MACHINE LEARNING AND CONTENT ANALYSIS

Final project / The opposites team

Abstract

Scope of this project is the implementation of ML models capable to visualize, analyze, and predict ratios or trends related to the activity of an industrial company operating in the food sector. Data provided are both quantitative and qualitative are not fictional, and have been extracted from entity's data base.

Dimitris Mantaos

P2822011@aueb.gr

Table of contents

TABLE OF CONTENTS

INTRODUCTION AND PROJECT GOALS – PROJECT MEMBERS

Current study consists of three subprojects as follows:

- Analyse and predict credit behaviour of customers related especially to payment delays. So, this model intends to adequately predict bad debts based on customers characteristics

- Calculate the COVID effect (if any) to customers purchases behaviour during the COVID crisis period. For this reason, we will try to find similarity ratio for the market basket of each customer, and on main segments, between years 2019 and 2020, meaning that we will try to investigate if crisis caused by the virus had effect in consumption as it concerns product mix (except the decline in volumes sold).

- Implement a sentiment prediction analysis model based on the comments derived from Telesales department and phone calls. This is the first time that the organization attempts such a move and data are still limited. In any case, project initiated, and results seem promising.

Team involved in this project (The opposites) had two members, but due to unexpected events, one of the members withdraw. So, project as it concerns data finding, transformation, methodology , code deployment and report writing , along with the presentation that will follow done by the other member of the team.

DATASETS OVERVIEW

DELAYED PAYMENTS PROJECT

The Dataset used for payments analysis consists on 19 main customers quantitative and qualitative characteristics, and has 2.487 observations. Column 0 is just the code of the customer in the database, so it is omitted on any analysis that will be performed

```
customer_base.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2487 entries, 0 to 2486
Data columns (total 20 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   CardCode            2487 non-null   object 
 1   CustomerRank        2487 non-null   int64  
 2   Domestic            2487 non-null   int64  
 3   Attika              2487 non-null   int64  
 4   Territory           2487 non-null   int64  
 5   U_GRSL              2483 non-null   float64
 6   U_ZeeKey2           2487 non-null   int64  
 7   SZSegmentation      2297 non-null   object 
 8   MarketSegmentation  2213 non-null   object 
 9   Subsegment          2399 non-null   object 
 10  SlpCode             2487 non-null   int64  
 11  AvgPayDays          2487 non-null   float64
 12  PaymentTerms        2487 non-null   int64  
 13  AvgLatePayDays      2487 non-null   int64  
 14  Tziros              2487 non-null   float64
 15  AvgSVal             2487 non-null   float64
 16  MaxSval             2487 non-null   float64
 17  invNo               2487 non-null   int64  
 18  AvgInvNoPerMonth    2487 non-null   int64  
 19  Basket              2487 non-null   object 
dtypes: float64(5), int64(10), object(5)
memory usage: 388.7+ KB
```

Figure 1 : Customers dataset information

Source file is in excel format and is named “Customers profile”. Below we give a description for each feature.

Column name	***Description***
Cardcode	Customer internal code
Customer Rank	Ranking of the customer based on sales
Domestic	Local or foreign customer
Attica	Customer in Attica or providence
Territory	Geographical Areas
U_GRSL	Trade
U_Zeekey2	Market category
SZsegmentation	Internal segmentation of customers
MarketSegmentation	Segmentation of customers based on Market
Subsegment	Sub market category
SlpCode	Responsible representative
AvgPadDays	Weighted average payment days
PaymentTerms	Agreed payment days
AvgLatePayDates	Average delay in days
Tziros	12 months turnover
AvgSVal	Sales per month
invNO	Number of invoices
AvgInvNoPerMonth	Invoices per month
Basket	Top products purchased

COVID CRISIS EFFECT

Then we have two datasets of the same format, one for year 2019 and one for year 2020 (which we call Covid data). Files are named covid_products_study_2019 and covid_products_study_2020.

```
products_2019.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2839 entries, 0 to 2838
Data columns (total 3 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   Cuscode      2839 non-null   object
 1   Domestic     2839 non-null   int64 
 2   Basket_2019  2839 non-null   object
dtypes: int64(1), object(2)
memory usage: 66.7+ KB
```

Figure 2 : Customers basket Y2019 information

Column name	***Description***
Cuscode	Customer internal code
Domestic	Local or export customer
Basket_2019 or Basket_2020	Top products purchased

Data set consists of 2.839 unique customer codes. Data set for Year 2020 is exactly of the same format and shape.

