APPLICATION OF DATA MINING IN MOVIE RECOMMENDATIONS

TAN YIN YEN

WQD180108

ANALYSIS GOAL

A movie streaming company (Netflix) seeks to maximize customer's retention by recommending highly rated movies with DVD or streaming options available to their users. A movie can be predicted as a good or bad movie based on the movie information, box office data and sentiment score of user reviews on the movie.

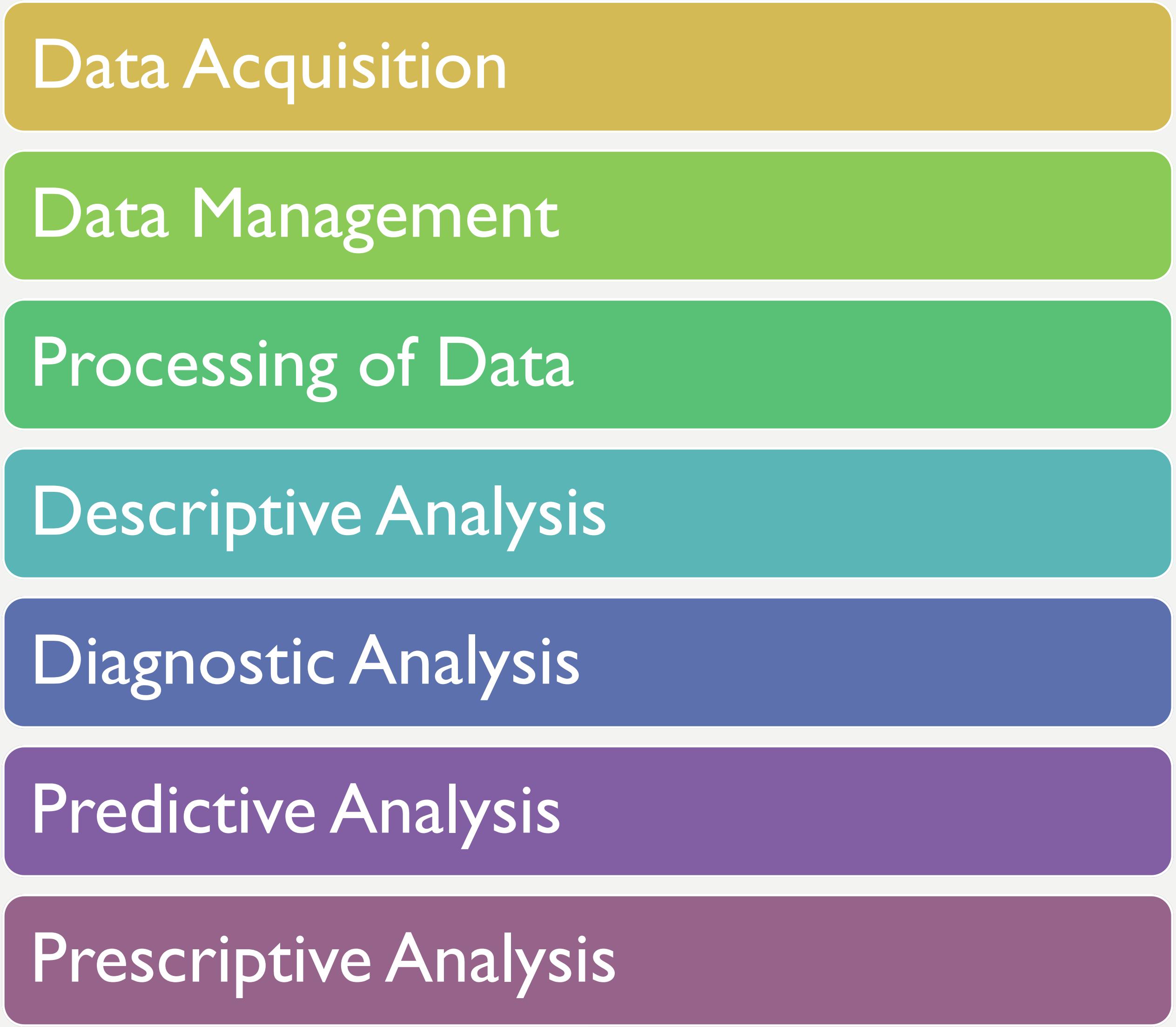
By predicting whether a movie is good or bad based on its reviews and box office achievement, the movie streaming company can filter out latest movies with DVD or streaming options available that are highly rated and recommend them to its users. Customers who are satisfied with the movie recommendations are more likely to subscribe to the movie streaming service in the next month.

OBJECTIVES

To come out with a model to predict whether a movie is good or bad based on its box office performance and user reviews.

To recommend users the top and latest movies with DVD or streaming options available based on the prediction.

