Classification Methods

T. Evgeniou, INSEAD

What is this for?

A bank is interested in knowing which customers are likely to default on loan payments. The bank is also interested in knowing what characteristics of customers may explain their loan payment behavior. An advertiser is interested in choosing the set of customers or prospects who are most likely to respond to a direct mail campaign. The advertiser is also interested in knowing what characteristics of consumers are most likely to explain responsiveness to the campaign. A procurement manager is interested in knowing which orders will most likely be delayed, based on recent behavior of the suppliers. An investor is interested in knowing which assets are most likely to increase in value.

Classification (or categorization) techniques are useful to help answer such questions. They help predict the group membership (or class - hence called **classification techniques**) of individuals (data), for **predefined group memberships** (e.g. "success" vs "failure" for **binary** classification, the focus of this note), and also to describe which characteristics of individuals can predict their group membership. Examples of group memberships/classes could be: (1) loyal customers versus customers who will churn; (2) high price sensitive versus low price sensitive customers; (3) satisfied versus dissatisfied customers; (4) purchasers versus non-purchasers; (5) assets that increase in value versus not; (6) products that may be good recommendations to a customer versus not, etc. Characteristics that are useful in classifying individuals/data into predefined groups/classes could include for example (1) demographics; (2) psychographics; (3) past behavior; (4) attitudes towards specific products, (5) social network data, etc.

There are many techniques for solving classification problems: classification trees, logistic regression, discriminant analysis, neural networks, boosted trees, random forests, deep learning methods, nearest neighbors, support vector machines, etc, (e.g. see the R package "e1071" for more example methods). There are also many R packages for everything developed in the past - including the "fashionable" methods on deep learning - see various news here or here or here for example. Microsoft also has a large collection of methods they they develop. In this report, for simplicity we focus on the first two, although one can always use some of the other methods instead of the ones discussed here. The focus of this note is not do explain any specific ("black box, math") classification method, but to describe a process for classification independent of the method used (e.g. independent of the method selected in one of the steps in the process outlined below).

An important question when using classification methods is to assess the relative performance of all available methods/models i.e. in order to use the best one according to our criteria. To this purpose there are standard performance **classification assessment metrics**, which we discuss below - this is a key focus of this note.

Classification using an Example

The "Business Decision"

A boating company had become a victim of the crisis in the boating industry. The business problem of the "Boat" case study, although hypothetical, depicts very well the sort of business problems faced by many real companies in an increasingly data-intensive business environment. The management team was now exploring various growth options. Expanding further in some markets, in particular North America, was no longer something to consider for the distant future. It was becoming an immediate necessity.

The team believed that in order to develop a strategy for North America, they needed a better understanding of their current and potential customers in that market. They believed that they had to build more targeted boats for their most important segments there. To that purpose, the boating company had commissioned a project for that market. Being a data-friendly company, the decision was made to develop an understanding of their customers in a data-driven way.

The company would like to understand who would be the most likely customers to purchase a boat in the future or to recommend their brand, as well as what would be the **key purchase drivers** that affect people’s decision to purchase or recommend.

The Data

With the aid of a market research firm, the boating company gathered various data about the boating market in the US through interviews with almost 3,000 boat owners and intenders. The data consisted, among others, of 29 attitudes towards boating, which respondents indicated on a 5-point scale. They are listed below. Other types of information had been collected, such as demographics as well as information about the boats, such as the length of the boat they owned, how they used their boats, and the price of the boats.

After analyzing the survey data (using for example factor and cluster analysis), the company managers decided to only focus on a few purchase drivers which they thought were the most important ones. They decided to perform the classification and purchase drivers analysis using only the responses to the following questions:

1 “Q16_1_Is a brand that has been around for a long time”
2 “Q16_2_Has best in class customer service”
3 “Q16_3_Has a strong dealer network”
4 “Q16_4_Is a leader in cutting edge technology”
5 “Q16_5_Is a leader in safety”
6 “Q16_6_Is known for its innovative products”
7 “Q16_7_Is a brand for people who are serious about boating”
8 “Q16_8_Is a good brand for people that are new to boating”
9 “Q16_9_Is a brand I see in the water all the time”
10 “Q16_10_Offers boats that provide a fast and powerful boating experience”
11 “Q16_11_Offers the best boats for socializing”
12 “Q16_12_Offers the best boats for water sports e g tubing ski wakeboard” 13 “Q16_13_Offers boats with superior interior style”
14 “Q16_14_Offers boats with superior exterior style”
15 “Q16_15_Offers boats that stand out from the crowd”
16 “Q16_16_Offers boats that look cool”
17 “Q16_17_Offers boats that can handle rough weather or choppy water”
18 “Q16_18_Offers boats that can handle frequent and heavy usage”
19 “Q16_19_Offers a wide breadth of product offerings and accessories”
20 “Q16_20_Offers boats that I can move around safely”
21 “Q16_21_Offers boats that are easy to maintain and or repair”
22 “Q16_22_Offers boats that are easy to use”
23 “Q16_23_Offers boats that are easy to clean up”
24 “Q16_24_Has low prices”
25 “Q16_25_Is a brand that gives me peace of mind”
26 “Q16_26_Makes me feel I made a smart decision”
27 “Q16_27_Is a brand that impresses others”

Let’s get the data and see it for a few customers. This is how the first 50 out of the total of 2813 rows look:

We will see some descriptive statistics of the data later, when we get into the statistical analysis.

A Process for Classification

It is important to remember that Data Analytics Projects require a delicate balance between experimentation, intuition, but also following (once a while) a process to avoid getting fooled by randomness in data and finding “results and patterns” that are mainly driven by our own biases and not by the facts/data themselves.

