Ecole Supérieure Privée
d'Ingénierie et de Technologies

SATISFACTION STUDY: TELECOMMUNICATION REPORT

DATA SCIENCE PROJECT

realized By :
Belkis Baccar
Donia Ksiaa
Issam Ben Moussa
Latifa Sassi
Med Khalil Chakroun
Med Fadhel Shel

CONTENTS

LIST OF FIGURES

LIST OF FIGURES

05

GENERAL INTRODUCTION

Data volumes have skyrocketed. More data were generated in the last two years than in the entire human history before that.
Big data holds the key to an amazing future. It reveals patterns and connections that significantly improve our lives. That's why the use of Data Science in the business context is constantly growing, the need of setting a business-oriented strategy is becoming a very important step in order to maximize profits, prevent loss and guarantee productivity.

Client satisfaction is the dilemma of companies in the 21 century; the consistent search for landing new clients, keeping the company's standards high and insuring satisfaction amongst customers is the new currency. Therefore, data science is the key to accomplish such understanding of client's behavior and wants. This powerful tool, a combination of statistics and IT , gives companies the ability to triumph.
It allows them to create insightful analysis or predictions on any type of data, it could reveal patterns that are hidden to the human understanding.

This research aimed to find out critical factors which mostly influence the client's satisfaction towards mobile operator services in Tunisia.
If the mobile operators such as Orange focus on those factors and improve those who are lacking they could sustain the market and make more profits.

BUSINESS UNDERSTANDING

Nowadays, with mobile telecommunication being the most relevant way of communication across the globe, Telecommunication companies find themselves facing a huge problem of market saturation which makes finding a new customer a much harder and costly task than keeping returning customers. This issue made the rivalry between companies intense and advanced due to the latest technologies used to predict accurate models of customer behavior and target the weak links in a company's client base. The investment in customer satisfaction issue and taking a close look into customer behavior is a prior necessity in the telecommunication industry as it could end up affecting the revenue numbers and influence policy decisions by the impact that could one customer apply to another to quit as well as adding extra costs especially in advertisement when keeping old subscribers is already cheaper.

Our main aim is studying the clients satisfaction regarding the services offered by Orange in comparison with with it's rivals.

BUSINESS UNDERSTANDING

1. *Business Objectives :*

Due to the ascending rivalry in the operational sector, it's becoming critical to study the satisfaction of clients regarding the services.
This study sights to analyze the customer's behavior through :

- Extracting divers sources of data from social media , vocal recordings, forms ...

- Storing the data in a data warehouse

- Modeling the architecture of the data warehouse

- Plotting Graphs visualization representing the satisfaction of the clients based on different axes

- Identifying satisfaction by different factors

- Extracting the external Data from different sources social media and vocal recording

08

ANALYTIC APPROACH

Our main goals are determining if our client is satisfied and for what reasons.
We can derive many insights from our data in order to answer those questions, like conducting descriptive analysis such as clustering in order to group similar individuals to show relationships and predictive analysis using classification in order to answer if a customer is satisfied or not.

On the other hand, social media provides a lot of feedbacks, like comments or interviews. So approaches such as Natural Language Processing, Face Emotion Recognition and Sound Emotion Analysis would be great assets.

The achievement of this project involved a series of steps and activities that have been managed in order to avoid delays, development problems and other issues.

09

DATA REQUIREMENT

The aim is the study of customer satisfaction from different aspects ranging from the services he uses to the emotions he expresses, thus the data required are his likeness to quality of the services he uses and his opinions towards the operator he is currently dealing with. So the data must present in different forms the satisfaction of every client on every services and the description of his usage of these services. These forms of data can be generated from forms that are sent to many clients to fill or data generated from public forms of satisfaction like forms created by national survey enterprises or the ones done by operators enterprises themselves.
For the study of satisfaction dealt by emotions we have multiple: Voice, Face and speech. The first one needs customer's recordings, second data in form of images and lastly texts derived from customer's voice recording.

DATA COLLECTION

We got a "Quality of Service business to business survey" done by Orange in 2019 in order to study the satisfaction of its clients compared to other operators (Ooredoo and Tunisie Telecom) .
We started by creating a diagram that describes our data perfectly and allows us to store it in a structured data Base 'SQL Server'.

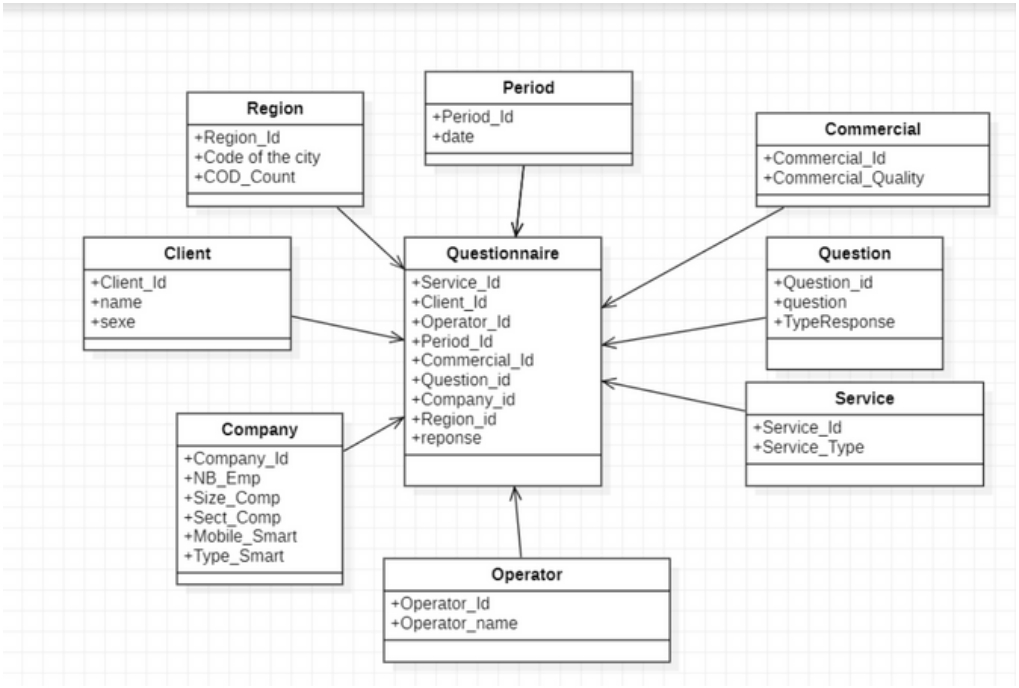


Figure 4.1: Data Model

DATA COLLECTION

Then we stored the data into a structured DataMart Using the 'SQL Server Integration Services'.
We started by extracting the data into a staging area and from it we filled our fact table and dimension tables.

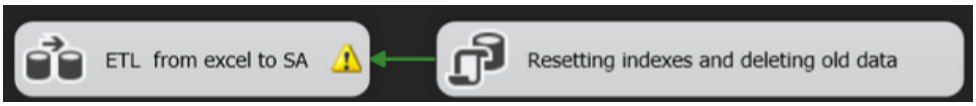


Figure 4.2: Staging Area Storing

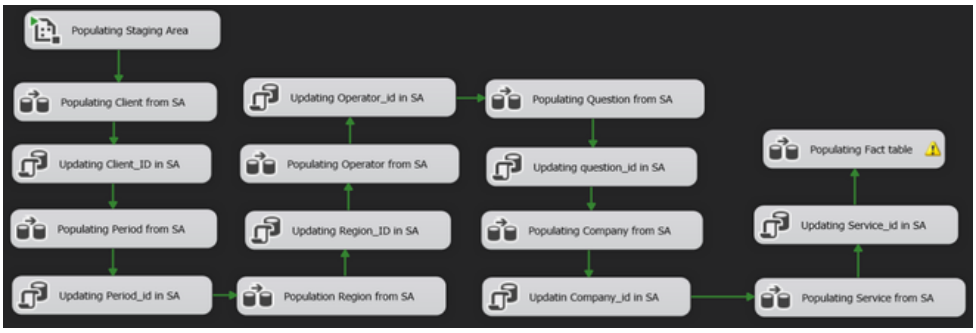


Figure 4.3: Data Integration Model

But relying on only one source of data is a mistake, that's why we scrapped data like comments and reviews from different social media outlets such as Youtube , Facebook and Instagram , because Tunisian mobile operators have really active channels. And since this data is not structured, we stored it in MongoDB.

DATA UNDERSTANDING

The Orange Survey provides answers to more than 130 questions related to different services that it and it's rival Ooredoo provide.
It englobes a lot of strong factors such as quality of service, price and ease of use.
It also provides additional information on the clients filling the form , the enterprises they work in , their regions ,etc...
Besides, The external data complements the internal data by depicting general thoughts of customers, feedbacks and reviews which can't be processed like the previous data due to the difference in structure so it must be processed differently from storing to modeling . It also includes information on other operators such as Telecom Operator.
As these different sources of data provide useful different information , this raises the need for using both of them for our models, each type of data will be useful for creating a model: the internal data will help us predict the satisfaction of users and rate the services of the operators by visualizing them, on the other hand the external data will be beneficial in predicting the satisfaction from videos , vocals and images as along as creating a fully automated chatbot that can provide all the needed information to the users and recommend services for them.

