Summer Engineering Internship

Name: ***Jaime Gutiérrez de Calderón Martinez***

Millikin ID: ***905640***

Professor: ***Dr. Daniel Miller, Professor of Mathematics***

Company*: NEOVANTAS CONSULTING S.L*

Location: FERNANDEZ DE LA HOZ STREET 33

28010 Madrid (Spain)

Tutor: ***Adrián González*** **(Business Analytic Senior)**

INDEX

1. INTRODUCTION

Since the beginning of the year, I have been looking to do an internship in a company during the summer to develop my skills and face real life problems. I was in different selection processes but the company that I finally selected was Neovantas. The reason to choose this one was that the projects that they have done before impressed me and the ones that they told me that I was going to be involved also attracted me. Another reason that helped me to take this decision was that Neovantas is a small consultancy, so I expected that I was going to be able to have some relation with most of the employers in the company.

2. ABOUT THE COMPANY

It was at the end of 2003 when Neovantas[1] was starting conversations with its first client in Spain and since then, it has remained faithful to our principles of being a management consultancy that brings a strong analytical component to its collaborations and that shows tangible and significant results in income, costs and/or quality to its customers.

Since 2009, they have been the first management consulting firm to systematically use sophisticated Speech & Text Analytics tools in Spain in order to incorporate unstructured information into our analyses.

Its main value proposition is aimed at achieving relevant RESULTS quickly and sustained over time. To do this, it provides two elements that differentiate us from the rest of the competition:

a) It exploits its clients data by incorporating the plus of UNSTRUCTURED INFORMATION (oral and/or written) to the analysis of the data to generate new intelligence that helps them outline better plans and concrete actions.

[1] Neovantas. "NEOVANTAS - Consultoría multinacional de alta dirección." *Neovantas - International Management Consultancy*, 11 Feb. 2021, www.neovantas.com.

b) And, in addition, it incorporates BEHAVIORAL OPTICS, systematically analyzing the psychological barriers (biases) in customers and employees to reduce the gap between their intentions (what they think) and their behavior (what they end up doing).

From its office in Madrid, Neovantas attends their commitments in Europe (Germany, Portugal, Poland and Italy, mainly), as well as in Latin America where it carries out projects in countries such as Mexico, Brazil, Peru, Chile and Argentina.

It works, fundamentally, with the main participants in the Banking, Insurance and Telco sectors, maintaining its relationship with several of them for many years continuously.

GENERAL OBJECTIVES

The company, Neovantas, is specialist in extracting nuances of great relevance (which often go unnoticed) for the objectives of its clients from the analysis of oral and unstructured information obtained from both end clients and employees.

It constantly searches and develops new tools.

It has recently created a Speech Analytics laboratory[2], one of the most advanced in the Spanish market, which facilitates the rapid exploitation of recordings. Its objective is to "read between the lines" everything indicated by end customers and employees. Exploiting the recordings in this way achieves a great impact on the results.

This company has developed JOY[3], which is a multi-device application that allows the exploitation of all the information available to an entity, both structured and unstructured (eg: recordings, emails), in an easy, fast and pragmatic way to improve customer satisfaction.

---

[2] Neovantas. "Laboratorio de Speech Analytics." *Neovantas - International Management Consultancy*, 12 Jan. 2021, www.neovantas.com/laboratorio-de-speech-analytics.

[3] Neovantas. "JOY - Satisfacción Clientes." *Neovantas - International Management Consultancy*, 12 Jan. 2021, www.neovantas.com/joy-satisfaccion-clientes.

Specifically, it calculates a score from 0 to 10 with its artificial intelligence engines, automatically weighting hundreds of variables and all this without asking the customer or the employee.

3. JOB DEVELOPED

CONTEXT DESCRIPTION

The internship was developed in the company facilities in the area of Analytics, which is located in the GOYA room on the 3rd floor. The company is located in C/ Fernandez de la Hoz, 33, 28010, Madrid, Spain.

SCHEDULE

- Monday-Thursday: 8.5 hours per day
- Friday: 6 hours, preferably between 9:00-15:00
- I had the flexibility to start at any time between 8:00-10:00 with 30 minutes break to have lunch

EQUIPMENT

Hp ProBook Laptop

- Model: HP ProBook 4530s
- Processor: Intel(R) Core(TM) i5-2410M CPU @ 2.30GHz 2.30 GHz
- Installed memory (RAM): 4.00 GB
- System type: 64-bit operating system
- Windows Edition
- Windows 7 Professional
- Copyright c 2009 Microsoft. Corporation.
- Service Pack 1

Desk

 Ergonomic desk and chair

Display

 18.5-inch Samsung LED monitor.

