Term Project Paper

Decision 520Q: Data Science for Business
Dr. Natesh Pillai
October 13, 2017

Davide Sgarbi
Flora Yang
Yin-Ta Pan
Aaron Rodriguez

URL FOR DATA SET

http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

CONTENTS

BUSINESS UNDERSTANDING

Marketing is any action taken to promote and sell products or services. Historically, companies have focused on acquiring customers through traditional outbound techniques. These are the techniques that one would learn in any introductory marketing class: commercials, coupons in the mail, billboards, cold-calls, etc. Traditional outbound techniques are typically defined as any non-digital marketing technique where companies reach out to prospective customers directly. Today, many companies are focusing on a different approach called in-bound digital marketing. This is where a company focuses on reeling customers by making sure customers can easily find them in non-intrusive way. A company will do this through a combination of search engine optimization and generating content (blogs, videos, etc.) to attract consumers to the website[1]. This method of marketing, while generally considered to be effective, has problems when attempting to measure its effectiveness. It is difficult to attribute incremental sales increases to these marketing techniques because there is generally a time lag between the investment and the sales and because it is difficult to predict the sales in the counter-factual where the marketing campaign was not run[2]. Thus, many companies smartly maintain a mix between the new-age inbound digital marketing techniques and some traditional outbound methods. In this particular case, the team studied a Portuguese banking institution's direct marketing phone call campaigns.

[1] Gupta, Sunil and Joseph Davin. "Digital Marketing." *Harvard Business Publishing*, 31 July 2015.
[2] Gallo, Amy. "A Refresher on Marketing ROI." *Harvard Business Review*, 20 Sept. 2017, hbr.org/2017/07/a-refresher-on-marketing-roi.

For any company, being able to evaluate the effectiveness of a marketing campaign is crucial. Because of the set-up of the campaign, this is fairly easy to do. This campaign featured direct phone-calls to potential customers in hopes of them subscribing to a bank term deposit. The direct call to action removes the aforementioned problems with attribution; either the campaign worked and the customer set up an account or it did not. Each call comes at a cost, meaning that every time a call fails the company lost money. The ability to target only the customers most willing to set up an account would add two benefits: adding the revenue generated from the customers while cutting the costs wasted calling customers who do not have any desire for one. Thus, the team decided to evaluate three things: the customers with the highest probability of signing up through a marketing phone call campaign, the probability of a customer signing up required to have a positive expected value for the call, and other miscellaneous insights related to marketing campaigns that can be derived from the data.

Understanding The Data

The dataset comes from the University of California – Irvine, Machine Learning Repository. It has 41,188 rows and 21 variables. The target variable is binary: whether or not the client created a bank term deposit. The other twenty input variables related to the target variable are separated into four different categories: bank client data, data related to the last contact of the current campaign, other attributes of the campaign, and social and economic context data. Other than the binary output variable, the rest of the variables are all either categorical or numeric. The categorical variables will be treated as dummy variables, with one being excluded to avoid perfect multicollinearity. The numeric variables are continuous and will be treated as such.

Cleaning The Data

It was important to spend a significant amount of time understanding each variable and the relationships between them to start cleaning the data. Luckily, for the most part the data was relatively clean and featured minimal instances of variables with NULL, NA, or unknown values. There were 990 rows featuring unknown values for both housing and loan data that also had a “0” value for previously contacted, meaning that these are individuals who are being cold-called.

In terms of removing irrelevant columns, there were only three columns that had to be deleted or ignored in the making of the model. The first column was called “Pdays”, which was

supposed to measure the amount of days that had passed since the bank had previously called that customer. However, it had said 96% of clients had not been previously contacted and directly contradicted the column evaluating previous outcomes. Because the data was inconsistent, saying that customers had previous outcomes but had never been contacted, we decided to delete the column. Further, we also determined that the duration column, which measures the length of the call, would not be helpful in figuring out which customers are most likely to sign up for the account. This is because duration is only known after the call is performed, meaning that it cannot be used in predicting the likelihood. Thus, the variable was only used in analyses of cost and other miscellaneous evaluations of the marketing campaign, but not in the predictive model. Lastly, the number of employees in the company, a measure of economic context, was constant throughout this time-period and thus was deleted.

