



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

EPFL

MGT-415

DATA SCIENCE IN PRACTICE

Optimizing cold calls for car insurance

Authors

Hugo Meyer

Jan Benzing

Laure Bruyère

Timothée Bornet dit Vorgeat

Veronica Pagano

Professor

Christopher Bruffaerts

April 5, 2018

Contents

1	Introduction	2
2	Business understanding	3
2.1	Opportunity identification	3
2.2	Strategy	5
2.3	Stakeholders	5
3	From business to analytics	6
3.1	Features of the model	7
3.2	From data-bank to dataset	8
3.3	Campaign timeline	8
4	Data fetching and profiling	9
5	Model analysis and interpretations	12
5.1	Data Preprocessing	12
5.1.1	Normalization	12
5.1.2	Missing values	12
5.1.3	Unused features	14
5.1.4	Outliers	14
5.1.5	Miscellaneous	17
5.2	Prediction model	17
5.2.1	Score metric	18
5.2.2	Model choices and performances	18
5.2.3	Back-wise feature selection	21
6	From analytics to business	23
6.1	Analytical results of the model	23
6.2	Benefits for the company	24
7	Deployment: how can our analysis be used on an ongoing basis?	26
7.1	Implementation in the company	26
7.2	Improving the model	27
7.3	Reusing the model	27
8	Conclusion	27
9	References	28
10	Appendix	28

1 Introduction

The goal of this study is to apply analytical data science tools on a car insurance dataset provided by an American bank. A car insurance principally provides financial protection against vehicle damages and physical injuries due to car accident, proportionally to the client responsibility. Additionally to insurance companies, it has been commonly used that banks provide those services as well.

One of the purposes of those companies is to have a maximum of their responsible clients subscribing to a vehicle insurance. Therefore, companies organize regular campaigns in order to attract new clients. This is called "customer development" as they are already registered at the bank. As employees initiate contact with their clients, they undertake outbound marketing.

During the campaign, the main procedure for recruiting new clients is by phone. In marketing, this is named "cold calling". It consists of soliciting potential customers who have not yet expressed any interest in the products or services that are being offered. Note that the bank retains relative information of potential new clients (that are already banking clients) for insurance. Bank employees can then call them for advertising available car insurance options. In the case of a customer already showing interest and the bank offering their service, we call this the "warm calls" in marketing.

Our main goal is to optimize the cold call statistics by finding, through the previous campaign's dataset, a profile that is most likely to subscribe to the bank's car insurance.

The goal is to target customers by establishing their probabilities of subscribing before the campaign. The clients with the highest probabilities would then be called preferably than others.

Hence, the main challenges of this project are :

- Understand and translate the scope of the business into an analytical one
- Assess a method in order to leverage the business value
- Based on data visualization and machine learning, formulate a model to determine customers' subscribing likelihood
- Translate the model into generated business benefits (added value of the project)
- Present a critical thinking with possible improvements

2 Business understanding

2.1 Opportunity identification

Today, the bank hiring us performs "call cold" campaigns every six months, to attract both old bank customers who did not subscribe yet to the insurance and the new bank customers who arrived between two campaigns. As the bank does not understand how to target their calls during a campaign and chooses them randomly, we saw that there was an opportunity to orient the calls towards specific clients. Today, the bank has an average success rate of 2%. So the bank calls a subsample of the data-bank (population) and by choosing randomly among this sample it will also have 2% of success as the distribution stays unchanged. Our goal is to take the same sample among a given population but having a much higher success rate. For our analysis, we chose to work for such a bank in the United States.

We assume that the bank organizes two campaigns per year in order to promote their car insurance; each one running for two weeks. During its campaign, the bank hires about 10 employees. The average call elapse time is deducted from the chosen dataset seen in section 3 and is about 220s for a successful call and 250s otherwise. Hence:

$$T_{avr_call} = 0.02 \cdot T_{succ_call} + 0.98 \cdot T_{fail_call} \quad (1)$$

$$= 0.02 \cdot 350 + 0.98 \cdot 220 \quad (2)$$

$$= 223s \quad (3)$$

Given that an employee works approximately 6 hours per day, we can infer the number of clients called:

$$Calls_{per_employ_per_day} = 6 \cdot 3600 / 223 \simeq 100 \quad (4)$$

$$Calls_{total} = Calls_{per_employ_per_day} \cdot NB_{days} \cdot NB_{employ} = 10'000 \quad (5)$$

Let's assume that the data-bank is composed of 100'000 clients, which means that 10% of the data-bank (see eq. 5) is called during each campaign.

Our goal is to evaluate the profitability of a campaign in the current situation by computing its expected value. This indicator is an alternative measure of the ROI and it takes into account the benefits of achieving to make a customer's subscription and the costs of failing. The expected value is expressed as follow (its generic formula is seen in section 6):

$$\begin{aligned}
 \text{Expected value} &= p(\text{subscribing}) \cdot \text{Benef}_{\text{subscribing}} + p(\text{not_suscribing}) \cdot \text{Costs}_{\text{not_suscribing}} \\
 &= p(\text{subscribing}) \cdot (PV_{\text{new_client}} + \text{Cost}_{\text{succ_call}}) + p(\text{not_suscribing}) \cdot \\
 &\quad \text{Cost}_{\text{fail_call}}
 \end{aligned}$$

With:

- $p(\text{subscribing})$: proportion of called customers that subscribe during one campaign is 2% (see above).
- $p(\text{not_suscribing})$: proportion of called customers that do not subscribe during one campaign is 98% (see above).
- $PV_{\text{new_client}}$: present value of all the annual benefits generated by one customer over an average period of twelve years [1].

$$PV_{\text{subscribing}} = \frac{500^1}{r_f} \left(1 - \frac{1}{(1 + r_f)^{12^2}}\right) = \frac{500}{0.05} \left(1 - \frac{1}{(1 + 0.05)^{12}}\right) \approx \mathbf{4'432\$}$$

- $\text{Costs}_{\text{succ_call}}$: The cost of one employee over one call period (350s) added to the communication cost, when the client subscribes. By approximating an employee salary of \$2'000/month and a communication cost of \$0.05/min, we have:

$$\text{Cost of call} = \text{Duration of call [s]} \cdot (\text{Cost of call [$/s]} + \text{Salary [$/s]})$$

$$\text{Cost of successful call} = 350 \cdot \left(\frac{0.05}{60} + \frac{1000}{80 \cdot 3600}\right) = \mathbf{1.51\$}$$

- $\text{Costs}_{\text{fail_call}}$: The cost of one employee over one call period (220s) added to the communication cost, when the client does not subscribe.

$$\text{Cost of unsuccessful call} = 220 \cdot \left(\frac{0.05}{60} + \frac{1000}{80 \cdot 3600}\right) = \mathbf{0.95\$}$$

We can now estimate the expected value generated by one campaign in the current situation of the bank (equality seen above):

$$\textit{Expected value} = 0.02 \cdot (4'432 + 1.51) - 0.98 \cdot 0.95 = \mathbf{85.68\$}$$

To find the profits, we just need to re-multiply the expected value by the total number of individuals:

$$\textit{Profits} = \textit{Expected value} \cdot 10'000 = \mathbf{856'800\$}$$

We will demonstrate in the following how can we increase these profits by targeting the individuals with a more competitive model than a random one.