SPECIFIC OBJECTIVES

The management of Neovantas Consulting, S.L, with the commercial name Neovantas, offered me during the summer of 2022 the opportunity to join their team as Business Analyst. My main goals during the internship were the following ones:

- Learn to carry out analysis concerning "problem solving" from the in-depth understanding of the situation to the analysis of relevant issues to end with drawing conclusions for resolution.
- Learn to analyze and manage different sources of information with a mostly unstructured profile (such as recordings, emails, social networks, etc.).
- Learn to prepare presentations and documents.
- Market studies.

PROJECTS

- Design a predicting model for customer satisfaction of the clients of Banco Santander Mexico from the different recording calls.
- Merge different databases with information about the clients of Mapfre Spain (insurance company).
- Design a predicting model to predict the customers that were going to renew their insurance policy of the company Verti (Insurance company).
-

4. DETAILED DESCRIPTION OF THE ACTIVITIES CARRIED OUT

THEORICAL BASIS OF THE TOOLS AND APPLIED KNOWLEDGE

I have done three different projects, two of them were really similar so I could apply the knowledge that I get in the first project in the second one, and the third one was a different project, and it was shorter than the first two.

For the predicting models I follow the following steps:

1. Descriptive Analysis

Descriptive analysis can be adjusted depending on the type of information you want to find or the destination of the data. Therefore, it is a flexible analysis that can have different applications and uses. Now, regardless of the reason why it is done, it usually complies with the following steps:

Collect data

The main thing is to collect the data. For which various instruments are used that are applied to the selected sample. For example, it can be done through a survey or questionnaire. Online communities are one of the instruments used to collect complete and accurate data. In Neovantas the majority of data collected was from the different recording calls that has the customer with the company service.

This instrument is very successful because it works online. In addition, it allows extracting qualitative information, generating virtual discussions, generating ideas, rewarding participants, etc.

Cleansing or purging data

As mentioned at the beginning, a huge amount of data is generated every day today. However, not all of them are useful for a certain purpose. This step cleans the data to remove data that will not be used. For example, those whose

formats do not allow access and therefore cannot be manipulated can be deleted.

When data is cleaned, its textual formatting, categorization, and outliers may change. This process can be summarized in the preparation of the data to get more out of it.

Apply methods

The last step is the application of statistical methods to be able to conclude what was planned with the analysis. There may be several methods depending on the type of data and what is expected to be determined. This is the most important step, since it involves the use of the data to meet certain objectives, make decisions, make corrections, etc.

The most important ones in my different projects were to find correlations of the variables with the target.

2. Creating the predictive models

The predictive model is a data model, based on inferential statistics, used to predict in my case, if a customer is going to be satisfied at the end of the year with his bank or insurance policy company or if he is going to decide to get another company.

The predictive model is typically created by data scientists and uses statistics to predict outcomes. Most of the time the event one wants to predict is in the future, but predictive modeling can be applied to any type of unknown event, regardless of when it occurred.

I have used different methods as Logistic regression, Decision trees, Random Forest, XGBoost, among others, and my objective was to use the one that predicts better the customers that were not happy with the company.

3. Conclusions

At the end of each project, I had to explain my point of view and show the most interesting results that I got.

PROJECTS IN WICH I HAVE BEEN INVOLVED

During the internship, my main focus were three important projects, two of them were related to creating predictive models and the other one was based in merging different databases to facilitate future descriptive analysis.

All my work was made in python using Jupyter, and the databases were sometimes .txt, others .csv and others .xlsx.

VERTI BTS (Monday-Tuesday 16 hours)

Verti[4] is the digital company of the MAPFRE Group. A young insurer born in January 2011, with 100% Spanish capital and backed by a leading group in the insurance market worldwide. Despite this short time, they are the direct insurance company that has grown the fastest in the history of Spain, already reaching almost 300,000 clients.

Its greatest challenge is to satisfy your needs with the maximum guarantees of protection at the best quality of service, creating products and services geared towards this.

They asked Neovantas to help them to improve the calls with their clients, so they were able to make them renew their contract with the company. Verti let Neovantas different recordings of telephone conversations with their clients, so we were able to get different data and let them know what were the common things between the conversations when the client decided to cancel the contract with the company.

[4] Verti Seguros. "VERTI | Los Mejores Seguros Online de Auto, Hogar y Moto." *Verti Seguros*, www.verti.es. Accessed 15 Aug. 2022.

I spent almost 16 hours (two days of job) in this project. These are the things that I did to develop a descriptive analysis of the database that they gave us and then find the best predicting model to know if the client was going to cancel.

