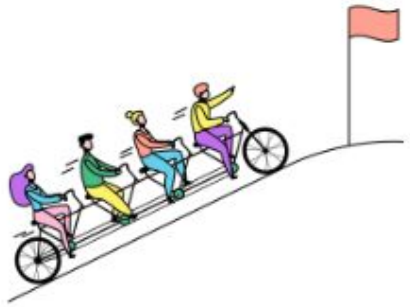
The first data analysis based on cross-data sources

1 What is Ignition Program ?

IGNITION PROGRAM
Leading startups and scaleups companies towards a peaceful and sustainable growth, thanks to 3 services :



RECRUITMENT

- 2000+ candidates/month
- 1500+ startups as clients
- 800+ love stories launched

One coach team dedicated to candidates
One *matchmaker* team dedicated to startups

One team dedicated to Tech-profile recruitment

One team dedicated to international development (mainly Portugal and Spain)

TRAINING

- 300+ managers trained
- NPS : 9,5/10 !!!

We teach management, non-violent communication, coaching, recruitment...

CONSULTING

- Led and facilitated 20+ off-sites
- Helped 10+ Boards towards better communication and collaboration
...

2 Context of the project

ignition program

PROBLEM

- multi-tool : Hubspot CRM, internal Back Office (matching tool), finance and accounting tool...
- dirty data
- never managed to connect our tools and lead cross-data analysis

OBJECTIVES

1. Collect data from multiple sources
2. Cleaning data about deals in our CRM
3. Cleaning data in our internal back office
4. Merging both databases
5. Lead the first cross-data analysis
6. Nice-to-have : building a predictive model to predict future turnover, considering opportunities open and candidates hired in the program

3

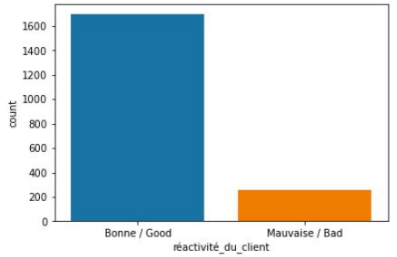
I started with... 😱

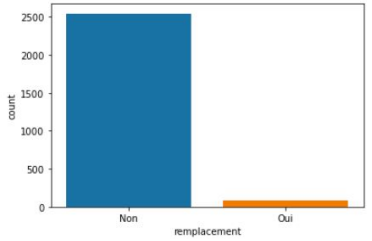
ignition program

86% of missing values in the CRM... 30% in the BO

150 columns to clean... but only 3800 rows

VERY dirty data





High class imbalance

VERY bad distributions, with manyyyyy outliers

well... REAL-LIFE DATASETS 😬

	Deal ID	Techno (L33T)	Pays / Job Country	Numéro de TVA (si entreprise basée à l'etranger)	Closed Won Reason	Tiering	Annual contract value	Last Modified Date	Montant du CA pour le propriétaire du deal / Turnover for the deal owner	ID Répondant Scan	Nom de facturation de l'entreprise (si différent)	Montant du CA pour l'apporteur d'affaires	Restitution orale faite	Salaire fixe final (en K€) / Final fixed salary
0	9008554681	NaN	NaN	NaN	NaN	NaN	NaN	2022-05-25 18:22	NaN	NaN	NaN	NaN	NaN	NaN
1	9007207702	NaN	NaN	NaN	NaN	NaN	NaN	2022-05-25 17:34	NaN	NaN	NaN	NaN	NaN	NaN
2	9007083660	NaN	NaN	NaN	NaN	Tiède	NaN	2022-05-25 17:18	NaN	NaN	NaN	NaN	NaN	NaN
3	9005498498	NaN	NaN	NaN	NaN	Tiède	NaN	2022-05-25 15:15	NaN	NaN	NaN	NaN	NaN	NaN

4 METHODOLOGY

ignition program

DATA CLEANING

a- Columns and format cleaning

b- Dropping columns and rows (but a few because I didn't have much at the beginning...)

c- Cleaning numericals with KNN Imputer

d- Grouping categorical values to reduce class imbalance

e- Encoding and cleaning categoricals with KNN Imputer

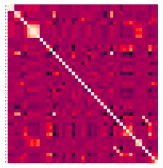
f- Exporting 2 final dataframes for each datasets : one for analysis purpose (with non-encoded categorical values) and one for model-building purpose (with encoded values)

DATA EXPLORATION

a- Merging both dataframes into one : OUR FIRST CROSS-DATABASE 🙌

b- Choose not to remove outliers (because not enough rows)

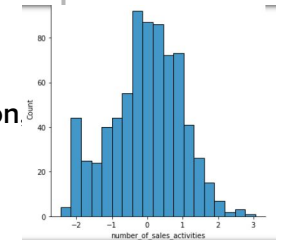
c- Mainly exponential distribution for continuous data and high multi-class imbalance for discrete



d- Multicollinearity

e- Scaled data with PowerTransformer to be as closest as possible to a normal distribution

f- AMOUNT (y, value to predict) already followed a normal distribution, no need for log transfo



PREDICTIVE MODEL

a- Objective : predicting AMOUNT c

b- Created a function to try different models : linear, knn-regressor and mlp-regressor

c- Bad results : R2 of 0.23 at best with knn method and 6 neighbors...

DATA VISUALISATION

a- Using Tableau for better outputs transmission :)

b- Answering the following questions (raised by internal teams)

➔ who is our typical client (BU Recruitment) ?
➔ who are our ambassadors ? Time wasters ?
➔ what is the seasonality of our sales ?
➔ why our interview/hiring ratio is decreasing ?
➔ how to optimize our chances to close a job (and win a deal) ?
➔ how are our sales team performing in terms of matchmaking (not sales) ?