2.2 Strategy

As we possess social information, customer behavior and previous campaign information we can use this historical data in order to score profiles that occur to respond more positively to a car insurance request. Thus, we are going to perform a multi-variable segmentation. This will be done through a rule-based approach and dimensionality reduction techniques. Given that the bank has a fixed budget for this campaign, we will use data science and machine learning in order to select profiles who have a high probability to subscribe. With this analysis, we will be able to bring to the bank and to stakeholders a more efficient way to increase the profit margin. At the same time, it will decrease the odds of disturbing a customer that is not interested in a new car insurance. However, in this analysis we will focus only on maximizing new subscribers. This is called a targeted marketing action. Finally, the strategy is to list the probability of doing a warm-call by descending order of each observation. Hence, if the bank runs our model on their entire database, they will be able to guide their employees on who they have to call first.

2.3 Stakeholders

Several parties are involved in this project and benefit directly or indirectly from it. The party for which the project is the most profitable is obviously the bank itself, by having its profits increased (as demonstrated in section 6) for the same campaign budget. Within this entity many employees will be involved:

¹Approximation of the annual benefit generated by one insurance customer inferred with AXA insurance data. It is expressed as the total benefits of the company divided by its number of customers [4]

²Note that we have discounted the positive cash flow coming from subscribers over 12 years[1]

- The *Cold caller*: Their yield increases and brings them more satisfaction and enthusiasm. This can be felt in their interaction with clients and give even better results.
- The employees in the administration whose job is to prepare the campaign and define the list of clients to call.
- The IT department: it will have to implement the model to replace the old one.
- The shareholders of the bank who will profit from its business performance and from an eventual dividend payout increase

The customers are also directly involved by having to respond to the the phone and eventually subscribe to this insurance. In this case, they would be financially involved by paying an annuity to the bank for the service.

Finally, the consulting company (us), will benefit from this project by being mandated by the bank. It will receive either a fixed amount of money (with a quote) or perceive a percentage of the benefits difference.

3 From business to analytics

The data used to cope with the problematic explained above, are under the form of a dataset containing 4000 observations (rows of the dataset), corresponding to 4000 customer profiles of the bank. Each profile is explained with 18 attributes detailed in section 3.1. Given that our goal is to predict whether a person will subscribe or not, we considered the corresponding feature named *CarInsurance* apart from the others. It will be used as a target for our model and not as a feature, what reduces the number of attributes to 17.

Feature	Description	Type
Age	Age of the client	Numerical
Job	Job of the client	Categorical (Multiclass)
Marital	Marital status of the client	Categorical (Multiclass)
Education	Education level of the client	Categorical (Multiclass)
Default	Has credit in default?	Categorical (Binary)
Balance	Average yearly balance, in USD	Numerical
HHInsurance	Is household insured?	Categorical (Binary)
CarLoan	Has the client a car loan?	Categorical (Binary)
Communication	Contact communication type	Categorical (Multiclass)
LastContactDay	Day of the last contact	Categorical (Multiclass)
LastContactMonth	Month of the last contact	Categorical (Multiclass)
NoOfContacts	Number of contacts performed during this campaign for this client	Numerical
DaysPassed	Number of days that passed by after the client was last contacted by a previous campaign (numeric; -1 means client was not previously contacted)	Numerical
PrevAttempts	Number of contacts performed before this campaign and for this client	Numerical
Outcome	Outcome of the previous marketing campaign	Categorical (Multiclass)
CallStart	Start time of the last call	Date
CallEnd	End time of the last call	Date

Table 1: Features description

3.1 Features of the model

Having a deep understanding of the features meaning is necessary to gain prior intuition about their importance, their interaction and how they could be used within the model.

The type of the features data is also critical, because it can lead to very different approaches during the preprocessing.

The table 1 lists the 17 attributes of the dataset with their description and their type.

Almost all the numerical features do not need much explanation except *DaysPassed* since it is a mix of numerical and categorical data. This feature either informs us about the

number of days passed between two last calling campaign (numerical) or labels with *-1* if the client was not contacted before the last campaign.

3.2 From data-bank to dataset

One important consideration is the distribution of the positive and negative cases, respectively whether a person subscribed or not. In the dataset we use, we no longer have the true population ratio of 98%/2% but a biased one of 60%/40%.

It is preferable to have a balanced dataset for training the model. Indeed, it often leads to a more robust predictions. Nevertheless, we will see in section 6 how to render the true ratio when testing, in order to have representative results.

3.3 Campaign timeline

We have a calling campaign of two weeks every six months. As seen if figure 1 a small number of the features of our dataset include informations about previous campaign (*i-1*) or campaigns as such we first wanted a model containing these ones. This model would have used informations one year old or more in order to predict behaviors for the next campaign (*i+1*). As we will see in section 5.1.2 we dropped these three features in part because we wanted to have a model which used data from the current campaign (*i*). We also thought of dropping all information about previous and current campaigns and only use the private informations the bank has on its clients as well as the target (*CarInsurance*) but we found an accuracy not satisfying enough.

Between every campaign the bank loses and acquires clients. They are automatically targeted by our campaigns in order to get the data needed to process them with our model. For example if a customer of the bank arrives between campaigns (*i-1*) and (*i*) he will be contacted in (*i*), if he subscribe to the insurance then great but if not then his data will be used by the model to determine if he will be ignored or targeted by the campaign (*i+1*).

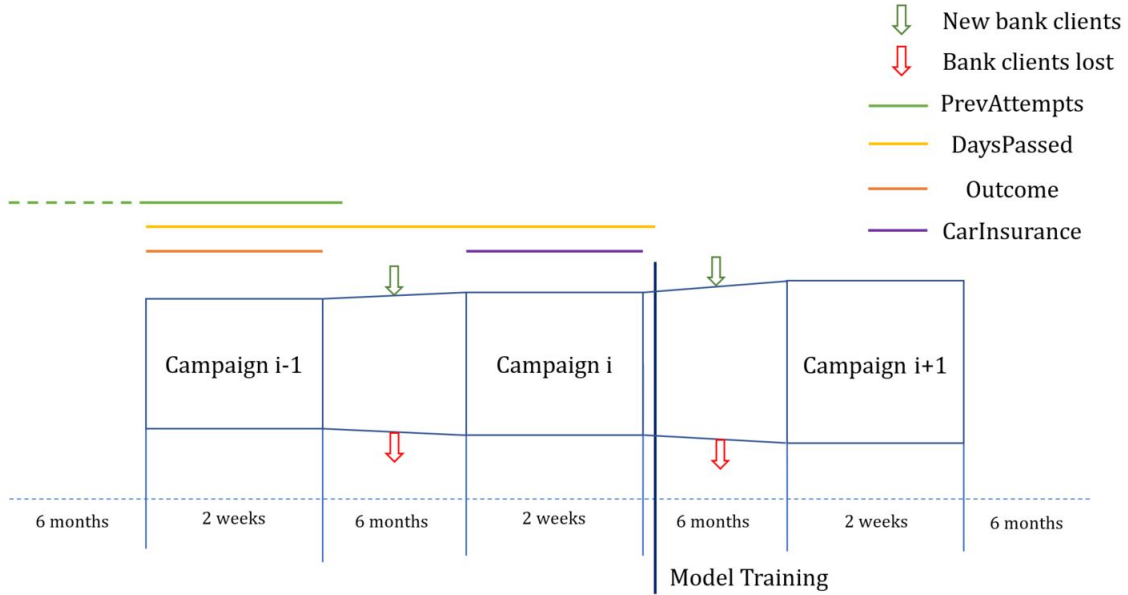


Figure 1: Diagram representing the time value of several features

4 Data fetching and profiling

The goal of this section was to get as familiar as possible with the dataset and with the informations it contains, mostly by visualizing data with *Tableau* software. Having an increased knowledge about the features and observations of the dataset allows us to take a step-back from the programming part (preprocessing and predictive model). It helped us to take some decisions before-hand and during the coding phase, sparing us a lot of time.

