



## 데이터마이닝 기법을 활용한 영화 흥행 실적 예측 기법

---

저자 (Authors)	강지훈, 박찬희, 도형록, 김성범
출처 (Source)	<a href="#">한국경영과학회 학술대회논문집</a> , 2014.5, 142-154 (13 pages)
발행처 (Publisher)	<a href="#">한국경영과학회</a> Korean Operations Research And Management Society
URL	<a href="http://www.dbpia.co.kr/Article/NODE07171410">http://www.dbpia.co.kr/Article/NODE07171410</a>
APA Style	강지훈, 박찬희, 도형록, 김성범 (2014). 데이터마이닝 기법을 활용한 영화 흥행 실적 예측 기법. 한국경영과학회 학술대회논문집, 142-154.
이용정보 (Accessed)	경희대학교 163.***.18.29 2018/07/30 14:22 (KST)

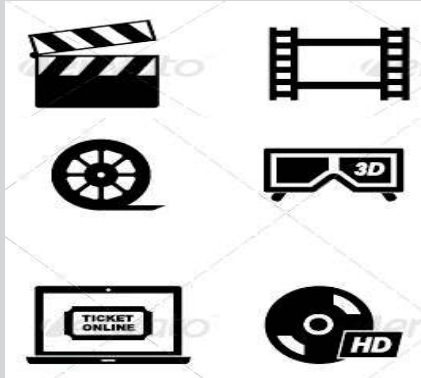
---

### 저작권 안내

DBpia에서 제공되는 모든 저작물의 저작권은 원저작자에게 있으며, 누리미디어는 각 저작물의 내용을 보증하거나 책임을 지지 않습니다. 그리고 DBpia에서 제공되는 저작물은 DBpia와 구독 계약을 체결한 기관소속 이용자 혹은 해당 저작물의 개별 구매자가 비영리적으로만 이용할 수 있습니다. 그러므로 이에 위반하여 DBpia에서 제공되는 저작물을 복제, 전송 등의 방법으로 무단 이용하는 경우 관련 법령에 따라 민, 형사상의 책임을 질 수 있습니다.

### Copyright Information

Copyright of all literary works provided by DBpia belongs to the copyright holder(s) and Nurimedia does not guarantee contents of the literary work or assume responsibility for the same. In addition, the literary works provided by DBpia may only be used by the users affiliated to the institutions which executed a subscription agreement with DBpia or the individual purchasers of the literary work(s) for non-commercial purposes. Therefore, any person who illegally uses the literary works provided by DBpia by means of reproduction or transmission shall assume civil and criminal responsibility according to applicable laws and regulations.



## 데이터마이닝 기법을 활용한 영화 흥행 실적 예측 기법

고려대학교 산업경영공학부  
강지훈, 박찬희, 도형록, 김성범

## 목차

- 서론 및 배경
  - 영화 수요 예측
  - 관련 문헌 조사
  - 예측 알고리즘 선정
  - 수집 데이터 소개
- 예측 기법
  - 예측 프로세스
  - 방법론
  - 예측 성능 검증
- 예측 결과 및 요약