There is *not a single best* process for classification. However, we have to start somewhere, so we will use the following process:

Classification in 6 steps

1. Create an estimation sample and two validation samples by splitting the data into three groups. Steps 2-5 below will then be performed only on the estimation and the first validation data. You should only do step 6 once on the second validation data, also called **test data**, and report/use the performance on that (second validation) data only to make final business decisions.
2. Set up the dependent variable (as a categorical 0-1 variable; multi-class classification is also feasible, and similar, but we do not explore it in this note).
3. Make a preliminary assessment of the relative importance of the explanatory variables using visualization tools and simple descriptive statistics.
4. Estimate the classification model using the estimation data, and interpret the results.
5. Assess the accuracy of classification in the first validation sample, possibly repeating steps 2-5 a few times in different ways to increase performance.
6. Finally, assess the accuracy of classification in the second validation sample. You should eventually use/report all relevant performance measures/plots on this second validation sample only.

Let's follow these steps.

Step 1: Split the data

It is very important that you finally measure and report (or expect to see from the data scientists working on the project) the performance of the models on **data that have not been used at all during the analysis, called "out-of-sample" or test data** (steps 2-5 above). The idea is that in practice we want our models to be used for predicting the class of observations/data we have not seen yet (e.g. "the future data"): although the performance of a classification method may be high in the data used to estimate the model parameters, it may be significantly poorer on data not used for parameter estimation, such as the **out-of-sample** (future) data in practice. The second validation data mimic such out-of-sample data, and the performance on this validation set is a better approximation of the performance one should expect in practice from the selected classification method. This is why we split the data into an estimation sample and two validation samples - using some kind of randomized splitting technique. The estimation data and the first validation data are used during steps 2-5 (with a few iterations of these steps), while the second validation data is only used once at the very end before making final business decisions based on the analysis. The split can be, for example, 80% estimation, 10% validation, and 10% test data, depending on the number of observations - for example, when there is a lot of data, you may only keep a few hundreds of them for the validation and test sets, and use the rest for estimation.

While setting up the estimation and validation samples, you should also check that the same proportion of data from each class, i.e. people who plan to purchase a boat versus not, are maintained in each sample, i.e., you should maintain the same balance of the dependent variable categories as in the overall dataset.

For simplicy, in this note we will not iterate steps 2-5. Again, this should **not** be done in practice, as we should usually iterate steps 2-5 a number of times using the first validation sample each time, and make our final assessment of the classification model using the test sample only once (ideally).

We typically call the three data samples as **estimation_data** (e.g. 80% of the data in our case), **validation_data** (e.g. the 10% of the data) and **test_data** (e.g. the remaining 10% of the data).

In our case we use for example 2250 observations in the estimation data, 281 in the validation data, and 282 in the test data.

Step 2: Choose dependent variable

First, make sure the dependent variable is set up as a categorical 0-1 variable. In this illustrative example, the "intent to recommend" and "intent to purchase" are 0-1 variables: we will use the later as our dependent variable but a similar analysis could be done for the former.

The data however may not be always readily available with a categorical dependent variable. Suppose a retail store wants to understand what discriminates consumers who are loyal versus those who are not. If they have data on the amount that customers spend in their store or the frequency of their purchases, they can create a categorical variable ("loyal vs not-loyal") by using a definition such as: "A loyal customer is one who spends more than X amount at the store and makes at least Y purchases a year". They can then code these loyal customers as "1" and the others as "0". They can choose the thresholds X and Y as they wish: a definition/decision that may have a big impact in the overall analysis. This decision can be the most crucial one of the whole data analysis: a wrong choice at this step may lead both to poor performance later as well as to no valuable insights. One should revisit the choice made at this step several times, iterating steps 2-3 and 2-5.

Carefully deciding what the dependent 0/1 variable is can be the most critical choice of a classification analysis. This decision typically depends on contextual knowledge and needs to be revisited multiple times throughout a data analytics project.

In our data the number of 0/1's in our estimation sample is as follows:

	Class 1	Class 0
# of Observations	1177	1073

while in the validation sample they are:

```
              Class 1   Class 0
```

# of Observations 100 181

Step 3: Simple Analysis

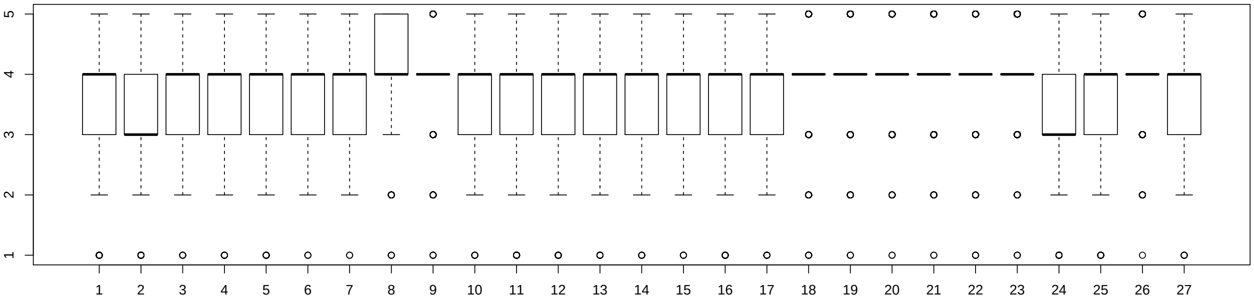
Good data analytics starts with good contextual knowledge as well as a simple statistical and visualization exploration of the data. In the case of classification, one can explore "simple classifications" by assessing how the classes differ along any of the independent variables. For example, these are the statistics of our independent variables across the two classes, class 1, "purchase" (first table), and class 0, "no purchase" (second table):

