Politechnika Krakowska im. Tadeusza Kościuszki w Krakowie

Wydział Fizyki, Matematyki i Informatyki

Kierunek: Matematyka

Porównanie wybranych metod klasyfikacji binarnej w problemie prognozowania ocen filmów

Joanna Zając

1 Opis problemu

1.1 Cel projektu

Celem projektu będzie stworzenie dwóch modeli klasyfikacji, których zadaniem będzie stwierdzenie, czy dany użytkownik oceni dany film jako `dobry (1)` lub `slaby (0)`. W tym celu arbitralnie dobrano podział skali ocen, w wyniku którego jako film `dobry` uznaje się taki, który uzyskał od danego użytkownika ocenę 4.0 lub wyższą (w skali 0.5-5 w krokiem 0.5). Decyzja ta będzie podejmowana na podstawie następujących atrybutów: `movieId, userId, genres, tag, relevance`. Nie każdy użytkownik określił film tagiem, z tego powodu analiza zostanie ograniczona do danych, które zawierają wszystkie powyższe atrybuty.

1.2 Pochodzenie danych

Zbiór danych `ml-20m` pochodzi z serwisu movielens.org, gdzie użytkownicy oceniają film w skali 0.5-5 co pół stopnia oraz przypisują mu dowolny tag w postaci tekstu, który ma opisać dany film.
Znajduje się w nim 20000263 ocen i 465564 tagów opisujących 27278 filmów. Dane te zostały stworzone przez 138493 użytkowników pomiędzy 09-01-1995 a 31-03-2015. Użytkownicy zostali wyłonienie losowo, każdy z nich ocenił co najmniej 20 filmów. Dane składają się z 6 plików: `‘genome-scores.csv‘, ‘genome-tags.csv‘, ‘links.csv‘, ‘movies.csv‘, ‘ratings.csv‘ i ‘tags.csv‘`.

1.3 Struktura danych

movies.csv

Zawiera dane o filmach. Każda linia reprezentuje jeden film w formacie `movieId,title,genres`. Tytuł zawiera rok premiery podany w nawiasach. `Genres` to lista gatunków filmu wybierana spośród

- Action
- Adventure
- Animation
- Children’s
- Comedy
- Crime
- Documentary
- Drama
- Fantasy
- Film-Noir
- Horror
- Musical
- Mystery
- Romance
- Sci-Fi
- Thriller
- War
- Western
- (no genres listed)

links.csv

Zawiera linki do filmów w innych serwisach. Każda linia reprezentuje jeden film w formacie `movieId,imdbId,tmdbId`. Ten plik nie będzie wykorzystywany w projekcie.

`movieId` to identyfikator używany na stronie movielens.org, np. www.movielens.org/movies/1.

`imdcId` to identyfikator używany na stronie www.imdb.com, np. www.imdb.com/title/tt0114709.

`tmdbId` to identyfikator używany na stronie themoviedb.org, np. www.themoviedb.org/movie/862.

tags.csv

Zawiera tagi dopisane do filmów przez użytkowników. Struktura: `userId,movieId,tag,timestamp`.

genome-scores.csv

Plik zawiera dane na temat dopasowania tagu do filmu. Dane są postaci `movieId,tagId,relevance`.

genome-tags.csv

Zawiera Id tagu oraz tag, dane są postaci `tagId,tag`.

ratings.csv

Zawiera oceny filmów przez danego użytkownika. Struktura: `userId,movieId,rating,timestamp`.