After removing and/or ignoring irrelevant columns, the team manipulated many of the personal categorical variables to attempt to generate more insights. First, we changed the marital status column into a binary variable, where married = 1 and anything else = 0. Our rational was that only 9% of the data was divorced or other and not all divorced people act in the same way. Someone who has been divorced for one year thinks and acts in a way that is different that someone who has been divorced for fifteen years (who probably acts more like a single person). Without being able to differentiate the two, we lumped them all together. For the education column, we simplified the data into University Education, Professional Education, and No Higher Education. Finally, we transformed the rest of the easier categorical variables into factor variables and began to do our analysis.

MODELING

EXPLORATORY DATA ANALYSIS

Before diving into the modeling process, the group did some exploratory data analysis to further understand the relationships between variables. The process began with a series of correlation matrixes. The first correlation matrix, in Plot1, featured the relationships between all variables featuring personal information and the dependent variable. From this, it became apparent that there are strong relationships between age and marital status, age and retirement, and education level and type of job (professional courses have blue collar jobs and university degrees tend to work in management). These relationships confirm logic and intuition. However, this plot also showed that few of the personal information variables have a strong

correlation with the target variable. However, age seems to have a parabolic relationship with y, according to Plot2. On the extremes of the age curve, people become more likely to subscribe to a term deposit, while middle-aged people are less likely. The group moved on to run a correlation matrix between all the variables regarding the last contact of the current campaign, in Plot3. It quickly became apparent that there are strong relationships between the type of contact, home versus cellular, and the month of previous contact. As we can see in Plot4, the bank tends to contact people's home phones more in May, June, and July and the performance during these months seem to be lower than average. These variables also have some correlation with the target variable, but because they are also correlated with themselves, it is hard to conclude whether it is the time of year or the device contacted that affects the outcome. Additionally, duration and the previous outcome both are strongly correlated with the outcome variable, which again confirms intuition. The group's final correlation matrix in Plot5 analyzed the economic indexes at the time. Each of these variables are highly correlated with each other and the outcome variable. Because of this, the variables are valuable but also must be used carefully.

By looking into the three types of variables separately, we need to check whether there is any correlation between the three. We applied principal component analysis to reduce the number of variables without losing important information. In Plot6, it suggests that we should take the top 10 factors from the information about our customers; take the top 10 factors from the information about the latest call; and take the top 2 factors from the economic indexes. Then we can further scrutinize whether there is any correlation between those principal component factors. As we can see in Plot7, there is no strong relationship between each factor, except the first component from latest call can the first component from economic indexes. Besides, the two components are related to the outcome variable. We can use the first component from latest call to indicate whether we should make a phone call to the clients, based on their previous performance. Also, we can use the first component from economic indexes as an indicator whether it is suitable for making a cold-call.

Creating the Models

In order to carry out a thorough analysis, we decided to create four different models on which to undertake a successive evaluation. Two of these models are relatively simple models, i.e. the logistic regression and the decision tree. These two models have been implemented

directly on the dataset, after cleaning up the data from irrelevant or highly correlated attributes and uninformative observations with unknown or NA values for different variables.

The third model we opted to introduce in our analysis is a Logistic model with Interactions (i.e. a logistic model in which every variable interacted with every other variable), and then regularized through Lasso by using the theoretical way to select the value of Lambda. We opted for the theoretical value as we were worried that a too-low value for Lambda (for instance, the value for which deviance would be minimized) would overfit our data, as research has been proven is often the case. Thus, after having let Lasso automatically select the relevant variables to include in our model, we performed an in-depth analysis of those variables and of the economic and intuitive relevance that each single one had into our model. We concluded that all the variables selected had reasons to be included, and that the model thus established appeared to be consistent. It's important to note here that we executed the variable selection through the Lasso model only on 10,000 observations, as greater amounts would cause R to crash and not produce any output; important implications and possible solutions of this issue are discussed below.

