SZKOŁA GŁÓWNA HANDLOWA W WARSZAWIE
WARSAW SCHOOL OF ECONOMICS

Studium magisterskie

Kierunek: Metody Ilościowe w Ekonomii i Systemy Informacyjne

Przedmiot: Ekonometria bayesowska

Michał Cisek

Prognozowanie łącznego dochodu filmu nominowanego do Oscara

Warszawa 2017

1. Wprowadzenie

Celem niniejszego raportu jest zbadanie determinant wpływających na całkowity dochód uzyskany z filmu. Określenie czynników wpływających na to zjawisko może okazać się przydatne dla osób pracujących w produkcji filmów czy marketingu, na przykład w celu dobrania odpowiedniej kampani marketingową. Dodatkowo w celu precyzyjniejszej wiedzy na temat parametrów, do estymacji zostaną włączone informacje pochodzące z innych badań na ten temat.

2. Opis danych

Dane użyte w pracy zostały pobrane za pomocą scrapera napisanego własnoręcznie przez autora w języku R, który jest załączony do pracy, a także dostępny na profilu github[1] Dokładniej rzecz ujmując dane na temat filmów zostały pobrane z serwisu IMDb. Zbiór danych zawiera 526 instancji dla filmów nominowanych do nagrody Akademii Filmowej od pierwszej edycji, czyli roku 1928. Zmienne dostępne dla każdej produkcji przedstawia Tabela 1.

Tabela 1 Opis zmiennych

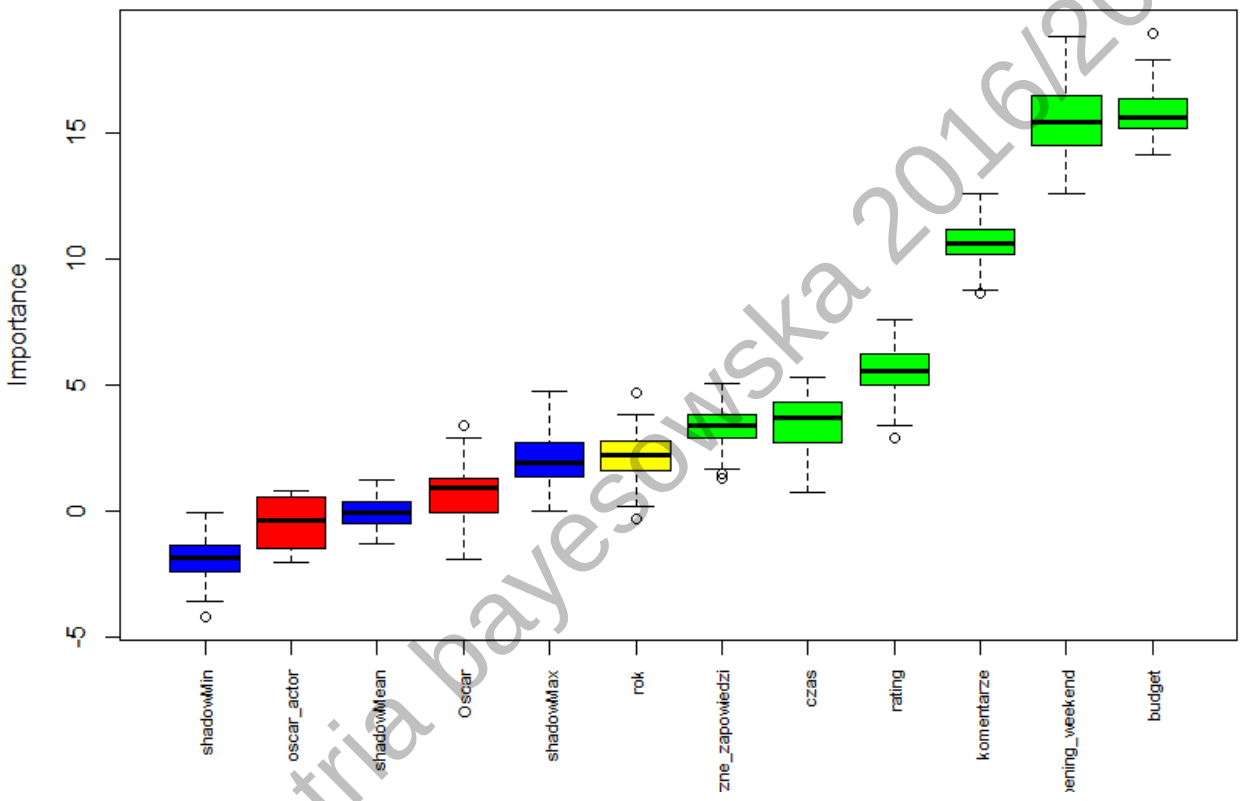
Nazwa zmiennej	Opis zmiennej	Typ zmiennej
Oscar	Czy dany film zdobył Oscara	Binarna
Rok	Rok nominacji	Nominalna
Budget	Wielkość budżetu (w mln USD)	Nominalna
Opening_weekend	Przychód w pierwszy weekend po premierze (w mln USD)	Nominalna
Gross	Łączny przychód (w mln USD)	Nominalna - zmienna zależna
Czas	Czas trwania filmu (w minutach)	Nominalna
Rating	Ocena filmu przez użytkowników (w skali 1-10)	Nominalna
Komentarze	Liczba komentarzy	Nominalna
Zewnetrzne_zapowiedzi	Liczba zapowiedzi filmu poza stroną IMDb	Nominalna
Oscar_actor	Czy w obsadzie filmu znajdował się aktor, który zdobył nagrodę dla najlepszego aktora (zarówno pierwszo- jak i drugoplanowego)	Binarna
Gatunek	Gatunek filmu	Dyskretna

