Mikołaj Jaworski

Analiza biznesowa wpływu różnych czynników na oceny użytkowników aplikacji mobilnych w Google Play Store z zastosowaniem metod uogólnionych modeli liniowych oraz imputacji wielokrotnej.

1. Wstęp

Na podstawie danych z końca 2017 r. Google Play Store był najsilniej rozwijającą się platformą do sprzedaży aplikacji. W swoich zasobach mieścił ponad 3,5 miliona ofert[1]. W stosunku do swojego największego rywala App Store, Google Play Store charakteryzuje się wysokim odsetkiem aplikacji darmowych. Większość deweloperów oferujących swój produkt oczekuje zysku w postaci dochodu z reklam lub zakupów dodatkowych funkcjonalności do aplikacji. W takim wypadku ich głównym założeniem jest skłanianie coraz większej liczby użytkowników do ściągnięcia aplikacji w celu zainteresowania coraz większej liczby reklamodawców i zwiększenia szans na sprzedaż płatnych dodatków.

Z drugiej strony jednym z głównym kryterium wyboru aplikacji przez użytkownika pozostaje jej ocena. W przedstawionej pracy podjęto próbę analizy czynników wpływających na uśrednioną wartość oceny aplikacji za pomocą uogólnionych modeli liniowych. W tym celu wykorzystano dane pozyskane metodą web-scrappingu, które zostały udostępnione publicznie za pośrednictwem strony Kaggle.com[2]. Zbiór zawiera podstawowe informacje o aplikacjach ściągniętych z Google Play Store wykorzystywanych do analizy rynku aplikacji oferowanych na platformie Android. Przedstawiony zbiór zawierał jednak liczne braki danych, które zostały wyeliminowane metodą imputacji wielokrotnej, a wyniki zastosowanej metody przedstawiono w dalszej części pracy. Do analizy wykorzystano dwa narzędzia: Python do przygotowania danych oraz SAS Base do analizy wstępnej i właściwej.

2. Opis zbioru danych

Analizowany zbiór danych zawiera 13 zmiennych o łącznej liczbie rekordów 10 840. Wśród danych uwzględnionych w zbiorze można znaleźć informacje m.in. na temat: nazwy aplikacji, kategorii tematycznej, uśrednionej oceny, liczby pobrań, liczby opinii, rozmiaru aplikacji, ceny oraz podstawowych informacji na temat kompatybilnego androida.

W wyniku wstępnej selekcji, czyszczenia oraz transformacji danych z wykorzystaniem języka Python wybrano 4 zmienne do modelu:

- *Rating* – zlogarytmizowana wartość uogólnionej oceny aplikacji
- *Reviews* – zlogarytmizowana wartość opinii na temat aplikacji
- *Installs* – zlogarytmizowana wartość pobrań aplikacji
- *Free* – zmienna binarna określająca, czy aplikacja jest darmowa.

Resztę zmiennych odrzucono na podstawie braku logicznego uzasadnienia wpływającego na ocenę aplikacji lub zbyt dużej liczby kategorii.

---

[1] https://blog.appfigures.com/ios-developers-ship-less-apps-for-first-time/
[2] https://www.kaggle.com/lava18/google-play-store-apps/version/5

3. Analiza właściwa

Hipoteza badawcza została sprawdzona z wykorzystaniem uogólnionych modeli liniowych. Do wypełnienia braków danych użyto metody wielokrotnej imputacji, w której wykorzystuje się rozkład zaobserwowanych danych, aby oszacować wartości wprowadzanych danych z uwzględnieniem różnych czynników losowych. Celem nadrzędnym tego procesu jest odtworzenie macierzy wariancji i kowariancji przy założeniu posiadania pełnego zbioru. Dla każdej imputacji budowany jest oddzielny model, a zebrane wyniki są uśredniane. Poniżej szczegółowo został przedstawiony proces analityczny składający się z trzech etapów wraz z interpretacją otrzymanych rezultatów.

1. Model imputacji danych

Procedura MI pozwala na stworzenie 100 wariantów kompletnych zbiorów danych (wypełniona zmienna *Rating*). Wysoka liczba imputacji pozwala na m.in. zwiększenie dokładności estymacji FMI (*Fraction of Missing Information*) oraz obniżenia średniego błędu kwadratowego. Ze względu na niespełnione założenie o łączności rozkładów wszystkich zmiennych, zastosowano metodę FCS (*fully conditional methods*), która opiera się na warunkowych rozkładach dla każdej zmiennej oddzielnie. Wypełnianie wykonano za pomocą regresji logistycznej ze skumulowanym logitem jako funkcją łączącą, która jest odpowiednią metodą w przypadku danych skategoryzowanych i uporządkowanych, jak właśnie oceny aplikacji przez użytkowników. Rezultaty głównej procedury MI zostały omówione w analizie wstępnej (wykresy śladowe nie są generowane dla zmiennych skategoryzowanych).