Finally, the last model we decided to include in our analysis is a PCA model. The reason why we decided to include such a model in our analysis is that we realized our attributes' dimension could be easily reduced to three main groups: customer data, previous contact information and macroeconomic data. This means that by grouping the variables belonging to each different cluster we identified, we could reduce the complexity of the model. Moreover, we realized that running a PCA would give us some key advantages such as low noise sensitivity, decreased requirements for capacity and memory (which turned out to be a main issue in our case, for example in the Lasso model), and increased efficiency (given that the processes take place in smaller dimensions).

EVALUATION

In order to evaluate our different models, we decided to employ a 10-fold cross-validation technique, according to the best industry practices. In this way, by training our models ten different times on ten different training sets and testing these models on holdout samples, we can get a true representation of whether our model are overfitting (or underfitting) the data. This is rendered possible by the inclusion of two relatively simple models (a logistic regression and a decision tree including only the variables present in the model) and two more complex models

(the logistic interaction model regularized via Lasso Theory and the PCA model). Finally, we chose Out of Sample R^2 as performance measure, as this showed to be the most effective measure in depicting the accuracy of different models.

By visualizing the results of the 10-fold cross-validation through a boxplot (Plot 8), it is easy to see that the PCA model outperforms all the other models both in terms of median OOS R^2 and in terms of lowest variance. We can also see that the Lasso model underperforms, and it even performs worse than the "simple" logistic model; finally, the tree model seems to be the one with lowest accuracy among the others.

It is important to note here that the low performance of the Lasso model is probably biased due to the necessity of selecting the relevant variables from the logistic interaction only by using 10,000 observations (roughly 25% of our data). This might have caused a biased selection of variables through regularization, and this may explain the underperformance when we try to use these variables when explain the bigger dataset through cross-validation. As noted above, any computation utilizing an amount of observations greater than that would result with R crashing and not producing any output. Availability of additional computational power would easily solve the problem.

Finally, the last step in our model evaluation is the fitting of the model on the whole dataset and the evaluation of its prediction power through the use of True Positive Rates and False Positive Rates for different thresholds of probability (i.e. the plotting of the ROC curve, Plot 9). Since our dataset does not present a balanced division between the success and failure of a call (i.e., the success rate among all the observation is only around 11%), the probabilities generated by our model will be affected in that direction too. Thus, we will have to set lower thresholds of probability in order to obtain the best performance metrics.

By looking at the ROC curve, we can see that the TPR increases steeply for the first thresholds of probability, then flattening for values of $p > .20$. At this threshold in particular, it seems that the tradeoff between having a high TPR and a low FPR is optimized; we would expect this value to be the "sweet spot" at which our profit would be maximized had we had to deploy the model on the basis of a ranking of probability (as we will discuss in the next section).

DEPLOYMENT

Deployment of the algorithm is simple: for each new set of customer information gained, a bank employee can apply the PCA formula to the dataset and call any customer with a

probability of depositing greater than 20%. The beauty of this method is its flexibility. The PCA model can be further trained with any additional data that comes in. Further, if the cost per call or the customer lifetime value changes over time, either due to inflation or other factors, the parameters in the PCA can be changed in order to obtain ever more accurate predictions of the profits derived from carrying out an additional campaign.

