FINAL PROJECT

INTRODUCTION

The present work is a project scope elaborated with the dataset collected by Moro et al. (2011), from a Portuguese bank from the process of 17 marketing campaigns from May 2008 to November 2010, which had the increment of long term deposits, as it main objective of study. The original dataset englobes 59 variables relate with the client information, some information about the first and last contact in the campaign, historical information of previous campaigns, and the results of each campaign in terms of the invested resources. The total number of contacts made by the bank during this time was of 79,354, with a success rate of 8%, when measuring the number of contacts that terminated in deposits effectuated. There are only 17 variables, and 45211 number of instances, which can be understood as clients contacted by the bank in the one campaign, public available.

After a preprocess of the data set which included the division of the dataset in training and testing, selection of variables, treatment of the outliers, and transformation of variables, the final dataset englobes 14 variables, with 33,875 observations for the training part and 8,621 for the testing partition.

Statistical learning is related to a large set of tools that allow the understanding of data. These tools can be classified as supervised or unsupervised. Unsupervised statistical learning can provide an understanding of the relationships and the structure of a dataset, but it does not give a supervised outcome or response (James, G. et al., 2015, p.15). On the contrary, supervised tools, predict or estimate a response from various inputs that can be understood as predictors. Seeing that there is an interested in a prediction method, because there is an associated response variable, which is the historical result of one of the campaigns for each of the contacted clients, called "y", this is a problem that can be solve through supervised tools. Moreover, cause the response is a qualitative variable, the problem can be treated as a classification one, since each observation must be assigned to a category or class.

A model accuracy will be given by the proportion of mistakes that result when comparing the real response of the training data, vs the response obtained through the model. This could be explained by the next equation: $\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$. Where: the $I(y_i \neq \hat{y}_i)$ can be understood as the indicator variable, that takes values of 1 if the response variable predicted $\hat{y}_i$ differs from the veridic y_i response, or 0 if they are the same. This calculation gives a result that can be show in matrix form, and where the diagonal observations are the ones misclassified.

The best method will minimize this classification error and give a more precise response at an overall level considering also other metrics. In the case of the estimation of general liabilities from customer deposits, there can be identify two errors define as: one, the clients which the bank expects them to make a long-term deposit and spend resources by contact them, and in the end, did not effectuated a deposit, and in the other hand, one less shocking to the cost of the bank the ones that where estimated not to perform a deposit and in the end, make one. These errors will be the ones that the bank will take in account, in order to be able to predict the amount of clients that have to contact and estimate the amount of money that they could have available from the clients at a certain period. Deriving these errors in terms of this marketing campaign point of view, can be resume as: the clients that the bank classified as proper to contact for this campaign and in the end, did not make a deposit, and on the contrary, the clients that did not call and make a deposit.

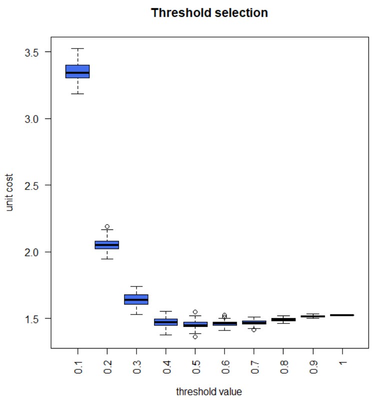
It is no easy to determine the economic cost derived from these errors, because they will be derived from the actual cost of contact and the opportunity cost of contact someone else, which profile is most accurate to be an elected candidate for the campaign and could make a deposit. Besides this, a matrix cost has been determined just for an example purpose for this analysis and which will be used after defining the best method to predict the probability of a contact client making a long-term deposit.

In the previous section logistic regression and Bayes classifiers were used for this purpose, having as a result that the best method from this type is Penalized Linear Regression. In the following section, machine learning tools have been used in order to compare them to PLR and to get the most accurate predictions for the problem.