PROCESS FLOW



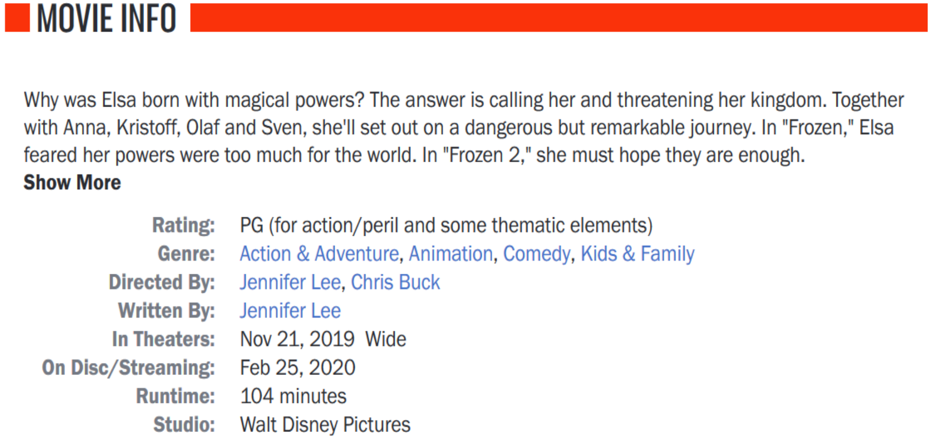
- Data Acquisition
- Data Management
- Processing of Data
- Descriptive Analysis
- Diagnostic Analysis
- Predictive Analysis
- Prescriptive Analysis

DATA ACQUISITION

Scope: Scrape data of latest 5000 movies with DVD or streaming options available.

Data acquired:

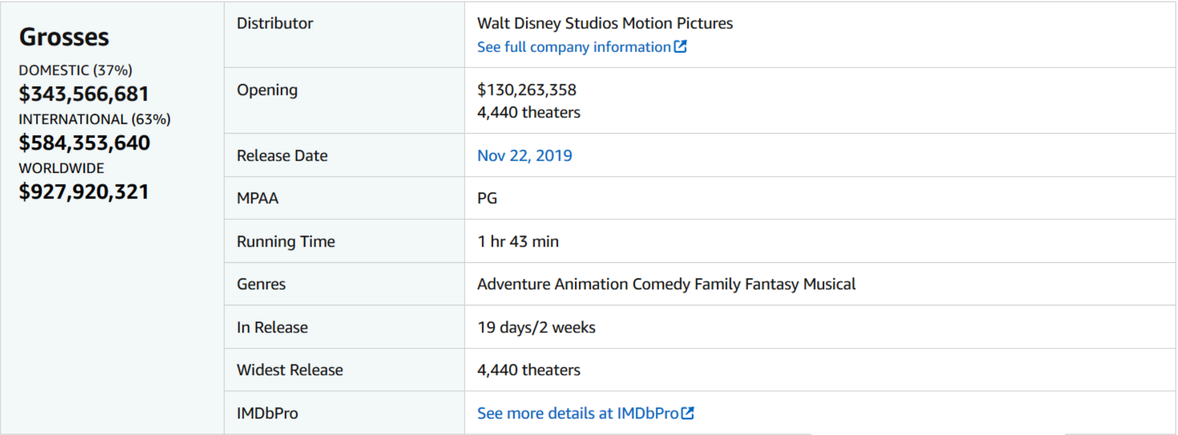
- Structured data
  - Movie information scraped from *rottentomatoes.com*
  - Movie box office data scraped from *boxofficemojo.com*



MOVIE INFO

Why was Elsa born with magical powers? The answer is calling her and threatening her kingdom. Together with Anna, Kristoff, Olaf and Sven, she'll set out on a dangerous but remarkable journey. In "Frozen," Elsa feared her powers were too much for the world. In "Frozen 2," she must hope they are enough.
Show More

Rating:	PG (for action/peril and some thematic elements)
Genre:	Action & Adventure, Animation, Comedy, Kids & Family
Directed By:	Jennifer Lee, Chris Buck
Written By:	Jennifer Lee
In Theaters:	Nov 21, 2019 Wide
On Disc/Streaming:	Feb 25, 2020
Runtime:	104 minutes
Studio:	Walt Disney Pictures