In order to monitor the clients and to adapt our model with any possible profile shift of the population, we could use the data to profile customers. To that end, we use different graphs created in *Tableau* containing a maximum of information and where the shifts in characteristics of the clients are easily observable. For this goal, we have to carefully study figure 2, 3 and 6.

First, we observe that the category that in average is most likely to subscribe to a car insurance is *Student* followed by *Retired*. Surprisingly, the *Unemployed* have also a high average subscription rate. One hypothesis could be the fact that these persons have more free time and are more likely to answer the phone. They will tend to have longer

conversations and be easier to persuade.

Besides, looking independently at the average annual balance (*Balance* feature in \$), it indicates that this feature has no tremendous importance on the likelihood of subscribing to a car insurance. However, it could still be useful in a predictive model that would take features interaction into account. Combined with other one it could still lead to an explanation of the target. It is important to mention that in figure 2, we show the average and not the median because our dataset is already cleaned and the potential biased from tail-values (outliers) has been removed.

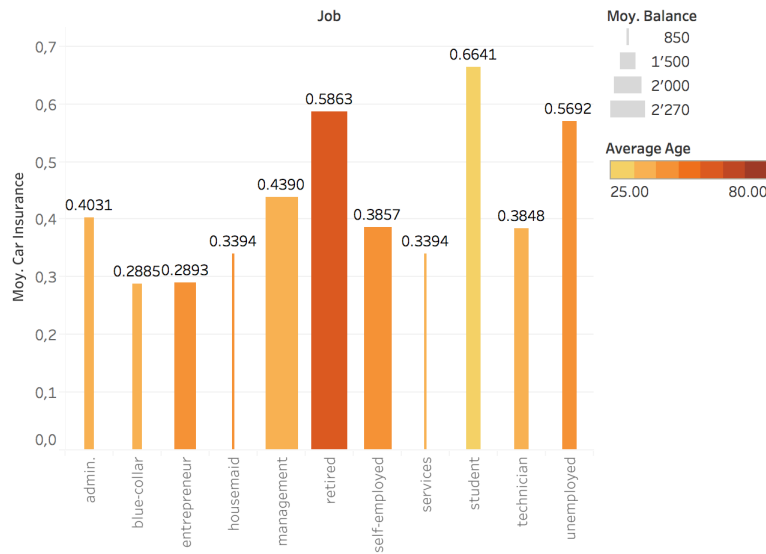


Figure 2: Presenting Average car insurance subscription, Balance and Age versus Job

Additionally, we can observe the influence of marital status by looking at (figure 3). It reveals that a single person is more likely to subscribe than a married one. Another interesting observation is the average subscription for divorced people at retirement which is at 0.7288, whereas the divorced entrepreneurs are at an average of 0.1538. Indeed, single retired people are more easy to reach due to their free time and are surely more willing to talk unlike entrepreneurs. Old people may also be easier to influence. These considerations permit to reveal a meaningful profile for monitoring the outcome of our model.

At a first glance (figure 4), we could think that the last month of contact is a particularly useful feature to optimize our advertisement strategy. Indeed, in average, the months of March, September, October and December have the highest rate of success in calls. But when we zoom in, we noticed that those months were also those with the least data and so the average was not a significant information. As such the last month of

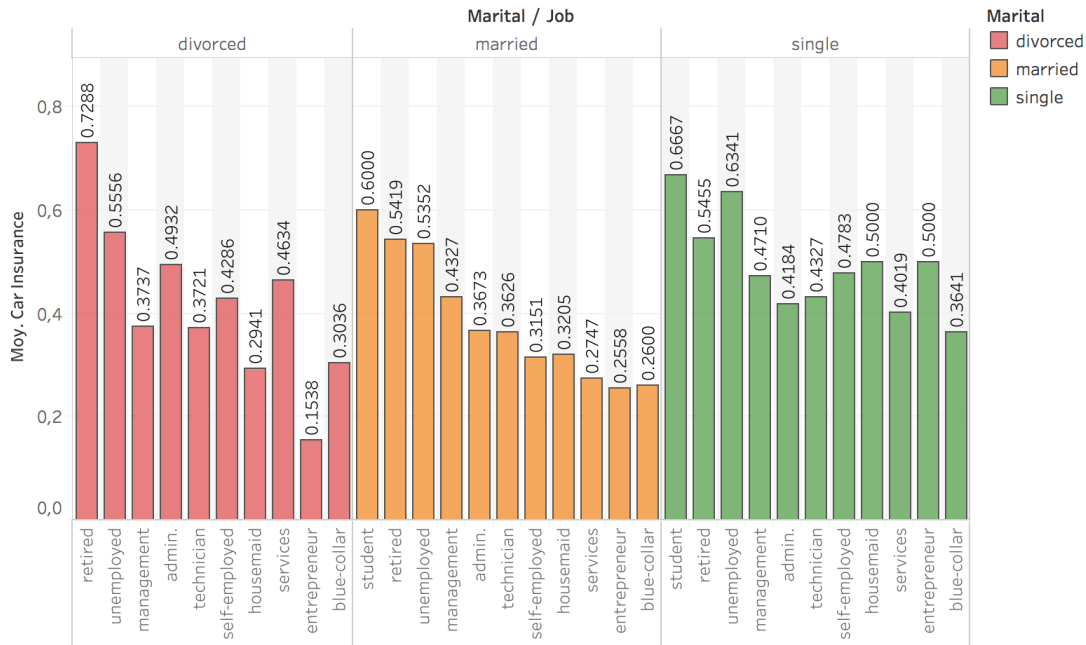


Figure 3: Presenting Average car insurance subscription over job and marital status

contact is not a useful feature but in interactions with the other ones and with the help of our algorithms, we will definitely determine its usefulness.

In order to access the data of the clients during every months of the year the bank has to shift the timing of the campaigns by two weeks each years.

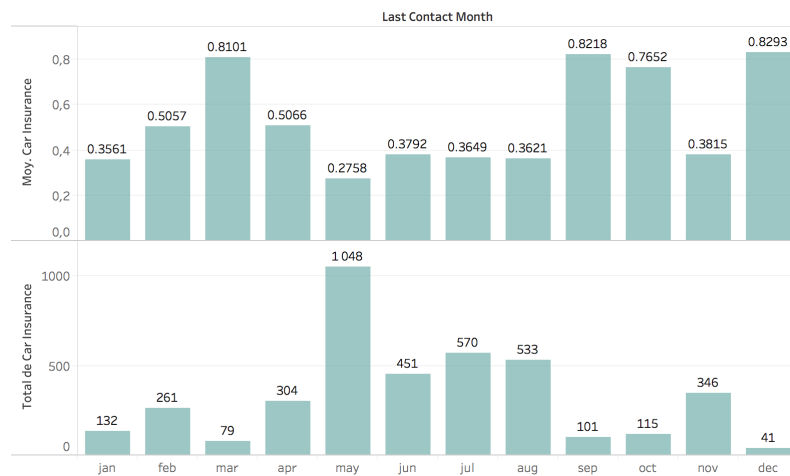


Figure 4: Presenting Average and Total car insurance subscription over Month of the last contact

5 Model analysis and interpretations

5.1 Data Preprocessing

5.1.1 Normalization

In data science the usual algorithms need to normalize the data such that the units of the features do not influence the analysis and their importance in the algorithms are the same.