SENSITIVITY COST ANALYSIS

The sensitivity cost analysis will prove exactly how much cost implies the misclassification of the observations, for this, the bank will have to be aware of the costs that implies each error of its call center. For example, we could determine the cost of the marketing campaign if we say that the cost of contact a client is 1€, if this client does not make a deposit at the end, in order to replace this client we will have to contact at least 13 more if we consider the success rate of deposit that the bank is known to have which is 8%, so the cost of this error will be 14€. On the other hand, if we know that we did not have to call a client because he is no likely to make a long-term deposit, the bank will not lose anything. Additionally, if we do not contact a client but in fact he did a deposit, the bank do not lose, but maybe if the bank had contacted him, the deposit will be effectuated sooner and 13 less calls could have been done, so the cost will be 13€. Finally, the cost of contacting a client that the bank though it will make a deposit and, in fact, did it will be equal to the cost of the call, then it was determined to be 1€.

By effectuating 100 test and train divisions and modeling for each train part a penalized logistic regression, which was tested after, and calculating the cost derivate to the misclassification of the prediction error with every possible threshold from 0.1 to 1, we got an estimated mean and median cost for each threshold. The results tell us that the threshold which minimizes the cost taking as a measure the median is 0.5, moreover, the optimal threshold for the mean is also 0.5.



Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Median	3.343667	2.051373	1.642161	1.473871	1.449218	1.464275	1.468556	1.493062	1.516534	1.523767
Mean	3.347225	2.052486	1.640216	1.472826	1.452105	1.461709	1.468795	1.491627	1.516867	1.523767

The final prediction that will be used is a combination of both thresholds resulting in 0.5, which gives an average final cost for each client of 1.45 € per client after testing the linear penalized regression with the testing dataset the results show good metrics, the sensitivity which measures the rate of success on determining the clients that will not make a deposit is high. Additionally, the specificity is better than other methods tested in the previous part of this project, and finally the type one error rate, which in this case corresponds to the clients that were catalog as suitable to make a long-term deposit and in the end didn't, is low, and the most important thing, it minimizes the cost involve, which is the final goal. Through the profile determined as the most likely to make a deposit, in the logistic regression, it is possible for the bank to tell who are the best clients and to improve the selection of costumers to contact, in that way they will get a higher success rate and with less resources involved.

THRESHOLD 0.5 PLR

		Reference		
		no	yes	*Total*
Prediction	no	7427	735	8162
	yes	195	264	459
	Total	7622	999	8621

	no	yes
no	97.44%	73.57%
yes	2.56%	26.43%
Total	100.00%	100.00%

Accuracy	89.21%
Kappa	31.19%
Sensitivity	97.44%
Specificity	26.43%
Precision	90.99%
F1	94.11%
Cost	1.4556 €

MACHINE LEARNING TOOLS

Machine Learning tools have more emphasis on large scale applications and prediction accuracy. As well with the help of Machine Learning we could examine data, learn from it and predict in order to make decisions. We will test different tools, starting by K- Nearest Neighbors, then SVMs, Decision Trees, Random Forest, Gradient Boosting, and finally, Neural Networks and Deep Learning. Many of the methods will be performed several times in order to get the most accurate metrics. Additionally, the analysis includes the optimization of hyperparameters by a 5 fold cross-validation and as a measure we have incorporate the implicit cost of the errors detailed before.

NEAREST NEIGHBORS

In order to make a prediction for an observation X = x, the K training observations that are closest to x are identified and will be assigned to the class to which these observations belong. This approach is expected to dominate LDA and logistic regression when the decision boundary is highly non-linear because it is a completely non-parametric approach (James, G. et al., 2015, p.163). The Nearest Neighbors requires the previous definition of the hyperparameter K, which is the number of classes for the dataset.

The value of K was chosen automatically using a 5 fold cross-validation approach with a metric created as an Economic function which involves the cost derived by each error. The incorporation of this function allows us to replicate the analysis of cost sensitivity for all the methods that will be perform in the next sections and create comparable results for all methods.

The results show that a K=11, after searching for the best k over 10 options, seems to be best parameter in order to reduce the average cost for the marketing campaign.

KNN ECONOMIC COST

		Reference		
		no	yes	*Total*
Prediction	no	7464	820	8284
	yes	158	179	337
	Total	7622	999	8621

	no	yes
no	97.93%	82.08%
yes	2.07%	17.92%
Total	100.00%	100.00%

Accuracy	88.66%
Kappa	22.25%
Sensitivity	97.93%
Specificity	17.92%
Precision	90.10%
F1	93.85%
Cost	1.5139 €

We can se that the results are like Penalized Regression, but accuracy, kappa and cost are higher, which means that this cannot be consider as the best model.