## 목차

### ■ 서론 및 배경

- 영화 수요 예측
- 관련 문헌 조사
- 예측 알고리즘 선정
- 수집 데이터 소개

### ■ 예측 기법

- 예측 프로세스
- 방법론
- 예측 성능 검증

### ■ 예측 결과 및 요약

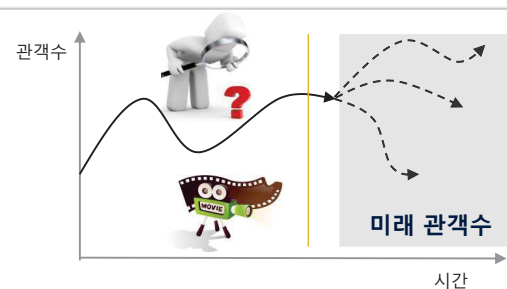
3

## 서론 및 배경 > 영화 수요 예측

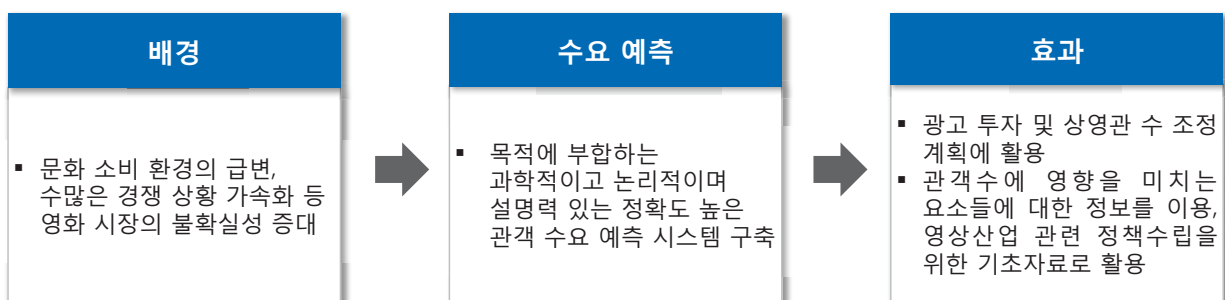
### 영화 관객수 (수요) 예측의 정의 및 중요성

#### [ “영화산업” 에서의 관객수 예측 ]

- 특정 영화가 개봉 전/후, 정해진 기간 동안의 전국 누적 관객수 예측



#### [ 관객수 예측의 중요성 ]

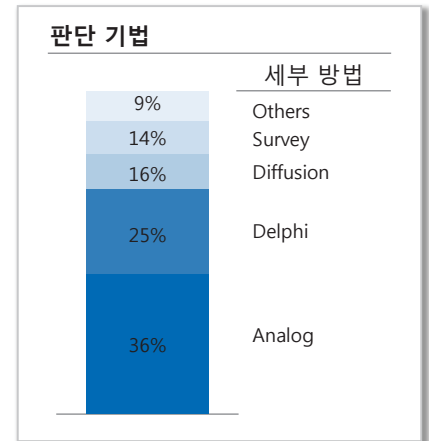
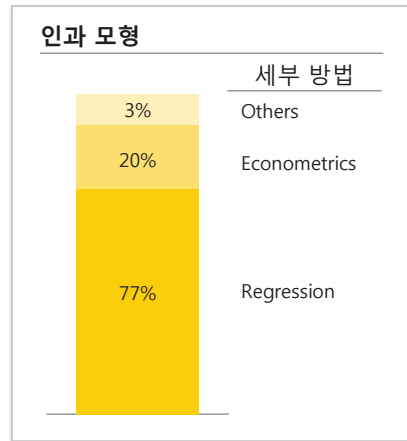
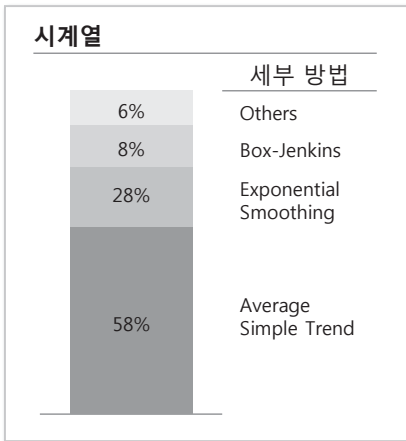
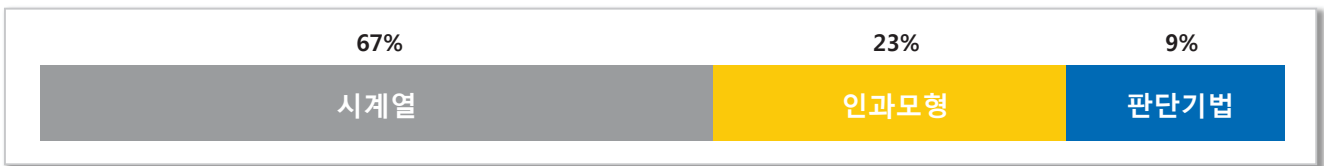


4

## 서론 및 배경 > 관련 문헌 조사

### 기존의 대표적인 수요 예측 방법론들

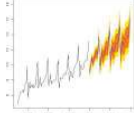
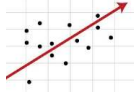

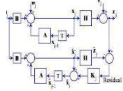
#### 예측 기법



Reference : 2004 IBF (Institute of Business Forecast) 주최 컨퍼런스 & 교육 참가자 대상 설문조사 결과 (316개 社)

## 서론 및 배경 > 관련 문헌 조사

### 기존의 수요 예측 방법론들의 한계점

	모델 설명	한계점
Time Series Prediction	 <ul style="list-style-type: none"> <li>시계열 기반의 접근법은 과거의 수요 패턴을 기반으로 모델을 구축하여 향후 시간에 따른 수요를 예측함</li> <li>주로 단기 예측에 활용</li> </ul>	<ul style="list-style-type: none"> <li>모델 설계 단계에서 설정해야 할 파라미터들이 많음</li> <li>모델 수립에 걸리는 시간이 길며, 일반적으로 강건(robust)하지 않음</li> </ul>
Regression	 <ul style="list-style-type: none"> <li>회귀 모델은 종속변수와 한 개 또는 여러 개의 독립변수 간의 인과 관계를 가정하고 모델을 구축함</li> <li>주로 장기 예측에 활용</li> </ul>	<ul style="list-style-type: none"> <li>종속변수와 독립변수 간의 비선형 관계가 있을 경우 모델 성능이 떨어짐</li> </ul>
Knowledge-Based Prediction	 <ul style="list-style-type: none"> <li>지식 기반의 모델 구축은 시스템에 대한 사전 지식과 경험을 바탕으로 예측을 위한 규칙을 만드는 방법임</li> </ul>	<ul style="list-style-type: none"> <li>예측을 위한 규칙을 만들기 위해서는 모델 설계자가 반복적으로 시행 착오를 거치면서 규칙을 찾아내야 함</li> <li>Ex) Case-Based Reasoning</li> </ul>
Noise Filtering	 <ul style="list-style-type: none"> <li>잡음이 포함되어 있는 데이터에서 잡음을 제거해주는 재귀적인 필터를 이용해 선형 모델을 구축하는 방법임</li> </ul>	<ul style="list-style-type: none"> <li>데이터의 주기성 가정이 성립해야 하며 비선형 모델을 만들 때에는 성능이 떨어짐</li> <li>Ex) Kalman filter</li> </ul>

Reference : 1) Cascaded Artificial Neural Networks For Short-Term demand Forecasting, A. S. AlFuhaid et. al.

2) Financial time series forecasting using independent component analysis and support vector regression, Chi-Jie Lu, Tian-Shyug Lee, Chih-Chou Chiu

## 서론 및 배경 > 관련 문헌 조사 > 관객 수요 예측 관련 연구

- Mestyán, Márton, Taha Yasseri, and János Kertész. "Early prediction of movie box office success based on Wikipedia activity big data." *PloS one* 8.8 (2013): e71226.
- Asur, Sitaram, and Bernardo A. Huberman. "Predicting the future with social media." *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*. Vol. 1. IEEE, 2010.
- Krauss, Jonas, et al. "Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis." *ECIS*. 2008.
- Sharda, Ramesh, and Dursun Delen. "Predicting box-office success of motion pictures with neural networks." *Expert Systems with Applications* 30.2 (2006): 243-254.
- K. Dave, S. Lawrence, & D. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings WWW 2003*, 2003.

7

## 서론 및 배경 > 관련 문헌 조사

### 데이터마이닝 및 다변량 분석 방법론의 장점

#### ✓ 모델 구축에 필요한 과정 최소화



- 전통적 통계 모형과 달리 변수들의 분포, 파라미터에 대한 가정에서 자유로움
- 일반적으로 강건(robust)한 모델 구축 가능

#### ✓ 복잡한 구조의 데이터에 대한 활용



- 복잡한 패턴을 보이는 현상을 모델링 하는데 적합한 방법론
- 일반적으로 비선형 패턴 및 다양한 형태의 데이터 타입을 가지고 있는 경우에 높은 예측 정확성을 보임

