Karta kredytowa - dla małego i dużego

autor: Paulina Kulczyk

Grudzień 2022

Contents

1 Wstęp **3**

2 Dane **3**
2.1 Źródło . . . 3
2.2 Import . . . 3
2.3 Opis . . . 3

3 Braki danych **10**

4 Niezbilansowane obserwacje **11**

5 Przetwarzanie danych **11**

6 Tworzenie zbioru treningowego i testowego **11**
6.1 Analiza łączy . . . 11
6.2 Wagi zmiennych . . . 13
6.3 Korelacja zmiennych . . . 14
6.4 Wykresy zliczeń . . . 14
6.5 Podsumowanie danych treningowych . . . 18

7 Wybór modeli i uczenie **18**
7.1 Inicjalne modele - z "defaultowymi" hiperparametrami . . . 19
7.2 Modele o innych wartościach hiperparametrów . . . 21
7.3 Walidacja . . . 22
7.4 Tworzenie ensemblingów . . . 24
7.5 Podsumowanie modeli . . . 27

8 Redukcja wymiarowości i testowanie na zredukowanych danych najlepszych modeli **27**

9 Podsumowanie **30**

10 Bibliogradia **30**

1 Wstęp

W obecnych czasach coraz więcej osób korzysta z kart kredytowych, banki także zachęcają swoich klientów do tej usługi. Jednak wciąż jest pewna grupa osób, które nie chcą się przekonać do tego rozwiązania, ponadto niektórzy po próbie decydują się na rezygnację z posiadania karty. Utrata posiadaczy kart bankowych jest dla banku dużą stratą, dlatego menadżerowie robią co mogą, aby klienci, którzy zdecydowali się na tą usługę nie zrezygnowali z niej.

Poniższy projekt jest skierowany właśnie do takich menadżerów banków. Ma on na celu predykcję jaki typ klientów chce odstąpić od korzystania z karty kredytowej. W tym celu zbadam tabelę zawierającą informację o klientach banku oraz zbuduję model decyzyjny. Projekt pomoże w ten sposób właścicielom banków na lepsze dostosowanie ofert do klientów oraz w przeciwdziałaniu utraty użytkowników kart kredytowych.

2 Dane

2.1 Źródło

Dane do projektu charakteryzujące użytkowników kart kredytowych zostały pobrane z serwisu Kaggle (plik: 'BankChurners.csv'). Są one dostępne pod adresem:
https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers?resource=download.

2.2 Import

Plik 'BankChurners.csv' zaimporotwałam do projektu utworzonego w programie SAS Enterprise Guide. Tam też dokonałam wstępnej analizy danych oraz podziału na zbiory testowe i treningowe.

2.3 Opis

Zaimportowane do programu SAS Enterprise Guide dane przetrzymywałam w postaci tabeli sasowej o nazwie 'DF'. Poniżej prezentuję pierwsze 5 wierszy tej tabeli:

Obs	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon
1	768805383	Existing Customer	45	M	3	High School	Married	$60K - $80K	Blue	39	5	1	3
2	818770008	Existing Customer	49	F	5	Graduate	Single	Less than $40K	Blue	44	6	1	2
3	713982108	Existing Customer	51	M	3	Graduate	Married	$80K - $120K	Blue	36	4	1	0
4	769911858	Existing Customer	40	F	4	High School	Unknown	Less than $40K	Blue	34	3	4	1
5	709106358	Existing Customer	40	M	3	Uneducated	Married	$60K - $80K	Blue	21	5	1	0

Obs	Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal	Avg_Open_To_Buy	Total_Amt_Chng_Q4_Q1	Total_Trans_Amt	Total_Trans_Ct	Total_Ct_Chng_Q4_Q1	Avg_Utilization_Ratio	Naive_Bayes_Classifier_Attritio	VAR23
1	3	12691	777	11914	1.335	1144	42	1.625	0.061	0.000093448	0.99991
2	2	8256	864	7392	1.541	1291	33	3.714	0.105	0.000056861	0.99994
3	0	3418	0	3418	2.594	1887	20	2.333	0	0.000021081	0.99998
4	1	3313	2517	796	1.405	1171	20	2.333	0.76	0.00013366	0.99987
5	0	4716	0	4716	2.175	816	28	2.5	0	0.000021676	0.99998

Jak widać zawiera ona bardzo dużo zmiennych. Przy pomocy procedury contents sprawdziłam liczbę obserwacji oraz typy danych. Oto wyniki:

Observations	10127
Variables	23

W naszej tabeli sasowej mamy 23 zmienne oraz 10127 obserwacji.

#	Variable	Type
2	Attrition_Flag	Char
16	Avg_Open_To_Buy	Num
21	Avg_Utilization_Ratio	Num
1	CLIENTNUM	Num
9	Card_Category	Char
13	Contacts_Count_12_mon	Num
14	Credit_Limit	Num
3	Customer_Age	Num
5	Dependent_count	Num
6	Education_Level	Char
4	Gender	Char

8	Income_Category	Char
7	Marital_Status	Char
12	Months_Inactive_12_mon	Num
10	Months_on_book	Num
22	Naive_Bayes_Classifier_Attritio	Num
17	Total_Amt_Chng_Q4_Q1	Num
20	Total_Ct_Chng_Q4_Q1	Num
11	Total_Relationship_Count	Num
15	Total_Revolving_Bal	Num
18	Total_Trans_Amt	Num
19	Total_Trans_Ct	Num
23	VAR23	Num

Dane opisujące zmienne występują zarówno w postaci numerycznej jak i tekstowej. Zmienną celu jest

- Attrition_Flag - informuje ona nas o tym czy użytkownik nadal korzysta z usługi karty kredytowej czy zrezygnował z niej

Pozostałe zmienne to:

- Card_Category - opisująca rodzaj karty jaką użytkownik posiada/posiadał
- Education_Level - wykształcenie
- Gender - płeć
- Income_Category - przychody
- Marital_Status - status cywilny
- Avg_Open_To_Buy - środki, które pozostały do wykorzystania na karcie (różnica między limitem na karcie a wydanymi pieniędzmi)
- Avg_Utilization_Ratio - współczynnik wyliczany poprzez stosunek sumy wydatków na karcie i pieniędzy wypłaconych z kart do sumy limitu wydatków na karcie i limitu wypłacania pieniędzy z karty $(credit\_card\_spent + money\_withdrawal)/(Total\_available\_limit\_credit\_card\_ + Total\_money\_withdrawal\_limit)$
- CLIENTNUM - numer klienta
- Contacts_Count_12_mon - liczba kontaktów banku z klientem, np. w celach reklamowych w ciągu roku
- Credit_Limit - limit kredytowy
- Customer_Age - wiek
- Dependent_count - liczba osób będących na utrzymaniu użytkownika
- Months_Inactive_12_mon - liczba miesięcy nie korzystania z karty kredytowej w ciągu roku
- Months_on_book - liczba okresów interakcji użytkownika z bankiem w ciągu roku
- Total_Amt_Chng_Q4_Q1 - pokazuje o ile zwiększyły się wydatki dokonane za pomocą karty kredytowej przez klienta w 4 kwartale w stosunku do 1 kwartału
- Total_Ct_Chng_Q4_Q1 - pokazuje o ile zwiększyła się liczba transakcji dokonanych za pomocą kart kredytowej przez klienta w 4 kwartale w stosunku do 1 kwartału
- Total_Relationship_Count - Liczba usług w banku, które posiada klient - np. karta, konto oszczędnościowe
- Total_Revolving_Bal - suma debetów na koncie w przeciągu korzystania z karty, mówi nam, którzy klienci 'lubią się' zadłużać

- Total_Trans_Amt - liczba wydatków całkowitych z karty w ciągu roku
- Total_Trans_Ct - liczba transakcji wykonanych za pomocą karty w ciągu roku
- Naive_Bayes_Classifier_Attritio
- VAR23

Dwie ostatnie zmienne zawierają dane liczbowe. Na stronie serwisu Kaggle otrzymujemy tylko ich skrócony opis 'Naive Bayes'. Taki opis nic nam nie mówi o zmiennych, może jedynie sugerować, że są wynikami wcześniejszego szkolenia jakiegoś naiwnego klasyfikatora Bayes'owskiego. Te zmienne zatem nic nie wnoszą do charakterystyki naszych klientów banków, ponadto, jak sugeruje sam autor bazy danych, mogą one negatywnie wpłynąć na wyniki naszych badań. Rozsądne zatem wydaje się nie branie ich pod uwagę w dalszym badaniu i ich porzucenie.

Wielkim atutem tej bazy danych jest brak braków danych (sprawdzone za pomocą funkcji CMISS i NMISS). Przyjrzyjmy się jednak dokładniej wartościom tabeli. Sprawdźmy jakie wartości przyjmują zmienne tekstowe. Pozwoli nam to na wykrycie obserwacji bezsensownych (jak na przykład pytanie o ciąże i odpowiedź pozytywna dla użytkownika o płci męskiej) oraz zbadanie rozkładu zmiennych:

Variable	Label	Value	Frequency Count	Percent of Total Frequency
Attrition_Flag		Existing Customer	8500	83.9340
		Attrited Customer	1627	16.0660

Jak widać niestety nasz cel jest nierównomiernie rozłożony pomiędzy obserwacjami. Ponad 83 % obserwacji dotyczy klientów obsługujących nadal kartę kredytową.

Variable	Label	Value	Frequency Count	Percent of Total Frequency
Card_Category		Blue	9436	93.1767
		Silver	555	5.4804
		Gold	116	1.1455
		Platinum	20	0.1975

Wśród naszych obserwacji mamy 4 rodzaje kart kredytowych dostępnych dla użytkowników banku. Także i tu dominuje jedna obserwacja (karta 'blue' stanowi 90 % obserwacji). Warto jednak żauważyć, że karta 'blue' jest podstawową kartą, którą na początek w ramach zwykłej usługi karty kredytowej dostaje każdy użytkownik. Pozostałe rodzaje kart stanowią wersje premium i należy za nie dopłacić lub można je dostać, jeśli się jest zaufanym/długoletnim klientem.

Variable	Label	Value	Frequency Count	Percent of Total Frequency
Income_Category		Less than $40K	3561	35.1634
		$40K - $60K	1790	17.6755
		$80K - $120K	1535	15.1575
		$60K - $80K	1402	13.8442
		Unknown	1112	10.9805
		$120K +	727	7.1788

Dane przychodów dla wartości między 40K a 120K rozkładają się w miare równomiernie. Wartości występujące częściej to obserwacje, dla których zarobki użytkowników kart wynosiły poniżej 40K, a wartości występujące rzadziej przypadają na użytkowników zarabiających powyżej 120K. Jednak patrząc na statystyki zarobków w Stanach Zjednoczonych, taki rozkład jest zgodny z występującym w świecie rzeczywistym.

Dodatkowo należy zauważyć, że w tej kategorii odnotowano obserwacje 'Unknown' - nie znamy zarobków użytkownika. Takich obserwacji jest aż 10,98 procent.

Variable	Label	Value	Frequency Count	Percent of Total Frequency
Education_Level		Graduate	3128	30.8877
		High School	2013	19.8776
		Unknown	1519	14.9995
		Uneducated	1487	14.6835
		College	1013	10.0030
		Post-Graduate	516	5.0953
		Doctorate	451	4.4534

Wartości dotyczące zmiennej opisującej poziom edukacji mają podobny rozkład do wartości zmiennej Income_Category. Tu wartością dominującą jest obserwacja 'Graduate'. Natomiast w mniejszym stopniu niż pozostałe pojawiają się zmienne 'Post-Graduate' oraz 'Doctorate'. Tutaj również pojawiają się obserwacje 'Unknown'.

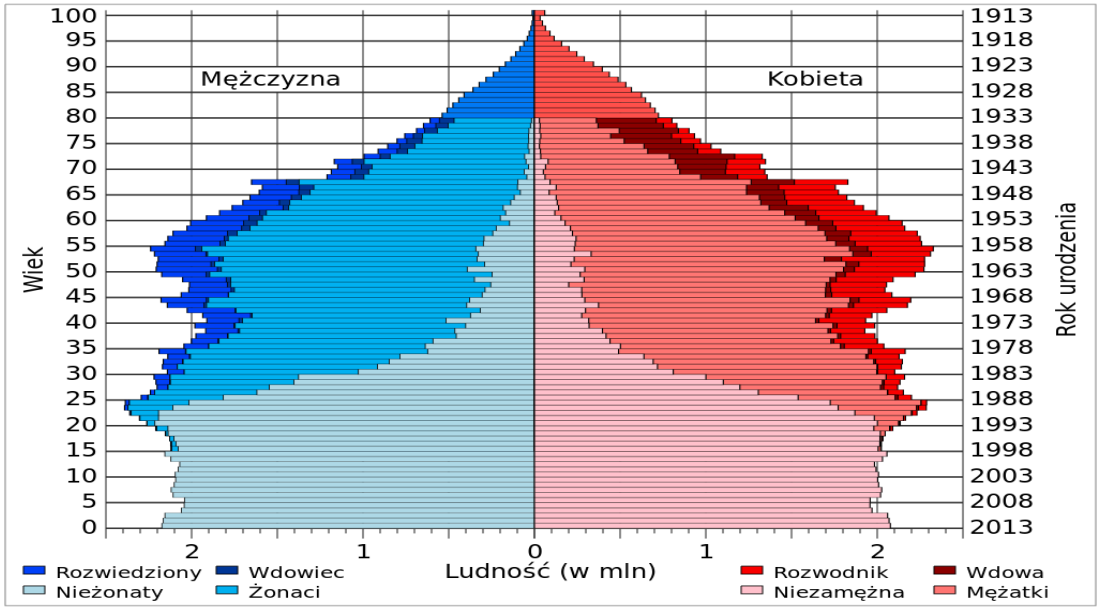
Variable	Label	Value	Frequency Count	Percent of Total Frequency
Marital_Status		Married	4687	46.2822
		Single	3943	38.9355
		Unknown	749	7.3961
		Divorced	748	7.3862

Najcześciej występujące obserwacje w kolumnie 'Martial_Status' to: 'Single' lub 'Married'. W mniejszym stopniu pojawia się obserwacja 'Divorced'. W ilości zbliżonej do częstości występowania zmiennej 'Divorced' pojawia się także zmienna 'Unknown'.

