dr Karol Przanowski

[kprzan@sgh.waw.pl](mailto:kprzan@sgh.waw.pl)

**Opis nowego projektu zaliczeniowego do zajęć**

**nr przedmiotu: 223110-1234**

**„Podstawowe i zaawansowane programowanie oraz statystyka w SAS”**

# Ogólny opis zadań

Celem projektu jest zapoznanie studenta z budową w pełni automatycznych procesów raportowych wykorzystując zarówno metody przetwarzania danych jak i analizy statystyczne. Projekt oparty jest na podstawowych sposobach raportowania i analizowania portfela kredytów detalicznych przydatnych w zarządzaniu ryzykiem kredytowym. W tym wypadku analizy koncentrują się na etapie weryfikacji danych wejściowych, analizie jakości i pierwszych identyfikacjach predyktorów ryzyka kredytowego.

# Teoretyczne podstawy

W obecnych czasach przeprowadzania analizy jakości danych staje się coraz bardziej istotnym elementem każdego projektu analitycznego. Śmiało można założyć, że nie uda się zgromadzić danych o dobrej jakości, ale trzeba umieć zmierzyć ich jakość i potem wiedzieć, jak to wpłynie na finalne wyniki analiz. Najczęstszymi elementami jakości są:

* Identyfikacja braków danych
* Identyfikacja wartości odstających, nietypowych
* Stabilność w czasie.

# Rzeczywistość projektowa

Niestety coraz częściej pojawiają się sytuacje, kiedy w projekcie nie posiadamy wystarczającej informacji. Z reguły trzeba wtedy organizować wiele dodatkowych spotkań, czy warsztatów, a także robić wywiady z różnymi uczestnikami procesu. Zawsze pozostaje jednak pewien brak. Należy zatem przyjąć, że opis projektu zawiera wszystkie najważniejsze informacje, a wszelkie niejasności i luki w informacji należy uzupełnić samodzielnie poprzez ustalenie swoich założeń i wybranie dla nich najlepszej metody. Można też zadawać pytania prowadzącemu podczas zajęć.

# Sposób oceny projektu

Należy jednoznacznie podkreślić, że nie istnieje jedno i najlepsze rozwiązanie projektu. Można go wykonać na wiele sposobów, co oznacza, że oceniane będą różne elementy: wykonanie, prezentacja, założenia, metodyka, ciągłość logiczna wszystkich analiz i formułowanych wniosków.

Projekt z formalnego punktu jest prezentowany i broniony.

Punktowane są następujące elementy:

|  |  |
| --- | --- |
| **Kryterium** | **Maksymalna liczba punktów** |
| Prezentacja | 3 |
| Logika i poprawność argumentacji | 5 |
| Szczegółowość raportów | 4 |
| Interaktywność i wizualizacja raportów | 3 |
| Finalne wnioski | 5 |
| **Razem** | **20** |

# Opis danych

Do projektu dołączone są dwa zbiory danych Abt\_sam\_beh\_train (52841 obserwacji) i Abt\_sam\_beh\_valid (53070 obserwacji) posiadające 2310 zmiennych każdy z nich i reprezentujące podobny okres czasu losowo wygenerowanych danych z większego zbioru danych reprezentującego zbierane dane podczas procesu akceptacji kredytowej.

Zmienne dzielimy na:

* Identyfikacyjne: cid – identyfikator klienta i period – moment czasowy, na który liczymy wszystkie cechy klienta (inaczej punkt obserwacji lub data danych).
* Funkcje celu: default\_cusX – czy klient od punktu obserwacji w ciągu pierwszych X miesięcy wszedł w opóźnienia więcej niż 3 (dla X=3 – więcej niż 2) raty.
* Charakterystyki klienta: zmienne zaczynające się od app\_ - cechy z aplikacji, act\_ - opisujące jakiś stan z danego punktu czasowego, agr\_ i ags\_ - zmienne agregujące informacje z wielu miesięcy przed punktem obserwacji.

Szczegółowy opis zmiennych w języku angielskim znajduje się w Excelu o nazwie: labels.xlsx.