1) **Descriptive analysis.**

   a) **Import the database.**

      I used *pandas* to import the database and then be able to do a descriptive analysis

   b) **Watch the variables**

      Then I look the variables and see what type of information they give me, there were variables about the time of the conversation, the time that the client and the customer were in silence, also if the increase their volume during the conversation, if there were any noise during the conversation, who was the customer and who was the client and finally I had our target that was a column called ‘renovación’ that was a binary column, and it was 0 if the client cancel the contract after the conversation or if he renew it and continue with the company was 1.

   c) **Types of variables**

      Before starting to do the predicting model I had to review if the Data Base had object variables which the majority of predicting models are not able to analyze. In fact this database had these type of variables.

```
RangeIndex: 22254 entries, 0 to 22253
Columns: 141 entries ID_CONTACTO to Categoría de sentimiento
validada manualmente
dtypes: datetime64[ns](3), float64(19), int64(96),
object(23)
memory usage: 23.9+ MB
```

   d) **Change the variables that were not float or int**

      I changed all the variables that were not float or int into codes using the function “astype('category').cat.codes”

e) Fill the Nan cells

There were some columns that have some cells that were Nan so I changed to 0

f) Describe our target

I thought that it was important to know how our target was distributed and how many 0 and how many 1 we have, so it would be possible to know if later I will have to use any imbalanced method, and the result was

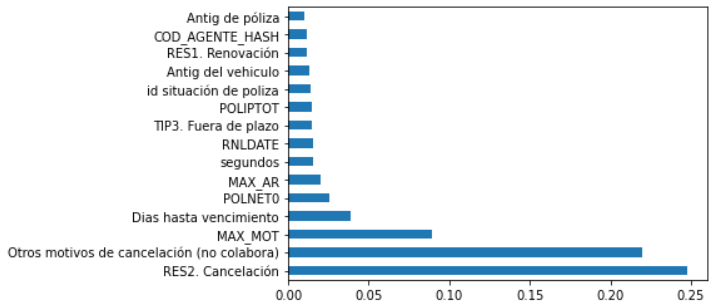
```
0    14101
1     8153
```

g) Correlation

I watched how the variables were correlated with our target, also I noticed that some variables were correlated between them more than 0.9 so I decided to clear one of them. But there was not any variable correlated with our target more than 0.5

h) Shap Values

The last thing that I did was to get the Shap values to show the most important variables for our predicting model.



2) Predicting Models

A predictive model is a set of processes carried out through computational data analysis techniques that help to infer the probability that certain situations will occur prior to their achievement.

I start with the models that I learned during my class Data Analysis and then my tutor showed me other type of models that were more effective.

a) Dividing the data in training and testing.

I used the 80% of the data as training and the other 20% to test if the model was okay.

b) Logistic Regression

It is a regression method that allows estimating the probability of a binary qualitative variable as a function of a quantitative variable. One of the main applications of logistic regression is binary classification, in which observations are classified into one group or another depending on the value of the variable used as predictor.

It is one of the most basic models, but it allows us to see how much work we will have to do in the next models.

The results with this method were quite bad. The accuracy was just 64% and the recall of the 1s were just 6% so it was not really useful for our purpose.

	precision	recall	f1-score	support
0	0.64	0.98	0.77	2807
1	0.64	0.06	0.11	1644
accuracy			0.64	4451
macro avg	0.64	0.52	0.44	4451
weighted avg	0.64	0.64	0.53	4451

c) Decision Tree Classifier

A decision tree is a map of the possible outcomes of a series of related decisions.

The results with this method were much better than with the Logistic Regression.

	precision	recall	f1-score	support
0	0.89	0.89	0.89	1818
1	0.82	0.82	0.82	1076
accuracy			0.86	2894
macro avg	0.85	0.86	0.85	2894
weighted avg	0.86	0.86	0.86	2894

As we can see the accuracy was 86% and the recall for the 1s was 82%.

Then I tried with the other methods that I have learnt in class, but they did not improve the results of the Decision Tree Classifier. The ones that I tried were:

d) Decision Tree Regressor
e) Random Forest Classifier
f) Random Forest Regressor
g) Bernoulli Naïve Bayes
h) Gauss Naïve Bayes
i) Multinomial Naïve Bayes

My Tutor asked me to try with new methods and they were able to improve the results.

j) XGBoost

XGBoost is a popular and efficient open-source implementation of the Boosted Trees Gradient algorithm. Gradient boosting is a supervised learning algorithm, which tries to accurately predict a target variable by combining the estimates of a set of simpler and weaker models.

The results with this method were much better.

	precision	recall	f1-score	support
0	0.93	0.94	0.94	2832
1	0.89	0.88	0.88	1619
accuracy			0.92	4451
macro avg	0.91	0.91	0.91	4451
weighted avg	0.92	0.92	0.92	4451

The accuracy and the recall were bigger than the ones with the other methods.