COMMENTS SENTIMENT ANALYSIS

For this project we used a data set named “telesales calls.csv” with the following features:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3346 entries, 0 to 3345
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   cardcode      3346 non-null   object
 1   calldate      3346 non-null   object
 2   Result        3346 non-null   object
 3   ShortComment  725 non-null    object
 4   LongComment   2710 non-null   object
 5   Found         1782 non-null   float64
 6   Interest      1042 non-null   float64
dtypes: float64(2), object(5)
memory usage: 183.1+ KB
```

Figure 3 : Comments file description

Data set has 3.346 records, but there are some missing values which will be cleaned during the process.

Column1	cardcod	calldat	Result	ShortComment	LongComment	Found	Interes
0	A.06446	4/9/2019	Παραγγελία	Επανάκληση	ΕΔΩΣΕ ΠΑΡΑΓΓΕΛΙΑ ΕΝΑ MAMAS CAKE ΛΕΥΚΟ ΚΑΙ	1	1
1	A.01775	4/9/2019	Ενημέρωση	Αντιπρόσωπος μόνο	ΘΕΛΕΙ ΚΑΤΑΛΟΛΟ ΜΕ ΤΑ ΠΡΟΙΟΝΤΑ ΨΩΝΙΖΕΙ ΑΠ	1	0
2	A.03996	4/9/2019	Ενημέρωση	Αντιπρόσωπος μόνο	ΨΩΝΙΖΕΙ ΑΠΟ ΝΟΥΣΗ ΧΡΗΣΤΟ ΚΑΙ ΘΑ ΔΩΣΕΙ ΠΑΡΑ	1	0
3	A.04698	4/9/2019	Ενημέρωση	Ανταγωνισμός	ΨΩΝΙΖΕΙ ΑΠΟ ΖΑΦΕΙΡΑΤΟ ΚΑΙ ΔΟΥΛΕΥΕΙ ΤΟ ΜΑΜ	1	0
4	A.01653	4/9/2019	Ενημέρωση	Κατάλογος προϊόντων	ΘΕΛΕΙ ΔΕΙΓΜΑ ΓΙΑ ΨΩΜΙ oLIKIS ΜΕ ΠΡΟΖΥΜΙ ΠΑΙΡ	1	0
5	A.04465	4/9/2019	Ενημέρωση	Ελλειπε ο υπεύθυνος	ΔΕΝ ΑΠΑΝΤΑ ΣΤΟ ΤΗΛ.	0	
6	A.04023	4/9/2019	Ενημέρωση	Ελλειπε ο υπεύθυνος	ΕΛΕΙΠΕ Ο ΥΠΕΥΘΥΝΟΣ.	0	
7	A.04002	4/9/2019	Ενημέρωση	Λαθασμένα στοιχεία ε	ΛΑΘΟΣ ΤΗΛ	0	
8	A.05475	4/9/2019	Ενημέρωση		ΨΩΝΙΖΕΙ ΑΠΟ ΠΥΡΡΟ ΑΛΛΑ ΘΕΛΕΙ ΚΑΛΥΤΕΡΕΣ ΤΙΜ	1	0
9	A.01767	4/9/2019	Παραγγελία Δείγμο	Κατάλογος προϊόντων	ΕΝΗΜΕΡΩΘΗΚΕ ΓΙΑ ΝΕΑ ΠΡΟΙΟΝΤΑ ΚΑΙ ΕΠΙΘΥΜΕ	1	0
10	A.03972	5/9/2019	Ενημέρωση	Αντιπρόσωπος μόνο	ΨΩΝΙΖΕΙ ΑΠΟ ΝΟΥΣΗ ΒΑΣΙΛΗ ΚΑΙ ΕΙΝΑΙ ΕΝΗΜΕΡΩ	1	0

Column name	Description
Cardcode	Customer internal code
Calldate	Date of the call
Result	Outcome of the contact
ShortComment	Predefined comment
Long comment	Free text
Found	Check point if the concact was made
Interest	Field used for the sentiment analysis

DATA VISUALISATION AND TRANSFORMATION

We visualize datasets to get an idea about the structure of the data. Fist we create a histogram to see how delays of payments are distributed.

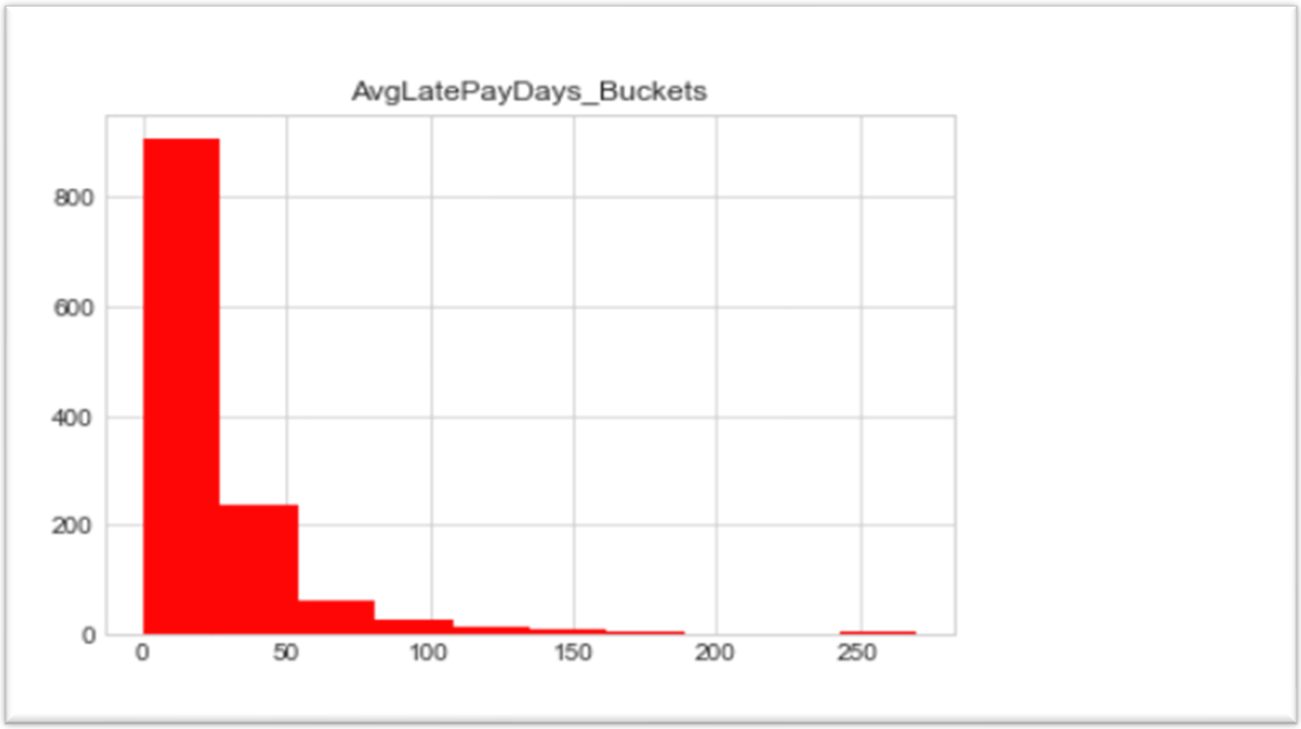
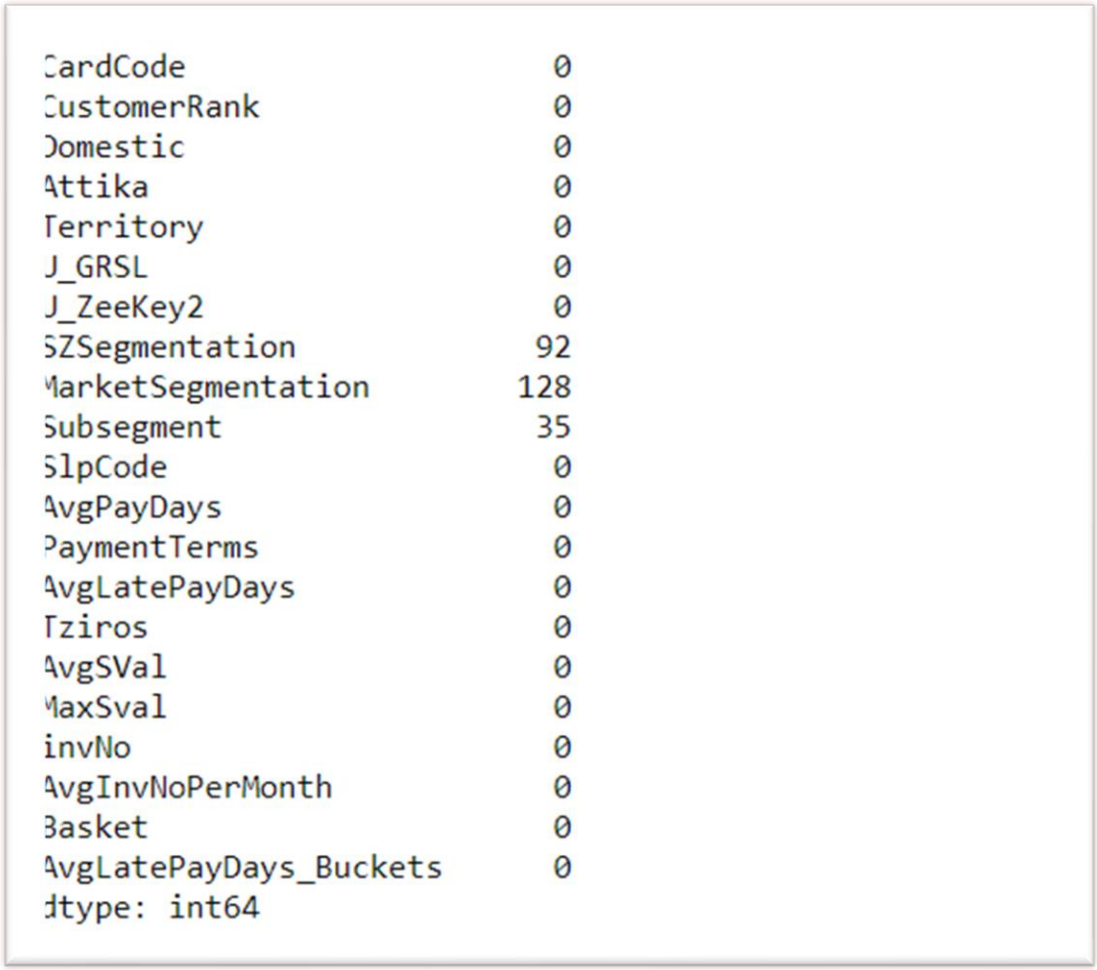


Figure 4 : Analysis of payment delays

It is noticeable that most of the delays are concentrated in the range between 0 and 60 days. Presence of delays more than 60 days is limited.



.

Figure 5 : Missing values in customers profile dataset

Data set contains missing values which we removed.