#### ✓ 모델 구축에 소요되는 사용자의 주관적 노력 최소화



- 데이터 기반의 객관적 모델을 반복 수립 가능
- 매번 예측 모델을 구축하는 데 걸리는 시간/노력 감소

#### ✓ 다양한 정보를 반영할 수 있는 모델 구축



- 예측하고자 하는 변수의 자체적인 시계열 패턴 뿐만 아니라 다양한 인자를 사용 가능
- 좀 더 많은 정보를 반영하여 유연하고 정확한 모델 구축

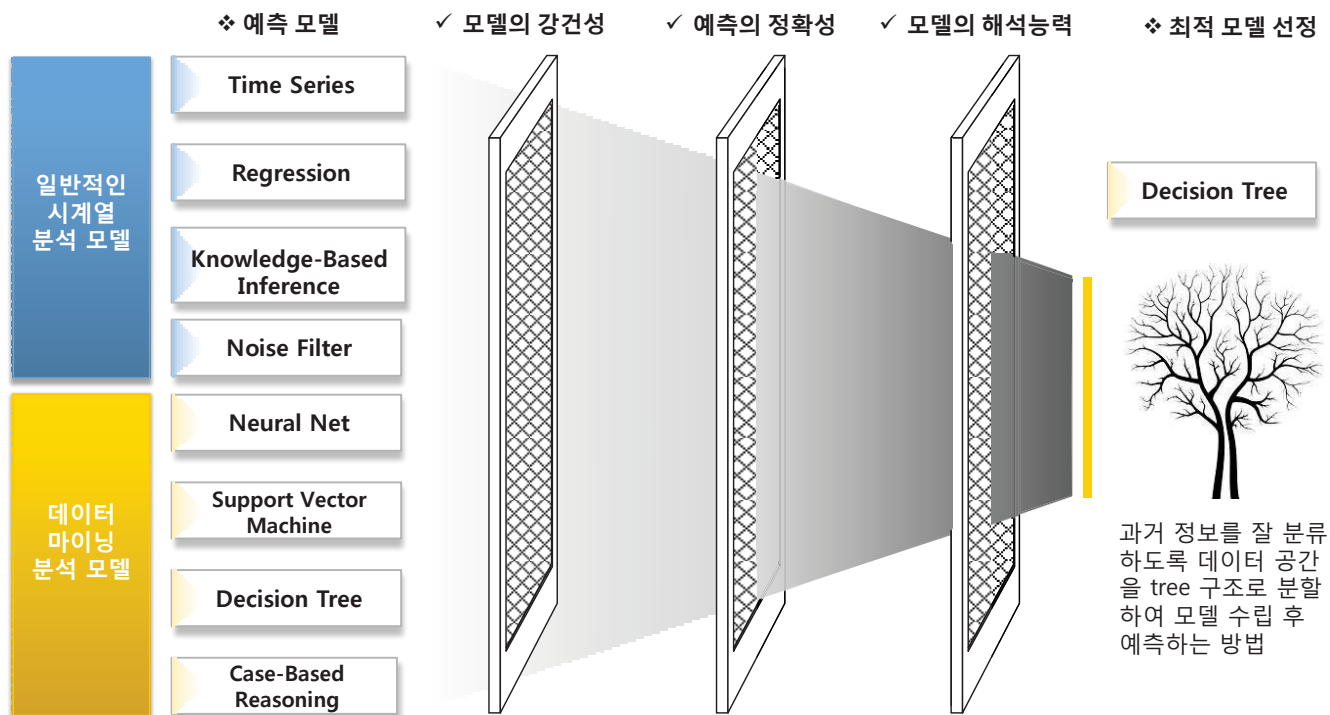
Reference : Univariate modeling and forecasting of monthly energy demand time series using abductive and neural networks, R.E. Abdel-Aal

8

## 서론 및 배경 > 예측 알고리즘 선정

### 알고리즘 선정 과정

- 다면적인 비교 분석을 통해 예측 문제 해결을 위한 최적의 분석 모델로써 Decision Tree 알고리즘 활용

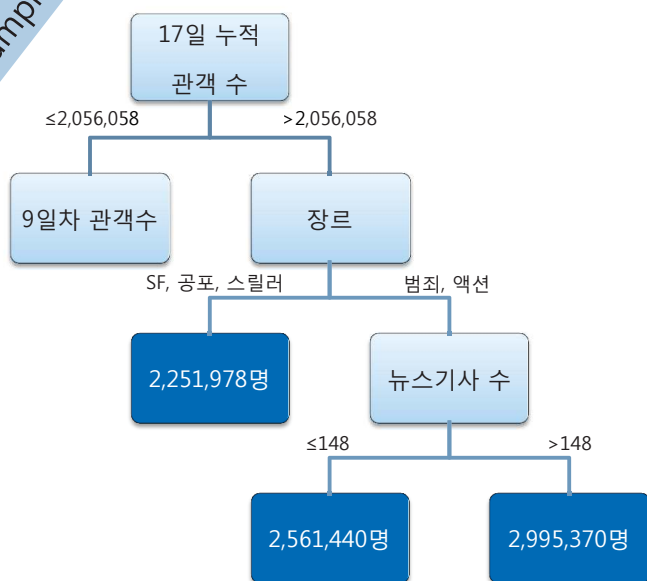


9

## 서론 및 배경 > 예측 알고리즘 선정

### Decision Tree의 상세 알고리즘 및 특징

Example



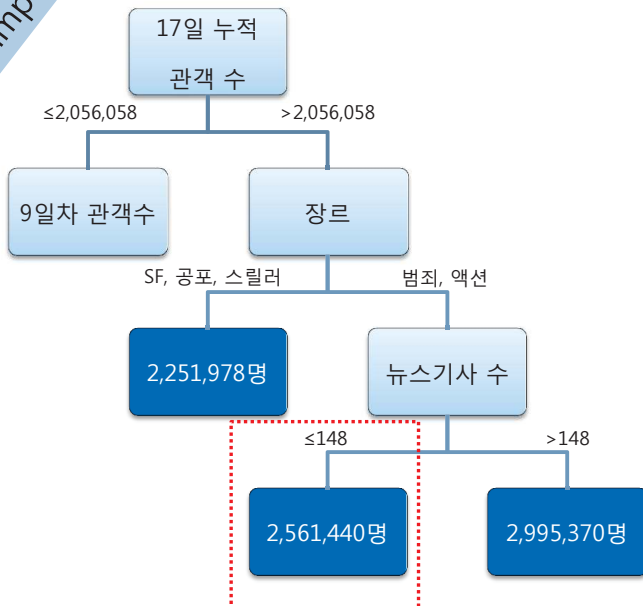
- 데이터 **분포**에 대한 가정 불필요
- 예측하고자 하는 변수(종속변수)와 관계가 있는 설명변수들의 정보 이용
- 예측에 큰 영향을 미치는 소수의 **설명변수**를 찾을 수 있음
- 설명변수가 **연속형**과 **범주형**이 섞여있는 데이터에도 효과적으로 활용가능
- 모델 **결과의 해석**이 매우 용이 (If-then 룰의 형태)

