Dimitris Mantaos

P2822011@aueb.gr

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.*

Machine Learning and content analysis

Final project / The opposites team

***Table of contents***

Table of Contents

[Introduction and project goals – project members 2](#_Toc81740890)

[Datasets overviεw 3](#_Toc81740891)

[Delayed payments project 3](#_Toc81740892)

[COVID crisis effect 4](#_Toc81740893)

[Comments sentiment analysis 5](#_Toc81740894)

[Data VISUALISATION AND transformation 8](#_Toc81740895)

[Project results 11](#_Toc81740896)

[Delayd payments 11](#_Toc81740897)

[COVID effect on market basket 16](#_Toc81740898)

[Comments sentiment analysis 20](#_Toc81740899)

[Discussion on lessons learned and future steps 26](#_Toc81740900)

[Bibliography and resources 27](#_Toc81740901)

# 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 overviεw

## 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

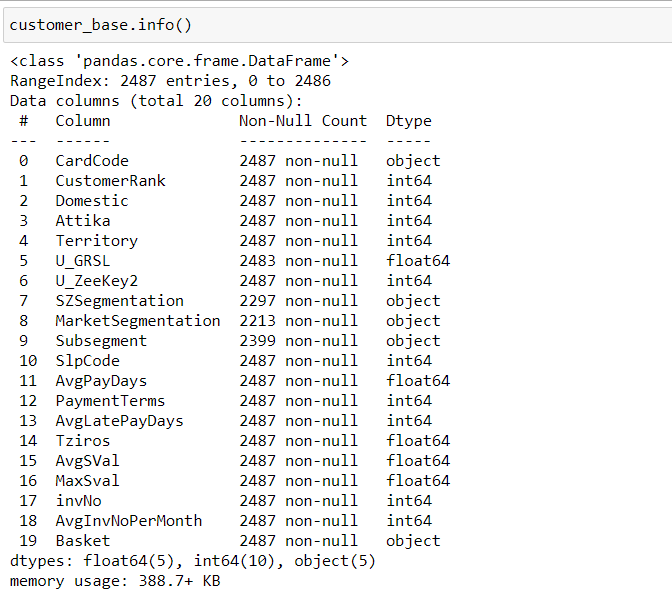


Figure : Customers dataset information

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.

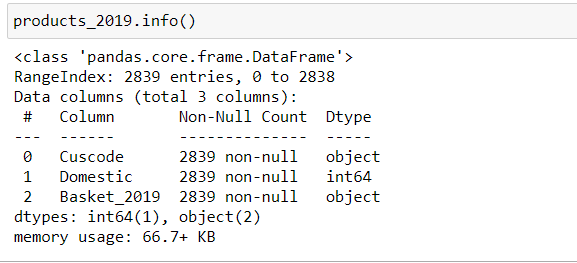


Figure : Customers basket Y2019 information

|  |  |
| --- | --- |
| ***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:

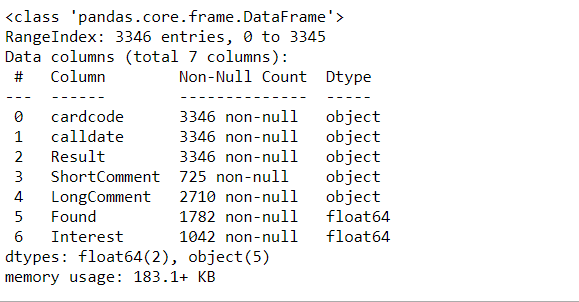


Figure : Comments file description

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



|  |  |
| --- | --- |
| 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.

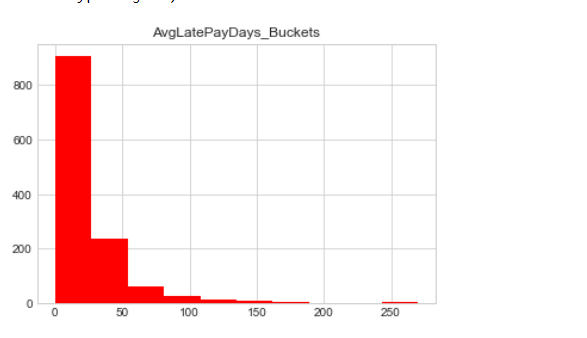


Figure : Analysis of payment delays

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.