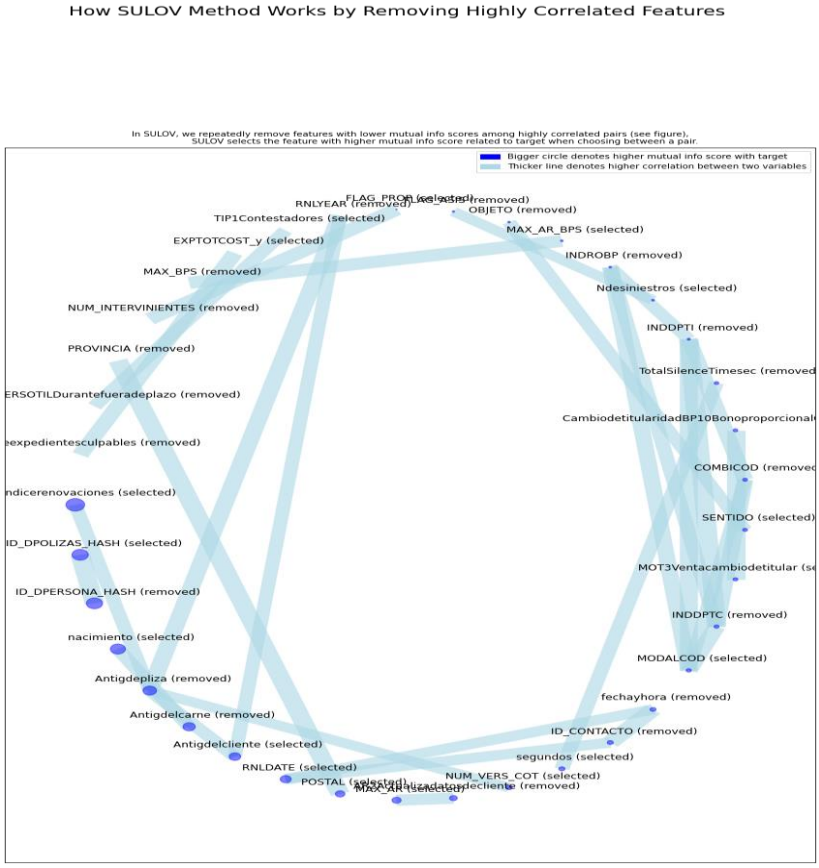
k) LightGBM

Unlike XGBoost (among others) which uses presort-based algorithms, LightGBM uses histogram-based algorithms (ie, bin continuous attribute values into discrete bins) to speed up training and reduce memory usage.

This method got similar results as XGBoost but it was much faster.

	precision	recall	f1-score	support
0	0.93	0.94	0.93	2832
1	0.89	0.88	0.88	1619
accuracy			0.91	4451
macro avg	0.91	0.91	0.91	4451
weighted avg	0.91	0.91	0.91	4451

To try to improve the models I watched in a webpage called cienciadedatos.com that there was a function that select the most representative variables of the database, the method is called Select K Best.



But after doing this I test the XGBoost method and the LightGBM and results did not improve.

3) Conclusions:

The best predicting model resulted XGBoost.

The variables more correlated with our target were:

- Cancelación,
- Otros motivos de Cancelación
- and MAX_MOT

and these ones were the ones where the company (Verti) has to be focused in future calls.

MAPFRE SPAIN (Wednesday 8 hours)

Mapfre[5] is a global insurer made up of a team of professionals, women and men, from all over the world. At MAPFRE they work to create value for all the people with whom we interact. They are a company also committed to sustainability and the development of people in the countries where they are present.

They have 18 databases of each month since January 2021 until April 2022, they have information about the clients and they asked us to merge all the databases and to observe, per month, how many clients cancel their subscription, how many clients renew and how many new clients they get each month

To be able to do this I follow the next steps:

1) **Observe the data bases variables**

   I watch which variables were common between all the databases and my tutor told me that the way to watch if the client renews or not, I had to observe the column "Policy ID" and check if it exists in every month or there was any month where it does not appear.

2) **Merge the databases**

   I used pandas to merge all the data, I was combining the databases one by one, first I merge January 2021 with February 2021 next I merge this new database with March 2021 and I continue doing this until I join all the databases.

I had some problems of time, the computer that I had just have 4GB of RAM and these databases have at least one million lines each of them so to do this I spent a lot of time and I was not able to do another thing while the program was running because if I try to do anything more the computer gets overheat.

[5] MAPFRE. "Home." *Grupo MAPFRE Corporativo - Acerca de MAPFRE*, 8 Aug. 2022, www.mapfre.com.

3) **Checking the ID Policy that did not repeat at the next month**

I used a for loop to do this assignment, I checked if the ID Policy were repeated in the next month, also I observed how many new Policy were new in each month.

4) **Problem**

At the middle of this process my tutor and I observed something strange. There were to many new ID Policy and to many cancellations, but there were few renewals.

So, we decided to explore the reasons for this and we realized that there was another column that was Client ID, and that sometimes the Client have different ID Policy in each month. This was because when they renew their contract, sometimes they change their conditions with the company so the ID Policy change.

5) **Solving the problem and Conclusion**

At the end we decided to do the analysis with the Client ID column and the results were really different. 90% of the clients renew each month, there was almost 100.000 new clients each month.

The months with more new clients were January and September and the months where there were more cancelations were June and December.