Variable	Label	Value	Frequency Count	Percent of Total Frequency
Gender		F	5358	52.9081
		M	4769	47.0919

Wsród naszych obsereacji ilość kobiet w stosunku do ilości mężczyzn jest bliska 1:1. Oznacza to, że klasy są dobrze reprezentowane przez tą zmienną.

Zatrzymajmy się tu jednak na chwilę i spójrzmy z ciekawości na demografię Stanów Zjednoczonych:

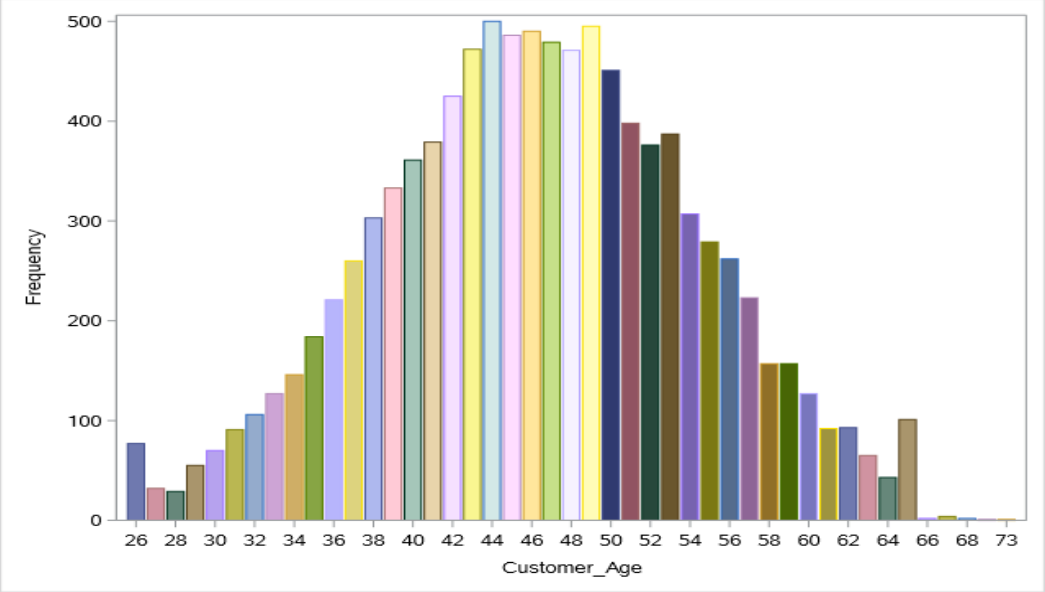


grafika pobrana z Wikipedii

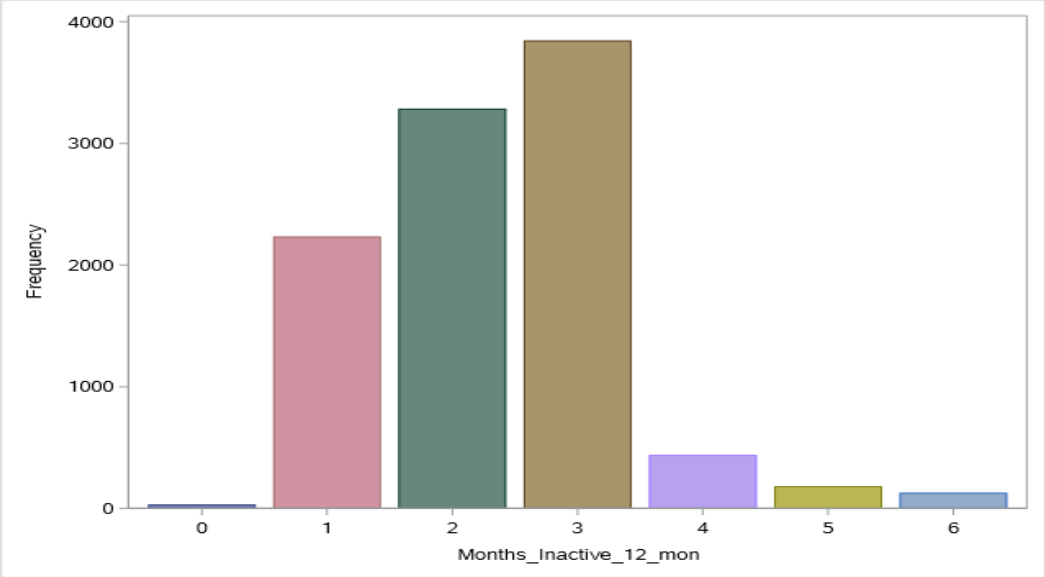
Jak widzimy w Stanach stosunek mężczyzn do kobiet jest również bliski 1:1. Miło widzieć, że nasze dane tak dobrze oddają rzeczywistość :).

Powyżej przyjżeliśmy się wszystkim zmiennym kategorycznym. Aż 3 zmienne tego typu posiadają obserwacje 'unknown'. Zauważmy, że jest to pewnego typu brak danych. Sprawdźmy czy obserwacje te pokrywają się na wysokości tych samych wierszy? Jeśli tak to dobrym pomysłem do dalszej analizy danych oraz konstruowania modelu wydaje się usunięcie tych wierszy, ponieważ mogą być one "fałszywymi obserwacjami". Po zaaplikowaniu polecenia WHERE, widzimy że wierszy, w których wszystkie 3 zmienne mają wartości 'Unknown' jest tylko 7. Ponadto wszytkie obserwacje są przyporządkowane dla kobiet, których było nieznacznie więcej w naszym zbiorze. Warto też zauważyć, że 6 z 7 tych obserwacji jest dla 'Existing Customer', których mieliśmy znacznie więcej w naszym zbiorze danych. Tym bardziej sugeruje to, że usunięcie ich nie wpłynie znacząco na istotność naszych danych, a może ulepszyć funkcjonowanie modelu. Pozostaje pytanie jak postapić z pozostałymi obserwacjami zawierającymi wartość 'Unknown'. Zauważmy, że ta wartość występuje przy zmiennych dotyczących statusu matrymonialnego, zarobków i edukacji - zmienne te powszechnie są uważane za wrażliwe w życiu publicznym. Stąd też w przyszłości będzie się zdażać, że nasi klienci nie będą chcieli odpowiedzieć na pytania dotyczące tych kategorii. Algorytm powinien nauczyć się więc radzić z takimi obserwacjami, stąd obserwacji, gdzie wartość 'Unknown' nie występuje we wszystkich 3 kategoriach nie będziemy usuwać.

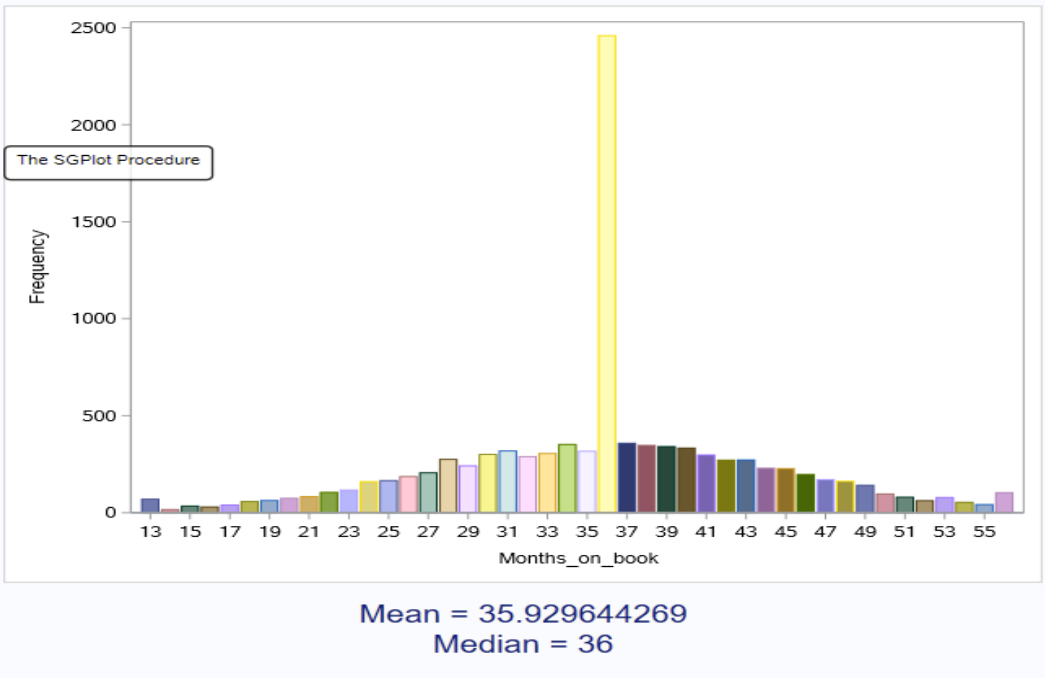
Teraz przystąpmy do badania zmiennych numerycznych. Ich analizy będzimy dokonywać na podstaweie tabeli sasowej 'df_fin' - jest to tabela, która nie posiada już tych 7 obserwacji, gdzie zmienna 'Unknown' występowała przy wszystkich 3 kategoriach
Rozkłady zmiennych liczbowych:



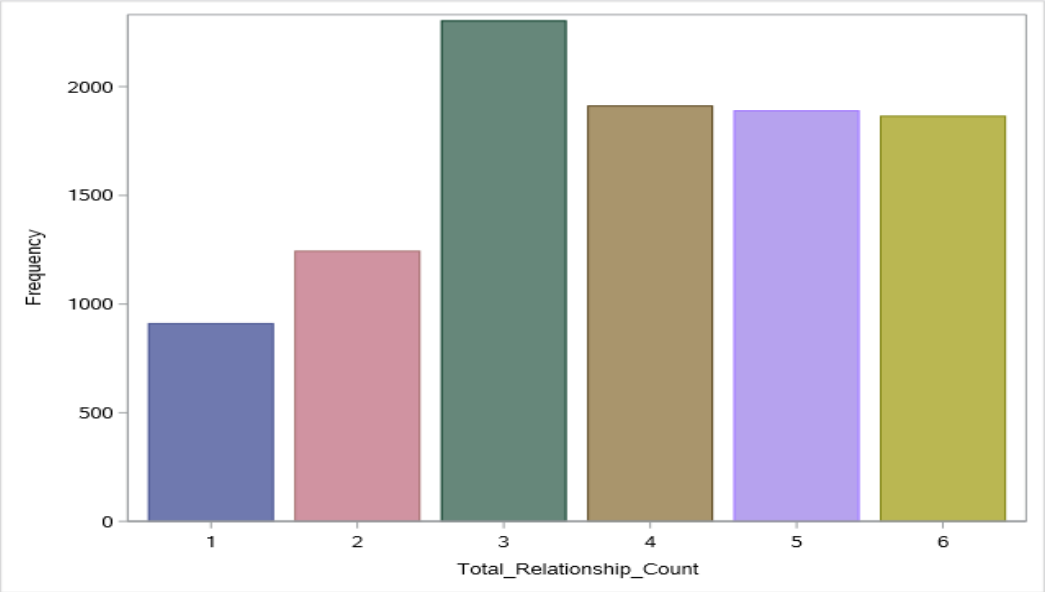
W tabeli sasowej 'df_fin' posiadamy obserwacje dotyczące klientów między 26 a 73 rokiem życia. Najwięcej obserwacji dotyczy użytkowników w średnim wieku.



Powyższy wykres dotyczy liczby miesięcy niekorzystania z karty kredytowej w ciągu 12 miesięcy. Niektórzy użytkownicy potrafili z niej niekorzystać aż przez 6 miesięcy!

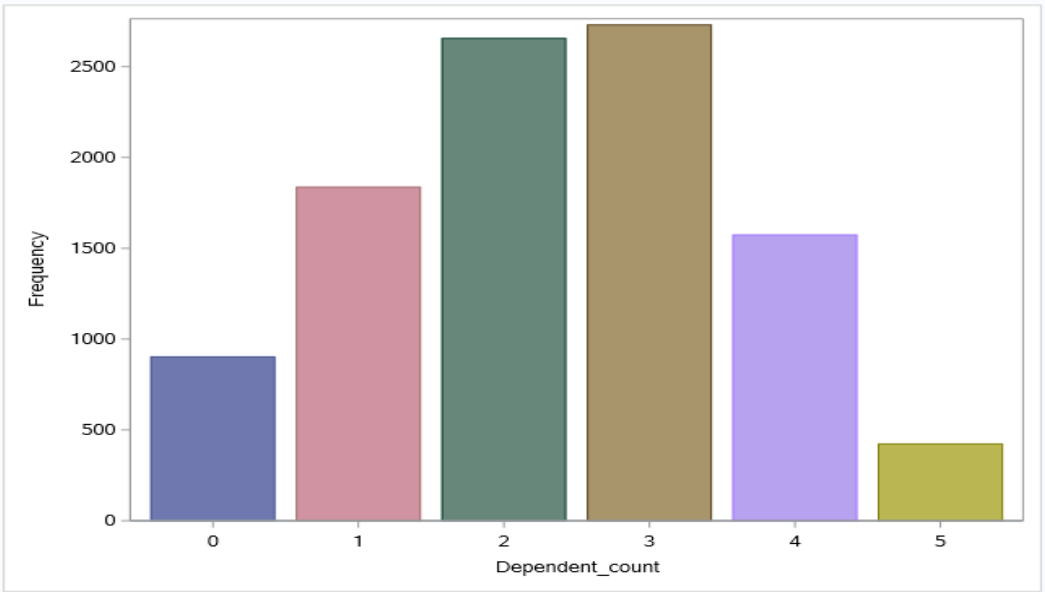


Użytkownicy kart kredytowych wchodzili w ciągu roku w inerakcje z bankiem od 13 do 56 razy . Jednak zdecydowanie dominującą wartością jest liczba 36 kontaktów. Występuje ona prawie 2500 razy, gdzie pozostałe wartości występują średnio poniżej 100 razy. Zaskakujący jest ten nierównomierny rozkład! Zauważmy, że 36 razy oznacza średnio 3 kontakty w ciągu miesiąca. Może to oznaczać, że pracownicy banku mają zadnaie kontaktować się z klientami 3 razy na miesiąc. Możliwe, że jest to zabieg przeciwdziałania odchodzenia od posiadania karty.



