05.03.19

Credit risk scoring modeling

PwC Business Case Competition 2019
SGH

Agenda

1. Background
2. Rules
3. Data
4. Timeline
5. Output file format
6. Prizes
7. Q&A

1. Background

1. Background

Our Client, one of the leading retail bank in Poland, wants to improve its risk assessment process in private individuals segment

Overview of current situation

Bank's new business strategy for the upcoming years assumes significant increase in private individuals lending, with the main focus on cash loans for existing Bank customers . The Risk Department, however, is struggling with the quality of their current credit portfolio. In order to meet the business expectations and simultaneously limit the possible credit losses, CRO[1] demands to redesign the current scoring model that could more precisely asses the credit risk of new potential borrowers.

The Bank expects the following

To help the Risk Department, **PwC is asked to develop the new scoring model based on:**

- Application data (information provided by the client during the credit application process)
- Behavioral data (data regarding customers' previous loans and their performance)
- Geolocation data (characteristics of the region of the customer's permeant residence)

1. Chief Risk Officer

2. Rules

2. Rules

1. The main objective of the competition is **to develop scoring model predicting the possible default of the customer**

2. Students will be divided into working groups, each of which will be receiving **the same data set**
   Data set consists of **100k credit exposures** and all relevant information that can be used in modeling
   Data set is divided into two samples; **for 70% of the observations** students can see the target variable (default of the customer) – ***development sample***; **for the remaining 30% of observations** target variable is unknown for the competitors – ***test sample***

3. The quality of the models will be measured by:
   1. Maximization of **GINI index measured on the full data set**

4. Additionally, students will have to **present their best model during final classes**. Presentation (preferably in Power Point) can include description of chosen approach, parts of used code, interesting findings, justification behind choosing the final model etc.

5. Each working group will have to provide **the codes for their best model**. The codes should allow for development of the final model, including all possible data transformation operations

6. The codes should be prepared using R. Minimum software requirements should be assured:
   R version: >= 3.5.0, Platform: x86_64-w64-mingw32/x64 (64-bit), Running under: Windows >=8 x64

7. **Every 2 weeks**, each working group will be able to check GINI index of their model by sending results (score) of the test sample. **(PL_Risk.Competition.2019@pwc.com)**

3. Data

3. Data

Behavioral data

- Aggregated data from the period before the credit application
- Delays in payments, due amounts, book values of exposures
- Indicates how the customer pays his installments

Application data

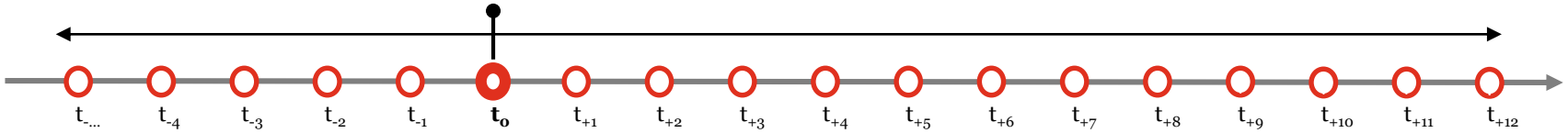
- Data from the credit application form
- Customer characteristics, income, age, marital status, etc.
- Information is provided by the client and then verified by the Bank

Geolocation data

- Data from external data sources; e.g. public registries, national statistics bureau
- Data on employment, spendings, housing, demographics
- Determinates the specific characteristics of customer's residence region

