Konspekt 02-12-2020

Table of Contents

#### O co chodzi w pracy? “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 ilościowej

1. Predykcja, jakie jest prawdopodobieństwo drugiego zakupu klienta (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
   * Ogólne statystyki rynku e-commerce, może duzi gracze – znaleźć jakiś raport
   * Ogólne charakterystyki Brazylii – takie rzeczy jak liczba ludności/pkb per capita, struktura geograficzna
2. E-commerce
   * Ogólne fakty – wartość rynku itd.
   * Specyfika klienta (procent stałych klientów vs. Inne branże, charakterystyka zakupów itd.)
3. Churn prediction
   * 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
   * Metody (rozszerzenia klasycznych modeli ML pod to zadanie)
   * 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
   * Generalne informacje – po co się wykonuje, klasyczne podejścia (tu dużo można z „Principles of Marketing" Kotlera i ogólnie teorii marketingu)
   * Algorytmiczne podejście do customer segmentation
   * Case studies – oparte o model RFM, ale też inne metody
   * Algorytmy mrówkowe
5. Redukcja 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:
   * Znacząco więcej kategorii niż produktów w 1 koszyku – rzadka macierz
   * Tylko binarne zmienne
   * Bardzo wiele zmiennych – tyle ile rodzajów produktów, nie do włożenia do modelu.
6. Text mining recenzji w formacie tekstowym
   * Case studies z literatury
   * Analiza sentymentu – ogólnie istniejące podejścia.
   * Topic modeling, w szczególności topic modeling recenzji
   * Połączenie analizy sentymentu i analizy tematu recenzji. Sentyment powie czy recenzja jest pozytywna, topic modeling o czym jest recenzja, ale dopiero z połączenia tych danych można dostać wartościową informację (np. klient jest zadowolony z ceny, ale nie zadowolony z dostawy).

#### Prace nad modelem

Sposoby

* Najprostszy model z podstawowymi zmiennymi, GBM, XGB i Logistic (jest zrobione, wstępne wyniki poniżej)
* Dane geo/demograficzne (jest zrobione, wstępne wyniki poniżej)
* Dane o recenzjach (do wytrenowania modelu text mining, potrzebny duży nakład pracy)
* dane o koszyku zakupów (metoda w stylu PCA jest łatwa do zastosowania, ale w literaturze można znaleźć inne sposoby redukcji wymiarów utworzone specjalnie pod dane transakcyjne. To pociąga za sobą więcej pracy)
* Analiza istotności zmiennych – najbardziej podstawowy to relatywna ważność zmiennych (zrobione), ale w dalszej kolejności warto też sprawdzić bardziej zaawansowane metody, na przykład wykresy ceteris paribus

#### Prace nad klastrowaniem

* PAM, k-means na tych samych danych co w modelowaniu - zrobione dla danych które przygotowałem do tej pory pod predykcję - wyniki są mierne.
* Myślę że warto wrócić do klastrowania dopiero wtedy, kiedy przetworzę dane o recenzjach i o koszykach produktów - obecnie dane które wkładałem do modelu są dosyć “zdepersonalizowane”. W metodach które testowałem (PAM, k-means, DBSCAN) dostaję 2 klastry - jeden większość klientów i drugi który jest “szumem” - głównie outliery.
* Bardziej zaawansowana metoda klastrowania – np. algorytm mrówkowy. Do pewnego stopnia zaimplementowałem metodę, ale zostało jeszcze trochę pracy tak aby działało tak dobrze jak standardowe funkcje do klastrowania.

#### Co powinno być w części z modelowaniem? - ogólny zarys

* Opis podejścia dla każdej „części" danych (tekstowe, geo itd.)
* Tabelka z modelami na podstawowych danych (wyniki, rodzaj modelu, na których danych był testowany)
* wykresy ROC dla dla najlepszego modelu, porównanie modeli
* Tabela z najlepszymi hiperparamaterami w najlepszym modelu
* Bootstrap wyników AUC - przedziały ufności
* Wykresy relatywnej ważności zmiennych w najlepszym modelu
* Explainable AI - z paczki DALEX, np. Wykresy wpływu lokalizacji geograficznej na wynik z modelu.

#### Modelowanie

Tutaj wrzuciłem bardzo podstawowy opis modelowania które wykonałem do tej pory - oczywiście do dorobienia porządne opisy, na tą chwilę tylko wrzuciłem pojedyncze wykresy/informacje które myślę że powinny się znaleźć w jakiejś formie w końcowej pracy

##### Dataset preparation description - conspect

Ways of joining different types of information to the main dataset:

1. Basic features
   1. Order value, review 1-5 etc.
   2. no special preparation - this is main table of the dataset
2. Geographic info
   1. Demographic data from sttistical office
   2. Different resolution than main dataset - description of data transformation
   3. TODO - opis jakie dane dokładnie
3. Transaction data - items in the orders

Transformations used on the joined dataset from the previous spot:

* On demographic data - lots of features (30), can be mapped into way smaller no. of features as proxies (eg. rich vs. poor regions). Use PCA with number of components to catch.

