Rotterdam School of Management
Erasmus University

Individual Assignment

Khoi Gia Pham 523755kg

Course: Big Data Management and Analytics (BM04BIM)

MSc Business Information Management

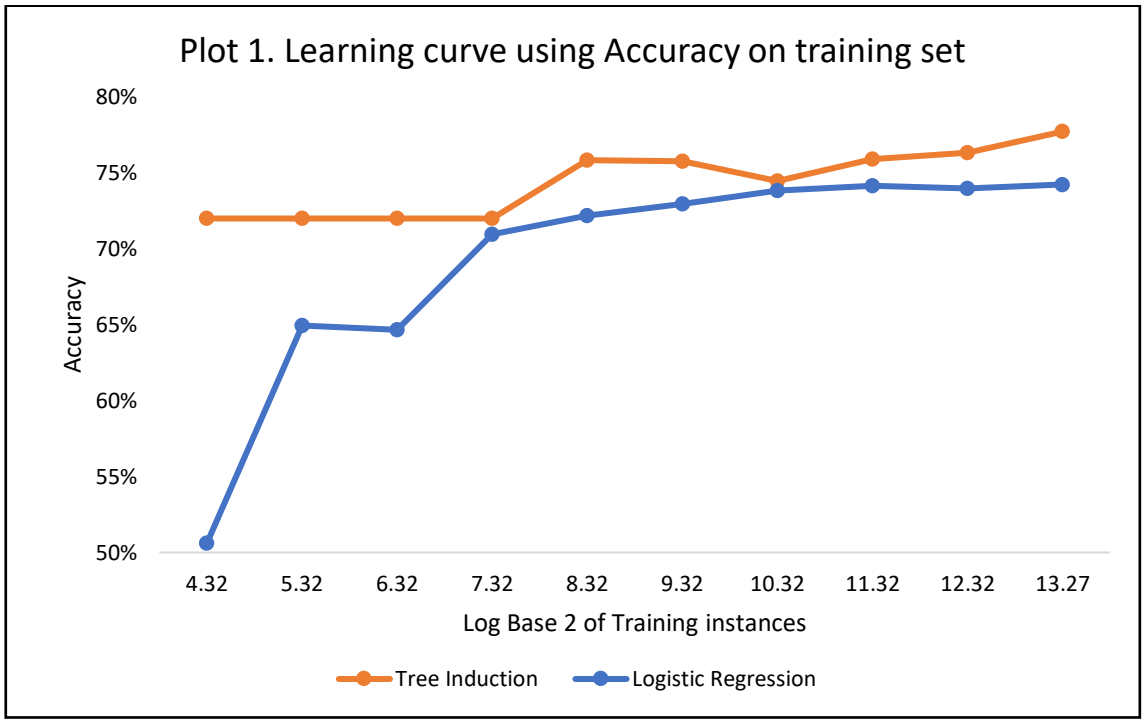
Rotterdam School of Management, Erasmus University

