Big Data Management and Analytics
Individual Assignment

November 2, 2022

INSTRUCTIONS

- This is an individual assignment.
- Submit your answer digitally as a single **PDF file** through Canvas.
- We recommend you to use RapidMiner Studio (download from www.rapidminer.com) for this assignment. You can get familiar with Rapidminer by following the **Tutorial Videos page on Canvas** (Pages > RapidMiner Tutorial Videos).
- On Canvas, as an attachment to this assignment, you will find two data sets (`churn_hw2_train.csv` and `churn_hw2_test.csv`), corresponding to the churn prediction problem.
- Read this entire assignment before beginning. Follow the directions and answer all questions. Provide graphs/charts where appropriate. Your grade will reflect the informativeness and conciseness of your results and quality of presentation.

Deadline: Wednesday, November 16 at 23:59.

- Late submissions are not allowed

Warning: The detection of any form of plagiarism in your work means the assignment will be graded with ZERO points.

HINTS

1. Remember that your task is to present your analysis as good as you can - hence your tables and graphs should be easy to read for a person who is checking your work and should be presenting your work in the best way possible. The quality of execution matters!
2. Make use of formal, concise language and sufficiently elaborate concepts - your assignments as future work-related tasks require of you a standard of language formality. Regardless of your content, usage of casual language creates a bad impression to the reader that undermines your effort. Avoid ambiguous, blank, not scientific or logical statements that weaken your arguments, e.g. "The optimal complexity is X minimum leaf size, because it's a good number".
3. When we ask you to describe patterns in the Exercise, we want you to tell us something more than general - tree performs better/worse than logistic regression. Analyze the outputs of your analysis as thoroughly as possible.
4. Utilize knowledge from your classes/book as a support to your statements and properly cite them.

Targeting likely churners

Congratulations! You have been hired by Double Zero Data, a prestigious analytics consulting firm. Your first project is to develop and evaluate a prediction model for MegaTelco, a telecommunications firm. This is a context familiar to you as you have had previous experience with the churn problem from your BIM Master. The goal of MegaTelco is to identify likely churners so that they can target them with an attractive offer: a grace period of six months with all communications at half price, no strings attached.

Exercise 1

Design, execute, report and interpret on an evaluation to assess and compare the performance of two models – **tree induction** and **logistic regression** – as a function of the size of the training set.

Questions:

1. `[2 points]` You should produce a total of 2 learning curve plots to show how the two approaches (tree induction and logistic regression) compare for the various sizes of the data set:

   Plot 1: two learning curves on the same plot, one for tree induction and one for logistic regression, using **accuracy on the training set** as the performance measure. Use **only the training set** to both train and test your models.

   Plot 2: two learning curves on the same plot, one for tree induction and one for logistic regression, using the **Area under the ROC curve (AUC) on the training set**

as the performance measure. Use **only the training set** to both train and test your models.

Provide a **comprehensive interpretation** of each plot, and an interpretation of the differences between each of the plots. Which patterns do you observe? Which model performs better and why? Be concise and try to provide the fullest answer possible. Your answer should not be longer than 240 words.

2. `[1 point]` Describe the ideal learning curve plot (presented models, x-axis, y-axis) to assess whether it is worth to collect more data to improve our predictions. Justify your answer.

Detailed instructions:

You should start with 10 data points and double the number of observations at each step until you include the entire training data set. You can get various data sizes by using the "Sample" operator.

For tree induction, use the following parameters:

- set `criterion` to `information_gain`;
- set `maximal depth` to `20`;
- uncheck `apply pruning` (so that the tree is not automatically pruned);
- leave `apply prepruning` checked;
- set `minimal gain` to `0`;
- and force moderately large leaves: set `minimal leaf size` to `30`.

In your plots, in order to see the dynamics of the learning curve, **use a log scale on the axis representing the size of the data set.** (You should use your normal plotting environment, such as Excel, to make the plots.)

Hint: In RapidMiner, to use AUC as a measure of performance you need to use the operator "Performance (Binominal Classification)".

Exercise 2

For this exercise (and only for this exercise) use only the training data to calculate performance, and use cross validation. You can use the "Cross Validation" operator. You can use the minimum leaf size as a way to control model complexity.