Jako że dla filmów z XX wieku często brakowało danych na temat wielkości budżetu czy przychodu w pierwszy weekend po premierze, dlatego też dane zostały ograniczone do nominacji przyznanych od 1990 roku, tak że ostateczny zbiór zawierał 151 obserwacji.

[1] https://github.com/michalcisek/Filmy-oscarowe

3. Selekcja zmiennych

W celu ograniczenia liczby zmiennych objaśniających i włączenia do końcowego modelu tylko istotnych regresorów dokonana została selekcja zmiennych. Użyty został algorytm Boruta, który ocenia istotność opierając się na lesie losowym[2]. Wynik działania tego algorytmu przedstawia Rysunek 1. Widzimy że tylko zmienne określające wygranie Oscara, występowanie aktora nagrodzonego tą nagrodą, oraz rok nominacji uznane zostały za nieistotne.



Rysunek 1 Istotność zmiennych na podstawie algorytmu Boruta

[2] https://cran.r-project.org/web/packages/Boruta/Boruta.pdf

4. Estymacja modelu KMNK

Pomimo że w artykułach wyszczególnionych w literaturze autorzy wskazywali, że lepsze efekty do analizy przedstawionego problemu daje zastosowanie klasteryzacji (np. za pomocą metody k-średnich), a następnie wielomianowej regresji logistycznej, w naszym badaniu użyty został klasyczny model regresji liniowej. Oszacowania takiego modelu widoczne są w Tabeli 2.

Tabela 2 Oszacowania parametrów KMNK

Zmienna objaśniająca	Oszacowanie parametru
Stała	100.823
Opening_weekend	1.122
Budget	1.249
Rating	-8.485
Komentarze	0.045
Zewnetrzne_zapowiedzi	-0.0321
Czas	-0.1497

Przykładowo, wraz ze wzrostem o 1 milion dolarów budżetu, łączny przychód filmu wzrasta o 1.25 miliona *ceteris paribus*. Warto zwrócić uwagę, że tylko zmienne *Opening_weekend, Budget* oraz *Komentarze* zostały uznane za statystycznie istotne. Współczynnik determinacji wyniósł 57%.