SVMS

SVMs perform well in a variety of settings and are often consider one of the best classifiers intended for the binary classification setting. When the classes are well separated, SVMs tend to behave better than logistic regression (James, G. et al., 2015, p.357).

SVMs are expensive to train for large problems because they require the testing of various combinations of kernels and model parameters but can deal with non-linear data in high dimensions by the means of this kernel function.

SVMs are very sensitive to the choice of the kernel parameters, so we tested a linear and a radial or Gaussian type to see which one is more accurate and for each we tune some important parameters to find the best, based on the metric of the Economic Function explained before.

SVM LINEAR

		Reference		
		no	yes	*Total*
Prediction	no	7455	812	8267
	yes	167	187	354
	Total	7622	999	8621

	no	yes
no	97.81%	81.28%
yes	2.19%	18.72%
Total	100.00%	100.00%

Accuracy	88.64%
Kappa	22.97%
Sensitivity	97.81%
Specificity	18.72%
Precision	90.18%
F1	93.84%
Cost	1.5173 €

Moreover, the SVM from a radial approach or a RBF kernel, which implies the tuning of at least two parameters the gamma and c. Gamma or sigma in caret determines the reach of a single training instance, with a high gamma value, the boundary for the classification will depend on just the points closer to it, in contrast, a low gamma will result in a more linear decision boundary that will consider points that are further. The parameter c will trade off the correct classification of training examples against the maximization of the decision function`s margin. Large values of c imply that smaller margins will be accepted, while a lower c implies that a larger boundary will be consider (Scikit Learn, 2020). For choosing the best parameters, we change several times the wide of the grid created, in order to be able to find where the best parameters laid. The results show a lower and close to zero gamma (0.05) which means that the boundaries behaved almost linearly but not completely linearly. As the results show, this method performs better that the linear one, the measure of accuracy is higher, the kappa is almost 9% better and the cost is lower in 0.10 cents. Moreover, this method is even better than the PLR, which recall had as a cost 1.45 €, consequently, until now this is by far the best model that we have got.

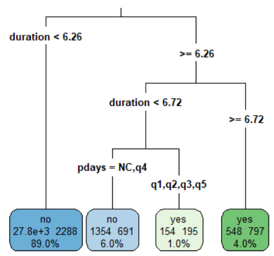
SVM Radial

Prediction	Reference		
	no	yes	*Total*
no	7446	736	8182
yes	176	263	439
Total	7622	999	8621

	no	yes
no	97.69%	73.67%
yes	2.31%	26.33%
Total	100.00%	100.00%

Accuracy	89.42%
Kappa	31.75%
Sensitivity	97.69%
Specificity	26.33%
Precision	91.00%
F1	94.23%
Cost	1.4262 €

DECISION TREES



Decision trees solve classification problems through the prediction of each observation and assign it to the most commonly occurring class of training observations in the region to which it belongs.

As the results show, this method uses just two variables in order to classify the clients: "duration", which is the time of the last contact, and "pdays", which englobes the number of days that passed after the client was contacted for previous campaigns and that has been categorized in 6 groups based on the quintiles of the variable, where the first group are the clients that were not contacted before. When testing this method we can see that the results are not so good, they are better than KNN and Linear SVM but not as good as PLR and Radial SVM, the average cost lays in 1.48€, and the accuracy is lower what is reflected in the sensitivity and specificity rates reach by this model.

DT

Prediction	Reference		
	no	yes	*Total*
no	7437	767	8204
yes	185	232	417
Total	7622	999	8621

	no	yes
no	97.57%	76.78%
yes	2.43%	23.22%
Total	100.00%	100.00%

Accuracy	88.96%
Kappa	27.84%
Sensitivity	97.57%
Specificity	23.22%
Precision	90.65%
F1	93.98%
Cost	1.4839 €

Trees can be displayed graphically and are easily interpreted but do not have the same level of predictive accuracy and they are not as robust as other classification approaches. However, when combining a large number of models, the accuracy predicted can often present a great improvement, this is the main reason why random forest and gradient boosting are good ways to improve the results of tree predictions, which be develop in further sections. Additionally, some improvements can be reach through some simpler modifications, one example is the C05 methodology which considers boosting. Through C05 we could reach a cost of 1.42€ when applying it to the training dataset. If we test this tool with the partition for testing there is an improvement from simple Decision Trees, as results shown, there is a lower average cost per client of 1.47€ and there is a mayor improvement in the kappa measure and which is showed also in the increase of the specificity rate as more clients that will make a deposit are identify in advanced by the algorithm.

