Contents

List of Figures

General introduction

Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.

The process of learning begins with observations of data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. In turn, these systems generate insights which analysts and business users can translate into tangible business value.

Nowadays,more and more companies are coming to realize the importance of data science, AI, and machine learning. Regardless of industry or size, organizations that wish to remain competitive in the age of big data need to efficiently develop and implement data science capabilities or risk being left behind. And since , the insurance industry is regarded as one of the most competitive and less predictable business spheres and it is instantly related to risk and dependent on statistics ,data science is considered as the best solution to change this dependence forever.

As we have mentioned before the dependence of insurance companies on data science has been increasing for years , for that reason we were approached as beginner data scientists to create a model that will help the General Committee of insurance detect numerous types of fraud in their existing data . Furthermore , the model that we provide well help affect the convenient Bonus Malus class to all new insured costumers.

Chapter 1

Project Context

“

The use of cars in the world and more precisely in Tunisia has evolved . According to the "Agence Technique des Transports Terrestres" (ATTT), the number of cars in Tunisia is estimated at the end of 2016, at nearly 2 million vehicles on Tunisian roads and is growing by 70 to 80,000 vehicles per year. the total new vehicle market was 63,685 registrations, an increase of 5 compared to 2016. It is then predictable that by increasing the rate of cars present in Tunisia, the legal problems will also increase. Moreover, legally speaking, any natural person or any legal person, whose civil liability may be incurred in connection with the operation of a land motor vehicle and its trailers, must conclude an insurance contract guaranteeing the liability it may incur as a result of damage to people and property caused by the vehicle. The insurance contract covers the civil liability of the contractor, the owner of the vehicle and any person having custody or control of the vehicle. This obviously pushes us to reconsider problems in this regard. That is to say the increase in fraud, and traffic compared to the bonus malus class.

”

1.1 Project context :

Following the exponential growth in the volume of data and information must be insured, the CGA (General insurance committee) has encountered follow-up problems and analysis in the field of automobile insurance. Among these, we quote the calculation of some data is done manually which can give incorrect results. So the data analysis and reporting is done at the end of each month and not on the day of the day, which causes a great delay in taking action. Moreover,the interface offered by Excel is limited. Since the monthly monitoring data of the insurance companies are stored in several separate Excel files, this results in a loss of time. All these limits and this lack leads us to opt for a project that aims to collect all the information of all the insurance companies in a data by ensuring their persistence and preventing them from being lost. In addition, providing reliable, consistent and relevant information through the automation of calculating, disseminating data (internal and external) and updating reports on a daily basis, offering a manager the opportunity to conduct appropriate analyses along several axes. The goal is of course to facilitate decision-making through relevant indicators and to reduce the costs of the current analysis and reporting procedure.

Keywords: Class bonus-malus system, claims, frauds.

1.2 Data Science methodology :

Data science is essential in the insurance field. And to develop a solution that will help us solve the problems of the CGA, while relying on well-determined methodology. We have used CRISP which is The Cross Industry Standard Process for Data Mining (CRISP-DM). It's a process model with six phases that naturally describes the data science life cycle. It's like a set of guardrails to help you plan, organize, and implement our data science project. The 1.1 contains the methodology :

1-Business understanding – What does the business need?

2-Data understanding – What data do we have / need? Is it clean?

3-Data preparation – How do we organize the data for modeling?

4-Modeling – What modeling techniques should we apply?

5-Evaluation – Which model best meets the business objectives?

6-Deployment – How do stakeholders access the results

Figure 1.1 – CRISP Methodology

Conclusion :

Understanding the project and its hidden flaws, reviewing all applications was a crucial step for us. Putting ourselves in the context of data science while choosing the work methodology that we will follow, was also an indispensable second step. All this for one purpose only; to succeed in the project. But, it did not stop there. On the contrary, it is now that everything will start correctly.

Chapter 2

Business Understanding

Introduction :

> *The first step to every data science project based on the Crisp DM methodology is Business Understanding which focuses on understanding the objectives and requirements of the project. How to succeed without understanding the trade in question.In this context, we launched our research to be able to discover the world of automobile insurance first. Precisely, the CGA (the General Insurance Committee), how it works with it. We also worked to analyze all the CGA requests, all their shortcomings, their limitations, etc. We gradually discovered the technical words that are related to insurance, the traffic, the evolution in the world of cars, and everything that will help us to end up with a solution that is well suited to this area.*

2.1 CGA description :

The CGA (General Insurance Committee) in its first definition is an agreement by which an insurance company undertakes, in the event of realisation of the risk or at the term fixed in the contract, to provide the insured with a financial benefit in return for remuneration known as a premium or contribution.

It shall ensure the protection of the rights of policyholders and beneficiaries of insurance contracts and the soundness of the financial base of insurance undertakings and reinsurance undertakings and their ability to honour their commitments. Within the framework of the tasks assigned to it, the Committee shall be responsible in particular for: The control of insurance undertakings, reinsurance undertakings and professions related to the insurance sector and the monitoring of their activities, consideration of legislative, regulatory and organizational matters relating to insurance and reinsurance transactions, insurance undertakings and reinsurance undertakings submitted to it by the Minister for Finance and the preparation of draft texts relating thereto at his request, the study of technical and economic matters relating to the development of the insurance sector and its organisation and the submission of proposals to the Minister of Finance for this purpose, and in general, to study and give an opinion on any other matter within its remit.

Figure 2.1 – CGA logo

Conclusion

The stage of understanding the business and the whole field helped us to immerse ourselves in history. It has enabled us to tackle the next steps, which are solid and knowledgeable in the field. Without this passage, everything that comes, would be much more difficult for us to attack.

Chapter 3

Data understanding and data analytics

> *In developing the CRISP methodology, a question arose : What data do we have need? Is it clean? And now we have to answer. We must not forget to look at the business objectives, or on other terms the demands of the CGA (General Insurance Committee) and how the objectives of data science will go in parallel. Going further, the data collection step will be required. We will end up with a detailed description of everything that is provided to us in our data set and in our external data collected to succeed in our decision-making at the end.*

3.1 Business objectives :

Our main objective is to help the CGA (General Insurance Committee) to successfully detect fraud so that it can be minimized afterwards. It will also be able to predict the Bonus-Malus class of its upcoming customers according to the history of their old ones.

3.2 Data Science objectives :

The technical objectives are parallel to the business objectives. Nothing is accomplished without the other. Therefore, we used different techniques:

- *Featuring engineering:*

This step consists of cleaning up the data provided in order to give them meaning and make them more relevant.

- *Web scrapping:*

In order to have more data that will help us do things more credibly, we used the collection of data on the sites.

- *Machine learning:*

This is reflected in the application of several learning models (supervised and not supervised) to classify our data.

- *Reporting:*

Visualizations that summarize data and help to better understand.

3.3 Data collection :

What has been provided to us will certainly do the job, but within limits. It is preferable that we add more data. The research was long, we asked ourselves the question :

What will strengthen our decision-making?