Then we analyze datasets having the information about market basket. Products are presented with their internal codes in a list containing the top products purchased

	Cuscode	Domestic	Basket_2019
0	A.00001	0	0015027000, 0018007003, 0025006010, 0025006011...
1	A.00008	0	0025022020, 0094013002, 0094079001, 0095004000...
2	A.00023	0	0005001010
3	A.00028	0	0001026020, 0015035000, 0068001001, 0105003001...
4	A.00032	0	0009001002, 0015027000, 0015042001, 0025006000...

	Cuscode	Domestic	Basket_2020
0	A.00001	0	0015027000, 0015035001, 0018007003, 0025006011...
1	A.00008	0	0025022020, 0094013002, 0094079001, 0095004000...
2	A.00023	0	0005001010
3	A.00028	0	0001026020, 0015035000, 0068001001, 0135000000
4	A.00032	0	0015027000, 0015042001, 0025006000, 0025006010...

Data sets differ just by 4 customers, which possibly are accounts that do not exist in Year 2020 but were present during Year 2019.

DELAYD PAYMENTS

First, we use visualizations to investigate possible relations between delay in payments and customer characteristics. For this reason, we used the fields “Tziros”, “SZSegmentation”, and “Payment Terms”

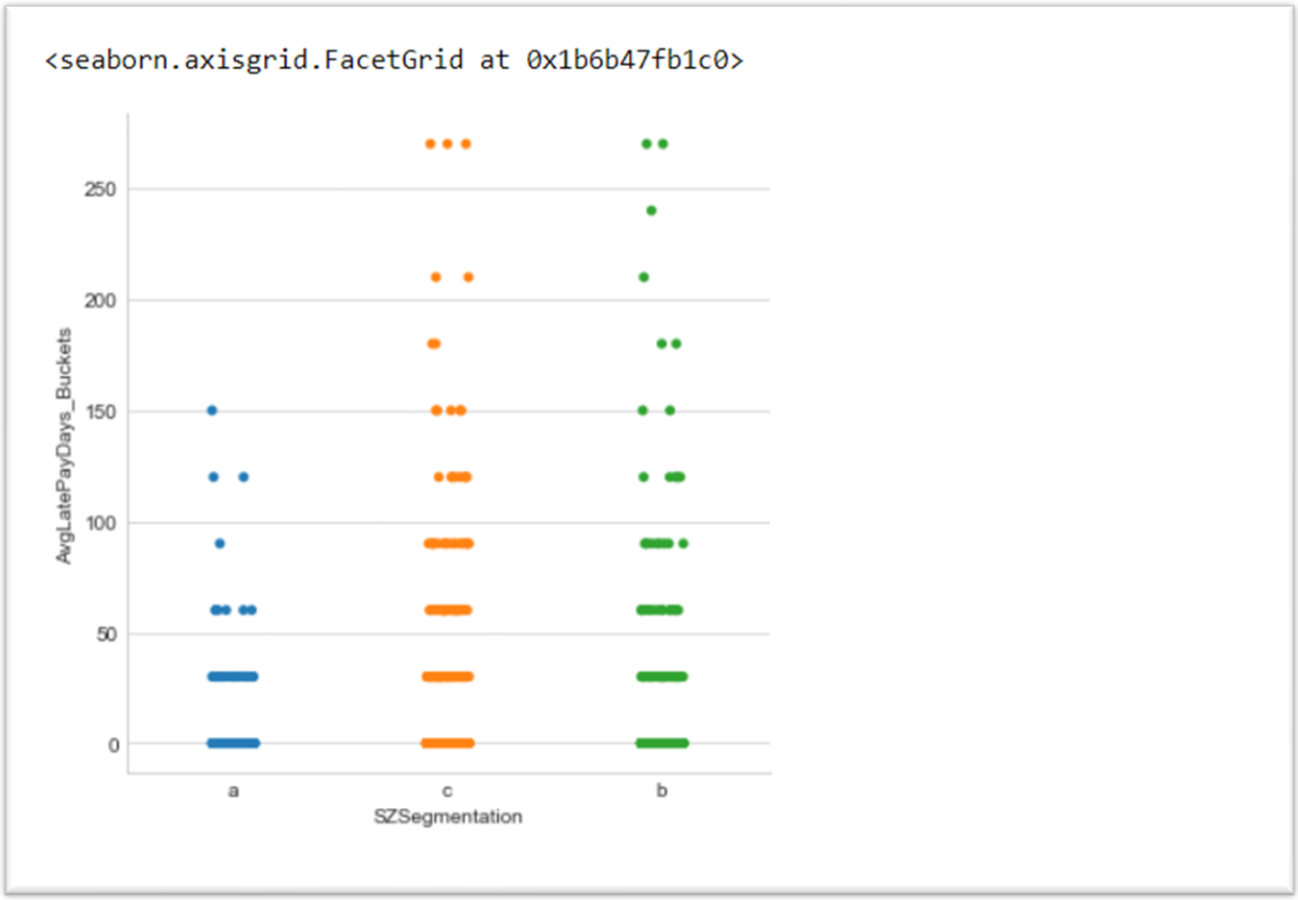
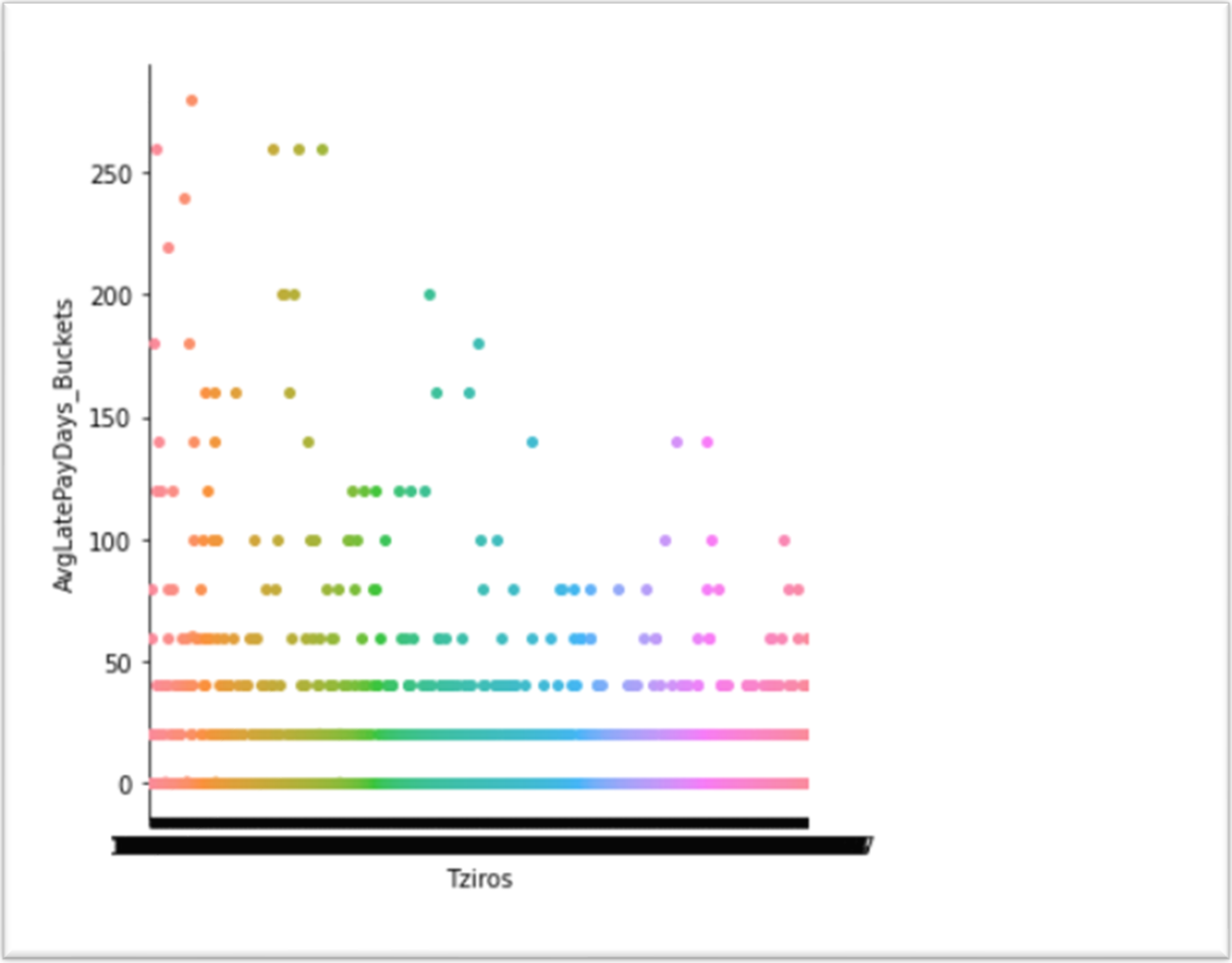


Figure 6 : Correlation between Market Segmentation and payment delays

We notice that there is no relation between lateness in payments and customers segmentation. Specific segmentation is based to customers size and importance for the organization. So customers of category “a” are considered having financial strength and big size. The only result coming from the exhibit is that delays more than 150 days are coming from customers categorized as “b” or “c”



We perform the same analysis comparing turnover with late payments. From visualization we come to the same result. There is not direct relation with the level of sales to late payments.

From these exhibits we assume that is difficult to find a prediction model with high accuracy ratio.

After a lot of experiments, we chose a decision tree model as the best option for this case. Then we grouped days of delay in buckets of 30. After running the model o lot of times the following features chosen for the model construction

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1137 entries, 9 to 2484
Data columns (total 4 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   SZSegmentation   1137 non-null   object
 1   PaymentTerms     1137 non-null   int64
 2   Tziros           1137 non-null   float64
 3   invNo            1137 non-null   int64
dtypes: float64(1), int64(2), object(1)
memory usage: 44.4+ KB
```