10

## 서론 및 배경 > 예측 알고리즘 선정

Decision Tree의 상세 알고리즘 및 특징

Example



영화 : 내가 살인범이다

24일차 누적 관객 수

예측: 2,561,440명  
실제: 2,348,097명



규칙

17일차 누적관객수 >2,056,058  
&  
장르 = 범죄, 액션  
&  
뉴스기사 수 <148

SCI-FI  
FANTASY  
STOR  
ANIMAL  
MYSTERY  
FAMILY MOVIES  
FRIENDS  
PRINCESS  
ACTION  
MAGIC  
SCARY  
SPORT



- 영화의 흥행요인을 미리 가늠하고 흥행 영화 제작에 있어서 가이드라인으로 활용 가능

11

## 서론 및 배경 > 수집 데이터 소개

관객수 예측을 위해 사용한 변수들

	변수 종류	선정 이유
영화 속성 정보를 반영한 변수	<ul style="list-style-type: none"> <li>감독과 주연배우의 흥행력=평균 누적 관객수</li> <li>제작사, 등급, 장르</li> </ul>	<ul style="list-style-type: none"> <li>감독과 주연배우의 영향력은 영화흥행에 긍정적인 영향을 미침<sup>1)</sup></li> <li>제작사의 규모, 장르에 따라 흥행성적에 영향을 미침<sup>2)</sup></li> </ul>
관객 반응을 고려한 변수	<ul style="list-style-type: none"> <li>*평점</li> <li>*댓글수</li> </ul>	<ul style="list-style-type: none"> <li>소셜 네트워크의 발달로 인터넷 상에서 관객들의 입소문이 영화흥행에 큰 영향을 미침<sup>3)</sup></li> </ul>
마케팅 요소를 반영한 변수	<ul style="list-style-type: none"> <li>배급사의 영향력 = 총 누적 관객수/총 영화 수</li> <li>*개봉 前 일주일 기사 수</li> </ul>	<ul style="list-style-type: none"> <li>배급사 규모는 스크린 수와 직접적인 관계가 있음<sup>4)</sup></li> <li>개봉 前 기사 수는 잠재 관객들의 영화 관심도를 증가시킴</li> </ul>
파생 정보 변수	<ul style="list-style-type: none"> <li>회귀식 기울기</li> <li>누적 관객수</li> </ul>	<ul style="list-style-type: none"> <li>관객 반응의 확산 속도를 반영하기 위한 변수</li> <li>보다 정확한 누적 관객 수 예측을 위해 관객 규모를 정규화 시키는 변수</li> </ul>

1)김휴중, "한국 영화스타의 스타파워 분석," *문화경제연구* 1.1 (1998): 165-200

2)송현주, "영화의 흥행성과 제작비 규모와의 관계-2011 년 한국영화의 흥행결정요인 분석," *사회과학연구* 51.1 (2012): 45-79.

3)임현정, "온라인 구전의 특성과 영화흥행에 관한 연구-구전량과 영화 개봉일을 중심으로," *한국경영정보학회 학술대회* (2013): 425-430.