13

DATA PREPARATION

INTERNAL DATA

After finishing with data collection, we started with visualizing the data to understand the problems in it to start the cleaning phase . At first we found out that our data contains a non regular shifting : a shifting that isn't formal so every line and it's own shifting . We found out that this problems is due to filling the datasheet regularly by humans and it's due to filling a number of different choices for some questions , so the shifting is proportional with the number of additional choices in those questions . Besides the some values were stored in 2 cells due to clicking twice on the "Tab" button .So the first step is to correct the shifting by creating new columns for each choice of the features representing the multiple choice questions.

```
dchoices = {'Dans la presse' :'', 'sur le site internet':'' , 'en boutique':'' , 'via bouche a oreille':'' ,
   'via un moteur de recherche':'' ,'en appeleant le service':''}
dataChoices = pd.DataFrame(dchoices)
dataChoices
```

Figure 6.1: Multi choice columns Creation

```
j=0
for i in range(1418):
    dchoices['Dans la presse'][i]=0
    if data['40 - [49] Q05B. Par quels principaux moyens avez-vous trouvé les informations que vous recherchiez ?']
        dchoices['Dans la presse'][i]=1

    dchoices['sur le site internet'][i]=0
    if data['40 - [49] Q05B. Par quels principaux moyens avez-vous trouvé les informations que vous recherchiez ?']
        dchoices['sur le site internet'][i]=1
```

Figure 6.2: Multi choice columns filling

14

DATA PREPARATION

INTERNAL DATA

Additionally, we found out that many features are unnecessary holding unneeded information for our models and visualization. So we dropped them on the floor.

```
df=df.drop(["116 - [143] Q29. Dans les 6 derniers mois, avez-vous eu besoin d'aide auprès de [V15] concernant les offres souscrites ?",
            "117 - [144] Q31. Et pour cette démarche vous avez : ",
            "118 - [145] Q32.1 La compréhension de votre demande ou réclamation",
           "119 - [146] Q32.2 La rapidité de traitement de votre demande ou réclamation",
            "120 - [147] Q33. Après cette démarche, avez-vous dû recontacter un commercial pour la même demande / le même problème ?",
            "121 - [148] Q34. Vous avez appelé le service client. Avez-vous réussi à parler à un conseiller ?",
            "122 - [150] Q35.1 …Le temps d'attente avant de parler à un conseiller",
           "123 - [151] Q35.2 …L'accueil téléphonique, l'amabilité de votre interlocuteur …",
           "124 - [152] Q35.3…La compréhension de votre demande / réclamation",
           "125 - [153] Q35.4 La solution proposée",
            "126 - [154] Q35.5 La résolution de votre réclamation / demande lors de cet appel",
```

Figure 6.3: Dropping Columns

Then we proceeded with correcting the data types of the feature "9 S1. Taille de l'entreprise" and change the data type into string one representing if the enterprise's employees number is between 0 and 24 as this information was represented by a dataFrame.

```
: #column '9 S1. Taille de l'entreprise' changing values
  c="9 S1. Taille de l'entreprise"
  df[c].unique()
  df[c]=df[c].replace(datetime.datetime(2017, 10, 24, 0, 0),"0-24")
  df[c].unique()

: array(['0-24', '25-50', nan], dtype=object)
```

Figure 6.4: Replacing Values

DATA PREPARATION

INTERNAL DATA

For the next step, we continued with correcting the typos or misspellings in every feature separately in order to make sure to correct all misspellings as they represent different values to our models even though they hold the same information.

```
#column '4 - [12] S0.2. Sexe :' changing values
df['4 - [12] S0.2. Sexe :'].unique()
df['4 - [12] S0.2. Sexe :']=df['4 - [12] S0.2. Sexe :'].replace('1. Homme','Homme')
df['4 - [12] S0.2. Sexe :']=df['4 - [12] S0.2. Sexe :'].replace('1. HOMME','Homme')
df['4 - [12] S0.2. Sexe :']=df['4 - [12] S0.2. Sexe :'].replace('HOMME','Homme')
df['4 - [12] S0.2. Sexe :']=df['4 - [12] S0.2. Sexe :'].replace('HOM%ME','Homme')
df['4 - [12] S0.2. Sexe :']=df['4 - [12] S0.2. Sexe :'].replace(0,'Homme')
df['4 - [12] S0.2. Sexe :']=df['4 - [12] S0.2. Sexe :'].replace('2. Femme','Femme')
df['4 - [12] S0.2. Sexe :']=df['4 - [12] S0.2. Sexe :'].replace('2. FEMME','Femme')
df['4 - [12] S0.2. Sexe :'].unique()
```

Figure 6.5: Replacing Values-Sex Column

```
#column c changing values
c='21  Q01A. Recommanderiez-vous les services de téléphonie mobile de'+
'votre opérateur principal à vos collègues ou partenaire commercial ?'
df[c].unique()
df[c]=df[c].replace('1 (Ne recommanderait certainement PAS)',1)
df[c]=df[c].replace('10 (Recommanderait certainement)',10)
df[c]=df[c].replace('11 (Ne sait pas/Non applicable)',11)
df[c].unique()
```

```
array([ 6.,  5.,  3.,  1.,  4.,  8., 10.,  7.,  9., 11.,  2., nan])
```

Figure 6.6: Replacing Values-Note Column

DATA PREPARATION

INTERNAL DATA

Furthermore , we filled missing values with the -1 value and we renamed our features with appropriate names for an easier understanding and to ease using them in the modeling and the questions categorization .

```
df.rename(columns={"89 - [193] Orange money":"89 Raison pour utilise Orange money"})
df.rename(columns={"90 - [194] Mobicash (ooredoo)":"80 Raison pour utilise Mobicash Ooredoo"})
df.rename(columns={"91 - [195] Mdinar (TT)":"91 Raison pour utilise Mdinar TT"})
df.rename(columns={"92 - [117] Orange Money":"92 facilite d'utilisation et fonctionnement Orange Money"})
df.rename(columns={"93 - [118] Mobicash (ooredoo)":"93 facilite d'utilisation et fonctionnement Mobicash (ooredoo)"})
df.rename(columns={"94 - [119] Mdinar (TT)":"94 facilite d'utilisation et fonctionnement Mdinar (TT)"})
```

Figure 6.7: Renaming Columns

For the questions categorization , we transposed all columns names , generated 4 new columns : 'service name', 'operator', ,'type question', 'type response' where we store the service of the question , the operator if it's a question about a specific operator or if it's a global question , the question type and the response type, if it's a yes or no, numerical or a choice question.

	questions	service name	operator	type question	type reponse
0	16 S6. Depuis combien de temps votre entrepri...		operateur principale	specifique	String
1	17 S7 Pour Orange, quel énoncé décrit le mieu...		Orange	specifique	String
2	18 S7Pour Ooredoo, quel énoncé décrit le mieux...		Ooredoo	specifique	String
3	19 S7 Pour Tunisie Telecom, quel énoncé décri...		Tunisie Telecom	specifique	String
4	20 S8. Quel type d'offre votre entreprise ou ...		operateur principale	specifique	String
...	...	...	...	...	...
122	151 Q46. Pour quel autre opérateur principal c...		operateur principale	specifique	String
123	153 Prénom		operateur principale	specifique	String
124	154 Nom		operateur principale	specifique	String
125	155 Numéro de téléphone		operateur principale	specifique	String
126	156 Nom de l'entreprise		operateur principale	specifique	String

Figure 6.8: Questions Table

DATA PREPARATION

EXTERNAL DATA

Our critical objective is to look into different data provided from different sources for an effective satisfaction analysis, thus internal data stays insufficient to make final insights.
As a result we scrapped client's comments on posts belonging to different operators :

- using a **google extension** able to extract targeted data and make a structured output, besides to Youtube we used a **pre-implemented API**.

Moreover, we implemented a python script that aims to convert the **speech to text** :

- we used it to convert vocal samples.

To improve our data, we made a python script to prepare extracted data:
most of the comments came in Tunisian dialect

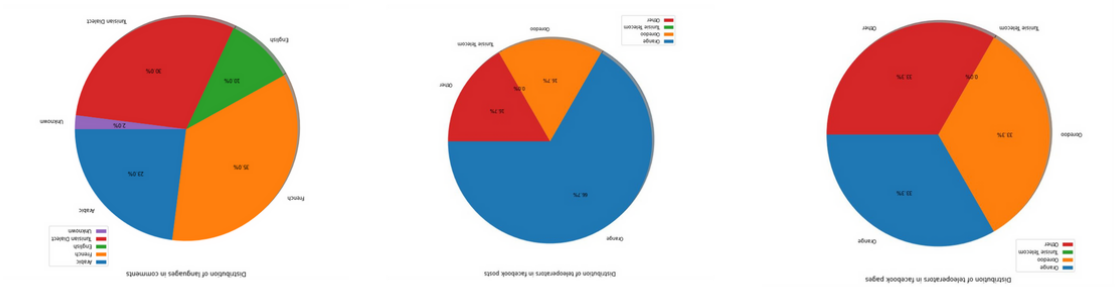


Figure 6.9: Comments Languages

18

MODELING AND EVALUATION

INTERNAL DATA

Clustering and classification

Now our data is ready to be processed and analyzed. In our case, since our data is a survey about quality of service, we would like to deduct if our client is satisfied or not.
We used unsupervised and supervised machine learning to achieve our goal.

- Unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention.
- Supervised machine learning is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately, we used classification.

For our purpose, we proceeded using two steps :

- We calculated the client's satisfaction using a formula which is the average of marks given to different services. in order to confirm it's firmness, we applied clustering.
- We applied a classifier to predict futur clients satisfaction.

19

MODELING AND EVALUATION

INTERNAL DATA

Formula approach

The formula led to these results :

```
for k,v in d_cluster.iterrows():
    if(v.count()>0):
        if (v.sum()/v.count())>=5:
            sat.append(1)
        else:
            sat.append(0)
```

Figure 7.1: Satisfaction formula

```
Satisfied with formula :  851
Unsatisfied with formula :  559
```

Figure 7.2: Total Counts

20

MODELING AND EVALUATION

INTERNAL DATA

Clustering

We used Agglomerative clustering, which is a from of hierarchical clustering that builds nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample.

Agglomerative clustering is useful in our case because :

- It can generate many clusters, we need two.
- It deducts connectivity constraints, the marks can be connected.
- It calculates non Euclidean distances, and it is irrelevant for us.

We also used the ward method for linkage because minimizes the sum of squared differences within all clusters.
The cophenet distance (the distance between two disjoint clusters under a parent cluster) equals to 0.649 which is a good score considering the variety of answers and the missing data.