Figure 7 : Features used for predition modeling

```
array([[147,  41,   6,   3,   0,   2,   1,   0,   0,   0],
       [ 38,  10,   2,   3,   0,   1,   0,   0,   0,   0],
       [  8,   3,   1,   1,   0,   0,   1,   0,   0,   0],
       [  3,   1,   1,   0,   0,   0,   0,   0,   0,   0],
       [  2,   2,   0,   0,   1,   0,   0,   0,   0,   0],
       [  2,   0,   0,   0,   0,   0,   0,   0,   0,   0],
       [  1,   0,   0,   0,   0,   0,   0,   0,   0,   0],
       [  1,   0,   1,   0,   0,   0,   0,   0,   0,   0],
       [  1,   0,   0,   0,   0,   0,   0,   0,   0,   0],
       [  1,   0,   0,   0,   0,   0,   0,   0,   0,   0]], dtype=int64)
```

Figure 8 : Confusion matrix for Decision tree model

We calculated the accuracy ratio and got as a result using *sklearn.metrics* Accuracy score of the predictions is 0.56.

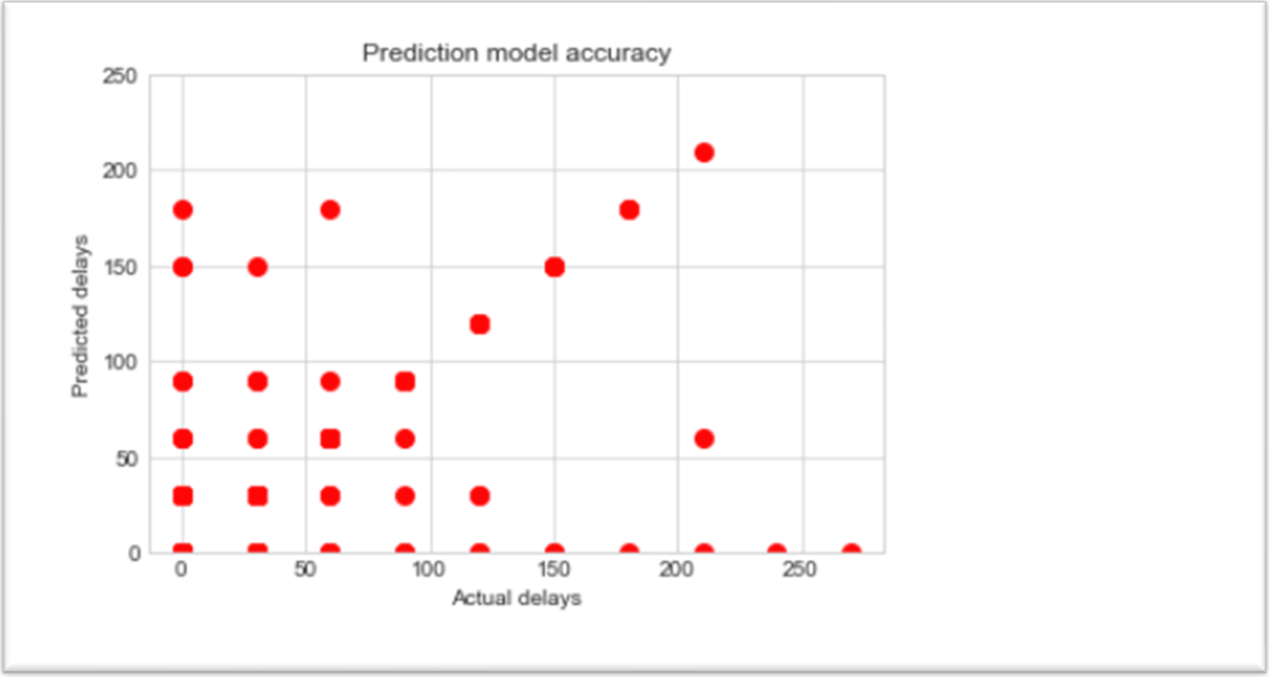


Figure 9 : Actual VS predictions for Decision Tree model

As seen in exhibit above, model performs better when delays are high, whilst it has poor performance in the lower parts.

Then we tried to predict the same ratio using a MLP classifier Neural Network. Results seemed better since accuracy score was 0.72 instead for the decision tree model that was 0.56

```
irray([[200,   0,   0,   0,   0,   0,   0,   0,   0,   0],
       [ 54,   0,   0,   0,   0,   0,   0,   0,   0,   0],
       [ 13,   1,   0,   0,   0,   0,   0,   0,   0,   0],
       [  5,   0,   0,   0,   0,   0,   0,   0,   0,   0],
       [  5,   0,   0,   0,   0,   0,   0,   0,   0,   0],
       [  2,   0,   0,   0,   0,   0,   0,   0,   0,   0],
       [  1,   0,   0,   0,   0,   0,   0,   0,   0,   0],
       [  2,   0,   0,   0,   0,   0,   0,   0,   0,   0],
       [  1,   0,   0,   0,   0,   0,   0,   0,   0,   0],
       [  1,   0,   0,   0,   0,   0,   0,   0,   0,   0]], dtype=int64)
```

Figure 10 : Confusion Matrix for MLP classifier model

Furthermore, MLP classifier performed better in the lower bucket of days (0-30) but failed totally in the upper levels as shown in exhibit below

It was not scope of this study any further analysis so this will be performed in later stages.