# Lista zadań

Najważniejsze zadania w projekcie sprowadzają się do następujących analiz:

1. Analiza braków danych, ich udziałów w czasie (czyli stabilności w czasie) i porównywania udziałów pomiędzy zbiorami train i valid (czyli stabilności na zbiorach) w postaci szczegółowego raportu tabelarycznego.
2. Analiza braków danych w postaci zwizualizowanych graficznych raportów w celu szybkiego identyfikowania zmiennych z większą i mniejszą liczbą braków danych oraz z ich różną stabilnością.
3. Analiza wartości nietypowych, odstających lub ogólnie nieregularności rozkładów, ich udziałów w czasie (czyli stabilności w czasie) i porównywania udziałów pomiędzy zbiorami train i valid (czyli stabilności na zbiorach) w postaci szczegółowego raportu tabelarycznego.
4. Analiza nietypowych danych w postaci zwizualizowanych graficznych raportów w celu szybkiego identyfikowania zmiennych z większą i mniejszą nietypowością i nieregularnością danych oraz z ich różną stabilnością.
5. Analiza zależności pomiędzy funkcją celu defaut\_cus12 a charakterystykami klienta. Identyfikacja zmiennych wpływających na funkcję celu, ich współzależności, korelacji, zależności itp. W tym wypadku finalny raport może składać się i z raportów tabelarycznych, i graficznych wizualizujących dobroć predyktorów.

Raporty mogą być wykonane w dowolnym formacie: Excel, RTF, HTML, PDF lub cokolwiek innego. Preferowane są raporty interaktywne, by można było szybko przełączać się pomiędzy nimi.

Finalny zestaw raportów powinien składać się z kilku warstw:

* Warstwa zbiorcza, gdzie przedstawione są najważniejsze wskaźniki zmiennych i ich porządek.
* Warstwa szczegółowych analiz zmiennych.

# Dodatkowe zagadnienia

Podczas projektu mogą się pojawić następujące problemy, które warto rozwiązać:

1. Czy bak danych jest informacją, czy tylko losową utratą informacji?
2. Czy pokazać nazwę zmiennej, czy wygenerować pełny jej opis na podstawie labels.xlsx?
3. Czy w liczeniu wpływu cech na funkcję celu uwzględniać braki danych w cechach i czy w funkcji celu brać dwie wartości (0, 1) czy tylko trzy (0, 1, .i)?
4. Jak badać stabilność w czasie? Jaką statystykę wyznaczyć?
5. Jak badać stabilność na zbiorach train i valid?
6. Czy zmienną dzielić na kategorie, czy może tylko na grupy: typowe i nietypowe?
7. Jak policzyć dobroć predyktora?
8. Czy można wyznaczyć benchmarki dla udziału braków danych i wartości odstających? Czy przydatny będzie raport z sygnalizacją świetlną?
9. Czy benchmarki powinny być zależne od skośności rozkładu zmiennej?
10. Czy mogą się pojawić zmienne, które są prawie puste i nie warto ich analizować?
11. Czy mogą się pojawić zmienne, które posiadają inną nazwę a wartości takie same?
12. Czy istnieją zmienne, które są podobne do siebie i być może wystarczy wybrać tylko ich reprezentanta?

# Kreatywność

Istotą projektu jest nie tylko napisanie właściwych kodów w SAS, ale także umiejętne rozwiązanie problemu dużej liczby raportów i wielu obliczeń. Wykonawcy projektu powinni postawić się w roli osób uczestniczących w projekcie budowy modelu predykcyjnego na etapie badania jakości danych wejściowych i starać się najlepiej przygotować wnioski, by móc jednoznacznie odpowiedzieć na pytanie: czy jakość otrzymanych danych jest zadowalająca do budowy modelu predykcyjnego prognozującego default\_cus12? Finalne zestawy raportów muszą być czytelne i w powinny być tak dobrane do danych, by można było w szybki sposób zauważyć najważniejsze własności danych.

Projekt może wykonywać kilka osób odpowiednio dzieląc się zadaniami, byleby każdy student miał mniej więcej tyle samo pracy do wykonania, preferowana liczba uczestników w projekcie to 5.