1.4 Typ danych i dane statystyczne

Dane z plików `'genome-scores.csv', 'genome-tags.csv', 'movies.csv', 'ratings.csv', 'tags.csv'.` zostały połączone w jedną tabelę z wyselekcjonowanymi kolumnami. Po zaimportowaniu każda kolumna zawierała dane typu `string`, należało więc odpowiednio je przekształcić. Dodano także kolumnę `label`, która zawiera klasę, czyli informację, czy dany użytkownik uznał film na `dobry` czy `słaby`. Sytuację "przed" i "po" zamianie typów danych przedstawia poniższy schemat:

```
root
 |-- movieId: string (nullable = true)
 |-- tagId: string (nullable = true)
 |-- tag: string (nullable = true)
 |-- userId: string (nullable = true)
 |-- rating: string (nullable = true)
 |-- title: string (nullable = true)
 |-- genres: string (nullable = true)
 |-- relevance: string (nullable = true)

root
 |-- movieId: integer (nullable = true)
 |-- tagId: integer (nullable = true)
 |-- tag: string (nullable = true)
 |-- userId: integer (nullable = true)
 |-- rating: double (nullable = true)
 |-- title: string (nullable = true)
 |-- genres: string (nullable = true)
 |-- relevance: double (nullable = true)
 |-- label: double (nullable = true)
```

Zobaczmy jak wyglądają dane poprzez wyświetlenie pierwszych rekordów.

movieId	tagId	tag	userId	rating	title	genres	relevance	label
100010	null	null	24994	4.0	Battle of Los Ang...	Action|Sci-Fi	null	1.0
100010	null	null	113075	4.0	Battle of Los Ang...	Action|Sci-Fi	null	1.0
100010	null	null	41267	2.0	Battle of Los Ang...	Action|Sci-Fi	null	0.0
100010	null	null	49817	3.0	Battle of Los Ang...	Action|Sci-Fi	null	0.0
100010	null	null	46470	3.0	Battle of Los Ang...	Action|Sci-Fi	null	0.0
100010	null	null	16693	3.5	Battle of Los Ang...	Action|Sci-Fi	null	0.0
100010	null	null	102118	0.5	Battle of Los Ang...	Action|Sci-Fi	null	0.0
100010	null	null	106476	2.5	Battle of Los Ang...	Action|Sci-Fi	null	0.0
100010	null	null	12131	1.5	Battle of Los Ang...	Action|Sci-Fi	null	0.0
100010	null	null	61728	5.0	Battle of Los Ang...	Action|Sci-Fi	null	1.0
100010	null	null	127063	3.0	Battle of Los Ang...	Action|Sci-Fi	null	0.0
100010	null	null	30507	3.0	Battle of Los Ang...	Action|Sci-Fi	null	0.0
100010	null	null	44101	2.0	Battle of Los Ang...	Action|Sci-Fi	null	0.0
100010	null	null	73026	2.5	Battle of Los Ang...	Action|Sci-Fi	null	0.0
100010	null	null	53478	4.0	Battle of Los Ang...	Action|Sci-Fi	null	1.0
100010	null	null	4347	3.0	Battle of Los Ang...	Action|Sci-Fi	null	0.0
100010	null	null	75603	5.0	Battle of Los Ang...	Action|Sci-Fi	null	1.0
100010	null	null	94445	2.0	Battle of Los Ang...	Action|Sci-Fi	null	0.0
100010	null	The Asylum	67075	0.5	Battle of Los Ang...	Action|Sci-Fi	null	0.0
100010	null	null	135806	3.0	Battle of Los Ang...	Action|Sci-Fi	null	0.0
100010	null	null	5352	3.5	Battle of Los Ang...	Action|Sci-Fi	null	0.0
100010	null	null	20180	1.0	Battle of Los Ang...	Action|Sci-Fi	null	0.0
100248	789	plot twist	22178	3.5	Lo (2009)	Comedy|Horror|Rom...	null	0.0
100553	null	null	55414	5.0	Frozen Planet (2011)	Documentary	null	1.0
100553	null	null	72860	5.0	Frozen Planet (2011)	Documentary	null	1.0
100553	null	null	16085	2.5	Frozen Planet (2011)	Documentary	null	0.0
100553	null	null	70984	5.0	Frozen Planet (2011)	Documentary	null	1.0
100553	null	null	87931	3.0	Frozen Planet (2011)	Documentary	null	0.0
100553	null	null	57709	5.0	Frozen Planet (2011)	Documentary	null	1.0
100553	null	null	1705	4.0	Frozen Planet (2011)	Documentary	null	1.0