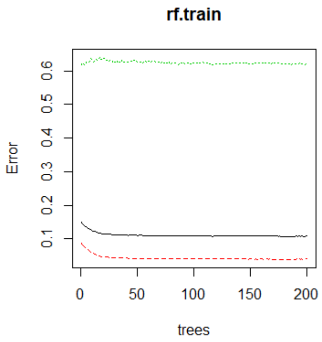
DT C05		Reference		
		no	yes	*Total*
Prediction	no	7394	712	8106
	yes	228	287	515
	Total	7622	999	8621

	no	yes
no	97.01%	71.27%
yes	2.99%	28.73%
Total	100.00%	100.00%

Accuracy	89.10%
Kappa	32.60%
Sensitivity	97.01%
Specificity	28.73%
Precision	91.22%
F1	94.02%
Cost	1.4772 €

RANDOM FORESTS

Random Forest, implies the construction of multiple trees and at each split in each tree, the algorithm is not even allowed to consider a majority of the available predictors. Hence, it avoids correlation problems by forcing each split to consider only a subset of the predictors, so when doing Random Forest, it is necessary to define the m size of predictors that will be consider each time. Since we have 14 variables, we will start by performing a random forest with 10 variables as a default number of parameters selected for each tree division and with 200 trees to see the first results of this method. The error rate which can be simply interpreted as the fraction of the training observations in that region that do not belong to the most common class, will act as the criterion for making the binary splits. This can be defined as:



$$E = 1 - \max_{k}(\widehat{p_{mk}})$$

Where $\widehat{p_{mk}}$ is the proportion of training observations in the mth region that are from the kth class (James, G. et al., 2015, p.312). As seen in the plot, the error decreases in concordance with the addition of more trees, then more than 200 will be the necessary number of trees, at least.

By looking at the prediction results, we can conclude that this method is the second best with a resulting average cost per client of 1.4533€ which is really similar to PLR, it even has an accuracy measure greater than Radial SVM of 0.04% more, which can be related to the better balanced of the specificity and sensitivity. Thus, this method implies a higher cost because it sacrifices sensitivity for more specificity.

RF		Reference		
		no	yes	*Total*
Prediction	no	7317	605	7922
	yes	305	394	699
	Total	7622	999	8621

	no	yes
no	96.00%	60.56%
yes	4.00%	39.44%
Total	100.00%	100.00%

Accuracy	89.44%
Kappa	40.76%
Sensitivity	96.00%
Specificity	39.44%
Precision	92.36%
F1	94.15%
Cost	1.4533 €

Although, Rain Forest is computationally expensive allows us to have more accurate and variance reduce predictions that can handle categorical variables and non-linear relationships like the ones that conform this dataset, so we have tried to achieve better metrics through parameter tuning.

First, we did a hyper parameter tuning of the number of variables that will be selected, and we proved thought the Economic Function the best by checking 300 trees realizations. The results point that the optimal number is 4, given the next results.

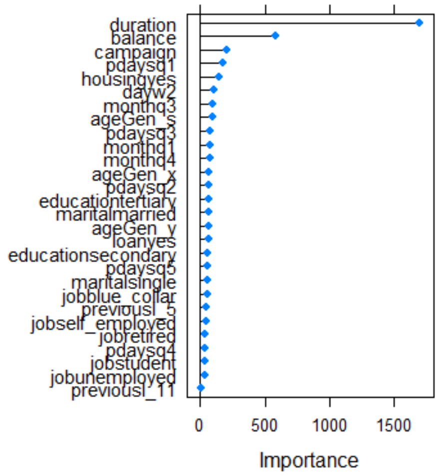
RF ECONOMIC COST

		Reference		
		no	yes	*Total*
Prediction	no	7472	778	8250
	yes	150	221	371
	Total	7622	999	8621

	no	yes
no	98.03%	77.88%
yes	1.97%	22.12%
Total	100.00%	100.00%

Accuracy	89.24%
Kappa	27.73%
Sensitivity	98.03%
Specificity	22.12%
Precision	90.57%
F1	94.15%
Cost	1.4424 €