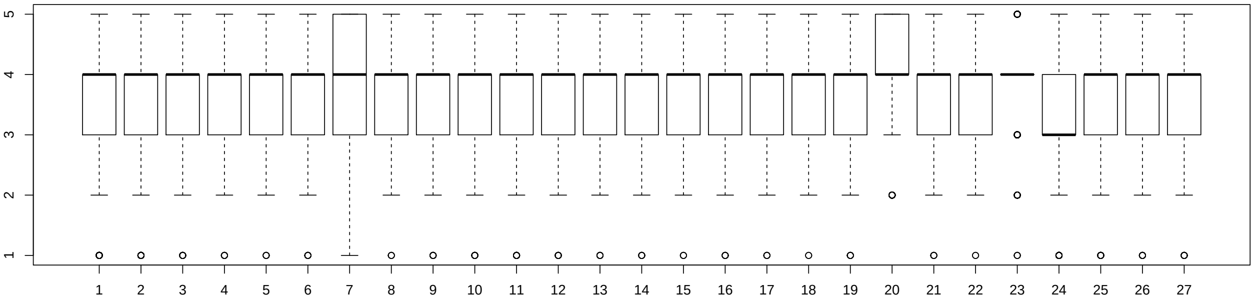
The purpose of such an analysis by class is to get an initial idea about whether the classes are indeed separable as well as to understant which of the independent variables have most discriminatory power. Can you see any differences across the two classes in the tables above?

Notice however that:

Even though each independent variable may not differ across classes, classification may still be feasible: a (linear or nonlinear) combination of independent variables may still be discriminatory.

A simple visualization tool to assess the discriminatory power of the independent variables are the **box plots**. These visually indicate simple summary statistics of an independent variable (e.g. mean, median, top and bottom quantiles, min, max, etc). For example consider the box plots for our data for the class 0 (top) and class 1 (bottom):





Can you see which variables appear to be the most discrimatory ones?

Step 4: Classification and Interpretation

Once we decide which dependent and independent variables to use (which can be revisited in later iterations), one can use a number of classification methods to develop a model that discriminates the different classes.

Some of the widely used classification methods are: classification and regression trees, boosted trees, support vector machines, neural networks, nearest neighbors, logistic regression, lasso, random forests, deep learning methods, etc.

In this note we will consider for simplicity only two classification methods: **logistic regression** and **classification and regression trees (CART)**. However, replacing them with other methods is relatively simple (although some knowledge of how these methods work is often necessary - see the R help command for the methods if needed). Understanding how these methods work is beyond the scope of this note - there are many references available online for all these classification methods.

CART is a widely used classification method largely because the estimated classification models are easy to interpret. This classification tool iteratively "splits" the data using the most discriminatory independent variable at each step, building a "tree" - as shown below - on the way. The CART methods **limit the size of the tree** using various statistical techniques in order to avoid **overfitting the data**. For example, using the rpart and rpart.control functions in R, we can limit the size of the tree by selecting the functions' **complexity control** paramater **cp** (what this does is beyond the scope of this note. For the rpart and rpart.control functions in R, smaller values, e.g. cp=0.001, lead to larger trees, as we will see next).

One of the biggest risks when developing classification models is overfitting: while it is always trivial to develop a model (e.g. a tree) that classifies any (estimation) dataset with no misclassification error at all, there is no guarantee that the quality of a classifier in out-of-sample data (e.g. in the validation data) will be close to that in the estimation data. Striking the right balance between "over-fitting" and "under-fitting" is one of the most important aspects in data analytics. While there are a number of statistical techniques to help us find this balance - including the use of validation data - it is largely a combination of good statistical analysis with qualitative criteria (e.g. regarding the interpretability or simplicity of the estimated models) that leads to classification models which can work well in practice.

Running a basic CART model with complexity control cp=0.01, leads to the following tree (**NOTE**: for better readability of the tree figures below, we will rename the independent variables as IV1 to IV27 when using CART):

The leaves of the tree indicate the number of estimation data observations that belong to each class which "reach that leaf" as well as the pecentage of all data points that reach the leaf. A perfect classification would only have data from one class in each of the tree leaves. However, such a perfect classification of the estimation data would most likely not be able to classify well out-of-sample data due to over-fitting of the estimation data.

One can estimate larger trees through changing the tree's **complexity control** parameter (in this case the rpart.control argument cp). For example, this is how the tree would look like if we set cp = 0.005

One can also use the percentage of data in each leaf of the tree to have an estimated probability that an observation (e.g. person) belongs to a given class. The **purity of the leaf** can indicate the probability an observation which "reaches that leaf" belongs to a class. In our case, the probability our validation data belong to class 1 (e.g. the customer is likely to purchase a boat) for the first few validation data observations, using the first CART above, is:

In practice we need to select the **probability threshold** above which we consider an observation as "class 1": this is an important choice that we will discuss below. First we discuss another method widely used, namely logistic regression.

