses at her work. Also, she work a lot of through excel tools – she created pivot tables and use sorting that is installed in excel. However, we decided to apply got knowledge of python and a powerful library ‘pandas’. It is faster than use excel tools. On the other hand, we need to use excel and learn its implementations through online resources on YouTube that helps to understand how to operate with excel. Modern data analysis uses clustering algorithms based on separation. Our algorithm is the McKeen average method (k-means clustering; MacQueen, 1967), in which each of the k clusters are represented by a centroid. This method is spread and common use. Thus, we will try to implement it and get the result. What do companies need this for? To get to know their clients better. To find an individual approach to each client, rather than working with everyone in the same way.

Despite the fact that many companies use loyalty programs and have colossal data, their analysts first determine the person of the customer, and then analyze its behavior. As part of the task, we had to perform clustering without a teacher. One of the most popular methods is K-Means. This method allows us to do:

- Segmentation of the market (types of buyers, loyalty).

- Association of close points (shops, customer addresses, buildings) on the map.

- Analysis and markup of new data

- Image compression/

- Identify abnormal behavior detectors.

There are also Mean-Shift, DBSCAN models, but their result is difficult to interpret.

**3. Selection of methods, algorithms, and models for project implementation**

* Information about filtering of data:

Firstly, I got unfiltered data called ‘data.csv’ in the format of .csv which I opened in Microsoft Excel, and my first step was to convert the file with data to the readable file for Mac OS as our mentor works with Microsoft formats. I worked with Python3 in Jypiter Notebook as it is powerful tool to work with datasets:

“ df = pd.read\_csv('filtered\_dataAll.csv', sep=';', encoding='utf-8-sig') “

The data consists of 797192 rows × 38 columns, where rows are orders and columns different ratings, including economic and user data of the order. Then, I checked that there are some bags in a table and I need to filter it. I used commands:

df = df[(df['Количество'] > 0)]

df = df[(df['Отменено'] != 'Да')]

Rows in the 'Cancelled' column where the 'yes' is standing - we are not interested in customers who have canceled the order to compile an analytical base table.

df = df[(df['СуммаЗаказаНаСайте'] != 0) | (df['СуммаДокумента'] != 0)]

df = df[(df['КоличествоПроданоКлиенту'] > 0)]

Got the table consists of 627124 rows × 38 columns. It is table of all placed orders. This is cleaned table without null and uninterested rows. Then, I saved the table in another encoding to work with:

df.to\_csv('filtered\_dataAll.csv', index=False, sep=';', encoding='utf-8-sig')

I am creating a new table. However, to simplify my work I used excel commands to make some tasks. In the excel, I edit and remove the empty values in the "Region" column, replacing them with values from the "Store City" column. I save the result in the table via excel. The next step is removing goods without a region as it is not interesting for us to analyze such orders as they do not gibe necessary information about users. I decided to delete all the lines where it says "Delivery" in the column "Nomenclature" and respectively in the column "Group2", because delivery can not be a complete order. The result is the filtered table which is prepared to make an analytical base table and make clustering. It is final size is 473504 rows × 42 columns called 'filtered\_dataAll.csv'

* Creating pivot table

First of all, I checked unique orders - #127 113. Further, I wrote a code for a function that will generate pivot tables with unique orders placed on the given site. There I passed on different column names each time to generate separate summary tables on different indicators. I am creating a prototype function to create a spreadsheet, so that I can then change the parameters and get the data. In the function we prescribe methods to search for one or another parameter. In this Dataset, the numbers are presented through a separator "," which prevented the function from working. For this purpose, in the table we used the function in the exсel: command in excel is "=ПОДСТАВИТЬ (T2;",";".")". So, thus I generated several pivot tables for placed orders. The next step is create some visualizations to these tables. I do it with excel tools like pie charts, graphs.

* Analyze pivot tables

1. Analysis of pivot tables by month:

In the data we got only two months: July (07) 2017 and August (08) 2017. For both months, the percentage of revenue for redeemed goods relative to placed goods is approximately 60%, which is a good indicator. A little more than half of the goods (57% on average) were redeemed from all placed items, which may be due to the fact that the number of redeemed items and unique cheques almost halved when filtering the table due to the removal of many items (orders) with erroneous data.

The percentage of unique cheques and clients is slightly higher in August (70%) than in July (68%), which is quite a high figure. This means that every month new customers come, demand and revenue increases.

*Hypothesis:* In the summer months, the buyout rate and other relative indicators are satisfactory.

Изображение выглядит как снимок экрана

Автоматически созданное описаниеИзображение выглядит как снимок экрана

Автоматически созданное описание

1. Analysis of summary tables by days of the week:

We noticed from the summary tables of purchased goods that the beginning of the week is the most profitable for our shop in terms of maximum revenue. The leaders of the days of the week with maximum revenue and margin are Monday, Tuesday, Wednesday and Thursday. The average revenue on these days is around 46,000,000 ₽. The share of redemption on these days fluctuates around 18%. However, it is interesting that the highest rate of redemption from our online store is on Sunday 66%. The number of checks aims at 18,000 ₽ these days. On the least profitable days in terms of revenue - Friday, Saturday and Sunday. Revenue is 29 669 190 ₽, 27 939 397 ₽ and 32 493 680 ₽, respectively. The number of receipts is 8,000 - 9,000 per day. The number of unique clients is half as much. And because of this, the value of customers these days is the highest. The small number of clients may be due to the fact that on weekends people rest, spend less time in the mail and websites (do not read the mailing lists), rest from computers and gadgets and spend less time at home shopping in general. While on weekdays people are more likely to check their mail and place orders (business). Also, it may be justified by the fact that many people know and believe that online stores at the weekend do not work and will deliver anyway on weekdays - so do not make orders. It may be worth making a mailing list and alerts to work at the weekend as usual and raise sales levels.