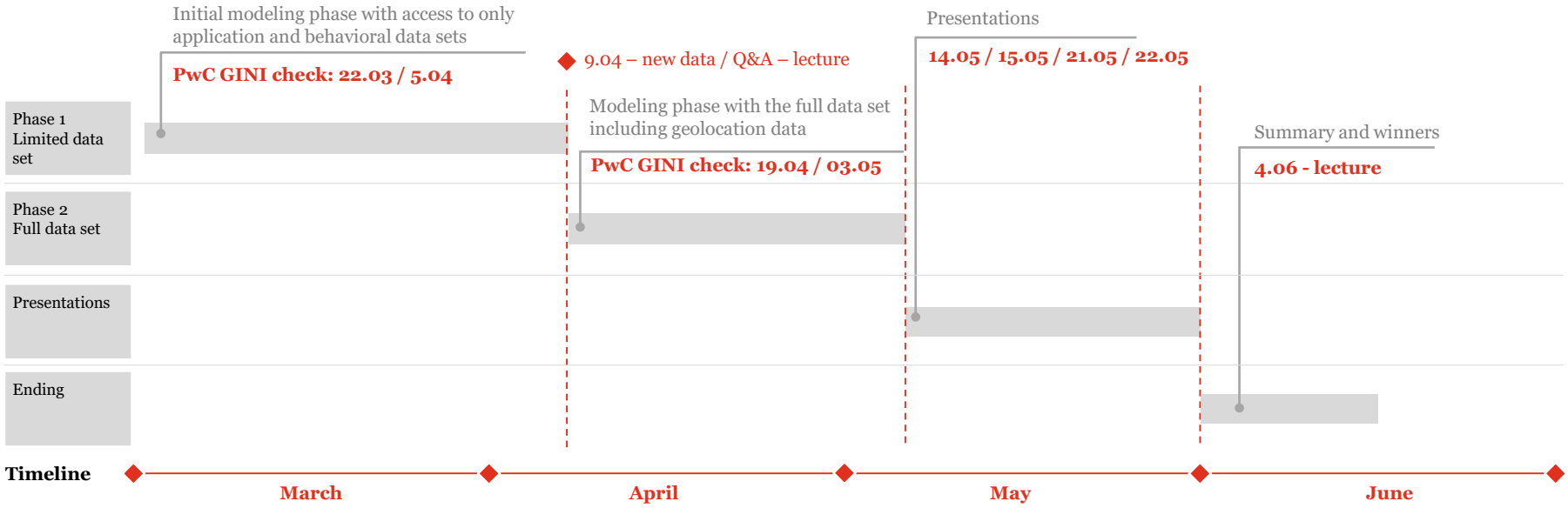
Default flag

- Default flag indicates whether the exposure defaulted during 12 month time from the origination period
- Usually it is assumed that the exposure which is due more than 90 days is defaulted – such exposure is unlikely to be repaid by the customer



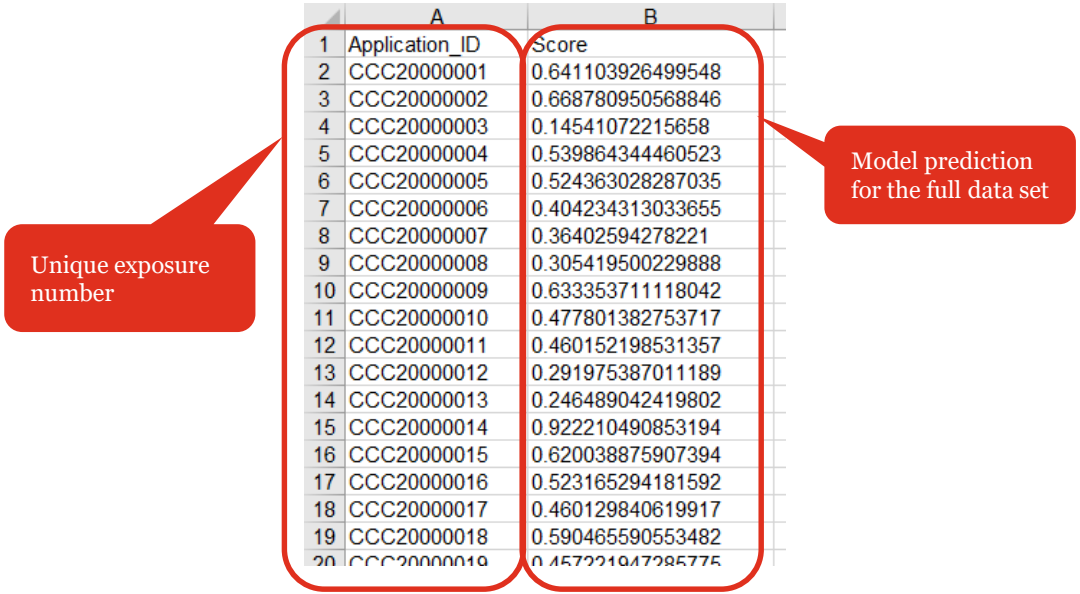
4. Timeline

4. Timeline


Initial modeling phase with access to only application and behavioral data sets
PwC GINI check: 22.03 / 5.04
9.04 – new data / Q&A – lecture
Modeling phase with the full data set including geolocation data
PwC GINI check: 19.04 / 03.05
Presentations
14.05 / 15.05 / 21.05 / 22.05
Summary and winners
4.06 - lecture
Phase 1 Limited data set
Phase 2 Full data set
Presentations
Ending
Timeline
March
April
May
June

5. Output file format

5. Output file format



	A	B
1	Application_ID	Score
2	CCC20000001	0.641103926499548
3	CCC20000002	0.668780950568846
4	CCC20000003	0.14541072215658
5	CCC20000004	0.539864344460523
6	CCC20000005	0.524363028287035
7	CCC20000006	0.404234313033655
8	CCC20000007	0.36402594278221
9	CCC20000008	0.305419500229888
10	CCC20000009	0.633353711118042
11	CCC20000010	0.477801382753717
12	CCC20000011	0.460152198531357
13	CCC20000012	0.291975387011189
14	CCC20000013	0.246489042419802
15	CCC20000014	0.922210490853194
16	CCC20000015	0.620038875907394
17	CCC20000016	0.523165294181592
18	CCC20000017	0.460129840619917
19	CCC20000018	0.590465590553482
20	CCC20000019	0.457221947285775

Requirements for the format and sharing process of the output file:

- File should be in .csv format
- File name – „Output”
- File should include two variables: exposure number and prediction value
- File should include 100k observations without any blank values (important!), i.e. predictions should be also calculated for the test sample where Default flag is not available
- Output file should be sent in agreed timeline (p. 4. Timeline) on: PL_Risk.Competition.2019@pwc.com
- Email title: **[Case_study][SGH][group][ID]**. Full list with ID numbers of groups/people should be delivered until March 12

6. Prizes

6. Prizes

Authors of the best models will be rewarded with the following prizes:

Internships in PwC

Official diplomas

PwC branded gifts

7. Q&A

Thank you!