MODELING AND EVALUATION

INTERNAL DATA

Clustering

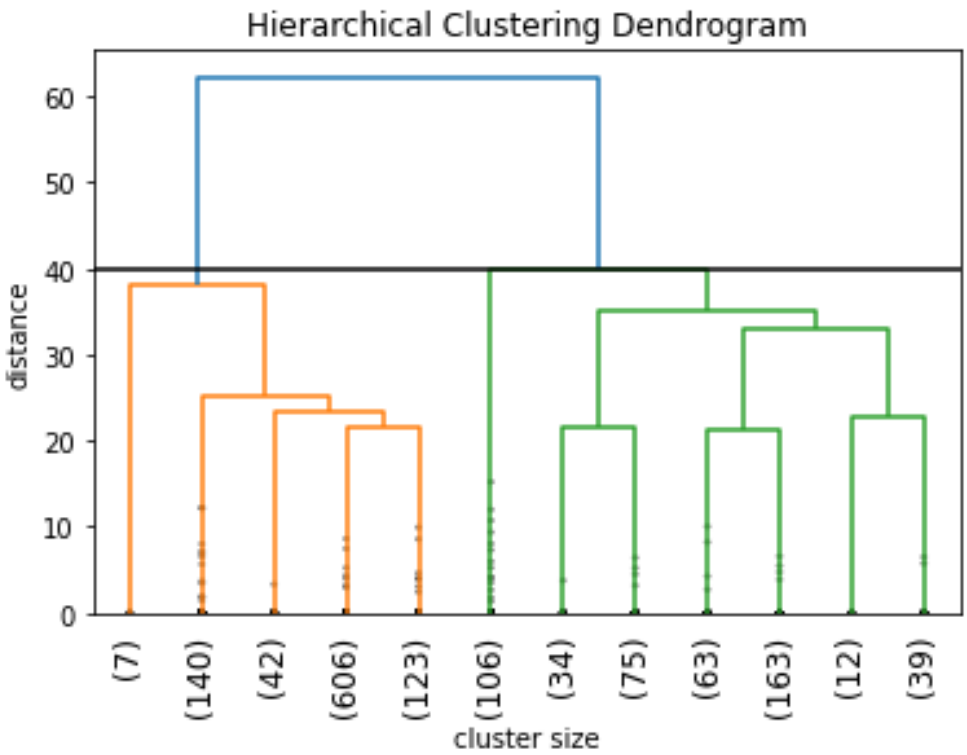


Figure 7.3: Dendogram

With a distance 40, we divided our data in two major clusters.

22

MODELING AND EVALUATION

INTERNAL DATA

Clustering

If we plot points for two service axis, we can see a pattern

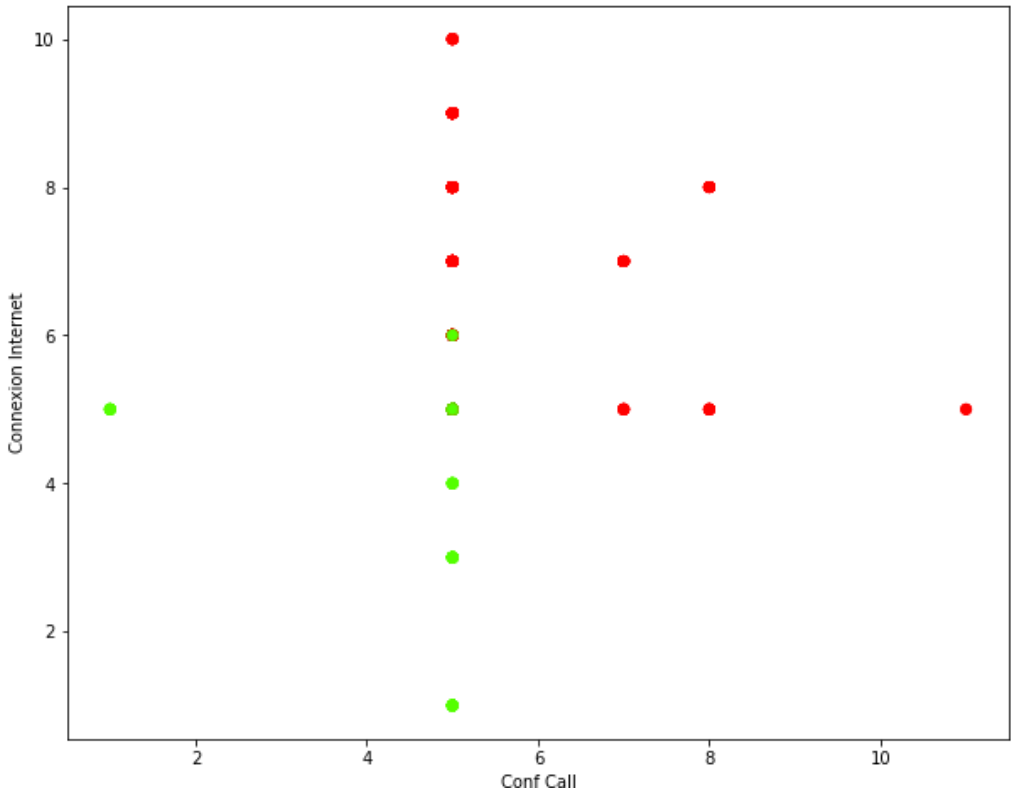


Figure 7.4: Clusters

The green points seem to be low for both services unlike the red ones that are high for both, from this visualisation we can deduct that the green cluster are unsatisfied clients and the red one are satisfied clients.

23

MODELING AND EVALUATION

INTERNAL DATA

Clustering

No we are going to compare our two approaches for determining satisfaction

```
clusters = fcluster(Z, max_d, criterion='distance')
for i in clusters:
    if(i==1):
        s_sat+=1
    else:
        s_unsat+=1
c_sat=len(d_cluster[d_cluster['satisfied']==1])
c_unsat=len(d_cluster[d_cluster['satisfied']==0])
print('Formula :')
print('Satisfied :',c_sat,'Unsatisfied :',c_unsat)
print('Clusters :')
print('Satisfied :',s_sat,'Unsatisfied :',s_unsat)
```

```
Formula :
Satisfied : 851 Unsatisfied : 559
Clusters :
Satisfied : 918 Unsatisfied : 492
```

Figure 7.5: Formula and Clusters comparaison

Our formula is 95.2% accurate

MODELING AND EVALUATION

INTERNAL DATA

Classification

Since we have a lot of categorical features, we need to calculate the correlation between these features and our target so our data doesn't explode in dimensions.
Since our output is binary, we used the Chi² independency test.
We create the contingency matrix for each feature and compute it's p-value.
Consider this example in our data

Figure 7.6: Contingency table for sex feature

The chi² independency test relies on comparing the observed values against the expected ones with the following coefficient
coef = (n(i,j) - f(i)*n(j))² / (f(i)*n((j)))
The p-value is the sum of that coefficient, if p-value exceeds alpha (1% in our case), it means that this feature doesn't influence the result, thus we can drop it.

MODELING AND EVALUATION

INTERNAL DATA

Data preparation for models

Before we can apply any classifying algorithms, we have to prepare our data, these are the steps :

1. Divide data into train and validation
2. Encode categorical features using One Hot Encoder
3. Impute missing values with the mean
4. Divide train data into train and test
5. Scale data

The reason why scaling must be done after splitting data intro train and test is that there shouldn't be any external interference.

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

Figure 7.7: One Hot encode

$$\frac{x_i - Q_1(\boldsymbol{x})}{Q_3(\boldsymbol{x}) - Q_1(\boldsymbol{x})}$$

Figure 7.8: Robust scaler formula

MODELING AND EVALUATION

INTERNAL DATA

Classification

We applied various classification algorithms starting with : **KNN**
K Nearest Neighbor(KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. It is based on feature similarity approach. It calculates the distance between our sample to predict and existing samples, finds the closest K samples and then votes for it's label.

Pros ✅

- The training phase of K-nearest neighbor classification is much faster compared to other classification algorithms
- KNN can be useful in case of nonlinear data.
- Simple and easy to understand

Cons ❌

- KNN is not suitable for large dimensional data.
- It requires large memory for storing the entire training dataset for prediction.
- It requires scaling of data because it uses distances between two data points to find nearest neighbors. Distances are sensitive to magnitudes.

27

MODELING AND EVALUATION

INTERNAL DATA

Classification

Decision Tree Classifier

A decision tree is a flowchart-like tree structure where an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning.

Pros

- Decision trees are easy to interpret and visualize.
- Requires little data preparation. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed. However that sklearn does not support missing values.

Cons

- Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This problem is mitigated by using decision trees within an ensemble.
- Decision-tree learners can create over-complex trees that do not generalise the data well.

28

MODELING AND EVALUATION

INTERNAL DATA

Classification

Random Forest Classifier
A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Pros ✅

- Random forests is considered as a highly accurate and robust method because of the number of decision trees participating in the process.
- No overfitting problem
- You can get the relative feature importance, which helps in selecting the most contributing features for the classifier.

Cons ❌

- Random forests is slow in generating predictions because it has multiple decision trees.
- The model is difficult to interpret compared to a decision tree, where you can easily make a decision by following the path in the tree. You have to select your based on a particular score.

MODELING AND EVALUATION

INTERNAL DATA

Classification

XGBOOST

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance that is dominative competitive machine learning. Gradient boosted trees use regression trees in a sequential learning process as weak learners. These regression trees are similar to decision trees, however, they use a continuous score assigned to each leaf (i.e. the last node once the tree has finished growing) which is summed up and provides the final prediction.

Pros ✅

- Lightning speed compared to other algorithms
- Great ability to generalize
- One of the best algorithms when speed as well as high accuracies are of the essence

Cons ❌

- XGBoost is more difficult to understand, visualize and to tune compared to other algorithms because it requires a lot of tuning and needs more time and expertise from the user

MODELING AND EVALUATION

INTERNAL DATA

Classification evaluation and results

A confusion matrix is an N X N matrix, where N is the number of classes being predicted. For the problem in hand, we have N=2, and hence we get a 2 X 2 matrix.

