Case Study - Machine Learning

This case consists of a supervised learning example, similar to what some ML teams are working with on a daily basis.

Your task is to predict the probability of default for the data points in the attached file where that variable is missing. Your solution should contain your predictions in a .csv file with two columns, *uuid* and *pd* (probability of *default==1*). Once done, you should expose this model with an API endpoint on a cloud provider of your choice. We (mostly) use Python for modeling so please use this programming language to craft your submission. There's a bonus point if you use AWS in your deployment.

When you're ready, send us your solution. Include the code, details on how to query the endpoint, and a 1-pager that summarizes how you built and validated your model.

Our advice is to avoid spending too much time optimizing your prediction *results*. We want to evaluate how you structure, reason and evaluate the problem. Showing off your skills in model building, analysis, and software engineering is more important than maximizing predictive performance. Similarly, further information about the variables below is **not** available, as your domain knowledge and skills in business translation will be assessed at a later stage. 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 `uuid`, the target variable `default` and a number of features with somewhat different data types. Missing values are denoted as `NA`. The features are as follows:

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