Group 8 Members: Yijing Zhou, Chanelle Reynolds, Yidi Zhou, Kaan Tuncel

Data Mining Final Project

Professor Foster Provost

May 12, 2018

Bank Telemarketing

Business Understanding

We are the newly minted Data Science division of a prominent Portuguese bank. Our domain-experts assure us that competition in the retail sector of Portuguese banking is cut-throat and so growth comes from fighting over market share. (Pinho, 2000) Additionally, banking is dominated by "relationship value" where a higher relationship means that a single consumer has more accounts and products open with a single bank. Higher relationships mean, ceteris paribus, that a customer is likely to have a lower default rate, lower churn and have higher utilization rates. (Agarwal, 2009) As a result, the marketing program of the Bank is very important in extending products and accounts and deepening these relationships. As a first test, the Bank has tasked the DS group with researching potential targeting in the direct call marketing for Term Deposit (TD) products.

Each client may have different responses to the marketing phone call: some may immediately subscribe after receiving one call; some may refuse when initially contacted, but ultimately become attracted the offer and subscribe after consideration; some may never subscribe even after tons of calls. The business problem here, for the bank, is to predict the probability that a certain customer is going to subscribe, so as to be able to efficiently allocate resources in its marketing.

The data science formulation of this problem is to be able to predict whether a particular customer is worth calling. This represents a supervised learning problem, as the target variable is well-defined ("How likely is customer X to respond well to a targeted call?"). We are taking the approach of classifying the population into two classes: a "yes" class who will respond to a call by signing up to a TD product and a "no" class who will not. This is an approximate solution, as it leaves out a potentially significant middle ground (for example, customers who need multiple calls to be convinced or those who would be nudged with a special offer). However, we believe this simple two-classification solution is adequate as the foundations of this program and represent a significant enough step-up from what we assume is a non-targeted legacy program, that we could show the benefits of targeting in the form of improved profit. Framing the problem in terms of expected value will be helpful for demonstrating this and will form an integral part of our modelling. Therefore, we think the approximation is more than adequate and the data science formulation does a good job of potentially solving the business problem.

Data Understanding

The bank provided our DS team with a full dataset of approximately 45,000 instances which reflected demographics about their targeted clients and past campaign. These instances were collected between May 2008 to Nov 2010 when the customer attrition team ran its campaign aimed at increasing subscriptions to the term deposit service at the bank. This is a supervised model with target variable being: the probability that customer "X" will subscribe to the bank's term deposit after the last contact of the campaign. There is no time limit between the last contact and the result (i.e. the result will not be different whether customer "X" subscribed immediately after the phone call or he/she subscribed three months after last phone call). The target variable is a binary variable:

either “yes” or “no”. There are 20 features in total with three segments: bank client features; contact features and social and economic context attributes.

We accessed this data from the banks internal server. The dataset was labelled as the "Bank Marketing" UCI dataset (please check the description at: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing). Credit for sharing the original sources should be given to: [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014.

Bank client features:

1) Age: the age of the client (numeric);
2) Job: type of job (categorical: “admin”, “blue-collar”, “entrepreneur”, “housemaid”, “management”, “retired”, “self-employed”, “services”, “student”, “technician”, “unemployed”, “unknown”);
3) Marital: marital status (categorical: “divorced”, “married”, “single”, “unknown”)
   Note: “divorced” means divorced or widowed;
4) Education (categorical: “basic.4y”, “basic.6y”, “basic.9y”, “high.school”, “illiterate”, “professional.course”, “university.degree”, “unknown”);
5) Default: has credit in default or not (categorical: “no”, “yes”, “unknown”);
6) Housing: has housing loans or not (categorical: “no”, “yes”, “unknown”);
7) Loan: has personal loan or not (categorical: “no”, “yes”, “unknown”);

Contact features:

1) Contact: contact communication type (categorical: “cellular”, “telephone”);

2) Month: last contact month of year (categorical: “jan”, “feb”, “mar”, …, “nov”, “dec”);
3) Day_of_week: last contact day of the week (categorical: “mon”, “tue”, “wed”, “thu”, “fri”);
4) Duration: last contact duration, in seconds (numerical);
5) Campaign: number of contacts performed during this campaign and for this client, includes last contact (numerical);
6) Pdays: number of days that passed by after the client was last contacted from a previous campaign (numerical, 999 means client was not previously contacted);
7) Previous: number of contacts performed before this campaign and for this client (numerical);
8) Poutcome: outcome of the previous marketing campaign (categorical: “failure”, “nonexistent”, “success”);