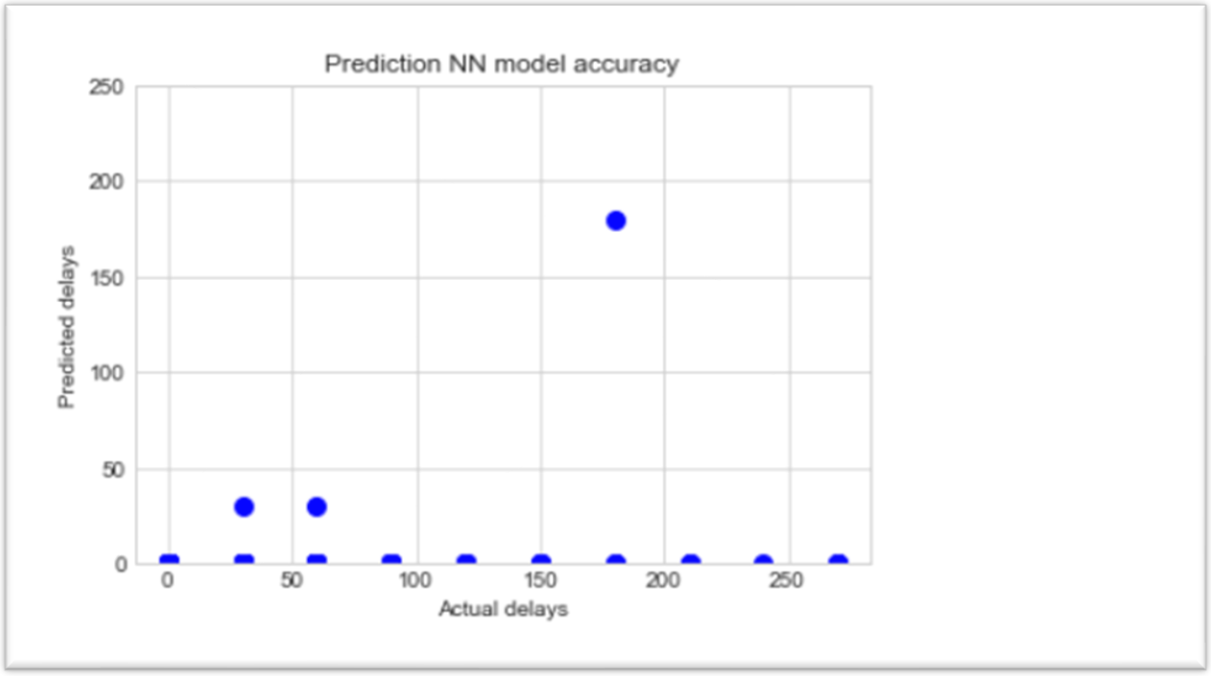


Figure 11 : Actual VS prediction for MLP classifier

A question raised is if we could use a combination of models to enhance prediction accuracy, but this was left for later stages.

As a conclusion, none of the models seemed adequate to predict bad debt. This mainly has to do mostly with the existing segmentation of customers. Current groups are made based on commercial characteristics and ignore the financial position of the customers.

To construct a prediction model, we should seek financial data and cluster the customers based on these. Then we should expect a model that will have a better accuracy ratio.

COVID EFFECT ON MARKET BASKET

We merged datasets for Year 2019 and Year 2020 in one new dataset.

	Cuscode	Domestic_x	Basket_2019	Domestic_y	Basket_2020
0	A.00001	0	0015027000, 0018007003, 0025006010, 0025006011...	0	0015027000, 0015035001, 0018007003, 0025006011...
1	A.00008	0	0025022020, 0094013002, 0094079001, 0095004000...	0	0025022020, 0094013002, 0094079001, 0095004000...
2	A.00023	0	0005001010	0	0005001010
3	A.00028	0	0001026020, 0015035000, 0068001001, 0105003001...	0	0001026020, 0015035000, 0068001001, 0135000000
4	A.00032	0	0009001002, 0015027000, 0015042001, 0025006000...	0	0015027000, 0015042001, 0025006000, 0025006010...

To evaluate the similarity between Basket_2019 and Basket_2020 we use from difflib library the Sequence Matcher method.

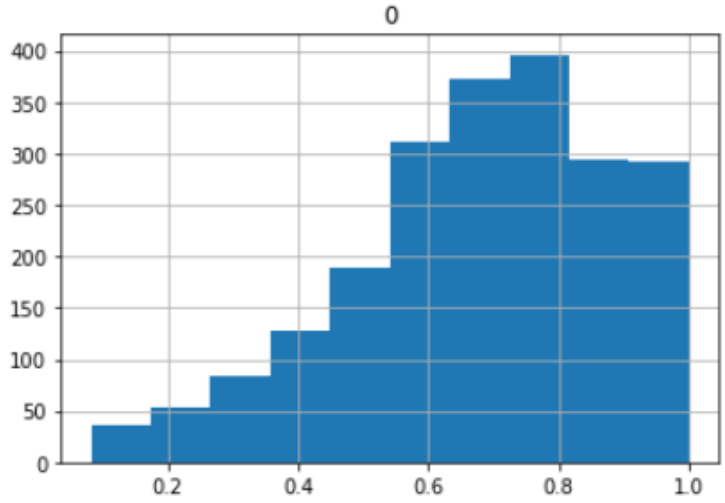


Figure 12 : Similarity for the total customers list

Initialy, we get the similarity ratio for the full dataset. Main statistics are:

Descripion	Value
Mean	.68
Standard deviation	.21
Maximum	1.0
Minimum	.08

Then we get the same ratio by applying the same method in the customers segmented by “SZsegmentation” feature. So we cluster customers according the values in this field (“a”,”b”,”c”) and we got the following results :

Customers “a”

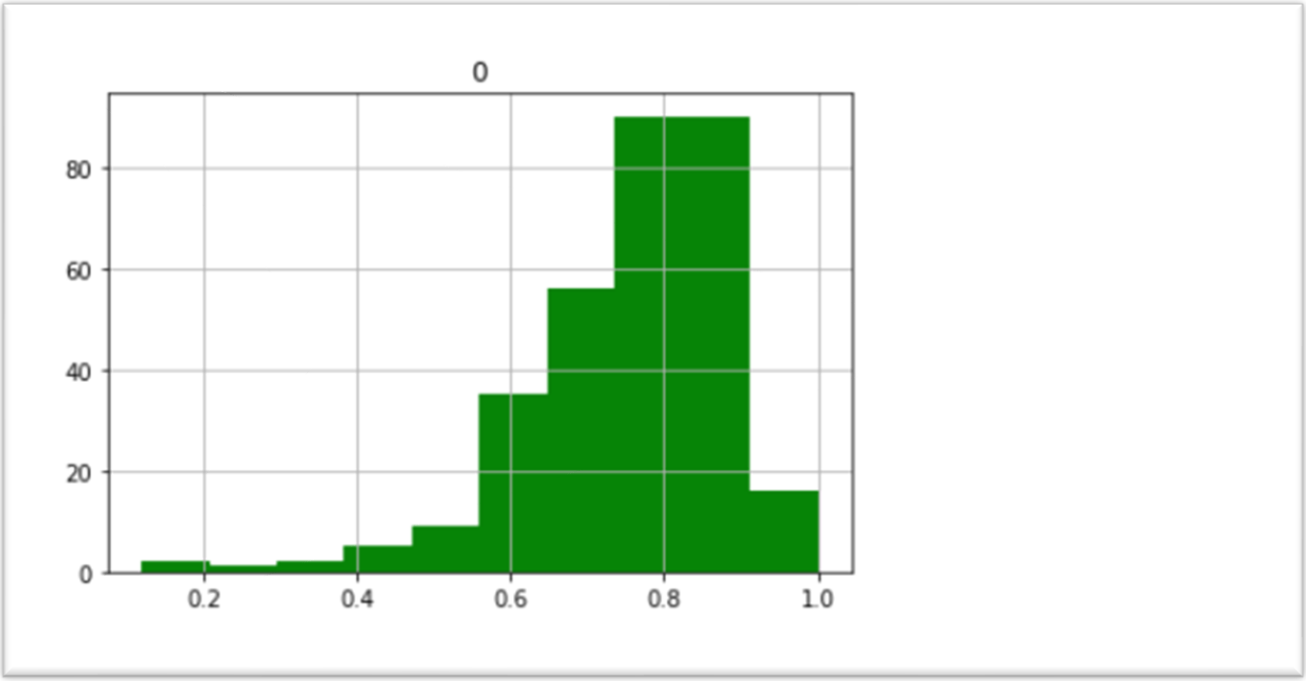


Figure 13 : Similarity for customers "a"