SANTANDER BANK MEXICO (Thursday-Friday 16 hours)

Santander[6] develops products and services adapted to the needs of customers, taking advantage of global experiences but taking local characteristics into account. Thus, they offer financial services to all types of clients: Individuals, Companies, Institutions, Corporations, Private Banking and Universities.

This Project was very similar to the first one. We had a database with data of the telephone calls between customer and client and we had to generate predicting models to know if the client was satisfied or not with the call. Our target column was one called "NPS" and there were 3 types of answers, if the client was satisfied the number was 1, if he was not the number was -1 and if he did not have a clear position about how the call was, the answer was 0.

I started doing the same things as I did with the first project and the process was faster because I just have to watch what I did before and copy it, just changing the target and the data base.

1) **Descriptive analysis.**

   a) **Import the database.**

      I used pandas to import the database and then be able to do a descriptive analysis.

   b) **Watch the variables**

      There was a lot of different data about the conversations like how much time was he conversation, what type of conversation (it was the customer who called or the client), about what they talked (problem with app, want to buy some actions, want to change euros to another type of money, etc), there was a column with all the conversation transcript and our target that was the NPS.

[6] Santander, Banco. "PARTICULARES." *Banco Santander*, www.bancosantander.es/particulares. Accessed 15 Aug. 2022.

c) Types of variables

In this database there was different types of data there was object, float and int variables so I had to take this in count to later make a properly predicting model.

```
Int64Index: 22222 entries, 0 to 22221
Columns: 179 entries, ID_CONTACTO to Columna1
dtypes: datetime64[ns](3), float64(19), int64(132), obje
ct(25)
memory usage: 30.5+ MB
```

d) Change the variables that were not float or int

I hanged all the variables that were not float or int into code category using this “astype('category').cat.codes”

e) Fill the Nan cells

There were some columns that have some cells that were Nan so I changed them and instead of Nan there were 0

f) Describe our target

It was important to know how our target was distributed so I show how many different possibilities were in each case in the NPS column

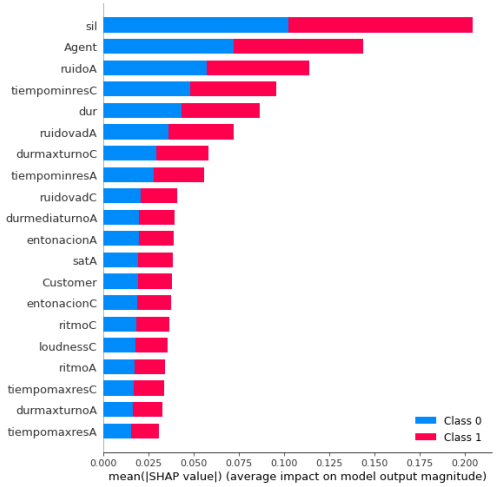
```
0    15121
1     4879
-1    2222
```

g) Correlation

I observed if there was any variable that was extremely correlated with the target but none of them have a correlation greater than 0.5. However there were some of them that were really correlated between them so I removed one of each to speed up the process.

h) Shap Values

The last thing that I did was to get the Shap values to show the most important variables for our predicting model.



1) Predicting Models

As I did in the first project I started to try different predicting models to observe which one was the best to predict the unsatisfied client after the telephone call.

a) Dividing the data in training and testing.

I used the 85% of the data as training and the other 15% to test if the model was okay. I selected more data to train because I was told that if the data base is not to long is better to use more data to train.

Because of the results of the I decided to use directly the best predicting models of the first project that I did.

b) XGBoost

The results were not bad at all, I put that it has to predict correctly the -1 results in the target column so I told the program to predict the answers that were < 0. This ones were the results.

	precision	recall	f1-score	support
False	0.90	0.90	0.90	2470
True	0.87	0.88	0.87	1975
accuracy			0.89	4445
macro avg	0.89	0.89	0.89	4445
weighted avg	0.89	0.89	0.89	4445

c) Gradient Boosting

This one was also one of the best predictors in the last project, so I also used it. The results were these ones.

	precision	recall	f1-score	support
False	0.95	0.71	0.81	2470
True	0.72	0.95	0.82	1975
accuracy			0.81	4445
macro avg	0.83	0.83	0.81	4445
weighted avg	0.85	0.81	0.81	4445

As we can see the recall was much better to our purpose but the precision was worse.

d) LightGBM

The last predict model that I used in this project was this one, as before it was much faster and the results were similar to the GradientBoosting.

	precision	recall	f1-score	support
0	0.95	0.76	0.84	2427
1	0.77	0.95	0.85	2018
accuracy			0.85	4445
macro avg	0.86	0.85	0.85	4445
weighted avg	0.87	0.85	0.84	4445

It has more precision but it still being a little bit bad, so I looked some ways to improve the model.