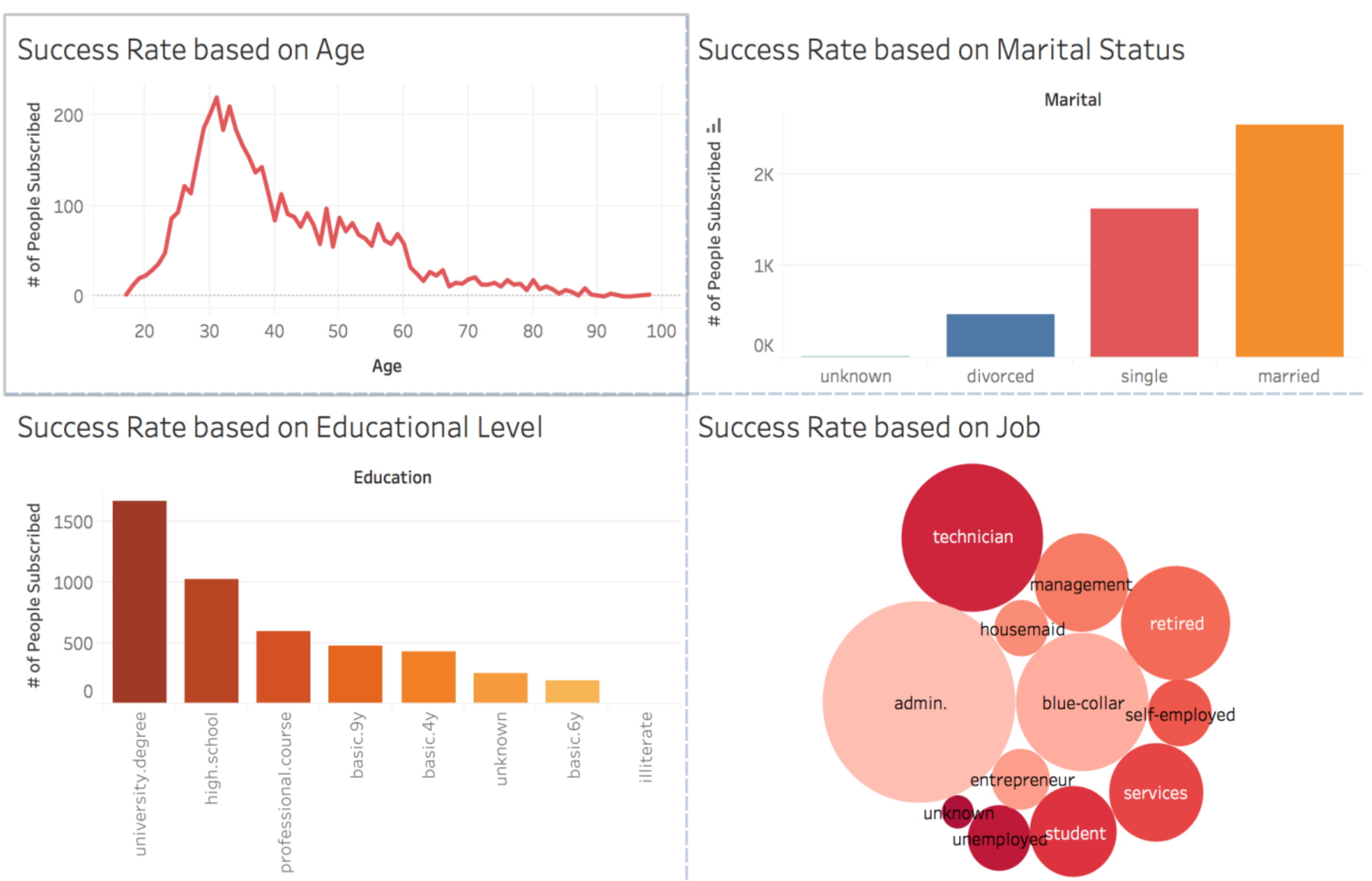
Social and economic context attributes:

1) Emp.var.rate: employment variation rate (numerical, quarterly indicator);
2) Cons.price.idx: consumer price index (numerical, monthly indicator);
3) Cons.conf.idx: consumer confidence index (numerical, monthly indicator);
4) Euribor3m: euribor 3-month rate (numerical, daily indicator);
5) Nr.employed: number of employees (numerical, quarterly indicator).

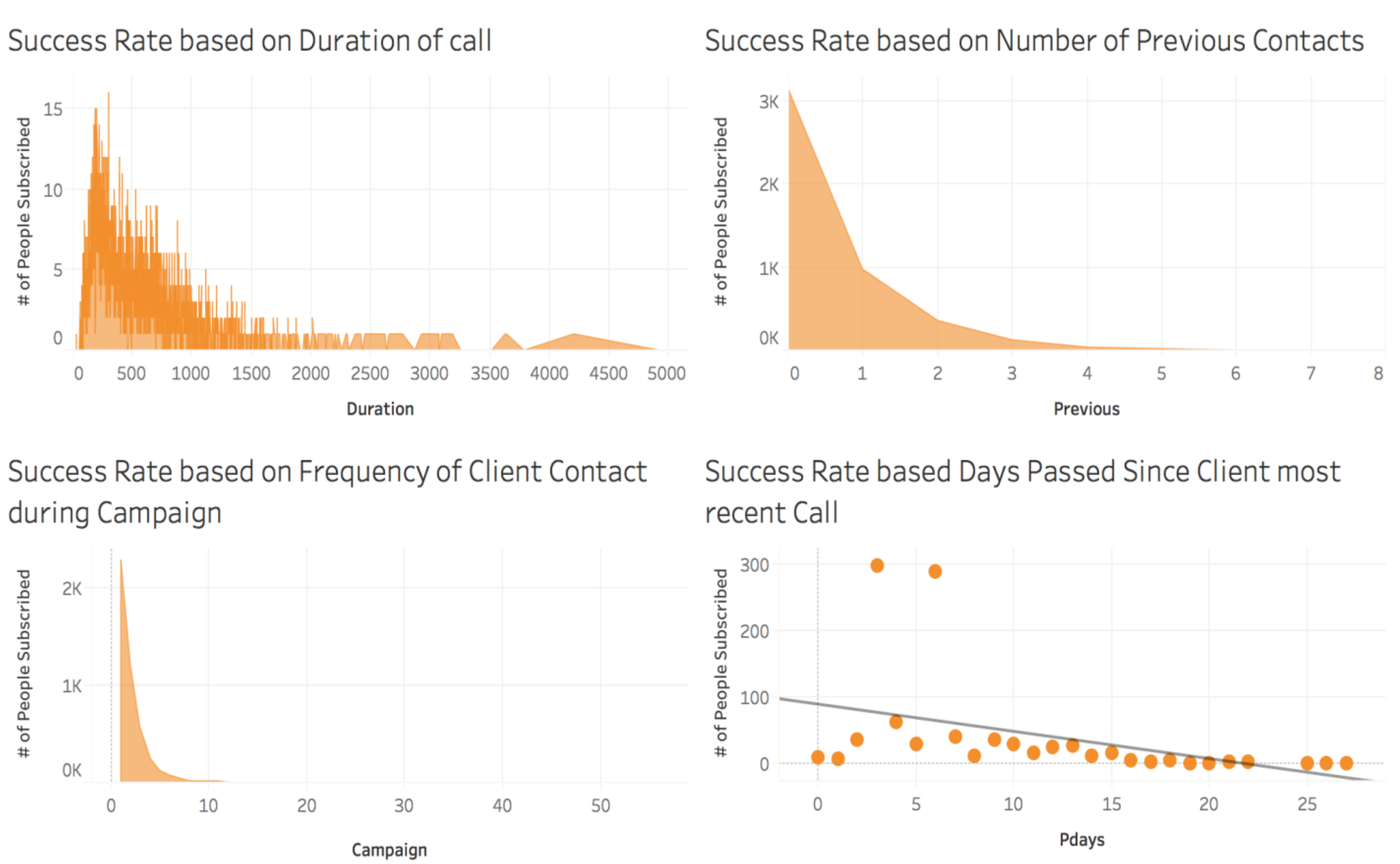
Note: These social and economic context attributes are updated with the last contact of each client. For instance, if a client’s last contact is May 5th, let’s say Monday, the corresponding context attributes will be recorded as May 5th’s data, whatever daily, monthly or quarterly adjusted.