In our case the normalization is necessary for all the algorithms linked to a metric (distance system) so principal component analysis and Mahalanobis Distance calculation. But it is not the sole reason, indeed for our neural network the normalization is needed to keep the homogeneity of the weights to avoid a runaway of their values.

We also used standardization for our our logistical regression because it was necessary for the use of regularization. The support vector classifier also needed this kind of standardization.

5.1.2 Missing values

Our process of dealing with the missing values depends on the features but first let's enumerate the possibilities that we have :

- Dropping attributes with a majority of missing value. This helps our model because it would be wrong to impute the majority of the values from the minority.
- Dropping observations containing at least a missing value. This results in a loss of information, as such we will not use this method.
- Replacing the missing value with the median or the mean value (if numerical) of the feature or the mode (if categorical). This creates a bias in the feature distribution.[2]
- Replacing the missing values with respect to the proportion of the categories before the replacement. Unlike the mode, inferring missing values by keeping categories proportions does not bring bias.
- Inferring the missing values by using correlation between features if there are ones to be found. Given that in our case, the missing values concerns only categorical features, instead of using correlation, we use *Tableau* to visualize the distribution of the sparse feature against another. The more they are correlated and the more heterogeneous distributions we will have, what helps for inferring. In a second step, we infer by using the mode or the proportion (as explained before). It is then an

indirect way of inferring.

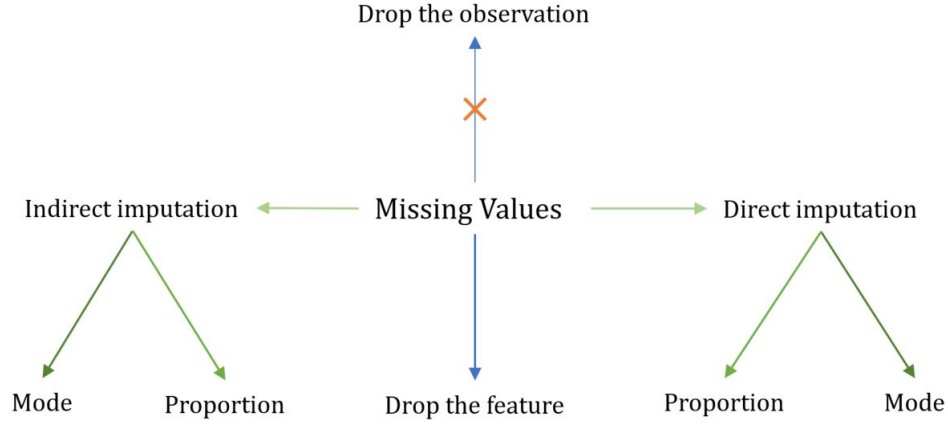


Figure 5: Diagram representing the choices we have when dealing with missing values

In our case we dropped three of our features. The first one is the *Outcome* of the previous campaign since it has more than three fourth of its values missing (NA).

Using the mode of a feature is the easiest solution but has to be used only when the final proportion has not significantly changed. For instance when the amount of missing values is low or when the mode represent the extreme majority of cases.

We used the direct imputation method only for the *Communication* feature. And because the amount of its missing values is low but especially because the proportion of cellular contact makes the extreme majority of the feature (91.4%). When using the mode as replacement value for the missing one we get a new proportion of cellular contact of 93.3%. Confirming the absence of significant effect.

Indirect imputation is only done between features with evident relation. Education and Jobs is the only one of such case. We can see this relation in figure 6. It also sufficient to inform use of which methods of imputing is adapted. For example, the observations which had Management as Job but no Education value, were set as having tertiary education since its mode is so big. With the same idea, secondary education was imputed for the people whose Jobs are Technician and Services. The mode was also used to infer the Jobs of the ones whom we know their Education since this group was really small. With the same justification the level of Education of the ones whose Jobs were *Admin.*, *Entrepreneur*, *Housemaid*, *Self-employed*, *Student* and *Unemployed*. The ones whose Education and Job were unknown were put as having *Secondary Education* (the mode of *Education*) then being *Technician* (the mode in *Secondary Education* category). We used the proportional method on the *Blue-collar* (12 to *Primary* and 19 to *Secondary*) and *Retired* people (10 to *Primary* and 9 to *Secondary*) since the number of missing value was

too high and the mode not dominant enough.


Education	Job												CarInsurance
	admin.	blue-coll..	entrepre..	housema..	manage..	NA	retired	self-emp..	services	student	technician	unemplo..	
NA	14	31	3	4	26	8	19	7	9	25	21	2	
primary	9	281	13	56	22	4	93	6	27	8	16	26	
secondary	374	430	50	33	94	5	100	43	269	67	446	77	
tertiary	62	17	55	16	751	2	37	84	25	31	177	25	

Figure 6: Linking jobs to education

5.1.3 Unused features

By studying and understanding the meaning of the attributes, we realized that some of them were irrelevant for our specific application or impossible to process. We then decided to remove them from our model:

- *DaysPassed*: This feature contains two drawbacks, making it very difficult to handle. On one hand, the feature is numerical but the (-1) indicates a category for the people that have not been contacted yet. Unless we built a new binary feature indicating whether a client was contacted in previous campaigns or not, there was not many options left. We decided to drop the feature from the dataset. Furthermore, as explained in 3.3, this feature complicated the collection of campaign informations for further monitoring or outcomes prediction (using the model). Indeed, it gave informations about the previous campaign, what is six months before the other features. It was also problematic when running the model on new arriving bank clients, given that we had to wait two campaigns before being able to get these data.
- *PrevAttempts*: this feature was also problematic because it informed on the total attempts to reach a client through all the previous campaign. It complicated the updating of the data for the same reasons as the feature *DaysPassed*. As an additional justification we also compared the accuracy of our models with and without this feature and found no significant difference ($< 1\%$).

5.1.4 Outliers

In this section, we are going to analyze and detect outliers in order to remove them. First, an outlier is an "observation which deviates so much from other observations as to arouse suspicion it was generated by a different mechanism", Hawkins (1980).

The trade-off when removing outliers is between removing too little of them thus modeling exceptional data and removing too much resulting in over-fitting.

One approach could be to take all features into account and perform a 2D Principal Component Analysis (PCA). This will help us to get a systematic visual representation of outliers. The results are shown in figure 7.

The PCA projects all used features on the two first eigenvectors, representing the highest variance coming from the original dataset. We can see on the right bottom a point representing an out-lier. His Id. had been found and thus the feature could be identified through *Tableau*. The feature corresponding to this outlier was identified through *Tableau* in figure 9.