In our case, and to go further with the causes of the claims made, we opted for information that we consider solid. And because no one can deny the effect of weather on the accident rate, we ended up focusing on this point. As long as the work is done in our country, research must import data from Tunisia. In this context, a reliable site was used. We subsequently collected everything we needusing **web scrapping** from this web site :

https://www.historique-meteo.net/?fbclid=IwAR1RfVx68zdgVfPMv1lI327i-szMuLwunBnyM0rr6_g8DhWML-oBWwuwjYA

	lieu	date	temperatureMaximale	temperatureMinimale	vitesseVent	Humidite	Visibilite	etat	ajuste
0	Tunis	01/01/2017	12°	11°	12km/h	61%	10km	Nuageux	0mm
1	Tunis	02/01/2017	14°	11°	9km/h	10km	4%	Ciel dégagé pleinement ensoleillé	63%
2	Tunis	03/01/2017	15°	12°	23km/h	54%	10km	Soleil et partiellement nuageux	0mm
...	...	...	...	...	...	...	...	...	...
1092	Zarzis	29/12/2019	15°	13°	34km/h	61%	9.875km	Faibles averses de pluie	5mm
1093	Zarzis	30/12/2019	13°	11°	35km/h	67%	9km	Faibles averses de pluie	11mm
1094	Zarzis	31/12/2019	14°	11°	33km/h	67%	8.875km	Pluie forte ou modéré	14mm

Figure 3.1 – Data collection

3.4 Data description:

Despite the first step in understanding the business, which has clarified so much about the insurance field, we are still unclear until we discover the data.

What does each mean?

What is the relationship between them?

We have summarized all the relationships that connect the tables in this diagram 3.2

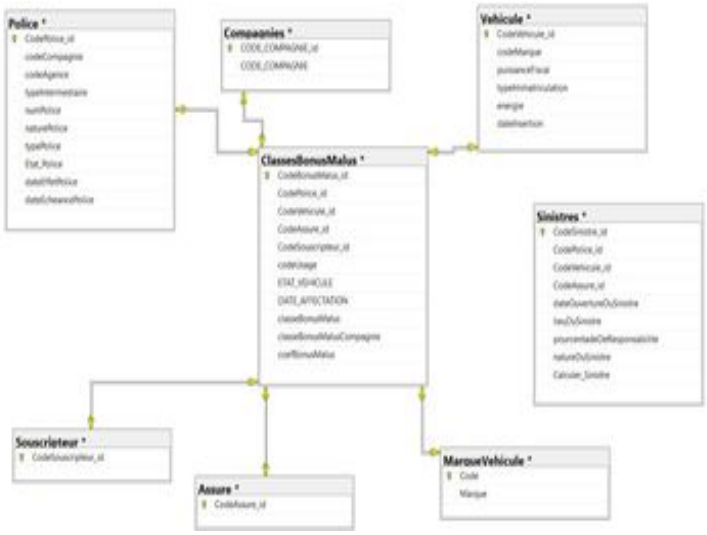

Figure 3.2 – Global schema

For more details on each attribute of every table :

- **The table 'Assure' contains the following attributes :**

  -**Id**: Insured ID

  -**odetypepieceidentite**: 1 (cin), 2 (tax number), 3 (passport) and 4 (residence card)

  -**CodePostal**: Insured City Postal Code

- **The table 'ClassBonusMalus' contains the following attributes :**

  -**Id**: The bonus-malus class id

  -**CodePoliceId**: Contract id

  -**CodeVehciuleId**: Vehicle id

  -**CodeAssureId**: Insured id

  -**CodeSouscripteur**: Agent id

  -**CodeUsage**: type of use

  -**Vehicle type**: R (Terminated) or V (in force)

  -**Dateaffectation**: Date of assignment of the bonus-malus index

  -**ClasseBonusMalus**: This is the degree that varies between 1 and 11 and indicates the percentage of the premium thereafter that varies from 70 to 350 .

  -**ClasseBonusMaluesCompany**: Class of vehicle relative to company

  -**CoefBonusMalus**: Payment percentage relative to initial amount

- **The table 'Compagnies' contains the following attributes :**

  -**Codecompagnieid**: Company id

  -**Codecompagnie**: Company code

- **The table 'MarqueVehicule' contains the following attributes :**

  -**Code** :Vehicle code

  -**Marque**: The car brand

- **The table 'Police' contains the following attributes :**

  -**Id**: police id

  -**CodeCompagnie**: company number

  -**CodeAgence**: Agency office number

-**TypeIntermediary**: 1 (agent) 2 (office) and 3 (broker)

-**CodeCourtierCGA**: If the intermediary type is broker it takes a code otherwise it takes null

-**NumPolice**: Contract number

-**NaturePolice**: R (renewable) and T (temporary)

-**Dateeffetpolice**: Date of commencement of guarantee for the insured, when does insurance have an effect

-**Dateecheancepolice**: The date on which the insurance contract ends or it is automatically renewed when the policy nature is renewable

-**Dateexpiration**: Date on which the insurance contract ends when the nature of the temporary policy

-**Verouillagemodifpolice**: The change is made by adding a guarantee, removing a guarantee exclusion, changing the deductible or the ceiling, etc.

-**Etatpolice**: V to say Vigor which means that the vehicle still works, R to say terminated or wreck and S to say suspended

-**Dateresilation**: The date when the status is of the terminated type

-**Datesuspension**: The date when the status is suspension type

-**Dateremiseenvigueur**: The date when the status is of the force type

-**PoliceType**: I(individual) or F(floating)

-**DateEffetPolice**: Start date of guarantee for the insured

- **The table 'Sinistre' contains the following attributes :**

  -**Id**: Sinister ID

  -**Cgapoliceid** : CGA ID

  -**Cgavehiculeid** : CGA vehicle ID

  -**Cgaassureid** : CGA insurer ID

  -**Date**: Sinister date

  -**Numerodusinistre**: Sinister number

  -**datedesurvenanceDusinistre**: Sinister date

  -**Heuesurvenancedusinistre**: Specific Time

  -**Lieudusinistre**: Place where the sinister occured

  -**Pourcentageresponsabilité**: Percentage of responsibility

  -**Pourcentagecompagnieadverse**: Percentage of liability for the other party

-**Numerodepolicecompagnieadverse**:Opposing Company Policy Number

-**Codecompagnieadverse**: Adverse company code

-**NatureDuSinistre**: Material or body (M or C)

-**CalculerSinistre**: Takes the value 0 or 1, 1 for calculated and 0 for not calculated

- **The table 'Souscripteur' contains on attribute :**

  -**Codesouscripteurid** : Subscriber ID

- **The table 'Vehicule' contains the following attributes :**

  -**Id** :Vehicle id

  -**NumChassis**: Chassis number

  -**Codemarque** : Vehicle code

  -**Puissancefiscale**: Fiscal power

  -**Typeimmatriculation**:TU or others

  -**Energie**: The fuel

  -**Dateinsertion**: Date of entry into circulation

  -**Datedernierevisite**: Date of last visit

  -**DateAjout**: The date it was added to the table

  -**Etatvehicule**:V or E

  -**DateMisenepave**: Date the car no longer works

  -**DateRetrait**: Date of removal

  -**Datedemiseencirculation**:Date when the car drove

  -**Datemiseajourvehicule**: Date when the contract is updated

To answer the question of the relationship between our tables, here are some examples in the graphs 3.3and 3.4 below:

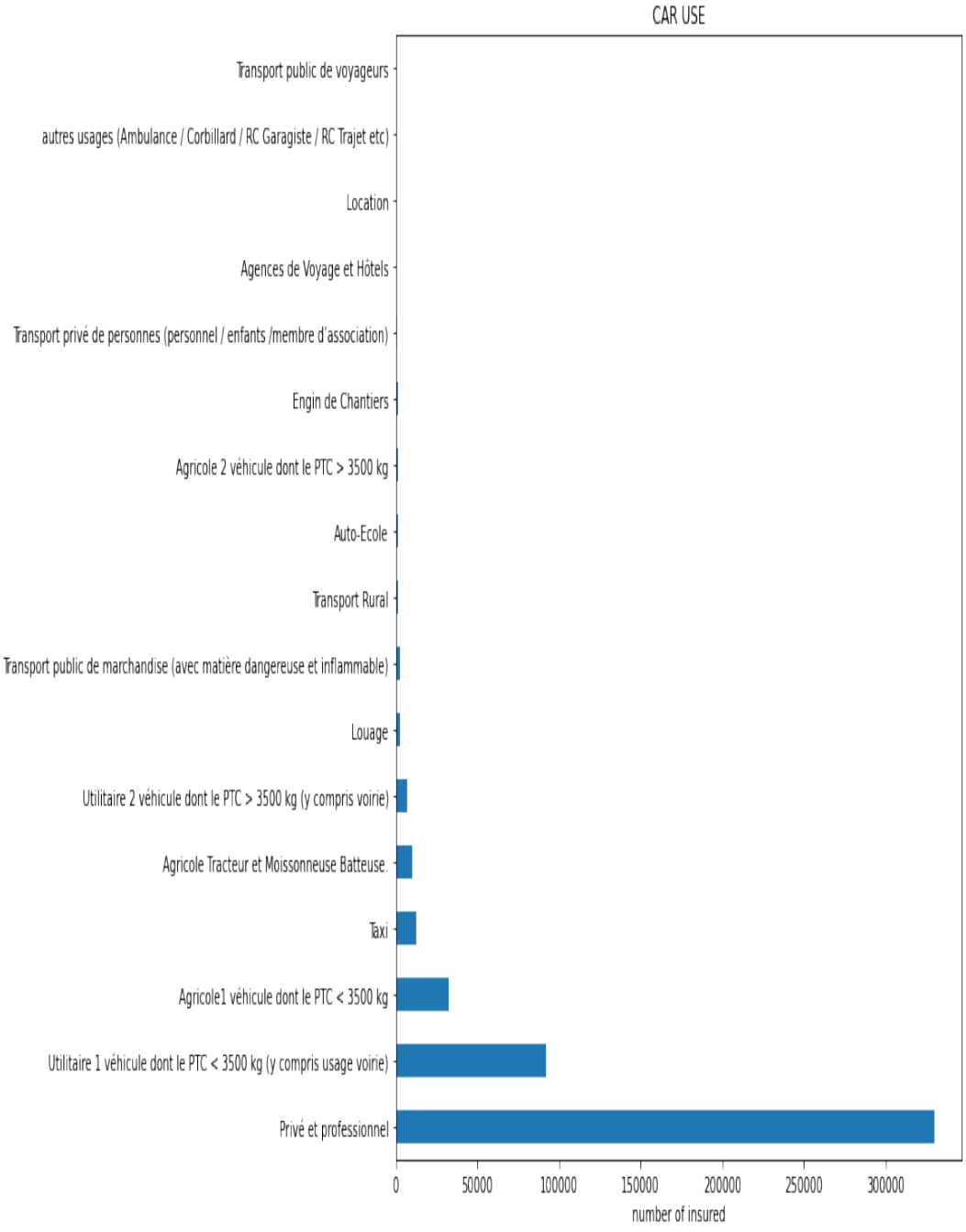

Figure 3.3 – Number of insured per use

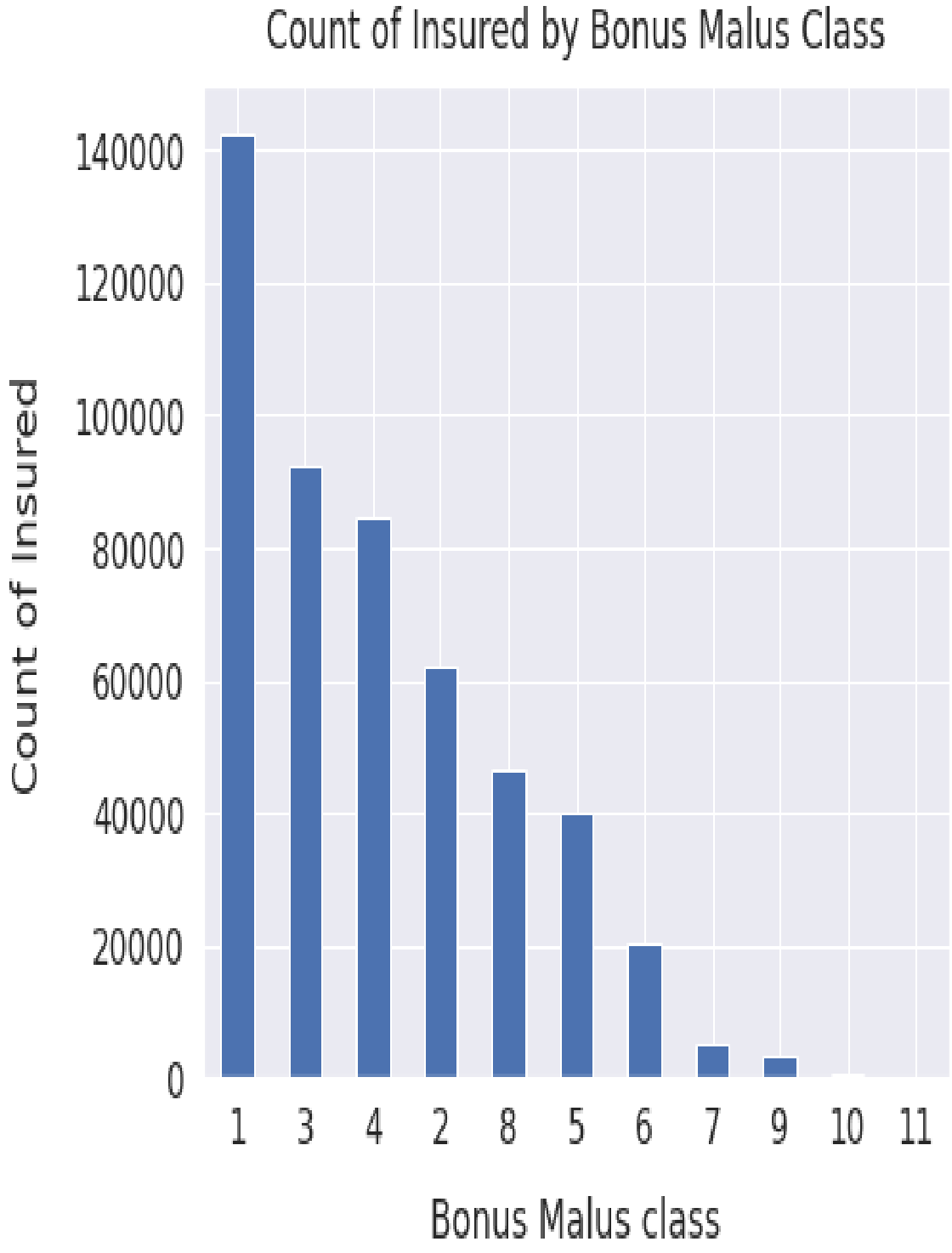

Figure 3.4 – Count of insured by bonus-malus class

Conclusion:

It is true that throughout this stage, we managed to understand our data, collect others and prepare all the favorable ground for the work. But anything we save is not enough to make a perfectly accurate prediction, because not all the data are relevant. There are missing data, others that are useless. All this leads us to a next step, very important in data science projects which is the preparation of data.