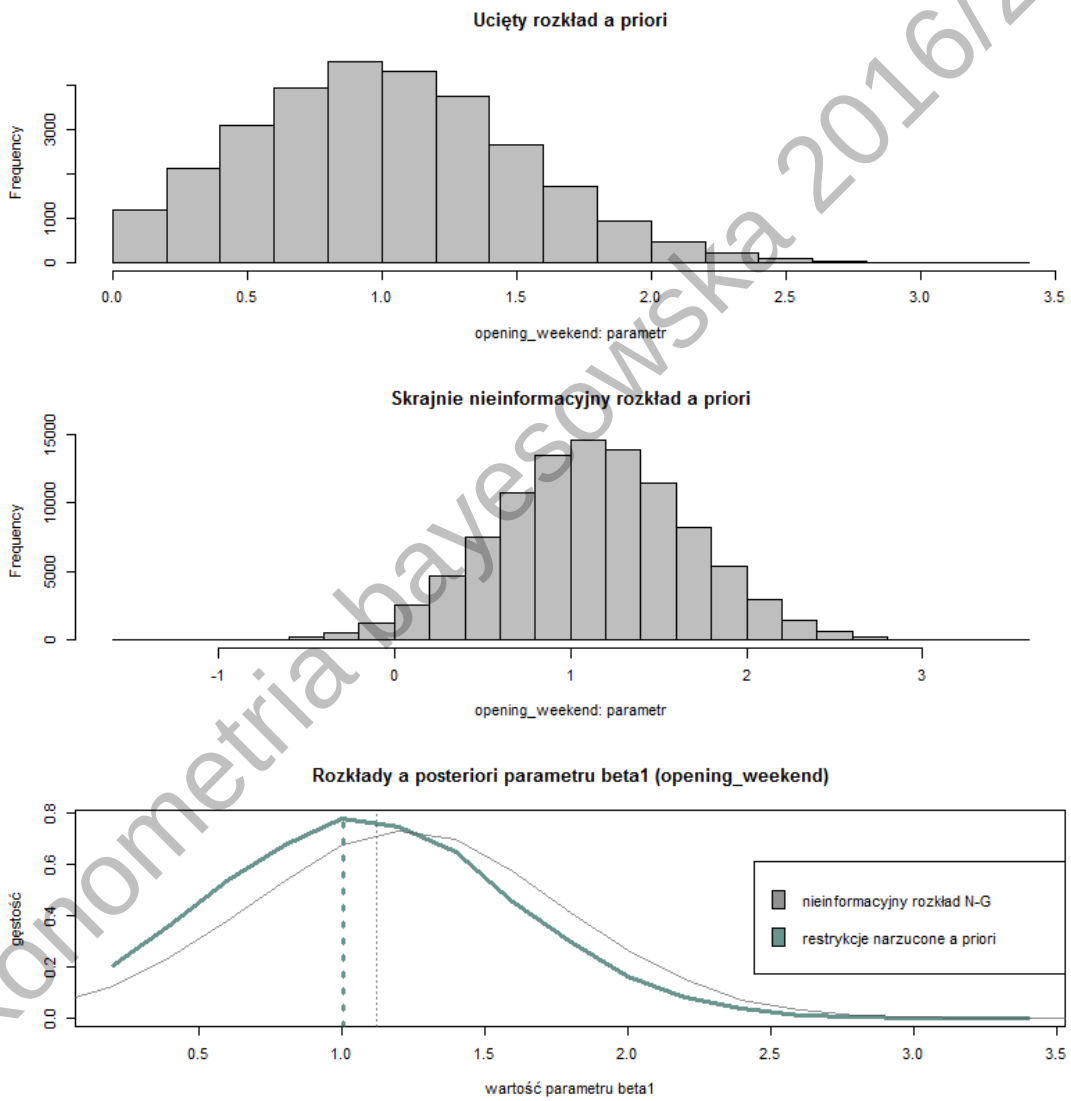
5. Wiedza apriori

Problem określenia determinant finansowo udanych produkcji nie został wyczerpująco opisany w ogólnodostępnych artykułach w internecie. Dlatego też wiedza wniesiona na podstawie dwóch artykułów dostępnych w literaturze jest dosyć skromna i dotyczy tylko znaku parametrów 3 zmiennych użytych w modelu regresji. Szczegółowiej, została wykazana istotna dodatnia korelacja między łącznym dochodem, a dochodem uzyskanym w pierwszy weekend po premierze. W związku z tym zakładamy, iż parametr przy zmiennej *Opening_weekend* będzie dodatni. Podobny związek został potwierdzony w stosunku do budżetu. Ostatnią informacją uzyskaną w wyniku innych badań jest występowanie pozytywnego powiązania między ratingiem udzielonym w ankiecie prowadzonej po seansie filmowym. Taką ankietę możemy utożsamiać z ratingiem wystawionym przez użytkowników serwisu IMDb, dlatego zakładamy że parametr przy zmiennej *Rating* będzie większy od zera.

