Dimitris Mantaos

P2822011@aueb.gr

Abstract

*Scope of this project is the implementation of ML models that will be 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 and are extracted from entity’s data base.*

ML and content analysis

Final project / The opposites team

***Table of contents***

Table of Contents

[Introduction and project goals – project members 2](#_Toc81572289)

[Datasets overviεw 3](#_Toc81572290)

[Delayed payments project 3](#_Toc81572291)

[COVID crisis effect 4](#_Toc81572292)

[Comments sentiment analysis 5](#_Toc81572293)

[Data Visualisation and transformation 7](#_Toc81572294)

[Project results 10](#_Toc81572295)

[Delayd payments 10](#_Toc81572296)

[covid effect on market basket 15](#_Toc81572297)

[Comments sentiment analysis 19](#_Toc81572298)

[Discussion on lessons learned and future steps 25](#_Toc81572299)

[Bibliography and resources 26](#_Toc81572300)

# Introduction and project goals – project members

Current study consists of three subprojects as follows:

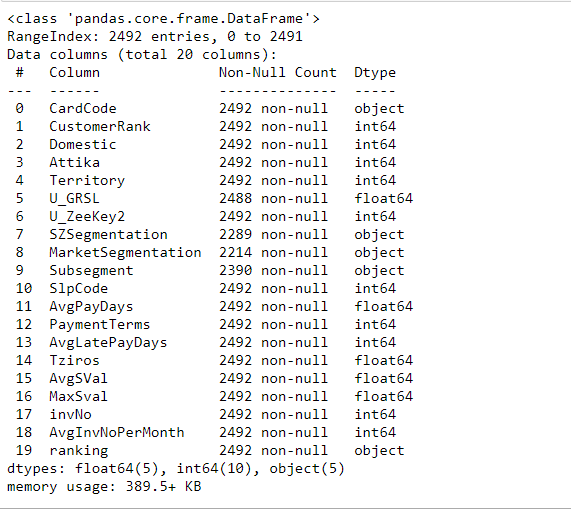
* Analyse and predict credit behaviour of customers related especially to payment delays. So, this model intends to safely predict bad debts based on customers characteristics
* Calculate the COVID effect (if any) to customers purchases behaviour during the COVID period. For this reason, we will try to find similarity ratio for the market basket of each customer, and on main segments
* 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.

Project initiated with 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 wade just by Dimitris Mantaos.

# Datasets overviεw

## Delayed payments project

The Dataset used for payments analysis consists on 19 main customers quantitative and qualitative characteristics, and has 2.492 observations. Column 0 is just the code of the customer in the database, so it is omitted on any analysis that will be performed

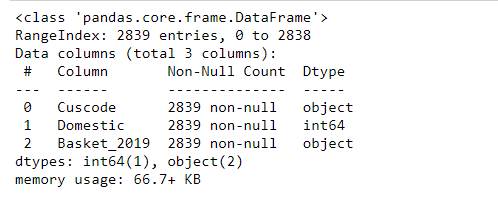


Source file is in excel format and is named “Customers profile” . Below we give a description for each feature.

|  |  |
| --- | --- |
| ***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.

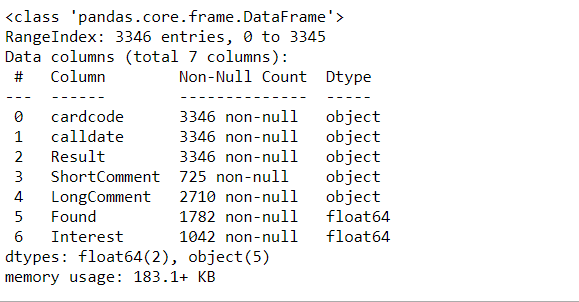


|  |  |
| --- | --- |
| ***Column name*** | ***Description*** |
| Cuscode | Customer internal code |
| Domestic | Local or export customer |
| Basket\_2019 or Basket\_2020 | Top products purchased |

Data set consists of 2.839 unique customer codes.

