Assignment Business Analytics and Data Science

Prof. Stefan Lessmann & Johannes Haupt
WS 2017/18

Assignment

This assignment will allow you to apply your skills in business analytics on real-world data from the field of e-commerce and customer targeting and practice the scientific methods for rigorous testing and documentation. It will also determine your grade for the class *Business Analytics and Data Science*. The assignment consists of the applied, "hands-on" development of a prediction model for real-world data and the scientific documentation of your approach. For the first, you will apply the machine learning techniques studied in class by building a predictive model. For the second, you will document, explain and justify your methodology, experiments, and results in a term paper. The main part of term paper including relevant graphs and tables excluding the reference list should not exceed 16 pages. You are required to complete the task in a group of 3-4 students for which you must register on moodle.

For the assignment, you are highly encouraged to go beyond the standard methods taught in class as well as make use of the scientific literature and conduct and document your own experiments with the data. Make sure to consider all stages of a typical modeling process: - research the relevant technical and task-related knowledge in the literature - gather, clean and preprocess the relevant data - select the best model and model parameters - deploy and assess the model in terms of performance and plausibility with possibly revision of any step or the whole process.

To faciliate easy communication and work distribution within your group, we recommend Github (and its RStudio integration) for version control and Slack or a similar messenger for communication. We will discuss GitHub in class and there is a short setup guide on Moodle.

Timeline:
- November 1, 2017: Data and task description available on moodle
- December 1, 2017: Submission of individual predictions
- February 5, 2018: Submission of final group predictions
- February 12, 2018: Submission of group term paper

Individual assignment:
Please submit two files via the moodle system: your code (in a zipped folder if more than one file) and the prediction. Every student will have to submit a prediction to qualify for the term paper. Please make sure to submit one `.csv` file with the order identifier and your prediction: order_item_id; return. You can upload your submissions at any time *before the deadlines above*. The file name for the prediction must be in this format: 'matriculation number'_'last name'.csv, e.g. 551234_haupt.csv

Group term paper:
Please submit a total of three files via the moodle system: your code (in a zipped folder), prediction and the written report. Only one group member will have to submit for the group. For the prediction, please make sure to submit one `.csv` file with the order identifier and your prediction: order_item_id; return. You can upload your submissions at any time *before the deadlines above*.

Setting

Customers send back a substantial part of the products that they purchase online. Return shipping is expensive for online platforms and return orders are said to reach 50% for certain industries and products. Nevertheless, free or inexpensive return shipping has become a customer expectation and de-facto standard in the fierce online competition on clothing, but shops have indirect ways to influence customer purchase

behavior. For purchases where return seems likely, a shop could, for example, restrict payment options or display additional marketing communication.

For this assignment, you are provided with real-world data by an online retailer. Your task is to identify the items that are likely to be returned. When a customer is about to purchase a item, which is likely to be returned, the shops is planning to show a warning message. The warning message has been found in pre-tests to lead approx. 50% of customers to cancel the item. In case of a return, the shop calculates with shipping-related costs of 3 euros plus 10% of the item value in loss of resale value. Your task is to build a targeting model to balance potential sales and return risk in order to optimize shop revenue. The data you receive is artificially balanced (1:1 ratio between (non-)returns). Since the real data contains substantially more non-returns than returns, the misclassification costs include a correction factor of 5.

Data

You are provided with two data sets containing 13 variables, which will require creative data preparation to reach their potential. Data set `known` also includes information about one target variable (`return`) and should be used to build a predictive model. The target values for data set `class` are not provided and need to be predicted. Be aware that the data has not yet been pre-processed and will need some cleaning, so pay attention to variable types, missing values, and plausibility of values.

Model assessment

You are expected to provide a binary estimate (0/1) if the customer will return the item within the order. The performance of your prediction model will be evaluated by the net revenue gain. In this case, costs and gains are asymmetric. One one hand, warning a customer that an item might not suit their taste or size will save some costs if the customer chooses not to order the item. On the other hand, if a customer does not order an item that they would not have returned, then a higher revenue is lost.

		True value	
		item kept (0)	item returned (1)
Prediction	item kept (0) / no intervention	0	$0.5 \cdot 5 \cdot -(3 + 0.1 \cdot itemvalue)$
	item returned (1) / warning	$0.5 \cdot -itemvalue$	0

Table 1: Cost matrix for model assessment