Chapter 4

Data Collection and Preparation

> *We have reached the third phase of this project which is the data collection and preparation phase .This is the stage of the project in which we decides on the data that we're going to keep for analysis. The main criteria that helped us make a decision is the relevance of the data , its quality.*

4.1 Internal Data :

4.1.1 Bonus Malus table :

These are the steps we followed to have clean , coherent data : We started by merging tables from csv and xls files leading with changing the columns names so both tables can have the same headers. After that, we treated each resulted table (police,vehicule..) individually and checked for empty/nearly empty columns then we deleted them , we also changed the format of some date provided . In order to detect all types of fraud and affect the convenient bonus malus class , we needed to have all the data gathered in one table reassembling the information that we needed . To make that happen we merged the tables "Vehicule" , "Police", "Assure","UsageVehicule","Marque vehicule" with the class that contained the variable to study "Bonus malus class", we decided not to merge the table "Sinistre" because adding it to the previous ones will deviate the model later since only insured clients with accidents will be taken into consideration. The 4.1 containes the columns that we were left with after the merge and the cleaning.

CodeMarqueVehicule	0
Marque	0
Code	0
Usage	0
CodeAssure_id	0
CodePolice_id	0
codeCompagnie	0
codeAgence	0
typeIntermediaire	0
numPolice	0
naturePolice	0
typePolice	0
Etat_Police	0
dateEffetPolice	0
dateEcheancePolice	52889
dateRemiseEnVigueure	497869
dateRemiseEnVigueureajuste	497869
CodeVehicule_id	0
numChassis	488071
codeMarque	0
puissanceFiscal	0
numImmatriculation	488071
typeImmatriculation	0
energie	0
dateInsertion	0
dateAjout	488071
etatVehicule	488071
dateMiseCirculation	497869
dateMiseAjourVehicule	497865
CodeBonusMalus_id	0
CodeSouscripteur_id	0
codeUsage	0
ETAT_VEHICULE	0
DATE_AFFECTATION	0
classeBonusMalus	0
classeBonusMalusCompagnie	0
coefBonusMalus	0

Figure 4.1 – Bonus Malus table columns

Now that we have all the data gathered in one place ,we started to clean the table as a whole : **Firstly** ,We deleted all the columns that were unnecessary to realise our prediction, such as :"CodeAssureid", "DATEAFFECTATION", "ETATVEHICULE", "dateAjout" and many others , the mentionned columns were removed due to the huge lack of needed information or simply its irrelevance . **Secondly**,the next step was calculating the driving experience for each client simply by subtracting the driving licence obtaining date from the current year ,we then filled all the missing data using the KNNImputer method and we made sure that the values describing that variable didn't change in the process. **Thirdly**, To go even further and to guarantee a more accurate result in the modelling phase we converted weak modalities in each column with the value "autres" to insure that all that weak information can be transformed into a stronger piece of data. **Fourthly**, yet following the same spirit of changing modalities , we noticed that we had more that 350 car brands which inspired us to gather car having the same car group together for example (BMW=["MINI","ROLLSROYCE","BMW"]) following this strategy we were finally left with only 12 groups which can be considered as a huge improvement. **Finally**,after completing these steps we performed categorical variables encoding in order to facilitate the task during the modeling phase.

4.1.2 Sinister table :

An essential and primordial phase in our project dealing with data preparation is the detection of all types of scenarios leading to a possibility of fraud committed by clients. We went through the table and their columns multiple times to guide us towards ideas about frauds. The four years range between 2016 and 2019 was the one with the most available data , for that reason we decided to gather "police" by their annual number of "sinistres". A column fraud was added with (0=no fraud and 1=fraud) as possible values and or each fraud type we added a new column labelled as the number of that fraud containing the same values. This will help distinguish clients having different types of fraud simultaneously.

- **First Type of fraud :** It's quite illogical for a single insured client to have more than 5 accidents in which he is faulty (per year) , this explains why we searched for accidents where the percentage of responsibility is greater than 50 , and divided the number of accidents by the number of owned cars and classified each result given that was greater to 5 accidents as fraud .
- **Second Type of fraud:** Reviewing all the data we had we couldn't help but notice a great amount of incoherent data in the column describing the percentage of responsibility of each both people taking part of the accidents . If one person is completely faulty that the percentage affected will be 100 for that client and 0 for the other , if

both are responsible that percentage will be divided to 50 50 per client. To resume each line of data where the percentage is greater than 100 was classified as fraud.

- **Third Type of fraud :** How many cars can a single person own ? Scrolling down the number of vehicles per client , we have noticed that many illogical values can be found , huge number such as 200 and 300 cars for a person with a "police type" = individual . For that reason we decided to classify each illogical value (greater than 5) as fraud .
- **Fourth Type of fraud :** In table "police" there are two possible values for the "type police" one of them is individual , a single person
- **Fifth Type of fraud :** Since April 2007, the car insurance rate has been calculated with the application of a Bonus Malus system this information was gathered in a table connecting each bonus malus class to the percentage that will be paid by the clients , based on this table , we were able to detect all the clients with a bonus malus class incoherent with the percentage affected , all eleven bonus malus classes were studied and we classified the incompatible ones as fraud.

The figure 4.2 below is a bar chart that illustrates the number of data we have for each type of fraud.

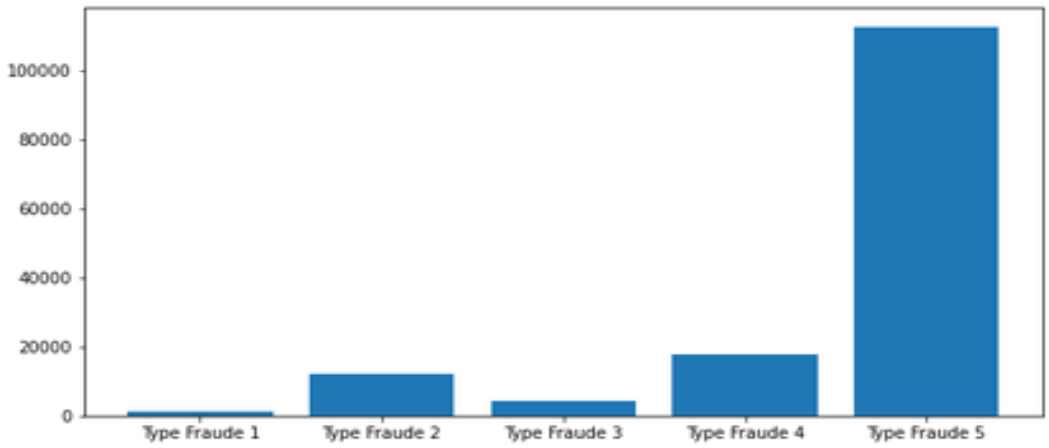

Figure 4.2 – Types of fraud

4.2 External Data :

Web scraping is a term for various methods used to collect information from across the Internet. Generally, this is done with software that simulates human Web surfing to collect specified bits of information from different websites. We were looking for relevant

information that will be used can external data for our model. Since many accidents are caused by poor weather conditions we decided to scrape a weather site having multiple useful data , such as the date , the location, wind speed ... . After that we adjusted the location provided , buy gathering the information we had into states ,doing that we reassembled numerous amount of lost data . Since we had the new affected location in a new column , the old one has become useless , for that reason it was deleted accompanied by all the irrelevant columns such as maximum and minimum temperature. To help get the best results in the modelling phases we decided to merge the table we had with the table sinister by simply adding the number of accidents occurred daily.As a last step we did categorical variables encoding in order to lead the modelling phase . At the end , we left with 11 encoded columns as shown in figure 4.3.