Logistic Regression is a method similar to linear regression except that the dependent variable can be discrete (e.g. 0 or 1). **Linear** logistic regression estimates the coefficients of a linear model using the selected independent variables while optimizing a classification criterion. For example, this is the logistic regression parameters for our data:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.8	0.4	-2.1	0.0
Q16_1_Is.a.brand.that.has.been.around.for.a.long.time	0.0	0.0	0.4	0.7
Q16_2_Has.best.in.class.customer.service	-0.2	0.1	-3.4	0.0
Q16_3_Has.a.strong.dealer.network	0.0	0.1	0.2	0.9
Q16_4_Is.a.leader.in.cutting.edge.technology	-0.1	0.1	-0.9	0.3
Q16_5_Is.a.leader.in.safety	0.1	0.1	1.1	0.3
Q16_6_Is.known.for.its.innovative.products	-0.1	0.1	-1.3	0.2
Q16_7_Is.a.brand.for.people.who.are.serious.about.boating	0.0	0.1	-0.1	1.0
Q16_8_Is.a.good.brand.for.people.that.are.new.to.boating	0.1	0.1	1.8	0.1
Q16_9_Is.a.brand.I.see.in.the.water.all.the.time	0.1	0.1	2.1	0.0
Q16_10_Offers.boats.that.provide.a.fast.and.powerful.boating.experience	0.2	0.1	3.3	0.0
Q16_11_Offers.the.best.boats.for.socializing	-0.1	0.1	-2.0	0.0
Q16_12_Offers.the.best.boats.for.water.sports..e.g...tubing..ski..wakeboard.	0.2	0.1	3.2	0.0
Q16_13_Offers.boats.with.superior.interior.style	-0.2	0.1	-2.4	0.0
Q16_14_Offers.boats.with.superior.exterior.style	-0.1	0.1	-1.4	0.2
Q16_15_Offers.boats.that.stand.out.from.the.crowd	0.1	0.1	1.3	0.2
Q16_16_Offers.boats.that.look.cool	-0.3	0.1	-4.4	0.0
Q16_17_Offers.boats.that.can.handle.rough.weather.or.choppy.water	-0.1	0.1	-1.7	0.1
Q16_18_Offers.boats.that.can.handle.frequent.and.heavy.usage	-0.1	0.1	-1.0	0.3
Q16_19_Offers.a.wide.breadth.of.product.offerings.and.accessories	0.1	0.1	1.8	0.1
Q16_20_Offers.boats.that.I.can.move.around.safely	0.0	0.1	0.3	0.8
Q16_21_Offers.boats.that.are.easy.to.maintain.and.or.repair	0.2	0.1	2.2	0.0
Q16_22_Offers.boats.that.are.easy.to.use	0.1	0.1	0.9	0.4
Q16_23_Offers.boats.that.are.easy.to.clean.up	0.0	0.1	-0.5	0.6

	Estimate	Std. Error	z value	Pr(>|z|)
Q16_24_Has.low.prices	0.1	0.1	1.5	0.1
Q16_25_Is.a.brand.that.gives.me.peace.of.mind	-0.1	0.1	-1.1	0.3
Q16_26_Makes.me.feel.I.made.a.smart.decision	0.1	0.1	1.6	0.1
Q16_27_Is.a.brand.that.impresses.others	0.1	0.1	1.8	0.1

Given a set of independent variables, the output of the estimated logistic regression (the sum of the products of the independent variables with the corresponding regression coefficients) can be used to assess the probability an observation belongs to one of the classes. Specifically, the regression output can be transformed into a probability of belonging to, say, class 1 for each observation. In our case, the probability our validation data belong to class 1 (e.g. the customer is likely to purchase a boat) for the first few validation data observations, using the logistic regression above, is:

The default decision is to classify each observation in the group with the highest probability - but one can change this choice, as we discuss below.

Selecting the best subset of independent variables for logistic regression, a special case of the general problem of **feature selection**, is an iterative process where both the significance of the regression coefficients as well as the performance of the estimated logistic regression model on the first validation data are used as guidance. A number of variations are tested in practice, each leading to different performances, which we discuss next.

In our case, we can see the relative importance of the independent variables using the "variable.importance" of the CART trees (see help(rpart.object) in R) or the z-scores from the output of logistic regression. For easier visualization, we scale all values between -1 and 1 (the scaling is done for each method separately - note that CART does not provide the sign of the "coefficients"). From this table we can see the **key drivers** of the classification according to each of the methods we used here.

	CART 1	CART 2	Logistic Regr.
Q16_1_Is.a.brand.that.has.been.around.for.a.long.time	0.0000000	0.0110197	0.0909091
Q16_2_Has.best.in.class.customer.service	-1.0000000	-0.8994543	-0.7727273
Q16_3_Has.a.strong.dealer.network	0.1781076	0.3994461	0.0454545
Q16_4_Is.a.leader.in.cutting.edge.technology	-0.2729842	-0.7653571	-0.2045455
Q16_5_Is.a.leader.in.safety	0.1948052	0.6577938	0.2500000
Q16_6_Is.known.for.its.innovative.products	-0.1320648	-0.1039517	-0.2954545
Q16_7_Is.a.brand.for.people.who.are.serious.about.boating	0.0000000	-0.0545403	-0.0227273
Q16_8_Is.a.good.brand.for.people.that.are.new.to.boating	0.0723126	0.6581984	0.4090909
Q16_9_Is.a.brand.I.see.in.the.water.all.the.time	0.0000000	0.0056923	0.4772727
Q16_10_Offers.boats.that.provide.a.fast.and.powerful.boating.experience	0.0000000	0.4917861	0.7500000
Q16_11_Offers.the.best.boats.for.socializing	-0.3209647	-0.2301927	-0.4545455
Q16_12_Offers.the.best.boats.for.water.sports..e.g…tubing..ski..wakeboard.	0.1404807	0.7367238	0.7272727
Q16_13_Offers.boats.with.superior.interior.style	-0.1508388	-0.3358969	-0.5454545
Q16_14_Offers.boats.with.superior.exterior.style	-0.2136343	-0.2026567	-0.3181818
Q16_15_Offers.boats.that.stand.out.from.the.crowd	0.1340070	0.0961085	0.2954545
Q16_16_Offers.boats.that.look.cool	-0.5755178	-0.6404725	-1.0000000
Q16_17_Offers.boats.that.can.handle.rough.weather.or.choppy.water	-0.5041929	-0.8468072	-0.3863636
Q16_18_Offers.boats.that.can.handle.frequent.and.heavy.usage	-0.0994017	-0.4040648	-0.2272727
Q16_19_Offers.a.wide.breadth.of.product.offerings.and.accessories	0.2862924	0.2973946	0.4090909
Q16_20_Offers.boats.that.I.can.move.around.safely	0.2652031	0.2999890	0.0681818
Q16_21_Offers.boats.that.are.easy.to.maintain.and.or.repair	0.9530697	1.0000000	0.5000000
Q16_22_Offers.boats.that.are.easy.to.use	0.6291439	0.7394180	0.2045455
Q16_23_Offers.boats.that.are.easy.to.clean.up	-0.4365717	-0.4120430	-0.1136364
Q16_24_Has.low.prices	0.5189457	0.3721827	0.3409091
Q16_25_Is.a.brand.that.gives.me.peace.of.mind	-0.1706865	-0.9380984	-0.2500000
Q16_26_Makes.me.feel.I.made.a.smart.decision	0.3420844	0.2453395	0.3636364