These show that a lower cost measure is achieved by sacrificing specificity to gain sensitivity, this can be seen as a better method that the last presented. Moreover, decision trees can also give us information about the most important variables in order to performed classification. The next plot presents the variables organized by importance, as we can see the ones that highlighted are "duration", which as explained before gives the time of contact, "balance" which is the average yearly balance position of the client in the bank. Next, we can see some variables that appear to be less important that the two described: "campaign" that gives the number of contacts to the client that the bank has made before, "pdays" specifically if the clients were contacted in less than 110 days after the last contact and if the clients have a housing loan. This will be the 5fith variables mostly used to classified clients in the two categories "yes" if they are expected to perform a deposit and "no" if not.



Now, we will performed a second parameter tuning in the sample size, since there is evidence that generally adjusting the biased trade off, could lead to a better performance of the method. Then, if the sample is greater, the results will be less random, in contrast, if we decrease the sample size, the variation in the individual trees will increase preventing overfitting. After reviewing the results, we can conclude that this is in fact an improve model and that by far is the best that we have gotten. By comparing this, to the Radial SVM which is the second best model, we can see that the sensitivity is lost by 0.66%, but instead the specificity presents an important growth of 6.21%, what makes the kappa go up to 36.72% and the accuracy to 89.56%, that are the best measurements that we have reach. Moreover, the price is the lowest with an average per customer of 1.4211 €.

RF SAMPLESIZE

		Reference		
		no	yes	*Total*
Prediction	no	7396	674	8070
	yes	226	325	551
	Total	7622	999	8621

	no	yes
no	97.03%	67.47%
yes	2.97%	32.53%
Total	100.00%	100.00%

Accuracy	89.56%
Kappa	36.72%
Sensitivity	97.03%
Specificity	32.53%
Precision	91.65%
F1	94.26%
Cost	1.4211 €

Another option that we could try is using a subsampling technique. Given that there is un unbalanced class in the dataset, since there are more people that did not make a deposit we could use the "up" subsampling which samples with a replacement the minority class to be the same class as the majority class. As we can see, there is an improvement in the past results in terms of specificity and kappa measures, but the accuracy achieved is lower with a rate of 89.56%. However, this model could help us to see how the cost change will if there is a more balance behavior in terms of deposits.

RF SUBSAMPLING

		Reference		
		no	yes	*Total*
Prediction	no	7020	445	7465
	yes	602	554	1156
	Total	7622	999	8621

	no	yes
no	92.10%	44.54%
yes	7.90%	55.46%
Total	100.00%	100.00%

Accuracy	87.86%
Kappa	44.52%
Sensitivity	92.10%
Specificity	55.46%
Precision	94.04%
F1	93.06%
Cost	1.7129 €

GRADIENT BOOSTING

Like Random Forest, Gradient Boosting implies the construction of a set of decision trees, but while Random Forest builds each tree independently, Gradient Boosting builds one tree at a time. This additive model or what is called ensemble works in a forward manner in contrast to Random Forest. Additionally, Gradient Boosting combines results along the way instead of combining results at the end of the process using by averaging or "majority rules". Gradient Boosting can result in a better performance when hyper tuning parameters, but it implies a lot of time in order to get a more accurate result than Random Forest.

The first model will consider the definition of the following parameters:

- The total number of trees to fit (400)
- The number d of splits in each tree, which controls the complexity of the boosted ensemble (2)
- The learning rate or how quickly the algorithm proceeds down the gradient descent. Smaller values reduce the chance of overfitting, so we are going to use a 0.01.
- And, subsampling which controls whether you use a fraction of the available training observations. Using less than 100% of the training observations means that we are implementing stochastic gradient descent, and this could help to This can help to minimize overfitting.

This model presents some more accurate results than models like KNN, Lineal SVM, DT, and the first Random Forest, but it is not comparable to Radial SVM, and the last Random Forest.