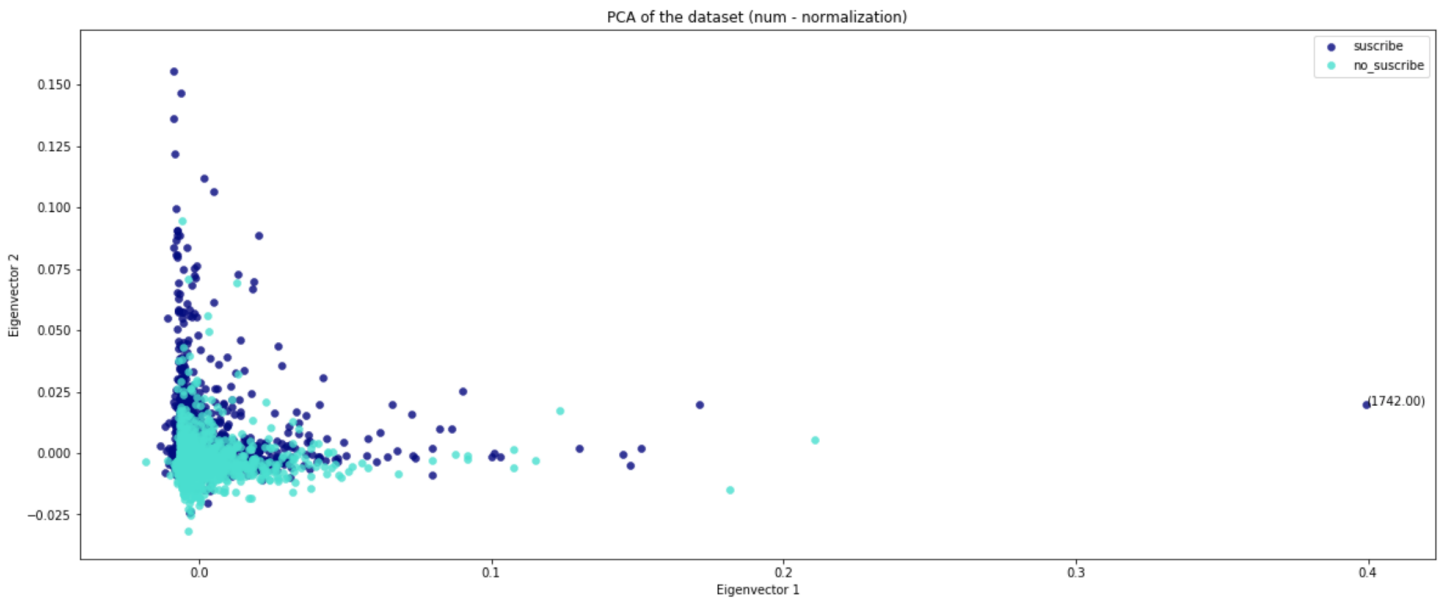


Figure 7: Presenting feature projection on the first two eigenvector

A second approach would be to use the Mahalanobis Distance. This method enabled us to combine every numerical feature. Hence, we could determine outliers that may not be visible by looking at each feature separately. The results are displayed in figure 8.

By observing this plot, we first thought of removing four outliers given by the four bar counting from the right because of their high Mahalanobis Distance. Once we have identified them, we analyzed them with *Tableau* and noticed that only the individual 1742 and 163 are removed as outliers. The other profiles do not have a feature extreme enough to be removed.

Finally, we looked independently at different features in order to detect additional outliers. For example, we found that in certain job categories such as housemaid and technician, we could identify one out-lier (Id:1959) as seen visually in figure 10. With this approach

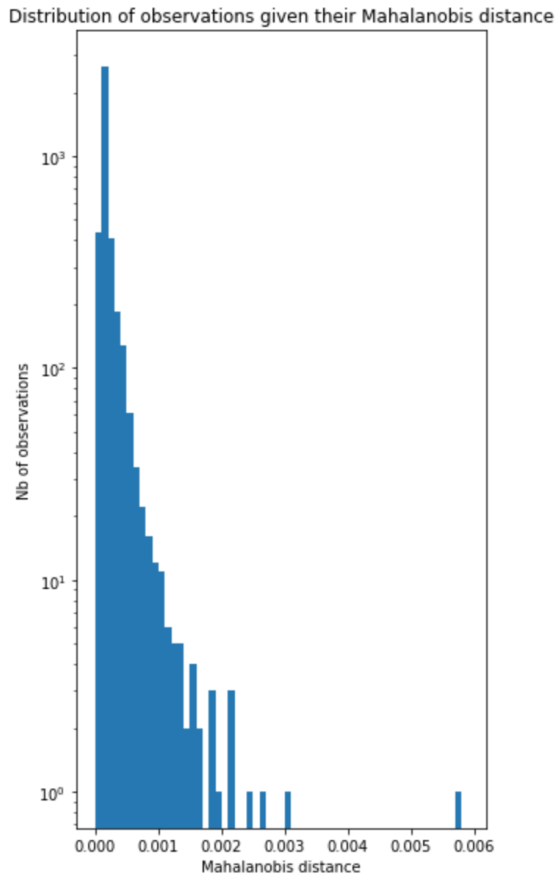


Figure 8: Plot of Mahalanobis distance

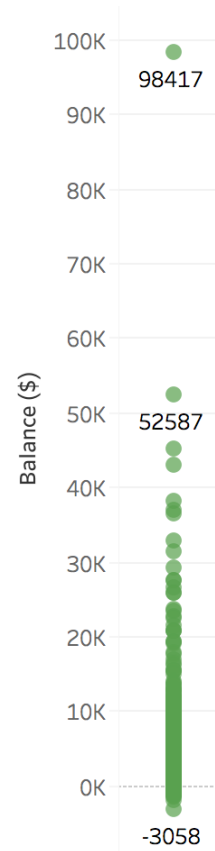


Figure 9: Highlighting out-lier observation in feature: balance

we have to be careful as the decisions are made without any computation. Therefore, we also showed in this same figure a case where an observation was not taking as an out-lier (Id. 25).

Ultimately, we can conclude on the fact that the Mahalanobis Distance is a robust and systematic approach for detecting outliers, especially for the *Balance* feature. This measure allowed us to detect them by taking into account multiple numerical dimensions. However, this approach has to be combined with data visualization such as *Tableau* in order to include human intuition on one or two specific features. Hence, we can verify the outliers selected by the Mahalanobis Distance and gain further information about outliers in categories that could not be exploited as a distance.

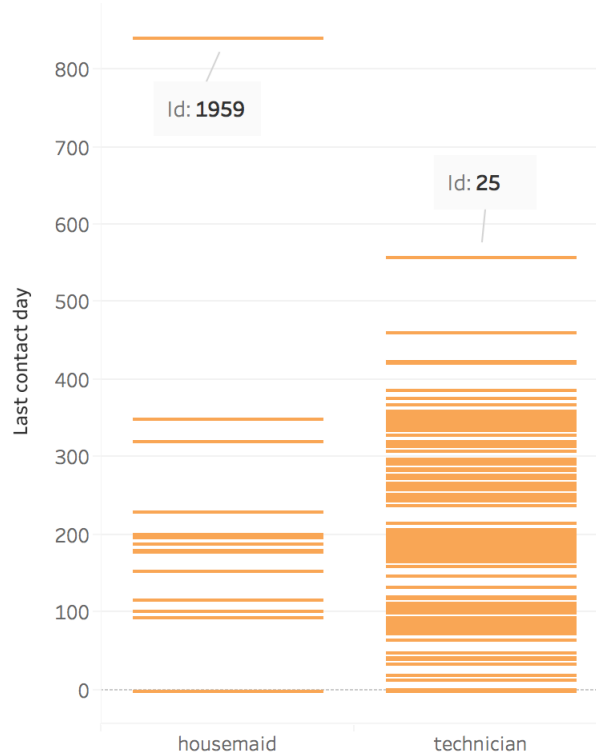


Figure 10: Presenting last contact day for two jobs and showing extreme values

5.1.5 Miscellaneous

Since the date cannot be directly used by our algorithms, we created another feature, the time difference (in second) between the *CallEnd* and the *CallStart*, and called it *TimeElapsed*. We dropped *CallEnd* but kept the hour part of *CallStart*. As we did not know if we should use this feature as numerical or categorical we tried our algorithms with both and found no significant difference so we kept it as numerical. Effectively we transformed two dates into two different numerical features.