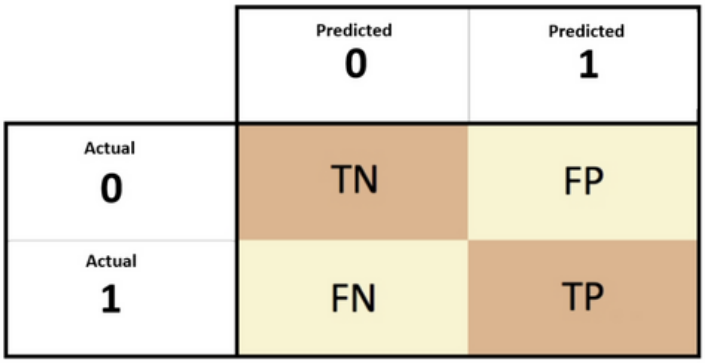


Figure 7.9: Confusion Matrix

- true positives (TP): These are cases in which we predicted correct 1
- true negatives (TN): We predicted 0, and they are 0.
- false positives (FP): We predicted 1, but are not 1
- false negatives (FN): We predicted 0, but they are not 0

From this matrix we can calculate different scores :

- **Precision** : the proportion of positive cases that were correctly identified. Precision = tp/(tp+fp)
- **Recall** : the proportion of actual positive cases which are correctly identified. Recall = tp/(tp+fn)
- **F-score :** harmonic mean between precision and recall
  F-score= 2 * precision * recall / (precision + recall)

MODELING AND EVALUATION

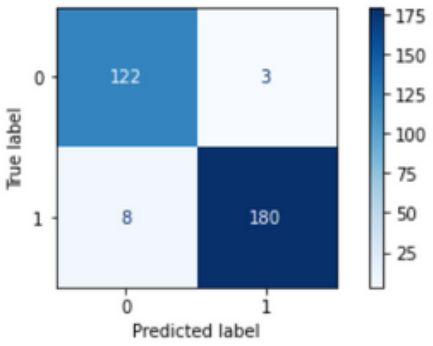
INTERNAL DATA

Classification evaluation and results

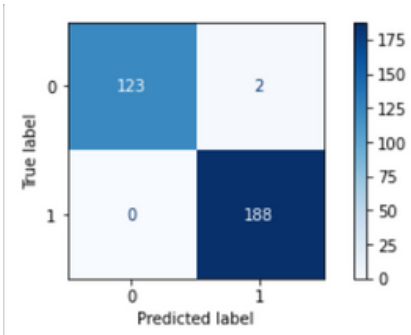
The ROC curve is the plot between sensitivity and (1-specificity). (1- specificity) is also known as false positive rate and sensitivity is also known as True Positive rate.

Confusion matrix for each model

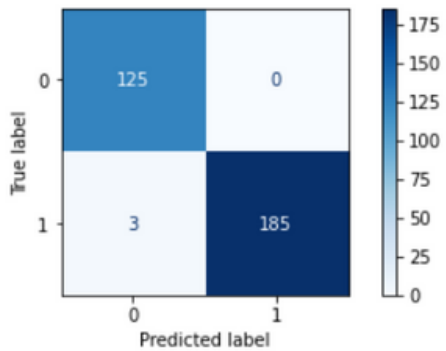
KNN



Decision tree



Random Forest



XGBoost

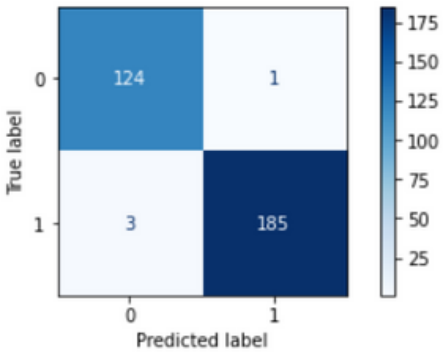
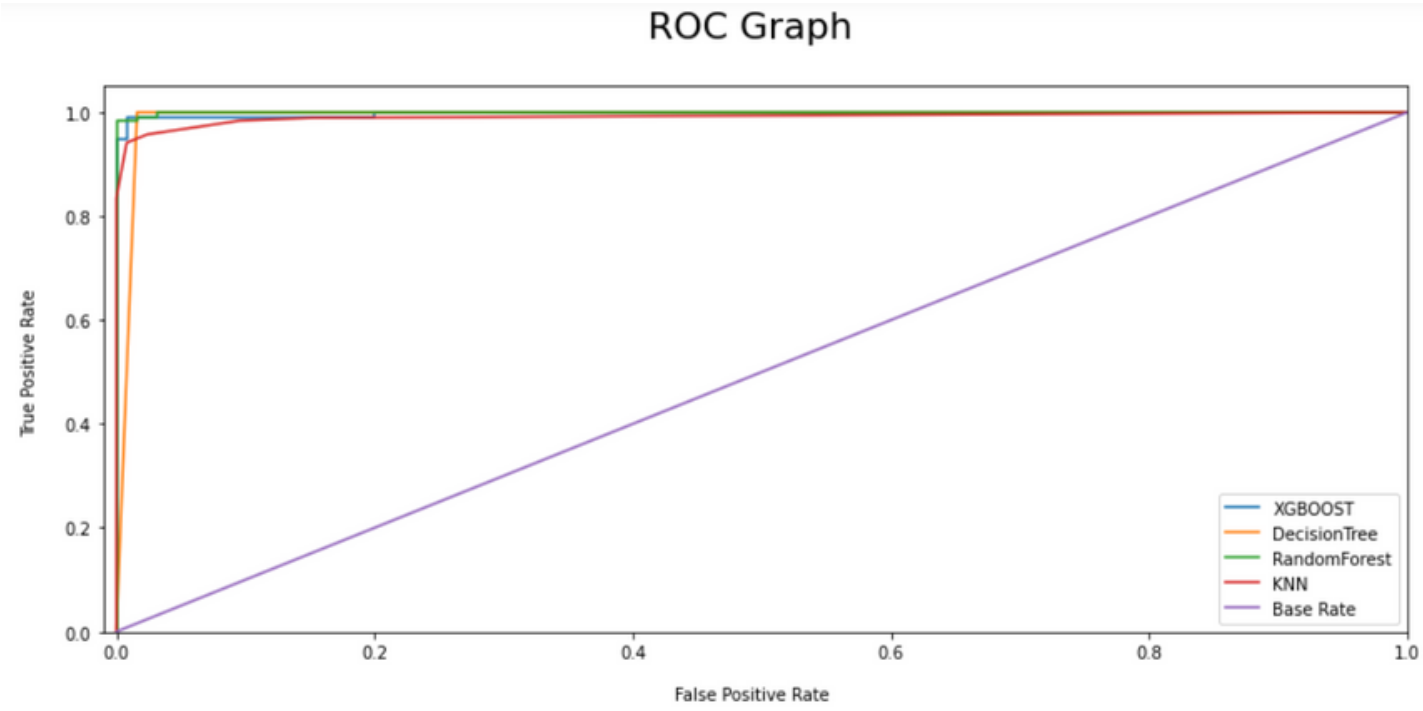


Figure 7.10: Confusion Matrix of every model

MODELING AND EVALUATION

INTERNAL DATA

ROC curve and scores



	Precision	Recall	F1	AUC	Train Score	Test Score
Random Forest	1.000000	0.984043	0.991957	0.999745	0.998874	0.990415
XGBOOST	0.994624	0.984043	0.989305	0.997532	0.996622	0.987220
Decision Tree	0.989474	1.000000	0.994709	0.992000	1.000000	0.993610
KNN	0.983607	0.957447	0.970350	0.990936	0.985360	0.964856

Figure 7.11: Roc Curve and Models Scores

The Random Forest model has the best AUC and F-Score so it is our choice for deployment.

3 3

MODELING AND EVALUATION

EXTERNAL DATA :

Natural language processing

Natural language processing (NLP) refers to the branch of computer science and more specifically, the branch of artificial intelligence or AI concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

NLP combines computational linguistics rule-based modeling of human language with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to ‘understand’ its full meaning, complete with the speaker or writer’s intent and sentiment.

We used NLP on our external data consisting of social media comments to first of all detect the comment's language, secondly translate all comments to english, then conclude the satisfaction of the customers with NLTK and TextBlob and finally create a prediction model with sequential algorithms.

MODELING AND EVALUATION

EXTERNAL DATA :

