**NLP Project 2: Insurance reviews**

The aim of this project is to work on an insurance reviews Dataset that contains information about the rating of an insurance by a user. This work will be divided into 3 big parts: First of all, we will have to explore the data and process it with some cleaning, as well as some data visualization. Then, we will create a supervised model that will categorize a rating as positive or negative, using clusters. Finally, we will work with an unsupervised model to try to predict the rating, given the description the user gave to the rating.

# Data exploration, cleaning, and visualization

Une image contenant texte

Description générée automatiquement

The first step to understanding how we will work with this data and which processes we’ll have to make is to visualize it very quickly. The most obvious cleaning that appears is cleaning the date to make it usable, getting rid of the “suite à une…”. I also created a column named “date\_difference” to see the difference between the date of the review and the date of the experience.

Une image contenant texte

Description générée automatiquement

The next step was to process the review column:

Une image contenant texte

Description générée automatiquement

These were pretty basic steps for NLP projects, which aim to tokenize the reviews, lemmatize them to finally created new columns “produit\_cat” and “assureur\_cat” for further work on this column later.

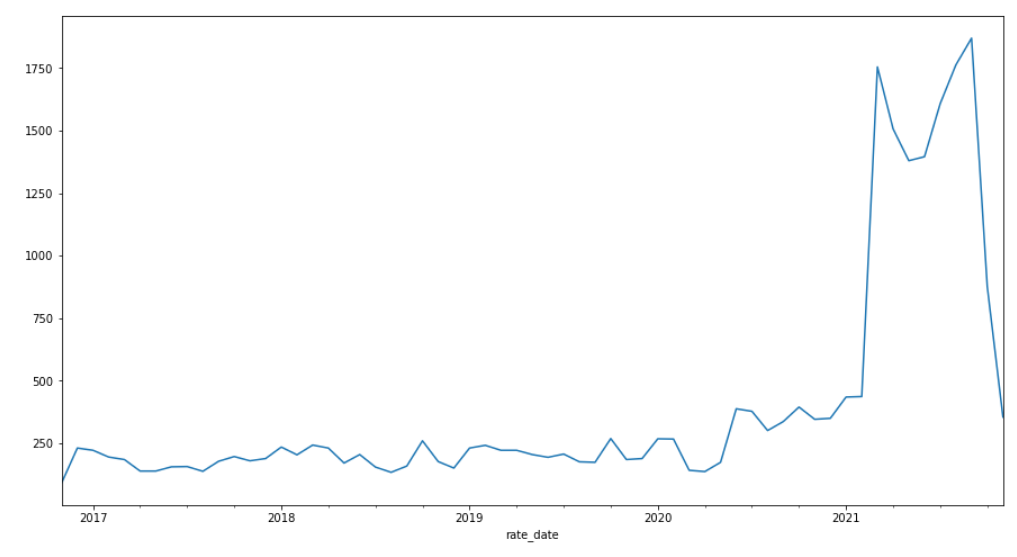
Let’s now do some data exploration on the columns.

Une image contenant texte

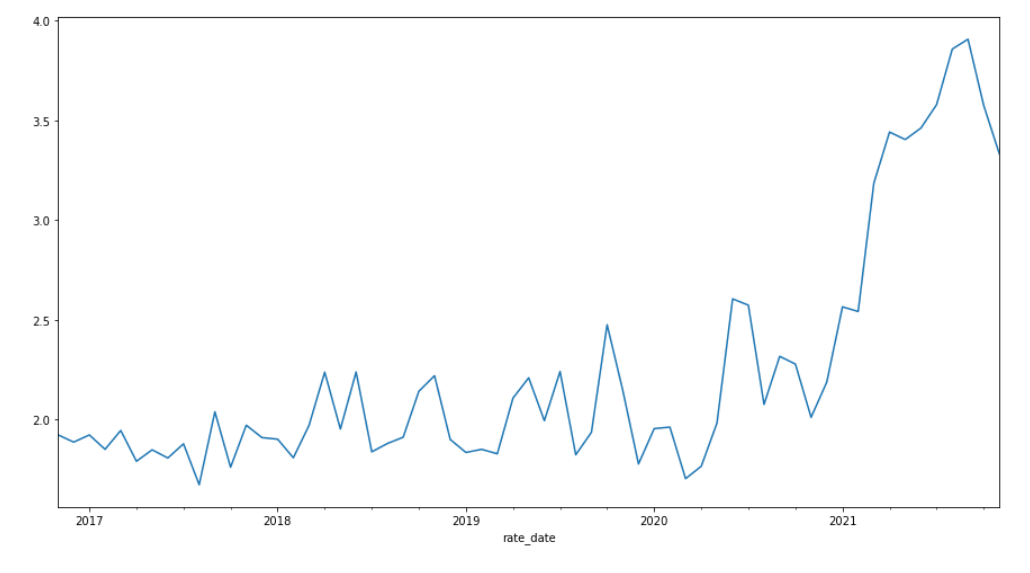
Description générée automatiquement

We first make sure that no column has null values, to get rid of problems that might come up later when working on our models.