Keep the same parameters as for Exercise 1:

- set `criterion` to `information_gain`;

- set `maximal depth` to `20`;
- uncheck `apply pruning` (so that the tree is not automatically pruned);
- leave `apply prepruning` checked;
- set `minimal gain` to `0`;
- and force moderately large leaves: set `minimal leaf size` to `30`.

Questions:

1. `[1 point]` Plot a fitting graph for the tree induction model. Use both generalization accuracy and AUC as performance measures.

   Provide a comprehensive interpretation of this plot. Which patterns do you observe? What is the optimal model complexity and why? Do you observe overfitting and why? Your answer should not be longer than 120 words.

2. `[1 point]` In order to answer this question, you need to change your operator "Decision Tree" to a "Random Forest" (as you will be using another operator that works only with a Random Forest model). A "Random Forest" model is equivalent to a decision tree if:

   - We set the parameter "number of trees" to 1;
   - We set the parameter "subset ratio" to 1.

   Using your best model in the previous question, create a table with the Variable Importance (based on information gain criterion). You can prepare it using the "Weight by Tree Importance" operator in RapidMiner.

   Provide an interpretation of the table with variable importance. Based on the table you created, which variables would you consider the most important and why?

Exercise 3

Use the Expected Value Framework to devise a plan to decide which customers to target with the offer. Use the best model you obtained in the previous questions.

You can use Excel to perform these calculations. You will need to export the predictions of the model (in the test set) and perform the calculations yourself in Excel. To export the predictions of a model you can use the "Write CSV" operator, which writes the predictions in a new file that you can open in Excel.

With all the predictions you can assign a cost to each type of misclassification and calculate the costs for each different threshold. Try to find the optimal threshold by calculating profits for different targeting thresholds and plotting a profit curve (in Excel).

Assume the following for your analysis:

- Once a customer is targeted, s/he automatically gets six months with 30% discount without the need of signing a new contract. There are no strings attached, so targeted customers can churn at any time after they get the offer without any penalty (to them).
  - The offer has an effectiveness of 68%, i.e., 68% of the actual churners who receive the offer end up not churning anymore;
  - Customers spend on average 100 euros per month; they would spend 70 euros per month for six months if targeted (due to the 30% discount) and go back to 100 euros after the six-month period (in case of not churning).

In case there is information missing (which is likely), make sensible assumptions and report them in question 2 below.

Questions:

1. `[1 point]` Which model have you chosen as the one that performs best? Justify your choice. Your answer should not be longer than 60 words.
2. `[1 point]` Think of assumptions you need to make in order to perform the EVF calculations. Enumerate each of them separately.
3. `[1 point]` Create a benefit-cost matrix, and justify the value in each cell.
4. `[1 point]` Report a **table with the necessary information to build a profit curve.** This table should have the following format:

Threshold (% of customers selected for targeting)	True Positives	False Positives	True Negatives	False Negatives	Revenues	Costs	Profits
0 customers - 0%	0	0	XXXX	XXXX	...	...	...
...	...	...	...	...	...	..	...
XXXX customers - 100%	...	...	0	0	...	...	...

5. `[1 point]` Create a profit curve based on the information in the table described above. How many customers should be targeted? Justify your answer.

Data Description

For this project, MegaTelCo gave us a historical data set of 14,088 customers, from which we have already created two data sets: a train set (`churn_hw2_train.csv`) with 9,856 customers and a test data set (`churn_hw2_test.csv`) with 4,232 customers. Use these data sets as your train and test sets, respectively. At the point of collecting the data, each customer either had stayed with the company or had left (churned). Each customer is described by the variables listed in the table below.

VARIABLE	DESCRIPTION
COLLEGE	Is the customer college educated?
INCOME	Annual income
OVERAGE	Average overcharges per month
LEFTOVER	Average number of leftover minutes per month
HOUSE	Estimated value of dwelling
HANDSET_PRICE	Cost of handset
LONG_CALLS_PER_MONTH	Average number of long calls – 15 mins or over – per month
AVERAGE_CALL_DURATION	Average duration of a call
REPORTED_SATISFACTION	Reported level of satisfaction
REPORTED_USAGE_LEVEL	Self-reported usage level
LEAVE (Target variable)	Did the customer stay or leave (churn)?