Assignment Business Analytics and Data Science

Prof. Stefan Lessmann & Johannes Haupt
WS 2016/17

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 should not exceed 20 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, from gathering, cleaning and preprocessing relevant data over algorithm and model selection, to model deployment, assessment, and possibly revision.

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.

Timeline:
- October 26, 2016: Data and task description available on moodle
- February 5, 2017: Submission of prediction on unknown data
- February 12, 2017: Submission of 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: ID | return_customer. You can upload and also change your submissions at any time *before the deadlines above*.

Setting

Online customers often order in a specific online shop only once. One goal of customer relationship management (CRM) is to maximize customer lifetime value, in this case by incentivizing customers to return to the shop. A common method to do so are coupons sent to customers some time after an order. However, a coupon poses a cost of foregone profit to the retailer when it is used. In cases where a customer would have made a follow-up purchase even without the coupon incentive, the coupon value is effectively wasted. For this reason, rather than sending coupons to all customers, only specific, promising ones are targeted.

For this assignment, you are provided with real-world data by an online retailer. Your task is to identify the customers that can be expected to purchase again in the next 90 days based on customer characteristics, order conditions and ordered products. The predicted return customers will not receive a coupon. The shop estimates that sending a coupon to a customer, who does not plan to return, will convince her to place another order in 20% of cases with an average order value of 20 €. Your job is to maximize revenue by providing a list of promising customers to be targeted.

Data

You are provided with two data sets containing 37 predictive variables. Data set `known` also includes information about one target variable (`return_customer`) 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 naturally within 90 days following the original purchase. The performance of your prediction model will be evaluated by the net revenue gain. In this case, costs and gains are asymmetric. Sending a coupon to a customer that would have returned anyway, i.e. a false negative, entails an effective loss of 10 €. Not sending a coupon to a customer that does not plan to return foregoes an expected profit of 3 €. The resulting cost matrix is depicted in the table below.

		True value	
		non-repurchaser (0)	repurchaser (1)
Prediction	non-repurchaser (0) / coupon	3	-10
	repurchaser (1) / no coupon	0	0

Table 1: Cost matrix for model assessment