Exercise 1

Question 1

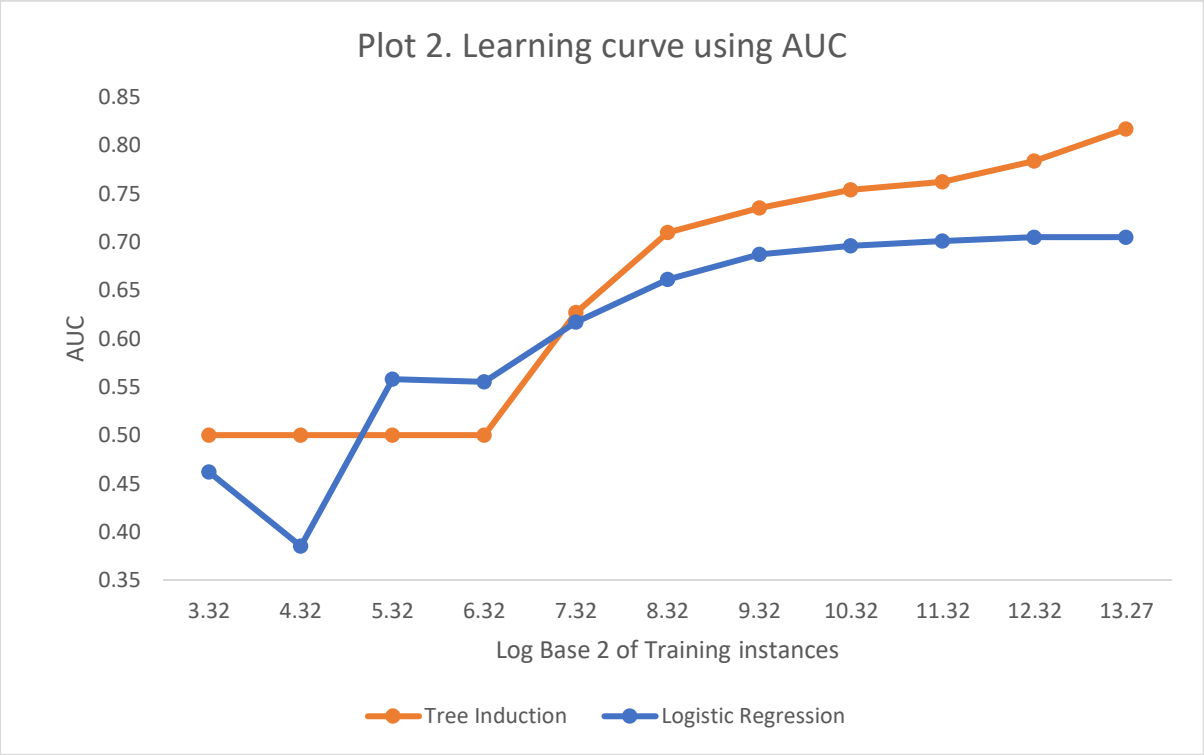
The two plots compare the performance on the training set between the Tree Induction and Logistic Regression models. The training data size is illustrated in log base two and is sampled with a random seed of 2001.

Plot 1



In the first plot, *Accuracy on the training set* is measured against the *Binary logarithm of Training instances*. Overall, more training instances improve the performance of both models. The improvements are significant in the early stage but become modest later on. When the training data is minimal (<100 instances), Induction tree accuracy stables at 71.98% since the model tends to overfit small-sized data sets. However, when the training data increases to around 200 to 1000 instances, Logistic Regression accuracy surges nearly to tree induction accuracy. For a larger training set (>1000 instances), the Tree induction model becomes more accurate due to its flexibility (complexity) characteristic.

Plot 2



The second plot measures *AUC* against the *Binary logarithm of Training instances.* Similarly, the learning curves follow that pattern in the first plot, with the same reasons mentioned above. Tree induction tends to perform better when the training set becomes larger large (>400 instances). However, a difference between the two plots is that Logistic Regression outperforms Tree induction when the dataset is relatively small (50-100 instances).