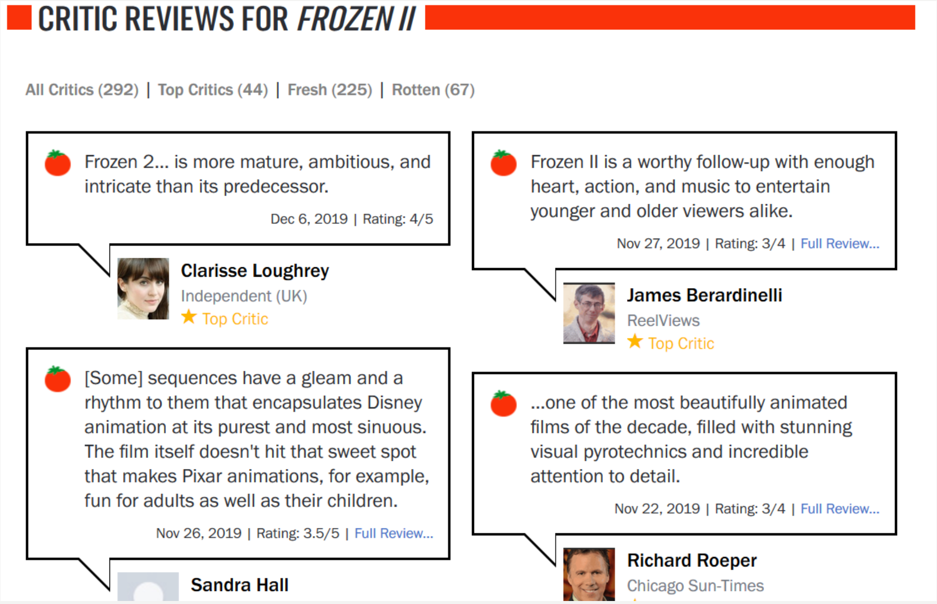


Grosses

DOMESTIC (37%)
$343,566,681
INTERNATIONAL (63%)
$584,353,640
WORLDWIDE
$927,920,321

Distributor	Walt Disney Studios Motion Pictures See full company information
Opening	$130,263,358 4,440 theaters
Release Date	Nov 22, 2019
MPAA	PG
Running Time	1 hr 43 min
Genres	Adventure Animation Comedy Family Fantasy Musical
In Release	19 days/2 weeks
Widest Release	4,440 theaters
IMDbPro	See more details at IMDbPro

- Unstructured data
    - Movie reviews scraped from *rottentomatoes.com*



CRITIC REVIEWS FOR *FROZEN II*

All Critics (292) | Top Critics (44) | Fresh (225) | Rotten (67)

Frozen 2... is more mature, ambitious, and intricate than its predecessor.

Dec 6, 2019 | Rating: 4/5

Clarisse Loughrey
Independent (UK)
★ Top Critic

Frozen II is a worthy follow-up with enough heart, action, and music to entertain younger and older viewers alike.

Nov 27, 2019 | Rating: 3/4 | Full Review...

James Berardinelli
ReelViews
★ Top Critic

[Some] sequences have a gleam and a rhythm to them that encapsulates Disney animation at its purest and most sinuous. The film itself doesn't hit that sweet spot that makes Pixar animations, for example, fun for adults as well as their children.

Nov 26, 2019 | Rating: 3.5/5 | Full Review...

Sandra Hall

...one of the most beautifully animated films of the decade, filled with stunning visual pyrotechnics and incredible attention to detail.

Nov 22, 2019 | Rating: 3/4 | Full Review...

Richard Roeper
Chicago Sun-Times

Rotten Tomatoes

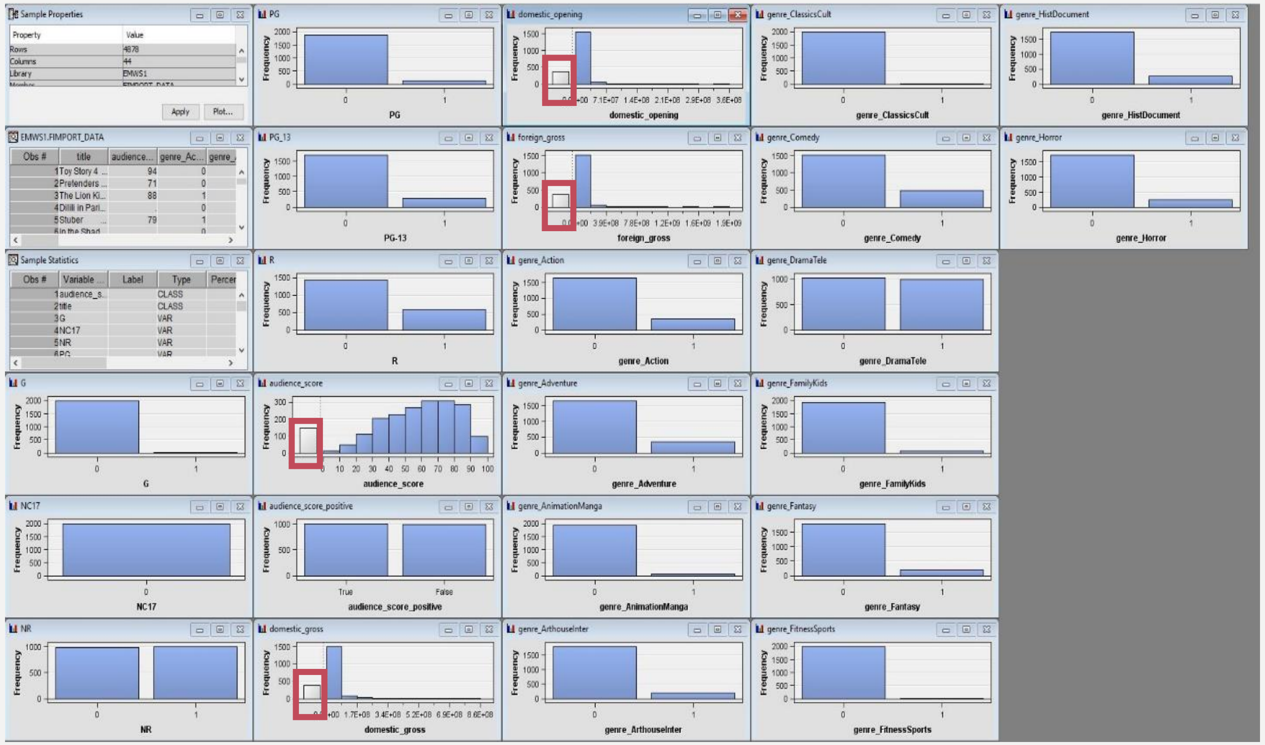
DATA MANAGEMENT

- Store the data scraped into hive data warehouse.