Descripion	Value
Mean	. 76
Standard deviation	.13
Maximum	1.0
Minimum	.12

Customers o category “a” have higher similarity ratio and values are more concentrated around the mean than the other categories

Customers "b"

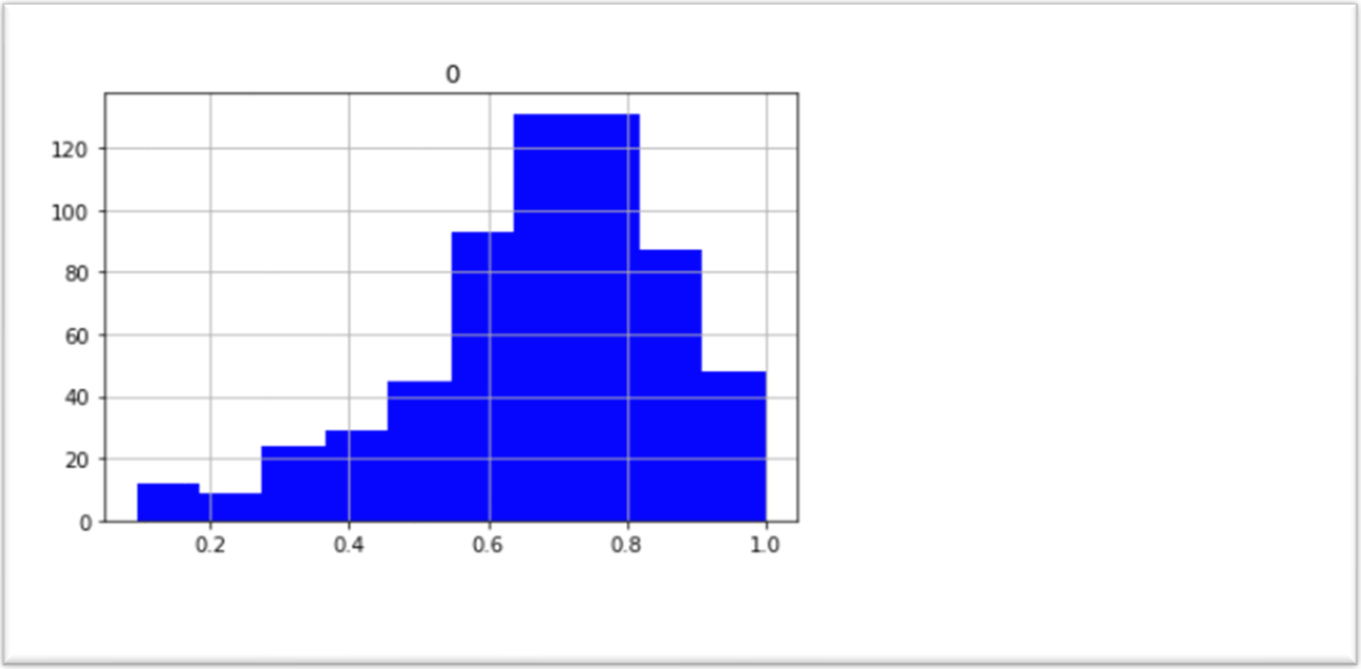


Figure 14 : Similarity of customers category "b"

Descripion	Value
Mean	.68
Standard deviation	.18
Maximum	1.0
Minimum	.09

Customers "c"

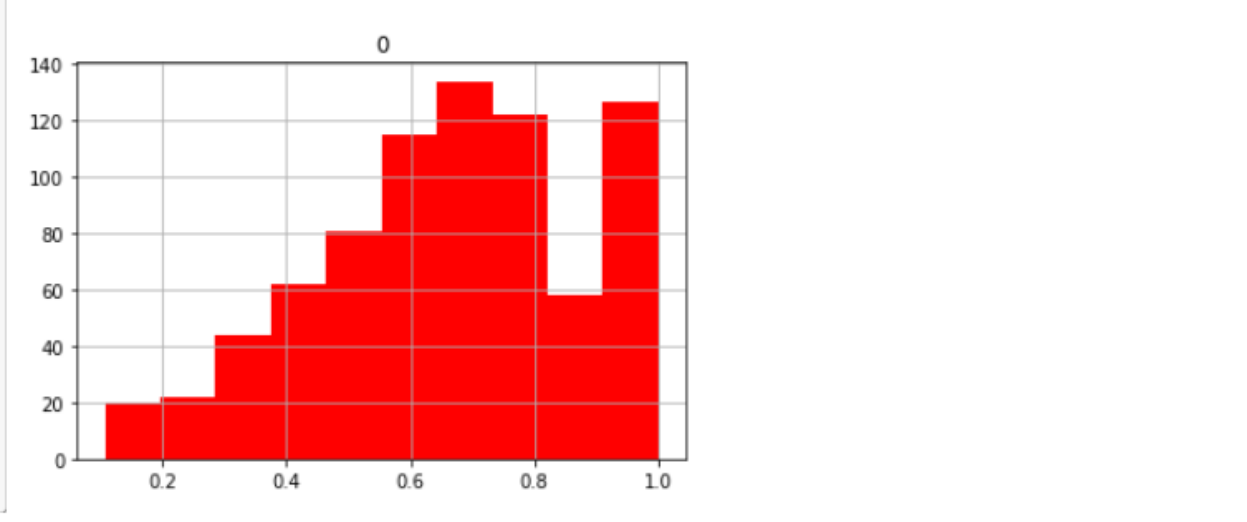


Figure 15 : Similarity for customers category "c"

Descripion	Value
Mean	.66
Standard deviation	.22
Maximum	1.0
Minimum	.11

Average similarity ratio for category "c" is .67 which is significant lower from both categories "a" and "b".

From statics above, we concluded that first there was a significant movement in customer behavior, and that customers of category "a" are the most stable. On the other hand, customers of category "b" and "c" affected the most.