Natural language processing

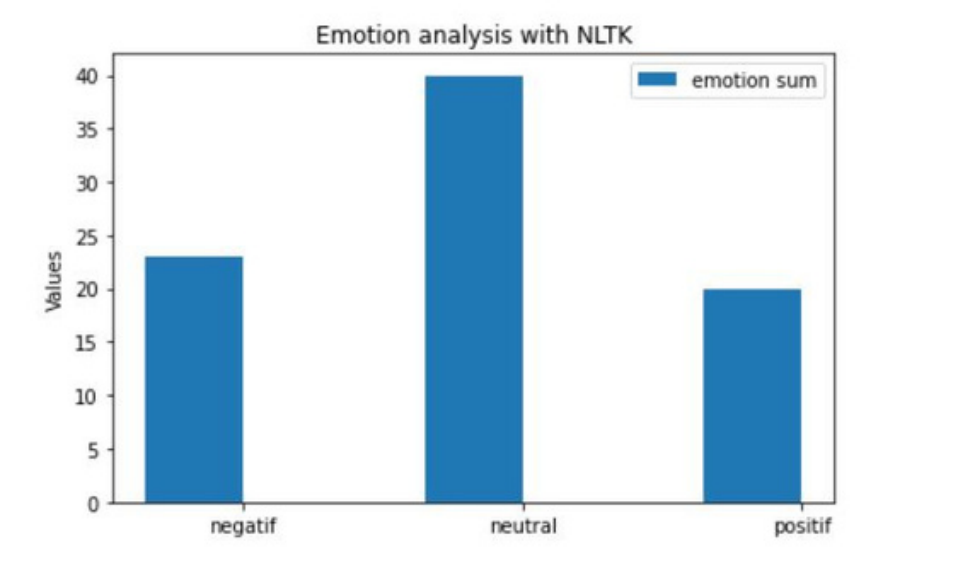


Figure 7.12: Emotion analysis with NLTK

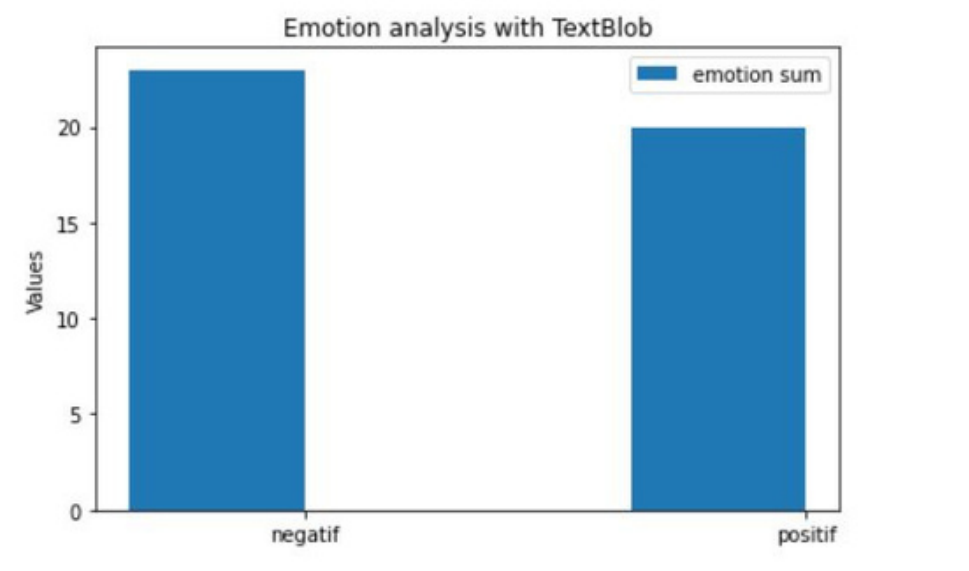


Figure 7.13: Emotion analysis with TextBlob

MODELING AND EVALUATION

EXTERNAL DATA :

Natural language processing

Tokenization

```
[5]: vocab_size = 202
     embedding_dim = 16
     max_length = 100
     trunc_type='post'
     padding_type='post'
     oov_tok = "<OOV>"
     training_size = 32

[6]: sentences = []
     labels = []
     for item in short1:
         sentences.append(item['Comment'])
         labels.append(item['satisfied'])

[7]: print(sentences)
     print(labels)
```

Figure 7.14: Tokenization

Sequencing

Train Set

```
training_sequences = tokenizer.texts_to_sequences(training
print(training_sequences)
```

```
[[9, 48, 16, 8, 49, 50, 51, 12, 52, 17, 53, 13, 54, 55, 4,
17, 4, 29, 63, 64, 2, 65, 8, 66, 13, 14, 4, 67, 18, 2, 68,
3, 35, 15, 4, 25, 26, 75, 2, 21], [15, 12, 36, 76, 77, 14,
88, 22, 89, 90, 40, 13, 91, 23, 6, 92, 17, 22, 41, 93, 40,
3, 12, 36, 6, 1, 11, 1, 20, 1, 41, 1, 9, 1, 1, 1], [6, 7,
37, 1, 1, 1, 27, 8, 9, 1, 1, 1, 11, 1, 1, 1, 9, 19, 1, 1,
0, 39, 8, 10, 12, 1, 47], [6, 7, 3, 5, 1, 10, 1, 21, 4, 25
```

```
from tensorflow.keras.preprocessing.sequence import pad_se
training_padded = pad_sequences(training_sequences, maxler
print(training_padded)
print(training_padded.shape)
```

Figure 7.15: sequencing

Training :

```
[24]: model = tf.keras.Sequential([
          tf.keras.layers.Embedding(vocab_size, embedding_dim, input_length = max_length),
          tf.keras.layers.GlobalAveragePooling1D(),
          tf.keras.layers.Dense(24, activation='relu'),
          tf.keras.layers.Dense(1, activation='sigmoid')
      ])
      model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])

[25]: model.summary()
```

```
Model: "sequential_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_3 (Embedding)      (None, 100, 16)           3232
_________________________________________________________________
global_average_pooling1d_3 ( (None, 16)                0
_________________________________________________________________
dense_6 (Dense)              (None, 24)                408
_________________________________________________________________
dense_7 (Dense)              (None, 1)                 25
=================================================================
Total params: 3,665
Trainable params: 3,665
Non-trainable params: 0
```

Figure 7.16: Training the model

MODELING AND EVALUATION

EXTERNAL DATA :

Recommendation system

A recommendation system is a machine learning model that involves predicting the user responses to the options meaning their wants.

It is an algorithm which aims to provide the most relevant and accurate services to the user by filtering useful stuff from of a huge pool of information base. Recommendation engines discovers data patterns in the data set by learning consumers choices and produces the outcomes that co-relates to their needs and interests.

We developed a custom recommendation system to be integrated in our contextual AI assistant. It predicts the best services fit to the customer interacting with the assistant, based on the services asked about in the conversation evolves around it recommends the services with the highest probability rate to fit our customer's profile

MODELING AND EVALUATION

EXTERNAL DATA :

Recommendation system

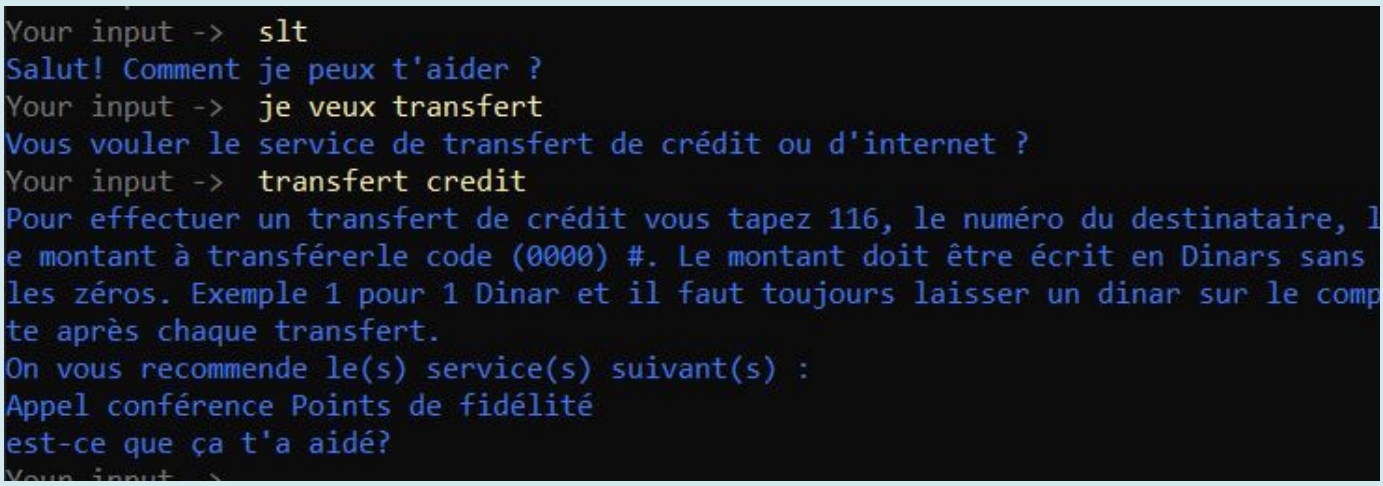


Figure 7.17: recommendation example 1

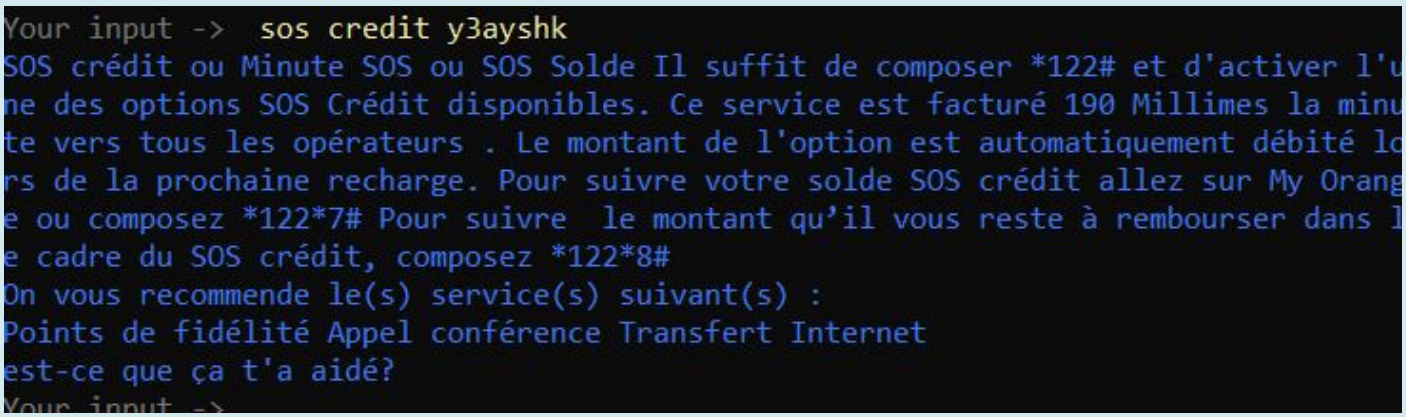


Figure 7.18: recommendation example 2

MODELING AND EVALUATION

EXTERNAL DATA :

AI contextual assistant : rasa

Rasa Open Source supplies the building blocks for creating virtual assistants. we used Rasa to automate human-to-computer interactions anywhere from websites to social media platforms. Rasa Open Source provides three main functions. Together, they provide everything our client needs to build a virtual assistant:

Natural Language Understanding
Convert raw text from user messages into structured data. Parse the user's intent and extract important key details.