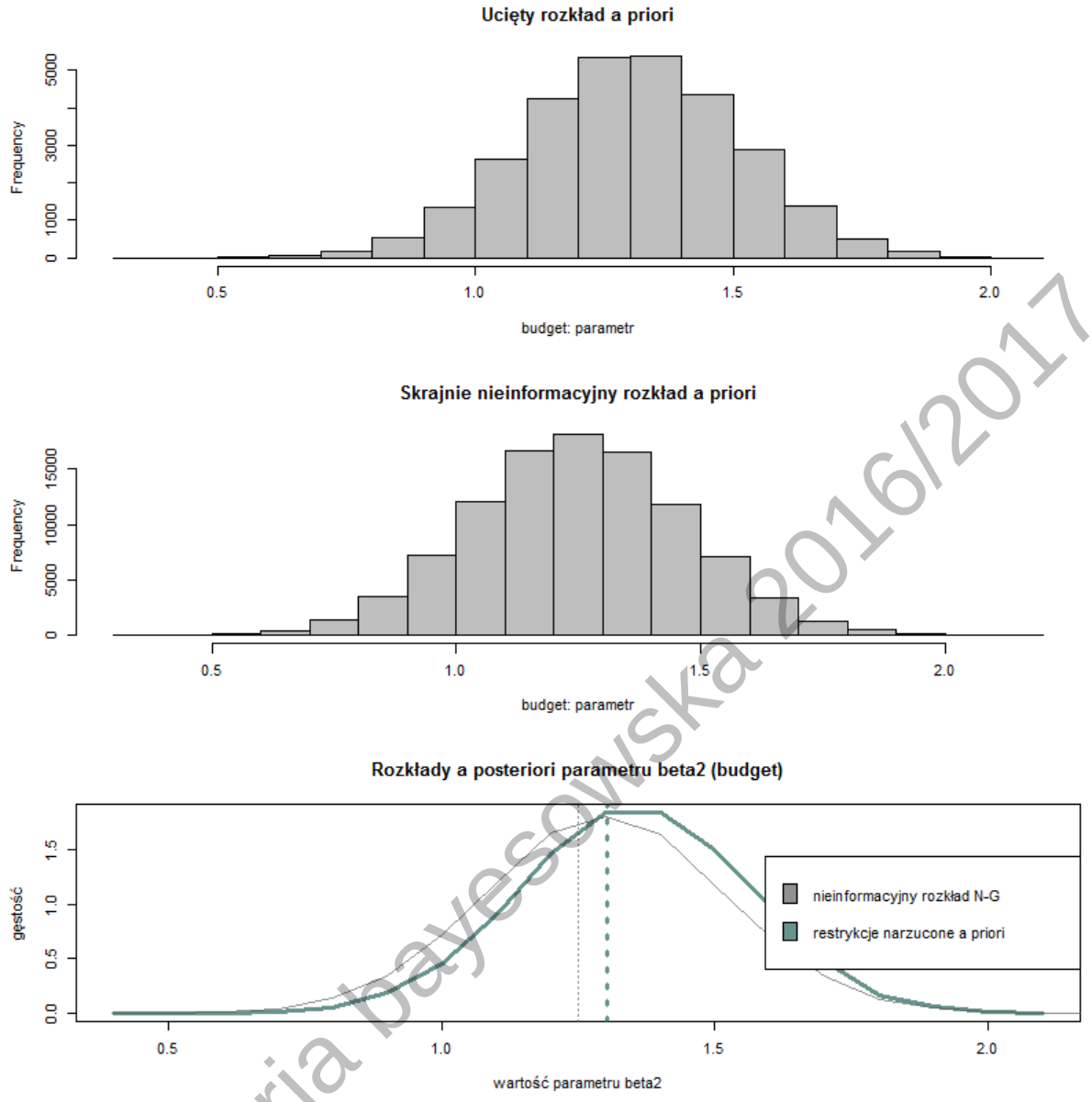
Nie udało się natomiast znaleźć wyników badań pozwalających na założenie *a priori* wielkości parametru przy zmiennej określającej długość trwania filmu. Jeśli chodzi o zmienne *Komentarze* oraz *Zewnetrzne_zapowiedzi,* jest to również utrudnione ze względu na specyfikę problemu. Nie możemy bowiem zakładać, że większa ilość komentarzy czy zapowiedzi przekłada się na to, że film jest bardziej lubiany, a co za tym idzie bardziej zyskowny. Należałoby przykładowo dokonać text miningu tych komentarzy i za pomocą analizy sentymentu stwierdzić jakie emocje towarzyszyły tym tekstom.