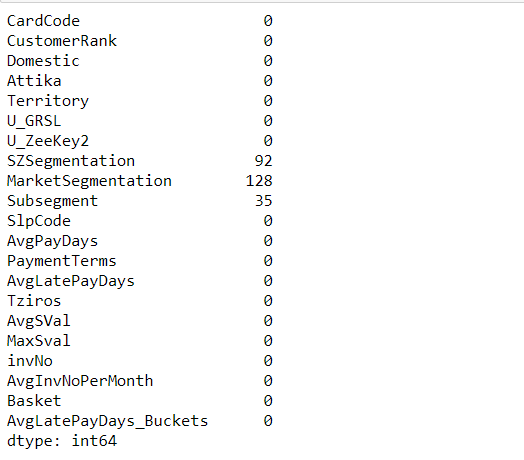
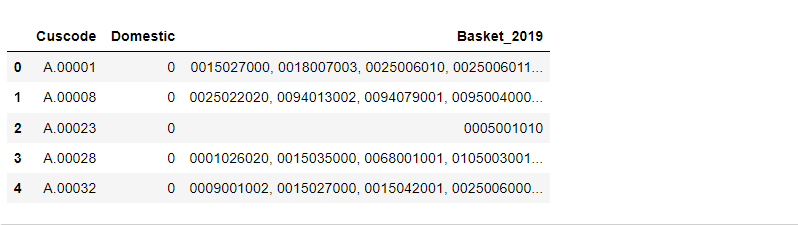
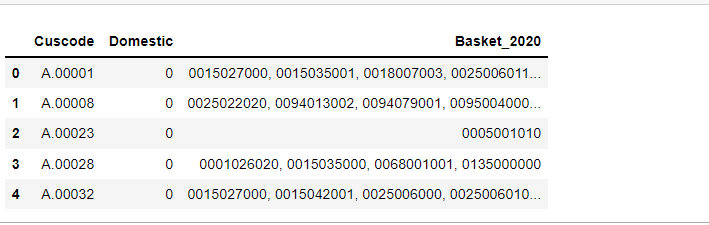
.

Figure : 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





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

# Project results

### 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”

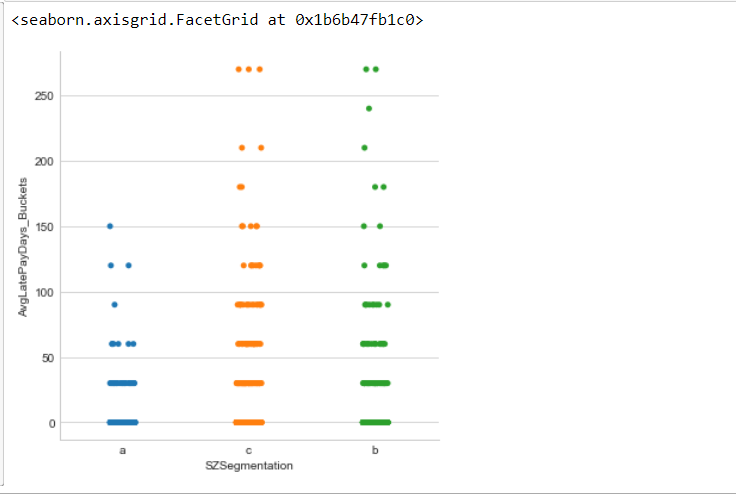
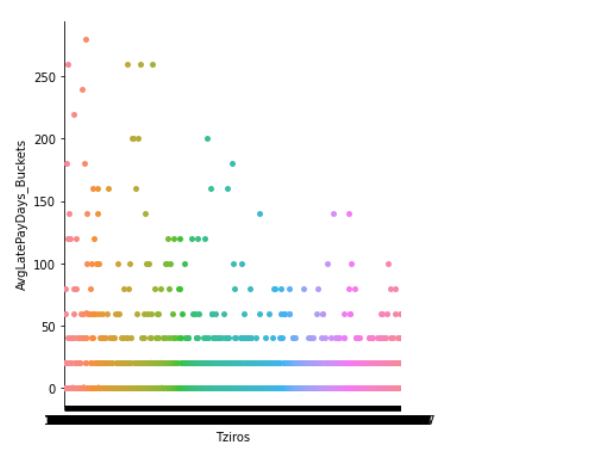


