Assignment - Credit Scoring

Philipp Borchert

Case description

Introduction

- 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, as well as a description of the dataset.
- 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 final model, i.e., credit scorecard, should be presented and discussed, and clearly a thorough evaluation of the performance of the model is required.
- 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%.
- In addition, you are asked to explore the development of a survival analysis model.

Deliverable

A detailed report:

- The report is to be written as a **Jupyter Notebook** and should include all the code that is required to preprocess the data and to arrive at the final scorecard, as well as to evaluate this scorecard.
- Describing the steps and motivating the choices you made in developing the various models.
- Presenting the resulting scorecard, as well as discussing relevant intermediate results.
- The general guideline here is that you should provide all the code required for a colleague to be able to replicate the model and sufficient documentation so as for your colleague to understand the steps and choices you made.
- **You can choose to implement your model using R or Python, as you prefer.**
- There is no '*page limit*' to your report but you are advised to be both concise and to provide sufficient insight and detail.
- The report on the development of a survival analysis model can be less elaborate, but still should reflect a good basic understanding. The report should be written in a **separate Jupyter Notebook** file.

Hand-in

- **Deadline: April 26th, 2022, 20:00h.**
- The report must be uploaded to IESEG Online (or sent by email to philipp.borchert@ieseg.fr)
- Include yourself in cc of your email to us so that you can verify whether your email was sent correctly and arrived.

Guidelines

- Three datasets are provided at the course platform (IESEG-Online – Credit Scoring):
  - RealEstateLoans_accepts.xlsx
  - RealEstateLoans_rejects.xlsx

    - Mortgage.csv
- A detailed description of the variables in the datasets is provided below.
- Make sure to avoid overfitting your model on the data that is provided using the appropriate methods. It is the performance of your scorecard on future applications which truly matters. Therefore, just like in Kaggle competitions (www.kaggle.com), your model will be evaluated on a separate dataset for a final, out-of-time evaluation of its discrimination power.
- Try to display in the report both technical mastering of the various steps in the model development process, maturity in making choices with respect to data preprocessing, model fitting and evaluation, and importantly also understanding of the exact application you are developing and insight in the results you obtain.

Specific requirements

In developing the application scorecard:

- Visually explore the data and evaluate univariate relationship between the target variable and the individual predictor variables. Report upon relevant findings of your initial analysis.
- Use weights-of-evidence encoding, test the sensitivity of the number of bins or categories.
- Compare a number of predictive methods in terms of performance and make a choice for one of the obtained models.
- Report upon various performance measures as discussed in the course.
- Explain the model (to the extent possible, given that variables EXT1 to EXT7 have been anonymized).
- I expect you to apply a **reject inference method** to make sure the scorecard will perform well on future applications similar to rejected applicants in the past.

In developing the survival analysis model:

- Visually explore the 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

Dataset description

1. **RealEstateLoans_accepts.xlsx, RealEstateLoans_rejects.xlsx**

Variable	Name	Meaning	Type	Remarks
1	Total_income	Income of the client	Continuous	
2	Loan_amount	Credit amount of the loan	Continuous	
3	Term	Term in months	Continuous	
4	Interest_rate	Loan interest rate in %	Continuous	

5	Own_car	Client owns a car	Binary	
6	Own_house	Client owns a house	Binary	
7	Nr_children	Number of children the client has	Continuous	
8	Income_type	Clients income type (businessman, working, maternity leave,…)	Categorical	
9	Education_type	Level of highest education the client achieved	Categorical	
10	Family_status	Family status of the client	Categorical	
11	Housing_type	What is the housing situation of the client (renting, living with parents, ...)	Categorical	
12	Region_population_ratio	Normalized population of region where client lives (higher number means the client lives in more populated region)	Continuous	normalized
13	Days_birth	Clients age in days	Continuous	
14	Days_employed	How many days before the application the person started current employment	Continuous	
15	Days_registration	How many days before the application did client change his registration	Continuous	
16	Mobile_number	Client provided mobile number	Binary	
17	Phone_number	Client provided landline	Binary	
18	Email	Client provided email	Binary	
19	Days_phone_change	How many days before application did client change phone	Continuous	
20	Occupation_type	What kind of occupation does the client have	Categorical	
21	Family_count	How many family members does client have	Continuous	
22	EXT1	Anonymized variable 1	Continuous	normalized
23	EXT2	Anonymized variable 2	Continuous	normalized
24	EXT3	Anonymized variable 3	Continuous	normalized

25	EXT4	Anonymized variable 4	Binary	
26	EXT5	Anonymized variable 5	Binary	
27	EXT6	Anonymized variable 6	Binary	
28	EXT7	Anonymized variable 7	Binary	
29	Default	Target variable indicating default (90+ days in payment arrears)	Binary	Not available for rejects

2. Mortgage.csv

Variable	Name	Meaning	Type	Remarks
1	ID	Borrower ID	Identifier	
2	time	time stamp of observation	Continuous	
3	orig_time	time stamp for origination	Continuous	
4	first_time	time stamp for first observation	Continuous	
5	mat_time	time stamp for maturity	Continuous	
6	balance_time	outstanding balance at observation time	Continuous	
7	LTV_time	loan to value ratio at observation time, in %	Continuous	
8	interest_rate_time	interest rate at observation time, in %	Continuous	
9	hpi_time	house price index at observation time, base year=100	Continuous	
10	gdp_time	GDP growth at observation time, in %	Continuous	
11	uer_time	unemployment rate at observation time, in %	Continuous	
12	REtype_CO_orig_time	real estate type condominium: 1, otherwise: 0	Binary	
13	REtype_PU_orig_time	real estate type planned urban developments: 1, otherwise: 0	Binary	

14	REtype_SF_orig_time	single family home: 1, otherwise: 0	Binary	
15	investor_orig_time	investor borrower: 1, otherwise: 0	Binary	
16	balance_orig_time	outstanding balance at origination time	Continuous	
17	FICO_orig_time	FICO score at origination time, in %	Continuous	
18	LTV_orig_time	loan to value ratio at origination time, in %	Continuous	
19	Interest_Rate_orig_time	interest rate at origination time, in %	Continuous	
20	hpi_orig_time	house price index at observation time, base year=100	Continuous	
21	default_time	default observation at observation time	Continuous	
22	payoff_time	payoff observation at observation time	Continuous	
23	status_time	default (1), payoff (2) and non-default/non-payoff (0) observation at observation time	Categorical	