Machine Learning Case Study

This case consists of a supervised learning example, similar to what we are working with on a daily basis in Klarna . Your task is to predict the probability of default for the datapoints where the `default` variable is not set. The answer should contain the resulting predictions in a csv file with two columns, `uuid` and `pd` (probability of `default==1`). Once done expose this model with an API Endpoint on a cloud provider of your choice. Bonus points if you use AWS. Send us the details on how to query the endpoint, attach code used for modelling, a short (max one page) explanation of your model and how you validated it.

We mostly use Python for modeling at Klarna but you are free to use other languages if you prefer as long as they are easily obtainable for us.

Don't spend too much time on the prediction results. We evaluate how you structure and reason about the problem rather than the predictive accuracy of your model.

Good luck!!

Dataset

The data is located in the attached file dataset.csv. This is a simple semicolon separated CSV file containing a unique id, the target variable `default` and a number of features with somewhat different datatypes and meanings. Missing values are denoted as `NA` in the set. Here is a list of the variables and their types:

Column Name	Column Type
uuid	text
default	categorical
account_amount_added_12_24m	numeric
account_days_in_dc_12_24m	numeric

account_days_in_rem_12_24m	numeric
account_days_in_term_12_24m	numeric
account_incoming_debt_vs_paid_0_24m	numeric
account_status	categorical
account_worst_status_0_3m	categorical
account_worst_status_12_24m	categorical
account_worst_status_3_6m	categorical
account_worst_status_6_12m	categorical
age	numeric
avg_payment_span_0_12m	numeric
avg_payment_span_0_3m	numeric
merchant_category	categorical
merchant_group	categorical
has_paid	boolean
max_paid_inv_0_12m	numeric
max_paid_inv_0_24m	numeric
name_in_email	categorical
num_active_div_by_paid_inv_0_12m	numeric
num_active_inv	numeric
num_arch_dc_0_12m	numeric
num_arch_dc_12_24m	numeric
num_arch_ok_0_12m	numeric
num_arch_ok_12_24m	numeric
num_arch_rem_0_12m	numeric
num_arch_written_off_0_12m	numeric
num_arch_written_off_12_24m	numeric
num_unpaid_bills	numeric
status_last_archived_0_24m	categorical
status_2nd_last_archived_0_24m	categorical
status_3rd_last_archived_0_24m	categorical
status_max_archived_0_6_months	categorical
status_max_archived_0_12_months	categorical
status_max_archived_0_24_months	categorical
recovery_debt	numeric
sum_capital_paid_account_0_12m	numeric

sum_capital_paid_account_12_24m	numeric
sum_paid_inv_0_12m	numeric
time_hours	numeric
worst_status_active_inv	categorical