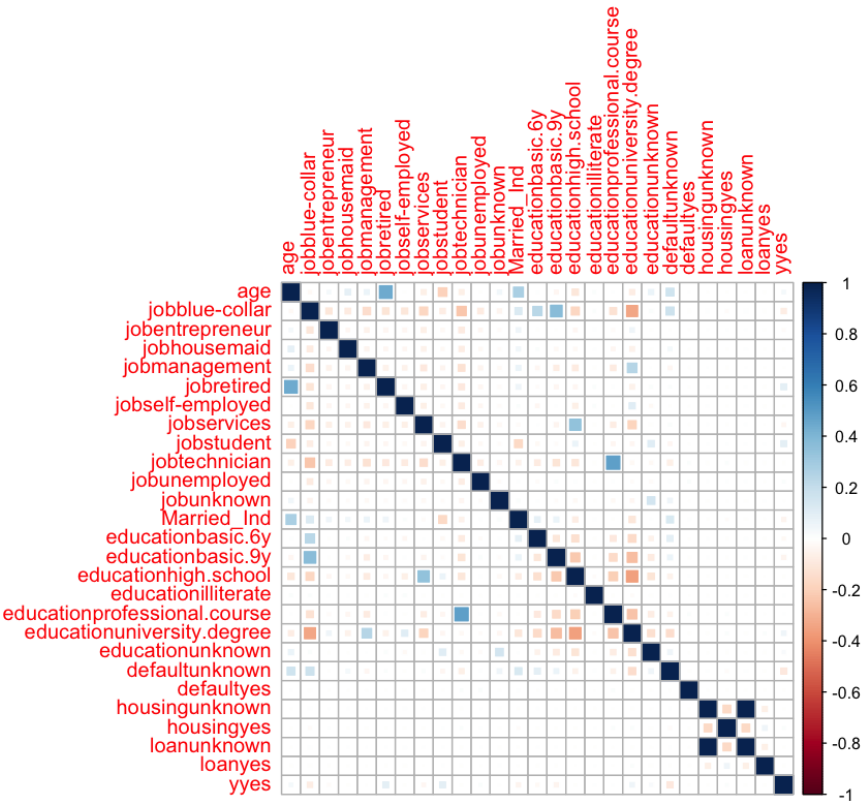
The key in the deployment of this model is patience. The first reason for patience is that the overall probability for a customer depositing is low. The rate of success previously was roughly eleven percent; phrased in another way, eight out of nine calls ended in failure. The improvement by using this methodology is significant, as we define the relevant threshold of probability to be twenty percent. As a matter of fact, such a threshold would achieve roughly a .70 True Positive Rate, meaning that almost five out of seven calls will be successful by employing our best prediction model; this amounts to a huge gain in terms of saving time and money. Specifically, it would roughly take 30% of the time to run the campaign, thus improving productivity of the sales/marketing department; moreover, by contacting only those who are the most likely to accept to open a new Term Deposit, the labor and overhead costs incurred to deploy the campaign would hugely decrease, leading to a five-fold increase in profit, from a baseline value of $54,315 to a much higher $265,680 (according to our research of $130 profit generation from a new Term Deposit[3] and $15 cost-per-call and overhead costs[4] of generating the campaign and administering the call center). More specifically, Plot 10 depicts a full representation of the expected profit generated by a targeted campaign addressing, in order, the customers most likely to obtain a term deposit to those less likely. As we can see, the expected profit maximize at around 20%, implying that targeting only those potential customers could entail, again, huge savings in terms of time and money.

People are generally skeptical of using algorithms, even if it is demonstrated to be more effective than previous methods[5]. If a new marketing campaign begins and gets off to a slow start, it is important to be persistent and patient with the methodology. Furthermore, the algorithm will become more precise as it is trained with more data. As the saying goes, "patience is a virtue"; in this case, one that will be rewarded with both time and savings.

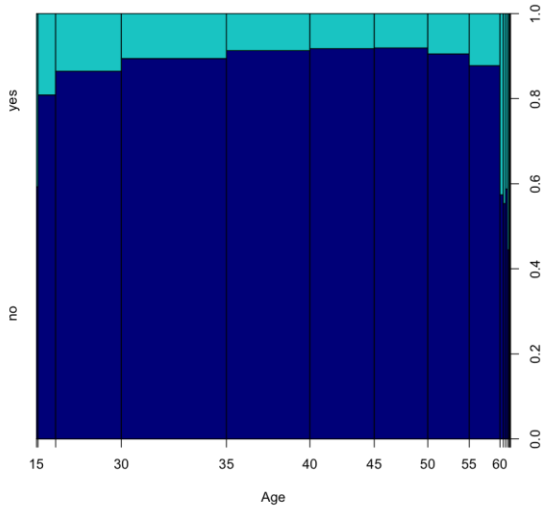
[3] https://www.columbiabank.com/resources/financial-calculators/business-marketing/customerlifetime-value
[4] https://www.worldwidecallcenters.com/call-center-pricing/
[5] Highhouse, Scott. "Stubborn Reliance on Intuition and Subjectivity in Employee Selection." *Industrial and Organization Psychology*, vol. 1, 2015, pp. 333–342.