## Comments sentiment analysis

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

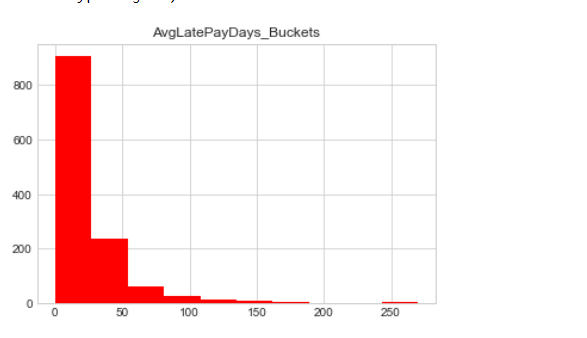


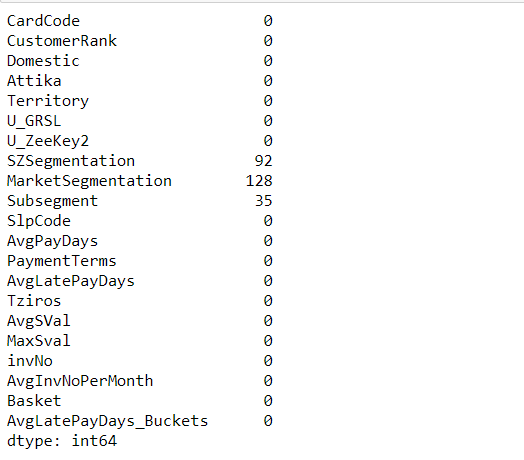
Data set has 3.346 records, but there are some missing values which will be cleaned during the process.



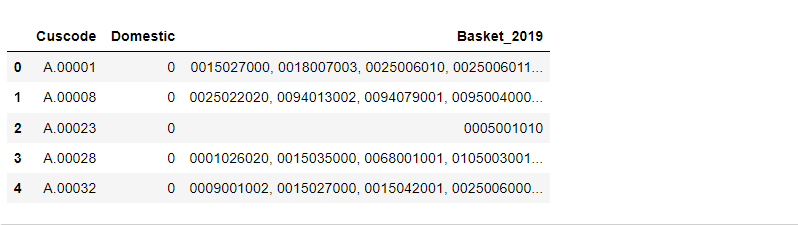
|  |  |
| --- | --- |
| Column name | Description |
| Cardcode | Customer internal code |
| Calldate | Date of the call |
| Result | Outcome of the contact |
| ShortComment | Predefined comment |
| Long comment | Free text |
| Found | Check point if the concact was made |
| Interest | Field used for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

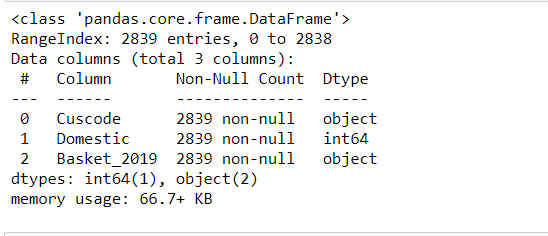
# Data Visualisation and transformation

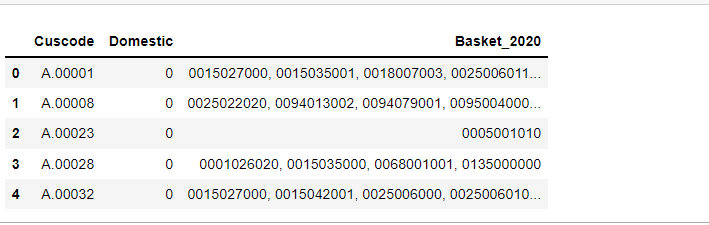


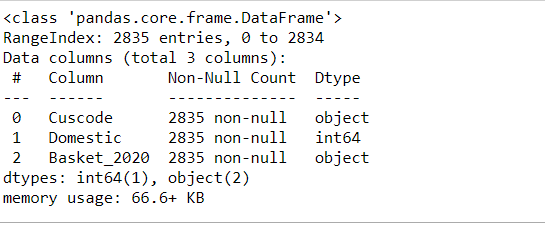
We notice that most of the delays are concentrated in the range between 0 and 60 days. Presence of delays more than 60 days is limited.

Data set contains missing values which we removed.







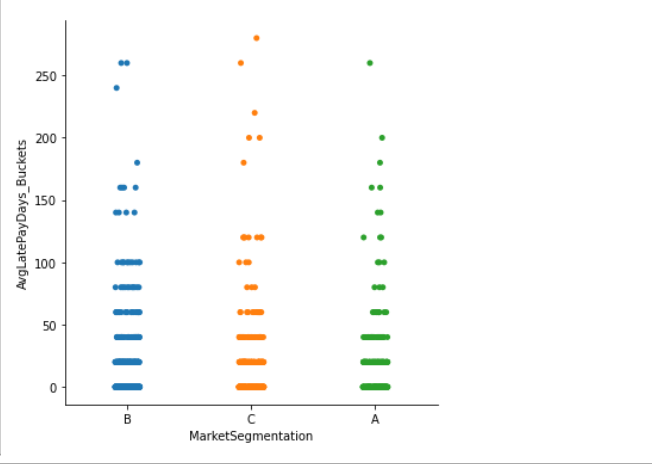


Data sets differ just by 4 customers, which possibly are accounts that do not exist in year 2020 , but were present during Year 2019