	temperature	vitesseVent	Humidite	Visibilite	Precipitations	lieu_encode	etat_encode	annee	mois	jour	NombreSinistre
0	10.0	8.0	66.0	10.000	-1.0	1	11	2017	1	1	2
1	10.0	10.0	64.0	10.000	0.0	2	4	2017	1	1	7
2	12.0	8.0	56.0	10.000	-1.0	7	11	2017	1	1	2
3	7.0	13.0	63.0	10.000	-1.0	8	16	2017	1	1	1
4	12.0	16.0	59.0	10.000	0.0	11	4	2017	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...
16930	27.0	17.0	64.0	10.000	-1.0	12	16	2019	9	13	1
16931	30.0	13.0	64.0	9.875	1.0	12	16	2019	9	18	3
16932	30.0	16.0	61.0	10.000	0.0	12	16	2019	9	19	2
16933	28.0	18.0	67.0	9.375	2.0	12	16	2019	9	20	1
16934	28.0	25.0	74.0	9.625	4.0	12	16	2019	9	21	1

16935 rows × 11 columns

Figure 4.3 – External Data

Conclusion

Data preparation tools have allowed us to clean data before analyzing it. This forms a solid and reliable foundation to help us is the next phases of our project specially the modelling by enabling our project to perform better and faster and seamlessly meet business objectives.

Chapter 5

Data modelling

> *In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed. So, now we move to the core of the project which is the modeling of our data that will allow us to reach our primary goals: the client classification, the fraud and the bonus-malus detection.*

5.1 Data modeling :

To ensure good modelling, steps are taken:

We begin by partitioning our data frame to two types of variables (explanatory and explaining). Moreover, we cannot escape this crucial step. We have exchanged a long search about the classification algorithms, especially the **decision trees**. All we were really looking for were powerful algorithms. The results were variable and it was up to us to decide, from where it was the choice of **DecisionTreeClassifier** ,**XGBClassifier** ,**RandomForestClassifier** , **LGBMClassifier** and **CatBoostClassifier** which will be explained in detail below.

Decision trees are non-parametric learning methods used for classification and regression problems. The goal is to create a model that predicts the values of the target variable, based on a set of decision rule sequences derived from the learning data. The tree approximates the target by a succession of if-then-else rules. This paradigm applies to both categorical and numerical data. The more complex the tree generated, the better the model "explains" the learning data but the higher the risk of over-fitting.

DecisionTreeClassifier is able to handle classification problems with multiple classes (for example, with labels 0, 1, ... K-1). The objective is to classify each instance into one of the three categories. One of the classes is linearly separable from the other two, but the other two are not separable one from the other.

XGBClassifier The XGBoost stands for eXtreme Gradient Boosting, which is a boosting algorithm based on gradient boosted decision trees algorithm.XGBClassifier is a scikit-learn API compatible class for classification.

RandomForestClassifier their core unit is the decision tree. The decision tree is a hierarchical structure that is built using the features (or the independent variables) of a data set. Each node of the decision tree is split according to a measure associated with a subset of the features. The random forest is a collection of decision trees that are associated with a set of bootstrap samples that are generated from the original data set. The nodes are split based on the entropy (or Gini index) of a selected subset of the features. The subsets that are created from the original data set, using bootstrapping, are of the same size as the original data set.

LGBMClassifier Light GBM grows tree vertically while other algorithm grows trees horizontally meaning that Light GBM grows tree leaf-wise while other algorithm grows level-wise. It will choose the leaf with max delta loss to grow. When growing the same leaf, Leaf-wise algorithm can reduce more loss than a level-wise algorithm.

CatBoostClassifiercomes from two words "Category" and "Boosting". As discussed, the library works well with multiple Categories of data, such as audio, text, image including historical data. "Boost" comes from gradient boosting machine learning algorithm as this

library is based on gradient boosting library. Gradient boosting is a powerful machine learning algorithm that is widely applied to multiple types of business challenges like fraud detection, recommendation items, forecasting and it performs well also. It can also return very good result with relatively less data, unlike DL models that need to learn from a massive amount of data.

So we tackled the modelling step by step: At first, we tried modeling with all 11 classes while waiting for the result of the accuracy which was *low* and *unsatisfactory* as shown in figure 5.1.

Out[246]:

	ModelName	Accuracy
0	DecisionTreeClassifier	0.474316
1	XGBClassifier	0.367917
2	RandomForestClassifier	0.454027
3	LGBMClassifier	0.348036
4	CatBoostClassifier	0.344507

Figure 5.1 – Step 1 outcome

We therefore opted to minimize the modality of the target variable in order to have good results. This means that we end up with only 3 classes: from 1 to 4 is **class 0**, from 5 to 8 is **class 1** and from 9 to 11 is **class 2**. The result 5.2 is shown opposite :

Out[350]:

	ModelName	Accuracy
0	DecisionTreeClassifier	0.754699
1	XGBClassifier	0.746353
2	RandomForestClassifier	0.769492
3	LGBMClassifier	0.737735
4	CatBoostClassifier	0.745335

Figure 5.2 – Step 2 outcome

The improvement exists, but, still far from being a good model. This pushed us to balance our data using undersampling. The goal is to make them more balanced in order to have better results. Except that we notice in this visualization 5.3 that the accuracy has decreased.

This is due to a large data loss . The visualizations before **??** and after 5.5 are clear here.

Out[365]:

	ModelName	Accuracy
0	DecisionTreeClassifier	0.516046
1	XGBClassifier	0.569961
2	RandomForestClassifier	0.554557
3	LGBMClassifier	0.582798
4	CatBoostClassifier	0.578947

Figure 5.3 – Step 3 outcome

```
Out[352]: 0    34807
          1    13017
          2     1298
          ..    ...
```

Figure 5.4 – Before

```
Out[355]: 1    1298
          2    1298
          0    1298
```

Figure 5.5 – After

After trying the undersampling which made us lose a lot of data, we decided to use the oversampling to be able to keep our data because thanks to them we could do a good job. Afterwards, things were concertized and the result is the following **??** where we notice the evolution of the accuracy. But, it is still not satisfactory. We continued the procedure, using the oversampling on classes 1 and 2 only. The choice of two classes was made because we are not obliged to have such classes in terms of individuals, but, it is enough that they are not distant. All this, not to lose any more data and not to exaggerate the oversampling.

Out[365]:

	ModelName	Accuracy
0	DecisionTreeClassifier	0.516046
1	XGBClassifier	0.569961
2	RandomForestClassifier	0.554557
3	LGBMClassifier	0.582798
4	CatBoostClassifier	0.578947

Figure 5.6 – After oversampling (3 classes)

The next step is the test on the two classes only where we notice that the result improved 5.7 . After, we tried what we call hyperparameter tuning, its role is to give the best parameters to ensure that the final result is the best of all. Finally, the best algorithm used is the RandomForestClassifier 5.8.