I found in a web called www.analyticslane.com the method Grid Search CV. GridSearchCV is a class available in scikit-learn that allows you to systematically evaluate and select the parameters of a model. By telling it a model and the parameters to test, you can evaluate the performance of the former against the latter using cross-validation.

So I used Grid Search with the models that I try before and it improve a little bit the models.

1) XGBoost

	precision	recall	f1-score	support
False	0.92	0.91	0.90	2470
True	0.88	0.90	0.90	1975
accuracy			0.90	4445
macro avg	0.90	0.90	0.90	4445
weighted avg	0.90	0.90	0.90	4445

2) Gradient Boosting

	precision	recall	f1-score	support
False	0.95	0.75	0.83	2470
True	0.75	0.95	0.83	1975

accuracy			0.85	4445
macro avg	0.85	0.85	0.85	4445
weighted avg	0.85	0.83	0.83	4445

3) LightGBM

	precision	recall	f1-score	support
0	0.95	0.78	0.84	2427
1	0.79	0.95	0.85	2018
accuracy			0.87	4445
macro avg	0.88	0.87	0.85	4445
weighted avg	0.87	0.87	0.84	4445

These results were not bad but I was able to improve the results considerably when I learnt the function Bag of Words and TfidfVectorizer. "Bag of words is a Natural Language Processing technique of text modelling. In technical terms, we can say that it is a method of feature extraction with text data. This approach is a simple and flexible way of extracting features from documents.

"A bag of words is a representation of text that describes the occurrence of words within a document. We just keep track of word counts and disregard the grammatical details and the word order. It is called a "bag" of words because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document."[7] An Introduction to Bag of Words in NLP using Python | What is BoW? (mygreatlearning.com)

[7] Great Learning Team. "An Introduction to Bag of Words in NLP Using Python | What Is BoW?" *GreatLearning Blog: Free Resources What Matters to Shape Your Career!*, 5 May 2022, www.mygreatlearning.com/blog/bag-of-words.

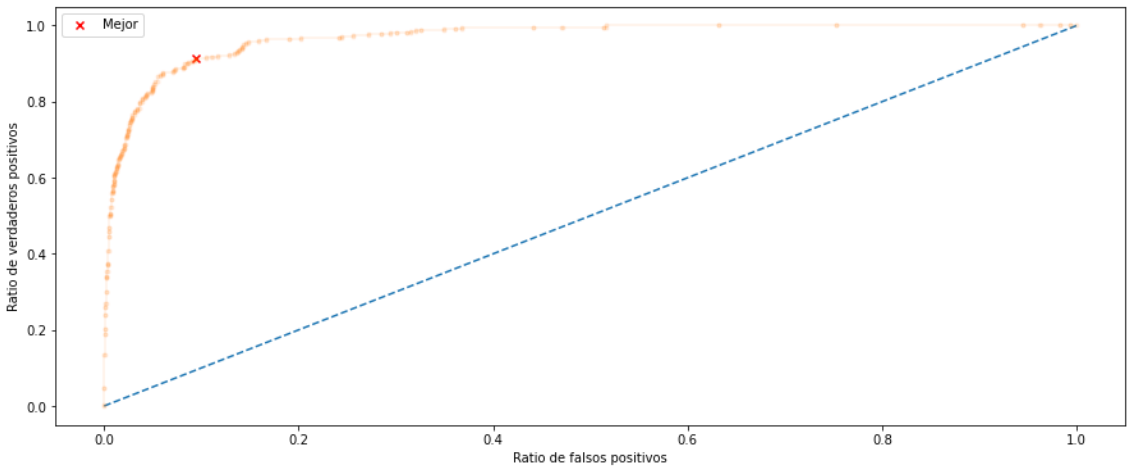
TF-IDF is a statistical measure that assesses how relevant a word is to a document in a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document and the inverse document frequency of the word in a set of documents.

These two new tools were really useful because one of the columns were the transcripts of the telephone calls and with this new thing I was able to improve considerably the results.

Using TFidf Vectorizer:

1) XGBoost

	precision	recall	f1-score	support
False	0.99	0.91	0.95	2470
True	0.89	0.91	0.90	1975
accuracy			0.94	4445
macro avg	0.94	0.91	0.90	4445
weighted avg	0.96	0.91	0.92	4445