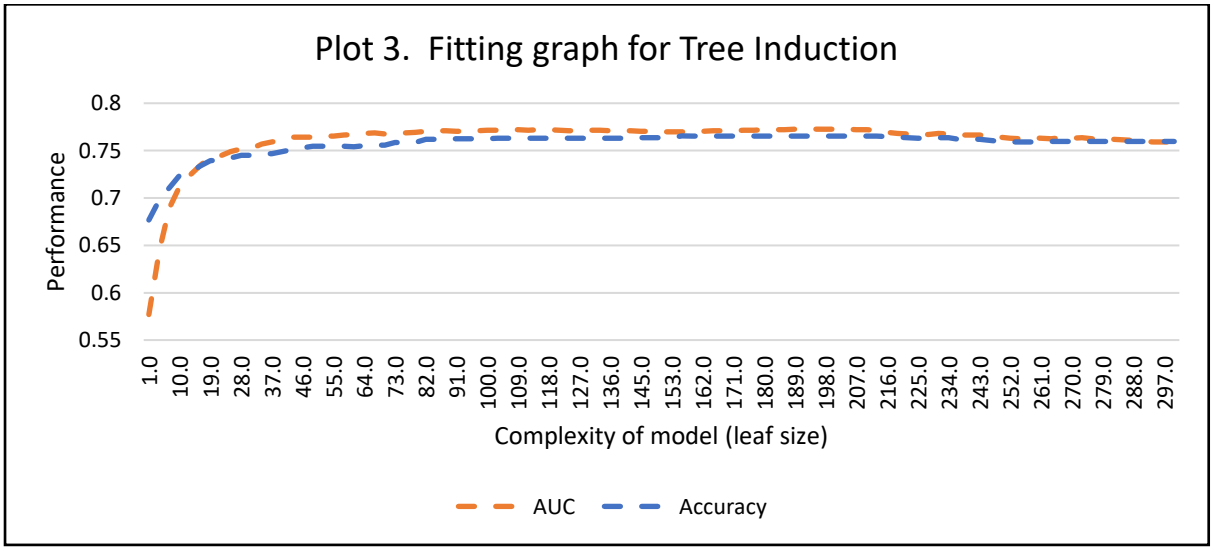
In brief, the performance (both accuracy and AUC) of both models increases as the number of training instances increases. In a smaller training set, Logistic regression performs better, while Tree induction tends to overfit the data. But as the training set becomes larger, Tree induction becomes more accurate due to its flexibility.

Question 2

A learning curve illustrates models' generalization performance as the training data grows. Ideally, the curve is abrupt at first and becomes flattered afterward. In our context, the improvements in the performance of the two models are flattering. Hence, collecting more data is of little value as its marginal advantage is becoming minor.

Exercise 2

Question 1



As illustrated from the plot, performance drastically increases when model complexity increases to 52 minimal leaf size. Afterward, the performance remains nearly unchanged, although the model gets more complex

The complexity of 198 minimal leaf size yields the highest pair of performance (AUC= 0.7726; Accuracy = 0.7654).

Overfitting occurs when the model becomes too complex and performs well on the training data but does not perform well on the testing data. Here, we used the Training data for both training and testing the model. And the resulting optimal performance is 77%. Hence, overfitting does not exist in this case.

Question 2

attribute	weight
overage	1
income	0.361886314
house	0.315443308
college	0.27518639
leftover	0.164706952
over_15mins_calls_per_month	0.071271895
average_call_duration	0.024719927
handset_price	0

Table 1. Variable Importance (based on Information Gain criterion)

In the table above, the attributes with higher weight are more relevant and important. And all the weights are normalized in a range from 0 to 1. As can be seen, *overage* is the most important attribute, followed by *income* and *house*. Their normalized weights are 1, 0.3619, and 0.3154, respectively.

Exercise 3

Question 1

From question 1, Tree Induction is selected as it outperforms Logistic Regression in almost all scenarios. From question 2, Tree Induction with a minimum leaf size of 198 is chosen as it yields the highest performance pair (AUC= 0.7726; Accuracy = 0.7654).

Question 2

Assumptions:

1. The Expected Value is measured within the one-year period (6 months with discounted price and 6 months with regular price).
2. Actual churners who receive the offer and end up not churning anymore will stay for at least one year. In other words, if he accepts the offer, he will not churn at any time during those 6 months with discount.
3. *"The offer has an effectiveness of 68%, i.e., 68% of the actual churners who receive the offer end up not churning anymore"*. This 68% is used to calculate the expected revenue of True Positive cases (*actual churners who receive the offer end up not churning anymore).*
4. *"Once a customer is targeted, s/he automatically gets six months with 30% discount without the need of signing a new contract"*. This means that 30*6=180 euros are realized as the targeting cost when the customer is targeted. And this cost remains unchanged even if the customers churn within the period with discount.