4)유현석, "영화흥행 변수에 관한 연구," *문화정책논총*, 제13집(2001): 231~254

\*영화 별 평점, 댓글수, 개봉 前 해당 영화 기사의 개수는 웹데이터 수집 기법인 crawling 기법 응용

12

## 목차

### ■ 서론 및 배경

- 영화 수요 예측
- 관련 문헌 조사
- 예측 알고리즘 선정
- 수집 데이터 소개

### ■ 예측 기법

- 예측 프로세스
- 방법론
- 예측 성능 검증

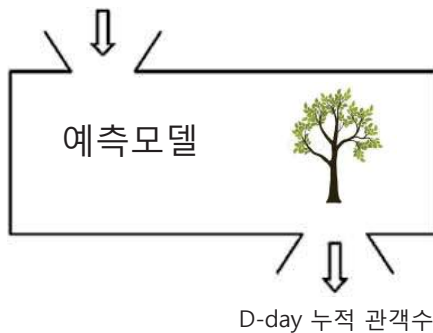
### ■ 예측 결과 및 요약

13

## 예측 기법 > 예측 프로세스

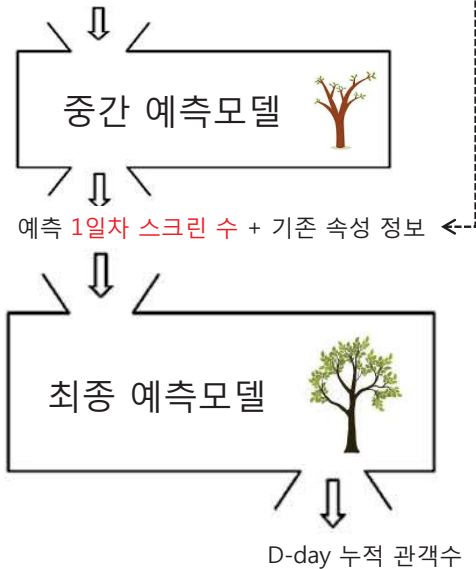
### Case 1: 상영 중 영화

영화 속성 변수  
관객 반응 변수  
마케팅 변수  
1일차 스크린 수



### Case 2: 개봉 前 영화

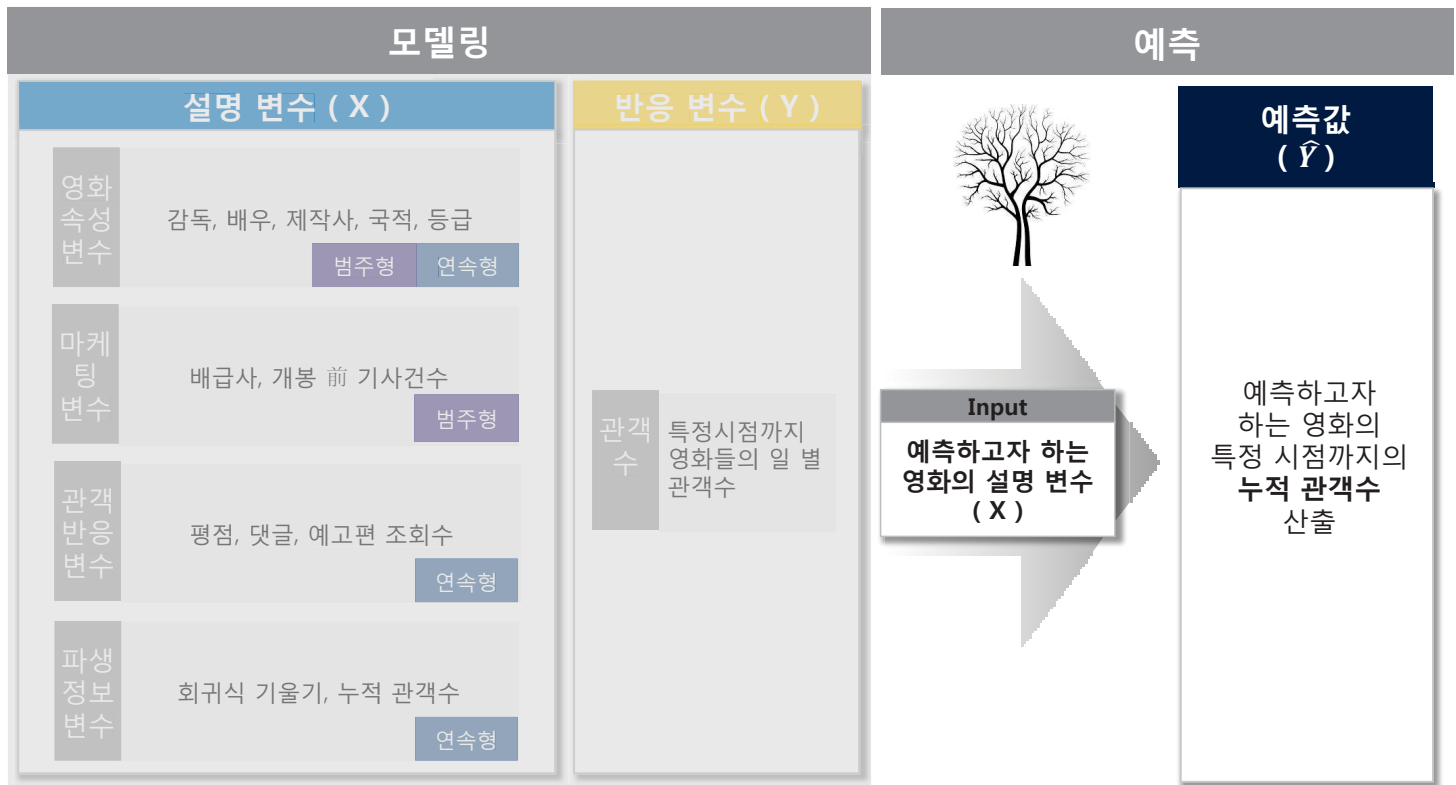
영화 속성 변수  
관객 반응 변수  
마케팅 변수