APPENDIX

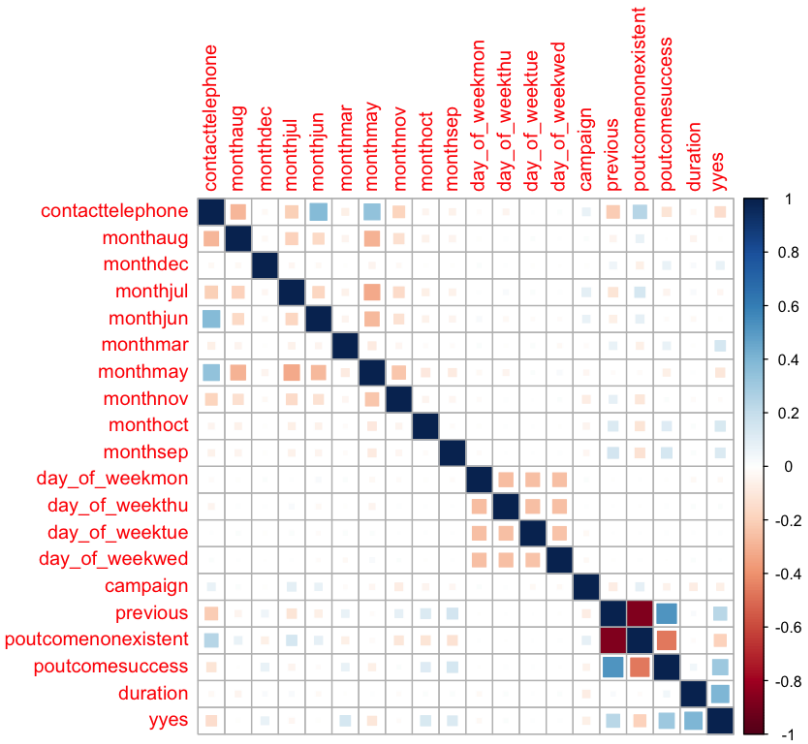
Plot1



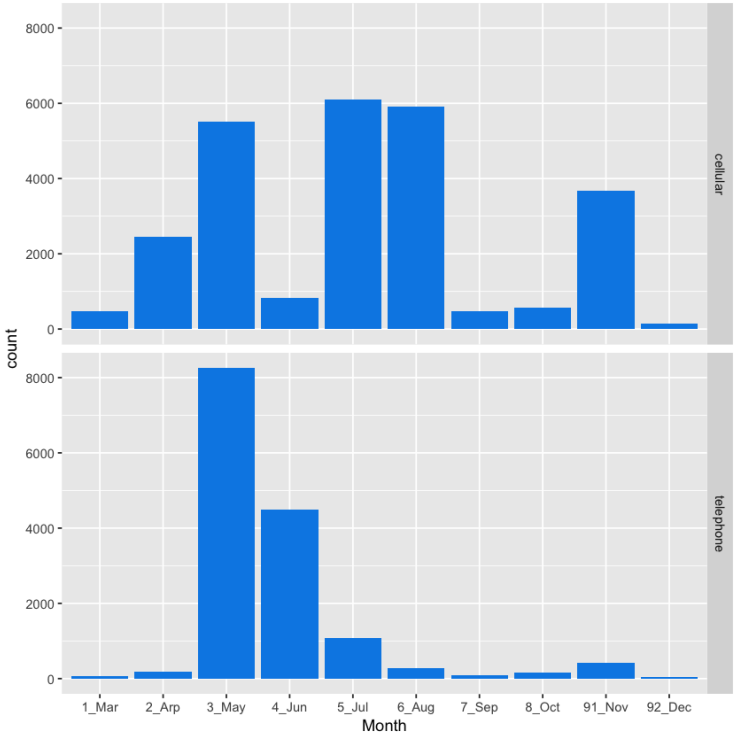
Plot2



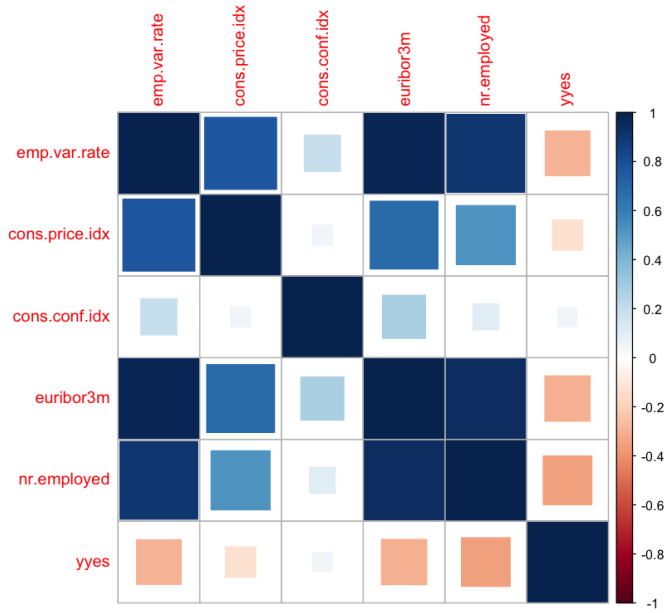
Plot3


contacttelephone
monthaug
monthdec
monthjul
monthjun
monthmar
monthmay
monthnov
monthoct
monthsep
day_of_weekmon
day_of_weekthu
day_of_weektue
day_of_weekwed
campaign
previous
poutcomenonexistent
poutcomesuccess
duration
yyes
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

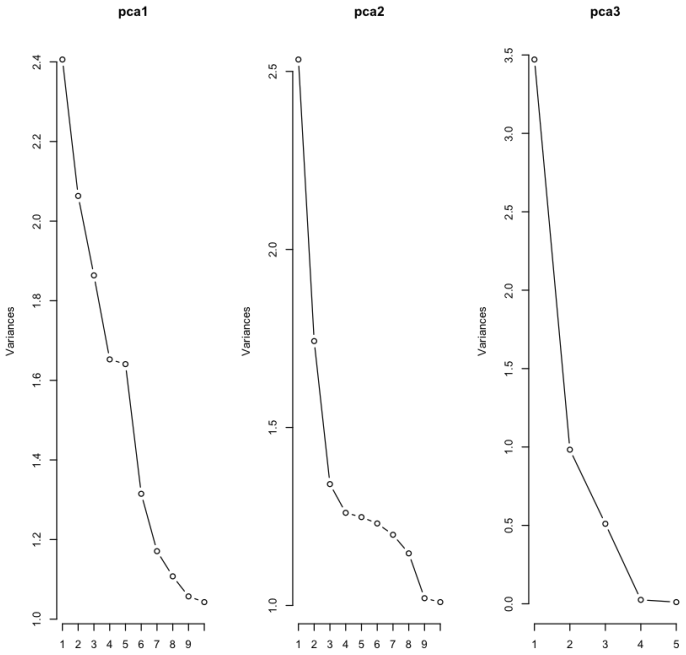
Plot4


8000
6000
4000
2000
0
count
cellular
telephone
1_Mar
2_Arp
3_May
4_Jun
5_Jul
6_Aug
7_Sep
8_Oct
91_Nov
92_Dec
Month