Figure : Correlation between Market Segmentation and payment delays

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

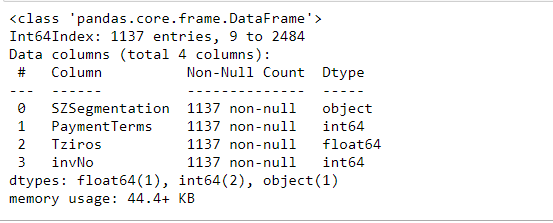


Figure : Features used for predition modeling

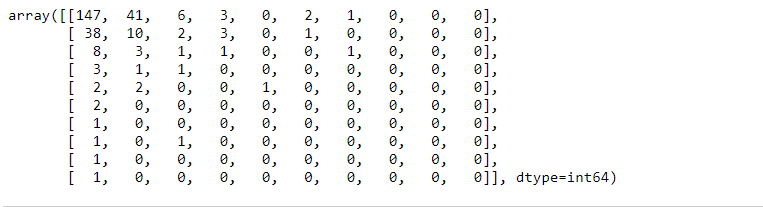


Figure : Confusion matrix for Decision tree model

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

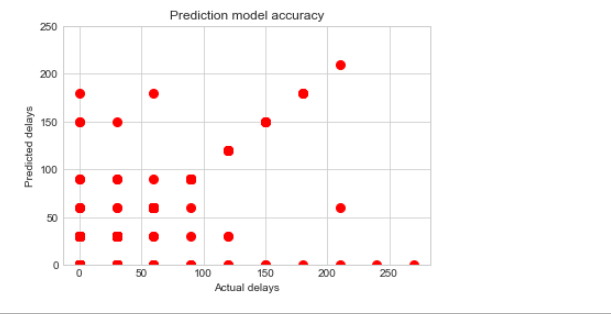


Figure : Actual VS predictions for Decision Tree model

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

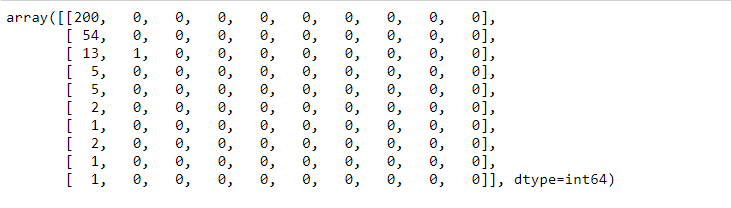


Figure : Confusion Matrix for MLP classifier model

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.

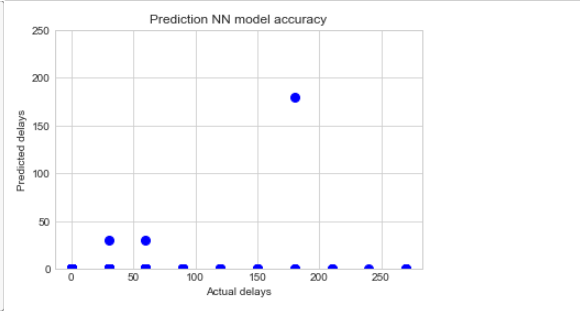


Figure : Actual VS prediction for MLP classifier

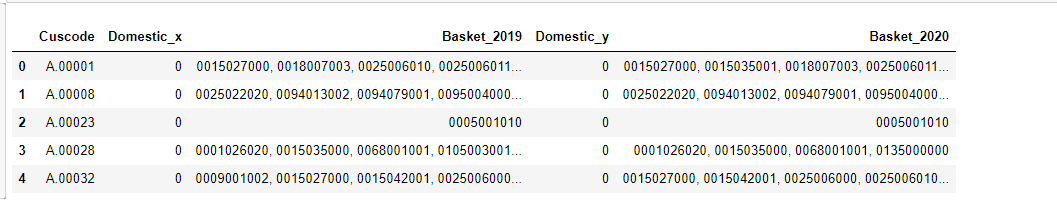
A question raised is if we could use a combination of models to enhance prediction accuracy, but this was left for later stages.

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.