14



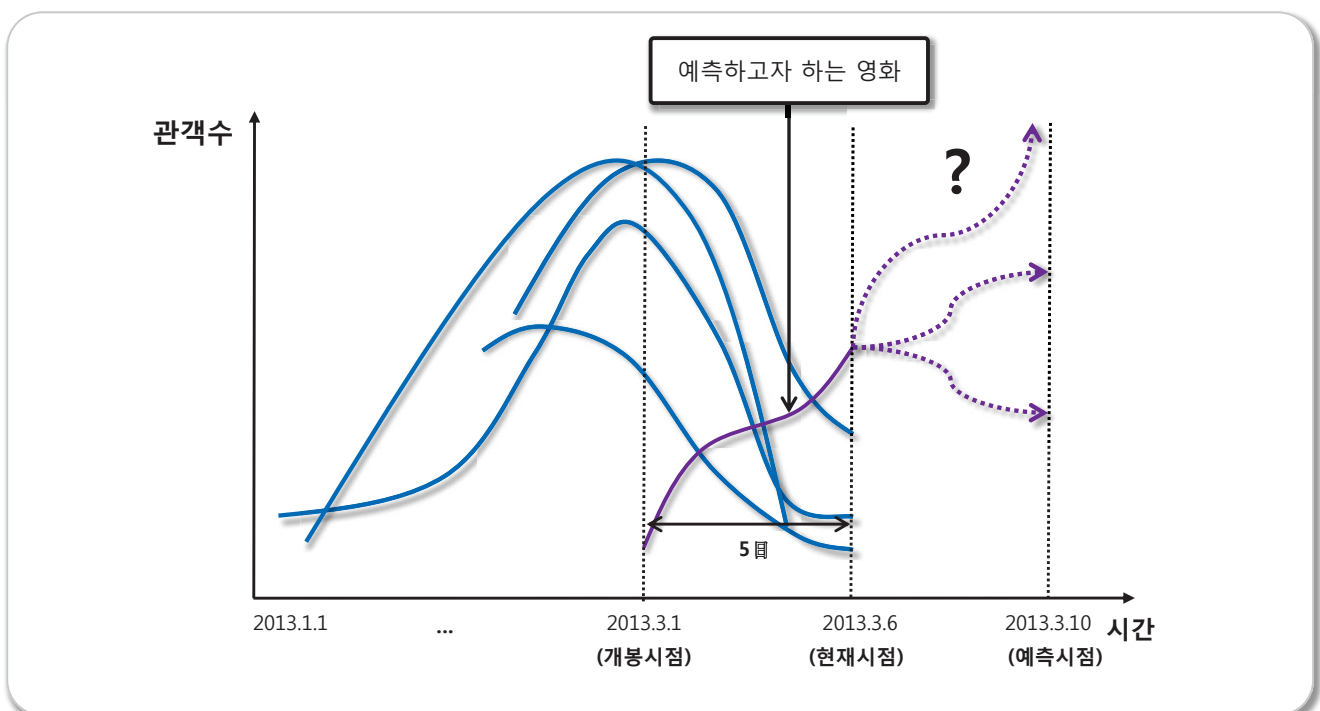
## 예측 기법 > 예측 모델 구축



15

## 예측 기법 > 방법론

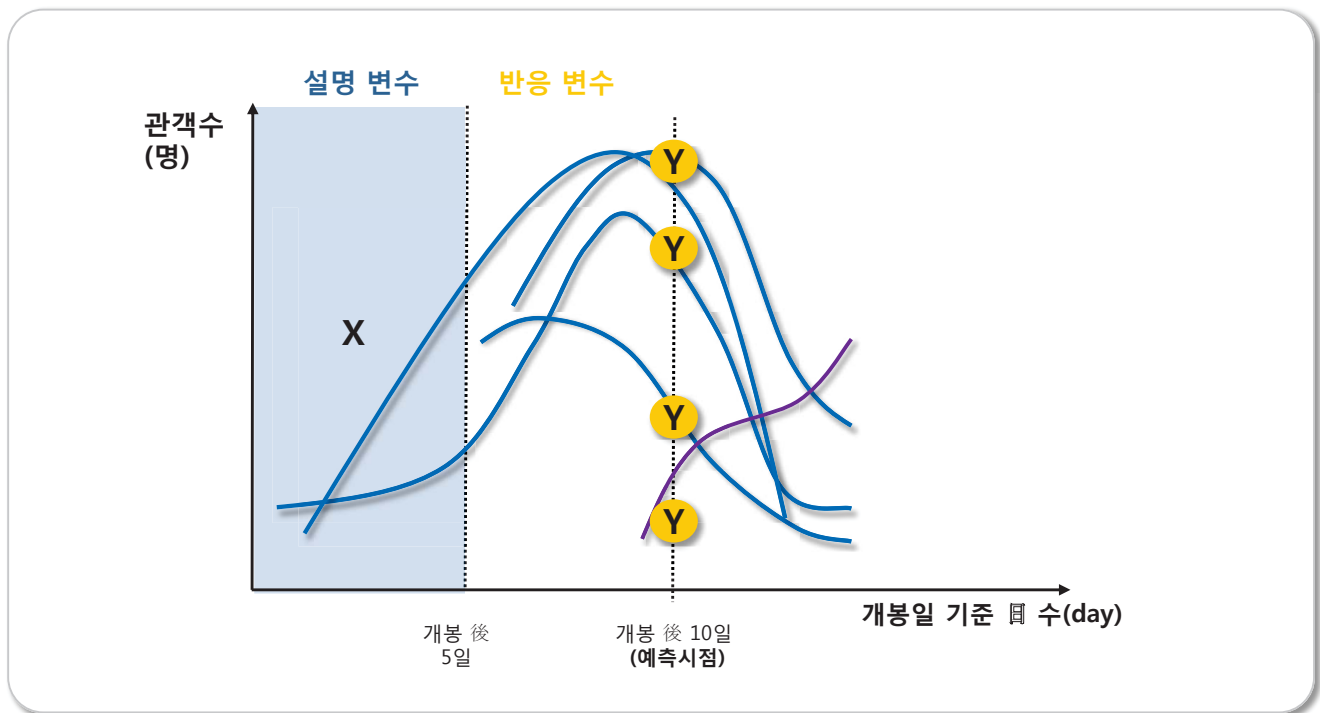
### 관객수 예측 모델



16

## 예측 기법 > 방법론

### ■ 관객수 예측 모델



- 예측하고자 하는 영화의 현재 상영기간을 기준으로 기존 유사 영화들의 패턴분석을 통해 최적의 예측값 추정 가능

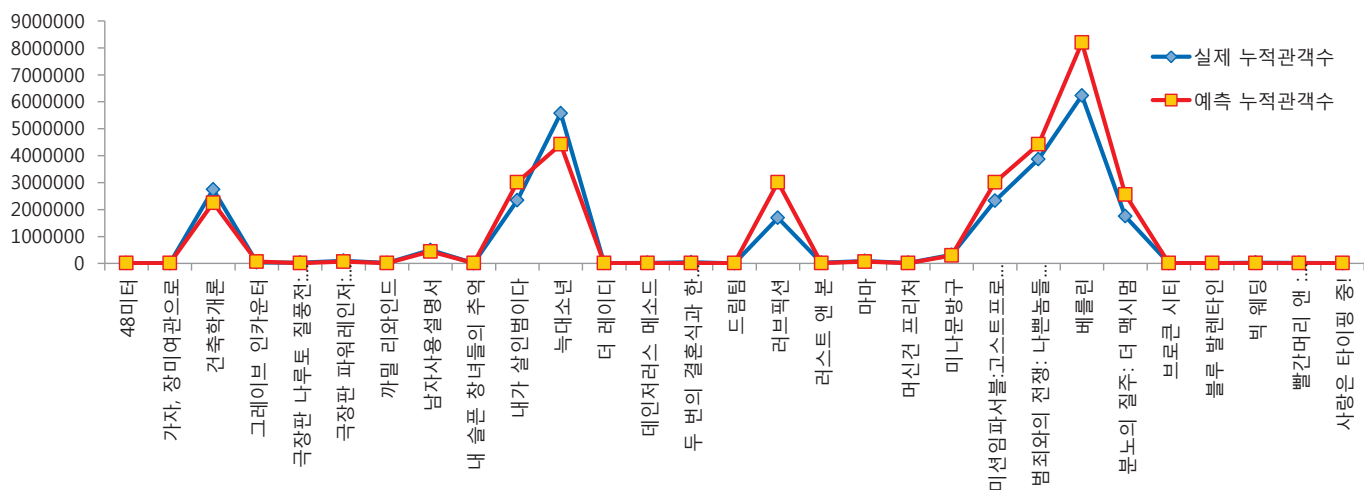
17

## 예측 기법 > 예측 성능 검증

### 개봉 후 24일 누적 관객수를 예측하는 Decision Tree 모델의 성능 검증

- 15일간의 관객 실적 패턴과 영화의 사전 정보를 토대로 예측한 결과
- 주어진 데이터를 9:1(학습:검증)로 나누어 구축한 모델의 예측 성능을 검증함
- Cross-validation 적용
- 예측 모델로써 다양한 영화들의 누적관객수를 비교적 정확히 예측함