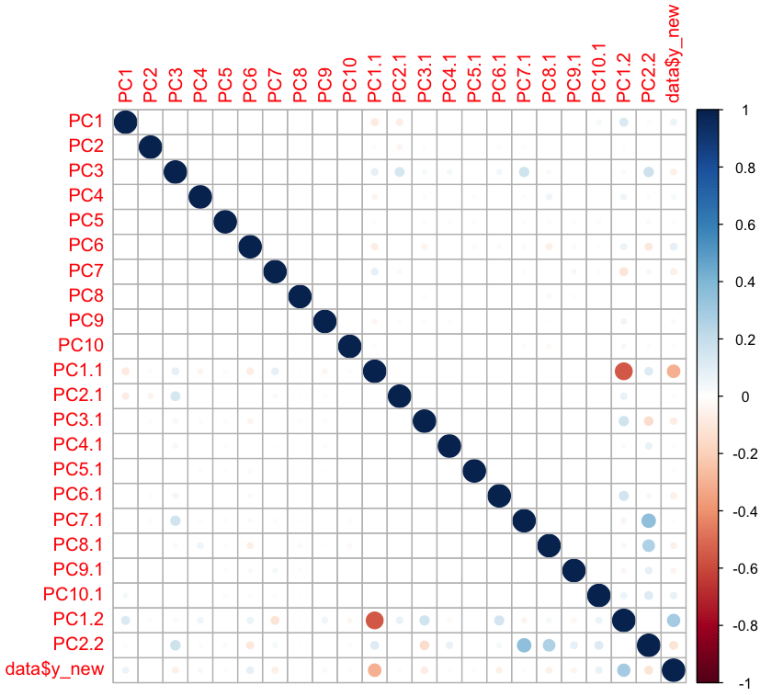
Plot5


emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed
yyes
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