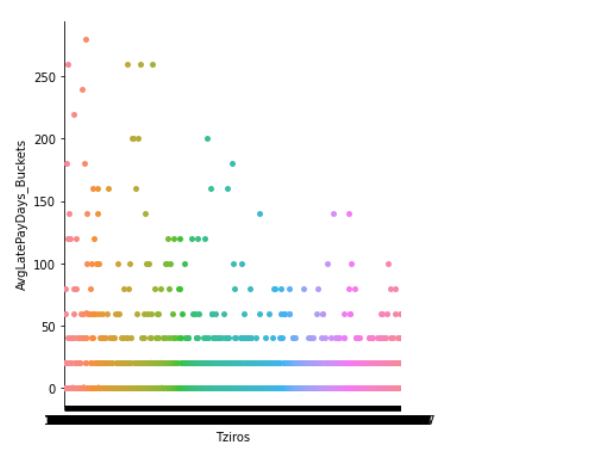
# Project results

### Delayd payments

First we use visualisations to investigate possible relations between delay in payments and customer characteristics. For this reason we used the field “Tziros”, “MarketSegmentation”, and “PaymentTerms”



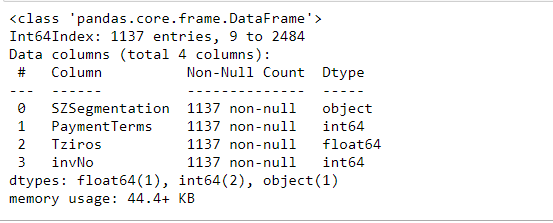
We notice that there is not relation between lateness in payments and Market Segment.

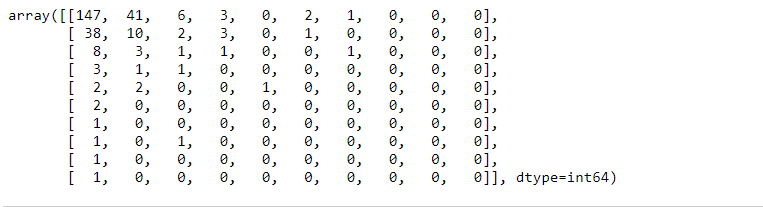


We perform the same analysis comparing turnover with late payments. From visualization we come to the same result. There is not relation with the level of sales to late payments.

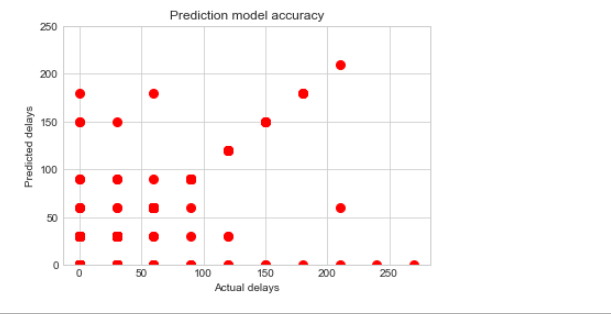
From these exhibits we assume that is difficult to find a prediction model with high accuracy ratio.

After a lot of experiments we chose a decision tree model as the best option for this case . Then we group days of delay in buckets of 20. After running the model o lot of times the following features chosen for the model construction



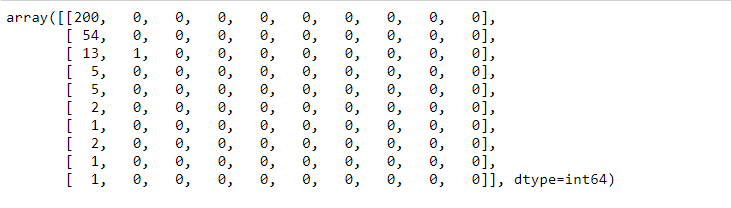


We calculated the accuracy ratio and got as a result using sklearn.metrics Accuracy score= 0.5578947368421052.