To evaluate the similarity between Basket\_2019 and Basket\_2020 we use from difflib library the Sequence Matcher method.

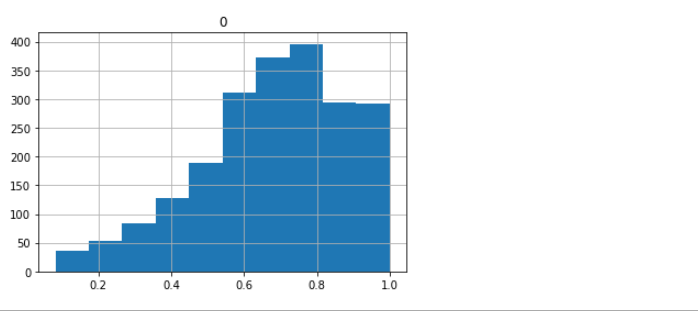


Figure : Similarity for the total customers list

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”***

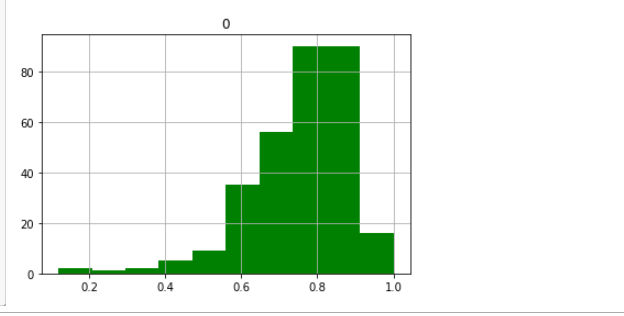


Figure : Similarity for customers "a"

|  |  |
| --- | --- |
| 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”***

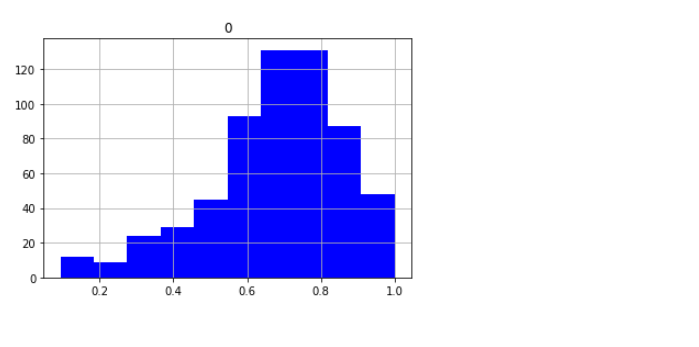


Figure : Similarity of customers category "b"

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

***Customers “c”***

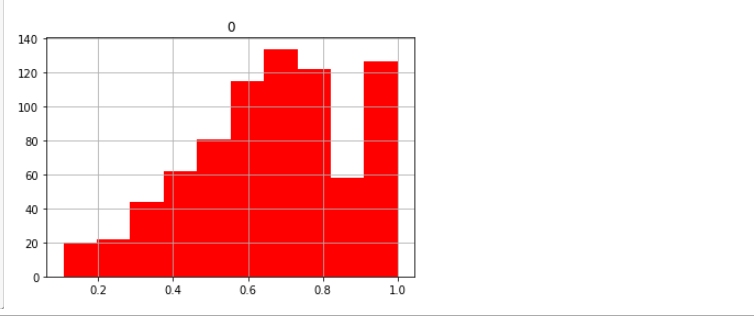


Figure : Similarity for customers category "c"

|  |  |
| --- | --- |
| 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.

### Comments sentiment analysis

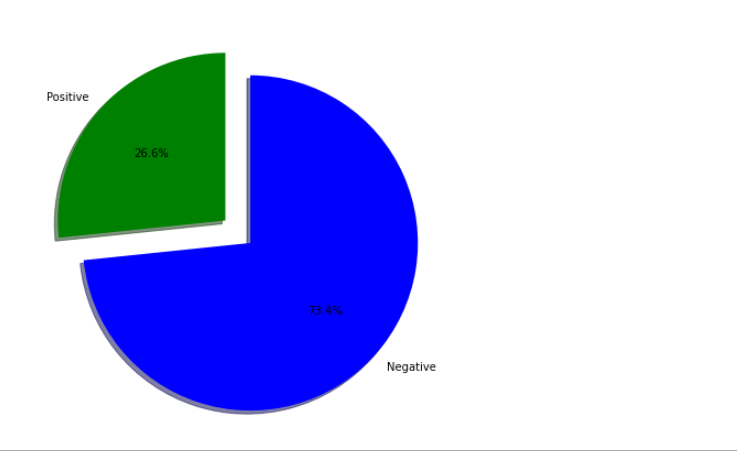


Figure : % of Negative and Positive comments

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”.