	CART 1	CART 2	Logistic Regr.
Q16_27_Is.a.brand.that.impresses.others	0.0000000	0.7199355	0.4090909

In general it is not necessary for all methods to agree on the most important drivers: when there is “major” disagreement, particularly among models that have satisfactory performance as discussed next, we may need to reconsider the overall analysis, including the objective of the analysis as well as the data used, as the results may not be robust. **As always, interpreting and using the results of data analytics requires a balance between quantitative and qualitative analysis.**

Step 5: Validation accuracy

Using the predicted class probabilities of the validation data, as outlined above, we can generate four basic measures of classification performance. Before discussing them, note that given the probability an observation belongs to a class, **a reasonable class prediction choice is to predict the class that has the highest probability**. However, this does not need to be the only choice in practice.

Selecting the probability threshold based on which we predict the class of an observation is a decision the user needs to make. While in some cases a reasonable probability threshold is 50%, in other cases it may be 99.9% or 0.01%. Can you think of such cases?

For different choices of the probability threshold, one can measure a number of classification performance metrics, which are outlined next.

1. Hit ratio

This is simply the percentage of the observations that have been correctly classified (the predicted is the same as the actual class). We can just count the number of the (first) validation data correctly classified and divide this number with the total number of the (fist) validation data, using the two CART and the logistic regression above. These are as follows for the probability threshold 50% for the validation data:

```
Hit Ratio
```

First CART 60.14235 Second CART 55.87189 Logistic Regression 54.44840

while for the estimation data the hit rates are:

```
Hit Ratio
```

First CART 59.64444 Second CART 63.95556 Logistic Regression 59.73333

Why are the performances on the estimation and validation data different? How different can they possibly be? What does this diffference depend on? Is the Validation Data Hit Rate satisfactory? Which classifier should we use? What should be the benchmark against which to compare the hit rate?

A simple benchmark to compare the performance of a classification model against is the **Maximum Chance Criterion**. This measures the proportion of the class with the largest size. For our validation data the largest group is people who do not intent do purchase a boat: 181 out of 281 people). Clearly without doing any discriminant analysis, if we classified all individuals into the largest group, we could get a hit-rate of 64.41% - without doing any work. One should have a hit rate of at least as much as the the Maximum Chance Criterion rate, although as we discuss next there are more performance criteria to consider.

2. Confusion matrix

The confusion matrix shows for each class the number (or percentage) of the data that are correctly classified for that class. For example for the method above with the highest hit rate in the validation data (among logistic regression and the 2 CART models), the confusion matrix for the validation data is:

	Predicted 1	Predicted 0
Actual 1	68.0	32.0
Actual 0	55.8	44.2

Note that the percentages add up to 100% for each row: can you see why? Moreover, a “good” confusion matrix should have large diagonal values and small off-diagonal oens: you see why?