PROCESSING OF DATA

SEMMA

- **SEM**MA
  - Sample: Import the data into csv from hive data warehouse
  - Explore: Identify missing values or patterns in the data



Sample Properties

Property	Value
Rows	4878
Columns	44
Library	EMWS1
Member	FIMPORT_DATA

Apply Plot...

EMWS1.FIMPORT_DATA

Obs #	title	audience...	genre_Ac...	genre_
1	Toy Story 4 ...	94	0	
2	Pretenders ...	71	0	
3	The Lion Ki...	88	1	
4	Dilili in Pari...	.	0	
5	Stuber ...	79	1	
6	In the Shad		0	

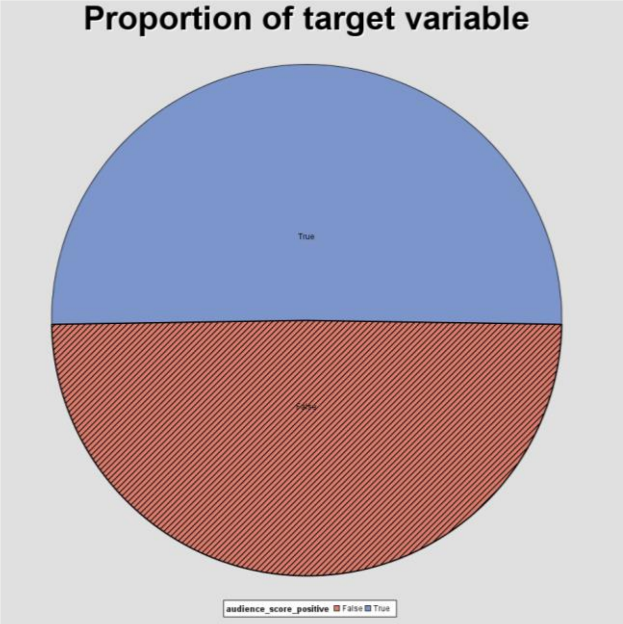
Sample Statistics

Obs #	Variable ...	Label	Type	Percer
1	audience_s...		CLASS	
2	title		CLASS	
3	G		VAR	
4	NC17		VAR	
5	NR		VAR	
6	PG		VAR	

Histograms of Variables

- Missing data is identified.

SEMMA



Pie Chart of Target Variable

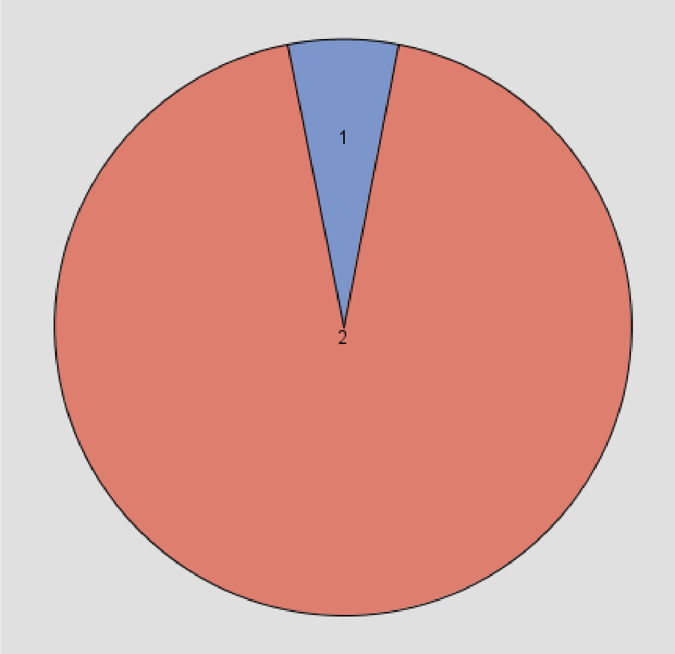
- Target variable has a proportion of approximately 50% True (good movie) and 50% False (bad movie)
- Target class is balance

- Modify: Create, transform or remove variables
  - Covert the reviews data to sentiment score
  - Impute missing values
  - Combine multiple genres into a single 'genre' cluster
  - Drop features that are not useful

S**E**MMA

DESCRIPTIVE ANALYSIS

- Hierarchical clustering is used for descriptive analysis as the number of clusters no need to be specified.