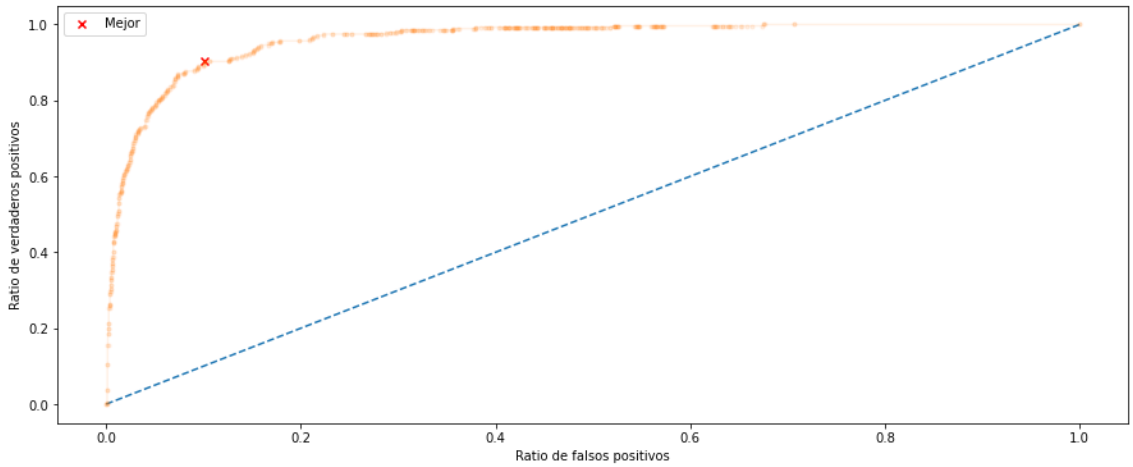


2) Gradient Boosting

	precision	recall	f1-score	support
False	0.99	0.90	0.94	2470
True	0.80	0.90	0.85	1975
accuracy			0.90	4445
macro avg	0.89	0.90	0.73	4445
weighted avg	0.95	0.90	0.92	4445

3) LightGBM

False	0.82	0.93	0.87	2427
True	0.90	0.75	0.82	2018
accuracy			0.85	4445
macro avg	0.86	0.84	0.84	4445
weighted avg	0.85	0.85	0.85	4445

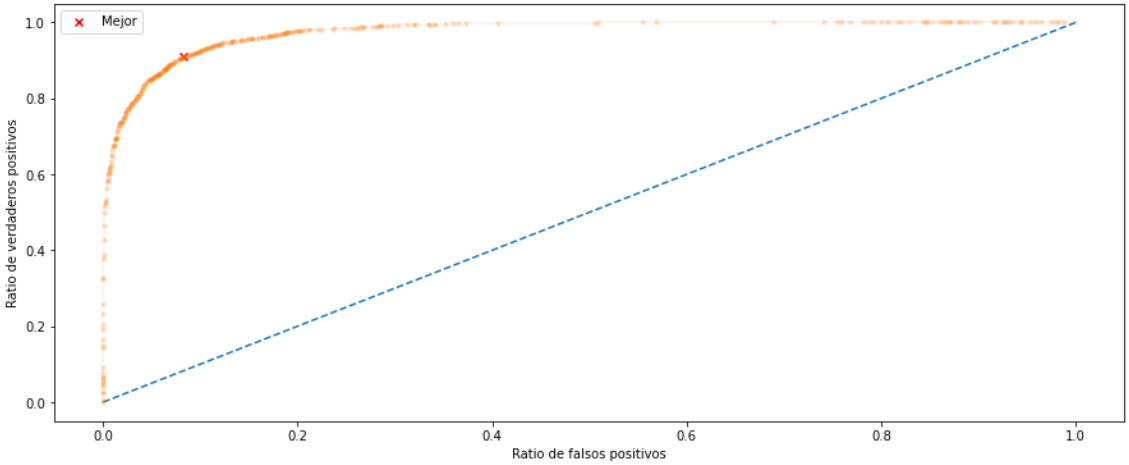


Using Bag of words:

1) XGBoost

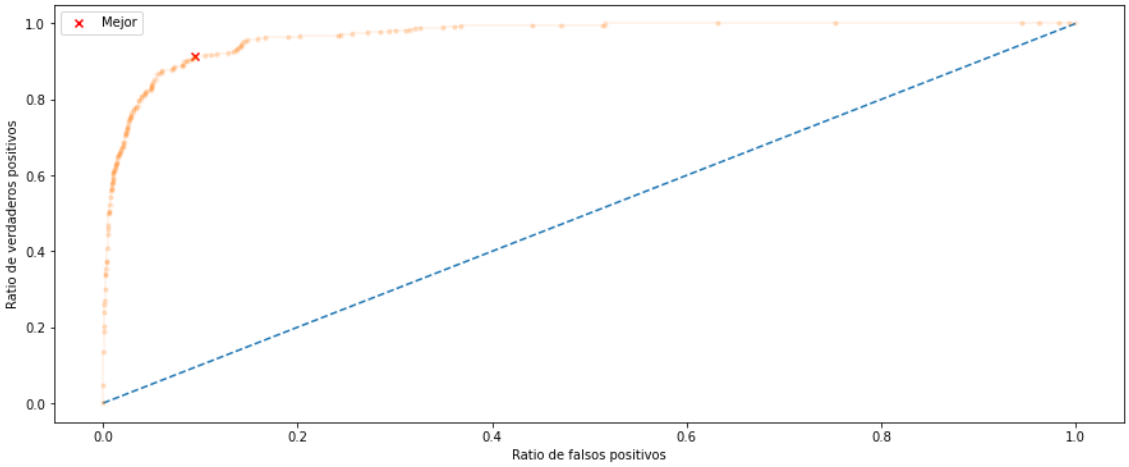
	precision	recall	f1-score	support
0	0.93	0.92	0.92	2533