*Hypothesis:* The first half of the week (Monday, Tuesday, Wednesday, Thursday) has the highest revenue and margin.

*Hypothesis:* Customers spend less time shopping at online stores at weekends because redemption rates are lower, revenue and margin are lower, and other indicators are lower.

*Hypothesis:* It is better to do stock alerts, mailings and notifications during the first half of the working week in order to get the most out of it.

*Hypothesis:* The client's value is the same on about all days of the week.

Изображение выглядит как снимок экрана

Автоматически созданное описаниеИзображение выглядит как карта, текст

Автоматически созданное описание

1. Analysis of summary tables by hours:

From the summary tables, we can see that the shopping time on the site is definitely changing and has trends during the day. The top hours that make the largest redeemed purchases (with the highest revenue, margin, absolute margin and high average check) include: 12:00, 14:00, 15:00, 17:00, 19:00, 22:00. Worst hours of shopping on our website: from about 1:00 to 7:00. These days, most customers are obviously sleeping and so no one is shopping. However, there are exceptions such as 1:00, 2:00. It is during these hours that the revenue is high - 3-4 times more than during the hours between 1:00 and 7:00.

*Hypothesis:* 1:00, 2:00 - hours of site overload and order status updates/processing that were made the night before.

*Hypothesis:* In the evening, starting from 18-00, the revenue from the site remains at a normal level and holds until 0:00 - this is the time when customers come from work / business. They do their homework, have a rest and also check the notifications. You can make an evening newsletter too, so that the customer will pay attention exactly.

*Hypothesis:* Since the customer categories on the children's goods website are based on parents - at this time 12:00, 14:00, 15:00, 17:00 the child is in kindergarten / school and the parent can make a purchase at a time not occupied by the child for him.

*Hypothesis:* It is better to make promotions, newsletters and notifications in the hours before 'rush hour shopping' on our website, i.e. in the morning from 8-00 to 12-00 while people wake up, check their mail, smartphones, read without being distracted by our homework/work offers. To attract more unique customers.

*Hypothesis:* In the afternoon from 12:00 to 15:00, many customers have a lunch break to place an order.

Изображение выглядит как снимок экрана

Автоматически созданное описание

* Clustering with k-means

**Analytical Base** (**ABT**) - a flat table that is used for building analytical models and scoring (predicting) the future behavior of a subject. It was done by Karina and gave to me to interpret and make a k-means algorithm.

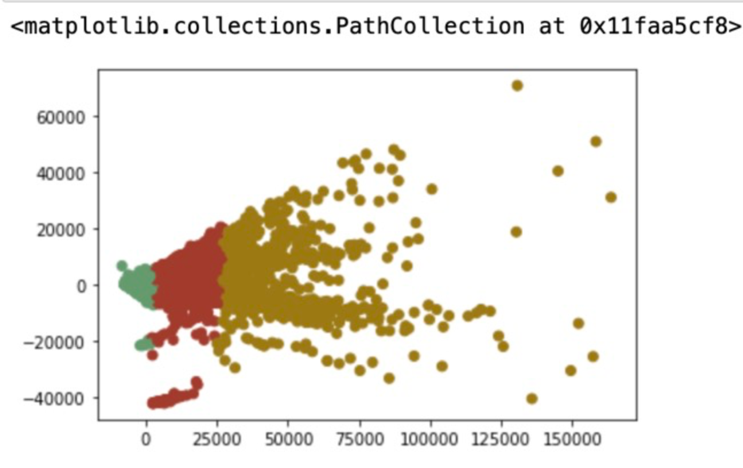
The shape of ABT was: (44056, 244)

I use Jupyter Notebook and special libraries matplotlib.pyplot, sklearn.cluster, sklearn.decomposition. The algorithm of k-mans could be used in the library sklearn.cluster. But first of all I faced with problem connecting with some specifics in Excel and Python. As ABT was made by excel tools – the data and float numbers include ‘,’ instead of ‘.’. The python cannot interpret the commas. Thus, I need to rewrite the ABT and rid of commas and strings instead of integers. I use excel tool "=ПОДСТАВИТЬ (T2;",";".")" and create additional tables in initial ABT. Then I convert strings to integers. In my code I use ‘df.drop’ not to read unnecessary columns with commas and strings. Then I get that there are some empty cells in the ABT. I decided to fill it with null values – ‘df = df.fillna(0)’. Then, I tried different values of ‘K’ – nember of clusters from 3 to 5 and implement kmeans algorithm:

kmeans = KMeans(n\_clusters=k, random\_state=42).fit(data)

So, to understand the size of my clusters I tried to visualize it with matplotlib library. I get such results:

**Изображение выглядит как снимок экрана

Автоматически созданное описание**

**Изображение выглядит как карта

Автоматически созданное описание**

The next step will be to understand which number of clusters is appropriate and suit our task. We need to normalize it and interpret the results.

1. **List of sources**

<https://habr.com/ru/company/datawiz/blog/248863/>

<https://habr.com/ru/company/sas/blog/430404/>

<https://compscicenter.ru/courses/data-mining-python/2018-spring/classes/> lecture from 1 to 8

<https://habr.com/ru/post/67078/>

<https://m.habr.com/ru/post/423597/>

<https://www.sql.ru/blogs/sasbi/1252>

presentations of our mentor