2. Model analityczny

Procedura GENMOD dopasowuje uogólnione modele liniowe do danych. Są one rozszerzeniem tradycyjnych modeli linowych (zakładających normalność rozkładu), które pozwala na modelowanie zależności liniowego predyktora poprzez nieliniową funkcję łączącą i funkcję wariancji[3]. Biorąc pod uwagę zmienną objaśnianą, przy modelowaniu przyjęto rozkład Poissona. Tak jak wspomniano wcześniej, dla każdej imputacji tworzony jest nowy model.

3. Ewaluacja modeli

Procedura MIANALYZE bierze pod uwagę oszacowania parametrów i błędy standardowe wszystkich wygenerowanych wcześniej modeli w celu stworzenia jednego, uśrednionego zestawienia statystyk do oceny otrzymanych modeli. Poniżej przedstawiono i poddano interpretacji wyniki całego procesu.

[3] https://support.sas.com/resources/papers/proceedings16/8380-2016.pdf

Tabela 7. Podsumowanie modeli – wariancja.

Variance Information (100 Imputations)								
Parameter	**Free**	**Variance**			**DF**	**Relative Increase in Variance**	**Fraction Missing Information**	**Relative Efficiency**
		Between	**Within**	**Total**				
Intercept		0.000005957	0.001080	0.001086	3.23E6	0.005570	0.005540	0.999945
Installs		0.000000299	0.000048977	0.000049279	2.64E6	0.006159	0.006122	0.999939
Reviews		0.000000338	0.000050694	0.000051035	2.22E6	0.006728	0.006684	0.999933
Free	**0**	0.000002569	0.001067	0.001070	1.68E7	0.002432	0.002427	0.999976
Free	**1**	0	0	0	.	.	.	.

Źródło: opracowanie własne.

Kolumna *Variance Between* wskazuje jaka zmienność występuje pomiędzy wyestymowanymi parametrami ze wszystkich 100 modeli. Przykładowo, wariancja oszacowanych parametrów przy zmiennej *Installs* wynosi 0.000000299. Można ją interpretować jako niepewność wynikającą z brakujących danych. *Variance Within* jest średnią arytmetyczną obliczoną z błędów standardowych. Mówi, jaka byłaby zmienność przy założeniu kompletności zbioru. *Variance Total* odzwierciedla sumaryczną wariancję ze wszystkich jej źródeł. Wraz ze wzrostem liczby imputacji rośnie precyzja oszacowania tym samym maleje wariancja sumaryczna. *DF* - liczba stopni domyślnie zbiega do nieskończoności (przy założeniu o normalności rozkładu estymatora) i wzrasta wraz z liczbą imputacji. W przypadku małych zbiorów konieczne jest dostosowywanie algorytmu, ale nie dotyczy to przypadku opisanego w pracy. *Relative Increase in Variance (RIV)* odzwierciedla wzrost zmienności wynikającej z braków danych. Przykładowo dla zmiennej *Free* parametr ten wynosi 0.006728 co oznacza, że zmienność jest o 0.67% wyższa niż w przypadku kompletnych danych. Kolumna *Fraction of Missing Information (FMI)* jest bezpośrednio związana z *RIV*. *Relative Efficiency (RE)* określa skuteczność liczby imputacji w stosunku do procentu braków danych. Dla zbioru użytego w pracy 100 imputacji zapewnia efektywność estymacji parametrów.

Tabela 8. Podsumowanie modeli – parametry.

Parameter Estimates (100 Imputations)											
Parameter	**Free**	**Estimate**	**Std Error**	**95% Confidence Limits**		**DF**	**Minimum**	**Maximum**	**Theta0**	**t for H0: Parameter=Theta0**	**Pr > |t|**
Intercept		0.429892	0.032958	0.36530	0.49449	3.23E6	0.424285	0.437877	0	13.04	<.0001
Installs		-0.027426	0.007020	-0.04118	-0.01367	2.64E6	-0.029102	-0.026035	0	-3.91	<.0001
Reviews		0.031349	0.007144	0.01735	0.04535	2.22E6	0.029919	0.032884	0	4.39	<.0001
Free	**0**	-0.011722	0.032704	-0.07582	0.05238	1.68E7	-0.016481	-0.007787	0	-0.36	0.7200
Free	**1**	0	0	.	.	.	0	0	0	.	.

Źródło: opracowanie własne.