\*실제값과 예측값의  
correlation : 0.97



18

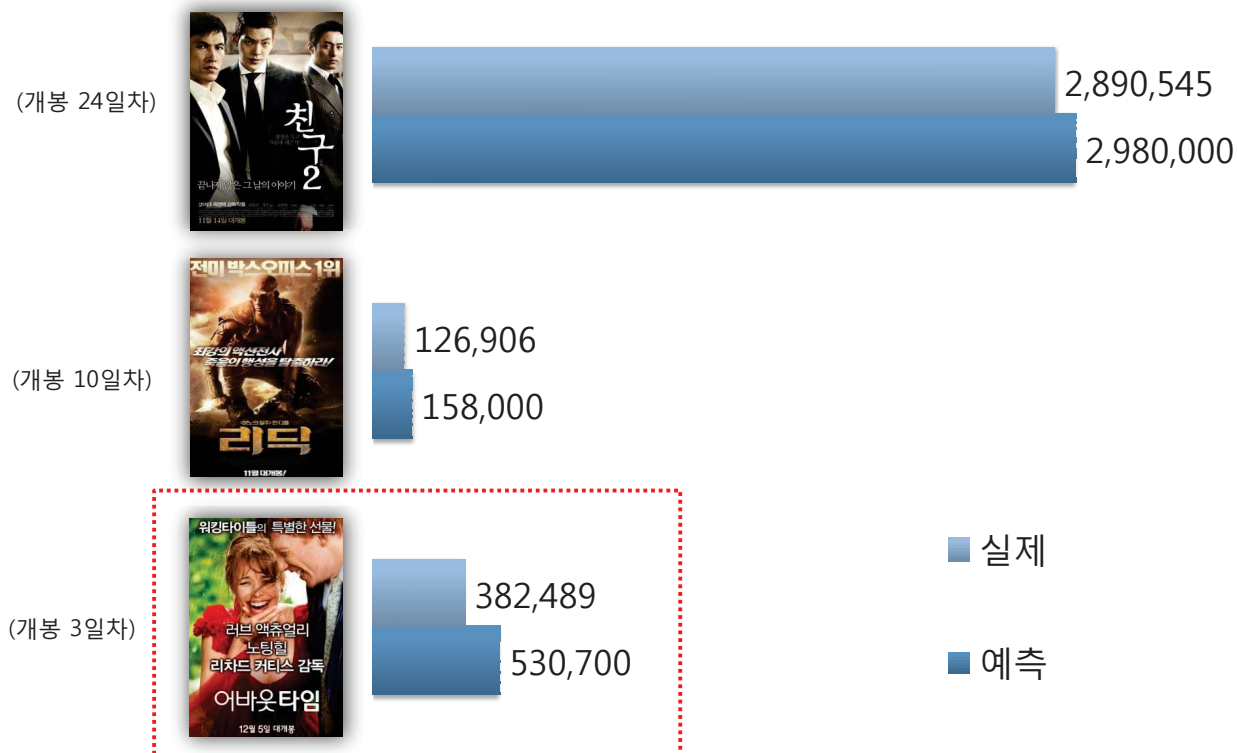
## 목차

- 서론 및 배경
  - 영화 수요 예측
  - 관련 문헌 조사
  - 예측 알고리즘 선정
  - 수집 데이터 소개
- 예측 기법
  - 예측 프로세스
  - 방법론
  - 예측 성능 검증
- **예측 결과 및 요약**

19

## 예측 결과

2013년 12월 7일까지 총 누적관객 수

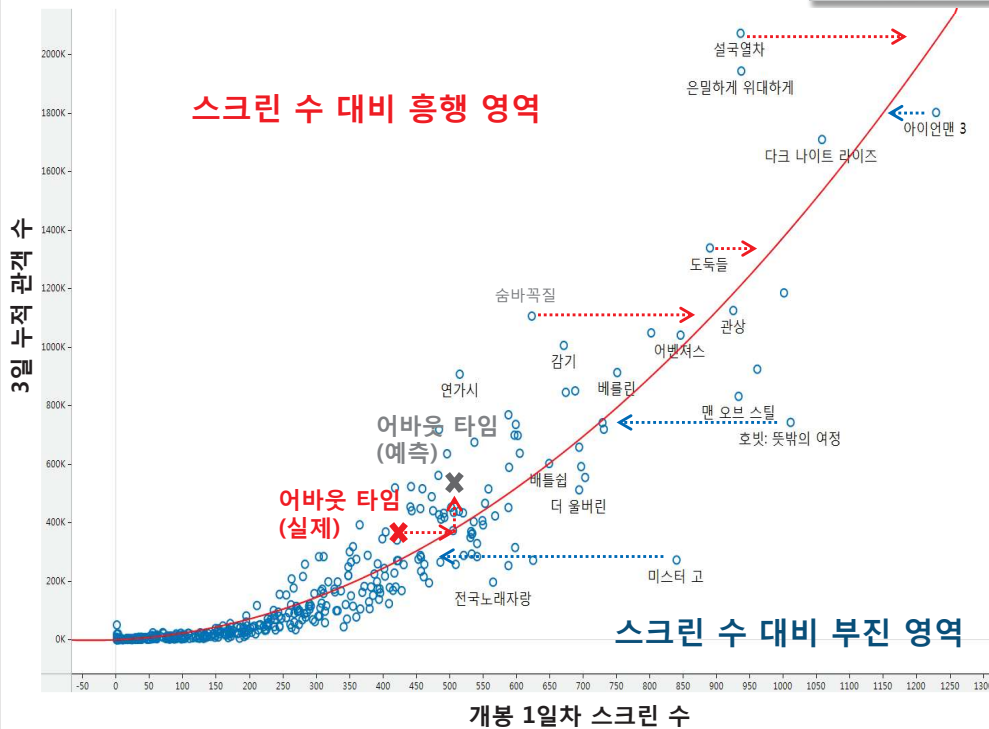


20

## 예측 결과의 재고 및 활용 방안

이차회귀방정식을 응용한 스크린 수 최적화 방안

- 3일 누적 관객수 =  $1.2758 * (\text{1일차 스크린수})^2 + 100.575 * (\text{1일차 스크린수})$
- $R^2 : 0.85$



- **어바웃 타임**의 경우, 약 380,000 명의 관객을 개봉 후 3일 내에 확보(개봉 당일 실제 스크린 수 442개, 예측 563개)
- 일반적으로 개봉 3일 이내에 380,000명 정도의 관객 확보를 위해서는 약 520개 스크린 이상 확보가 필요함 (즉, 어바웃 타임의 경우 역량 대비 흥행 확률이 높은 영화임을 파악 가능함)
- 저력 있는 영화의 스크린 수 확보와 부진영화의 과도한 스크린 수의 최적화를 위한 근거 마련 (주 단위의 스크린 수 조정을 위한 가이드 라인으로 활용 가능)