GRADIENT BOOST

		Reference		
		no	yes	*Total*
Prediction	no	7511	827	8338
	yes	111	172	283
	Total	7622	999	8621

	no	yes
no	98.54%	82.78%
yes	1.46%	17.22%
Total	100.00%	100.00%

Accuracy	89.12%
Kappa	22.89%
Sensitivity	98.54%
Specificity	17.22%
Precision	90.08%
F1	94.12%
Cost	1.4473 €

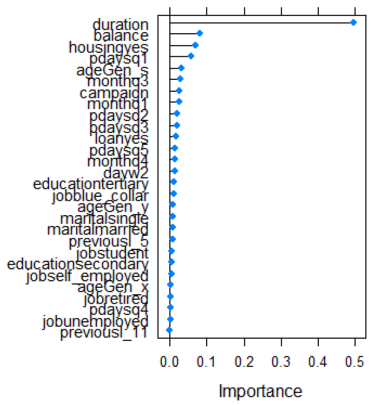
Next, we will perform an optimization of some hyperparameters by creating grids to test all the possible combination between parameters and see if the results get better. After reviewing the results, we can see that we got better and more accurate results. This, in fact is the best model that we have gotten, taking in account accuracy and kappa measures. When comparing the average cost, is not the best but it is really closed to the Random Forest results with an average value of 1.4225€, so 0.0014 cents higher.

GRADIENT BOOST OPT

		Reference		
		no	yes	*Total*
Prediction	no	7366	640	8006
	yes	256	359	615
	Total	7622	999	8621

	no	yes
no	96.64%	64.06%
yes	3.36%	35.94%
Total	100.00%	100.00%

Accuracy	89.61%
Kappa	39.11%
Sensitivity	96.64%
Specificity	35.94%
Precision	92.01%
F1	94.27%
Cost	1.4225 €



Also we have checked the importance of the variables in the classification method, as the plot shows, the two most important variables remain as “duration” and “balance”, but also we can see that campaign is no longer in the top 5 and that there is a smaller difference between the second and the third and four variables which are “pdays” the first quarter, and if the client has a housing loan. The fifth variable appears to be if the client is part of the S Generation or the ones that were born in 1945 and before.

NEURAL NETWORKS

Neural Networks is one of the most successful approaches that can handle non-linear data. This method consists of an artificial network of functions, called parameters, which allows the computer to learn, and to fine tune itself, by analyzing new data. Each parameter will act as a function which produces an output, after receiving one or multiple inputs. Those outputs are then passed to the next layer of neurons, which use them as inputs of their own function, and produce further outputs. This will be repeated until there is one final result for the model.

Results show that this method, for this case, is not as accurate as Random Forest, there is an interchange of specificity and sensitivity that increases the final average cost of the client.

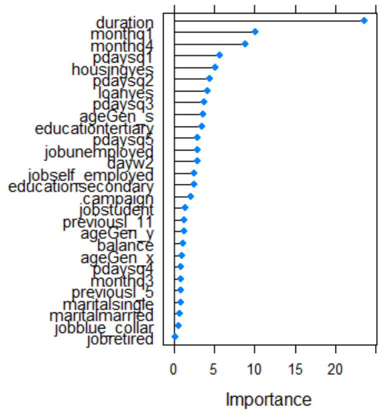
NEURAL NETWORK

		Reference		
		no	yes	*Total*
Prediction	no	7298	633	7931
	yes	324	366	690
	Total	7622	999	8621

	no	yes
no	95.75%	63.36%
yes	4.25%	36.64%
Total	100.00%	100.00%

Accuracy	88.90%
Kappa	37.41%
Sensitivity	95.75%
Specificity	36.64%
Precision	92.02%
F1	93.85%
Cost	1.5231 €

Moreover, if we check the importance of the variables in this method, we see, once again, that "duration" is the most important one by far, and there are other variables that appear now, like "month" which measures the moth when the called was made and then we see "pdays" and if the client has a housing loan.



DEEP NEURAL NETWORK

If a Neural Network has several hidden layers, then, it is said that this is a Deep Neural Network. Having more layers could be better to find more accurate predictions. This method is computationally expensive, but we can prove the optimal number of layers in small steps, by testing if the rest are zero at first so that we can check if the number just lays at the beginning.

DEEP NEURAL NETWO

		Reference		
		no	yes	*Total*
Prediction	no	7380	661	8041
	yes	242	338	580
	Total	7622	999	8621

	no	yes
no	96.82%	66.17%
yes	3.18%	33.83%
Total	100.00%	100.00%

Accuracy	89.19%
Kappa	35.10%
Sensitivity	96.82%
Specificity	33.83%
Precision	91.78%
F1	94.23%
Cost	1.4290 €

Results show good accuracy and kappa measures, there is an interchange of sensitivity and specificity that increments the cost a little when comparing it to Random Forest, but it is minimal and reaches a value of 0.0079 cents. This could be one of the best models that have been tested.