Out[450]:

	ModelName	Accuracy
0	DecisionTreeClassifier	0.877013
1	XGBClassifier	0.715836
2	RandomForestClassifier	0.888568
3	LGBMClassifier	0.650978
4	CatBoostClassifier	0.714849

Figure 5.7 – After oversampling (2 classes)

Out[471]:

	ModelName	Accuracy	Execution Time
0	DecisionTreeClassifier	0.814937	0.122716
1	XGBClassifier	0.664118	28.899746
2	RandomForestClassifier	0.881240	28.955833
3	LGBMClassifier	0.637344	10.802594
4	CatBoostClassifier	0.684693	48.579087

Figure 5.8 – Final result

Without forgetting the external data, in order to try to estimate the number of claims, we tried regression algorithms such as DecisionTreeRegressor, RandomForestRegressor, XGBRegressor and LGBMRegressor. We therefore found that LGBMRegressor is the best algorithm used with $R^2 = 89 : 5.8$ 5.8.

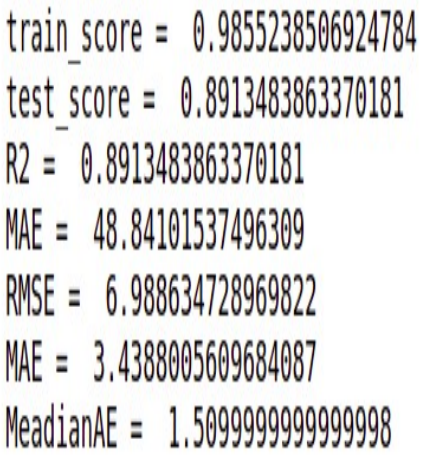

Figure 5.9 – RandomForestRegressor

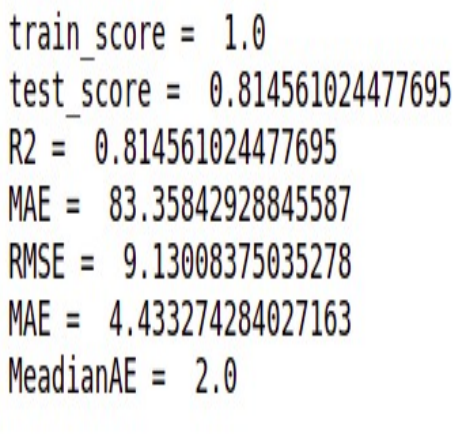

Figure 5.10 – DecisionTreeRegressor

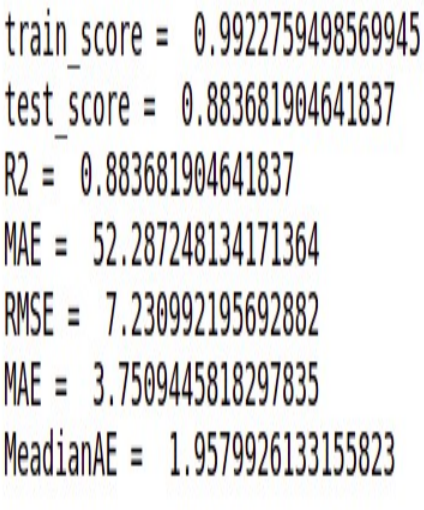

Figure 5.11 – XGBRegressor

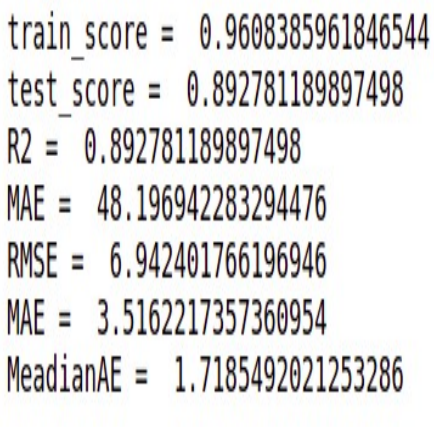

Figure 5.12 – LGBMRegressor

Conclusion

This phase was a mandatory passage to ensure a good model that will allow to predict at the end our needs for the project of CGA (General Insurance Committee). We now move on to the next steps to finish with a relevant and efficient solution.

Chapter 6

Evaluation

> *Whereas the main Model task of the Modeling phase focuses on technical model assessment, the Evaluation phase looks more broadly at which model best meets the business and what to do next. This phase has three tasks:*
>
> ***First*** *,evaluate results : Do the models meet the business success criteria? Which one(s) should we approve for the business?* ***Second*** *,review process: Review the work accomplished. Was anything overlooked? Were all steps properly executed? Summarize findings and correct anything if needed.* ***Third****,determine next steps: Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects.*

6.1 Evaluating results :

At this stage, we'll assess the value of our models for meeting the business goals that started the data-mining process. We'll look for any reasons why the model would not be satisfactory for business use. In our case , after trying different models and different methods we came to the conclusion that the Random Forest classifier is the most performing , it term of accuracy and execution time comparing to (The decision tree classifier , XGB classifier , LGBM classifier , CatBoost Classifier) as shown if the 6.1 below. Since we promised our client that the model we'll be fast and performing we were satisfied with the results we got from the Random Forest Classifier because it meets the business success criteria.

	ModelName	Accuracy	Execution Time
0	DecisionTreeClassifier	0.814937	0.122716
1	XGBClassifier	0.664118	28.899746
2	RandomForestClassifier	0.881240	28.955833
3	LGBMClassifier	0.637344	10.802594
4	CatBoostClassifier	0.684693	48.579087

Figure 6.1 – Accuracies and execution time

6.2 Reviewing the Process :

Now that we have explored data and developed models, we took time to review our process. This is an opportunity to spot issues that we might have overlooked and that might draw your attention to flaws in the work that we have accomplished while we still had time to correct the problem before deployment. We also thought about considering ways that might improve our process for future projects.

6.3 Determining the next steps:

The evaluation phase concludes with our recommendations for the next move. The model may be ready to deploy, or we may judge that it would be better to repeat some steps and try to improve it. Your findings may inspire new data-mining projects. After realizing that we came to the conclusion it quite ready to be deployed each and every demand of

our client was achieved with success , we responded to all the objectives that we have mentioned before if not more , adding our touch and ideas to every phase to make sure that the project has our identity.

Conclusion :

This phase was quite constructive , it made us review the whole project , check if it met the criteria we set in the beginning and especially if there are any failures to respond to the clients demands. Now that we are certain that every aspect of our product has been handled we can move to the next phase which is the deployment phase with great confidence .

Chapter 7

Deployment

> *This is the final step in the process. It involves putting the resulting models into production for the end users. Its objective is to put the knowledge obtained by modelling, in an adapted form, and integrate it into the decision-making process.*
>
> *The deployment can thus go, according to the objectives, from the simple generation of a report describing the knowledge obtained up to the implementation of an application, allowing the use of the obtained model, for the prediction of unknown values of an element of interest.*

7.1 Deployment :

We used the **Django** framework as a tool to ensure the deployment stage. We developed it with html, css and js.