Figure : 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.

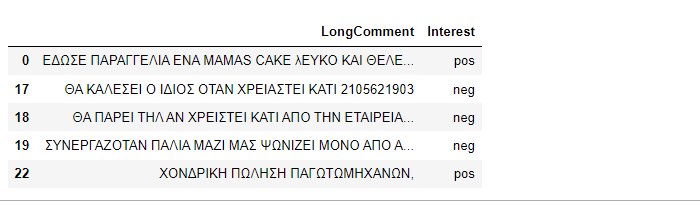


Figure : Final form of dataset that is used for prediction model

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

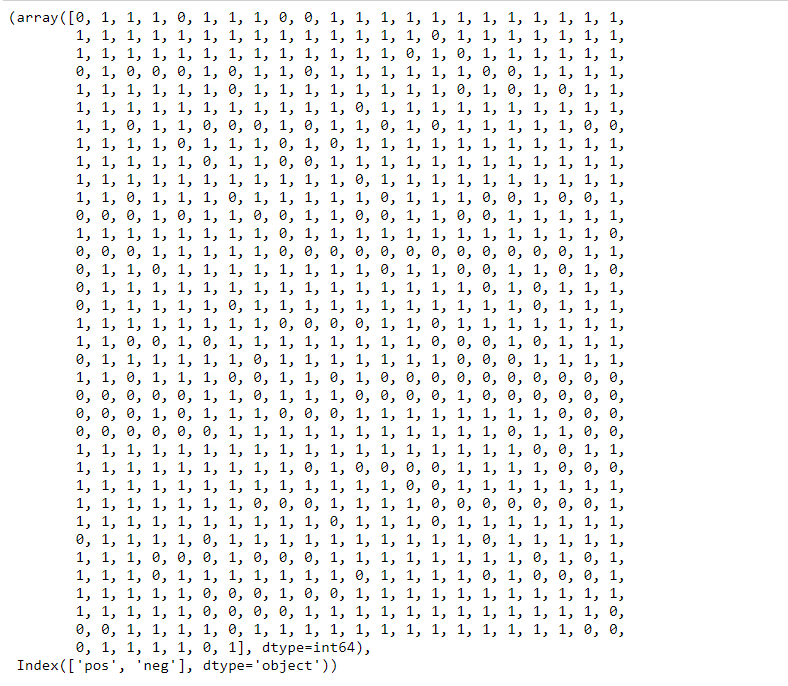


Figure : Factorization of responses

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

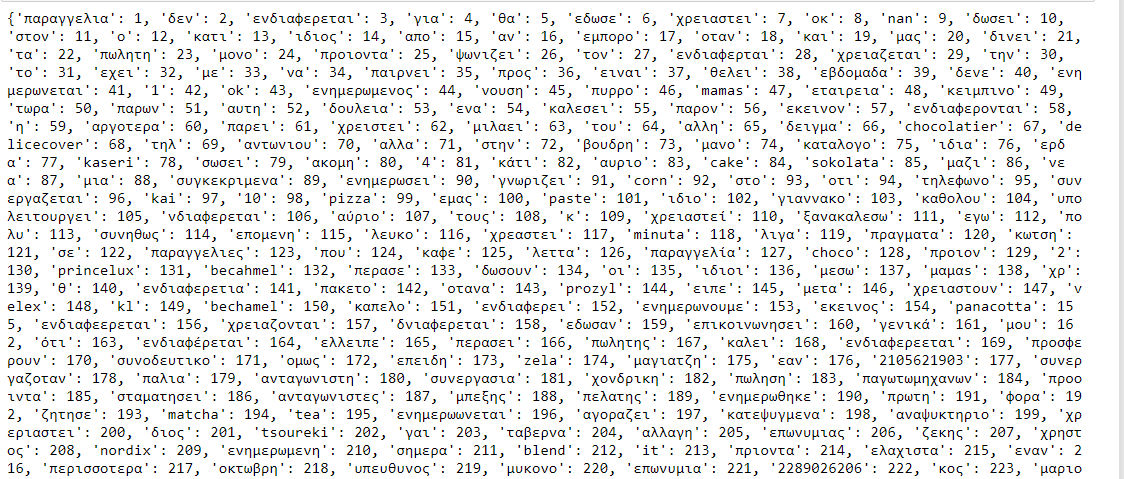
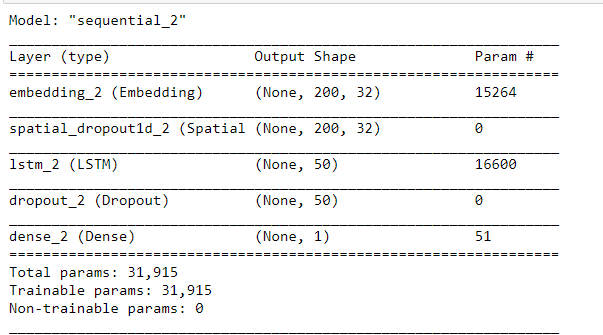


Figure : Tokenization of comments texts

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