Wartości *p-value* wskazują na istotność zmiennych *Installs, Reviews* oraz wyrazu wolnego. Natmiast zmienna *Free* jest nieistotna statystycznie. Jeżeli chodzi o wartości estymatorów, przykładowo wzrost zlogarytmizowanej zmiennej *Reviews* o jednostkę sprawia, że można się spodziewać, że zlogarytmizowana wartość *Rating* wzośnie o 3%.

4. **Porównanie wyników**

Dla porównania wyników przeprowadzono 3 rodzaje rozwiązań: imputację wielokrotną, usunięcie rekordów z brakami danych (*Complete Case Analysis*) oraz uzupełnienie braków danych medianą. Poniższa tabela przedstawia oszacowania parametrów dla modeli z różnymi metodami traktowania braków.

Tabela 9. Oszacowanie parametrów modelu dla 3 rodzajów imputacji

		Imputacja wielokrotna			Complete Case			Imputacja medianą		
Parameter		**Estimate**	**SE**	**P-value**	**Estimate**	**SE**	**P-value**	**Estimate**	**SE**	**P-value**
Intercept		0.429892	0.032958	<.0001	0.3985	0.0438	<.0001	0.4115	0.0329	<.0001
Installs		-0.027426	0.007020	<.0001	-0.0229	0.0084	0.0063	-0.0198	0.0070	0.0045
Reviews		0.031349	0.007144	<.0001	0.0282	0.0082	0.0006	0.0228	0.0071	0.0013
Free	0	-0.011722	0.032704	0.7200	-0.0034	0.0369	0.9259	-0.0080	0.0327	0.8061
Free	1	0	0	.	0	0	.	0	0	.

Źródło: opracowanie własne.

Dla każdego rodzaju imputacji istotnymi statystycznie zmiennymi są zmienne *Installs, Reviews* oraz wyraz wolny. Zmienna *Free* nie jest istotna statystycznie. Wartości błędów standardowych dla modelu dla imputacji wielokrotnej i imputacji medianą są zbliżone, najwyższe wartości błędu standardowego uzyskujemy dla modelu uzyskanego na danych po usunięciu rekordów z brakami danych.

Porównując otrzymane modele należy również wziąć pod uwagę miary dopasowania, które opisują jak dobrze model dopasowuje się do danych empirycznych. Poniższa tabela prezentuje miary dla 3 rodzajów imputacji. Najlepsze wartości miar dopasowania uzyskujemy dla modelu otrzymanego po usunięciu wszystkich braków danych (użyto 9366 rekordów). Natomiast najsłabsze wyniki pod względem tych statystyk otrzymał model po imputacji medianą.

Tabela 10. Miary dopasowania dla otrzymanych modeli.

	Imputacja wielokrotna	**Complete Case**	**Imputacja medianą**
Log Likelihood	-9970.3337	-8615.8407	-9957.0611
AIC	25131.3684	21714.5983	25152.2114
BIC	25160.5324	21743.1777	25181.3754
Pearson Chi-Square	157.3524	135.5239	139.4881

Źródło: opracowanie własne.

5. Bibliografia

1. iOS Developers Ship 29% Fewer Apps in 2017, the First Ever Decline – And More Trends to Watch. (2018). A*ppfigures*, https://blog.appfigures.com/ios-developers-ship-less-apps-for-first-time/ *(dostęp: 07.12.2018)*
2. Korczyński, A., *Własności estymatorów - porównanie kalibracji i imputacji wielokrotnej*, http://kolegia.sgh.waw.pl/pl/KAE/struktura/ISiD/struktura/ZAHZiAW/publikacje/Documents/Adam_Korczy%C5%84ski_Statystyk_zastosowania_biznesowe_i_spoleczne_pp_183_194.pdf *(dostęp 08.12.2018)*

3. Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data*. 2nd ed. New York: John Wiley
4. Stats.idre.ucla.edu. (2018). *Multiple Imputation in SAS Part 1*, https://stats.idre.ucla.edu/sas/seminars/multiple-imputation-in-sas/mi_new_1/?fbclid=IwAR3nT7I6gWRNr2YRF0-t1RaiighycamLabZKTjqLz_y7hdnQY88F5OJh_2s *(dostęp: 08.12.2018)*
5. Support.sas.com. *SAS/STAT(R) 9.3 User's Guide*, http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_mi_sect008.htm *(dostęp: 07.12.2018)*
6. SAS/STAT® 9.2 User's Guide The GENMOD Procedure (Book Excerpt). (2008). 1st ed. https://support.sas.com/documentation/cdl/en/statuggenmod/61787/PDF/default/statuggenmod.pdf *(dostęp: 08.12.2018)*