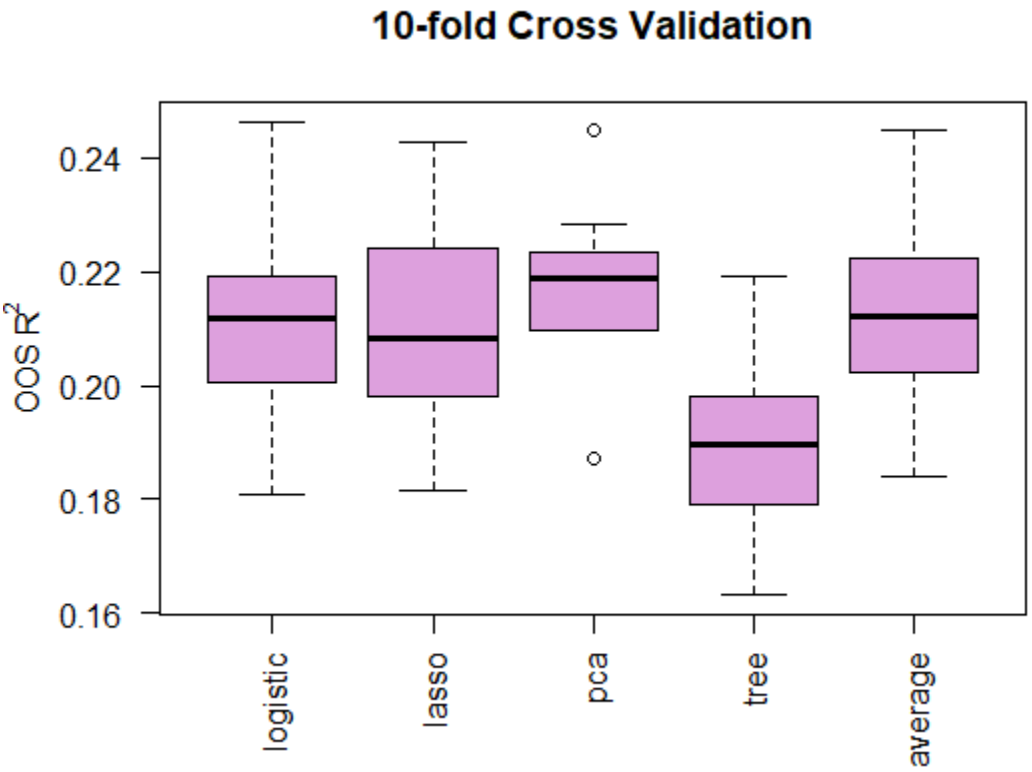
Plot6


pca1
pca2
pca3
Variances

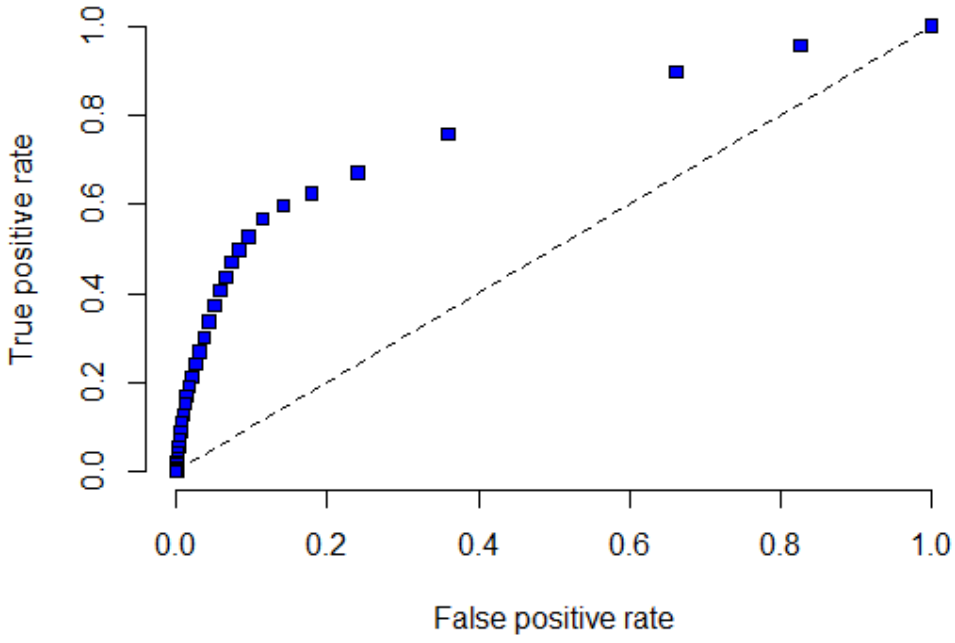
Plot7


PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
PC9
PC10
PC1.1
PC2.1
PC3.1
PC4.1
PC5.1
PC6.1
PC7.1
PC8.1
PC9.1
PC10.1
PC1.2
PC2.2
data$y_new
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