<u>Data Preparation (Data Visualization)</u>

From the 45,000 instances above, we take out 10% randomly as hold out (testing) data and the rest remains training data. Since the original data file is pretty “clean”, we didn’t spend too much time in data preparation. However, after initial modelling, we found the duration feature to be exceptionally predictive. As such, we removed it afterwards to create a more reflective model (please refer to modelling part for detail). Rather, we choose to focus on data visualization in order to better understand the data and analyze the relationship between predictors and the target variable—How much predicting power does the predictor have? Why does it have so much power? Would there be a leakage of data? (the latter two questions will be explored in modelling part)



The above graphs compare demographics or the "bank client" features to the number of people who subscribed after they were contacted as part of the campaign. As can be seen, 30-35 year olds are most likely to subscribe compared to all other age groups. This might be due to the fact that they are at a stage where they are thinking about their financial security and their future. However, as people get older, they are likely to have already settled with and are loyal to a specific bank. As such, they are less inclined to subscribe. Conversely, individuals under 30 may not be ready to settle down with a bank or may be turning to more technologically advanced alternatives such as saving apps. Additionally, married individuals were more likely to subscribe after being contacted as part of the campaign. Similarly, married couples are more likely to be concerned with financial planning, potentially due to the financial demands of having a family or starting one than single and divorced couples. Not surprisingly, a higher level of education was linked to an increased likelihood of someone subscribing to the bank. This is likely to be due to increased income, increased concern and knowledge about family planning and increased comprehension of the offer from the bank. Individuals with lower education also tend not to believe in banks and prefer to store their money in safe places at home. The only exception to this rule was people with basic 6 yr education being less likely to subscribe than those with a basic 4 year education. In relation to success based on jobs, technicians, admins and blue-collar workers were more likely to subscribe than other jobs such as students, retired individuals and managers. The latter jobs were likely to have less need for a banks.



The above graphs show the success rate based on the measures of the campaign. Interestingly, shorter durations on calls were more likely to result in a higher subscription to the bank. This is likely because people become hooked or decide to subscribe within the first few seconds of hearing an offer. Additionally, people who were contacted less on the campaign were more likely to subscribe to the bank. Similarly, clients that were contacted less times during the campaign were more likely to subscribe. This could be due similar reasoning given that people who were interested in the offer were likely to subscribe after the first or second contact. They didn't need constant convincing. For the fourth graph, we removed the data points for the people who never previously contacted denoted by '999' as it was distorting the pattern in the graph. As seen in the graph, as more time passed after the last contact, the lower the chances for contacted individuals to subscribed. If people are sold on an offer, they are more likely to subscribe immediately. If people wait a longer time, they are likely not interested in the offer. A noteworthy observation from working with the data is that when the ‘999’ data points were added, majority of the subscriptions as related to the number of days passed after last contact came from people who were not contacted during the campaign.

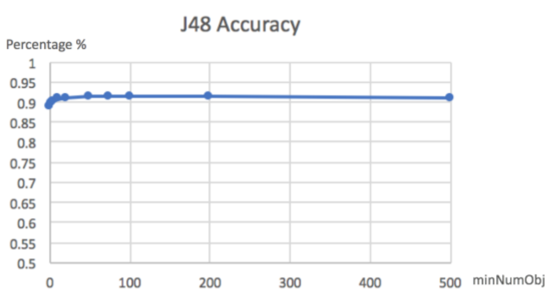
Modeling

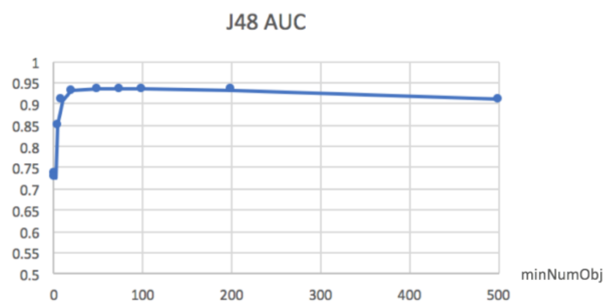
We decided to use WEKA and try three different classifiers: J48 decision tree; Random Forest and Logistic Regression. In order to make the results more accurate, we choose cross-validation with 10 folders for building all the models. There are two results we consider: accuracy and area under ROC curve (AUC). Accuracy gives us the percentage of prediction that is correctly made. It is quite intuitive and easy to understand. A ROC graph depicts relative tradeoffs that a classifier makes between benefits (true positives) and costs (false positives). We rely more on the result of ROC graph since it decouples classifier performance from conditions under which the classifiers will be used. Specifically, they are independent of the class proportions as well as the costs and benefits. It’s also more accurate for skewed or unbalanced data. We also use hold out (testing) data to test our models by comparing the accuracy, AUC as well as lift curves. Lift curves (cumulative response curve) represents the advantage the model provides over random guessing.