ENSEMBLE METHODS

Ensemble methods are a supervised learning algorithm that as the rest can be trained and tested with new data to reach predictions. Ensemble methods improve the predictive performance of a given model or fitting technique. There are several advantages derived to the use of what are called meta-learning methods: there is a variance reduce since results are less dependent of a single model, also the result could be less bias because multiple models can be able to result in more reliable predictions.

If we take the best models that we have tested, then there is a greater probability that the results will be more accurate. Consequently, we have taken the best models in term of accuracy and type and we have selected 7 models that will be combined to reach one final prediction.

In the first part of this project we reached to the results presented in the table, as said before the best model was Penalized Logistic Regression and after testing all the possible cutoffs it was proved that 0.5 was the best.

PROBABILISTIC METHODS	Accuracy	Kappa	Sensitivity	Specificity
LOGIT	89.22%	31.27%	97.45%	26.43%
PENALIZED LOGIT	89.15%	30.63%	97.44%	25.93%
LDA	88.92%	29.94%	97.18%	25.93%
QDA	83.44%	26.92%	89.03%	89.03%
NAÏVE BAYES	86.83%	30.06%	93.78%	33.83%
PENALIZED LOGIT ROC	89.21%	26.16%	98.23%	20.42%
THRESHOLD 0.3 PLR	78.47%	36.58%	77.92%	82.68%
THRESHOLD 0.5 PLR	89.19%	35.10%	96.71%	31.83%

Moreover, after performing Machine Learning tools, we can see that the probabilistic ones are less accurate, then we will just take PLR for the ensemble.

The results from Machine Learning tools have reached an accuracy over 88%, considering the different accuracies, kappa and the average cost per client predicted for each method we can compared them and select the best ones in order to perform the ensemble.

MACHINE LEARNING	Accuracy	Kappa	Sensitivity	Specificity	Average cost
KNN ECONOMIC COST	88.66%	22.25%	97.93%	17.92%	1.51 €
SVM RADIAL	89.42%	31.75%	97.69%	26.33%	1.43 €
SVM LINEAR	88.64%	22.97%	97.81%	18.72%	1.52 €
DT	88.96%	27.84%	97.57%	23.22%	1.48 €
DT C05	89.10%	32.60%	97.01%	28.73%	1.48 €
RF	89.44%	40.76%	96.00%	39.44%	1.45 €
RF ECONOMIC COST	89.24%	27.73%	98.03%	22.12%	1.44 €
RF SAMPLESIZE	89.56%	36.72%	97.03%	32.53%	1.42 €
RF SUBSAMPLING	87.86%	44.52%	92.10%	55.46%	1.71 €
GRADIENT BOOST	89.12%	22.89%	98.54%	17.22%	1.45 €
GRADIENT BOOST OPT	89.61%	39.11%	96.64%	35.94%	1.42 €
NEURAL NETWORK	88.90%	37.41%	95.75%	36.64%	1.52 €
DEEP NEURAL NETWORK	89.19%	35.10%	96.82%	33.83%	1.43 €

There are several approaches that we could consider when doing the ensemble. The first one was to consider the mode for each observation from all the 7 model, the results we get are the most accurate and the average cost is the lowest.

ENSEMBLE MODE

Prediction	Reference: no	Reference: yes	Total
no	7375	645	8020
yes	247	354	601
Total	7622	999	8621

	no	yes
no	96.76%	64.56%
yes	3.24%	35.44%
Total	100.00%	100.00%

Accuracy	89.65%
Kappa	38.93%
Sensitivity	96.76%
Specificity	35.44%
Precision	91.96%
F1	94.30%
Cost	1.4148 €

The second approach is to consider the mean for each observation of the probabilities assign for all the 7 models and then considering a threshold of 0.5, a classification has been done. This is the best result with a little lost in accuracy compare to the last method, but the cost 1.4122€ is smaller which is the final goal of the process.