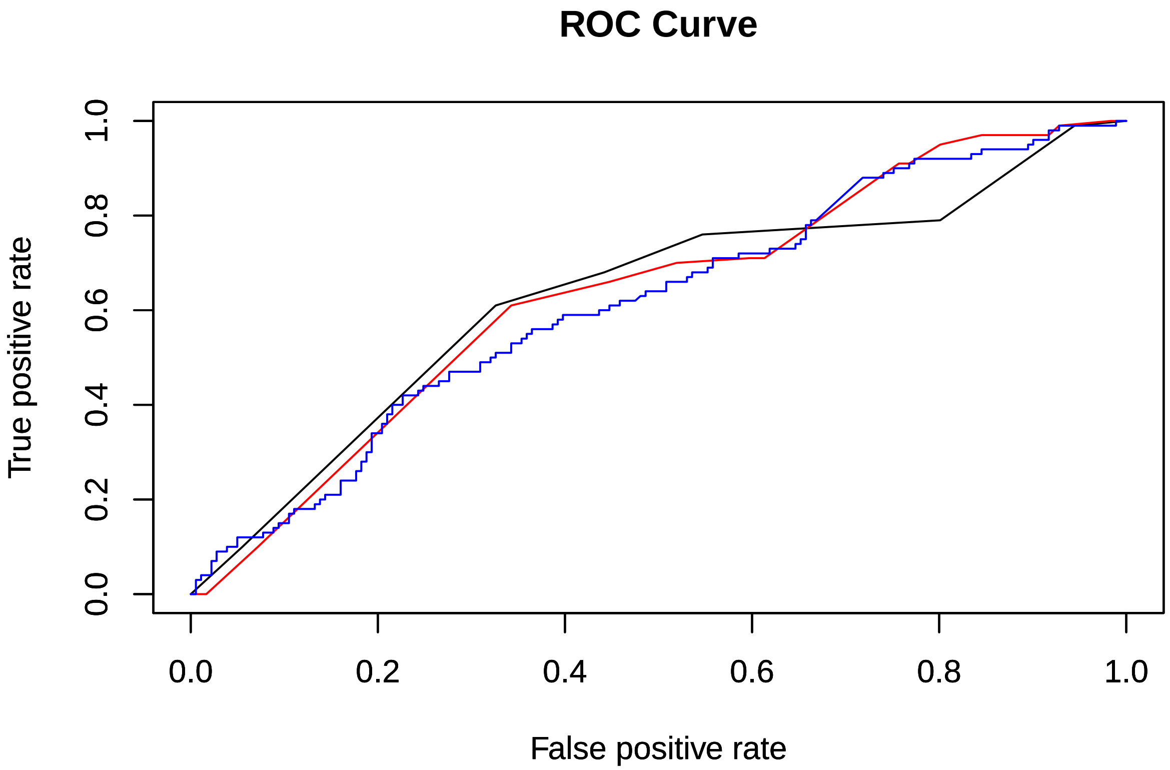
3. ROC curve

Remember that each observation is classified by our model according to the probabilities Pr(0) and Pr(1) and a chosen probability threshold. Typically we set the probability threshold to 0.5 - so that observations for which Pr(1) $>$ 0.5 are classified as 1’s. However, we can vary this threshold, for example if we are interested in correctly predicting all 1’s but do not mind missing some 0’s (and vice-versa) - can you think of such a scenario?

When we change the probability threshold we get different values of hit rate, false positive and false negative rates, or any other performance metric. We can plot for example how the false positive versus true posititive rates change as we alter the probability threshold, and generate the so called ROC curve.

The ROC curves for the validation data for both the CARTs above as well as the logistic regression are as follows:

which, if we plot all in the same graph for comparison, are (black: CART 1; red: CART 2; blue: logistic regression):



How should a good ROC curve look like? A rule of thumb in assessing ROC curves is that the “higher” the curve, hence the larger the area under the curve, the better. You may also select one point on the ROC curve (the “best one” for our purpose) and use that false positive/false negative performances (and corresponding threshold for P(0)) to assess your model. **Which point on the ROC should we select?**

4. Lift curve

By changing the probability threshold, we can also generate the so called lift curve, which is useful for certain applications e.g. in marketing or credit risk. For example, consider the case of capturing fraud by examining only a few transactions instead of every single one of them. In this case we may want to examine as few transactions as possible and capture the maximum number of frauds possible. We can measure the percentage of all frauds we capture if we only examine, say, x% of cases (the top x% in terms of Probability(fraud)). If we plot these points [percentage of class 1 captured vs percentage of all data examined] while we change the threshold, we get a curve that is called the **lift curve**.

The Lift curves for the validation data for our three classifiers are the following:

How should a good Lift Curve look like? Notice that if we were to randomly examine transactions, **the “random prediction” lift curve would be a 45 degrees straight diagonal line** (why?)! So the further **above** this 45 degrees line our Lift curve is, the better the “lift”. Moreover, much like for the ROC curve, one can select the probability threshold appropriately so that any point of the lift curve is selected. **Which point on the lift curve should we select in practice?**

5. Profit Curve

Finally, we can generate the so called profit curve, which we often use to make our final decisions. The intuition is as follows. Consider a direct marketing campaign, and suppose it costs $ 1 to send an advertisement, and the expected profit from a person who responds positively is $45. Suppose you have a database of 1 million people to whom you could potentially send the ads. What fraction of the 1 million people should you send ads (typical response rates are 0.05%)? To answer this type of questions we need to create the profit curve, which is generated by changing again the probability threshold for classifying observations: for each threshold value we can simply measure the total **Expected Profit** (or loss) we would generate. This is simply equal to:

Total Expected Profit = (% of 1's correctly predicted)x(value of capturing a 1) + (% of 0's correctly predicted)x(value of capturing a 0) + (% of 1's incorrectly predicted as 0)x(cost of missing a 1) + (% of 0's incorrectly predicted as 1)x(cost of missing a 0)

Calculating the expected profit requires we have an estimate of the 4 costs/values: value of capturing a 1 or a 0, and cost of misclassifying a 1 into a 0 or vice versa.

Given the values and costs of correct classifications and misclassifications, we can plot the total expected profit (or loss) as we change the probability threshold, much like how we generated the ROC and the Lift Curves. Here is the profit curve for our example if we consider the following business profit and loss for the correctly classified as well as the misclassified customers:

```
Predict 1   Predict 0
```

Actual 1 100 -75 Actual 0 -50 0

Based on these profit and cost estimates, the profit curves for the validation data for the three classifiers are:

We can then select the threshold that corresponds to the maximum expected profit (or minimum loss, if necessary).

Notice that for us to maximize expected profit we need to have the cost/profit for each of the 4 cases! This can be difficult to assess, hence typically some sensitivity analysis to our assumptions about the cost/profit needs to be done: for example, we can generate different profit curves (i.e. worst case, best case, average case scenarios) and see how much the best profit we get varies, and most important **how our selection of the classification model and of the probability threshold vary** as these are what we need to eventually decide.

Step 6: Test Accuracy

Having iterated steps 2-5 until we are satisfyed with the performance of our selected model on the validation data, in this step the performance analysis outlined in step 5 needs to be done with the test sample. This is the performance that "best mimics" what one should expect in practice upon deployment of the classification solution, **assuming (as always) that the data used for this performance analysis are representative of the situation in which the solution will be deployed.**

Let's see in our case how the **Confusion Matrix, ROC Curve, Lift Curve, and Profit Curve** look like for our test data:

Will the performance in the test data be similar to the performance in the validation data above? More important: should we expect the performance of our classification model to be close to that in our test data when we deploy the model in practice? Why or why not? What should we do if they are different?

