Task for Junior Quantitative Analyst in Credit Risk Model Validation

SEB

1 General

Purpose of exercise is to enable candidate to demonstrate ability to understand and manipulate data, combine different datasets, address data issues in a context of non-trivial statistical model application and make conclusions with sound reasoning level.

1.1 Glossary

Key terms and definitions used further on (simplified and task specific):

- **Credit arrangement** - an agreement between financial institution (bank) and person/business (borrower) under which bank issues credit (lends money) to borrower. Set of arrangements is referred to as a portfolio;
- **Default** - event within credit arrangement lifecycle where credit holder has not fulfilled agreed obligations under certain conditions or is recognized as potentially not being able to do so;
- **Obligor** - credit arrangement holder (borrower). Single obligor may hold several credit arrangements. Single credit arrangement can only contain single obligor in a context of given exercise;
- **Past due** - event within credit arrangement lifecycle where credit installment is not paid by agreed date on a payment schedule. Material amount past due for a certain amount of time may qualify as default event. There are more reasons for default events than past dues;
- **Scoring** - process (including methods) of assigning credit risk parameter estimates (scores) to credit arrangement portfolio records in order to use for further applications.

1.2 Context

Provided is dataset that describes usage of logistic regression model to estimate probability of default (PD) for a hypothetical credit arrangement portfolio. Model application is approached via considering a portfolio of valid credit arrangements (contracts) at the end of a calendar year and assigning probability (PD) that particular contract would go into default over the next 12 months. PD is estimated on a contract level for each calendar year (end) in a dataset and depends on various contract and contract holder related risk attributes. In the same manner, dataset also contains binary variable reflecting whether particular contract has actually defaulted.

1.3 Model

Design

PD is modelled in terms of logistic regression with the following risk attributes considered (see details in table 4):

- **Age** - age of credit arrangement holder, years;

- **Education** - education level of credit arrangement holder, categorical;
- **Remaining debt ratio** - ratio of outstanding credit amount in relation to whole borrowed amount at a point of time;
- **Number of past due days** - number of past due days during last 24 months (obligor level, maximum of all arrangements combined);
- **Months since last past due** - number of months since last past due days (obligor level, minimum of all arrangements combined)

Application

GIven portfolio is subject to an arrangement level risk scoring. Logistic regression model with risk attributes described above is further used to conduct actual scoring of actual portfolio (result in dataset `dat_scoring`). Model coefficients are provided in table 4. The following rules and transformations must be applied to raw risk attributes (in that order) before assigning coefficient:

- Missing values are replaced by field *MISSING_VALUE* value for all risk attributes
- Raw numeric risk attribute values are capped and floored by *MAX* and *MIN* values respectively corresponding to that attribute in table 4
- Raw numeric risk attributes are transformed by standardizing. Standardization for arbitrary raw risk attribute is conducted by using parameters *ST.DEV* and *MEAN* corresponding to that attribute in table 4

Given transformed and missing-value-managed risk attribute values, logistic regression model with coefficients (table 4) is applied and raw PD value is obtained (*PD* field in `dat_scoring`).
Raw PD estimate value is further grouped into categories by assigning PD Pool (table 5) dataset for details and pool boundaries). Both raw and pool level PD estimate might be used further for different purposes.

1.4 Data

Dataset consists of the following tables:

- `dat_scoring` - valid credit arrangement portfolio with obligor information and estimated PD and default flag assigned at year ends (see details at table 3);
- `dat_arrangement` - table of valid credit exposures (arrangements) over time (see details at table 1);
- `dat_obligor` - table of credit holder information over time (see details at table 2);
- `par_model` - risk attribute (variable) and coefficient structure of PD model used;
- `par_pools` - PD Pool assignment boundaries

Fields comprising a primary key are marked (pk) in each table respectively.

2 Task

Please approach the following questions to your best knowledge. Questions can be addressed independent from each other. Feel free to use software of your choice, as well as libraries, graphical tools, own built tools, etc. Result is expected to be delivered in a structured report, which is expected to include not only reasoning to decisions and conclusions made, but also brief description on data processing steps conducted (if any). Please also attach codes used. Results obtained are expected to be replicable.

1. Visualize and provide a brief description on qualitative portfolio structure in terms of obligor, arrangement and scoring features available, portfolio composition and performance. Assess consistency over time
2. Provide opinion on data quality in terms of completeness, accuracy, consistency.
3. Assess correctness of PD estimate calculation.
4. Using quantitative and/or statistical tools, measure PD model performance from at least one of the following perspectives:
   (a) PD model performance in terms of PD estimate sufficiency and accuracy
   (b) PD model performance in terms of risk differentiation on risk attribute raw PD estimate levels
   (c) PD model stability over time in terms of model coefficients and model result

3 Appendix

3.1 Data description

Field name	Field type	Description
AR_ID (pk)	character	Credit arrangement (credit contract) identifier
YEAR (pk)	integer	Calendar year (contract is valid at the end of)
DEBT_RATIO	double	Remaining debt ratio of underlying credit arrangement
DPD	integer	Number of past due days during last 24 months
M_LAST_DPD	integer	Months since last past due during last 24 months

Table 1: Structure and field attributes of data tables: `dat_arrangement`

Field name	Field type	Description
IP_ID (pk)	character	Person identifier
YEAR (pk)	integer	Reference year of information validity (year end)
AGE	integer	Person age (years) at reference time
EDUCATION	factor	Person education level at reference time

Table 2: Structure and field attributes of data tables: `dat_obligor`

3.2 Model parameters

Field name	**Field type**	**Description**
YEAR (pk)	integer	Calendar year (end of)
IP_ID (pk)	character	Person identifier of underlying credit arrangement holder (key in 2). Always consists of 3 letters and 3 numbers.
AR_ID (pk)	character	Arrangement identifier (key in 1). Always consists of 8 numbers.
PD	double	Estimated PD at the end of a year for given credit arrangement
PD_POOL	factor	PD pool (key in 5)
DFLT_FLAG	binary	Indicator whether contract has defaulted during next 12 months

Table 3: Structure and field attributes of data tables: `dat_scoring`

Field name	**Field type**	**Description**
VARIABLE	character	Risk attribute
LABEL	character	Risk attribute explained
MEAN	double	Standardization parameter of raw value
ST.DEV	double	Standardization parameter of raw value
MIN	double	Floor value of raw value
MAX	double	Cap value of raw value
MISSING_VALUE	double	Missing value handling rule
ESTIMATE	double	Beta coefficient in corresponding to particular (category of) risk attribute

Table 4: Structure and field attributes of data tables: `par_model`

Field name	**Field type**	**Description**
LABEL (pk)	factor	PD Pool
START	double	Lower boundary of raw PD
END	double	Upper boundary of raw PD

Table 5: Structure and field attributes of data tables: `par_pools`