ENSEMBLE MEAN

Prediction	Reference: no	Reference: yes	Total
no	7392	663	8055
yes	230	336	566
Total	7622	999	8621

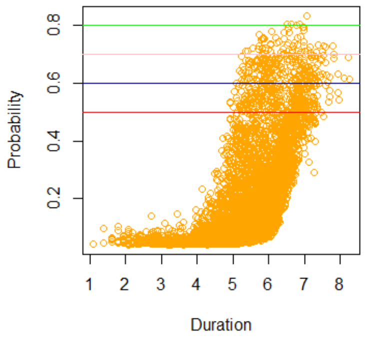
	no	yes
no	96.98%	66.37%
yes	3.02%	33.63%
Total	100.00%	100.00%

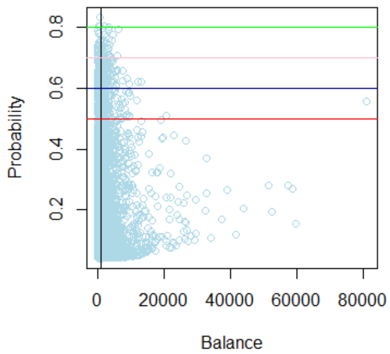
Accuracy	89.64%
Kappa	37.72%
Sensitivity	96.98%
Specificity	33.63%
Precision	91.77%
F1	94.30%
Cost	1.4122 €

After we have reach this final prediction we can conclude that the average cost derived from misclassification per client will be 1.4122 €, this means that if we consider that in the future this model is used then the average cost determine for the marketing campaign will be 112,077.84 € if the same amount of clients (79,354) is estimated to be contacted.

Moreover, when looking at the importance of the variables for the classification methods we see that "duration" and "balance" have a higher impact. The "duration" of the previous contact will be linked to the previous acceptance that the client has shown, consequently, if we do not have contacted the client before this campaign, is less likely that the contact has a success. Additionally, the variable "balance" is related with the position of the client in the bank. This variable keeps an interesting information that it is relevant for the campaign; if the yearly position is higher, then the probability of doing a deposit is higher. The times that the client was contacted previously will depend on the previous campaign and will not change, but the yearly balance is most likely to continue to change and so it is imperative that the information will be update, so more opportunities of identifying clients that are likely to make a long term deposit are taken. One analysis that could be performed is the classification and determination of the clients that involve a higher possibility of making a deposit based on these two variables, in order to determine groups of priority based on probability bands. When plotting duration and balance against the probability of deposit linked to every client through the predicted ensemble model mentioned, we could identify the clients that are most likely to make a deposit that have already been contacted and the ones that have a higher probability and also have positive balance in their positions.

Based on duration we could start by contact the ones above the green line, which correspond to 1% of the total tested, this will be the ones with higher probabilities, next the ones above the pink line (1%) and followed by the purple line (2%) and red band (3%). If we performed the contacts by groups is more efficient since they are small in number and we could reach one first goal budget early. Then we could go to the ones that are in grey zone, and that have been contacted early but they probability does not seem to be high, in order to convince them, this group will imply a harder work and a higher cost but with some special approaches maybe based on rate or services the success rate could increase.





On the other hand, based on the measure of Balance, the clients can be also scattered based on their success probabilities. Again, we could divide the population tested in groups. The groups with higher probabilities will be the ones that will be contact first, since they are expected to perform a deposit without higher marketing efforts. The first group will be the ones above the green line (less than 1% of the total population), next the ones above the pink and below the green (1%), followed by the purple (2%) and the red one (3%). Moreover, in this case there could also be a discrimination based on the clients that have a certain amount of money as balance for example those with capitals more than 1000€.

There could be various ways of addressing the predictions from a classification analysis, we have explore some options as an example, but it is clear that they are a powerful tool to achieve a clearer picture of the clients of a bank and start from there to see which ones will be more likely to make a deposit involving less costs in the process.

BIBLIOGRAPHY

Moro, S., Laureano, R., and Cortez, P. (2011). *Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology.* In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October 2011. EUROSIS. Retrieved from: https://core.ac.uk/download/pdf/55616194.pdf

Kuhn, M. (2019). *The caret Package*. Retrieved from: https://topepo.github.io/caret/index.html

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2015). An Introduction to Statistical Learning with Applications in R. Springer. Heidelberg Dordrecht, London.

Scikit Learn (2020). *RBF SVM parameters.* Retrieved from: https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html