We had a ratio of 40% unlabeled data so in order to optimize the performance of our classifiers we oversampled our dataset to a 50% ratio.

We used the one-hot encoding for all categorical data.

5.2 Prediction model

In this project we tested different prediction models in order to find the optimal one for our specific problem. We will discuss their performances, their pros and cons and the

scoring method used to assess them.

Here is a table summing up all these results:

Model	LR	LR (Ridge)	Decis. Tree	Rand. Forest	K-NN	SVC	NN
Accuracy	80.46%	79.31%	81.82%	86.62%	73.08%	84.0%	77.5%
Recall	79.71%	74.95%	86.75%	94.0%	77.02%	85.05%	-

Table 2: Score table of the different predictive models tested

5.2.1 Score metric

As we balanced the dataset with the target, the score resulting from the model is not biased.

Given that we reach 10% of the data-bank and only 2% of the clients subscribe (True Positives), we have a great margin to catch all the subscribers. We can then afford to call clients that will not subscribe at the end (False Positives). This case is not penalizing in terms of budget, the costs being far inferior to the reward of the true positive: 4330.5\$ » 0.95\$. On the other hand, the clients that would have subscribed but were not called are far more problematic (False Negatives). They represent a fictive gain that was not obtained of 4330\$. Hence, we want the less possible False Negatives. The score metric adapted to optimize the number of False Negative is the recall expressed as follows:

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

Our goal is then to maximize the recall to reduce as much as possible the number of FN. We also keep an assessment of the accuracy which is a general metric of the model and can be useful to compare the model with others. Its definition is:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (7)$$

For all the given results, the dataset was separated into a train set and a test set in order to avoid linkage.

5.2.2 Model choices and performances

Logistic Regression:

As a first and quick step, we used logistic regression, which is quick to set up. There are no parameters to tune and presents a fast execution time.

Even if this model does not present the best performances (recall of 79.71%), it is very convenient to have a quick performance feedback when a change is done upstream (e.g. in preprocessing).

Logistic Regression with ridge regularization:

In addition to the logistic regression, we tried to use regularization, a technique which is often used to simplify the complexity of the model. It allows by having a penalty term on the optimization function, to avoid any over-fitting on the train data.

The penalty term of the regularization was found by making a grid search and optimizing the recall. Unfortunately, it performed worse than logistic regression, maybe because of a too simplified model chosen by the regularization. The value of zero was not introduced in the grid search possibilities so it could not perform a simple logistic regression.

Decision Tree:

The decision tree method was tested because its model differs a lot from others by using decision rules with its attributes. It makes the separation between data with perpendicular boundaries and can be very efficient given the geometry of the problem. In the case of a binary classification, it can show very good performances. Indeed, we get a recall of 86.75%.

Once again, we used grid search and cross validation to find the optimal depth of the trees (max recall) without over-fitting.

Random Forest:

This ensemble method uses the decision tree method to generate a lot of trees (a forest). The depth used is the optimal one determined before by running the decision tree algorithm.

A grid search is also performed to find the max number of features to use, given that only a subset of features is used to find the trees.

This method being an enhancement of the decision tree method by decreasing the variance of the model, we note that the results are better. Moreover, this model gives the best results with a recall of 94% and the accuracy of 86.62%.

Given that we don't need any visualization of the rules used, we prefer having better results than the decision tree method instead of having a possibility a visualization.

This model is then the one that we retain and will implement in the bank to target the customers.

The model still allows by its separation metric to get the probabilities of clients subscrib-

ing, which will be used to find the optimal subset to call.

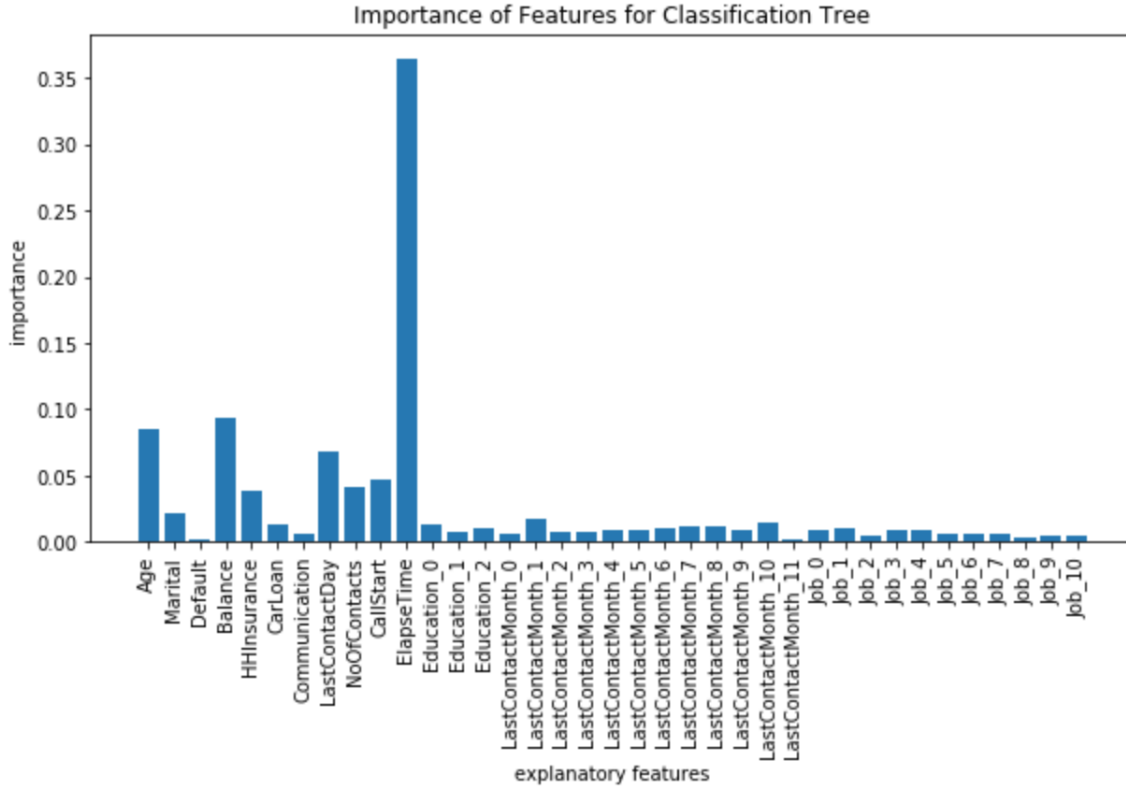


Figure 11: Importance of the feature provided by the random forest algorithm

Random forest allows also to assess the importance of the features by taking their interaction into account. We can see it on figure 11 the importance of each feature of the dataset. This is done by considering the number of decision rules used with each feature. We can see a high predominance of the feature *ElapsedTime*, which appears to explain the target a lot. Indeed a long call means that the client is interested by the offer and will be more likely than others to subscribe in the following campaigns (if he did not subscribe in current one).

We see also that the features *Default* and *Communication* are not very explicative. Intuitively, calling somebody on its cellular or its telephone should not have a significant effect on its likelihood to subscribe. Nevertheless, this plot cannot describe the importance of the one hot encoded features, which are divided in many ones.