##### Methods description conspect:

* Train-test 70/30 split of the dataset
* Hyperparameter search using 3-fold Cross-validation on training dataset, grid search of possible combinations
* Used models: XGBoost, GBM, Logistic Regression
* Upsampling for handling class imbalance (97/3) - only on training dataset
* Metric used - AUC, as there is class imbalance
* Bootstrap of test set with 100 replcations to get standard errors of AUC estimation
* Relative variable importance of variables in each model

Powyższe zrobiłem dla modeli które wytrenowałem do tej pory. Nadal do zrobienia:

* Explainable artificial intelligence - cateris paribus plots and others (DALEX package)
* Some features selection methods - to do on the most enhanced dataset with all possible features - now it doesn’t make sense

##### Results of modeling:

Models tried so far:

Models on 6 basic variables:

* Model 1 - GBM on standard hyperparameters with upsampling (AUC test 0.5921)
* Model 2 - GBM on standard hyperparameters without upsampling (Auc test 0.5899)
* Model 3 - Logistic with upsampling (Auc test 0.5575)
* Model 4 - Logistic without upsampling (Auc test 0.5563)
* **Model 5 - XGB extensive hyperparameters search with upsampling (Auc test 0.6159 - best)**

Models on 6 basic variables + demographic data:

* **Model geo 1- XGB extensive hyperparameters search with upsampling on geo data (Auc test 0.5546 - worse than model without demographic info but still good)**
* Model geo 2 - XGB extensive hyperparameters search with upsampling on geo data and with PCA on 8 components (Auc test 0.5289)

TODO: - create table and better summmary of above, list variables, table with tested hyperparameters for all (best?) models

ROC comparison of XGB models with and without geo features:

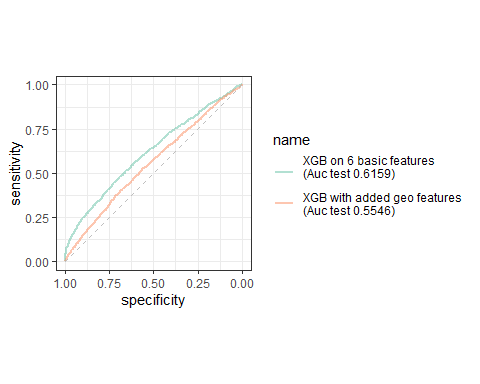
list(  
 'XGB on 6 basic features \n(Auc test 0.6159)\n' = calc\_roc(model5, to\_model\_test),  
 'XGB with added geo features \n(Auc test 0.5546)' = calc\_roc(model\_geo1, to\_model\_test\_geo)) %>%  
 pROC::ggroc(alpha = 0.5, linetype = 1, size = 1) +  
 geom\_segment(aes(x = 1, xend = 0, y = 0, yend = 1),   
 color = "grey",   
 linetype = "dashed") +  
 theme\_bw() +   
 coord\_fixed() +  
 scale\_color\_brewer(palette = "Set2")

## Setting levels: control = 0, case = 1

## Setting direction: controls > cases

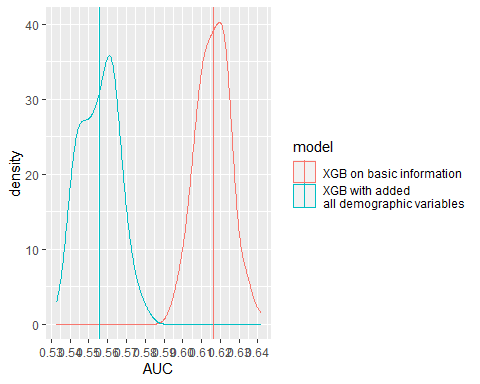
## Setting levels: control = 0, case = 1

## Setting direction: controls > cases



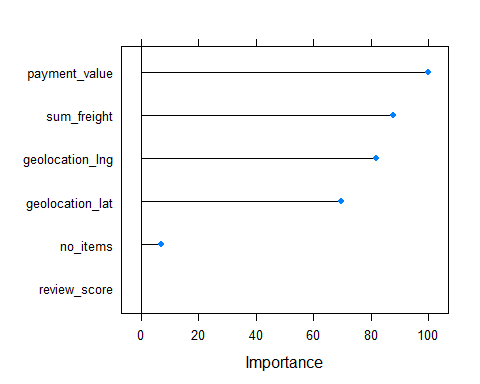
Comparison of bootstraped AUC distribution on test set for best model for geo data included and not:

rbind(  
 tibble(AUC = out\_roc5, model = 'XGB on basic information', mean\_auc = mean(out\_roc5)),  
 tibble(AUC = out\_roc\_geo1, model = 'XGB with added \nall demographic variables', mean\_auc = mean(out\_roc\_geo1))  
) %>%  
 ggplot(aes(x = AUC, color = model)) +  
 geom\_density() +  
 geom\_vline(aes(xintercept = mean\_auc, color = model)) +  
 scale\_x\_continuous(breaks = seq(0.5, 0.65, 0.01))



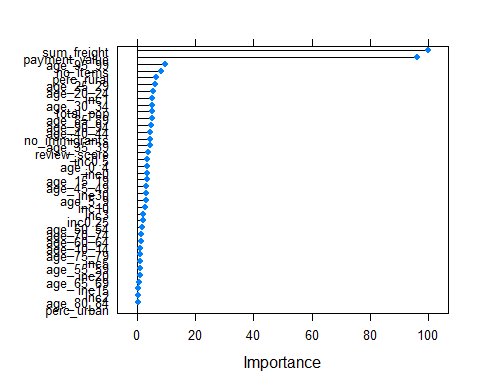
Most important features in XGB model - best model of all tested:

plot(varImp(model5))



Most important features in XGB model on geo data - demmographic features are less important than these connected with demography:

plot(varImp(model\_geo1))



Most important features in Logistic Regression model - most simple one:

plot(varImp(model3))