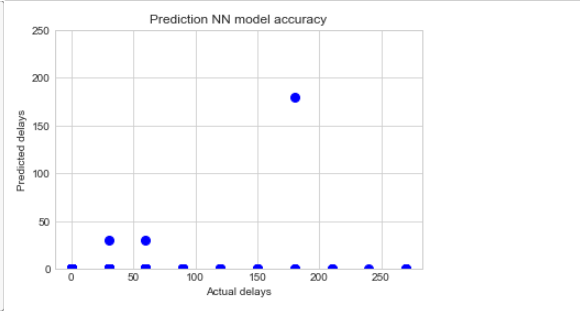
As seen in exhibit above, model performs better when delays are high, whilst it has poor performance in the lower parts.

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



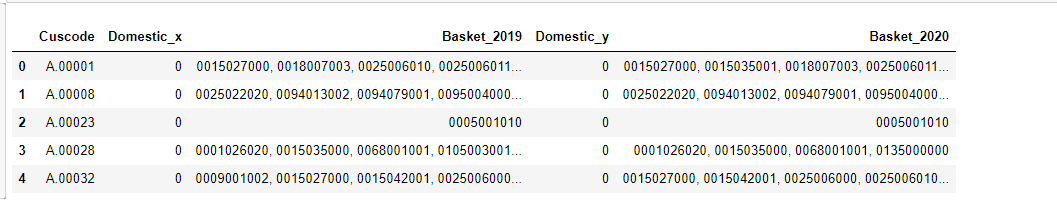
Furthermore MLP classifier performed better in the lower bucket of days (0-30) but failed totally in the upper levels as shown in exhibit below

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

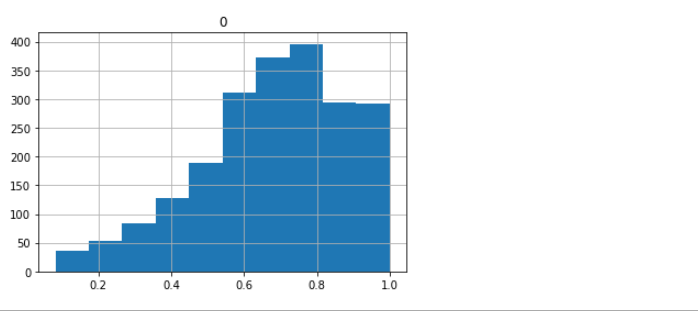


A question rised is if we could use a combination of models , but this was left for later .

### covid effect on market basket



To evaluate the similarity between Basket\_2019 and Basket\_2020 we use from difflib import SequenceMatcher.

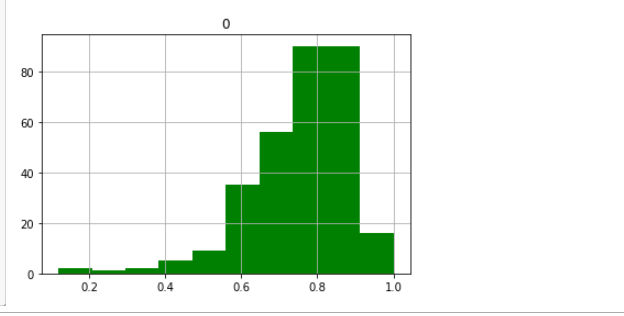


First we get the similarity ratio for the full dataset . Main statistics are :

|  |  |
| --- | --- |
| 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 :

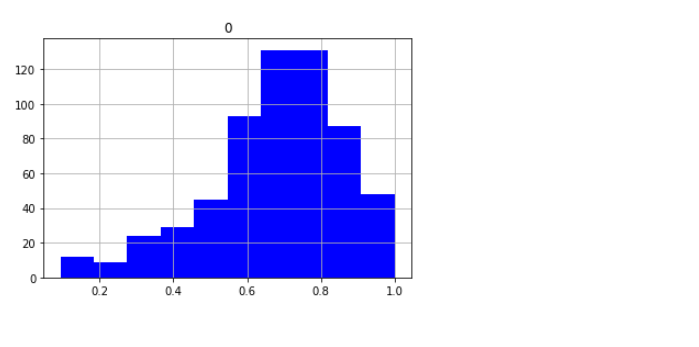
Customers a



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

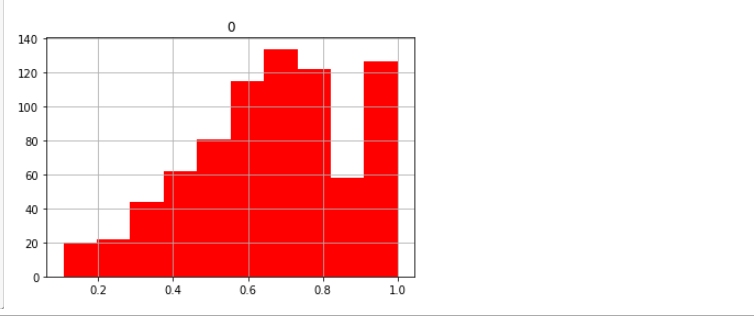
Customers o category “a” have higher similarity ratio and values are more concentrated around the mean than the other categories