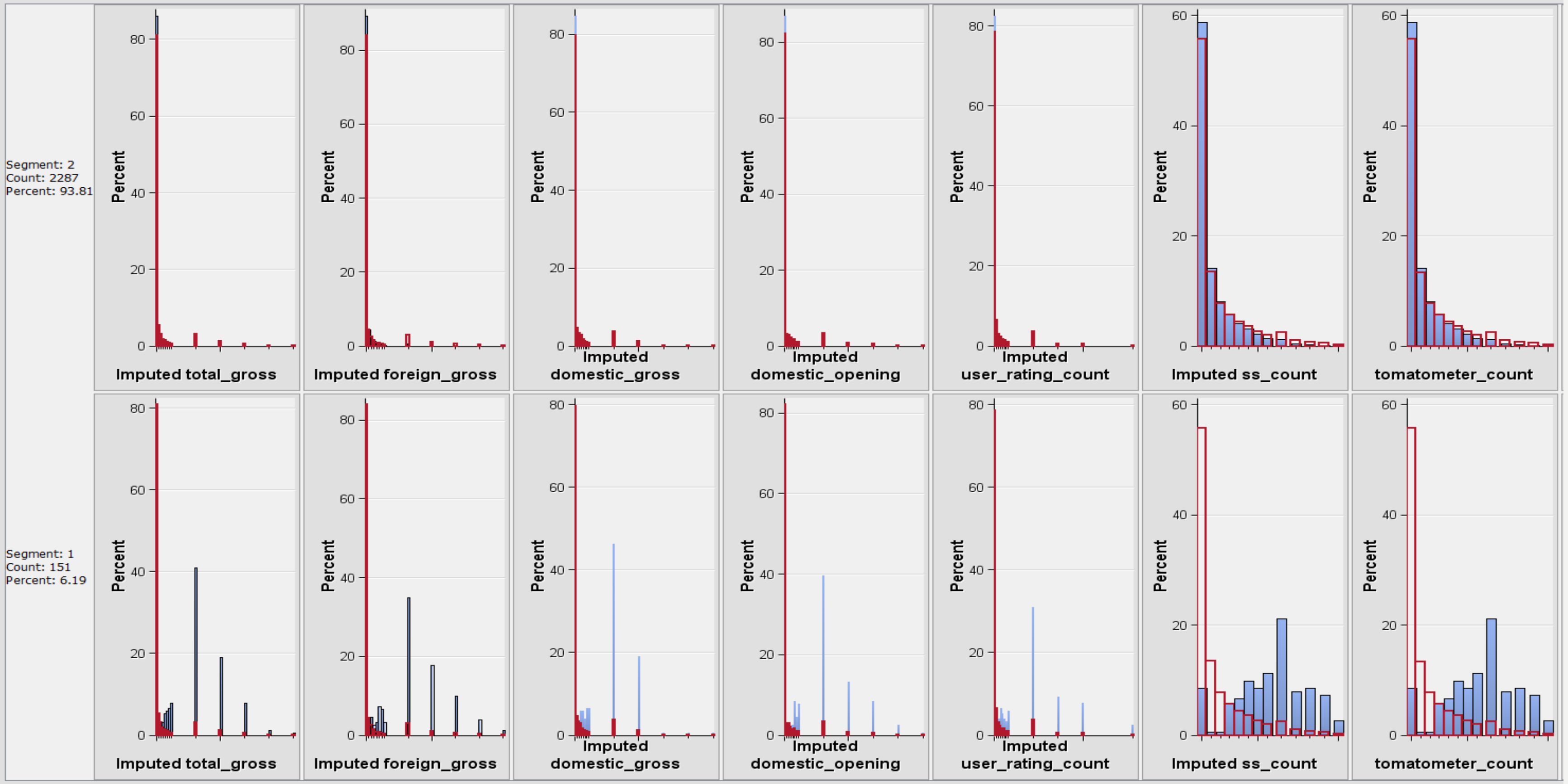
Segment Size Plot

- Two clusters are formed.
- Cluster 1 contains 151 cases.
- Cluster 2 contains 2287 cases.

🟦 Good movie

🟥 Bad movie

sEMMA



Profile plot of significant features

🟦 Clustered data distribution

🟥 Original distribution

- For segment 1, the overall box office distribution and number of reviews are higher than average.
- Contains very few cases (151 cases)

∴ **Good and recommended movies**

- For segment 2, the overall box office distribution and number of reviews are lower than average.
- Contains many cases (2287 cases)