Jak wspomniano wcześniej, przy wielu rekordach brakuje określenia tagu, a więc jednocześnie jego Id i relevance. Po usunięciu wierszy z brakującymi danymi sytuacja przedstawia się następująco:

movieId	tagId	tag	userId	rating	title	genres	relevance	label
1007	762	over the top	117270	3.5	Apple Dumpling Ga...	Children|Comedy|W...	0.1255	0.0
100729	887	sci-fi	102118	3.0	Starship Troopers...	Action|Animation|...	0.9757499999999999	0.0
101612	430	genius	4450	3.0	Admission (2013)	Comedy|Romance	0.18	0.0
103539	840	realistic	4450	3.0	The Spectacular N...	Comedy|Drama|Romance	0.7365	0.0
103539	840	realistic	54612	4.5	The Spectacular N...	Comedy|Drama|Romance	0.7365	1.0
103539	840	realistic	41273	4.5	The Spectacular N...	Comedy|Drama|Romance	0.7365	1.0
104241	168	brutal	47594	3.5	Kick-Ass 2 (2013)	Action|Comedy|Crime	0.88175	0.0
104241	168	brutal	10514	3.0	Kick-Ass 2 (2013)	Action|Comedy|Crime	0.88175	0.0
104241	168	brutal	122523	4.0	Kick-Ass 2 (2013)	Action|Comedy|Crime	0.88175	1.0
104241	168	brutal	88738	1.0	Kick-Ass 2 (2013)	Action|Comedy|Crime	0.88175	0.0
104374	375	family bonds	131900	4.0	About Time (2013)	Drama|Fantasy|Rom...	0.9470000000000001	1.0
104374	375	family bonds	3029	4.0	About Time (2013)	Drama|Fantasy|Rom...	0.9470000000000001	1.0
104374	375	family bonds	4450	3.0	About Time (2013)	Drama|Fantasy|Rom...	0.9470000000000001	0.0
104374	375	family bonds	10616	4.5	About Time (2013)	Drama|Fantasy|Rom...	0.9470000000000001	1.0
1094	110	based on a true s...	25737	5.0	Crying Game, The ...	Drama|Romance|Thr...	0.1355	1.0
1094	110	based on a true s...	9815	5.0	Crying Game, The ...	Drama|Romance|Thr...	0.1355	1.0
1094	110	based on a true s...	10573	4.0	Crying Game, The ...	Drama|Romance|Thr...	0.1355	1.0
109487	860	robot	33323	3.0	Interstellar (2014)	Sci-Fi|IMAX	0.8280000000000001	0.0
110	508	historical	96372	4.5	Braveheart (1995)	Action|Drama|War	0.9922500000000001	1.0
110	508	historical	63781	5.0	Braveheart (1995)	Action|Drama|War	0.9922500000000001	1.0
110	508	historical	1678	2.5	Braveheart (1995)	Action|Drama|War	0.9922500000000001	0.0
110	508	historical	72257	5.0	Braveheart (1995)	Action|Drama|War	0.9922500000000001	1.0
110	508	historical	131900	4.5	Braveheart (1995)	Action|Drama|War	0.9922500000000001	1.0
110	508	historical	84441	4.0	Braveheart (1995)	Action|Drama|War	0.9922500000000001	1.0
110	508	historical	23982	4.0	Braveheart (1995)	Action|Drama|War	0.9922500000000001	1.0
110	508	historical	47866	2.5	Braveheart (1995)	Action|Drama|War	0.9922500000000001	0.0
110	508	historical	96792	5.0	Braveheart (1995)	Action|Drama|War	0.9922500000000001	1.0
110	508	historical	76878	4.0	Braveheart (1995)	Action|Drama|War	0.9922500000000001	1.0
110	508	historical	119367	5.0	Braveheart (1995)	Action|Drama|War	0.9922500000000001	1.0
110	508	historical	40466	5.0	Braveheart (1995)	Action|Drama|War	0.9922500000000001	1.0