Dialogue Management
Machine learning-powered dialogue management decides what the assistant should do next, based on the user's message and context from the conversation.

Integrations
Built-in integration points for over 10 messaging channels, plus endpoints to connect with databases, APIs, and other data sources.

We built our own custom rasa AI contextual assistant that fits our client's needs. It provides seekers with answers about our client's services, how to put the in use and recommends the best package that fits the seekers based on their conversation with the help of our integrated recommendation system.

MODELING AND EVALUATION

EXTERNAL DATA :

AI contextual assistant : rasa

Figure 7.19: Rasa

```
Your input ->  comment m'inscrire
Tapez *112# depuis votre mobile Orange  ou tapez directement *112*2# et consultez l
r produit : Mobile, Clé 3G, Flybox et ADSL . Le solde de vos points cadeaux s'affic
ts cadeaux ont une validité de 24 mois.
est-ce que ça t'a aidé?
Your input ->  nheb transfert
Vous vouler le service de transfert de crédit ou d'internet ?
Your input ->  intarnet
Pour Transférer de l'internet depuis votre mobile est désormais possible via *129#
ge ! Vous pouvez dépanner un ami ou un proche en lui transférant de l'internet sur
est-ce que ça t'a aidé?
Your input ->  chneya nerbah
Grâce à vos points, bénéficiez d'un large choix de cadeaux * :
- Bonus valable vers tous les opérateurs - Internet Mobile - SMS valables vers tous
```

Figure 7.20: Chatbot example

MODELING AND EVALUATION

EXTERNAL DATA :

Image sentiment analysis : Face recognition emotion using Deep Face and openCV

Face recognition emotion is the process of detecting human emotions from facial expressions. The human brain recognizes emotions automatically, and software has now been developed that can recognize emotions as well. This technology is becoming more accurate all the time, and will eventually be able to read emotions as well as our brains do.

Which model of human emotions we accept and work with has important consequences for modeling them with Deep Learning using Deep Face and open CV. A categorical model of human emotion would likely lead to creating a classifier, where text or an image would be labeled as happy, sad, angry, or something else.

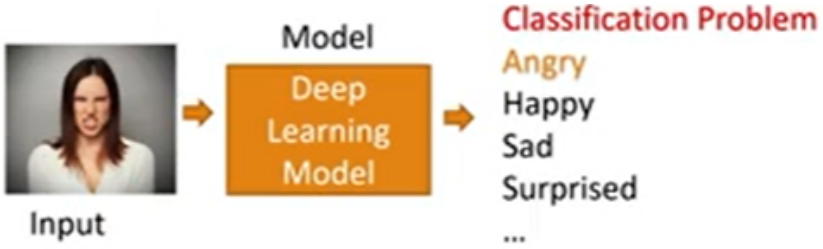


Figure 7.21: emotions classifications

41

MODELING AND EVALUATION

EXTERNAL DATA :

Image sentiment analysis : Face recognition emotion using Deep Face and open CV

We have not trained the data set rather we have used a pre-trained library which is **deep face.** It contains a lot of pre-trained Deep Learning architectures for face recognition emotion and we have detected the image with **open CV**.

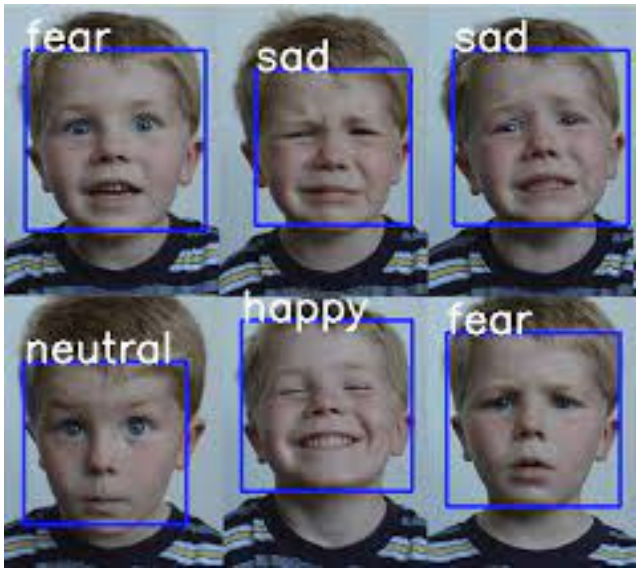


Figure 7.22: emotion examples

MODELING AND EVALUATION

EXTERNAL DATA :

Image sentiment analysis : Face recognition emotion using Deep Face and open CV


```
Out[7]: {'region': {'x': 122, 'y': 33, 'w': 45, 'h': 66},
 'emotion': {'angry': 61.26742362976074,
  'disgust': 1.1437006008918615e-06,
  'fear': 0.06510637467727065,
  'happy': 0.003444803587626666,
  'sad': 38.319358229637146,
  'surprise': 2.663711384798262e-06,
  'neutral': 0.34466739743947983},
 'dominant_emotion': 'angry',
 'age': 30,
 'gender': 'Man',
 'race': {'asian': 5.500579252839088,
  'indian': 4.10173237323761,
  'black': 0.6933250464498997,
  'white': 50.37513971328735,
  'middle eastern': 19.025160372257233,
  'latino hispanic': 20.304062962532043},
 'dominant_race': 'white'}
```


Figure 7.23: input description

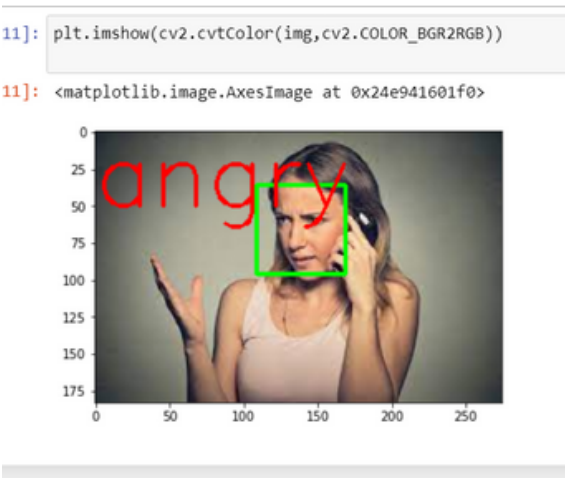


Figure 7.24: image result

MODELING AND EVALUATION

EXTERNAL DATA :

Voice sentiment analysis

The model works to detect the different emotions the human being can express from the tone of his voice (anger, fear, happiness, sadness, surprise) in order to determine wither the customer is satisfied or not.

The dataset is the combination of three other datasets RAVDESS,TESS,SAVEE containing the emotions we just mentioned

	source	actors	path	emotion2	emotion3
0	TESS	female	D://emotion_recognition_data\data3/OAF_Fear/OA...	negative_female	fear
1	TESS	female	D://emotion_recognition_data\data3/OAF_Fear/OA...	negative_female	fear
2	TESS	female	D://emotion_recognition_data\data3/OAF_Fear/OA...	negative_female	fear
3	TESS	female	D://emotion_recognition_data\data3/OAF_Fear/OA...	negative_female	fear
4	TESS	female	D://emotion_recognition_data\data3/OAF_Fear/OA...	negative_female	fear

Figure 7.25: voice input description

Using the Librosa library we can plot the sound wave

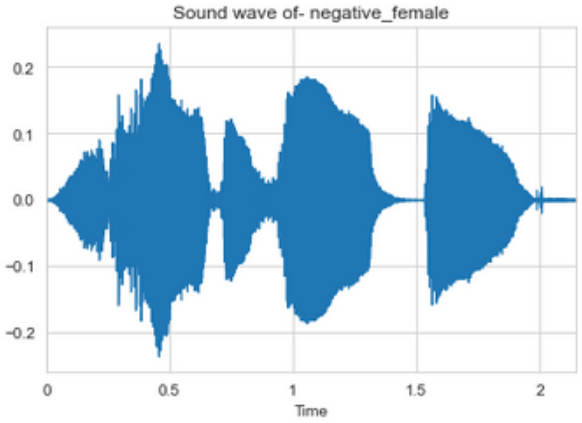


Figure 7.26: negative sound wave of female

MODELING AND EVALUATION

EXTERNAL DATA :

Voice sentiment analysis

By performing the Fourier transformation and log-scale we can display the same signal in a time-frequency domain, so that we could examine the different frequencies and amplitudes of our signal over time. this is called spectrogram.

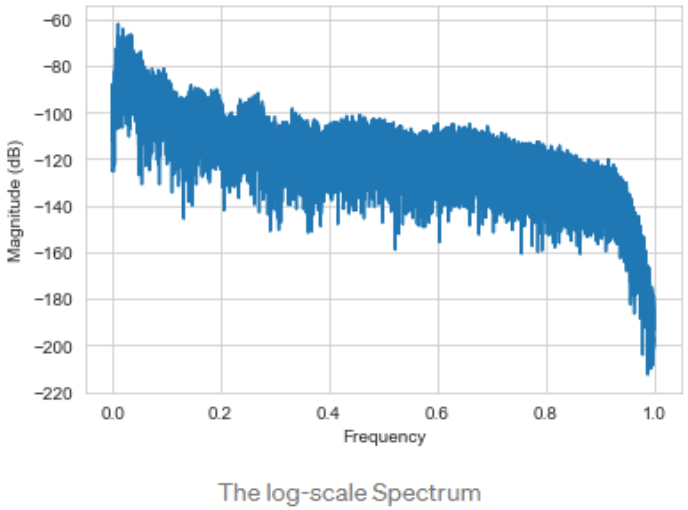


Figure 7.27: log scale spectrum

Now that we recognized our Vocal Tract, we need to find a way to extract it from our speech with Mel-frequency cepstrum coefficients, also known as the MFCC's!

```
mfccs = []
for i in tqdm(X):
    mfcc = librosa.feature.mfcc(y=i, sr=44000, n_mfcc=20)
    mfcc = mfcc.T
    mfccs.append(mfcc)
```

Figure 7.28: MFCC

MODELING AND EVALUATION

EXTERNAL DATA :

Voice sentiment analysis

Since we are dealing with images we can use Deep learning models such as CNN using 1-dimensional convolution layers and 1-dimensional pooling layers.

	precision	recall	f1-score	support
fear	0.83	0.83	0.83	75
disgust	0.85	0.89	0.87	57
neutral	0.82	0.91	0.86	76
happy	0.78	0.88	0.83	60
sadness	0.84	0.79	0.82	67
surprise	0.94	0.84	0.89	69
angry	0.92	0.81	0.86	68
accuracy			0.85	472
macro avg	0.85	0.85	0.85	472
weighted avg	0.85	0.85	0.85	472

Accuracy: 0.8496

Figure 7.29: voice sentiments results

DEPLOYMENT

For the final stage, the deployment of all the work previously done. We started by deploying the dashboards and reports created in a new website to provide our clients with visuals that illustrate the satisfaction of the clients based on many axes filtered by many variables .
These visuals Also describe the satisfaction distribution by many regions in Tunisia , the means of marketing used and the quality of information provided by them , it shows specific information on all of the services by of all the operators.