6. Importance Sampling

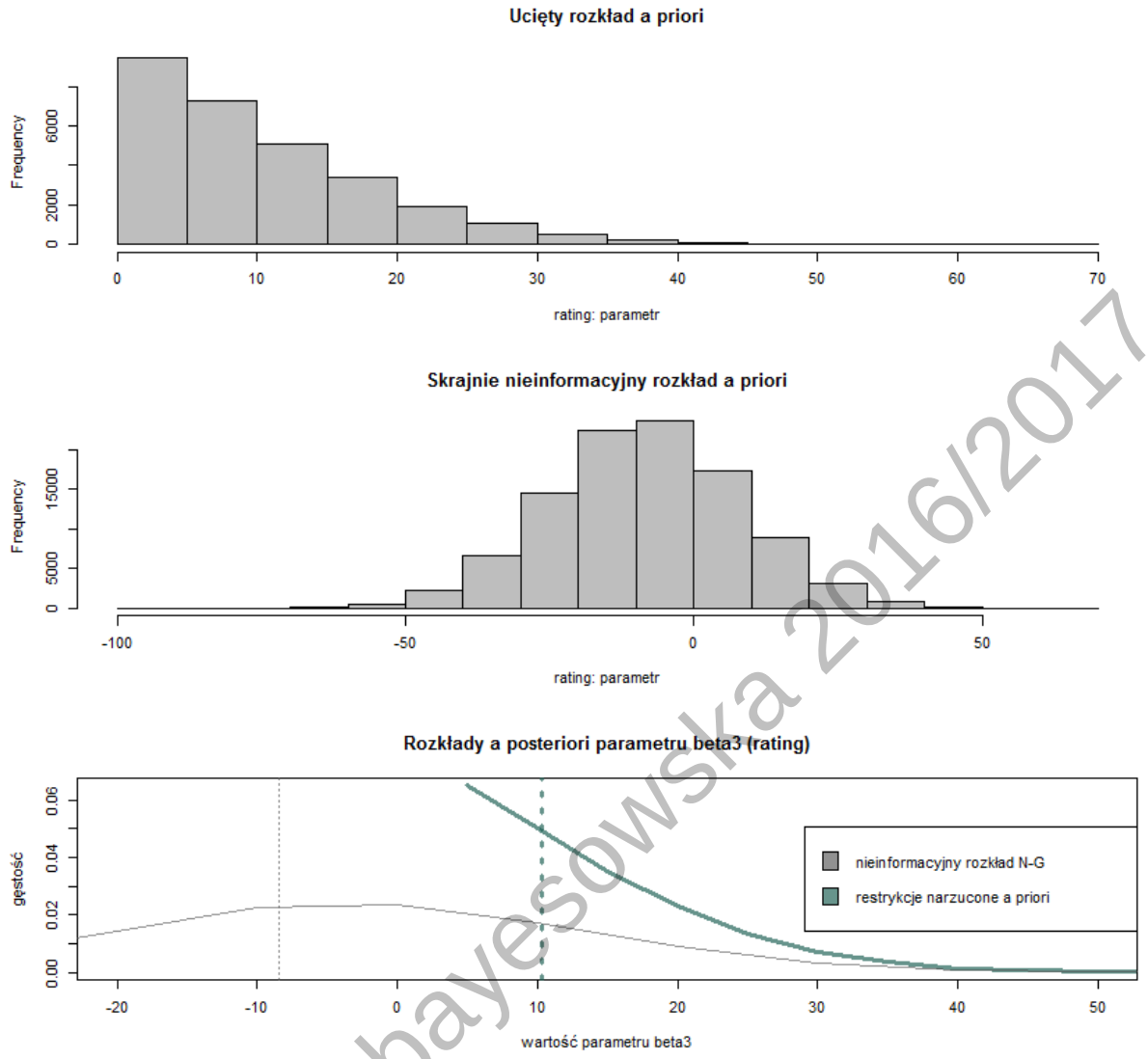
W celu oszacowania parametrów rozkładu *a posteriori* użyta została jedna z metod numerycznych, a dokładniej importance sampling. Parametry rozkładu *a posteriori* ustalone zostały ze skrajnie nieinformacyjnym rozkładem *a priori* N-G. Następnie nastąpiło 100000-krotne losowanie wektora parametrów za pomocą funkcji *q*, której dziedzina zawiera dziedzinę nieznanej funkcji gęstości *a posteriori*. Jak wcześniej zostało wspomniane narzuciliśmy restrykcję na parametry przy zmiennych *Opening_weekend*, *Budget* oraz *Rating*. Histogramy uciętych i skrajnie nieinformacyjnych rozkładów *a priori* oraz rozkładów *a posteriori* przedstawiają Rysunki 2-4.