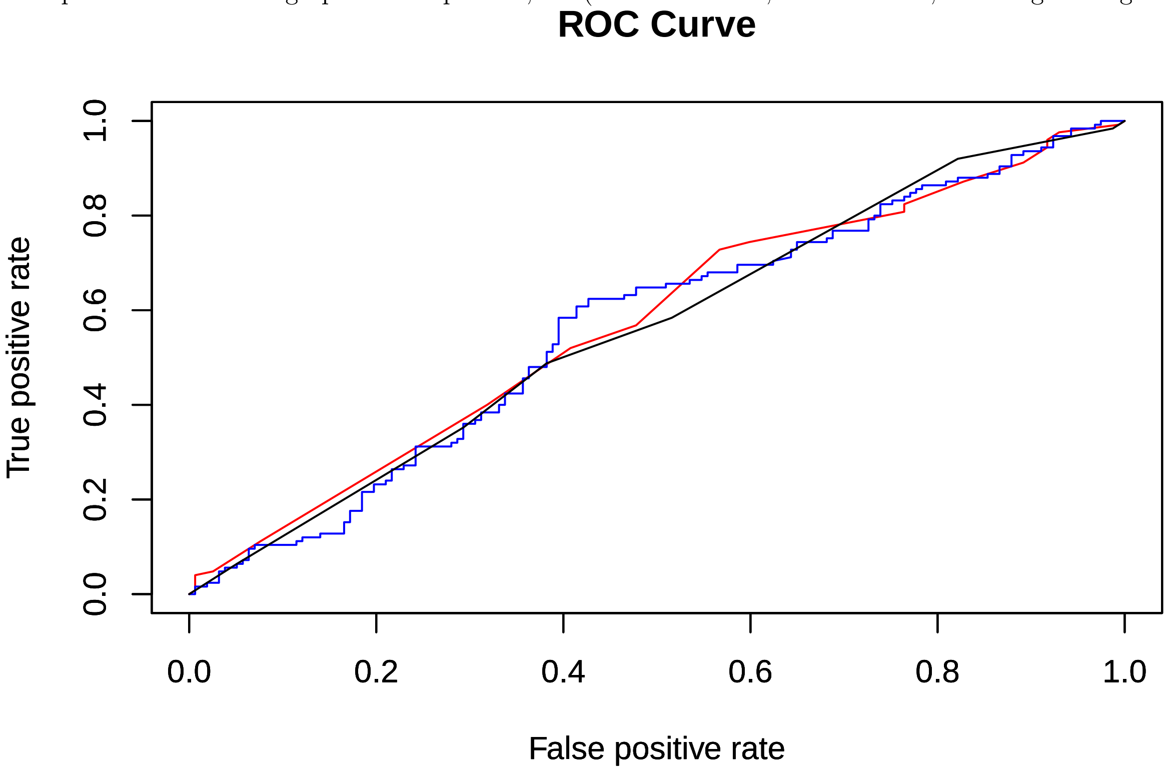
The Confusion Matrix for the model with the best validation data hit ratio above:

```
Predicted 1   Predicted 0
```

Actual 1 48.80 51.20 Actual 0 61.78 38.22

ROC curves for the test data

which, if we plot all in the same graph for comparison, are (black: CART 1; red: CART 2; blue: logistic regression):



Lift Curves for the test data:

Finally the profit curves for the test data, using the same profit/cost estimates as we did above:

What if we consider a segment-specific analysis?

Often our data (e.g. people) belong to different segments. In such cases, if we perform the classification analysis using all segments together we may not be able to find good quality models or strong classification drivers.

When we believe our observations belong in different segments, we should perform the classification and drivers analysis for each segment separately.

In this case, let's assume we found some customer segments based on earlier analysis. We consider two segmentation solutions:

First Segmentation: 5 Segments

Let's see first how many observations we have in each segment:

```
Segment 1   Segment 2   Segment 3   Segment 4   Segment 5
```

Number of Obs. 617 921 320 605 350

Using exactly the same analysis as above, let's see now the key drivers as well as the profit curve in the test data (only, for simplicity) when we perform the analysis for each segment separately (note: we only use segments with at least 100 datapoints). For simplicity we only show here the key drivers using only the logistic regression model for each segment.

	Segment 1	Segment 2	Segment 3	Segn
Q16_1_Is.a.brand.that.has.been.around.for.a.long.time	-0.0344828	0.4137931	0.4565217	-0.08
Q16_2_Has.best.in.class.customer.service	-0.4482759	-0.7931034	-0.1304348	-0.16
Q16_3_Has.a.strong.dealer.network	0.1034483	-0.1034483	0.1521739	-0.04
Q16_4_Is.a.leader.in.cutting.edge.technology	-0.5172414	-0.3448276	0.3043478	-0.04
Q16_5_Is.a.leader.in.safety	0.7241379	0.4137931	-0.4782609	-0.25
Q16_6_Is.known.for.its.innovative.products	-0.3103448	-0.1724138	0.0652174	0.25
Q16_7_Is.a.brand.for.people.who.are.serious.about.boating	-0.0689655	0.5172414	-0.0434783	-0.25
Q16_8_Is.a.good.brand.for.people.that.are.new.to.boating	0.1379310	0.1379310	0.1739130	0.29
Q16_9_Is.a.brand.I.see.in.the.water.all.the.time	0.4827586	0.1034483	0.4130435	0.58
Q16_10_Offers.boats.that.provide.a.fast.and.powerful.boating.experience	0.2068966	0.8965517	0.0652174	-0.04
Q16_11_Offers.the.best.boats.for.socializing	-0.2758621	-0.4482759	0.3478261	-0.58
Q16_12_Offers.the.best.boats.for.water.sports..e.g.…tubing..ski..wakeboard.	0.5517241	0.5862069	1.0000000	0.16
Q16_13_Offers.boats.with.superior.interior.style	-0.1379310	-0.3793103	-0.4565217	0.00
Q16_14_Offers.boats.with.superior.exterior.style	-0.3448276	-0.6206897	0.1521739	-0.08
Q16_15_Offers.boats.that.stand.out.from.the.crowd	0.5172414	-0.1379310	0.0652174	0.25
Q16_16_Offers.boats.that.look.cool	-0.3793103	-0.6206897	-0.2391304	-0.45
Q16_17_Offers.boats.that.can.handle.rough.weather.or.choppy.water	0.0000000	-0.2068966	-0.5434783	-0.41
Q16_18_Offers.boats.that.can.handle.frequent.and.heavy.usage	-0.4482759	0.2068966	-0.3695652	-0.29
Q16_19_Offers.a.wide.breadth.of.product.offerings.and.accessories	0.1379310	0.4137931	-0.0217391	1.00
Q16_20_Offers.boats.that.I.can.move.around.safely	0.0000000	-0.1724138	0.1304348	0.04
Q16_21_Offers.boats.that.are.easy.to.maintain.and.or.repair	0.0000000	1.0000000	-0.0434783	0.41
Q16_22_Offers.boats.that.are.easy.to.use	-0.2068966	-0.0689655	0.2826087	0.95
Q16_23_Offers.boats.that.are.easy.to.clean.up	-0.1379310	0.1724138	0.0652174	-0.79
Q16_24_Has.low.prices	0.3103448	0.8275862	-0.3478261	0.12
Q16_25_Is.a.brand.that.gives.me.peace.of.mind	0.2068966	-0.8275862	-0.2173913	-0.83
Q16_26_Makes.me.feel.I.made.a.smart.decision	0.6551724	0.3448276	-0.1304348	0.66
Q16_27_Is.a.brand.that.impresses.others	1.0000000	0.3448276	-0.0217391	0.33