∴ **More towards bad movies and requires further investigation**

DIAGNOSTIC ANALYSIS

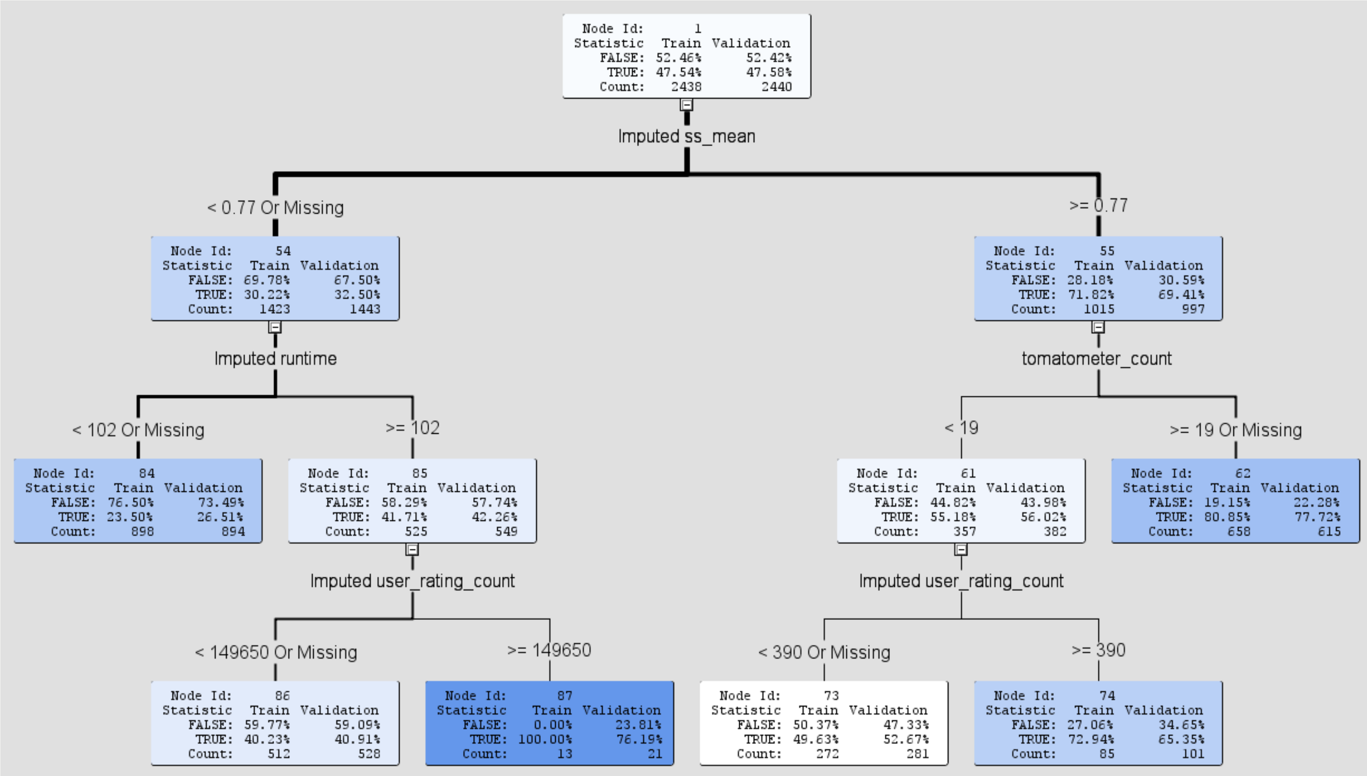
- Sentiment Analysis using Natural Language Processing

Positive Sentiment	Negative Sentiment
Excellent	Worst
Perfect	Awful
Great	Boring
Wonderful	Waste
Amazing	Bad
Superb	Poor
Enjoyable	Terrible
Best	Dull
Today	Poorly
Fun	Disappointment
Enjoyed	Disappointing
Brilliant	Unfortunately
Must see	Worse
Loved	Stupid
Fantastic	Horrible
Liked	Mess
Incredible	Nothing
Funniest	Lame
Wonderfully	Lacks
Better than	Save

Top 20 positive features and negative features extracted

PREDICTIVE ANALYSIS

- SEMMA
  - Model: Apply a model to the data
- A decision tree is created for predictive analysis as the model is simple and easier to interpret.



```
Node Id:        1
Statistic  Train  Validation
   FALSE:  52.46%     52.42%
    TRUE:  47.54%     47.58%
   Count:   2438       2440
```

```
Node Id:       54
Statistic  Train  Validation
   FALSE:  69.78%     67.50%
    TRUE:  30.22%     32.50%
   Count:   1423       1443
```

```
Node Id:       55
Statistic  Train  Validation
   FALSE:  28.18%     30.59%
    TRUE:  71.82%     69.41%
   Count:   1015        997
```

```
Node Id:       84
Statistic  Train  Validation
   FALSE:  76.50%     73.49%
    TRUE:  23.50%     26.51%
   Count:    898        894
```

```
Node Id:       85
Statistic  Train  Validation
   FALSE:  58.29%     57.74%
    TRUE:  41.71%     42.26%
   Count:    525        549
```

```
Node Id:       61
Statistic  Train  Validation
   FALSE:  44.82%     43.98%
    TRUE:  55.18%     56.02%
   Count:    357        382
```

```
Node Id:       62
Statistic  Train  Validation
   FALSE:  19.15%     22.28%
    TRUE:  80.85%     77.72%
   Count:    658        615
```

```
Node Id:       86
Statistic  Train  Validation
   FALSE:  59.77%     59.09%
    TRUE:  40.23%     40.91%
   Count:    512        528
```

```
Node Id:       87
Statistic    Train  Validation
   FALSE:    0.00%     23.81%
    TRUE:  100.00%     76.19%
   Count:      13         21
```

```
Node Id:       73
Statistic  Train  Validation
   FALSE:  50.37%     47.33%
    TRUE:  49.63%     52.67%
   Count:    272        281
```