Customers “b”



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

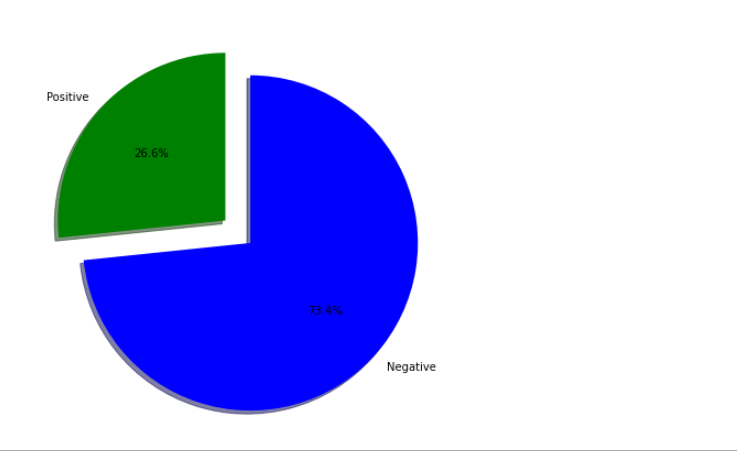
Customers “c”



|  |  |
| --- | --- |
| 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”.

### Comments sentiment analysis



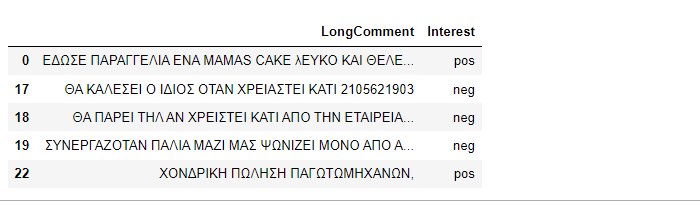
From dataset we calculate that Negative responses are 73,4% of the sample. This is fully explainable since calls were made to non active or old customers mainly of category c.

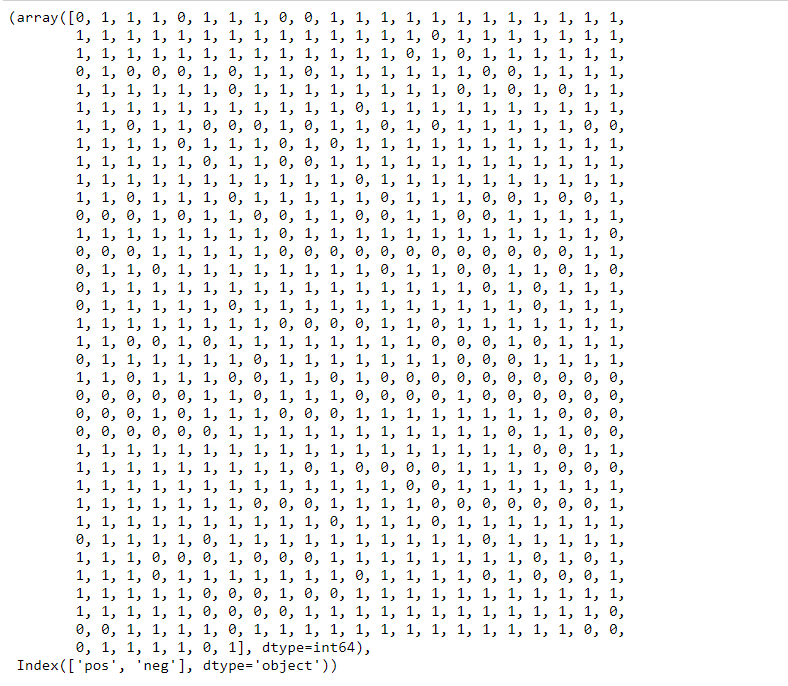
From the initial dataset we kept the columns “LongComment”, “Found”, “Interest”.