Sprawdźmy ilu unikatowych użytkowników i ile filmów pozostało w tabeli po przefiltrowaniu brakujących danych, ile tagów i gatunków filmowych danego rodzaju pojawia się w tabeli oraz policzmy podstawowe statystyki ocen.

```
Liczba użytkowników: 5102
Liczba filmów: 7061
Liczba wierszy: 183146
```

Liczba poszczególnych ocen wynosi

```
+------+-----+
|rating|count|
+------+-----+
|   5.0|40602|
|   4.5|33847|
|   4.0|45581|
|   3.5|24623|
|   3.0|16688|
|   2.5| 7749|
|   2.0| 6457|
|   1.5| 2671|
|   1.0| 2881|
|   0.5| 2047|
+------+-----+
```

zatem łatwo też określić liczbę poszczególnych rekordów w klasach.

```
+-----+------+
|label| count|
+-----+------+
|  0.0| 63116|
|  1.0|120030|
+-----+------+
```

Liczebność danych gatunków filmowych i tagów:

```
+--------------------+-----+
|              genres|count|
+--------------------+-----+
|               Drama|11870|
|              Comedy| 5995|
|        Comedy|Drama| 5984|
|       Drama|Romance| 5884|
|Comedy|Drama|Romance| 5236|
|         Crime|Drama| 4684|
|Action|Sci-Fi|Thr...| 3762|
|      Comedy|Romance| 3757|
|Action|Adventure|...| 3536|
|      Drama|Thriller| 2899|
|    Mystery|Thriller| 2631|
|Action|Adventure|...| 2626|
|Action|Crime|Dram...| 2547|
|Crime|Drama|Thriller| 2301|
|Drama|Mystery|Sci...| 2132|
|Action|Adventure|...| 2087|
|Comedy|Crime|Dram...| 1963|
|Drama|Mystery|Thr...| 1942|
|Action|Crime|Thri...| 1856|
|         Documentary| 1753|
+--------------------+-----+
```

```
+------------------+-----+
|               tag|count|
+------------------+-----+
|            sci-fi| 3256|
|       atmospheric| 2738|
|            comedy| 2406|
|            action| 2385|
|           surreal| 2280|
|      twist ending| 2241|
|   based on a book| 2225|
|             funny| 1924|
|          dystopia| 1897|
|       dark comedy| 1814|
|          stylized| 1801|
|            quirky| 1776|
|        psychology| 1662|
|           classic| 1659|
|           fantasy| 1587|
|       time travel| 1451|
|           romance| 1443|
|visually appealing| 1417|
| thought-provoking| 1370|
|        disturbing| 1349|
+------------------+-----+
```

Podstawowe dane statystyczne atrybutów ilościowych:

```
+-------+------------------+--------------------+
|summary|            rating|           relevance|
+-------+------------------+--------------------+
|  count|            183146|              183146|
|   mean|3.8990068033153875|  0.7957503248774205|
| stddev|0.9941301126725549| 0.22226520529827654|
|    min|               0.5|0.009000000000000008|
|    max|               5.0|                 1.0|
+-------+------------------+--------------------+
```

Średnia ocen to około 3.9, zatem wybór oceny 4.0 jako granicy pomiędzy klasami wydaje się uzasadniona. Zgodnie z tym faktem klasę `dobry` można zdefiniować jako film z oceną powyżej

średniej wszystkich ocen.
Struktura danych (liczba danych w kategoriach w każdym zbiorze) po podziale na zbiór treningowy i testowy w stosunku 70-30 przedstawia się następująco.

```
Liczba danych w zbiorze treningowym: 128088
Liczba danych w zbiorze testowym: 55058
Struktura w zbiorze treningowym:
+-----+-----+
|label|count|
+-----+-----+
|  0.0|44275|
|  1.0|83813|
+-----+-----+

Struktura w zbiorze testowym:
+-----+-----+
|label|count|
+-----+-----+
|  0.0|18841|
|  1.0|36217|
+-----+-----+
```

