Technical report

Introduction

This project corresponds to the final task of the course Machine Learning with Python offered by the Institute for Continuing Education in Science. The team will take the role of data scientists for an exclusive Hotel in the Bahamas. The goal is to screen applying Russian clients. The reason is that Russian clients represent a contrasting customer segment. On one hand, they spend a high quantity of money during their stay in the hotel. On the other hand, they also cause a high cost on damages that is not recoverable for the Hotel.

The work was divided in five points, following the recommendation of the professor. The different parts included: prepare the data set, predict the projected revenue per clients, predict which clients will cause damage, predict the damage that those clients would cause, and to create a measure of the expected value for each potential customer so that client list can be selected.

Problem definition

For this work the objective is to build a model and decision rules that will maximize outcome for t to select a client list of 200 persons, based on a data set of 500 potential customers. The client list should represent a balance between damage and profit.

To be able to achieve this, different algorithms were used. To train them, a second data set was utilized that contained information over former guests. This data set included variables such as profit, damage (from visits that excluded the last one), age , gender, among others. In the next sections more details will be discussed over the different parts of the work.

Preprocessing

For the preprocessing, the first step was visualizing the data

After the visualization, the missing values were handled. To accomplish this, the data set for the scores of the potential clients and the training data set of former clients were fused. This was done so that the changes made on the training data set would be the same as in the other data set. Also, the outcomes were dropped as these should not be changed or touched before the fitting of the model.

After this, the categorical variables were chosen. These were client_segment, sect_empl, gender, retired, gold_status, prev_stay, divorce, married_cd.

After selecting the categorical variables, a mode imputation was carried out to replace the missing values. The categorical variables that had values other than 0and 1, were dummified.

The next step was to drop the original features and one dummy category for the categorical variables. In this point the feature "profit per night" (the result of the division of the profit amount and the nights booked) was created.

Then the features that had more than 25% of missing values were found. In here, there was a choice: they could be dropped or another way to handle them could be selected. As they represented the scores of the clients, it was decided that instead dropping them, a mean imputation (as they were quantitative data), was chosen. Mean imputations were also chosen for the rest of th missing data.

As the last steps of the preprocessing, the data was rescaled and the score and data_train sets were again separated.

Modeling part

Profit Model

Prediction of Damage (binary)

Prediction of Damage

Results

Speak over the customer list, and over how it was selected.

Improvements/modifications

Concluding remarks