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

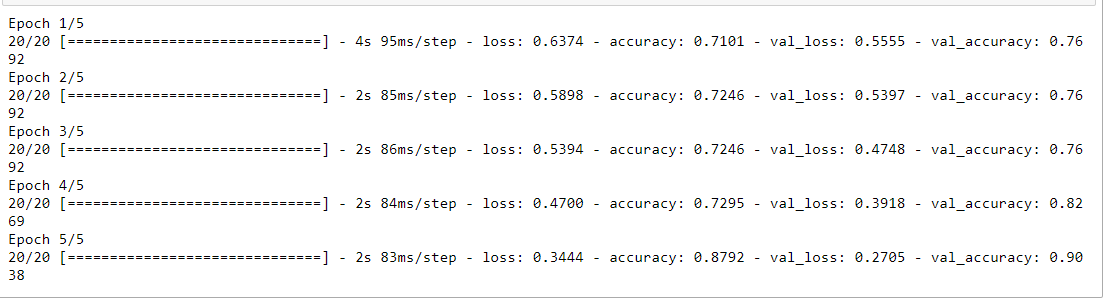


Figure : Prediction model accuracy

Accuracy ratio seems promising and according to our expectations.

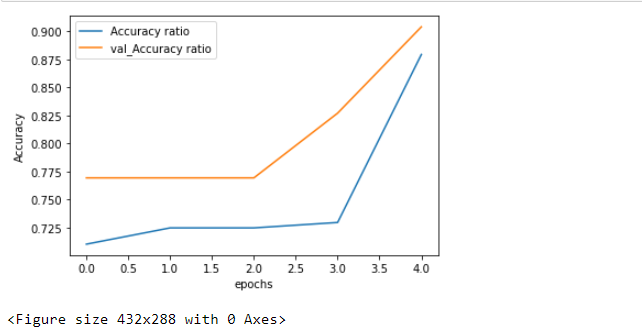


Figure : Validation and actual accuracy ratios

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.

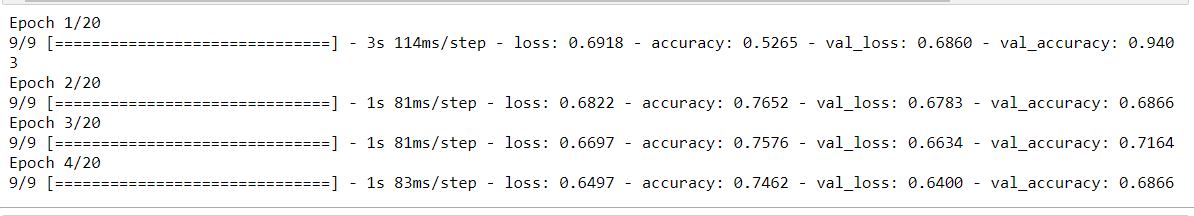


Figure : Prediction model with overfitting check

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/