2 Opis zastosowanych metod uczenia maszynowego

2.1 Drzewa decyzyjne

Drzewa decyzyjne są ważnym narzędziem w uczeniu maszynowym i eksploracji danych. Są wykorzystywane między innymi w problemie klasyfikacji. Główne składowe drzewa to korzeń oraz gałęzie, które łączą korzeń z kolejnymi wierzchołkami. Wierzchołki, z których wychodzi co najmniej jedna krawędź, są nazywane węzłami, natomiast wierzchołki z których nie wychodzą krawędzie to tzw. liście. Drzewo zaczyna od pojedynczego węzła reprezentującego cały zbiór treningowy. W każdym węźle sprawdzany jest pewien warunek dotyczący danej obserwacji, i na jego podstawie wybierana jest odpowiednia gałąź prowadząca do kolejnego wierzchołka. Klasyfikacja danej obserwacji polega na przejściu od korzenia do liścia i przypisaniu do tej obserwacji klasy zapisanej w danym liściu.

Zalety tej metody to między innymi

- czytelna dla człowieka forma reprezentacji (można je opisać graficznie - grafy skierowane),
- możliwość reprezentowania dowolnie złożonych pojęć pojedynczych lub wielokrotnych, jeżeli tylko ich definicje da się wyrazić w zależności od atrybutów,
- efektywność obliczeniowa - wyznaczenie kategorii przykładu wymaga w najgorszym razie przetestowania raz wszystkich jego atrybutów, często może wystarczyć ich niewielka część,
- możliwość przejścia od drzew decyzyjnych do reguł decyzyjnych.

Do wad metody drzew decyzyjnych można zaliczyć

- wysokie koszty reprezentacji alternatyw pomiędzy atrybutami (w przeciwieństwie do koniunkcji, która jest zapisywana jako pojedyncza droga od korzenia do liścia),
- możliwość testowania jednego atrybutu na raz - powoduje to rozrost drzewa dla danych gdzie poszczególne atrybuty zależą od siebie,
- trudności w aktualizowaniu - algorytmy udoskonalające gotowe już drzewa są bardzo złożone i zazwyczaj ich wynikiem jest drzewo gorszej jakości niż drzewo budowane od początku z kompletnym zestawem przykładów.

2.2 Regresja logistyczna

Regresja logistyczna jest jedną z metod regresji, którą stosuje się w przypadku, gdy zmienna zależna jest zmienną dychotomiczną, czyli przyjmuje tylko dwie wartości, najczęściej 0 i 1. Problemy, które możemy rozwiązać przy pomocy regresji logistycznej dotyczą klasyfikacji, czyli określenia prawdopodobieństwa sukcesu/porażki, śmierci/przeżycia itp. W takich przypadkach zwykła regresja liniowa nie jest dobrym wyborem, ponieważ dla dychotomicznej zmiennej objaśnianej regresja liniowa będzie szacowała wartości spoza akceptowalnego zakresu (poniżej 0 lub powyżej 1). Ponadto nie będą spełnione założenia modelu regresji liniowej takie jak rozkład normalny dla reszt oraz jednorodność wariancji.

W regresji logistycznej założenia są następujące

- zmienna Y podlega rozkładowi dwumianowemu, $Y\ B(1,p)$,
- wartości wyjściowe są statystycznie niezależne,
- wartość oczekiwana $\mathbb{E}(y|x) = P(x)$ jest obliczana na podstawie funkcji logistycznej

$$f(x) = \frac{e^x}{1+x^x}.$$

Funkcja logistyczna przyjmuje wartości z przedziału $[0,1]$, przy czym 0 i 1 są wartościami brzegowymi. Funkcja logistyczna dla początkowych argumentów przyjmuje wartości bliskie zera. Od momentu osiągnięcia wartości progowej następuje nagły wzrost wartości, a po osiągnięciu pewnej wartości, dla kolejnych argumentów funkcja przyjmuje wartości bliskie jedynki. W przeciwieństwie do regresji liniowej, w regresji logistycznej do estymacji parametrów używa się metody największej wiarygodności. Wiarygodność danego modelu jest określana jako łączne prawdopodobieństwo otrzymania obserwowanych wartości wyjściowych wyrażonych za pomocą funkcji wybranego modelu regresji.