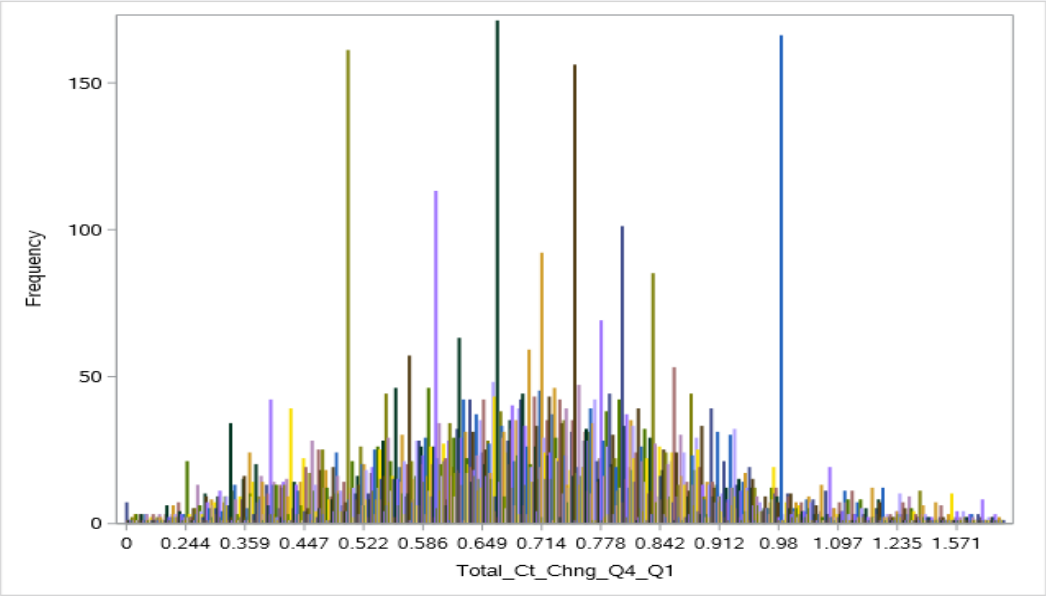
Mean = 3.812055336
Median = 4

Klienci zawarci w naszych obserwacjach korzystają z 1 do 6 usług oferowanych przez bank.



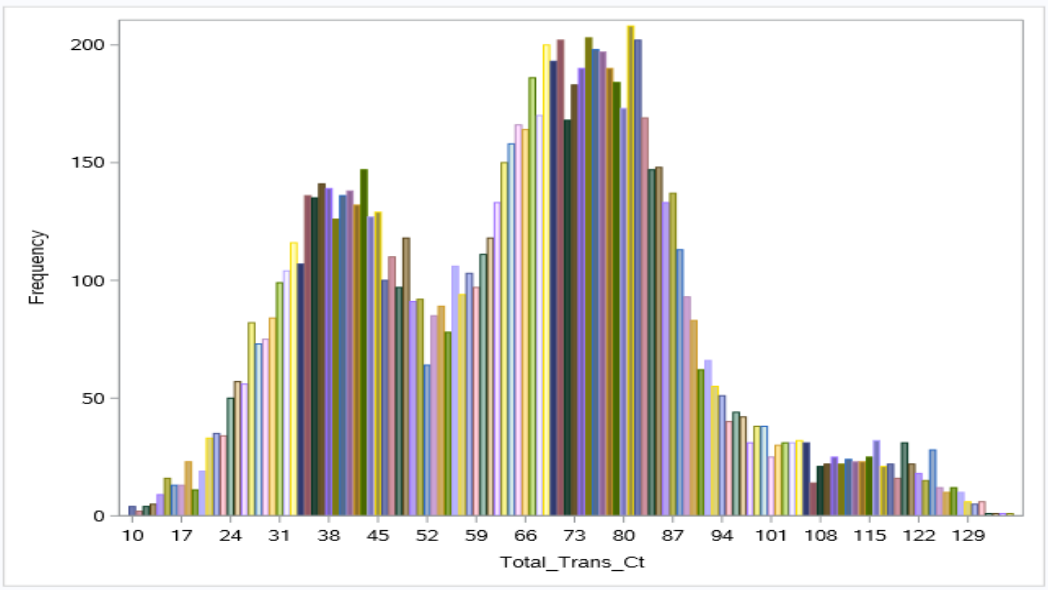
Mean = 2.3462450593
Median = 2

Dodatkowo nasi klienci mają na utrzymaniu od 0 do 5 osób



Mean = 0.712187253
Median = 0.702

Powyższy wykres ilustruje jak zwiększyła się liczba transakcji dokonanych za pomocą karty kredytowej przez klienta w 4 kwartale w stosunku do 1 kwartał. Widać, że niektórzy użytkownicy stabilnie nie zmieniają ilości transkakcji, niektorzy zaś zwiększają liczbę transakcji o ponad 100 %.



Mean = 64.86027668
Median = 67

Użytkownicy opisani w naszej tabeli sasowej potrafili wykonać od 10 do ponad 130 transakcji kartą. Najwięcej jednak obserwacji opisuje przypadki miedzy 66 a 83 transakcjami.

3 Braki danych

Na wstępie zauważyliśmy, że nasza tabela nie posiada braków danych. Jednak po dogłebnej anlizie, którą przeprowadziliśmy powyżej, zauważyliśmy, że w naszych obserwacjach pojawiają się rekordy 'Unknown', które "na pierwszy rzut oka" moglibyśmy potraktować jako zmienne, które można uznać za

wartości brakujące. Jednakże, zwróciliśmy uwagę, że dotyczą one zmiennych opisujących płeć, dochody i wykształcenie. Są to zmienne powszechnie uważane w społeczeństiwe za dane wrażliwe. Doszliśmy do wniosku, że rozsądne się wydaje pozostawienie tych wartości jako odzielnej kategorii. Takie podejście jest dobre, ponieważ możemy się spodziewać, że w przyszłości będą pojawiać się obserwacje 'Unknown' w tych kolumnach i model musi być przygotowany na nie, oraz umieć badać czy miedzy nimi a zmienną przewidywaną jest jakaś zależność.

4 Niezbilansowane obserwacje

W naszej tabeli sasowej (po usunięciu 7 rekordów zawierających dane 'Unknown' jednocześnie we wszytskich trzech kolumnach dotyczących płciu, zarobków i wykształcenia) jest zaobserwowanych 1626 klientów, którzy zrezygnowali z posiadania karty bankowej oraz 8494 klientów, którzy wciąż ją posiadają. Jak widzimy mamy nierównomierny rozkład danych pomiędzy obserwacjami pozytywnymi i negatywnymi. Taki rozkład danych dla zmiennej celu może prowadzić do złego wyuczenia modelu. Rozsądnym rozwiązaniem w tym przypadku wydaję się zrównoważenie danych - dokonanym tego poprzez losowe usunięcie nadmiarowych obserwacji poztywnych (dokładny opis tego procesu znajduje się w dwóch następnych rodziałach).

5 Przetwarzanie danych

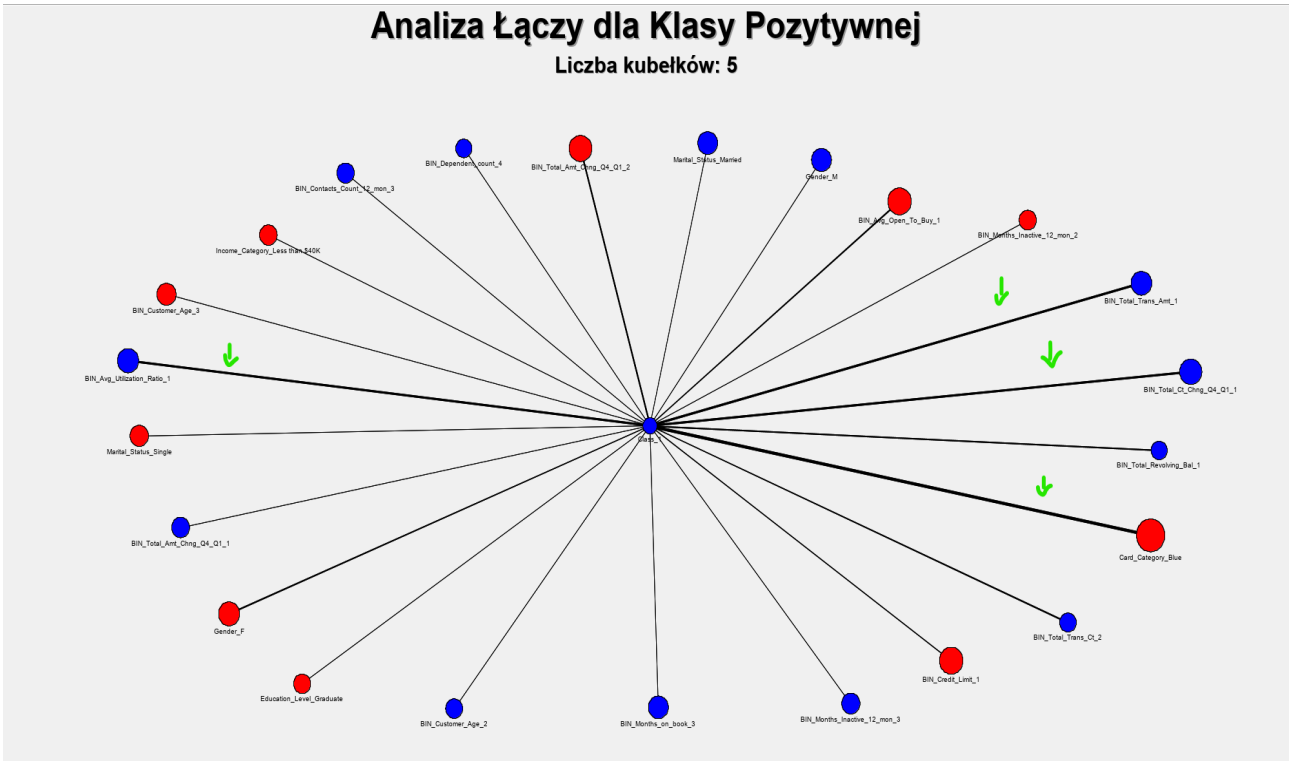
W czasie zapoznawania się z danymi dokonywałam różnych transformacji naszej pierwotnej tabeli sasowej o nazwie 'DF'. Pierwszą operacją, której dokonałam było porzucenie kolumn: Naive_Bayes _Classifier_Attritio i VAR23. Dokonałam tego za pomocą komendy drop (program 'porzucenie zmiennych'). Następnie dokonałam usunięcia wzpomnianych wyżej 7 obserwacji, które zawierały same dane 'Unknown' przy kategorii płeć, zarobki i wykształcenie. Uzasadnieniem dla tego procesu było to, że te odpowiedzi mogły być 'oszukane' (program 'obcinanie_obs_unknown'). Kolejnym krokiem było przyporządkowanie zmiennej celu wartości binarnych. W tym celu utworzyłam nową zmienna 'Class', która przyjmowała wartości 1 dla klientów, którzy zrezygnowali z karty i 0 dla tych, którzy nadal ją posiadają. Dokonałam takiego doboru wartości, ponieważ klienci, którzy rezygnują z usługi są dla nas "ważniejsi" do zaobserwowania (stanowią wartość alarmującą), dlatego lepiej, żeby przy trenowaniu modelu i jego interpretacji stanowili klasę pozytywną. (program 'tworzenie klas'). Ostatnią zmianą było stworzenie za pomocą procedury 'surveyselect' zmiennej 'Selected', która posłuży nam przy opisanym w następnej sekcji procesie tworzenie zbioru treningowego i testowego. Przyjmowała ona wartości 1 i 0 w stosunku 4:1. Wartości były przypisywane do danej obserwacji losowo.

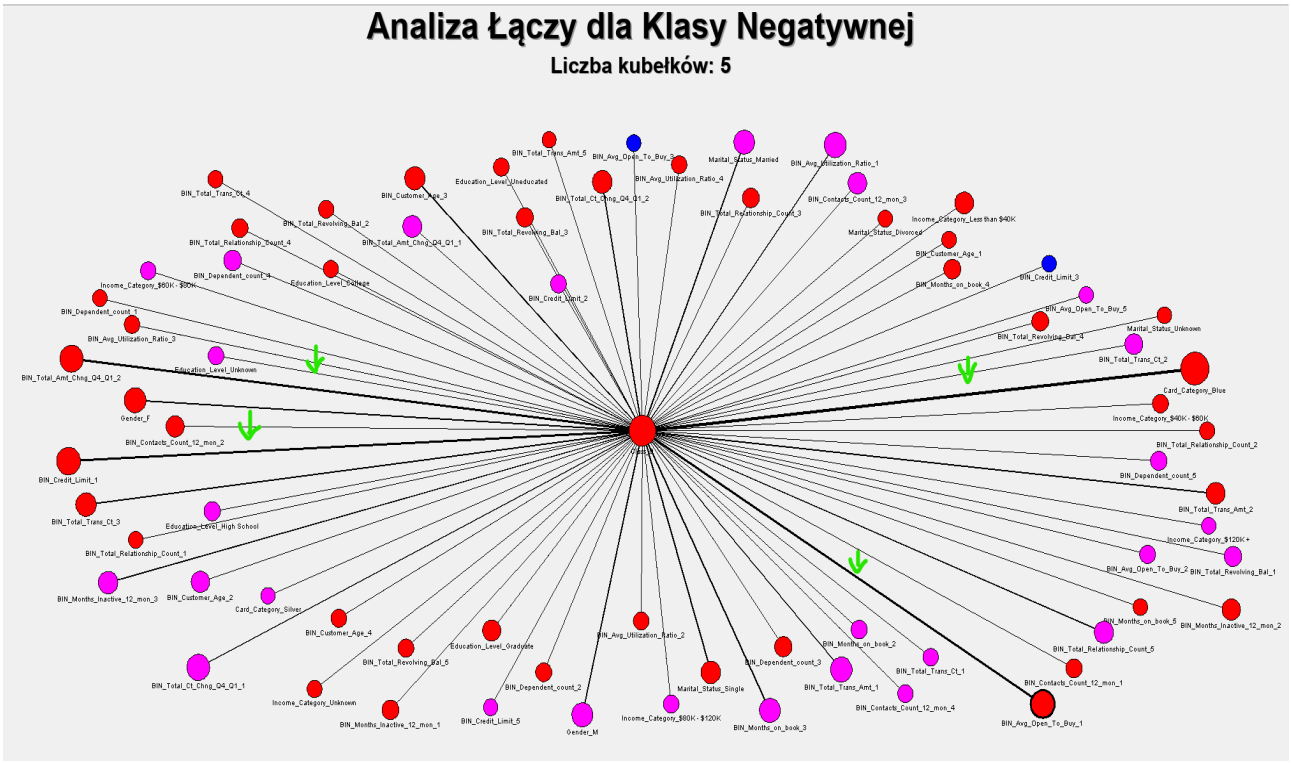
6 Tworzenie zbioru treningowego i testowego