1) J48 Decision Tree:

The complexity parameter for this model is MinNumObj. We set “Unpruned” to be true and change MinNumObj each time. The results are shown as below:

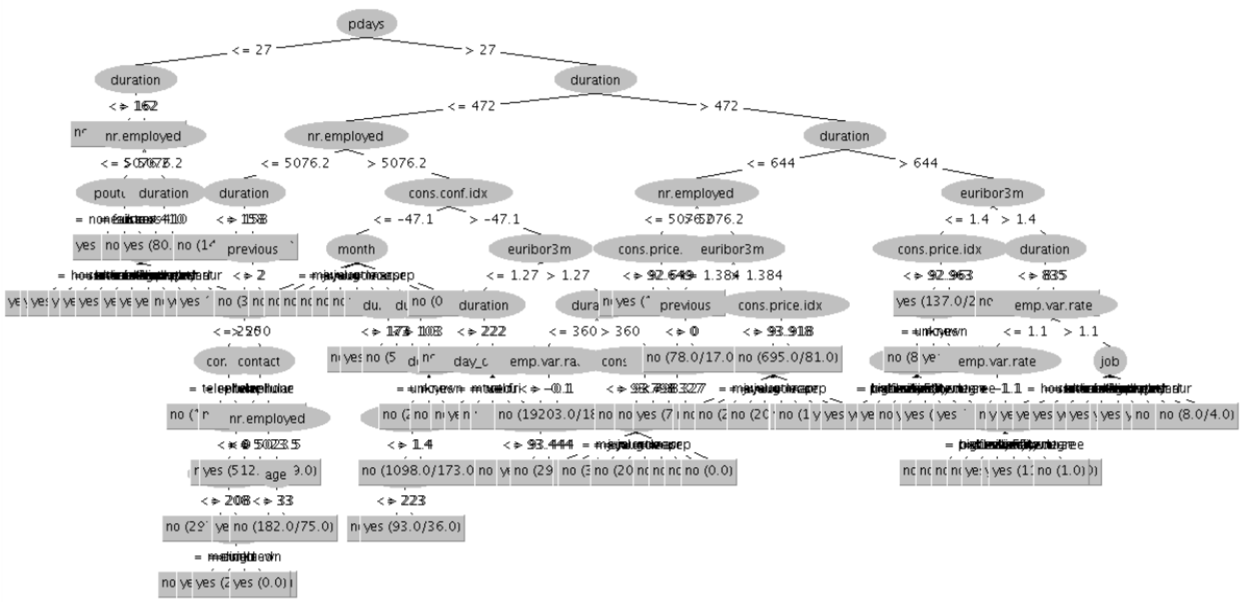
J48 Decision tree		
MinNumObj	Accuracy	AUC
1	88.80%	0.736
2	89.50%	0.727
5	90.11%	0.849
10	90.79%	0.908
20	91.02%	0.93
50	91.30%	0.934
75	91.34%	0.935
100	91.28%	0.934
200	91.28%	0.932
500	90.90%	0.91





By changing the complexity parameter from 1 to 500 (increase complexity by reducing MinNumObj), we can see that the accuracy almost remains the same while AUC first increases from below 0.75 to almost 0.95, and then slightly drop back. Therefore, we picked our best J48 model by setting MinNumObj to 75, with an accuracy of 91.34% and AUC of 0.935.

An advantage of decision tree is that we can visualize the model. Therefore, we could check what predictors are the most powerful and may arouse our further attention.

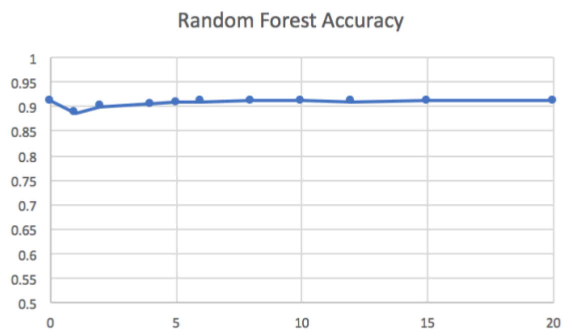


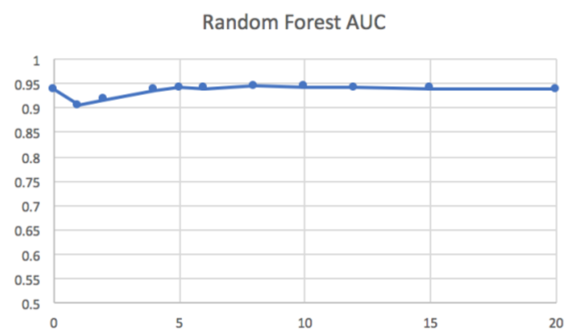
From the top of the tree, we can see that the most important features are: pdays, duration, nr.employed, euribor3m, cons.conf.idx and poutcome. We will discuss these features after building the three “best” models.