Główne zalety regresji logistycznej to

- prosta transformacja prawdopodobieństwa $P(y|x)$,
- dobrze radzi sobie nawet przy dużej liczbie obserwacji,
- ciągły wynik, sprowadzany do binarnego - z powodzeniem zatem może być stosowana w systemach rekomendacji, z dowolną metodą sortowania wyników,
- znany rozkład dwumianowy zmiennej objaśnianej.

W metodzie tej napotykamy też na kilka problemów.

- Problem selekcji zmiennych - zbyt wiele zmiennych może zmniejszać siłę dyskryminacji. Należy więc wybrać właściwy model.
- Zbyt mała próba może dawać niepewne współczynniki korelacji.
- Regresja logistyczna działa tylko dla problemów dwuklasowych - dla wieloklasowych należy użyć metody np. One vs Rest.

3 Ocena modeli

Ocena modeli oparta zostanie o następujące miary:

- `time` - czas potrzebny na dopasowanie modelu oraz przeprowadzenie przewidywań,
- `accuracy` - dokładność, miara dana wzorem $\frac{TP+TN}{TP+TN+FP+FN}$, określa jaką część wszystkich prognoz stanowią prognozy poprawne,
- `error`, czyli 1-`accuracy` - jaką część wszystkich predykcji stanowią te błędne,
- `area under ROC` - pole pod krzywą ROC (Receiver Operating Characteristic Curves),
- `recall` (in. sensitivity) - czułość, miara dana wzorem $\frac{TP}{TP+FN}$, określa prawdopodobieństwo, że klasyfikacja będzie poprawna pod warunkiem, że film jest `dobry`,
- `precision` - precyzja, miara dana wzorem $\frac{TP}{TP+FP}$, określa jakie jest prawdopodobieństwo, że film rzeczywiście jest `dobry`, gdy wynik predykcji to `dobry`,
- `specificity` - specyficzność, miara dana wzorem $\frac{TN}{TN+FP}$, określa prawdopodobieństwo, że klasyfikacja będzie poprawna pod warunkiem że film jest `słaby`,
- `f1` - średnia harmoniczna `precision` oraz `recall` dana wzorem $2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$. Miara ta daje ocenę balansu między czułością a precyzją, nie uwzględnia wyników prawdziwie negatywnych.

3.1 Drzewa decyzyjne

Po wyświetleniu początkowych predykcji model wygląda obiecująco.

```
+-----+----------+--------------------+
|label|prediction|         probability|
+-----+----------+--------------------+
|  1.0|       1.0|[0.33179608313649...|
|  1.0|       1.0|[0.33179608313649...|
|  1.0|       1.0|[0.33179608313649...|
|  1.0|       1.0|[0.33179608313649...|
|  1.0|       1.0|[0.33179608313649...|
|  1.0|       1.0|[0.33179608313649...|
|  1.0|       1.0|[0.33179608313649...|
|  1.0|       1.0|[0.33179608313649...|
|  1.0|       1.0|[0.33179608313649...|
|  1.0|       1.0|[0.33179608313649...|
+-----+----------+--------------------+
```

Miary jakości modelu jednak nie do końca to potwierdzają.

```
dt time 0:04:32.011660
True Positives: 34794
True Negatives: 1834
False Positives: 17007
False Negatives: 1423
Total 55058
DenseMatrix([[ 1834., 17007.],
             [ 1423., 34794.]])
recall 0.9607
precision 0.6717
specificity 0.0973
accuracy 0.6653
error 0.3347
weightedPrecision 0.6345
f1 0.5769
area under ROC curve 0.5317
```

Najwięcej wątpliwości w kwestii jakości modelu może budzić wartość `specificity` - model nie radzi sobie z wykrywaniem `słabych` filmów. Niskie wartości przyjmuje także powierzchnia pod krzywą ROC - wynik 0.5305 nie różni się praktycznie od przypadku, gdyby ktoś losowo dopasowywał kategorie `dobry` oraz `słaby` do każdego filmu.
Macierz pomyłek wskazuje, tak jak wartość `specificity`, że model ma problem z rozpoznawaniem `słabych` filmów, które często klasyfikuje jako `dobry`. Może być to wynikiem dysproporcji pomiędzy klasami. W obu zbiorach, treningowym i testowym, znajduje się dwa razy więcej etykiet `1` niż etykiet `0`.