21

## 요약

예측 프로세스 요약 및 방법론의 강점

### 1. 데이터 구성



- 영화 흥행에 영향을 미치는 4가지 주요 변수군으로 구성
- 1일차 스크린 수만을 이용
- 필요한 파생변수 생성

### 2. 예측 모델 구축



- 데이터의 특성과 분석 목적에 따라 설명변수와 반응변수 선택

### 3. 예측



- 예측하고자 하는 영화의 가용 정보와 예측 시점 확인
- 그에 상응하는 예측 모델을 선택하고 예측 시행

#### Adaptability

모델의 구축이 용이하고 체계적으로 모델을 설계, 적용 가능

#### Updatability

새로운 변수, 관측치 추가에 따라 모형이 진화하고 성능이 향상될 수 있음

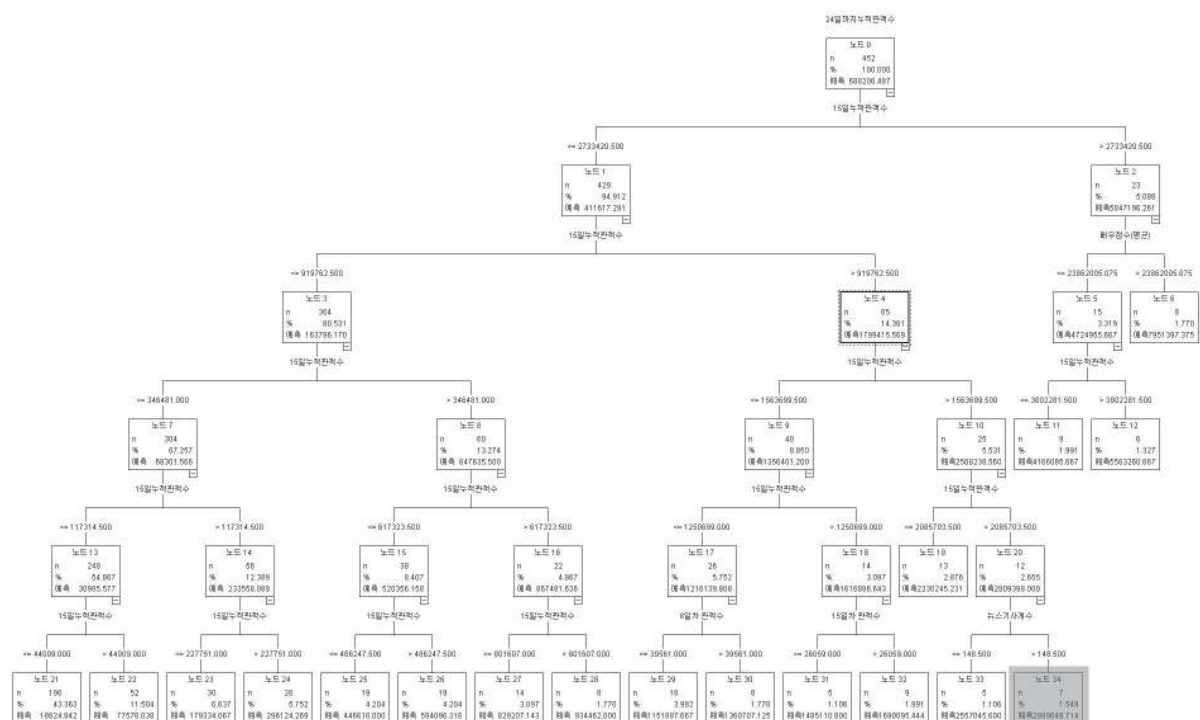
#### Practicality

실제 의사 결정 과정에 적용할 수 있을 만큼 정확하고 해석력이 뛰어나.

22



## 별첨 2. 관객수 예측 모델

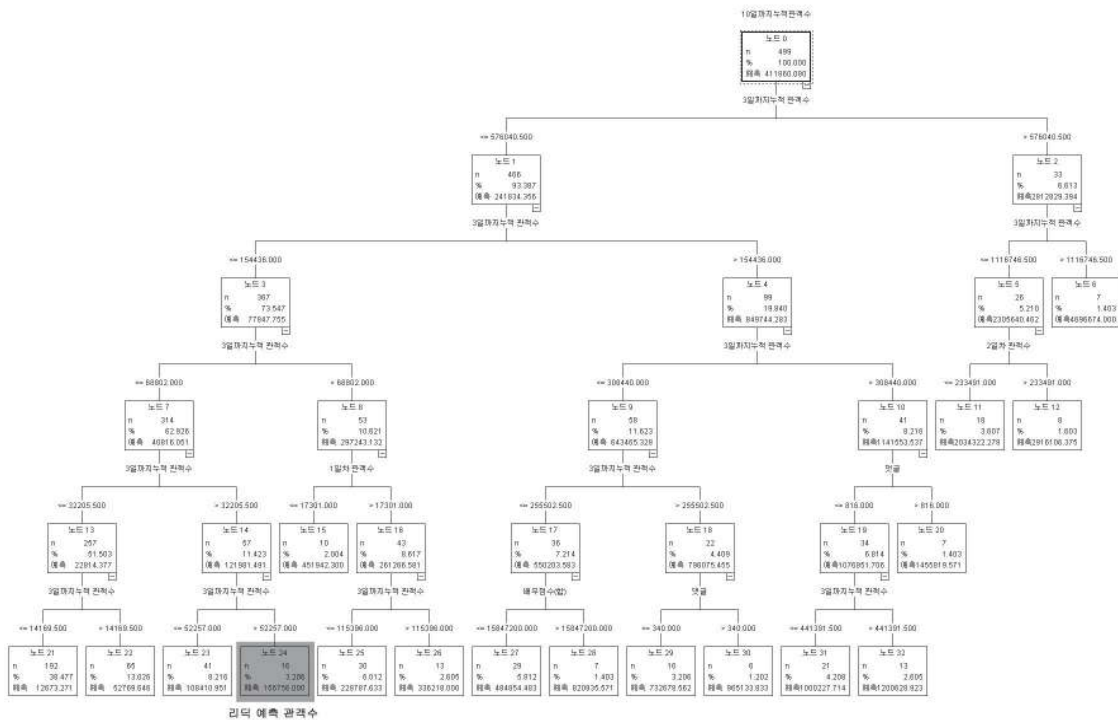


\*확대하여 확인 가능

친구2의 예측 관객수

## 별첨 2. 관객수 예측 모델

### Decision Tree 모델 결과물 (리딕)



\*확대하여 확인 가능

25

## 별첨 2. 관객수 예측 모델

### Decision Tree 모델 결과물 (어바웃 타임)

개봉 1일차 스크린수의 예측

개봉 後 3일간 누적 관객수 예측



\*확대하여 확인 가능

26