<u>Random Forest:</u>

For complexity parameter, we change maxDepth from 0 to 20 (increase complexity by increasing maxDepth) and get the following result:

Random Forest		
maxdepth	Accuracy	AUC
0	91.14%	0.939
1	88.73%	0.905
2	89.99%	0.917
4	90.55%	0.936
5	90.78%	0.941
6	91.10%	0.94
8	91.16%	0.945
10	91.19%	0.943
12	91.07%	0.942
15	91.13%	0.94
20	91.14%	0.939





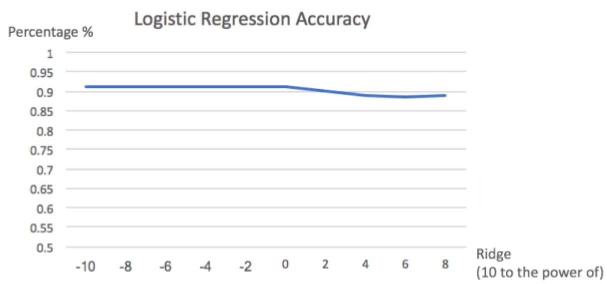
The accuracy and AUC doesn't change significantly. We choose maxdepth to be 8 where AUC is maximized to be 0.945 and there is an accuracy of 91.16%.

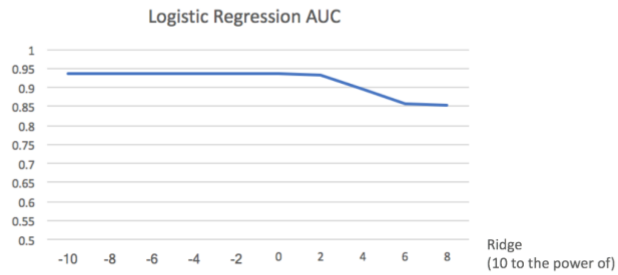
By choosing "ComputeAttributeImportance" to be true, we get the 10 most important features for this RF model: age, duration, housing, day_of_week, marital, pdays, education, month, job, previous.

Logistic Regression:

For logistic regression, we change ridge from 1^-10 to 1^-8 (increase complexity by decreasing ridge) and get the following result:

Logistic regression		
Ridge	Accuracy	AUC
1.00E-10	91.07%	0.935
1.00E-08	91.07%	0.935
1.00E-06	91.07%	0.935
1.00E-04	91.07%	0.935
1.00E-02	91.07%	0.935
1.00E+00	91.07%	0.935
1.00E+02	90.07%	0.933
1.00E+04	89%	0.895
1.00E+06	88.70%	0.856
1.00E+08	88.73%	0.852





We can see from the plot that the result remains the same from 1^-10 to 1 but slightly drop afterwards. Therefore, we would just pick ridge to be 1 for simplicity, with an accuracy of 91.07% and AUC of 0.935.

By comparing the three models above, Random Forest seems to be better. One thing we have to ask now is: Why are the results being so good? Are there any features being too predictive and may worth further detection? We then calculated the information gain in Weka and then get the 10 most important features: duration, euribor3m, cons.price.idx, cons.conf.idx, nr.employed, emp.var.rate, pdays, poutcome, month, previous.

Let's dig deeper into these repeatedly important features. There are five social and economic context features here (euribor3m, cons.price.idx, cons.conf.idx, nr.employed, emp.var.rate). Though we may ask at a first glance how might these features relate to customers' decision of subscription, we can actually think through it. With a higher bank offer rate, people are more willing to put their money in banks, which generate a high probability of subscription of term deposit. Also, economic factors like consumer price index and consumer confidence index would greatly influence people's decision of investment. Further, from the employment rate, we can see how much people's confidence in our bank fluctuates. What is interesting is the feature "Duration", which is also among the most important features of J48 decision tree model and Random Forest model. Let's check the definition of duration again: last contact duration, in seconds. You may think the relationship between this feature and our target variable could be: the longer the contact lasted, the more interested the customer might be and the higher probability he/she will subscribe. But wait a second, what if duration is zero? The result must be “No”—how could a customer subscribe without even being introduced to the product? Also, what’s more important, the duration is not known before a call is performed, but after the end of the call, the result (our target variable) is obviously known! Therefore, this feature contains information that we are going to predict—it is a leakage! This might explain why we get accuracy and AUC so high previously and we have to remove this feature to make our model accurate. (Actually we didn’t get to find this leakage after building our first three models. Hence, we think it might be better to show our thinking progress as we first figure out the most predictive features and then analyze what’s going on with them.)

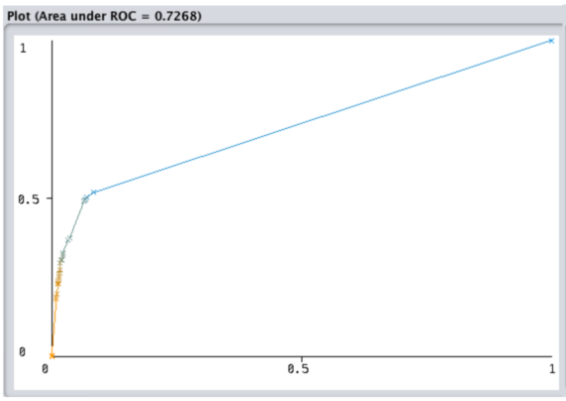
By removing the “Duration” feature, we go through all the steps above for each of the three classifiers. The results are shown in the table below:

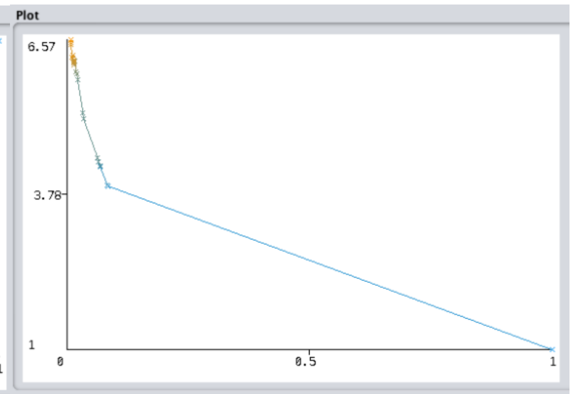
J48 Decision tree				Logistic regression				Random Forest		
MinNumObj	Accuracy	AUC		Ridge	Accuracy	AUC		maxdepth	Accuracy	AUC
1	89.93%	0.705		1.00E-10	89.99%	0.791		0	89.02%	0.767
2	90.00%	0.698		1.00E-08	89.99%	0.791		1	89.02%	0.767
5	90.05%	0.705		1.00E-06	89.99%	0.791		2	89.89%	0.784
10	90.08%	0.707		1.00E-04	89.99%	0.791		4	89.95%	0.795
20	90.09%	0.709		1.00E-02	89.99%	0.791		5	90.04%	0.798
50	90.07%	0.709		1.00E+00	90.01%	0.791		6	90.05%	0.797
75	90.04%	0.708		1.00E+02	89.99%	0.79		8	89.87%	0.792
100	90.07%	0.708		1.00E+04	88.91%	0.779		10	89.65%	0.785
200	90.13%	0.708		1.00E+06	88.73%	0.77		12	89.58%	0.776
500	89.80%	0.62		1.00E+08	88.73%	0.769		15	89.34%	0.77
								20	89.01%	0.769

We can see from the result that accuracy result doesn’t vary much for the three models, but there do exist some difference in AUC. The best model for J48 Decision tree is achieved when MinNumObj equals 20, with an accuracy of 90.09% and AUC of 0.709. Logistic regression generates higher AUC, with the best model being achieved when ridge equals 1. The corresponding accuracy is 90.01% and AUC is 0.791. For random forest, the results are even better—when max depth equals

5, the accuracy is 90.04% and the AUC is 0.798. Based on these “best” models for each classifier, we can use testing data to compare the performances of each selected model.

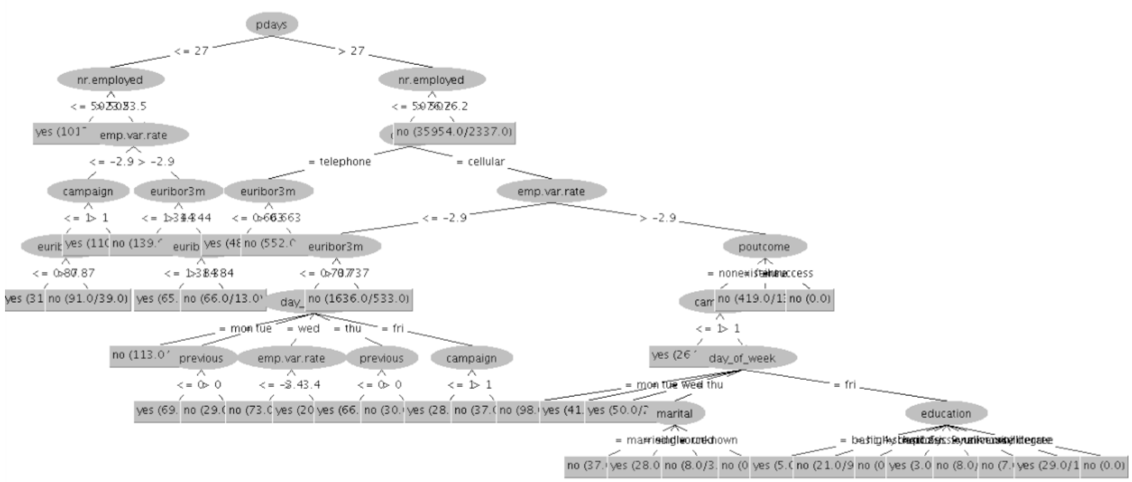
1. J48 Decision tree with MinNumObj = 20, we get a result of 90.6% accuracy and AUC equals 0.727.





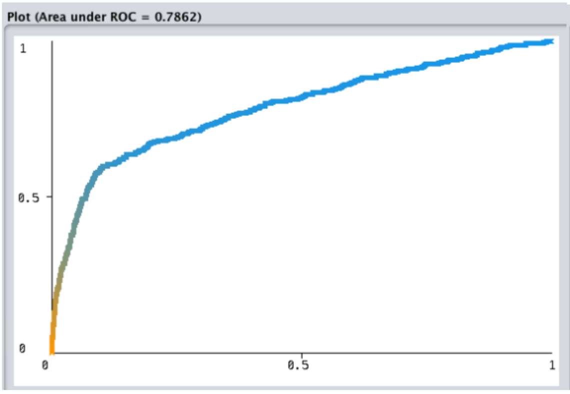
(ROC graph) (Lift curve)

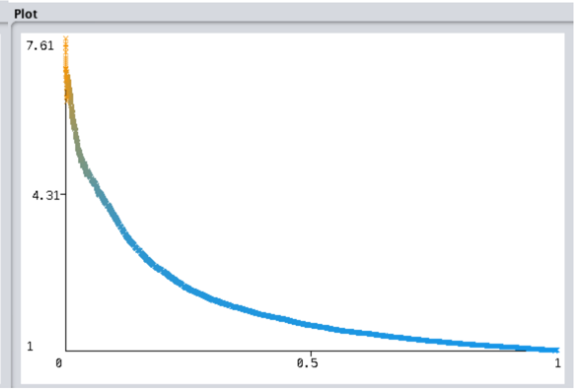
We can also visualize the tree structure:



The top features now becomes: pdays, nr.employed, emp.var.rate, euribor3m and campaign.