W programie SAS Interprise Guide dokonałam podziału tabeli sasowej 'fin_select' na zbiór treningowy i testowy ('df_final_train' i 'df_final_test'), w stosunku 4:1 (przydział 80% obserwacjiu do zbioru treningowego). Podziału tego dokonałam na podstawie wartości przyjmowanych przez opisaną w poprzednim rozdziale zmienną 'Selected'. Do zbioru treningowego trafiły obserwacje, gdzie zmienna 'Selected' przyjmowała wartość 1, natomiast do zbioru testowego pozostałe. Następnie w celu zrównoważenia wystąpień zmiennej celu stworzyłam 10 mniejszych zbiorów treningowych. Powstały one poprzez losowe wybranie 1800 obserwacji spośród 6775 obserwacji negatywnych ze zbioru 'df_final_train'. (przy pomocy pomocniczej tabeli sasowej 'only_existing_cus' zawierającej same obserwacje negatywne - 'Existing Customer'). Obesrwacji, dla których zmienna celu przyjmuje wartość pozytywną w zbiorze df_final_train nie modyfikujemy. W ten sposób dostajemy 10 zbiorów danych treningowych posiadających 1321 obserwacji pozytywnych i 1800 obserwacji negatywnych. Dodatkowo obserwacje w tych zbiorach mieszamy, aby nie miało to wpływu na przewidywania modelu. W ostateczności dostajemy tabele sasowej: 'df_final_shuffled1',..., 'df_final_shuffled10', które stanowią 'mniejsze zbiory treningowe'.

6.1 Analiza łączy

Przed przystąpieniem do wyboru i uczenia modelu dokonujemy analizy łączy na tabeli sasowej 'df_final_train'. Czy możemy zauważyć jakieś ciekawe powiązania?





Powyższe grafiki przedstawiają rezultaty operacji wykonanych przez węzeł 'Analiza łączy'. Szerokość połączenia odzwierciedla siłe powiązania pomiędzy zmiennymi, wielkość koła - częstość wystąpień zmiennej w zbiorze, kolor - klaster, do którego zmienna należy. Zielonymi strzałkami zostały oznaczone połączenia o największym znaczeniu. Mając te informacje na uwadze zobaczmy jakie wnioski możemy wyciągnąć z powyższych grafów.

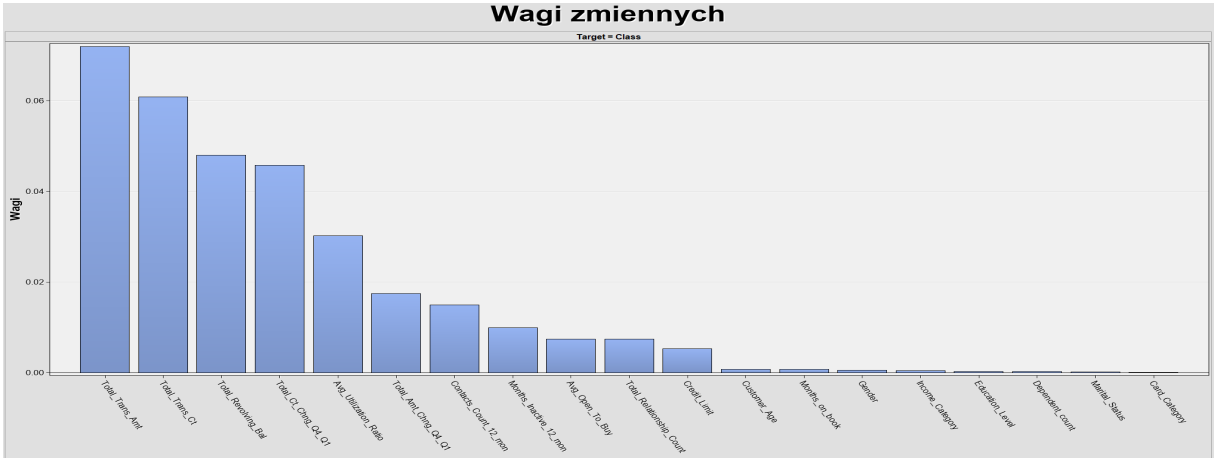
- Wartość klasy równa 1 - Attrited Customer - występuje często w powiązaniu z:

  - niebieską kartą (jednak pamiętajmy, że ona zdecydwanie dominowałą nad innymi wartościami zmiennej Card_Category, na co sugeruje także wielkość koła)
  - pewnymi wartościami zemiennej Total_Trans_Amt
  - pewnymi wartościami zmiennej Total_Ct_Chng_Q4_Q1
  - pewnymi wartsciami zmiennej Total_Avg_Utilization_Ratio
- Wartość klasy równa 0 - Existing Customer - występuje często w powiązaniu z:
  - niebieską kartą (tutaj ta sama uwaga co wyżej)
  - pewnymi wartościami zmiennej Avg_Open_To_Buy
  - pewnymi wartściami zmiennej Credit_Limit
  - pewnymi wartściami zmiennej Total_Amt_Chng_Q4_Q1

Zauważmy, że pomijając rodzaj karty, inne zmienne mają silne powiązanie z klasą pozytywną a inne z klasą negatywną. Może się okazać, że te zmienne odegrają dużą rolę w podejmowaniu decyzji przez model.

6.2 Wagi zmiennych

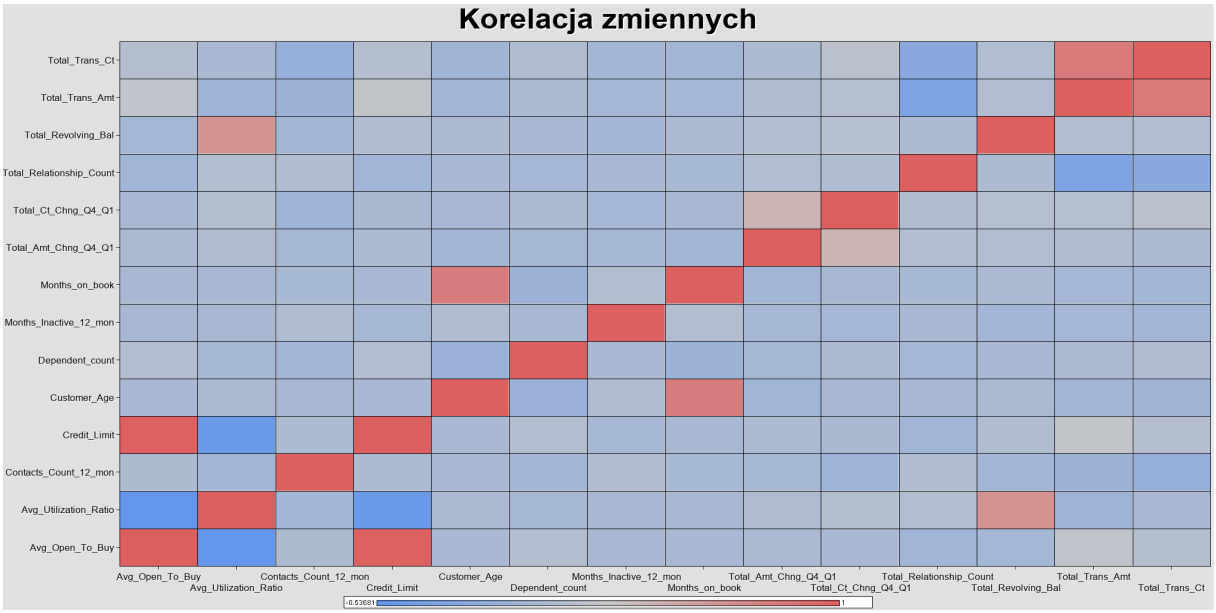
Na proces uczenia modelu duży wpływ mają wagi zmiennych. W programie Sas Enterprise Miner dostępny jest węzeł 'Eksploracja Statystyk', który pozwala na zbadanie "ważności" zmiennych. Zobaczmy, jakie wagi przypisał on dla naszych zmiennych:



Według programu największą wagę należy przywiązać do zmiennej Total_Trans_Amt. Zaraz za nią "na podium" znajdują się zmienne Total_Trans_Ct oraz Total_Revolving_Bal. Warto jeszcze wspomnieć o zmiennej Total_Ct_Chng_Q4_Q1, która znalazła się nieznacznie poza nimi. Zmienne te opisują częstość korzystania użytkowników z karty oraz mówią o wydatkach i skłonności do zadłużania. Rzeczywiście są to czynniki, które naturalnie byśmy wskazywali jako najważniejsze przy wyborze: korzystać z usługi karty czy nie. Ciekawą obserwacją jest to, że bardzo mała waga została przypisana zmiennej Months_on_book, która mówi nam o ilości kontaktów pomiędzy bankiem a klientem. Może to być sygnał, że większa ilość kontaktów z klientem, utrzymywana w formie jak dotychczas, nie wpłynie na decyzje o korzystaniu z usług. Jest to ważny alarm dla menadżera. Nie oznacza on raczej tego, że powinniśmy zaniechać kontaktów z klientem, tylko coś w nich zmienić. Może chodzi o treść komunikatów lub o rodzaj nawiązywania połączenia. Nie wybiegajmy jednak zbyt do przodu, dopóki nie mamy wyników uczenia modeli.

Jedną jeszcze wartą zwrócenia uwagi rzeczą w tym paragrfie jest to, że dwie zmiennem, które miały wysoką wartość połaczenia z klasą pozytywną znalzły się też w top 4 pod względem wag.

6.3 Korelacja zmiennych



Na podstawie powyższego wykresu możemy zauważyć, że Credit Limit jest silnie dodatnio skorelowany z Avg_Open_To_Buy co sugeruje, że zmienne te zachowują się podobnie, tzn gdy wartość jednej rośnie to drugiej też, analogicznie jeśli wartość jednej malej to drugiej też. Podobne zachowanie występuje między zmiennymi Total_Trans_Ct i Total_Trans_Amt.

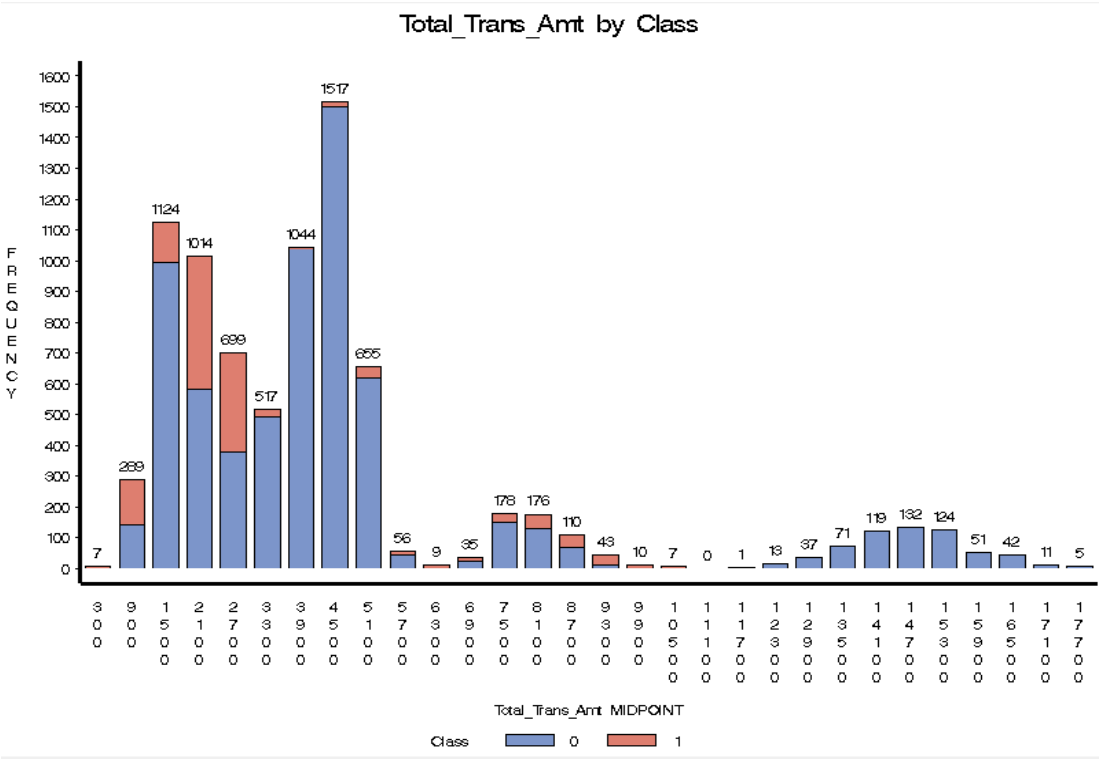
Na wykresie możemy też dostrzec silnie ujemne korelacje (tzn. zmienne zachowują się odmiennie - gdy wartość jednej rośnie to drugiej maleje i na odwrót). Takie zachowanie występuje między zmiennymi: Avg_Open_To_Buy i Avg_Utilization_Ratio oraz Credit_Limit i Avg_Utilization_Ratio.

Taka korelacja jest rzeczywiście dość intuicyjna - np.: jak rośnie liczba transakcji naturalnie będziemy się spodziewać, że rośnie suma wydatków na karcie, jak rośnie limit na karcie to będziemy się spodziewać spadku współczynnika utylizacji (tym bardziej, że w mianowniku zawiera on wartość limitu). Pozostałe dwa przypadki pozostawiam do rozpatrzenia samemu.

Warto zauważyć tutaj także, że korelacja dodatnia pomiędzy wyżej wymienionymi zmiennymi jest bliska 1. To sugeruje, że informacja wnoszona przez te pary zmiennch może być redundantna, tzn. druga zmienna w parze nie wnosi nowej informacji do analizy, ale powtarza informacje wniesione już przez pierwszą.

