Credit Scoring
Individual Assignment
Due Date: July 20th 2023, 23:59

Description:

You have been hired as a data scientist by a financial institution which extends loans to subprime customers and your first assignment is to develop an application credit scorecard. A partially anonymized dataset is provided for both a (representative) sample of accepted applicants and a sample of rejected applicants. There are several steps and decisions to be made in developing a credit scorecard, which you must make independently and report upon allowing a colleague from the validation team to exactly understand and reproduce the model you developed. Therefore, it is crucial to report the process in detail and motivate the various choices you made. The credit scorecard should be presented and discussed with a thorough performance evaluation. The bank requires you to develop the best possible scorecard that **minimizes operational risk and associated cost**. They provide you the following metrics: Loss given default is defined as 75%, risk-free interest rate is defined as 1.5%.

Requirements:

To complete this project, you are required to implement a demonstration of your credit scorecard and its capabilities in a notebook format. Your work should follow best ML practices, including using a training, validation and test split and tuning hyperparameters.

Formal requirements include:

- A short presentation of no more than 10 minutes. The recording should be included in the submission files along with your slides.
- Document your model development, evaluation, and testing in a notebook file. Please use comments as well as markdown cells to describe your thought process and intermediate results.
- Visually explore the dataset and report upon relevant findings of your initial analysis.
- Use weights-of-evidence encoding and test the sensitivity of the number of bins on the final scorecard.
- Compare different classification models. Clearly indicate the selected model for the application scorecard.
- Explain the model (to the extent possible, given that variables EXT1 to EXT7 have been anonymized) and provide insight in the scorecard points for each variable.
- Apply at least one reject inference method to make sure the scorecard will perform well on future applications.

An *additional objective* of the financial institution is to develop a survival analysis model using the `mortgage.csv` dataset.

- Evaluate whether the time to default for borrowers with low outstanding balance at origination time (30% quantile) significantly differs from borrowers with Average/High outstanding balance at origination time. Display and compare the probability of survival after 25 and 50 months for the two groups.
- Fit a Cox Proportional Hazard model (time varying) using the variables: ["default_time", "hpi_time", "gdp_time", "uer_time", "balance_time", "interest_rate_time", "FICO_orig_time", "LTV_time"]
- Interpret the results by specifically explaining the impact of variables regarding the general economy on the probability of default.

Deadline: July 20th, 2023, 23:59

The files (recorded presentation, slides, jupyter notebook) must be uploaded to IESEG-online. In case there are issues with the website (e.g. file size) you can submit your assignment by email to p.borchert@ieseg.fr

Data

`RealEstateLoans_accepts.xlsx`

Property	Description	DataFormat
total_income	Income of the client	Float
loan_amount	Loan credit amount	Float
term	Term in months	Integer
interest_rate	Loan interest rate in %	Float
own_car	Client owns a car	Binary
own_house	client owns a house	Binary
nr_children	Number of children	Integer
income_type	Client income type (businessman, working, maternity leave,etc.)	String
education_type	Level of highest education the client achieved	String
family_status	Family status	String
housing_type	What is the housing situation of the client (renting, living with parents, ...)	String
region_population_ratio	Normalized population of region where client lives (higher number means the client lives in more populated region)	Float
days_birth	Client age in days (converted from excel)	Integer

Property	Description	DataFormat
days_employed	Number of days in current employment (converted from excel)	Integer
days_registration	Number of days since last registration change (converted from excel)	Integer
mobile_number	Mobile number provided	Binary
phone_number	Landline number provided	Binary
email	Email provided	Binary
days_phone_change	Days since last phone number change (converted from excel)	Integer
occupation_type	Occupation type	String
family_count	Number of family members	Integer
EXT1	Anonymized variable 1	Float
EXT2	Anonymized variable 2	Float
EXT3	Anonymized variable 3	Float
EXT4	Anonymized variable 4	Binary
EXT5	Anonymized variable 5	Binary
EXT6	Anonymized variable 6	Binary
EXT7	Anonymized variable 7	Binary
default	Target variable	Binary
days_payment_arrears	Number of days in payment arrears	Integer

RealEstateLoans_rejects.xlsx

Property	Description	DataFormat
total_income	Income of the client	Float
loan_amount	Loan credit amount	Float
term	Term in months	Integer
interest_rate	Loan interest rate in %	Float
own_car	Client owns a car	Binary
own_house	client owns a house	Binary
nr_children	Number of children	Integer
income_type	Client income type (businessman, working, maternity leave,etc.)	String

Property	Description	DataFormat
education_type	Level of highest education the client achieved	String
family_status	Family status	String
housing_type	What is the housing situation of the client (renting, living with parents, ...)	String
region_population_ratio	Normalized population of region where client lives (higher number means the client lives in more populated region)	Float
days_birth	Client age in days (converted from excel)	Integer
days_employed	Number of days in current employment (converted from excel)	Integer
days_registration	Number of days since last registration change (converted from excel)	Integer
mobile_number	Mobile number provided	Binary
phone_number	Landline number provided	Binary
email	Email provided	Binary
days_phone_change	Days since last phone number change (converted from excel)	Integer
occupation_type	Occupation type	String
family_count	Number of family members	Integer
EXT1	Anonymized variable 1	Float
EXT2	Anonymized variable 2	Float
EXT3	Anonymized variable 3	Float
EXT4	Anonymized variable 4	Binary
EXT5	Anonymized variable 5	Binary
EXT6	Anonymized variable 6	Binary
EXT7	Anonymized variable 7	Binary

Mortgage.csv

Property	Description	Data Format
id	Unique identifier	Integer
time	observation time stamp	Integer
orig_time	time stamp origination	Integer
first_time	time stamp first observation	Integer
mat_time	time stamp maturity	Integer
balance_time	outstanding balance at observation time	Float
LTV_time	Loan to value ratio at observation time, in %	Float
interest_rate	interest rate at observation time, in %	Float
hpi_time	House price index at observation time, in %	Float
gdp_time	GDP growth at observation time in %	Float
uer_time	unemployment rate at observation time in %	Float
REtype_CO_orig_time	real estate type condominium: 1, otherwise: 0	Binary
REtype_PU_orig_time	real estate type planned urban developments: 1, otherwise: 0	Binary
REtype_SF_orig_time	single family home: 1, otherwise: 0	Binary
investor_orig_time	investor borrower: 1, otherwise: 0	Binary
balance_orig_time	outstanding balance at origination time	Float
FICO_orig_time	FICO score at origination time, in %	Float
LTV_orig_time	loan to value ratio at origination time, in %	Float
Interest_rate_orig_time	interest rate at origination time, in %	Float
hpi_orig_time	house price index at observation time, base year=100	Float
default_time	default observation at observation time	Float
payoff_time	payoff observation at observation time	Float
status_time	default (1), payoff (2) and non-default/non-payoff (0) observation at observation time	Integer