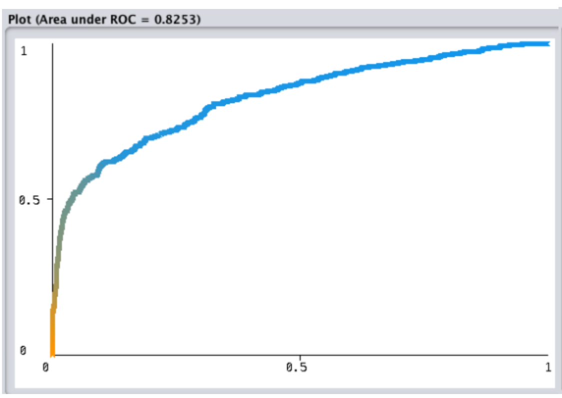
2. Logistic regression with ridges equals to 1, has an accuracy of 90.14% and AUC of 0.786.

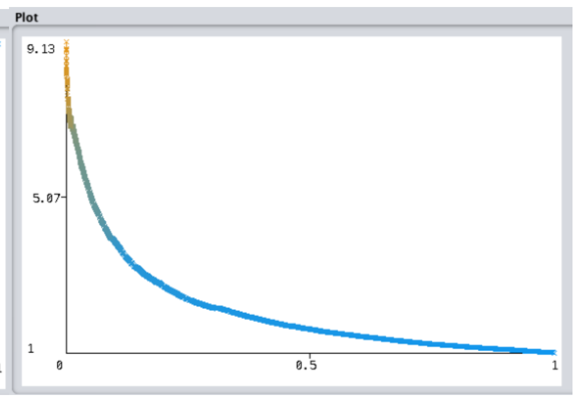




(ROC graph) (Lift curve)

3. Random forest with max depth equals to 5, has an accuracy of 91.04% and AUC of 0.825 (much better actually).





(ROC graph) (Lift curve)

From the results above, we can see that random forest classifier is doing much better than the other two classifiers, with an AUC even higher than the training data. However, we still need to look at the cost/benefit matrix to determine our final model.

Evaluation

We chose to evaluate our model based primarily on its expected value given a set of assumptions. We found a domain with very favorable characteristics for targeted marketing, given a very high multiple of benefit/cost and a very strong response rate. As a result the accuracy of the model was better captured by its true positive rate, as each successive true positive made a very strong impact to the bottom line. We found that each of the models provided an improvement on a Zero R baseline model except for the J48.

We attempted to base our cost/benefit matrix as much on domain specific knowledge as possible. As per the methodology of the model, we had a 2x2 matrix composed of classifying a potential call as “No” (won’t convert if called) and “Yes” (will convert if called). We conducted research into call centers in general, with the assumption that the costs for running a general call center were comparable to that of running one for a bank, and arrived at a $1 per min cost per call. (Booz & Co, p.4) We calculated the benefit by looking at the average net deposit value of a Portuguese consumer (net of any loans outstanding) and then applying a general profit margin of 1% which is the industry-wide average. (APB, p. 5) (Oliver Wyman, p. 2) This gave us an expected benefit of around $72. The Cost/Benefit matrix was then constructed with the assumption that only a False Positive and True Positive would incur any call center costs and the missed opportunity

General Cost/Benefit Matrix

		Classified As	
		No	Yes
Actual	No	$ -	$ (3.70)
	Yes	$ -	$ 63.47

would not be counted. This yielded the Cost/Benefit matrix below.

The overall results for the models are below. The baseline was a Zero-R model which would basically follow the most prevalent class. The baseline accuracy for this is 89%, as 89% of the instances are a "no". With this case there would be no profit, as no calls would have been placed.

Comparisons

Measures	Model: Zero-R (Baseline)	Model: J48	Model: Random Forest	Model: Logistic Regression
Accuracy	88.73%	90.09%	90.04%	90.01%
AUC	N/A	0.709	0.798	0.791
TP Rate	100.00%	24.59%	21.27%	22.87%
TP Rate (Threshold)		78.06%	85.02%	68.69%
Expected Value per customer	$ 3.87	$ 1.71	$ 1.48	$ 1.59
Expected Value per customer (Threshold)		$ 4.34	$ 4.41	$ 4.20
Total Expected Value (No threshold)		$ 70,267.23	$ 60,974.17	$ 65,352.59
Total Expected Value (Benefit Max Threshold)	$ 159,273.00	$ 178,876.41	$ 181,646.16	$ 172,853.66

This immediately explains the high accuracy ratings we found earlier, as the test set has a very high threshold for calling "no" correctly to begin with. The total expected value for the threshold maximized Zero-R was around $159,273, which is equivalent to calling every instance. This also provides a guide into the fundamental economics of the situation, as calling every single person of the data set provides a very strong profit. Even though only around 1 in 10 people successfully take up the product, the total profit from this whole group is nearly 3x the cost of reaching out to the other 9.

All of the models exhibited a strong improvement in expected profit from the whole program. All models were able to be optimized with a score threshold around 0.05, and were able to substantially improve upon both the baseline and a totally random sample. In a random targeting program, the program still generate on average $79k in profit, and the Zero-R was expected to generate $159k with targeting every person. All three programs made improvements on this scoring around $180k in total expected value.

These results demonstrate that the True Positive rate dominates the profit situation. Of all three models, the worst performing was the Logistic Regression model which only had a expected value of around $172k vs both other models being closer to around $180k. The best performing was the Random Forest model, which had an expected value of around $181k. And the difference in TP rate between the two is illustrative here. The LR model had only around a 69% true positive rate, vs the 85% rate of the RF model.