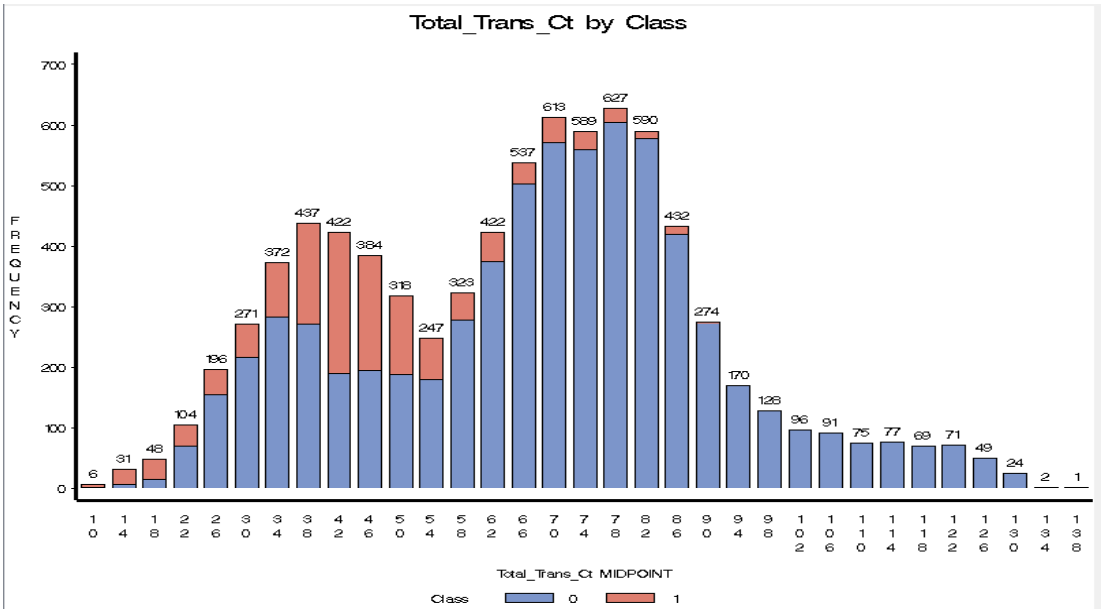
6.4 Wykresy zliczeń

Poniżej prezentuję wykresy zliczeń, które moim zdaniem niosą najciekawsze infomacje o zmiennych. Wykresy zliczeń prezentuję w postaci bar plotu, dodatkowo podział na słupku jest dokonany względem rodzaju zmiennej celu. Wykresy zostały wygenrowane przy pomocy węzła 'Wykresy różne'.



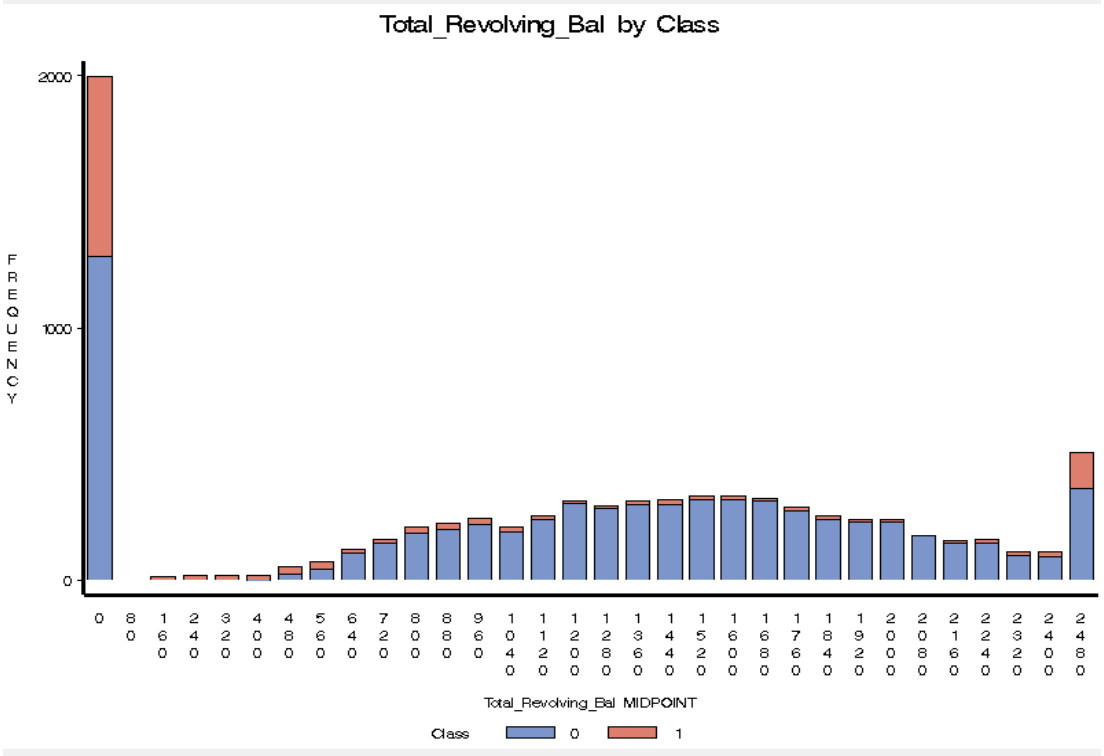
"Na pierwszy ogień" weźmy pod uwagę wykres odnoszący się do zmiennej o największej wyznaczonej wadze - Total_Trans_Amt. Widzimy tu, że większość uzytkowników, która zdecydowała się zrezygnować z usługi karty kredytowej miała małą łączną ilość wydatków zarejestriwanych na karcie.

Idąc dalej przyjrzyjmy się wykresowie przedstawiającemu zmienną zajmującą 2 miejsce na "podium wag":



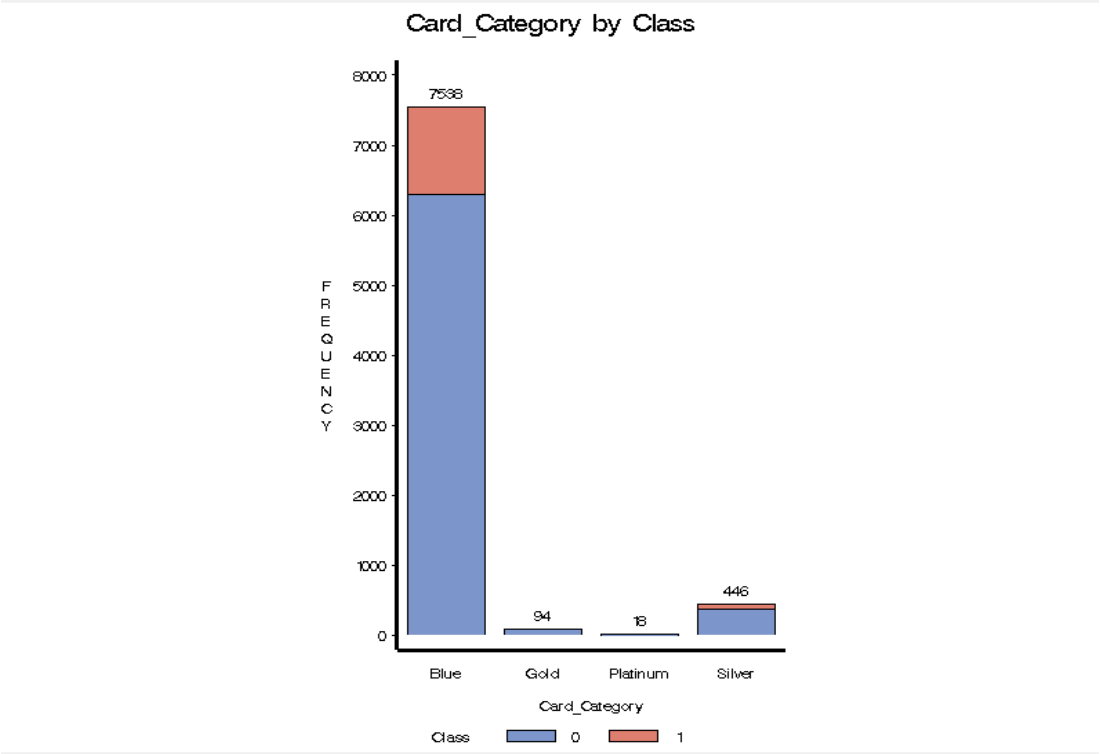
Można tu zauważyć, że "rezygnanci" z usługi karty, zdecydowanie żadziej dokonywali transakcji płatniczych za jej pomocą.

Ostatnie miejsce na "podium wag" zajmowała zmienna Total_Revolving_Bal. Myślę, że warto zerknąć na jej wykres:



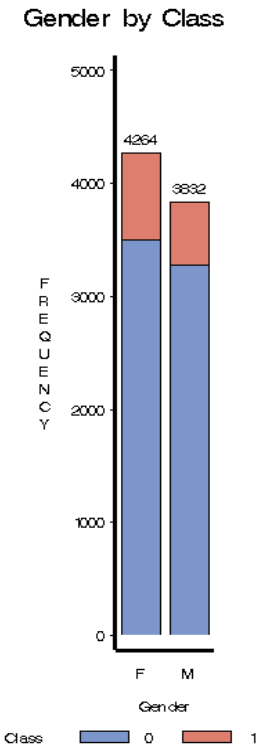
Widzimy tu wyraźny podział klientów należacych do klasy poztywnej. Ich łączna wartość debetu wynosi albo 0 albo wartość bliską wartości maksymalnej na wykresie. Jeśli pomyślimy nad tym dłużej to rzeczywiście taki rozkład ma sens, ponieważ wartość 0 może sugerować, że użytkownik nie korzysta z karty (dlatego brak debetu), co w konsekwencji prowadzi do rezygnacji z tej usługi. Natomiast duża liczność rezygnacji wśród wartości maksymalnej zmiennej sugeruje, że użytkownik często się zadłużał. Mógł on, więc dojść do wniosku, że wine za to ponosi kart kredytowa, ponieważ mając pieniądze "papierkowe" może być "na minusie" tylko wtedy, gdy oficjalnie zwróci się o pożyczkę. Zwrócenie o pożyczkę jest czynnością, którą wykonuje, więc rejestruje zdarzenie, natomiast "wpadnięcia" w debet na karcie może nie zauważyć. Możliwe więc, że bank powinien wprowadzić opcję alarmowania o debetach od pewnego progu.

Spójrzmy jeszcze na dwa, w moim odczuciu, ciekawe wykresy.



Tak jak mówiliśmy użytkowników posiadających niebieskie karty jest zdecydwoanie więcej niż pozostałych, dlatego wśród tych użytkowników jest sporo zarówno użytkowników posiadających nadal karte, jak i tych, którzy zrezygnowali z posiadania jej. Tłumaczy to wyniki analizy połączeń. Jak najbardziej obydwie kategorie są silnie związane z kartą niebieską. Ponadto, klasa 0 została przyporzadkowana do klastra wraz z tym rodzajem karty (kolor koła) - jest to uzasadnione dominacją licznościową tych użytkowników wśród wartości 'blue'. Jeszcze jedną obserwacją, którą warto wynieść z tego wykresu jest to, że wśród pozostałych rodzajów kart występują obydwie wartości zmiennej celu. Stąd rzeczywiście przyporzadkowanie małej wagi tej zmiennej jest zasadne, pomimo wysokich skojarzeń między kartą niebieską a użytkownikami.

Ostatni wykres odnosi się do zmiennej płci:



Jest on moim zdaniem ciekawy ponieważ idealnie obrazuje, dlaczego ta zmienna ma tak małą wagę. Biorąc poprawkę na to, że kobiet było nieznacznie więcej w badaniu, możemy stwierdzić, że klasy są niemalże równomiernie rozłożone pomiędzy płciami.

6.5 Podsumowanie danych treningowych

Podczas powyżej przeprowadzonej dogłebnej analizy zbioru treningowego dowiedziliśmy się jakie wagi są przypisywane zmiennym. Zrozumieliśmy też, dlaczego są one tak wybierane. Zwróciliśmy uwagę na powiązania pomiędzy zmiennymi objaśniającymi a zmienną celu. Zauważyliśmy, że mają one uzasadnienie w świecie rzeczywistym, co bardzo nas cieszy ponieważ nie ma "obserwacji dziwnych" co mogłoby sugerować na fałszywe obserwacje. Dodatkowo zauważyliśmy, że istnieją powiązania pomiędzy badaniami.

7 Wybór modeli i uczenie

Do przeprowadznia ucznenia maszynowego wybrałam 2 modele:

- drzewo decyzyjne
- regresję logistyczną

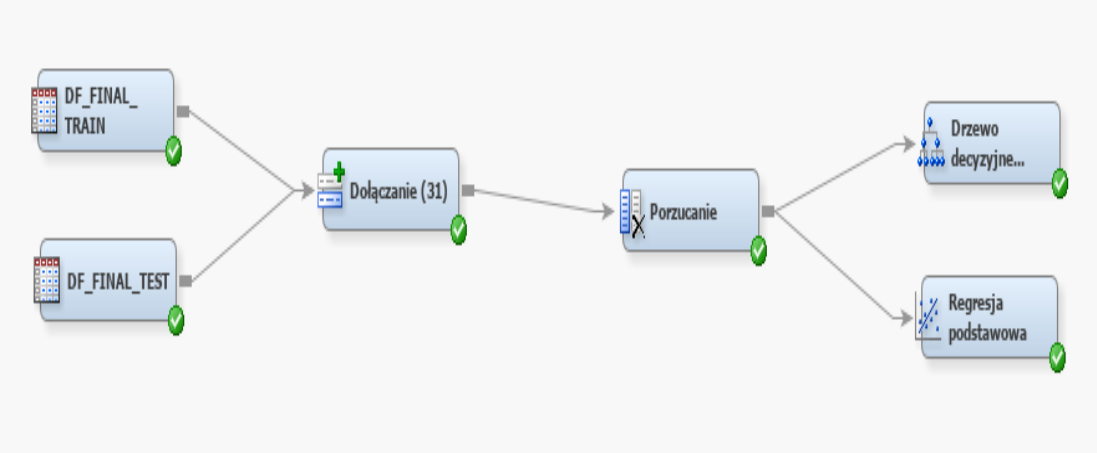
Dlaczego taki wybór?