Rysunek 2 Rozkłady dla zmiennej Opening_weekend



Rysunek 3 Rozkłądy dla zmiennej Budget



Rysunek 4 Rozkłady dla zmiennej Rating

Graficznie jesteśmy w stanie stwierdzić, że nałożenie restrykcji spowodowało znaczącą zmianę oszacowania parametru jedynie w przypadku zmiennej *Rating.*

Biorąc pod uwagę również 95% przedziały HPD przedstawione na Rysunku 5 możemy stwierdzić że nie doprowadziliśmy do zawężenia rozkładu ani w przypadku nie informacyjnego rozkładu N-G ani narzuconych restrykcji.

	Oszacowanie z ograniczeniami	Dolna granica HPD	Górna granica HPD
(Intercept)	-30.846	-171.861	100.689
opening_weekend	1.005	0.046	1.890
budget	1.303	0.896	1.699
rating	10.326	0.000	26.801
komentarze	0.039	0.021	0.056
zewnetrzne_zapowiedzi	-0.034	-0.105	0.040
czas	-0.229	-0.806	0.365

	Oszacowanie bez ograniczeń	Dolna granica HPD	Górna granica HPD
(Intercept)	100.767	-139.023	334.770
opening_weekend	1.124	0.073	2.220
budget	1.249	0.820	1.675
rating	-8.475	-40.338	24.026
komentarze	0.045	0.025	0.065
zewnetrzne_zapowiedzi	-0.032	-0.102	0.042
czas	-0.150	-0.746	0.461

	2.5 %	97.5 %
(Intercept)	-136.77507675	338.42284588
opening_weekend	0.04934213	2.19521344
budget	0.82314284	1.67681251
rating	-40.73531424	23.76346794
komentarze	0.02483633	0.06528315
zewnetrzne_zapowiedzi	-0.10402802	0.03977937
czas	-0.75151693	0.45205720

Rysunek 5 Porównanie przedziałów HPD

7. Podsumowanie

Podsumowując otrzymane wyniki możemy stwierdzić, że wniesienie informacji *a priori* na temat znaków parametrów nie doprowadziło do zawężenia rozkładów, w związku z czym można stwierdzić iż nie odnieśliśmy sukcesu w sensie Bayesowskim. Być może problemem było nieodpowiednie dobranie danych, tzn. ograniczenie się tylko do filmów nominowanych do nagrody Akademii Filmowej, co zniekształciło nam rzeczywiste zależności występujące między produkcjami filmowymi. Możliwe też, że bardziej adekwatne byłoby użycie innej metody statystycznej niż regresja logistyczna, tak jak zostało tp wskazane przez autorów innych badań.

8. Literatura

1. http://people.stern.nyu.edu/jsimonof/classes/2301/pdf/movies.pdf
2. http://cs229.stanford.edu/proj2011/YooKanterCummings-PredictingMovieRevenuesUsingImdbData.pdf
3. https://cran.r-project.org/web/packages/Boruta/Boruta.pdf
4. https://github.com/michalcisek/Filmy-oscarowe

9. Spis grafik