We also made sure to cross-validate our models with a testing data set we specifically kept as hold-out data. The hold-out data was of a similar make-up, with 11% yeses and 89% no's. Our summary data for the models tested on the hold-out data is below. All of the statistics were comparable indicating that our model does not suffer from over-fitting.

Tested Models Results

		Model		
		J48	Random Forest	Logistic Regression
Measures	Accuracy	90.75%	91.04%	90.14%
	AUC	0.804	0.825	0.786
	TP Rate	30.60%	25.90%	24.20%
	Score Threshold	10.12%	6.50%	5.88%
	Expected Value per customer (Threshold)	$ 2.68	$ 4.52	$ 4.14

As a next step, we'd take care to verify our assumptions with a domain-specific source of knowledge such as a VP of Marketing in a bank to confirm the cost assumptions and a VP of Finance to verify the value the product brings to the bank. In particular, we would try to flesh out the downside cases and the opportunity cost. As it is the models are heavily biased to the positive side because of the lack of penalties of a false negative. However, in the real banking world competition is fierce and our missing of the opportunity to sell a product might lead to a lost customer and the resulting decrease in relationship value between the bank and the customer as the customer brings a lot of her other financial products with them to the competitor. Unfortunately, we were unable to find the value of this with public information.

To summarise, we believe the results show our strategy of classifying the customers into two groups (likely to convert to a deposit product, and not likely to convert) has worked well when paired with a cost/benefit analysis. In particular, we think this approximates the real-world use of such a program. The biggest insight gained here was that the TP rate dominated the value here, as the value of a successful outweighed the cost of a false positive by a factor of around 6.

Deployment

Our plan of deployment is to apply the RF model to a continuously updated data-set based on a threshold to classify customers into the two buckets, and then to target the "yes" group in accordance with our resources.

We chose the RF model because it performed best on a profit and TP-rate basis, in accordance with our goals. We could also continue to run and update the other models, at least in a benchmarking capacity. We will definitely continue to run the Zero-R baseline to compare our efforts against. We think this baseline makes sense as the non-targeted marketing calls program is similar to calling everyone and hoping to capture positives that way. It's also an unusually strong targeting strategy as shown by the 11% positive population and great economics favoring this style, so it makes sense to compare against as it is likely to be a fall-back if our targeting failed. We will acquire new data by continuing to build upon the current set as the banks targeting program progresses.

Up until now we assumed that we would have an unlimited budget to pursue this targeting. However, we would consult with domain-experts to figure out a more realistic budget program. We would then establish a responsible threshold to run the program with, referring to the profit curves of each model (as a model other than the RF might perform better at a particular % of the population). We would recommend starting with the 5% threshold to begin with the RF model, but would also update this as the program continued. In week-to-week deployment, we would gather the data and analyze it in 15-day intervals. Every 15 days we would compute the probability a particular customer would respond to a targeted ad, classify them into "yes" and "no" groups with the threshold, and then target the "yes" group with calls.

We think this will best solve the problems of the business we hoped to solve with targeting. We also think we would be able to show the excess profit we were generating through targeting by comparing the results of our targeting to the data of the legacy program which had a cruder targeting program and lower TP rate. We would continue to place emphasis on presenting the TP rate and expected value generated from the program, as well as it's ROI, as these are likely to be the metrics most relevant to financial professionals who would be our stakeholders.

Conclusion

Our Random Forest model has provided a strong solution to both our data science problem and the underlying business problem. It is able to identify 85% of the positive classes and at the maximal threshold it improves the expected value per customer from $3.87 under a Zero-R model to $4.34 in a targeted case. It provides a more than 173% ROI which should give a strong impetus to managers above the DS level to implement the program. We think all of these results strongly indicate that targeting could provide a significant competitive edge for the bank in the very important area of developing customer relationships, and should be analyzed across more products and with contributions of more domain-specific knowledge to the assumptions.

Bibliography

Agarwal, Sumit, et al. “Benefits of Relationship Banking: Evidence from Consumer Credit Markets.” SSRN Electronic Journal, 2009, doi:10.2139/ssrn.1647019. Accessed at: http://finance.wharton.upenn.edu/~souleles/research/papers/BankRelationship8j_050509.pdf

Gupta, Amit, and Seamus McMahon. Redefining the Mission for Banks’ Call Centers Cut Costs, Grow Sales, or Both. 2008, Redefining the Mission for Banks’ Call Centers Cut Costs, Grow Sales, or Both, Accessed at: https://www.strategyand.pwc.com/media/file/Redefining_Mission_for_Banks_Call_Centers.pdf

Fine, Aaron, and Frank Rohde. DEPOSITS: A RETURN TO VALUE? 2013, DEPOSITS: A RETURN TO VALUE? Accessed at: http://www.oliverwyman.com/content/dam/oliver-wyman/global/en/files/insights/financial-services/2013/Dec/Deposits - A Return to Value.pdf

Pinho, Paulo Soares De. “Market Share Dynamics and Instruments of Competition in the Portuguese Deposits Market.” Strategic Challenges in European Banking, 2000, pp. 112–120., doi:10.1057/9780230377363_5

PORTUGUESE BANKING SECTOR OVERVIEW. Associo Potuguesa De Banco, Nov. 2015, www.apb.pt/content/files/Novembro_-_Overview_do_Sistema_Bancrio_Portugus_ING.pdf.

“UCI Machine Learning Repository: Bank Marketing Data Set.” UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set, archive.ics.uci.edu/ml/datasets/Bank Marketing#.