K-NN:

The question of implementing k-NN was raised because of its poor performance on high

dimensional data. Given that we one hot encoded our categorical features, we doubled the dimensionality of our dataset. We still wanted to check if the dimension was low enough to get good results.

Unfortunately, the theory was confirmed, we got the worse results with this method.

Support Vector Classifier (SVC):

This method was interesting because of its kernel trick which allows to classify non-linear problems. If we had a high interaction between features, it would lead to a significant non-linearity and this model would perform better than other ones.

Hence, we performed grid search to both optimize over the slack variable (C constraint) of the support vectors and the kernel. It appeared that the *radial basis function* (rbf) kernel was the one leading to the best results with a recall of 85.05%, confirming that our problem was surely non-linear.

This great performance is also explained by the fact that the SVC handles well high dimensional data: we have 38 dimensions.

Neural Network:

Neural networks being used more and more in the industry because of their capacity of describing very complex models. Their understanding and control remain today a wide field of research and their use is still very specific.

As it is a very different approach, we still wanted to try an implementation of a fully connected neural network to see if we could outperform the other models.

We used *pytorch* library of python allowing to tune the network as we want (nb of layers, nb of units, loss function, etc.).

As expected, the tuning of a neural network being not trivial at all, we couldn't manage with the time we had to perform better than the accuracy of 77.5%.

5.2.3 Back-wise feature selection

For each predictive model, We used the trick of the back-wise feature selection to simplify the dataset and/or enhance the performance.

After performing the predictive model and getting its score, we perform the feature backward selection as follows:

1. Remove feature with replacement one by one and save its associated accuracy
2. After iterating on all the features, keep the best accuracy

3. Compare the best accuracy with the previous one:
 - if higher: remove the associated feature from the dataset and go to 1.
 - else: end the algorithm

RESULT: new subset of features performing better

6 From analytics to business

After having determined the best model (random forest) maximizing the recall, the bank has to run it on its entire data-bank before every campaign in order to obtain a list of their clients in descending order of probabilities of answering positively to a cold call. The new bank clients arrived between two campaigns are automatically added to the targeted list, given that we only have bank informations about them. We still need their informations about cold calls. Hence, they are included with the most likely clients to form the 10'000 targets of the new campaign.

6.1 Analytical results of the model

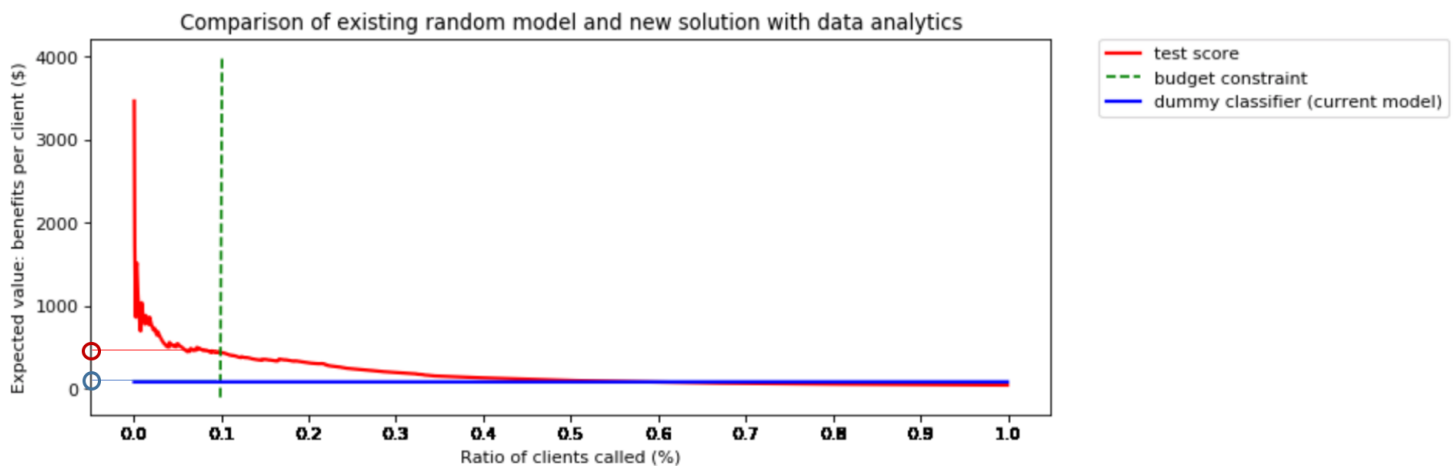


Figure 12: Expected value in function of the percentage of population called

We can see on the graph 12 that the current model (blue) performs the same whichever percentage of the data-bank is called. It shows pretty well that the situation is stagnant.

The graph was obtained by computing the expected value for different ratio of the population called (through different confusion matrices). This would allow to observe if there are optimal ratios for a maximized expected value. For further improvements, this could suggest to change the size of the targeted group and therefore the budget of the campaign.

In theory, the our model (red) should not look like this, especially the part with the low ratio of clients called (calling nobody or almost nobody makes the most money).

We split our dataset of 50/50 of targets 1 and 0 into 80%/20% for training/testing. The testing subset was oversampled on the targets 0 in order to reach the distribution of the true population (data-bank). Indeed, we wanted to have results in accordance with the reality and build a confusion matrix with representative values and compute the expected value with the benefit matrix. This can be explained by the fact that we oversampled our dataset in order to have a 98% ratio of null target. This oversampling of the testing dataset to fit the reality may explain why the curve on figure 12 is behaving weirdly for low percentage of observations.

As expected, the expected value for both models is the same when all the data-bank is called (% of observations equal to 1).

6.2 Benefits for the company

In this section we evaluate quantitatively, through the confusion matrix and benefits matrix (seen in figure 13) the impact of our work on the banks profit.

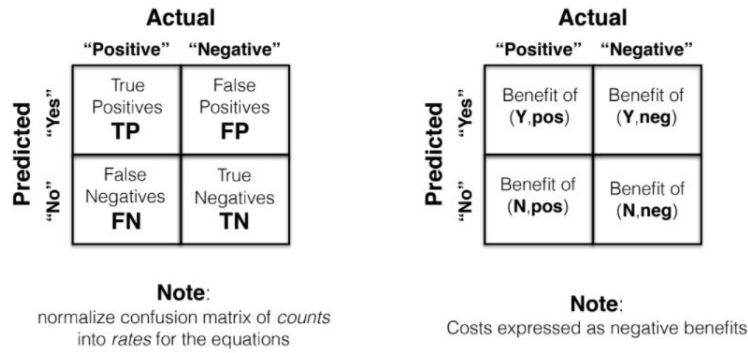


Figure 13: The Confusion Matrix and the Benefits Matrix [3]

A dummy classifier model will take randomly 10% (1645 observations) and have the following result:

As we can see its precision is $\frac{33}{1612+33} = 2\%$ while the one from our model is $\frac{168}{1477+33} = 10\%$.

$$ExpectedValue = p(Y, pos) \cdot b(Y, pos) + p(N, pos) \cdot b(N, pos) + p(N, neg) \cdot b(N, neg) + p(Y, neg) \cdot b(Y, neg) \quad (8)$$