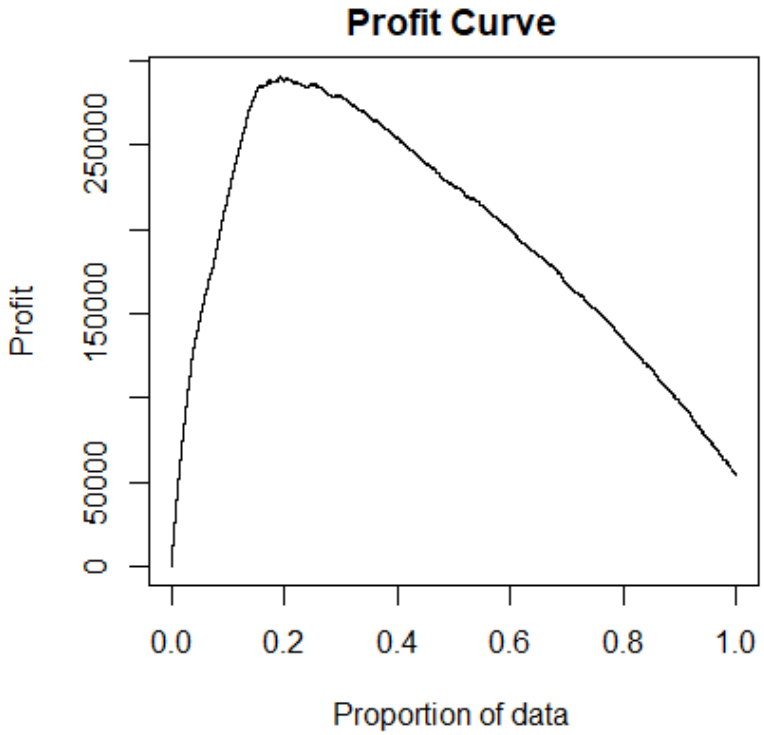
Plot8


10-fold Cross Validation
OOS R2
0.24
0.22
0.20
0.18
0.16
logistic
lasso
pca
tree
average

Plot9



Plot10



	Aaron Rodriguez	Flora Yang	Yin-Ta Pan	Davide Sgarbi
Business Understanding	x	x		x
Data Preparation	x	x	x	
Modeling		x	x	x
Evaluation	x		x	x
Deployment	x	x		x
R Script			x	x