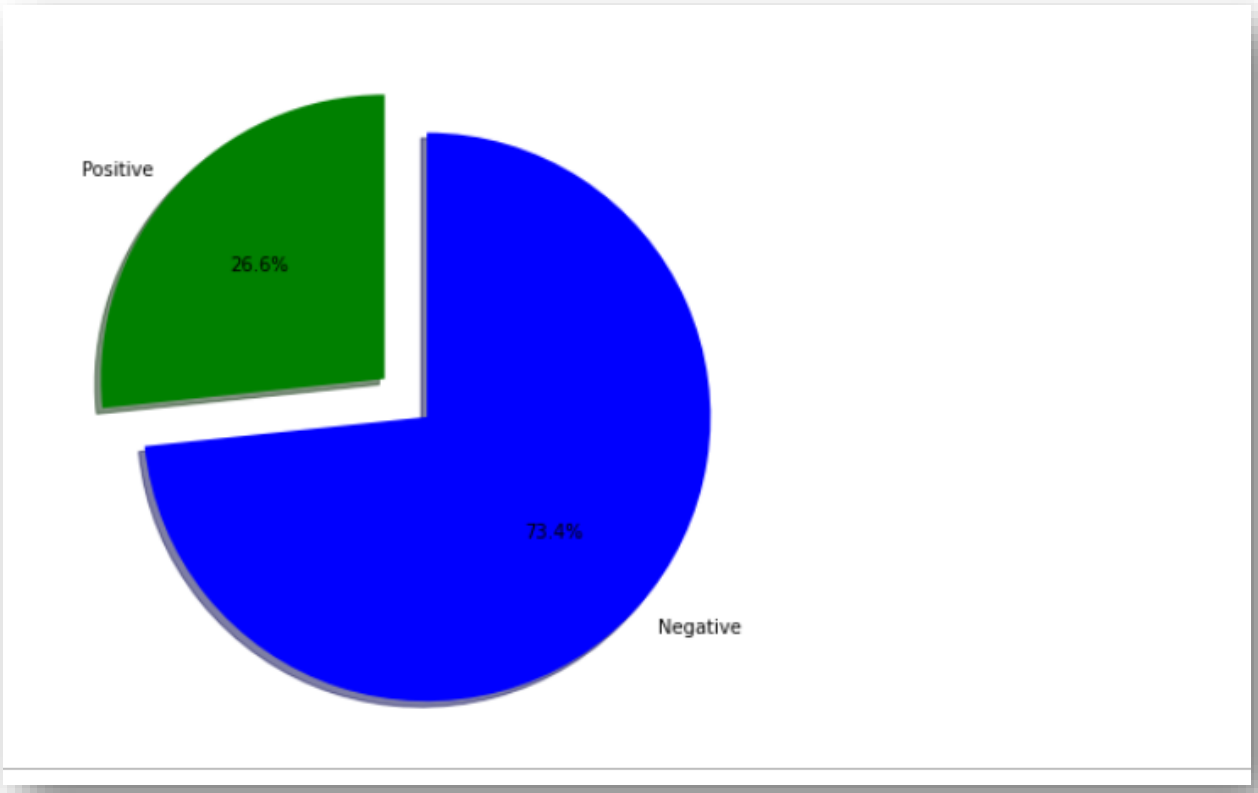


Figure 16 : % of Negative and Positive comments

From comments dataset we conclude that negative responses are 73,4% of the sample. This is fully explainable since calls were made to non-active or old customers mainly of category “c”.

From the initial dataset we kept the columns “Long Comment”, “Found”, “Interest”.

	LongComment	Found	Interest
0	ΕΔΩΣΕ ΠΑΡΑΓΓΕΛΙΑ ΕΝΑ MAMAS CAKE λΕΥΚΟ ΚΑΙ ΘΕΛΕ...	1.0	1.0
1	ΘΕΛΕΙ ΚΑΤΑΛΟΛΟ ΜΕ ΤΑ ΠΡΟΙΟΝΤΑ ΨΩΝΙΖΕΙ ΑΠΟ ΝΟΥ...	1.0	0.0
2	ΨΩΝΙΖΕΙ ΑΠΟ ΝΟΥΣΗ ΧΡΗΣΤΟ ΚΑΙ ΘΑ ΔΩΣΕΙ ΠΑΡΑΓΓΕΛ...	1.0	0.0
3	ΨΩΝΙΖΕΙ ΑΠΟ ΖΑΦΕΙΡΑΤΟ ΚΑΙ ΔΟΥΛΕΥΕΙ ΤΟ MAMAS CA...	1.0	0.0
4	ΘΕΛΕΙ ΔΕΙΓΜΑ ΓΙΑ ΨΩΜΙ oLIKIS ΜΕ ΠΡΟΖΥΜΙ ΠΑΙΡΝΕ...	1.0	0.0
5	ΔΕΝ ΑΠΑΝΤΑ ΣΤΟ ΤΗΛ.	0.0	NaN
6	ΕΛΕΙΠΕ Ο ΥΠΕΥΘΥΝΟΣ.	0.0	NaN
7	ΛΑΘΟΣ ΤΗΛ	0.0	NaN
8	ΨΩΝΙΖΕΙ ΑΠΟ ΠΥΡΡΟ ΑΛΛΑ ΘΕΛΕΙ ΚΑΛΥΤΕΡΕΣ ΤΙΜΕΣ. ...	1.0	0.0
9	ΕΝΗΜΕΡΩΘΗΚΕ ΓΙΑ ΝΕΑ ΠΡΟΙΟΝΤΑ ΚΑΙ ΕΠΙΘΥΜΕΙ ΔΕΙΓ...	1.0	0.0
10	ΨΩΝΙΖΕΙ ΑΠΟ ΝΟΥΣΗ ΒΑΣΙΛΗ ΚΑΙ ΕΙΝΑΙ ΕΝΗΜΕΡΩΜΕΝΗ...	1.0	0.0

Figure 17 : Sample from comments dataset

Then we dropped the rows where value of Found equals to 0. In these cases, customers were not available for a phone call

After we changed responses from -1 to “neg” and 1 to “pos”, and we discard observation with value 0, since these comments were considered neutral.

	LongComment	Interest
0	ΕΔΩΣΕ ΠΑΡΑΓΓΕΛΙΑ ΕΝΑ MAMAS CAKE λΕΥΚΟ ΚΑΙ ΘΕΛΕ...	pos
17	ΘΑ ΚΑΛΕΣΕΙ Ο ΙΔΙΟΣ ΟΤΑΝ ΧΡΕΙΑΣΤΕΙ ΚΑΤΙ 2105621903	neg
18	ΘΑ ΠΑΡΕΙ ΤΗΛ ΑΝ ΧΡΕΙΣΤΕΙ ΚΑΤΙ ΑΠΟ ΤΗΝ ΕΤΑΙΡΕΙΑ...	neg
19	ΣΥΝΕΡΓΑΖΟΤΑΝ ΠΑΛΙΑ ΜΑΖΙ ΜΑΣ ΨΩΝΙΖΕΙ ΜΟΝΟ ΑΠΟ Α...	neg
22	ΧΟΝΔΡΙΚΗ ΠΩΛΗΣΗ ΠΑΓΩΤΩΜΗΧΑΝΩΝ,	pos

Figure 18 : Final form of dataset that is used for prediction model

After construction of our final comments dataset as shown in Figure 18, we performed factorization of the responses.

```
(array([0, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
        1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1,
        1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1,
        0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1,
        1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1,
        1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
        1, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0,
        1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
        1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
        1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
        1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1,
        0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1,
        1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0,
        0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1,
        0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0,
        0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1,
        0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
        1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1,
        1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1,
        0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1,
        1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0,
        1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1,
        1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0,
        1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1,
        1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1,
        1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1,
        0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1,
        1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1,
        1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1,
        1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
        1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0,
        0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0,
        0, 1, 1, 1, 1, 0, 1], dtype=int64),
 Index(['pos', 'neg'], dtype='object'))
```

Figure 19 : Factorization of responses

Afterwards we tokenized the comments, so every word tied with a specific number.

