**Development of Probability of Default model**

The purpose of proposed assignment was development of Probability of Default model for a portfolio containing a large and diversified pool of customers. I was provide with raw data file and supportive materials with details of model development and validation process. The process of development also had to have sections dedicated to testing model's predictive power and calibration precision.

In this essay, I will briefly describe the approaches applied for development of the model, steps of the process and tools used for development. I will also discuss caveats I have encountered while developing the model, and of course, the results of the performed analysis.

The concept

The concept of Probability of Default model is analysis of so-called good and bad observations. In case of good/bad analysis observation is a state of a given client or contract at specific date (usually reporting date). Explanatory variables of this observation are features that describe its current state. Dependent variable (usually called target) is an indicator whether this contract or client defaulted within predertmined period (for the purpose of a current analysis default within 12 months is used). Set of such observations for one date comprise portfolio for a given reporting date. To obtain panel data for good/bad analysis portfolios for different dates are merged into one dataset. The dataset which was used in the analysis is exactly the one described above.

Tools for development

As a tool of development, I have used Python for its usability qualities. The environment to write and test the code Jupyter Notebook was used. Python has a great deal of libraries which are very essential while development of the model. Among libraries, I have used in the analysis are:

* pandas – for handling the data and performing various manipulations with it;
* NumPy – for algebraic expressions;
* SciPy – for statistical purposes;
* StatsModels and scikit-learn for running regressions and calculation of metrics;
* matplotlib and seaborn – for data visualization.

The approach

Now when we have discussed the concept of analysis let us dive into details of the approach for the analysis. In real world, there are just two possible states of the client or contract within 12 months - defaulted or not defaulted. We want to predict this state based on an information we have at the moment. Such type of a problem is called binary classification. In such problem object with unknown class needs to be classified into one of two possible classes. We need to construct an algorithm which will assign the class to an object based on information provided. For binary classification, the logic is following – calculate the probability that object belongs to a given class and determine the threshold for this probability. That probability is actually a probability of default in our case, i.e. probability that observation is bad (default within 12 months. For the purpose of Probability of Default, modeling usually logistic function is used:

where are weight for different factors (i.e. what is the strenght of factor's influence on probability).

After the choice of probability function, we can write likelihood function as a product of probabilities for entire sample and maximize it with respect to the set of weights.

In practice, the optimization problem to find the coefficient is solved via statistical packages. My experience during performing the assignment included application of statsmodels’ Logit method that outputs coefficient together with inference statistics.

Stages of the analysis

The high-level structure of the performed analysis is following:

* Data import and initial inspection
* Exploration of data descriptive statistics
* Single Factor Analysis (SFA)
* Multiple Factor Analysis (MFA)

This structure is common for any model development process and widely applied in financial sector. For instance, one can find quite good description of the process by Naeem Siddiqi (see sources). Actually, the author describes the entire lifecycle of the model from advising on development project management to implementation of models. In this essay, I will focus only in development process itself.

Data import and initial inspection

Data importing should not be underestimated since it is very crucial for proper functioning of the whole algorithm and correct estimations on further stages of model development. My experience with dataset provided for the analysis included following issues (but not limited to):

* Date formats for rating date and date of default differed from each other. Moreover the conversion to Python datetime format was needed;
* Total sales had missing values and string format when imported to Python. In addition, high conversion of string to number when there more than 20 symbols yielded errors. Thus some tricks were made before importing;
* Identification of some full duplicates and deleting them from the sample;
* Discovery of an issue with total sales missing values – several thousand observations with different total sales (one with number, another with missing value) figures given all other features fixed;
* Two columns describing high risk of region were absolute duplicates;
* 4 numerical and 9 categorical (one of them excluded) variables were identified.

