Wskazówki dotyczące Wykonania Projektu

1. Głównym zadaniem projektu jest przygotowanie raportów dotyczących ANALIZY JAKOŚCI oraz WSTĘPNY MODEL OCENY RYZYKA KREDYTOWEGO.

2. Preferowane (ale nie wymagane) są raporty interaktywne. Poniżej **przykład**:
   Building a Better Dashboard Using SAS
   To tylko przykład nie chce, żebyście reprodukowali wszystkie przedstawione tutaj rozwiązania, ale jest to dobry materiał do inspiracji.
   * Może to być też plik w Excelu z hiperłączami do poszczególnych zakładek.
   * Może to być strona HTML z łączami do tabel, grafów, itd.
   * Może przyjąć DOWOLNĄ elektroniczną formę, byleby użyć programowania w SAS.

   Nie zawieszajcie się na siłę na interaktywności raportów, gdyż może to być dość skomplikowane od strony programistycznej. Jeżeli się nie uda lepiej stworzyć bardziej pełny, bardziej informatywny ale statyczny lub jakiegoś rodzaju łączony statyczno – dynamiczny raport.

   W skrócie wynikiem projektu ma być zbiór wykresów, tabel, zestawień itd., który bada jakość danych i pokazuje jakiegoś rodzaju wstępny model oceny ryzyka kredytowego. Na obronie pokażecie to co wyprodukowaliście i opowiecie o tym.

   Nie piszczcie raportów w formie sprawozdań, kilkudziesięciu stronicowych tekstów z wykresami i tabelami. Za to przygotujcie jakiegoś rodzaju dashboardy, z analizy których będą płynąć określone konkluzje, z którymi się zgadzacie i będziecie w stanie je obronić.

3. WYJAŚNIENIA DO LISTY ZADAŃ:

   Ogólnie projekt ten przedstawia swego rodzaju przedwstępną analizę danych do modelowania ryzyka kredytowego. Dane musicie podzielić na zbiory:

   - treningowy (ten, na którym się uczy model)
   - walidacyjny (ten, na którym sprawdzamy, jak model został nauczony, dane te nie biorą udziału w procesie uczenia)
   * - testowy (służy do porównywania wyników pomiędzy różnymi modelami, jeżeli postanowilibyście tworzyć więcej niż jeden).

   Podział powinien być losowy istnieją różne metodologie jak dzielić dane popularne są na przykład:
   - 2/3 danych zbiór treningowy, 1/3 danych zbiór walidacyjny
   - 80% danych zbiór treningowy, 10% danych zbiór walidacyjny, 10% danych zbiór testowy

   Powyższe podziały są tylko PRZYKŁADOWE, bierzcie pod uwagę jak duże są zbiory, czy klasyczne podziały mają sens. Jeśli nie zaproponujcie inne. Wszystkie podejścia są dobre jeżeli da się je LOGICZNIE wyjaśnić.

   Model nie musi być skomplikowany! Może to być prosta regresja liniowa, regresja logistyczna, analiza kontyngencji, analiza korespondencji itp.

Przydatne książki z oficyny wydawniczej SGH:
Zaawansowane Metody Analiz Statystycznych

Modelowanie dla Biznesu

Statystyka od Podstaw z Systemem SAS

Przetwarzanie Danych w SAS

Wielowymiarowa Analiza Statystyczna

I oczywiście zachęcam do korzystania z Internetu!

Stabilność na zbiorach – jak braki danych lub nieregularności różnią się per zmienna per zbiór
Stabilność w czasie – jak baraki danych lub nieregularności różnią się per zmienna per punkty w czasie

AD1. Sprawdźcie braki danych na zmiennych (wybranych albo wybranych podzbiorach tematycznych przyjmijcie dowolną metodologię) i przedstawcie je w formie odpowiedniego zestawu tabel.

AD2. Stwórzcie wykresy pokazujące, które zmienne mają ile braków danych jak te braki rozkładają się w czasie.

Zastanówcie się jak duże są braki danych, czy jest sens imputacji danych, jeśli tak to w jaki sposób? Poprzez dominantę poprzez średnią, jakoś inaczej dlaczego?

AD3. Sprawdźcie czy istnieją wartości odstające, nietypowe i przedstawcie je w formie odpowiedniego zestawu tabel. Istnieje mnóstwo możliwości sprawdzenia czy istnieją wartości odstające, nietypowe (tabele kontyngencji, metoda trzech sigm, na podstawie rozstępu międzykwartylowego).

AD4. To co powyżej tylko w odpowiedniej formie graficznej.

Zastanówcie się co zrobić z takimi obserwacjami. Pozostawić, usunąć, przeprowadzić jakąś transformację?

AD5. Tutaj właśnie jest miejsce na model. Zwizualizowany za pomocą odpowiedniej liczby tabel i wykresów.

4. Warstwy

   Dobrze, jeżeli raport jest warstwowy. Czyli istnieje analiza zbiorcza, np. procentowy udział braków danych w zmiennych. Ale również szczegółowa analiza dla każdej analizowanej zmiennej. Ale nic na siłę! Ma być klarownie, przejrzyście I Z SENSEM 😊.

Podsumowując, projekt powinien być ustrukturyzowanym zbiorem tabel i wykresów, który będzie odpowiedzią na listę zadań oraz pomoże przedstawić wam jako zespołowi logiczny wywód na temat jakości otrzymanych danych, wstępnego modelu i rekomendacji co do jego użyteczności i dopasowaniu.

Przydatne procedury: proc freq, proc mean, proc reg, proc logistic, proc sgplot, proc sgrender, proc boxplot, proc ttest i wiele, wiele innych. Z uwagi na małą ilość zajęć musicie wiele doszukać sami.

Mam nadzieję, że to coś wyjaśni. Jak coś to piszcie, postaram się pomóc. Zależy my, żebyście napisali trochę kodu i przede wszystkim przedstawili mi spójny raport wyprodukowany w SAS. Lub też jego elementy wyprodukowane są w SAS (wykresy, tabele) a potem został złączony w inny sposób, pozostawiam dowolność 😊.

Zastrzegam, że to są moje uwagi i wskazówki! A nie wypracowane stanowisko wszystkich prowadzących!