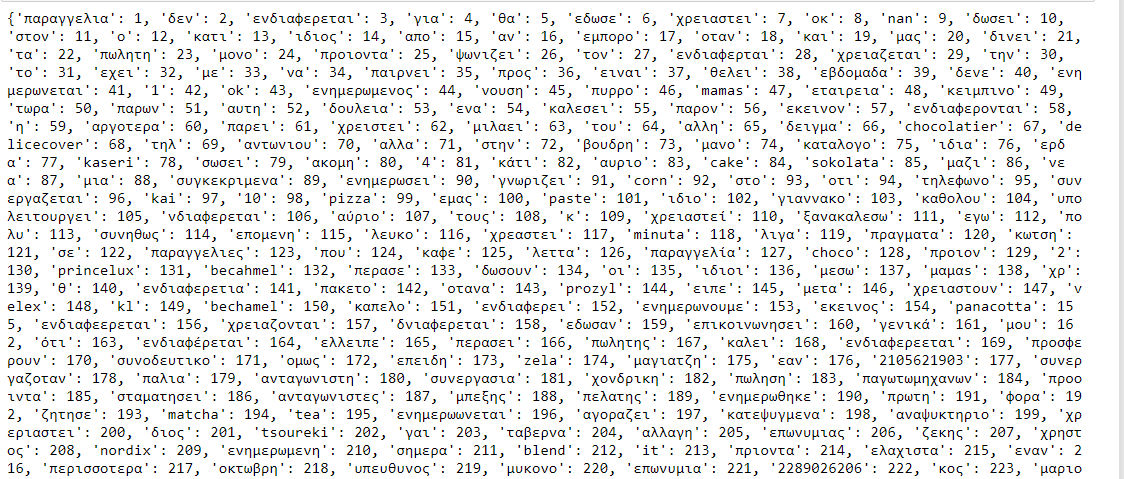


Then we dropped and the rows where value of Found equals to 1. In the opposite cases customers were not available for a phone call

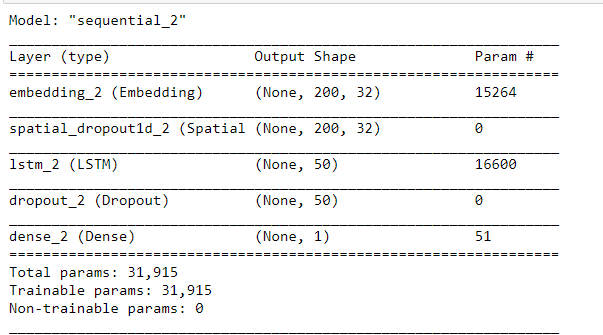
After we changed responses from -1 to “neg” and 1 to “pos”, and we discard observation with value 0 , since these comments were considered neutral.



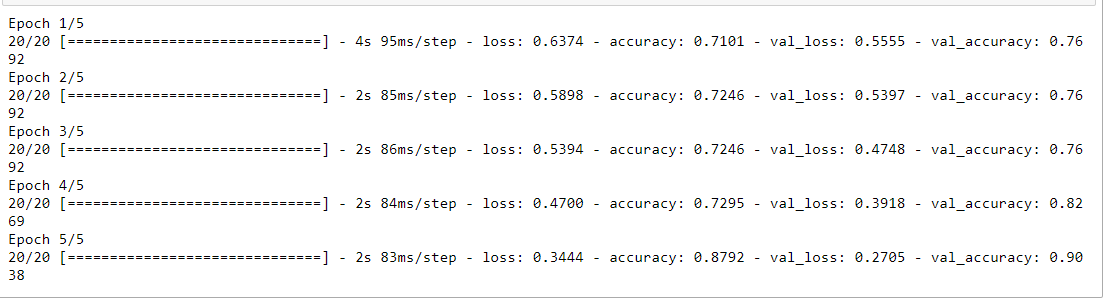


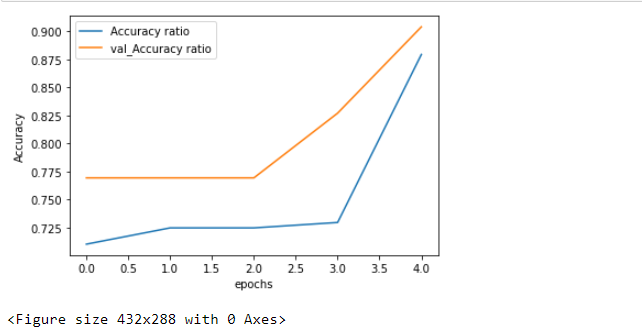


After all the above we were ready to build and fit a neural network model. We choose LSTM with the parameters as shown below :



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





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.

# 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. So 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.

# 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

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

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

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