- naszym zadaniem jest zadanie klasyfikacji binarnej, obydwa modele, potrafią obsługiwać ten problem
- modele są łatwe do zrozumienia, szczególnie drzewa decyzyjne są wysoce interpretowalne, co dla tego zadania jest bardzo ważne, dzięki temu, menadżer banku będzie mógł przewidzieć proces powstawania decyzji i na tej podstawie wprowadzić zmiany, które zmniejszą ilość klientów rezygnujących z usługi karty
- obydwa modele nie potrzebują skalowania danych
- nasza zbiór nie zawiera dużej ilości zmiennych, dzięki czemu trening powienien przebiegać szybko i skutecznie

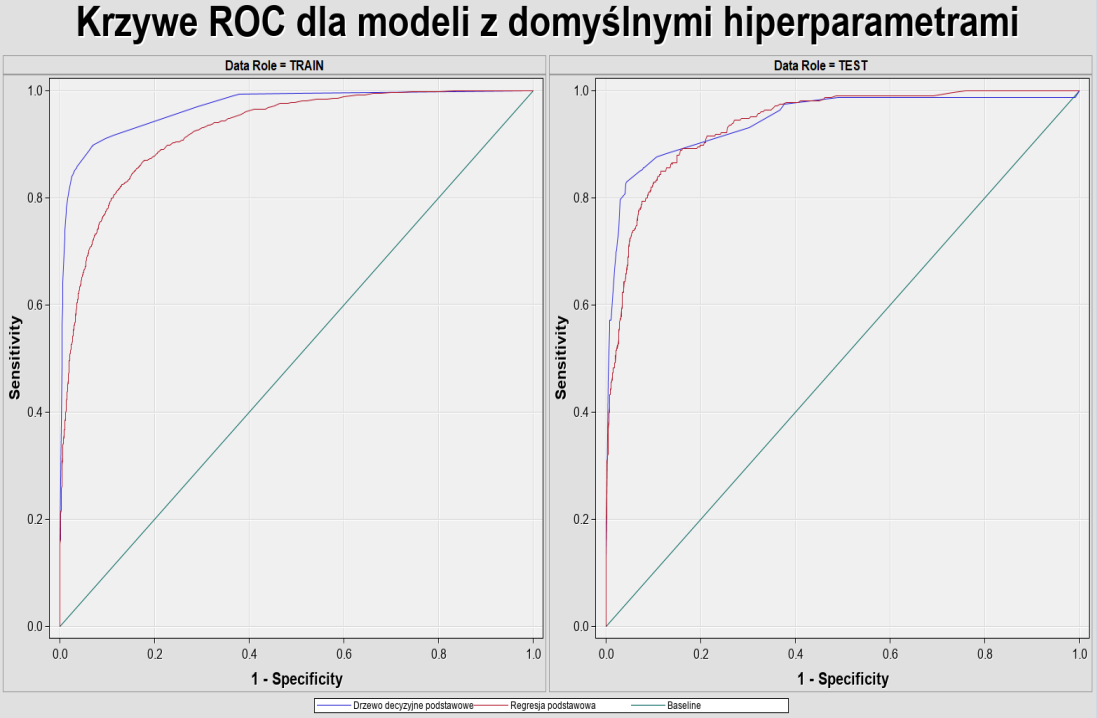
- obydwa modele cieszą się dużym uznaniem w zastosowaniu do zadań klasyfikacji binarnej w środowisku machine learningowym

7.1 Inicjalne modele - z "defaultowymi" hiperparametrami

Na początku szkolenie modeli dokonałam na tabeli 'df_final_train', z której usunęłam kolumny Attrition Flag, CLIENTUM i Selected. Uczenie przeprowadziłam przy ustawieniach domyślnych hiperparametrów.



Poniżej przedstawiam porównanie krzywych ROC dla modeli. Otrzymałąm je za pomocą kafelka 'Porównanie modeli':



Na podstawie powyższych wykresów widzimy, że nieznacznie lepiej prezentuje się model wyszkolony za pomocą drzewa decyzyjnego. Niestety różnica pomiędzy wykresem na danych treningowych i testowych może wskazywać na minimalne przeuczenie drzewa.
Poniżej widzimy dokładne wartości indeksu ROC, gdzie kolor zielony - odpowiada drzewu, różowy regresji logistycznej.

Train: Roc Index	Train: Gini Coefficient	Train: Kolmogorov-Smirnov Statistic	Train: Kolmogorov-Smirnov Probability Cutoff	Train: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	Train: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	Test: Roc Index
0.968	0.935	0.826	0.126	0.826	0.312	0.944
0.924	0.847	0.692	0.194	0.69	0.23	0.937

Rzeczywiście dobrze dostrzegliśmy na wykresie, że wartość indeksu dla drzewa spada na danych testowych. Jednak spadek ten jest na prawdę nieduży, ponieważ wynosi jedynie 0,024.

Spójrzmy teraz na macierz pomyłek:

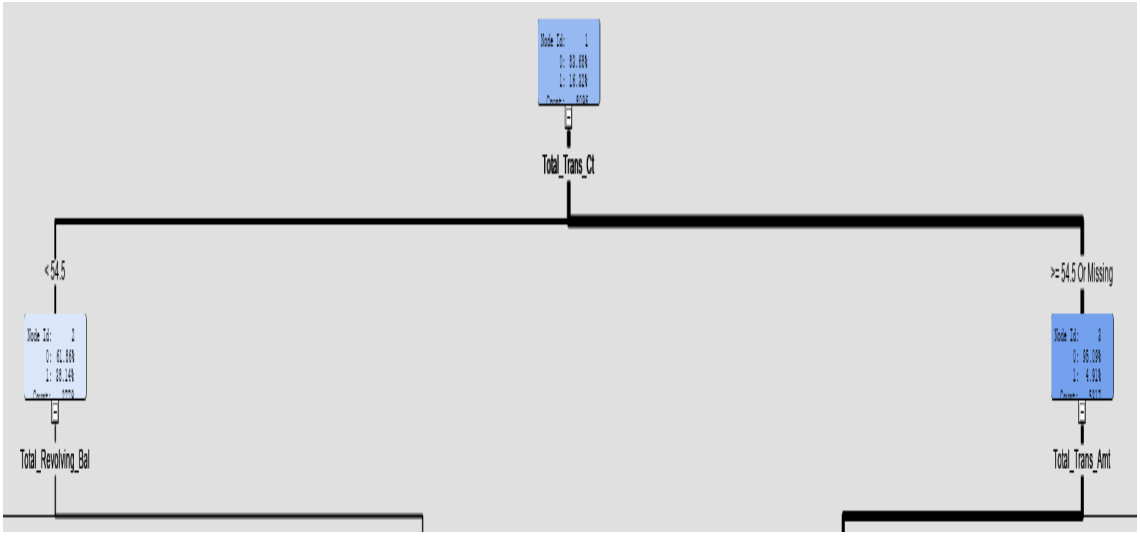
Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Drzewo decyzyjne podstawowe	TRAIN	Class		258	6653	122	1063
Regresja podstawowa	TRAIN	Class		544	6541	234	777

Widzimy, że drzewo decyzyjne mniej razy się myli, zarówno przy określaniu klasy pozytywnej, jak i negatywnej. Aby zmierzyć jak duża jest różnica pomiędzy podejmowanymi złymi decyzjami przez modele, spójrzmy jeszcze na wskaźnik pomyłek:

Model Description	Train: Misclassification Rate	Train: Average Squared Error	Train: Roc Index	Train: Gini Coefficient
Drzewo decyzyjne podstawowe	0.046937	0.038505	0.968	0.935
Regresja podstawowa	0.096097	0.071280	0.924	0.847

Wskaźnik ten jest o około 0,05 wyższy dla regresji.

Jak wspominaliśmy na wstępie drzewa decyzyjne są niesamowicie interpretowalnym modelem. Przyjrzyjmy się zatem jakie zmienne brało ono pod uwagę przy dokonywaniu decyzji oraz jakie były wartości decyzyjne.



Zauważmy, że podziały na wysokości 0 i 1 drzewa odbywają się w oparciu o zmienne z "podium wagowego". Ponadto pierwszym warunkiem podziału jest liczba transakcji wynosząca 54. Jeśli użytkownik wykonał więcej transakcji to drzewo przewiduje, że użytkownik nie odstąpi od korzystania z karty. Dalsze prześledzenie drzewa pozostawiam dla chętnych, ponieważ umieszczenie go tu jako statycznej grafiki stanowiłoby prawdziwe wyzwanie w interpretacji ze względu na jego rozmiar. Wspomnę tylko, że rzeczą, na którą warto zwrócić uwagę, jeśli zdecydujemy się mu przyjrzeć to to, że 5 ostatnich zmiennych z rankigu wag nie było brane pod uwage w czasie tworzenia drzewa, a zmienna Gender była bardzo rzadko występującym warunkiem.

7.2 Modele o innych wartościach hiperparametrów

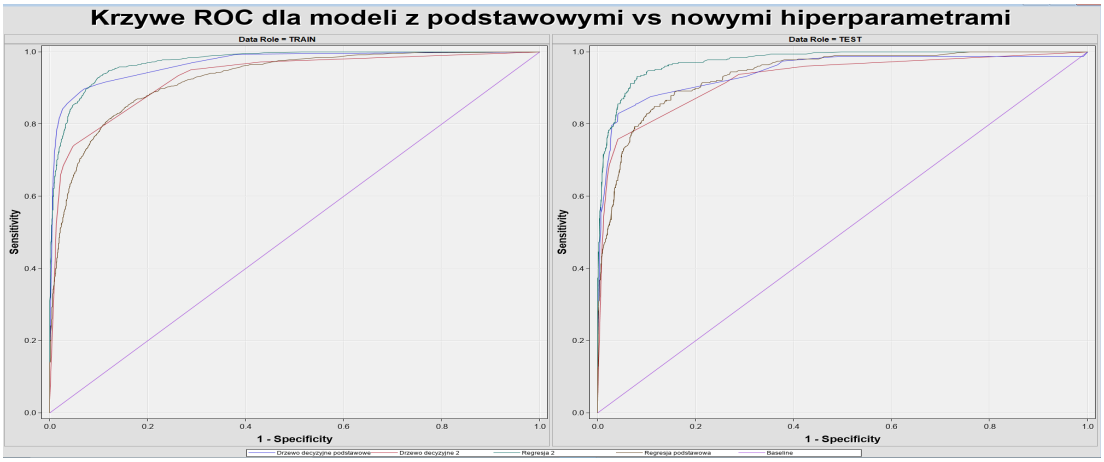
Powyżej zauważyliśmy, że drzewo mogło ulec minimalnemu przeuczeniu, natomiast regresja zachowuje się nieznacznie gorzej niż drzewo. Czy jeśli zmenimy hiperparametry to nasze przwidywania ulegną polepszeniu? Sprawdźmy! Dla drzewa zmieniam następujące hiperparametry:

- maksymalna głębia z 6 na 4
- minimalna wielkość zmiennej kategoryzującej z 5 na 10
- wielkość liścia z 5 na 10

Powyższe operacje służą przycięciu drzewa. Ponadto zdecydowałam się pozostawić entropię jako kryterium porządkowej zmiennej celu, ponieważ wskaźnik Giniego zdarza się izolować najczęściej występującą klasę w osobnej gałęzi, natomiast entropia generuje nieco bardziej zrównoważone drzewa. Dla regresji natomiast:

- włączyłam rozważanie wyrazów wielomianu
- stopień wielomianu ustawiłam na 4

Ten zabieg dokonałam w celu lepszego dopasowania modelu do danych, ponieważ przy istnieniu wielu cech, tak jak ma to miejsce w naszym zbiorze, model po takich zmianach jest w stanie znajdować lepsze powiązania między nimi. Oto wyniki:



Train: Roc Index	Train: Gini Coefficient	Train: Kolmogorov-Smirnov Statistic	Train: Kolmogorov-Smirnov Probability Cutoff	Train: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	Train: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	Test: Roc Index
0.972	0.945	0.829	0.154	0.825	0.19	0.975
0.968	0.935	0.826	0.126	0.826	0.312	0.944
0.924	0.847	0.692	0.194	0.69	0.23	0.937
0.929	0.857	0.691	0.15	0.688	0.247	0.926

Przy grafice pokazującyej wyniki wskaźnika ROC doszły dwa nowe kolory: żółty symbolizujący nowe drzewo decyzyjne ('Drzewo decyzyjne 2') oraz pomarańczowy - nową regresje logistyczną ('Regresja 2'). Widzimy, że dla drzewa pogorszyły się wyniki, przewidywań, chociaż już teraz nic nie wskazuje na przeuczenie. Dodatkowo nowe drzewo jako pierwsze kryterium podziału także brało pod uwagę wartość zmiennej Total_Trans_Ct. Pozostałe zmienne stanowiące warunek podziału też się powtarzają, a zmienne o wagach bliskich 0 nie występują. Natomiast dla regresji logistycznej wyniki uległy znacznemu polepszeniu. Decyzją moją jest więc w dalszym badaniu rozważać model 'Regresja 2' oraz 'Drzewo decyzyjne podstawowe'. Zobaczmy jeszcze ile poświęciliśmy czasu na szkolenie tych najlepszych modeli:

- 'Drzewo decyzyjne podstawowe' 0 godz. 0 min. 12,88 sek.
- 'Regresji 2' - 0 godz. 0 min. 11,46 sek.