We then looked at the ratings and the date they were made, and noticed that, in 2021, a big spike in the number of reviews as well as in rating appears:



*Number of ratings*

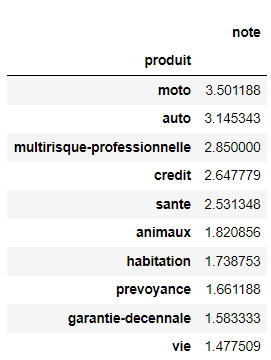


*Ratings*

We then looked at the products, to see which were the most present, and which were the best rated ones. It appears that “auto” products were the most present, which is pretty foreseeable, as insurance are necessary for them, however the best rated products were motorbikes, with 3.5 stars.

Une image contenant table

Description générée automatiquement



The same exploration was made on insurances themselves:

Une image contenant table

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

Finally, I decided to make a wordcloud on the reviews, to see which words were coming up the most in reviews:

Une image contenant texte

Description générée automatiquement

# Modeling

## Unsupervised Learning

The first model that we will work with is an unsupervised model. The aim of the model, as stated in the introduction, will be to categorize a review as positive or negative. Using the lemmatized reviews we processed before, I worked on word embedding:

Une image contenant texte

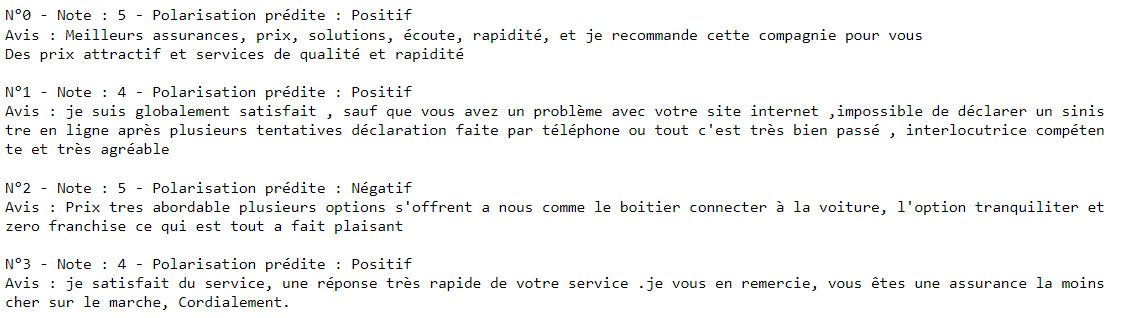
Description générée automatiquement

Thanks to this model, it is also possible to look at similar words. In these examples, I looked at “nul” to see similar negative words, and I looked at “satisfaisant” to see similar positive words. It looks like the similar words given are pretty accurate.

After this, the next step was to vectorize these reviews, to make them exploitable by algorithms such as the KMeans that I will use. This KMeans will have 2 clusters, which will correspond to Positive Reviews and Negative Reviews (1 or 0).

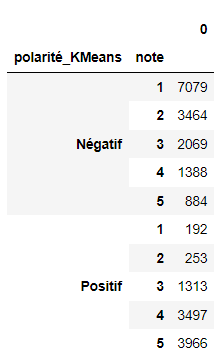


I looked at the first reviews to see if the clusters were working correctly:

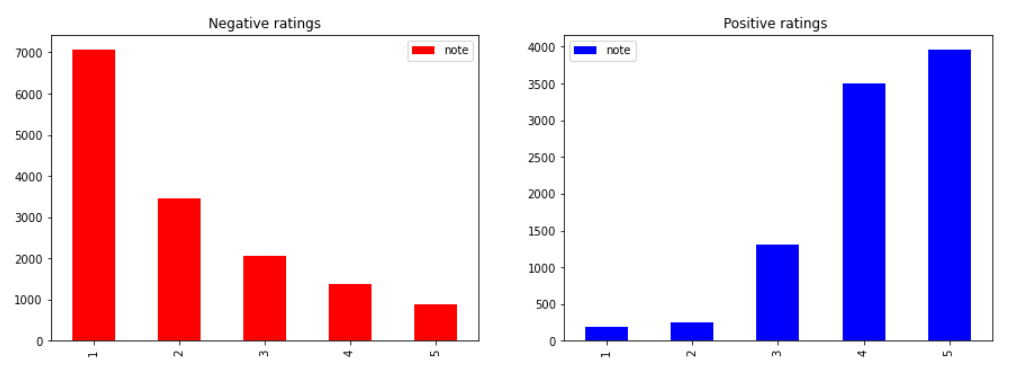


We can see that, other than the review N°2 that is a 5 but Negative (probably because of the word “zero” if we tried to speculate why), the clusters seem to work fine. I looked at the first 10 reviews and only the one we just talked about was wrongly clustered.