Figure 7.1 – Django logo

We deployed our developed solution and provided interfaces to make things happen.

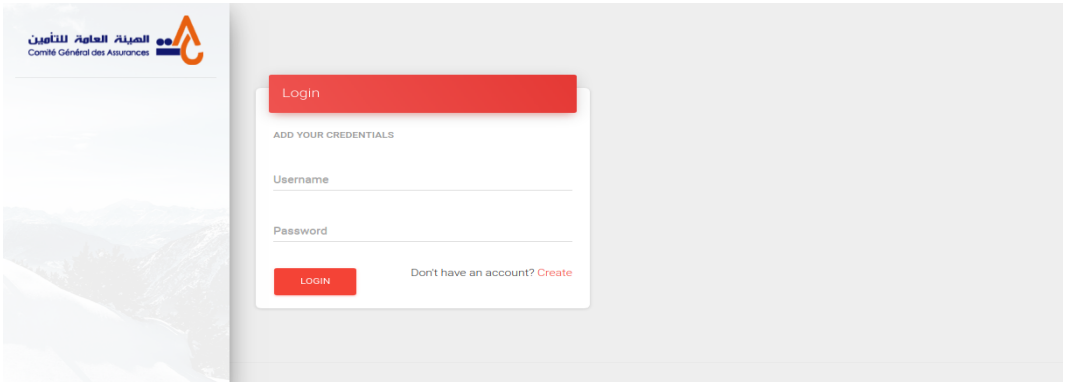

Figure 7.2 – Login interface

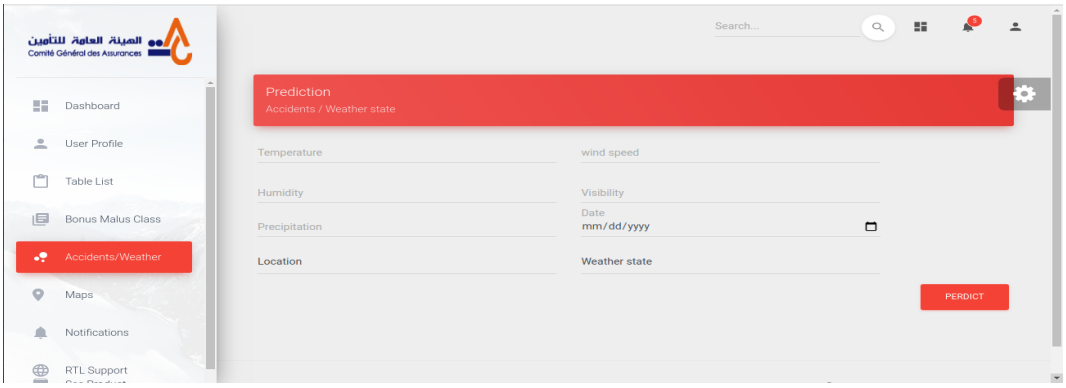

Figure 7.3 – Accidents / Weather interface

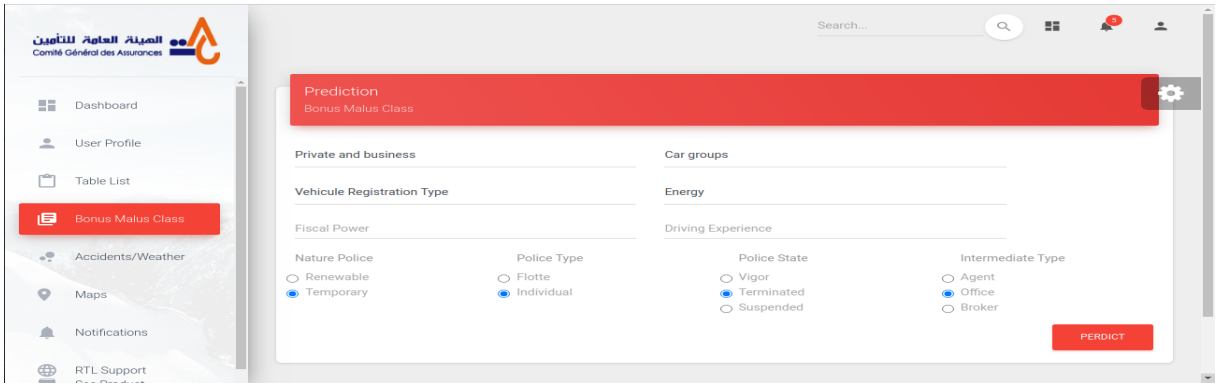

Figure 7.4 – Bonus-malus class interface

7.2 Visualisation :

We carried out a reporting by making precise choices to view the data and the results of our solution.