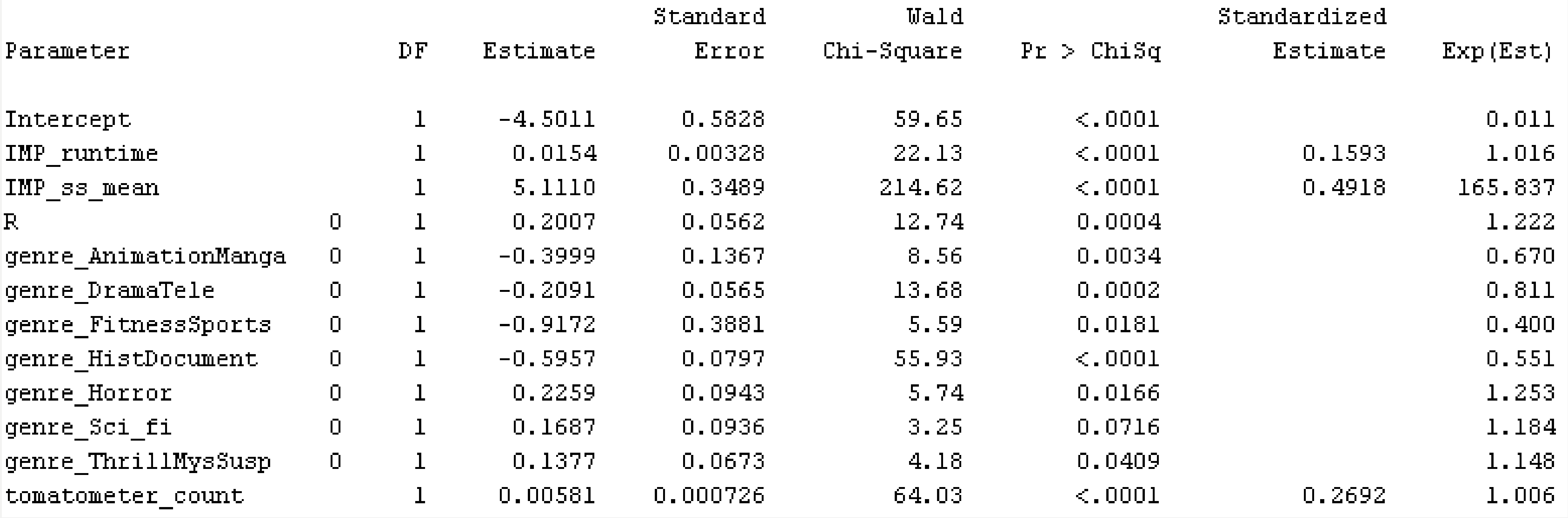
```
Node Id:       74
Statistic  Train  Validation
   FALSE:  27.06%     34.65%
    TRUE:  72.94%     65.35%
   Count:     85        101
```

The nodes are split based on

- Information gain: Higher information gain.
- Interpretability: Easier to understand

- First level of decision tree:
  - Movie with average sentiment score **higher** than or equal to 0.77 is likely to be a **good** movie
  - Movie with average sentiment score **lower** than or equal to 0.77 is likely to be a **bad** movie