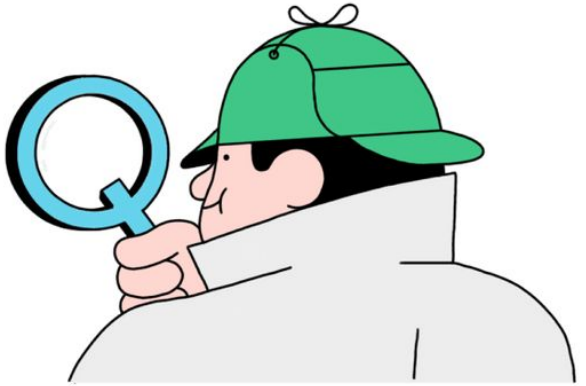
c- >> THE ANALYSIS IS HERE <<

5 Results - Predictive model

BAD PERFORMANCE OF OUR MODEL... 😭
Best R2 computed = 0.23

It could be caused by :

- insufficient size of datasets (ending with about 3000 rows)
- too many missing values in proportion, replaced by KNN : created bias
- too many outliers that we can't removed (due to insufficient number of rows)
- amount not correlated enough with other variables and too many multicollinearity
- high class imbalance between variables
- only had time to merge data on jobs from our two tools : to complete my model I will also need data from our candidates pipeline !

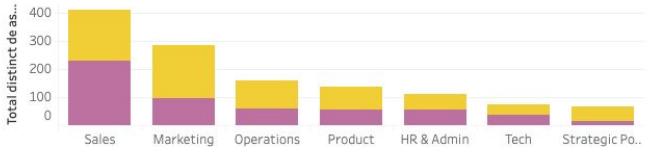


6 Results - Data Analysis : some outputs...

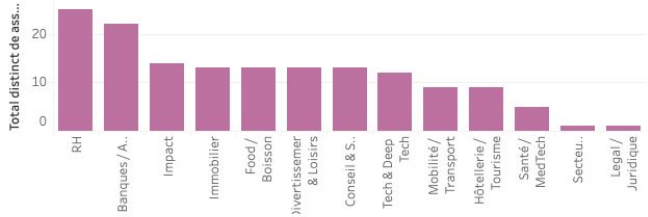
Typical client

A small-sized company in a high-paced and highly competitive environment, operating (mainly) in the service sector, and aiming at professionalizing itself.

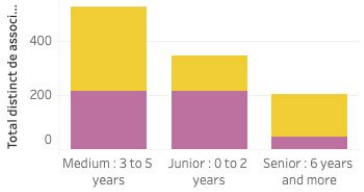
Our typical customer is looking for sales for growth purpose *(not very surprising...)*



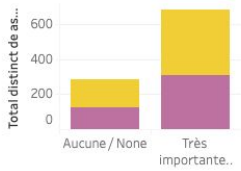
Our typical customers is in the HR & Education market. Impact companies enters the Top 3, mirroring candidates motivation !



Our typical customer looks for medium and senior-level for competencies purpose and for juniors for growth purpose



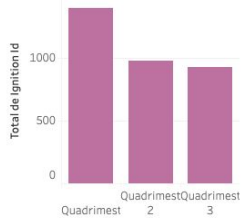
Our typical customer launches competitive tender on the job, but Ignition has a good capacity to answer



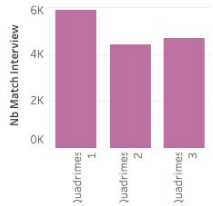
Seasonality

For opening jobs and integrating candidates into our program, high season is the first 4 months of the year.
To close jobs and earn $$, high season is from may to july.
The last 4 months of the year are far less profitable for our recruitment business !

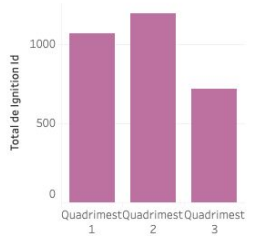
Jobs opening per quadrimestre since 2020



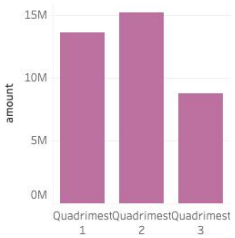
Interviews per quadrimestre since 2020



Jobs closing per quadrimestre since 2020



Turnover per quadrimestre since 2020



6 Results - Data Analysis : some outputs...

ignition program

THE FULL ANALYSIS RIGHT HERE

Next steps

ignition program

FOR THE FINAL PROJECT, I WOULD LIKE TO CONTINUE THIS WORK WITH TWO POSSIBLE OPTIONS :

SALES PREDICTIONS

Improving the current model by :

- ★ improving the cleaning of the current model : no KNN for every columns to avoid too much bias
- ★ adding candidates data to the current database (the success factor of a recruitment always depends on our sales variables, but also on our candidates pipeline
- ★ using a better performing model (that we will see in the following of the class 😉...)

SELECTION PREDICTIONS

Improving the selection process of our candidates :

- ★ CURRENT SITUATION : done manually today, based on selection grid and criteria, one full-time employee dedicated to it
- ★ Objective : building a model analysing previous applications (lost vs won vs admission criteria) to compute the chance of a new application to be admitted in the program + combining this output with our jobs pipeline (do we have opportunities for this candidate ? How much chance to recruit him after integrating him into the program ?)
- ★ Output : a probability for being accepted in the program, and then hired by a client, for each application ! 💪

ignition program

THANKS FOR LISTENING TO ME

😊