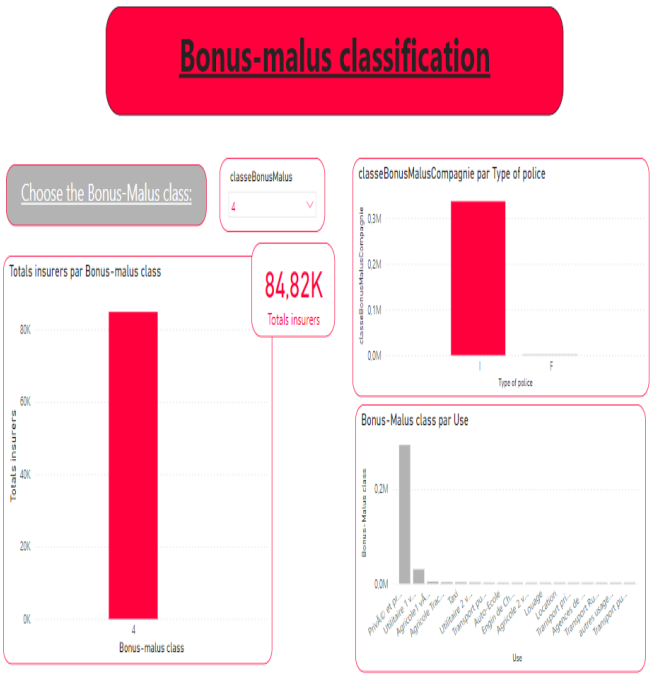

Figure 7.5 – Bonus-Malus classification visualisation

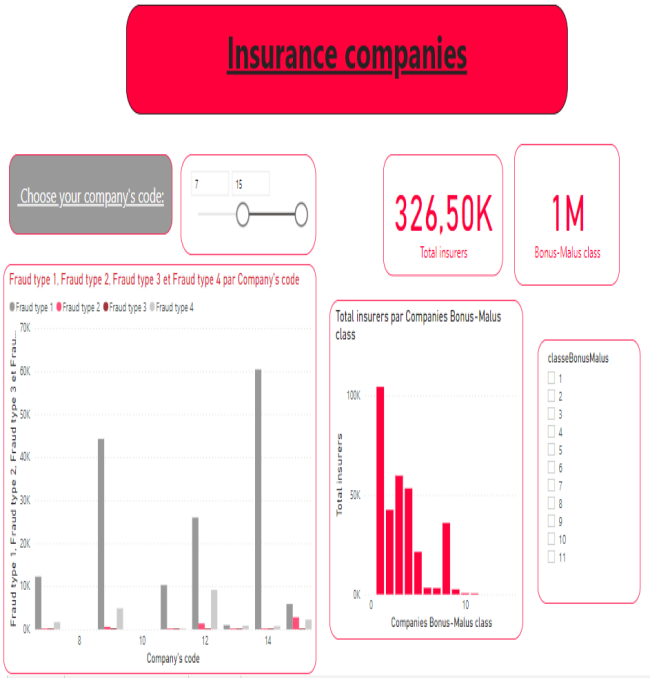

Figure 7.6 – Insurance visualisation

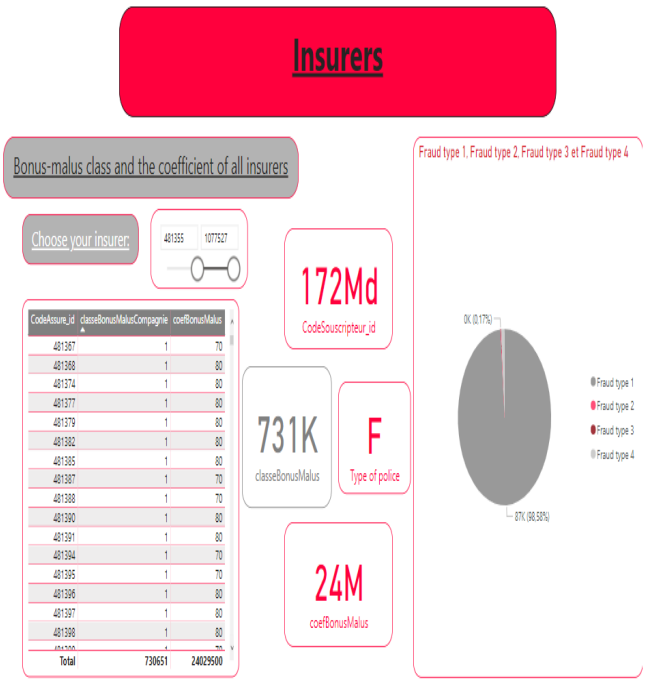

Figure 7.7 – Insurers visualisation

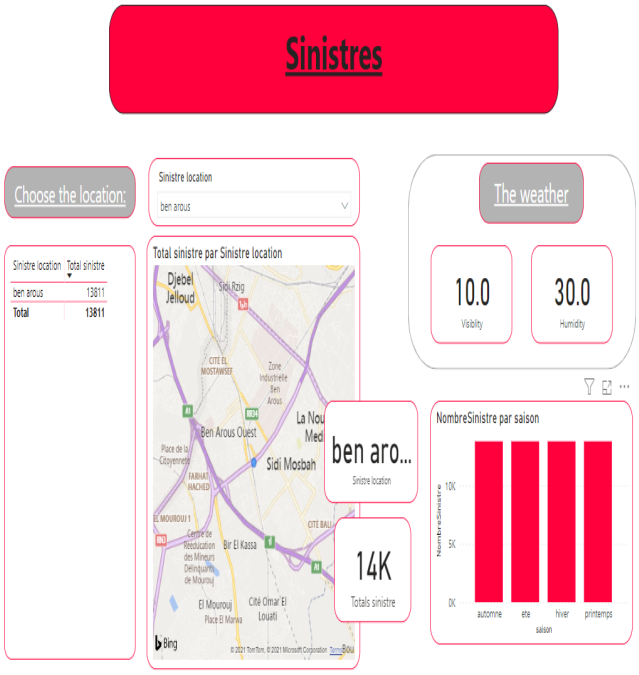

Figure 7.8 – Sinister visualisation

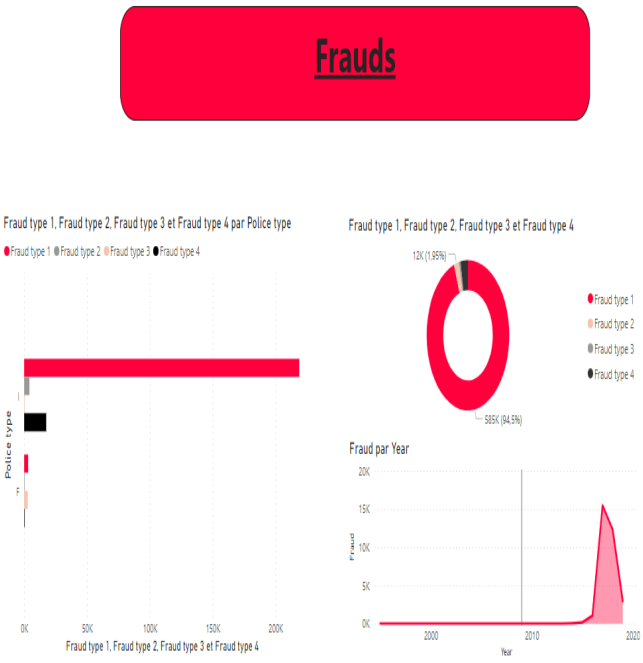

Figure 7.9 – Frauds visualisation

Conclusion

Throughout this previous step, everything has been concretized in interfaces. Just fill in the fields and preach. We were also able to visualize everything with the reporting done. Finally we finalized our project with this last step of our methodology.

General conclusion

This 4 month-long project has been the perfect opportunity for us to have a glance at the data science field and learn new ideas and strategies . It allowed us to understand the professional field from different angles, whether in terms of organization, work rigour or discipline. We realised this project for a company that needed our expertise in the field of data science the . In the process, we found ourselves facing professional learning situations in the field of machine learning. This professional status allows us to apply the knowledge acquired during the ESPRIT training. Group work is something we have become familiar with in recent years, but the coordination and harmony that we have experienced this year have created a pleasant atmosphere that has been the main reason why we have succeeded each phase of this project, we were able to share our knowledge, our ideas, our thoughts and our proposals to improve every little detail of each validation and presentation .

Furthermore, Crisp DM methodology that we applied during this whole semester , is one of the most common and most used methodologies in the machine learning field of study , for that reason now that we have mastered each phase and know exactly what steps to follow to get the best results, any project handed to us will be handled with great ease and expertise.

Moreover , the final product we have has responded to each and every client demand and more , we not only created a performing model that affect the convenient bonus malus class for every new client by classifying him as a good or a bad driver simply by filling a few fields in the form provided in our website , but also predict the number of accidents that will occur in a specific day based on the weather conditions.

Finally, we plan to improve all aspects of the final product we provide in the future, in order to better meet the needs of different customers, making it more intelligent,efficient, performing and up to date.

Bibliography

[1] https://www.datascience-pm.com/crisp-dm-2/

[2] https://www.ftusanet.org/wp-content/uploads/2016/11/FTUSA-RAPPORT-2015-fini.pdf

[3] https://www.ibm.com/docs/fr/spss-modeler/SaaS?topic=dm-crisp-help-overview

[4] https://www.tresor.economie.gouv.fr/Articles/2eb71a8c-735f-483e-8e91-e7e28b0fd449/files/3e1e0f96-90e9-4630-941b-db26fc76c3f7

[5] https://www.webmanagercenter.com/2016/04/27/169337/automobiles-2-millions-de-vehicules-circulent-en-tunisie/

[6] https://www.jurisitetunisie.com/tunisie/codes/assurance/ass1125.html/

[7] http://www.cga.gov.tn/index.php?id=104&L=0

[8] https://www.historique-meteo.net/?fbclid=IwAR1RfVx68zdgVfPMv1lI327i-szMuLwunBnyM0rg8DhWML-oBWwuwjYA