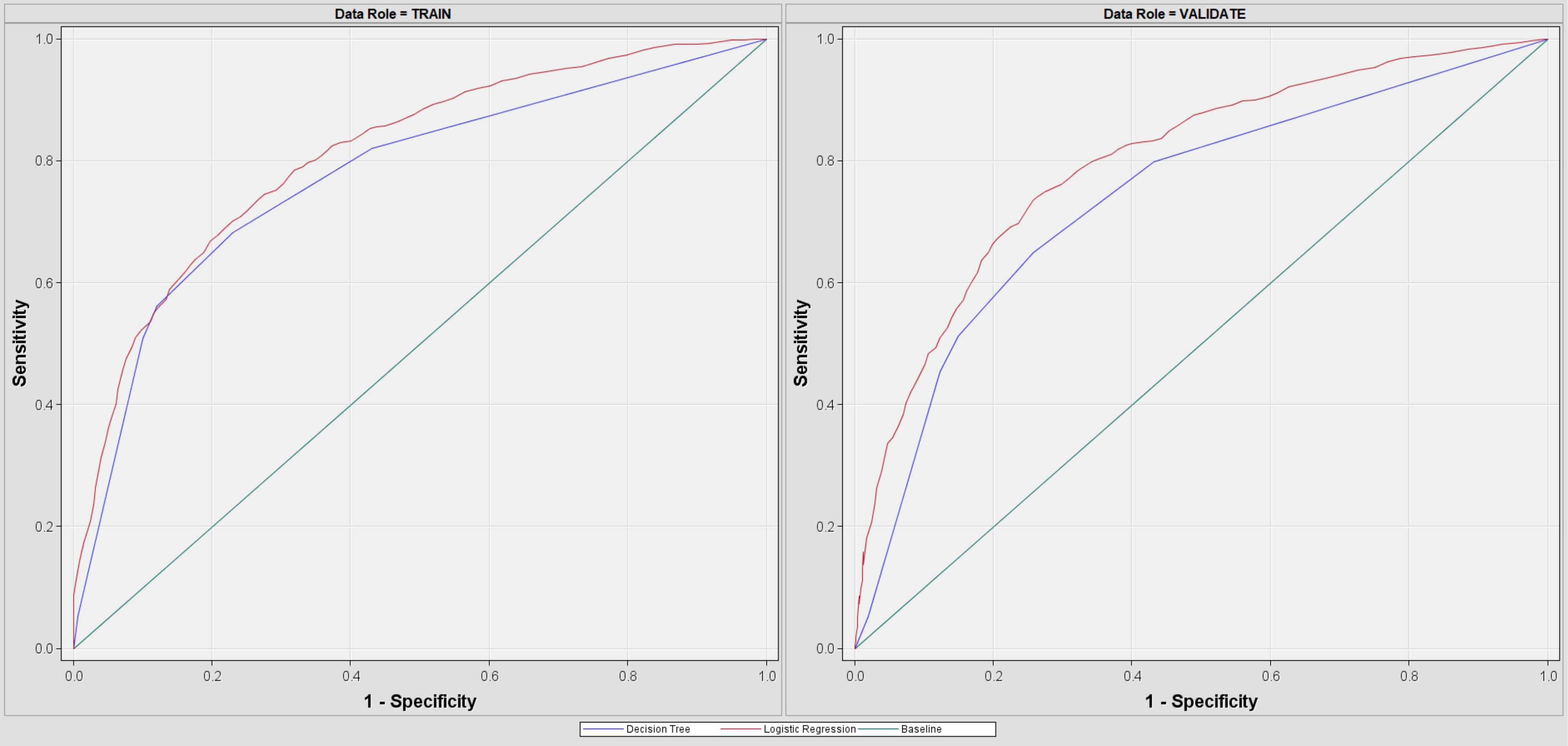
- A logistic regression is developed as the model is simple to implement and efficient to train.



```
                                             Standard          Wald                       Standardized
Parameter                   DF    Estimate       Error    Chi-Square    Pr > ChiSq            Estimate    Exp(Est)

Intercept                    1     -4.5011      0.5828         59.65        <.0001                           0.011
IMP_runtime                  1      0.0154     0.00328         22.13        <.0001              0.1593       1.016
IMP_ss_mean                  1      5.1110      0.3489        214.62        <.0001              0.4918     165.837
R                      0     1      0.2007      0.0562         12.74        0.0004                           1.222
genre_AnimationManga   0     1     -0.3999      0.1367          8.56        0.0034                           0.670
genre_DramaTele        0     1     -0.2091      0.0565         13.68        0.0002                           0.811
genre_FitnessSports    0     1     -0.9172      0.3881          5.59        0.0181                           0.400
genre_HistDocument     0     1     -0.5957      0.0797         55.93        <.0001                           0.551
genre_Horror           0     1      0.2259      0.0943          5.74        0.0166                           1.253
genre_Sci_fi           0     1      0.1687      0.0936          3.25        0.0716                           1.184
genre_ThrillMysSusp    0     1      0.1377      0.0673          4.18        0.0409                           1.148
tomatometer_count            1     0.00581    0.000726         64.03        <.0001              0.2692       1.006
```

- A positive regression coefficient indicates that the mean of the dependent variable (good or bad movie) increases with the value of independent variable.
- Higher regression coefficient indicates higher influence.
- Mean sentiment score has highest positive influence
  - Higher mean sentiment score is likely to be good movie.
- For genre, horror movie has highest positive influence, followed by science fiction movie.
  - Horror movie is often combined with sci-fi movie (e.g. Earth is threatened by Aliens)
- Sports movie has highest negative influence.
  - The 'plot' of the game is the same every time, hence is boring.
  - Waste of time watching people win their game.



- Compare decision tree and logistic regression.
- Accuracy (decision tree): 68.11%
- Accuracy (logistic regression): 73.36%
- Logistic regression has better performance compared to decision tree, hence is selected for prescriptive analysis.

PRESCRIPTIVE ANALYSIS

- SEMMA
  - Assess: Determine whether the result is useful
- Recommend movies with **high sentiment score** to users
  - Movie with a positive review are likely to be a good movie
- Recommend movies with **genre of horror and science fiction** to users
  - Movie with both of these genres are likely to be a good movie

CONCLUSION

- Box office and user reviews can be used to predict whether a movie is good or bad.
- Words like excellent, perfect and great normally appear in a good movie review.
- Words like worst, awful and boring normally appear in a bad movie review.
- Movies with positive reviews and genres of horror and science fiction are likely to be good.