Figure 20 : Tokenization of comments texts

After all the above we were ready to build and fit a neural network model. We choose LSTM with the parameters as shown below :

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_2 (Embedding)      (None, 200, 32)           15264
_________________________________________________________________
spatial_dropout1d_2 (Spatial (None, 200, 32)           0
_________________________________________________________________
lstm_2 (LSTM)                (None, 50)                16600
_________________________________________________________________
dropout_2 (Dropout)          (None, 50)                0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 51
=================================================================
Total params: 31,915
Trainable params: 31,915
Non-trainable params: 0
_________________________________________________________________
```

The we fit the model and run it with 5 epochs. Results are shown below

```
Epoch 1/5
20/20 [==============================] - 4s 95ms/step - loss: 0.6374 - accuracy: 0.7101 - val_loss: 0.5555 - val_accuracy: 0.76
92
Epoch 2/5
20/20 [==============================] - 2s 85ms/step - loss: 0.5898 - accuracy: 0.7246 - val_loss: 0.5397 - val_accuracy: 0.76
92
Epoch 3/5
20/20 [==============================] - 2s 86ms/step - loss: 0.5394 - accuracy: 0.7246 - val_loss: 0.4748 - val_accuracy: 0.76
92
Epoch 4/5
20/20 [==============================] - 2s 84ms/step - loss: 0.4700 - accuracy: 0.7295 - val_loss: 0.3918 - val_accuracy: 0.82
69
Epoch 5/5
20/20 [==============================] - 2s 83ms/step - loss: 0.3444 - accuracy: 0.8792 - val_loss: 0.2705 - val_accuracy: 0.90
38
```

Figure 21 : Prediction model accuracy

Accuracy ratio seems promising and according to our expectations.

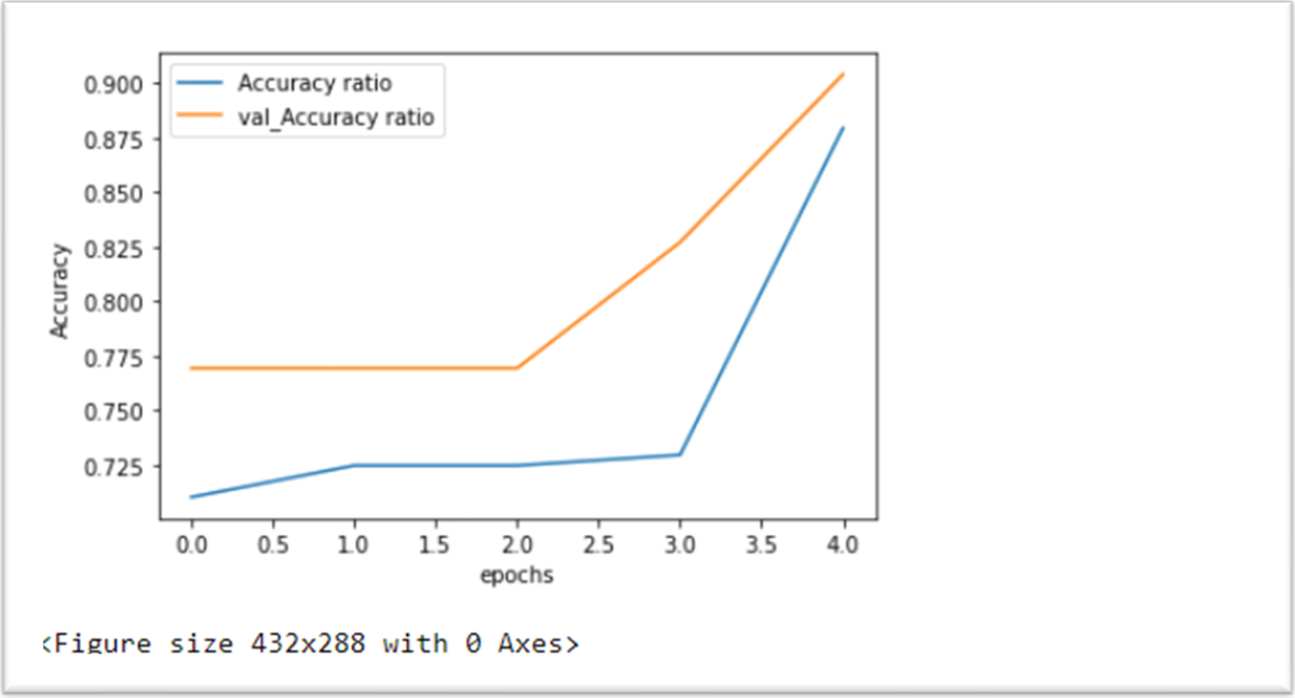


Figure 22 : Validation and actual accuracy ratios

Visualize improvement of Accuracy ratio through the sequential ratios. Just for experiment, we run the model with 10 epochs, but results were deteriorated somehow.

Since accuracy ratio close to 90% is adequate to our case for now, we consider this model as successful. Problem is that dataset was very small, so we should evaluate performance on a later stage when data are more.

Then we tried to check our model for overfitting . So for that we run the model again using the Nearmiss method.

```
Epoch 1/20
9/9 [==============================] - 3s 114ms/step - loss: 0.6918 - accuracy: 0.5265 - val_loss: 0.6860 - val_accuracy: 0.940
3
Epoch 2/20
9/9 [==============================] - 1s 81ms/step - loss: 0.6822 - accuracy: 0.7652 - val_loss: 0.6783 - val_accuracy: 0.6866
Epoch 3/20
9/9 [==============================] - 1s 81ms/step - loss: 0.6697 - accuracy: 0.7576 - val_loss: 0.6634 - val_accuracy: 0.7164
Epoch 4/20
9/9 [==============================] - 1s 83ms/step - loss: 0.6497 - accuracy: 0.7462 - val_loss: 0.6400 - val_accuracy: 0.6866
```

Figure 23 : Prediction model with overfitting check

Now prediction accuracy seems lower indicating that previous model was overfitted.

It is not part of this study further analysis on this subject. For sure this will be performed when we got a bigger dataset containing more that 5,000 records for example.

DISCUSSION ON LESSONS LEARNED AND FUTURE STEPS

It was the first time that such a project attempted within the organization. There was no problem in data collection, since data bases are in good shape, and data are mostly cleaned from origin. All files extracted from SQL tables and views, that were exported in Excel format.

Data are homogenized ,as for example, customer internal code is the same in all datasets. Thus, need for transformation was limited.

There are no double or redundant entries. There are some missing values, but no in big magnitude and not in important features.

So preprocessing of the datasets was in the bigger part already made by the ICT department of the entity.

As it concerns the sub-project results, we could comment the following:

- For delayed payments the models promoted failed to predicts with high accuracy the payment delays. This has to do mostly with the current customers portfolio segmentation than with models prediction power.
- Covid crisis effect on product market basket was clearly stated through the implementation of the similarity methods.
- As it concerns sentiment analysis on comments, results where surprising promising as our knowledge in the use of such tools is currently limited and dataset was very small

I would like to express many thanks to company ICT department for the quality of the datasets that made any attempt for analyzing and processing data much easier.

BIBLIOGRAPHY AND RESOURCES

Davies, A. (n.d.). *A Natural Language Processing (NLP) Primer*. Retrieved from https://towardsdatascience.com/a-natural-language-processing-nlp-primer-6a82667e9aa5

difflib — Helpers for computing deltas. (n.d.). Retrieved from https://docs.python.org/3/library/difflib.html#

Muller, A., & Guido, S. (n.d.). *Introduction to Machine Learning with Python.*

Perakis, G. (2021). Labs for NLTK , Keras, Tensorflow. AUEB Msc in Business Analytics.

Sharma, A. (2019, June 17). *Histograms in Matplotlib*. Retrieved from https://www.datacamp.com/community/tutorials/histograms-matplotlib

sklearn.metrics.jaccard_score. (n.d.). Retrieved from https://scikit-learn.org/stable/index.html: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.jaccard_score.html

TechVidan. (n.d.). *Sentiment Analysis using Python*. Retrieved from https://techvidvan.com/tutorials/python-sentiment-analysis/