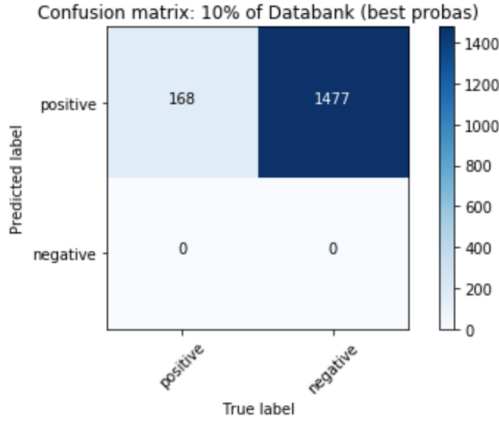


Figure 14: Confusion matrix of random forest model for 1'645 profiles

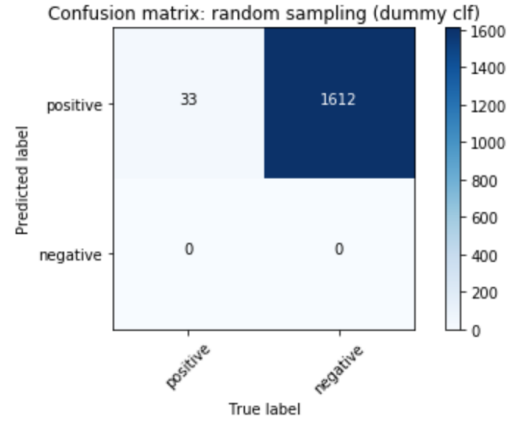


Figure 15: Confusion matrix of the dummy classifier

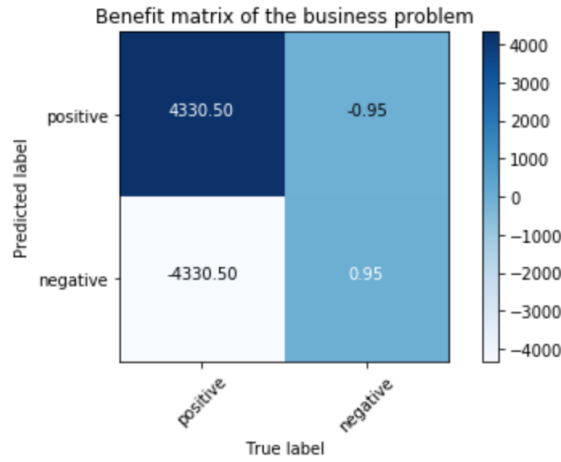


Figure 16: Benefice matrix of our model

The expected value that we provide to the bank is the sum of all elements of the cell-wise product of the benefits (figure 16) and confusion matrix (figure 14) divided by the sum of all cells of the confusion matrix. In the equation 8 the division part is implied by the probabilities which are the rates and not the counts of the confusion matrix.

In order to obtain the values of the benefit matrix, we must compute only three values:

- The benefit of an insured client during the average duration of a car contract
- The cost of a successful call including the salary of the caller
- The cost of an unsuccessful call including the salary of the caller

We reuse the values found in 2.1 to establish the benefit matrix of our problem presented on figure 16. Using its coefficient, we compute the expected value of our new model (figure 14):

$$\begin{aligned} \text{Expected Value} &= \frac{168}{1645} \cdot PV_{\text{Subscriber}} - \frac{1477}{1645} \cdot \text{Cost of unsuccessful call} \\ &= \frac{168}{1645} \cdot 4'431 - \frac{1477}{1645} \cdot 0.947222 \approx \$452 \end{aligned}$$

This value is direct money and not cost reduction since the false negatives and true negatives are null in our model. We could compute the value generated through this model. First, let's compute the increase in expected value we have:

$$\begin{aligned} \Delta \text{Expected Value} &= \text{Expected Value}_{\text{after}} - \text{Expected Value}_{\text{before}} \\ \Delta \text{Expected Value} &= 452 - 89.6 \approx \$362 \end{aligned}$$

This amount is significant and will led to a high increase in profitability. The profits are found by multiplying the expected value by the total number of observations: 16450 in our test, but 10000 in the reality. As mentioned before, it is possible to reuse the confusion matrix of the small sample (1645 observations) to the real one (10'000 observations) because the proportions should not vary much. This allows us to quantify the value generated by computing the profitability ratio. It measures earnings compared to campaign's expenses and other relevant costs.

$$\begin{aligned} \text{Profit}_{\text{ratio}} &= \frac{\text{Profit}_{\text{after}} - \text{Profit}_{\text{before}}}{\text{Profit}_{\text{before}}} \\ \text{Profit}_{\text{ratio}} &= \frac{4'520'000 - 896'000}{896'000} \\ \text{Profit}_{\text{ratio}} &\approx 4.0446 = \mathbf{404.45 \%} \end{aligned}$$

This means we outperform the previous overall productivity of the business campaign by more than 400%.

7 Deployment: how can our analysis be used on an ongoing basis?

7.1 Implementation in the company

Now that we have an effective software model, it still has to be physically incorporated within the company and adapted to existing equipment. One can imagine a more

user-friendly interface for a more intuitive use by the employees. This will be the role of the IT department of the company if it has one, or it will be outsourced to an engineering company.

Employees in the administration charged of establishing the list of the target clients for the next campaign have to be trained to use the model and sensitive to it.

7.2 Improving the model

In order to improve our model, we have to take into consideration two important points: the risk factor and the monitoring aspect. The risk is related to trust and match towards the clients. On the other hand, monitoring can be related to the business itself. In fact, an insurance company has to take into account changes in society or new threats coming from climate change that implies new costumes for car drivers. Our model would have to be updated and diversify over the years and through different campaigns.

The constant shifting of the campaign periods permits the acquisition and discovery of new data which will help improve our model but also find some optimal periods for the call campaigns.

7.3 Reusing the model

The key of reusing such a model is to adapt it to the customers' need. In other words, it has to correspond to the client's profile filled up with its freshly revised characteristics and preferences. It is also important to keep our data bank updated after each campaign. So that we can feed our model with new registered bank customers.

8 Conclusion

The aim of this project was to score every registered customer of the bank in order to find potential new clients for the bank's car insurance. In other words, we wanted to optimize the success of cold calls.

By applying our retained model, we could select profiles that presented the highest probability to subscribe. Thus, we reached through our best model a new much higher expected value per campaign. This led to more than 400% increase in gross profit margin.

However, it is important to remember that our model has its limits. We could improve it and make it more sophisticated. This would allow us to take into account the risk that a person makes a car accident according to his history. Thus, we could adjust our profit

increase into a more realistic value. Indeed, having a maximum of clients is not totally beneficial because some clients can have an important risk of insurance costs. Therefore, having a lot of risky clients can also make loose a lot of money to the bank and to all stakeholders in general.

Furthermore, if we want to go one step further we would have to include the value evolution of each profile. For that, we could segment each customer based on its value generated for the bank. This would mean that we could do a Customer Lifetime Value (CLV) in order to predict the net attributed profit according to the entire relationship with the bank's client.

Moreover, we could apply a similar approach in order to understand how we could lower the car insurance churn rate.

Another way to generate value for the bank would be to reduce the campaign time or the number of employees. This would mean a lower invested capital in marketing for the same outcome as before our involvement.

We could also operate our model on other distribution channels such as email, post or fairs. Ultimately, this study showed how powerful data science and machine learning can be when applied to business related topics. Today, it is certainly a studied field because of its many application and can leverage value in almost every modern company.

9 References

- [1] <https://www.valuepenguin.com/average-cost-of-insurance>
- [2] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4716933>
- [3] Data Science In Practice course
- [4] <https://www.axa.com/fr/investisseurs/rapports-annuels-et-semestriels>

10 Appendix

Find in separate pdf file.