1	0.89	0.91	0.90	1912
accuracy			0.91	4445
macro avg	0.91	0.91	0.91	4445
weighted avg	0.92	0.91	0.91	4445



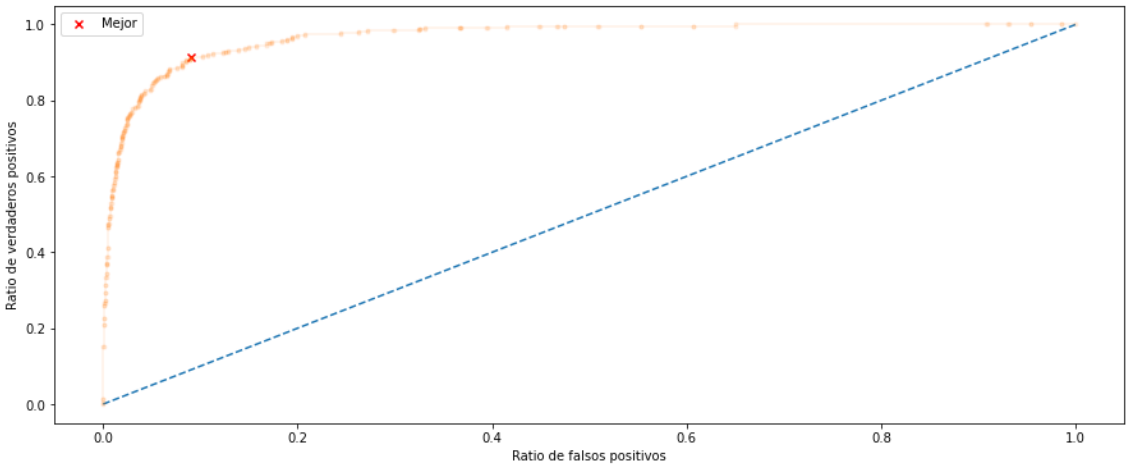
2) Gradient Boosting

	precision	recall	f1-score	support
0	0.99	0.91	0.95	4168
1	0.39	0.91	0.55	277
accuracy			0.91	4445
macro avg	0.69	0.91	0.75	4445
weighted avg	0.96	0.91	0.92	4445



3) LighGBM

False	0.82	0.93	0.87	2427
True	0.90	0.75	0.82	2018
accuracy			0.85	4445
macro avg	0.86	0.84	0.84	4445
weighted avg	0.85	0.85	0.85	4445



5. CONCLUSION

During my internship in Neovantas I have learnt a lot of new things and developed my skills in the Data Science area.

There have been some hard moments where I was blocked, and I did not know what to do to improve the programs that I was developing but I have had always the support of my tutor and my teammates.
The most relevant things that I have learnt have been:

Predicting models:

a) I have learnt new predicting models like XGBoost, LightGBM and Gradient Boosting that have better results that the ones that I knew before and there were also much faster.
b) I have been taught how to improve the predicting models changing their parameters with the function Grid Search that let you know what the best parameters for each model were.
c) I have developed new skills about showing my results of the predicting model, so the person that asked me to do the programs were able to understand better the results.
d) Also, I have learnt about the ROC and AUC that were terms that I did not learn before and there were really important to evaluate the performance of the model.

Data Bases:

a) I learnt how to manage different types of data base document like txt, csv or xlsx.
b) I have noticed that the data bases in real life usually are much more complicated that the ones that I was used to manage in the university classes, they have problems like having cells with no data, columns having data in a wrong form (like having the name of the person and his ID in the same cell, so you have to transform it and separate them, or there is some times that they have wrong information when you change an excel document to csv document).
c) Also, there is some Data Bases where the column target is really unbalanced so for example in one small project where the target were a binary response and there were 90% of 0 and the other 10% were 1. In this type of cases the predicting model was able to predict really good the 0s but not the 1s so I learnt how to create imaginary variables using oversampling and under sampling Imbalanced-Learn module in Python - GeeksforGeeks and this improve considerably the results.

Teamwork:

a) I have learnt to work in different projects with several teammates and I was able to adapt to each colleague and the way that they use to work.
b) I was able to get support of my tutor, specially, and my teammates whenever I had any problem, or I was blocked.

Finally, I want to thank Neovantas for the opportunity of doing this summer internship in their company and making it easy for me to work there. Also, I want to mention the importance of my tutor, Adrián Gonzalez, who has taught me most of the things that I have learnt during my internship and who was a big support since I began to work.

It has been a really good experience to work there, and I know that everything that I have learnt in this company will be really helpful in my future jobs.