O co chodzi w pracy? Quasi pytania badawcze

1. Sprawdzić co wpływa na to, że klienci kupują drugi raz – jako zysk dla firmy
2. Sprawdzić czy w tych klientach którzy kupili drugi raz da się wyróżnić jakieś grupy – w celu dostosowania strategii marketingowej pod konkretnego klienta

Obszary analizy:

1. Predykcja, jakie jest prawdopodobieństwo drugiego zakupu klienta (bliźniacze do churn prediction)
2. Klastrowanie klientów aby móc dostosować strategię marketingową do konkretnego klienta który ma szanse zostać

Tematy do zawarcia w literature review:

1. Rynek Brazylii
   1. Ogólne statystyki rynku e-commerce, może duzi gracze – znaleźć jakiś raport
   2. Ogólne charakterystyki Brazylii – takie rzeczy jak liczba ludności/pkb per capita, struktura geograficzna
2. E-commerce
   1. Ogólne fakty – wartość rynku itd.
   2. Specyfika klienta (procent stałych klientów vs. Inne branże, charakterystyka zakupów itd.)
3. Churn prediction
   1. Case studies – szczególny nacisk na to, z jakich danych korzystano – siłą mojej pracy ma być różnorodność „rodzajów” danych wejściowych do modelu – dane transakcyjne, demograficzne, przestrzenne, tekstowe itd
   2. Metody (rozszerzenia klasycznych modeli ML pod to zadanie)
   3. Klasycznie churn analizuje się dla modelu subskrypcji – czyli większość klientów zostaje. W tym datasecie jest odwrotnie – większość klientów odchodzi. Do sprawdzenia, czy są papery „w drugą stronę” niż zazwyczaj przy churnie
4. Customer segmenatation
   1. Generalne informacje – po co się wykonuje, klasyczne podejścia (tu dużo można z „Principles of Marketing” Kotlera i ogólnie teorii marketingu)
5. Algorytmiczne podejście do customer segmentation
   1. Case studies – oparte o RFM, ale też inne metody
   2. Algorytmy mrówkowe
6. (Opcjonalnie) Papery o redukcji wymiaru danych transakcyjnych. Żeby móc połączyć te dane do głównego zbioru (1 rekord – 1 klient), trzeba przetworzyć dane transakcyjne i zredukować wymiarowość macierzy zakupów. Specyfika takich danych:
   * 1. Znacząco więcej kategorii niż produktów w 1 koszyku – rzadka macierz
     2. Tylko binarne zmienne
     3. Bardzo wiele zmiennych – tyle ile rodzajów produktów, nie do włożenia do modelu.
7. (Opcjonalnie) Reviews text mining
   1. Case studies z literatury
   2. Analiza sentymentu – ogólnie
   3. Topic modeling, w szczególności topic modeling recenzji

Prace nad modelem:

* Basic model z podstawowymi zmiennymi, GBM, XGB i Logistic (jest zrobione)
* + dane geo (jest zrobione)
* + dane o recenzjach (do wytrenowania modelu text mining, duża robota)
* + dane o koszyku zakupów (puścić PCA byłoby łatwe, ale już bardziej fancy metody już więcej pracy)
* Do dorobienia analiza istotności zmiennych – najbardziej basic to varImpPlot, ale można się pobawić DALEXem

Prace nad klastrowaniem:

* Na tą chwilę chyba trzeba zaparkować – ciekawsze rzeczy mogą wyjść dla danych o recenzjach i produktach w koszyku, z danych o czasie dostawy i wartości transakcji nic ciekawego nie wyciągnę
* PAM na tych samych danych co w modelowaniu
* Bardziej fancy metoda – np. algorytm mrówkowy. Jest zaimplementowane, nie jest puszczone na zbiorze

Co powinno być w części z modelowaniem?

* Opis podejścia dla każdej „części” danych
* Tabelka z modelami na podstawowych danych
* ROC dla train i test dla najlepszego modelu
* Hiperparamatery najlepsze w najlepszym modelu
* Bootstrap wyników AUC
* Variable importance plots
* DALEX plots – conditional stuff
* Wykresy wpływu lokalizacji geograficznej na wynik