7.3 Walidacja

W celu dalszego badania wybranych modeli chciałam sprawdzić, czy zastosownie walidacji może poprawić wyniki predykcji, które są i tak juz bardzo wysokie. Ze zbioru treningowego przeznaczyłam 75 % danych na uczenie, a pozostałe na walidacje. Pozwoli to programowi na dostorjenie hiperparametrów. Rezultaty tym razem przedstawie tylko w postaci liczbowej, ponieważ z wykresów ciężko jest odczytać dane ze względu na niewielkie różnice. Teraz kolorami żółtym i pomarańczowym oznaczałam modele walidacyjne (żółty - dla drzewa, pomarańczowy - dla regresji):

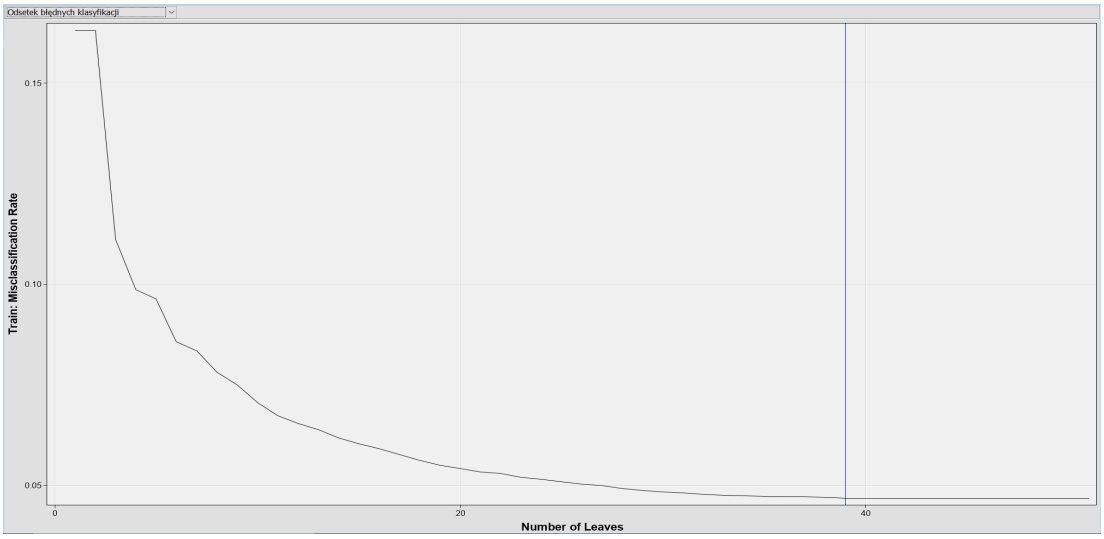
Train: Roc Index	Train: Gini Coefficient	Train: Kolmogorov-Smirnov Statistic	Train: Kolmogorov-Smirnov Probability Cutoff	Train: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	Train: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	Valid: Roc Index	Valid: Gini Coefficient	Valid: Kolmogorov-Smirnov Statistic	Valid: Kolmogorov-Smirnov Probability Cutoff	Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	Valid: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	Test: Roc Index
0.968	0.935	0.826	0.126	0.826	0.312	.	.	.	.	.	.	0.944
0.972	0.945	0.829	0.154	0.825	0.19	.	.	.	.	.	.	0.975
0.952	0.905	0.809	0.127	0.807	0.371	0.956	0.911	0.837	0.087	0.82	0.492	0.933
0.972	0.944	0.826	0.201	0.821	0.184	0.969	0.938	0.837	0.116	0.835	0.194	0.974

O dziwo jakość dla drzew się niznacznie pogorszyła, natomiast dla regresji pozostała na tym samym poziomie. Sprawdźmy jeszcze wskaźnik nieprawidłowych klasyfikacji:

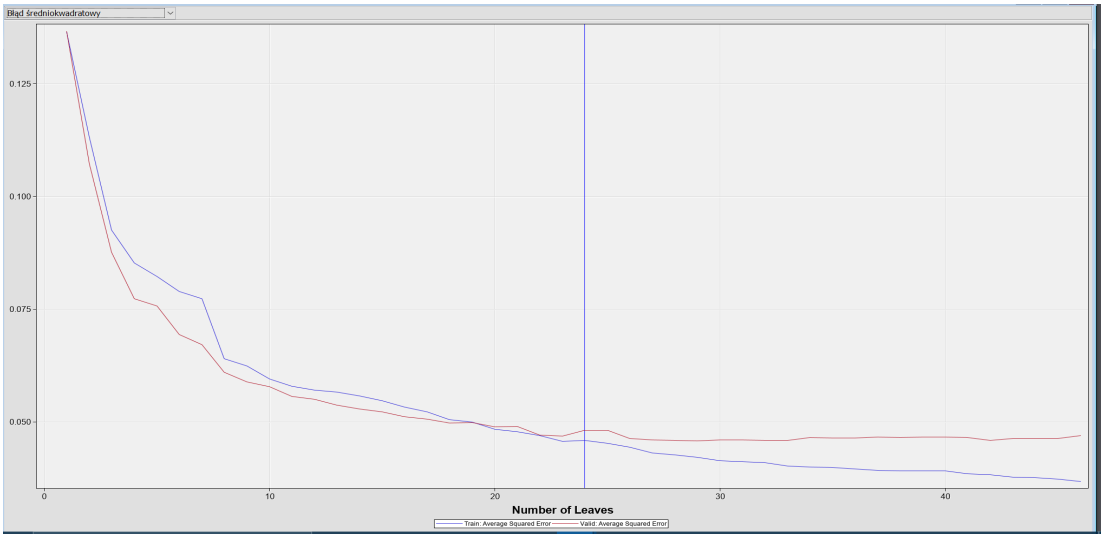
Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
.	0.038505	0.046937	.
.	0.044179	0.061141	.
0.057749	0.045866	0.053707	0.048124
0.068115	0.043947	0.061120	0.047363

Te wyniki pokrywają się z indeksem ROC.

Spójrzmy jak w takim razie zmieniły się ustawienia drzewa:
Drzewo przed walidacją:



Drzewo po walidacji:



Widzimy, że po zastosowaniu zbioru walidacyjnego algorytm dokonał dość znacznego podcięcia drzewa (oznaczone na obrazku pionową linią ciągłą). Przed walidacją drzewo miało blisko 40 liści po walidacji jest ich niespełna 25.

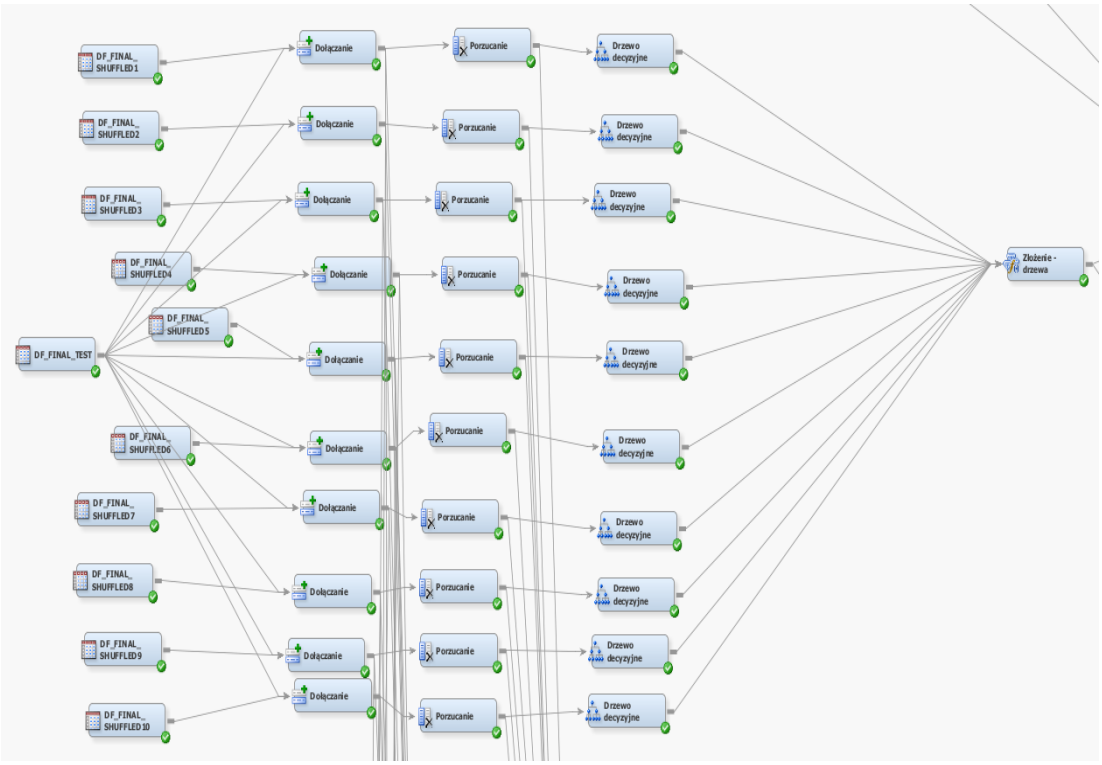
Wyniki te są dość zaskakujące, aczkolwiek spadek jakości modelu drzewa nie jest znaczny. Dla regresji nie doszło do zmian.

Spójrzmy jeszcze jaki był czas szkolenia modeli:

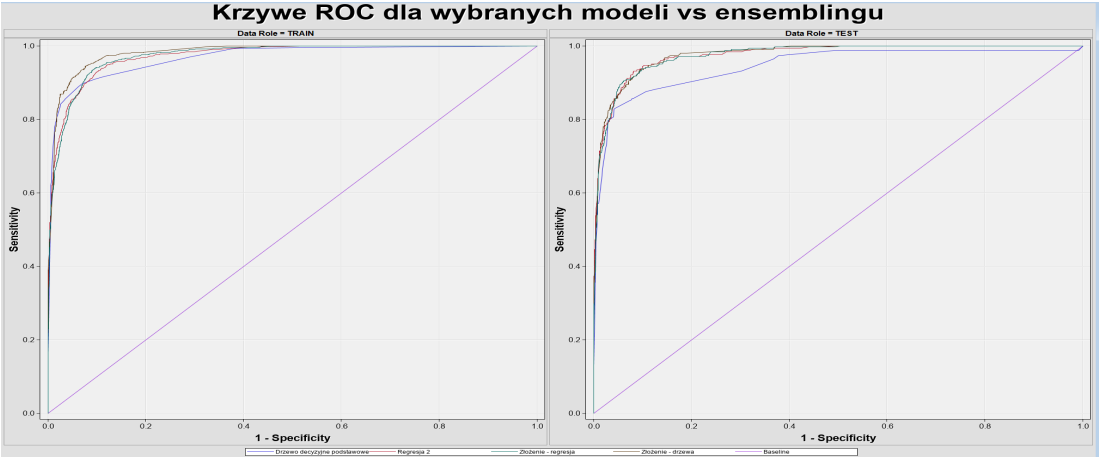
- 'Drzewo decyzyjne podstawowe walidacyjne' 0 godz. 0 min. 14,77 sek.
- 'Regresja 2 walidacyjna' 0 godz. 0 min. 13,16 sek

7.4 Tworzenie ensemblingów

W celu poprawy jakości modeli postawnowiłam zbalansować dane treningowe. Dzięki temu, że dane będą rozłożone równomiernie pomiędzy aktywnych klientów a tych którzy odeszli od usługi karty kredytowej modele nie będą częściej "strzelać" w klasę negatywną, która była bardziej liczna. W tej sekcji korzystam z tabel 'df_final_shuffled1',..., 'df_final_shuffled10'. Na tych danych zbuduję ensemblingi w skład, których będą wchodzić najlepsze modele przedstawione w powyższych sekcjach, tj. 'Regresja 2' i 'Drzewo decyzyjne podstawowe'. Tak wygląda proces tworzenia ensemblingu dla drzew:



Analogicznie wygląda uczenie dla modelu regresyjnego (zamieniamy kafelek modelu drzewa na kafelek odpowiednio 'Regresji 2') .
Możemy zauważyć, że tym razem nie wyodrębniałam zbioru walidacyjnego. Postąpiłam tak, ponieważ stworzyłam ensembling danych treningowych, na których szkoliłam odpowiednio modele drzewa/regresji logistycznej. Uważam, że tak obszerne rozbicie zbioru treningowego i algorytmów może zastąpić dokonywawnie walidacji. Poniżej przedstawiam wykresy krzywych ROC :



a także indeksy ROC:

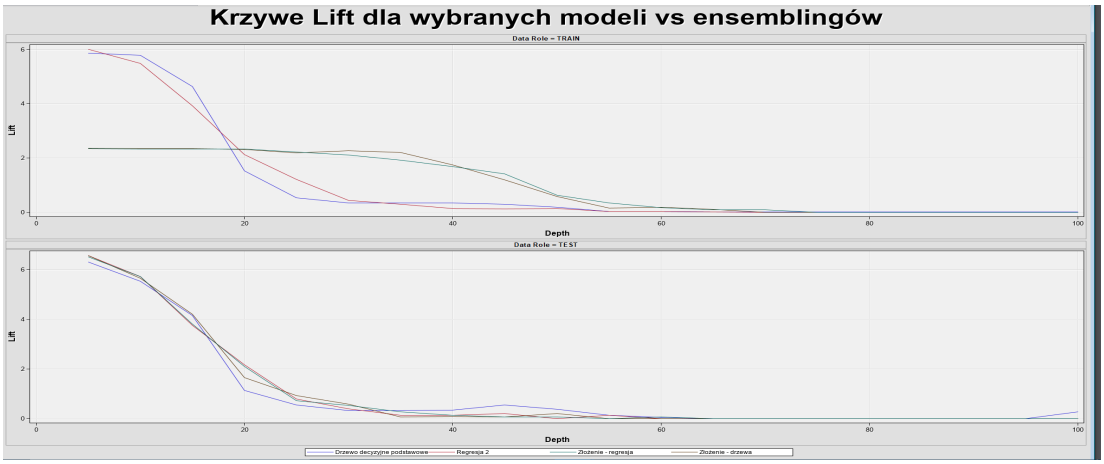
Train: Roc Index	Train: Gini Coefficient	Train: Kolmogorov-Smirnov Statistic	Train: Kolmogorov-Smirnov Probability Cutoff	Train: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	Train: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	Test: Roc Index
0.968	0.935	0.826	0.126	0.826	0.312	0.944
0.972	0.945	0.829	0.154	0.825	0.19	0.975
0.981	0.963	0.868	0.429	0.865	0.52	0.976
0.973	0.946	0.843	0.395	0.84	0.515	0.974

Oznaczenia na obrazku:

- kolor zielony - 'Drzewo decyzyjne podstawowe'
- kolor różowy - 'Regresja 2'
- kolor żółty - 'Złożenie - drzewa'
- kolor pomarańczowy - 'Złożenie - regresja'

Jak widzimy znacznie poprawiła się dokładność ROC AUC dla drzew. Zarówno na zbiorze treningowym jak i testowym. Dodatkowo mniejsza różnica pomiędzy wynikami dla zbiorów treningowych i testowych świadczy o elminacji przeuczenia (aczkolwiek było one i tak małe jak mówiliśmy wcześniej). Dla ensemblingów modeli regresyjnych wynik jest podobny. Jednak zauważmy, że wcześniej dla 'Regresji 2' wynik był już tak wysoki, że ciężko go bardziej poprawić.