while the profit curves are now:

Second Segmentation: 7 Segments

Let's see first how many observations we have in each segment:

```
Segment 1   Segment 2   Segment 3   Segment 4   Segment 5   Segment 6   Segment 7
```

Number of Obs. 365 921 201 252 119 605 350

Using exactly the same analysis as above, let's see now the key drivers as well as the profit curve in the test data (only, for simplicity) when we perform the analysis for each segment separately (note: we only use segments with at least 100 datapoints). For simplicity we only show here the key drivers using only the logistic regression model for each segment.

	Segment 1	Segment 2	Segment 3	Segn
Q16_2_Has.best.in.class.customer.service	-0.3214286	-0.7931034	-0.2727273	-0.29
Q16_3_Has.a.strong.dealer.network	0.0357143	-0.1034483	0.0606061	0.20
Q16_4_Is.a.leader.in.cutting.edge.technology	-0.5357143	-0.3448276	0.2121212	-0.08
Q16_5_Is.a.leader.in.safety	0.6785714	0.4137931	-0.4545455	0.58
Q16_6_Is.known.for.its.innovative.products	-0.3571429	-0.1724138	-0.2424242	-0.16
Q16_7_Is.a.brand.for.people.who.are.serious.about.boating	-0.0357143	0.5172414	0.0303030	-0.16
Q16_8_Is.a.good.brand.for.people.that.are.new.to.boating	0.3214286	0.1379310	0.0303030	-0.29
Q16_9_Is.a.brand.I.see.in.the.water.all.the.time	0.4285714	0.1034483	0.0606061	0.25
Q16_10_Offers.boats.that.provide.a.fast.and.powerful.boating.experience	0.1428571	0.8965517	-0.3030303	0.16
Q16_11_Offers.the.best.boats.for.socializing	-0.0357143	-0.4482759	0.6666667	-0.66
Q16_12_Offers.the.best.boats.for.water.sports..e.g... tubing..ski..wakeboard.	0.3571429	0.5862069	1.0000000	0.70
Q16_13_Offers.boats.with.superior.interior.style	-0.4285714	-0.3793103	-0.3333333	-0.12
Q16_14_Offers.boats.with.superior.exterior.style	-0.2500000	-0.6206897	0.0606061	-0.33
Q16_15_Offers.boats.that.stand.out.from.the.crowd	0.7857143	-0.1379310	0.4545455	-0.04
Q16_16_Offers.boats.that.look.cool	0.3928571	-0.6206897	-0.3333333	-1.00
Q16_17_Offers.boats.that.can.handle.rough.weather.or.choppy.water	0.1428571	-0.2068966	-0.7575758	-0.20
Q16_18_Offers.boats.that.can.handle.frequent.and.heavy.usage	-0.6785714	0.2068966	-0.4545455	-0.16
Q16_19_Offers.a.wide.breadth.of.product.offerings.and.accessories	-0.5357143	0.4137931	-0.2424242	0.66
Q16_20_Offers.boats.that.I.can.move.around.safely	0.2142857	-0.1724138	0.3333333	-0.50
Q16_21_Offers.boats.that.are.easy.to.maintain.and.or.repair	0.3928571	1.0000000	0.3030303	-0.62
Q16_22_Offers.boats.that.are.easy.to.use	-0.1785714	-0.0689655	0.0606061	0.25
Q16_23_Offers.boats.that.are.easy.to.clean.up	-0.3214286	0.1724138	-0.1818182	0.45
Q16_24_Has.low.prices	0.2500000	0.8275862	-0.5151515	0.08
Q16_25_Is.a.brand.that.gives.me.peace.of.mind	-0.3571429	-0.8275862	0.0606061	0.75
Q16_26_Makes.me.feel.I.made.a.smart.decision	0.5357143	0.3448276	-0.0606061	0.58
Q16_27_Is.a.brand.that.impresses.others	1.0000000	0.3448276	0.0303030	0.95

while the profit curves are now:

Does segment specific analysis help for our business decisions? Which solution should we use? Should we explore a different solution? Should we re-start from data collection, factor analysis, segmentation, or classification and drivers' analysis? How can we use the final results?

Of course, as always, remember that

Data Analytics is an iterative process, therefore we may need to return to our original raw data at any point and select new raw attributes as well as a different classification tool and model.

Till then...