Question 3

Base case scenario: Not sending out any discount offer

True Positive: Benefit from customers that I target and is an actual churner but ends up not churning

True Negatives: Benefit from customers that I do not target and is not an actual churner

False Positives: Loss from customers that I target and is not an actual churner.

False Negatives: Loss from customers that I do not target and is an actual churner.

	Cost	**Benefit**	**Profit**
True Positive	*30% discount for 6 months* 30*6= 180	*68% chance with revenue for 12 months* *12*0.68*100=816	**636**
True Negatives	0	*We earn revenue as usual* 0	**0**
False Positives	*30% discount for 6 months* 30*6= 180		**-180**
False Negatives	0	0	**0**

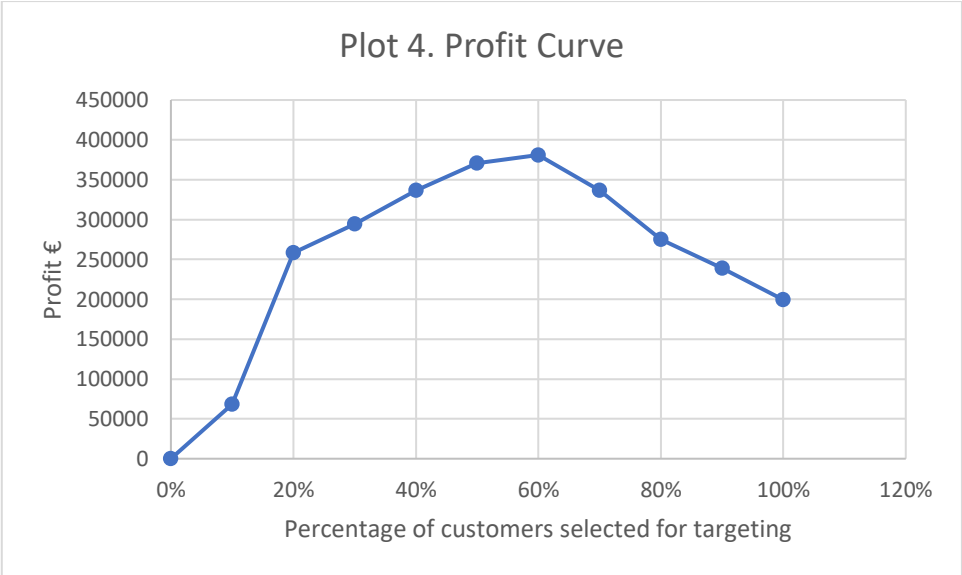
Table 2. The benefit-cost matrix

Question 4

Threshold	Customer Targeted	True Positives	False Positives	True Negatives	False Negatives	Revenues	Cost	Profit
0%	0	0	0	0	0	0	0	0
10%	423	177	246	2808	1001	112572	-44280	68292
20%	846	503	343	2711	675	319908	-61740	258168
30%	1270	641	629	2425	537	407676	-113220	294456
40%	1693	786	907	2147	392	499896	-163260	336636
50%	2116	921	1195	1859	257	585756	-215100	370656
60%	2539	1027	1512	1542	151	653172	-272160	381012
70%	2962	1066	1896	1158	112	677976	-341280	336696
80%	3386	1084	2302	752	94	689424	-414360	275064
90%	3809	1133	2676	378	45	720588	-481680	238908
100%	4232	1178	3054	0	0	749208	-549720	199488

Table 3. Expected Value table

Question 5



From the plot above, the threshold of 60% yields the highest profits of 381012 euros. Hence to maximize our profits, 2539 customers should be targeted. In more detail, if we target either less or more than 60% of the customers, the trade-off between revenue from the True Positive case and cost from the False Positive case is not optimal. However, it should be noted that the thresholds between 60%-70% were not considered. Thus, a more optimal threshold may exist within this range.