To further look into this, I tried to visualize the polarity of the KMeans associated with the ratings:

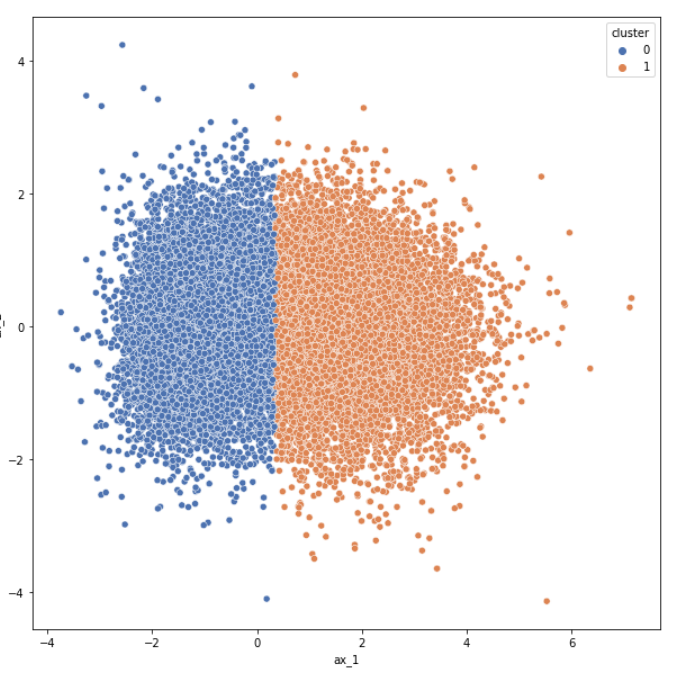


We can see that most of the negative reviews have [1;3] ratings, while most of the positive ones have [3;5]:



*Repartition of the ratings associated with Negative or Positive ratings*

I then worked with a PCA to try to visualize these clusters as well:



We can clearly see that both clusters are well separated, which means that the model worked well on the data given.

## Supervised Learning

To begin to work on the supervised learning model, I did a PCA to have 10 features instead of 100, to make the work less complex. I also added the column “Cluster” that we just worked on.

I decided to go with a linear regression model, as a first try with a basic model, fitting on the dataset, which has been split with a 0.2 test size ratio.

Une image contenant texte

Description générée automatiquement

The results on the train and on the test were pretty close, getting an R2 score of around 0.55 and an RMSE of 1:

Une image contenant texte

Description générée automatiquement

With those results in mind, I decided to create another model, to compare its results to the linear model. I decided to go with a Random Forest Regressor, as it is a pretty polyvalent solution to most problems. I used a gridSearch to look for the best parameters, and got the results:

Une image contenant texte

Description générée automatiquement

With those parameters now found, let’s look at the results of the Random Forest Regressor on the data:

Une image contenant texte

Description générée automatiquement

Finally, with 100 estimators being the best performing parameter for the train and the test, we get a R2 score of 0.61 and an MSE of 1.05 on the test data, with the mean result being a 2.87, which is predictable, since reviews range from [1;5]; and there are enough reviews to be close to the theoretical mean value (3).

## Working on the test data

The final step of this project is to predict ratings given reviews in a test data.

Une image contenant table

Description générée automatiquement

As we did previously, we tokenize the reviews, to then create a KMeans, to cluster these reviews, and then finally to predict the rating. This final step is combining both unsupervised learning and supervised learning steps into one. Because of that, the details of the tokenization, algorithms,… are the same as we previously saw. Given that, a final .csv sheet can be created: it will have the predicted ratings of each review of the test set:

Une image contenant texte

Description générée automatiquement

**Conclusion**

To conclude, in this project, we were able to explore a dataset about Insurance Reviews and created multiple Supervised and Unsupervised models. Finally, these models were used to cluster reviews into positive or negative, and to predict the rating given a review.

**Links**

<https://github.com/mehdinoureddine/NLP_insurance_ratings>

(You will find the notebook, as well as this document and the prediction sheet “pred\_nlp\_proj2.csv” in this repository)