Exploration of data descriptive statistics

Data exploration stage included but was not limited to analysis of factors’ values distributions proportion of default rates within different values etc. Among tasks performed on this stage are following:

* Bad and good cases – how factors’ values differ for default and non-default cases (to be more precise good and bad observations). Yet on current step «payment arrears within 12 months» was identified as potentially good predictor;
* Proportions of defaults within classes of dummy variables;
* Distributions of numerical factors’ values. This step was fruitful for discoveries of anomalies in the data. Among observations were (but not limited to):
  + «total sales» had very wild distribution;
  + «leverage percentage» had abnormal number of observations with one specific value (more than 60% of the sample);
  + «loan to book» and «solvency ratio» had abnormal numbers zeroes that obviously indicated those are actually missing values (was confirmed later);
* Pairwise distribution of numerical factors’ values;
* Observed default rates (upward trend was identified);
* Correlations between factors and multicollinearity analysis – «loan to book» and «solvency ratio» found to be perfectly correlated;

Single Factor Analysis

This stage was dedicated to study of factor’s influence on probability of default. Each factor was analyzed separately. This stage included division of entire sample to train and test subsets to analyze how model developed on train sample performs on test sample. On the stage of SFA two major task were executed:

* Running one factor logistic regressions, and evaluating statistical significance and factors’ influence magnitudes (on entire sample);
* Construction of Receiver Operator Characteristic (ROC) as a function of True Positive Rate (TPR) and False Positive Rate (FPR) for each factor on both train and test samples;
* Calculation of ROC-AUC (Area Under the ROC Curve) and Gini coefficient for each factor on train and test samples;
* For factors with missing and abnormal values, analysis was carried out with and without substitution.

As the result of analysis carried out while this stage, the shortlist of factors to be analyzed in MFA was compiled. The best factors according to statistical significance and Gini coefficients were «payment arrears within 12 months» and «leverage percentage». Poorly performing factors such as «total sales» and «high risk region» were excluded from further analysis. Totally 8 factors were selected.

Multiple Factor Analysis

Multiple factor analysis comprises selection of best performing combination of factors from SFA and testing the whole model with established set of factors on sample/test samples and development/out-of-time samples. Since in our case there were just 8 factors the step of optimal set was skipped. All factors were included in the model. In practice when there are much more factors selection algorithms such forward/backward/stepwise selection are used.

Steps performed on MFA:

* Running logistic regression with all factors on train and test samples, and testing joint hypothesis that all factors are not equal to zero. All regressions run on this stage were had low p-value for log-likelihood ratio statistic used to test joint hypothesis;
* Construction of ROC curves, and calculation of ROC-AUC and Gini for all samples;
* Inspection of distribution of scores assigned by estimated model and running Chi-square (Hosmer-Lemeshow) test to check calibration. Chi-square statistic had extreme value for development/out-of-sample case, thus final model had to be calibrated;
* Calibration of logit coefficients estimated on entire sample based on most recent observations.

Conclusions

Let us sum the analysis carried out to develop and validate the performance of the Probability of Default model. To develop the model I have gone through all necessary steps including data handling, descriptive statistics inference, SFA and MFA. As a result, model with sufficiently high Gini (56% – test set 54% – out-of-sample set). Due to change of default rate within sample (and probably further) model was calibrated to account for that.

Takeaways and possible next steps

It could be noticed from stages description that there were many caveats while development process. To overcome them various solutions were made. However there are several actions which could be done to enhance the current model:

* Missing values imputation. There are several methods to deal with the issue. In the analysis I have used mean (conditional and unconditional) substitution. Other option are described in ipynb file provided with this essay
* Good/bad analysis could be carried on reduced data set – that is when comparable number of good and bad observations are taken to run the analysis and build the model.

It is also possible to develop other types of predictive models, for instance one based on decision trees, and compare their performance with current one.

Sources used for the development process:

* Credit Risk Scorecards, Developing and Implementing Intelligent Credit Scoring, Naeem Siddiqi, 2006
* Studies on the Validation of Internal Rating Systems, Working Paper No. 14, 2005, BCBS
* A practical approach to validating a PD model, Lydian Medema et al., Journal of Banking & Finance, Volume 33, Issue 4, 2009
* How to Measure the Quality of Credit Scoring Models, Martin ŘEZÁČ & František ŘEZÁČ, Czech Journal of Economics and Finance, 61, 2011, no. 5
* Notes on courses of Econometrics for several modules taken in New Economic School in the period of 2010-2012, lecturer Stanislav Anatolyev, <https://scholar.google.com/citations?user=T2OAOI0AAAAJ&hl=ru>
* Materials and notes on course of Advanced Data Science led by Victor Kantor, <https://github.com/applied-data-science/ml2018jan_feb>