Zerknijmy jeszcze na krzywe Lift, które mówią nam jak użycie modelu wpłwya na przewidywanie obserwacji pozytywnych w porównaniu do wyznaczania obserwacji pozytywnych bez użycia jakiegokolwiek modelu :



Widzimy, że porównywane modele mają podobny przebieg krzywych, także będą z podobną skutecznością przewidywać klasę pozytywną. Na przykład wobec 10 procent. obserwacji o zwiększonym prawdopodobieństwie bycia klasą pozytywną w naszym zbiorze jesteśmy w stanie za ich pomocą przewidzieć prawie 6 razy lepiej klasę pozytywną niż nie używając żadnego modelu.

Na koniec rzućmy okiem na czas treningu tych modeli:

- 'Złożenie - drzewa' 0 godz. 0 min. 53,16 sek.
- 'Złożenie - regresja' 0 godz. 0 min. 49,48 sek.

Widać, że czas znacząco się zwiększył (w skali sekund).

7.5 Podsumowanie modeli

Na podstawie powyższych rozważań możemy zauważyć, że najlepiej prezentuje się model będący złożeniem 'Drzew decyzyjnych podstawowych', wytrenowany na zrównoważonych danych treningowych.

Jeżeli jednak kierujemy się jak najkrótszym czasem tworzenia modelu powiniśmy wybrać prostszy model. Jeżeli zależy nam na jak największej interpretowalności wybierzmy 'Drzewo decyzyjne podstawoe', natomiast jeśli wolimy postawić na jak największą dokładność w połączeniu z krótkim czasem szkolenia wybierzmy 'Regresje 2'.

W następnym rodziale sprawdzimy jednak, czy poprzez zmniejszanie wymiarowości jesteśmy w stanie przyspieszyć proces szkolenia, aby przy wyborze modelu kierować się zarówno dokładnością jak i szybkością.

8 Redukcja wymiarowości i testowanie na zredukowanych danych najlepszych modeli

W ciągu naszych rozważań dokonywaliśmy elminacji zmiennch, które by przeszkdzały w uczeniu modelu - zmienne nie mające sensu (np. 'Var23'), zmienne nie mające wpływu na decyzje klienta (np. numer klienta), możliwe fałszywe obserwacje (7 obserwacji z powtarzającymi się wartościami 'Unknown' w różnych kolumnach) . Na tak zredukowanych danych szkoliliśmy omówione powyżej modele.
Jednak w czasie dogłębnej analizy danych oraz analizy modelu wyciągaliśmy kolejne ciekawe wnioski mogące pomóc w wyelminowaniu kolejnych danych. Przypomnijmy sobie te wnioski:

- zauważyliśmy, że niektóre zmienne mają wagi bliskie 0
- dostrzegliśmy silnie dodatnią korelacje pomiędzy zmiennymi Credit_Limit i Avg_Open_To_Buy oraz Total_Trans_Ct i Total_Trans_Amt
- przy analizie modeli drzew zwróciliśmy uwagę, że w warunkach podziału węzła nie występują zmienne o wagach bliskich 0

Postanowiłam sprawdzić czy usunięcie tych zmiennych wpłynie na funkcjonowanie naszych najlepiej działajacych modeli -'Złożenie - drzewa' i 'Złożenie - regresja'. Procesu elminacji zmiennych dokonałam w dwóch etapach. Najpierw wyelminowałam 6 obserwacji o najmniejszych wagach i stworzyłam modele 'Złożenie - drzewa obc 1' i 'Złożenie - regresja obc 2'. Następnie wyelminowałam po jednej zmiennej z pary zmiennych o silnie dodatnich korelacjach - wybrałam tutaj Credit_Limit oraz Total_Trans_Amt. Dlaczego wybrałam te zmienne z pary? Wartość wagi Credit_Limit była nieznacznie niższa niż wartość zmiennej Avg_Open_To_Buy, natomiast o pozostawieniu Total_Trans_Ct a nie Total_Trans_Amt, stanowiło to, że zmienna Total_Trans_Ct stanowiła pierwszy warunek podziału węzła w drzewach.
Oto wyniki:

Train: Roc Index	Train: Gini Coefficient	Train: Kolmogorov-Smirnov Statistic	Train: Kolmogorov-Smirnov Probability Cutoff	Train: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	Train: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	Test: Roc Index
0.981	0.963	0.868	0.429	0.865	0.52	0.976
0.981	0.963	0.868	0.424	0.865	0.52	0.976
0.948	0.896	0.768	0.345	0.758	0.379	0.947
0.973	0.946	0.843	0.395	0.84	0.515	0.974
0.971	0.942	0.829	0.391	0.826	0.509	0.974
0.951	0.901	0.769	0.431	0.768	0.502	0.953

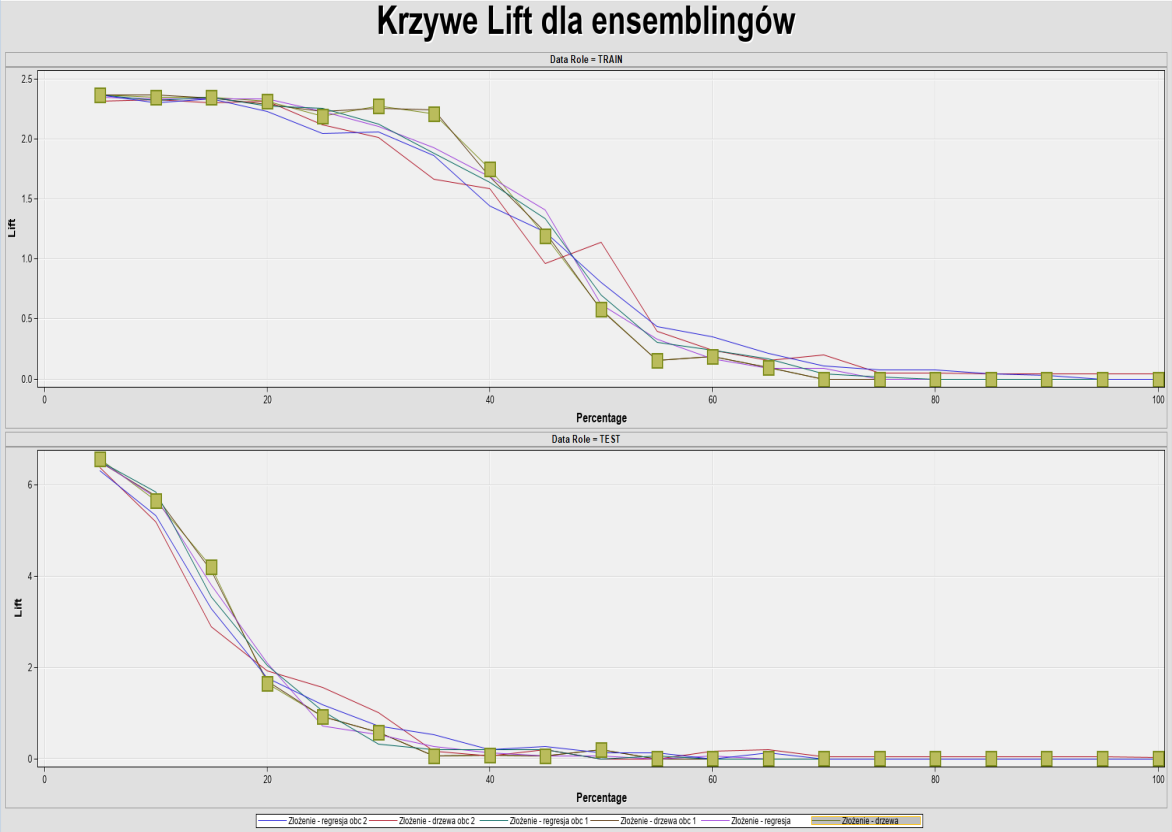
Kolorem zielonym są oznaczone wyniki dla drzew w kolejności odpowiednio: zwykłe złożenie, po pierwszej elminacji danych, po kolejnej elminacji dancyh. Kolorem różowym zostały oznaczone wyniki dla regresji w tej samej kolejności co dla drzew. Widzimy, że pierwszy etap redukcji nie wpłynął na wyniki indeksu ROC, natomiast drugi etap obniżył wartość indeksu. Sprawdźmy jak często myliły się algorytmy:

Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Złożenie - regresja obc 2	TRAIN	Class		182	1610	190	1139
Złożenie - drzewa obc 2	TRAIN	Class		182	1595	205	1139
Złożenie - regresja obc 1	TRAIN	Class		127	1651	149	1194
Złożenie - drzewa obc 1	TRAIN	Class		97	1683	117	1224
Złożenie - regresja	TRAIN	Class		111	1656	144	1210
Złożenie - drzewa	TRAIN	Class		97	1684	116	1224

Model Description	Train: Misclassification Rate	Train: Average Squared Error	Train: Roc Index	Train: Gini Coefficient
Złożenie - drzewa	0.06825	0.050715	0.981	0.963
Złożenie - drzewa obc 1	0.06857	0.050902	0.981	0.963
Złożenie - regresja	0.08170	0.062316	0.973	0.946
Złożenie - regresja obc 1	0.08843	0.065717	0.971	0.942
Złożenie - regresja obc 2	0.11919	0.085855	0.951	0.901
Złożenie - drzewa obc 2	0.12400	0.087579	0.948	0.896

Najbardziej satysfakcjonujące statystyki otrzymujemy znowu dla zwykłego złożenia i dla złożenia bazującego na zmiennych poddanych tylko pierwszemu etapowi redukcji.

Przyjrzyjmy się wykresą krzywych Lift:



Patrząc na krzywe dla zbioru testowego możemy zauważyć, że używając modeli '... - złożenie' oraz '... - złożenie obc 1' wobec 10 procent. obserwacji o zwiększonym prawdopodobieństwie bycia klasą pozytywną w naszym zbiorze jesteśmy w stanie przewidzieć prawie 6 razy lepiej klasę pozytywną niż nie używając żadnego modelu, a dla modeli '...obc 2' o około 5,25 raza lepiej. Natomiast wobec 20 procent obserwacji o zwiększonym prawdopodobieństwie bycia klasą pozytywną o około 2 razy lepiej dla wszystkich modeli.

Sprawdźmy jeszcze jak redukcja wymiarowości wpływa na czas szkolenia:

- 'Złożenie - drzewa' 0 godz. 0 min. 53,16 sek.
- 'Złożenie - regresja' 0 godz. 0 min. 49,48 sek.

- 'Złożenie - drzewa obc 1' 0 godz. 0 min. 51,68 sek.
- 'Złożenie - regresja obc 1' 0 godz. 0 min. 52,17 sek.
- 'Złożenie - drzewa obc 2' 0 godz. 0 min. 32,57 sek.
- 'Złożenie - regresja obc 2' 0 godz. 0 min. 28,90 sek.

Zauważmy, że przechodząc przez całą procedure redukcji wymiarowości możemy znacząco (w skali sekund ;)) zmniejszyć czas szkolenia modeli, jednak musimy się wtedy liczyć z utratą dokładności. Najlepszym rozwiązaniem wydaje się jednak przejście tylko przez etap 1 - w przypadku drzew zmniejszamy lekko czas, a w przypadku obydwu modeli zwiększamy interpretowalność nie tracąc na dokładności.

9 Podsumowanie

Podsumowując w przygotowanym przeze mnie projekcie na początku zapoznaliśmy się z problemem, którym było przewidywanie, którzy klienci banku mogą zrezygnować z usługi karty kredytowej. Na początku zapoznaliśmy się z danymi - dokonaliśmy ich wstępnej analizy za pomocą zliczeń, zbadania wartości jakie przyjmują oraz przeanalizowania braków danych. Zwróciliśmy uwagę na niezbilansowanie zmiennej celu. Następnie podzieliliśmy dane na zbiór treningowy i testowy. Dokonaliśmy potem też wtórnego podziału zbioru treningowego w celu strworzenia zbioru zbalansowanych zbiorów treningowych. Następnie wybraliśmy modele, które będziemy badać w projekcie - regresję i drzewo decyzyjne. W kolejnych krokach próbowaliśmy zwiększyć skuteczność modeli. Ostatecznie odkryliśmy, że największą skutecznością charakteryzuje się ensembling drzew decyzyjnych oparty na zrónoważonym zbiorze danych treningowych. Ostatnim krokiem w naszym projekcie było zajęcie się redukcją wymiarowości. Po jej dokonaniu przyjrzeliśmy się jeszcze raz modelom. Zauważyliśmy, że za jej pomocą możemy przyspieszyć czas szkolenia jednak musimy się liczyć z utratą dokładności.
Ostatecznie możemy stwierdzić, że wybór modelu przeznaczonego do przewidywania potencjalnych klientów rezygnujących z usługi karty kredytowej należy do menadżera banku. Każdy z modeli ma inne wady i zalety. W zależności od tego czy będzimy się kierować czasem szkolenia, dokładnością czy interpretowalnością wybierzemy inny model. Jednak wszystkie modele przedstawione w powyższym projekcie zachowywały się zachwycająco dobrze. Także obiektywnie możemy je polecić do użytkowania w sektorach bankowych wraz z uwzględnieniem całej analizy danych, która dostarcza także wielu informacji na temat klientów.

10 Bibliogradia

1 https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers?resource=download

2 https://turbotlumaczenia.pl/blog/jakie-sa-zarobki-w-usa-jakie-jest-srednia-pensja-w-usa/

3 https://towardsdatascience.com/meaningful-metrics-cumulative-gains-and-lyft-charts-7aac02fc5c

4 Wikipedia - demografia Stanów Zjednoczonych

5 https://howtolearnmachinelearning.com/articles/the-lift-curve-in-machine-learning/

6 "Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow", autor: Aurelien Geron, wyd. Helion

7 Wykłady z przedmiotu Wybrane algorytmy i systemy analizy przygotowane przez dr hab. inż. Maciej Grzenda

8 http://manuals.pqstat.pl/statpqpl:redpl