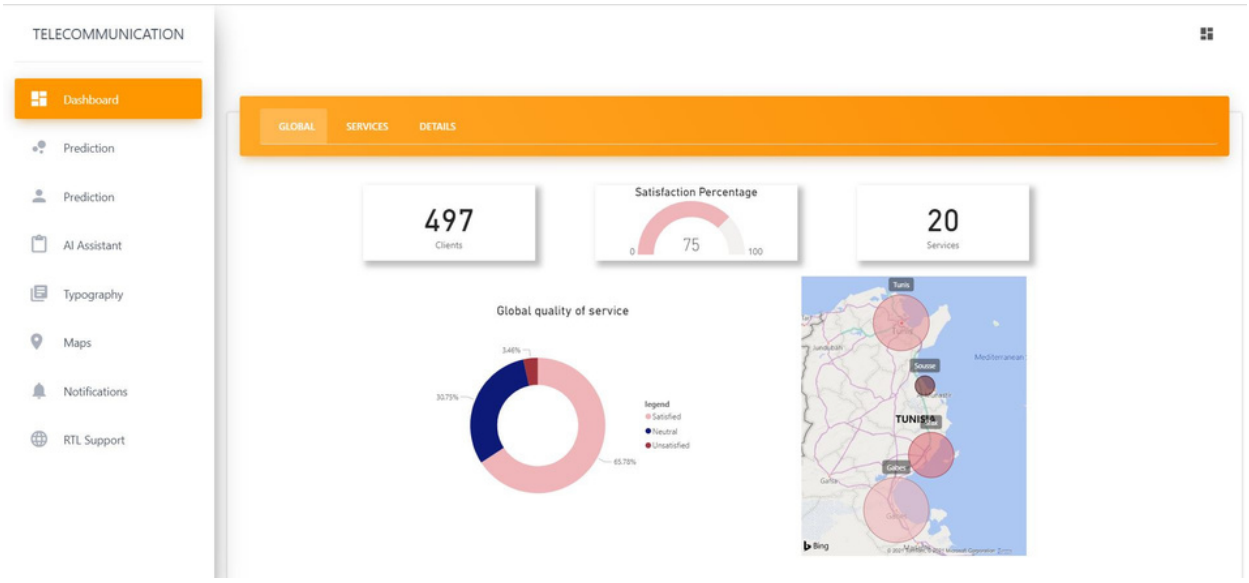


Figure 8.1: Main dashboard

47

DEPLOYMENT

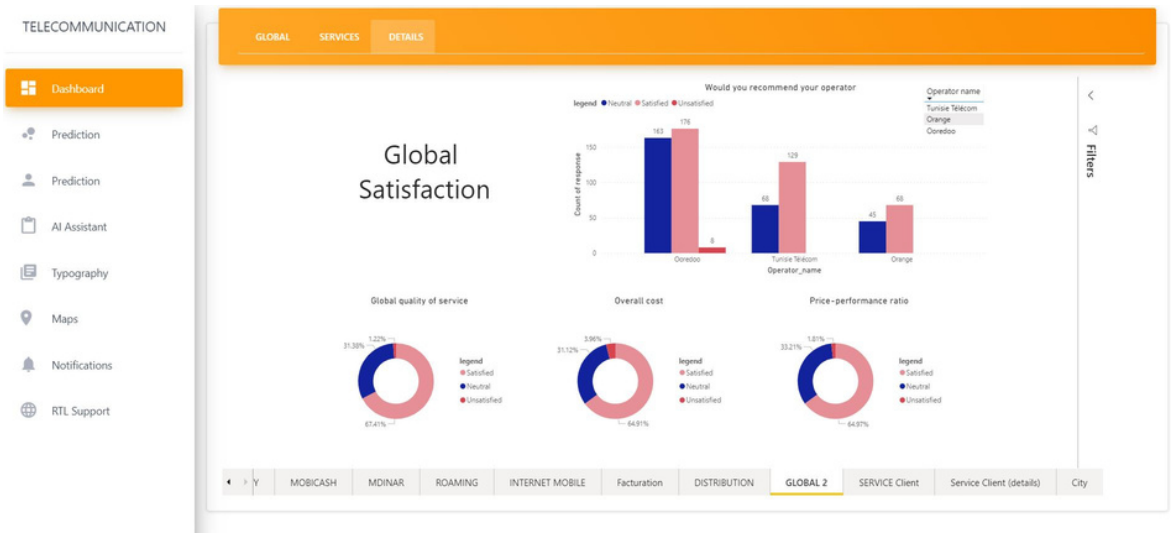


Figure 8.2: example of services dashboard

The visuals also provide information on the words that are used to describe the operators in addition to what the users actually expect from these operators and the problems they face .

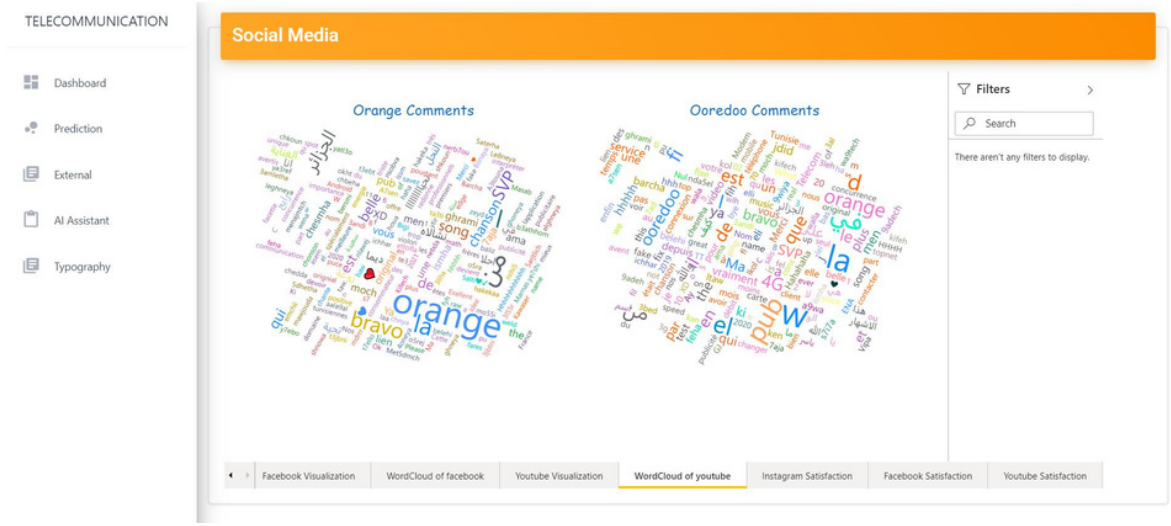


Figure 8.3: word clouds of orange and ooredoo comments

DEPLOYMENT

We added a chatbot to the site to create a useful information source that is always available to the users to help with unavailability and a quick access to said information .

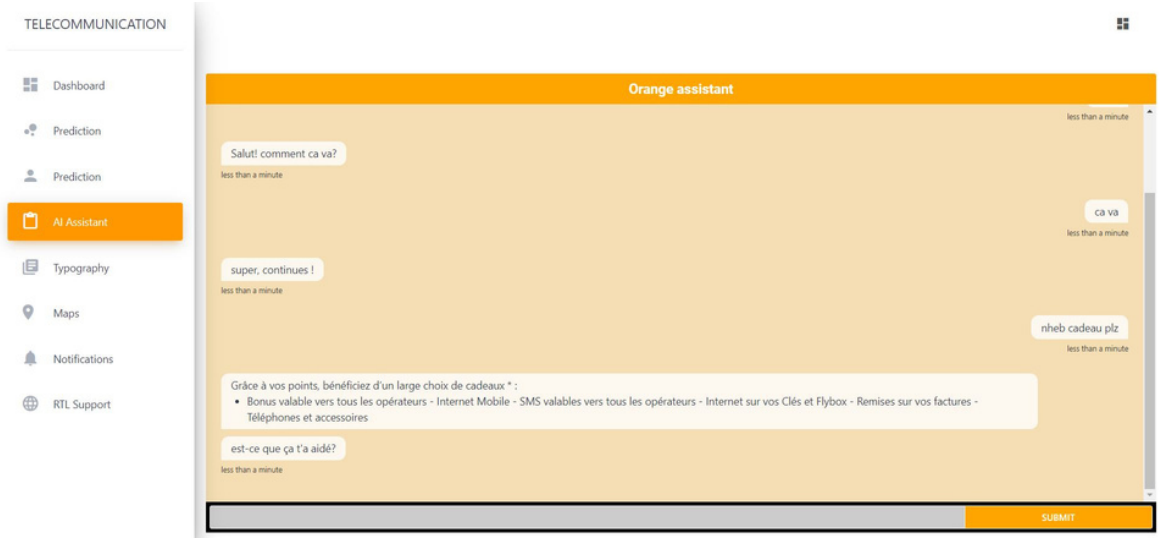


Figure 8.4: chatbot deployment

And for the final model , we provided our client with the power to predict the satisfaction of his users from all sources of data , videos , vocals or forms that our client post constantly with the possibility of multiple predictions using excel files instead of refilling the information provided by the client one by one .