3.2 Regresja logistyczna

Podobnie jak w przypadku drzewa decyzyjnego, po wyświetleniu kilku pierwszych predykcji model wygląda dobrze.

```
+-----+----------+--------------------+
|label|prediction|         probability|
+-----+----------+--------------------+
|  1.0|       1.0|[0.26743982737662...|
|  1.0|       1.0|[0.26753386074489...|
|  1.0|       1.0|[0.12711962288167...|
|  1.0|       1.0|[0.12714975586771...|
|  1.0|       1.0|[0.12717190317357...|
|  1.0|       1.0|[0.12722013760877...|
|  1.0|       1.0|[0.12726227316348...|
|  1.0|       1.0|[0.12730330333302...|
|  1.0|       1.0|[0.12730731463115...|
|  1.0|       1.0|[0.12730804261353...|
+-----+----------+--------------------+
```

Miary, na których oparta będzie ocena modelu, przyjmują wartości jak niżej.

```
lr time 0:05:03.563295
True Positives: 32836
True Negatives: 5943
False Positives: 12898
False Negatives: 3381
Total 55058
DenseMatrix([[ 5943., 12898.],
             [ 3381., 32836.]])
recall 0.9066
precision 0.7180
specificity 0.3154
accuracy 0.7043
error 0.2957
weightedPrecision 0.6904
f1 0.6715
area under ROC curve 0.7046
```

Wyniki wydają się nieco lepsze niż poprzednio, zwłaszcza widoczna jest duża poprawa powierzchni pod krzywą ROC oraz swoistości (`specificity`).

3.3 Porównanie

Aby łatwiej było zdecydować, który model okazał się lepszy poniżej sporządzono porównanie wyników w jednej tabeli. Na zielono zaznaczono lepszy wynik w danym wierszu.

	Decision Tree	Logistic Regresion
Time [min]	04:32	05:03
Recall	0,9607	0,9066
Precision	0,6717	0,7180
Specificity	0,0973	0,3154
Accuracy	0,6653	0,7043
Weighted precision	0,6345	0,6904
F1	0,5769	0,6715
Area under ROC	0,5305	0,7046

Regresja logistyczna uzyskała lepszą dokładność (i co za tym idzie, mniejszy błąd), oraz wyższą wartość f1. Ogromną poprawę widać w powierzchni pod krzywą ROC - tutaj poprawa wyniosła aż 17 punktów procentowych. Poprawa tych wskaźników nastąpiła jednak kosztem wydłużenia czasu o 12% (ponad pół minuty) oraz spadkiem wartości `recall`.

4 Podsumowanie

Podsumowując, w przedstawionym problemie lepiej sprawdził się model regresji logistycznej. Może wartości rzędu 60-70% nie są najlepsze, ale w porównaniu z wartościami 0,5317 oraz 0,5769 które uzyskał model drzewa decyzyjnego odpowiednio jako wartość pod krzywa ROC oraz f1, regresja logistyczna będzie lepszym wyborem. Analizując powyższe wyniki trzeba mieć na uwadze kontekst problemu. Ocena filmu tylko po jego gatunku i słowie go opisującym jest praktycznie niemożliwa (wiadomo przecież, że nawet po przeczytaniu całej recenzji trudno wyrobić sobie zdanie o tym czy film może się spodobać czy nie). Ponadto fakt, iż tagi były wpisywane przez użytkowników dowolnie sprawił, że są bardzo zróżnicowane, trudno zatem doszukać się w nich jakiś konkretnych zależności, które ułatwiłyby proces klasyfikacji.