DEPLOYMENT

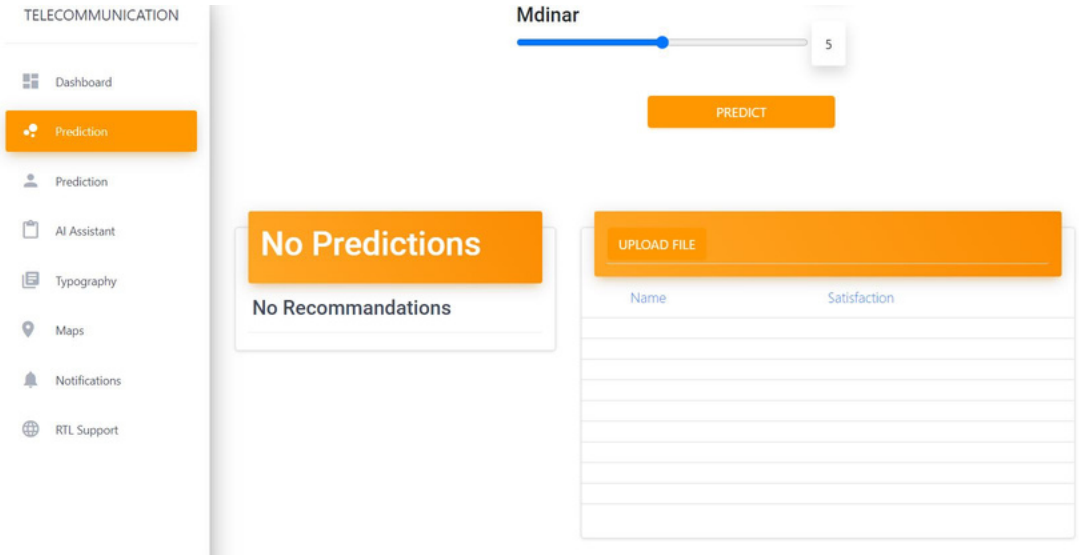


Figure 8.5: data input for prediction

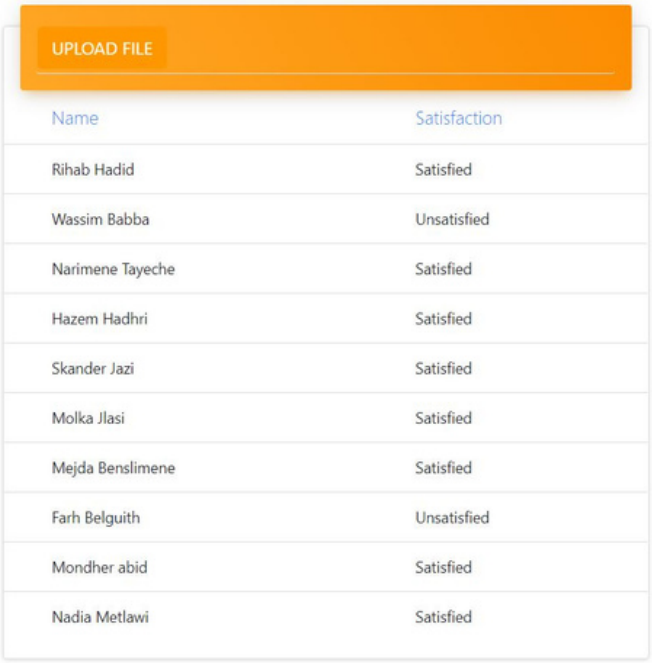


Figure 8.6: prediction results

5 0

DEPLOYMENT

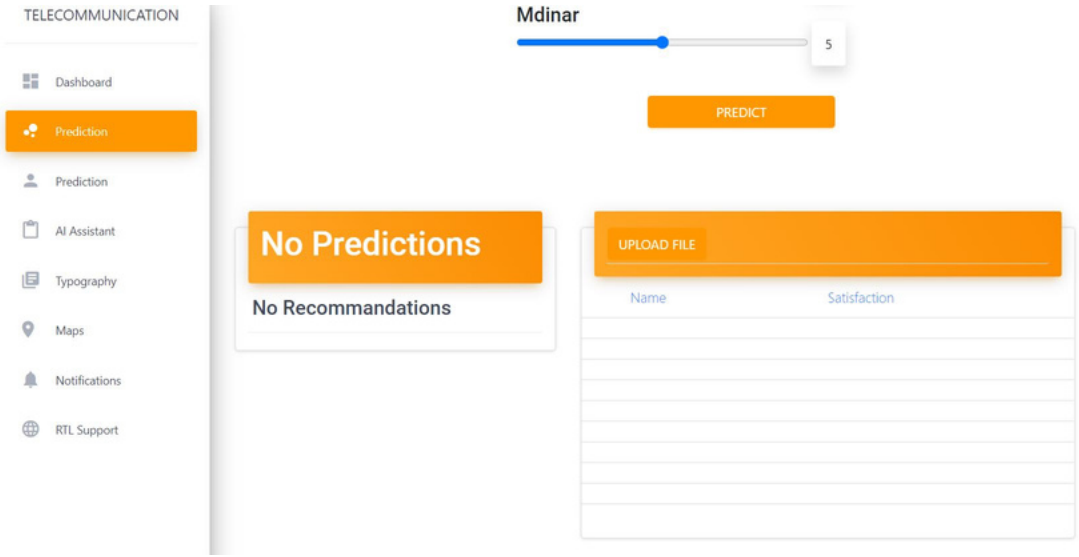


Figure 8.5: data input for prediction

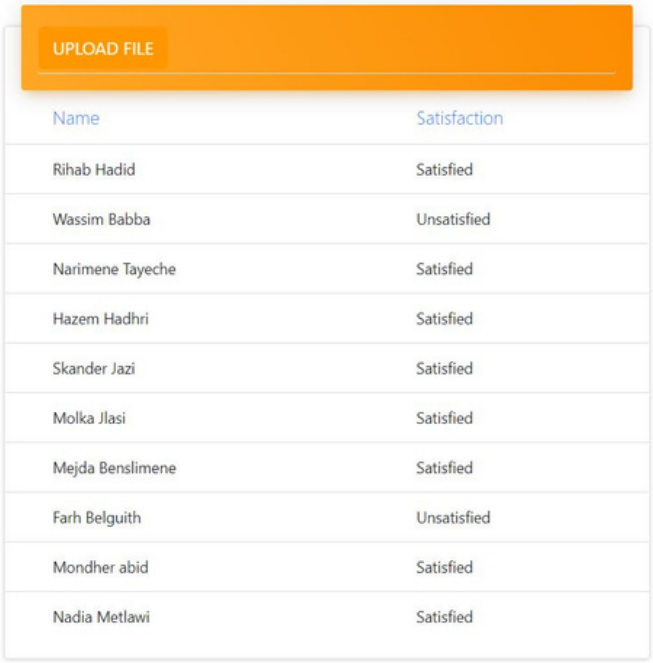


Figure 8.6: prediction results

5 1

DEPLOYMENT

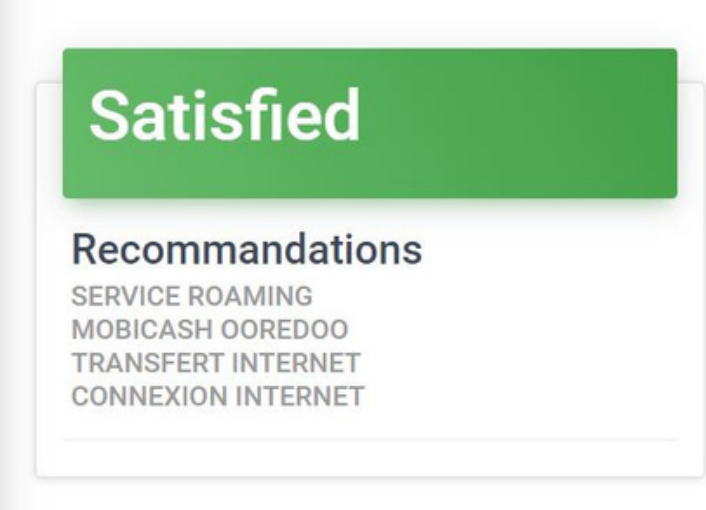


Figure 8.7: Satisfied result

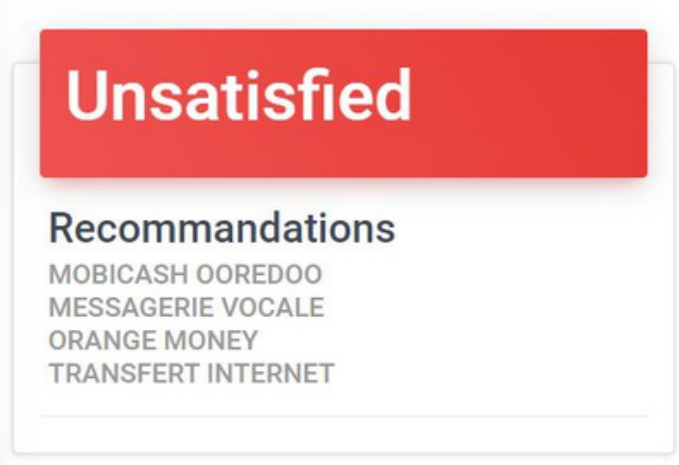


Figure 8.8: Unsatisfied result

5 2

CONCLUSION

Knowing the existing problems and knowing the services that are causing users to be unsatisfied are very useful information to our client and adjusting those services and predicting the number of churners will help our client in calculating the amount of money to spend on recruiting new clients as he will know approximately the exact number of clients leaving . This will help in achieving his financial balance .
Besides , providing a full time useful services like the chatbot will help in creating a better image and recruiting new clients .