이번 2주차에는

밑바닥부터 시작하는 데이터 사이언스를 공부하려 했지만..

빅콘테스트 제출일과 겹쳐

일단 빅콘테스트에 집중하기로 했다.

빅콘테스트에 사용한 모델에 대한 설명을 첨부하였다.

후에 비슷한 예측을 하는 사람들에게 부디 반면교사하는 계기가 되길 빈다.
(반면교사: 다른 사람의 잘못을 보고 자신의 교훈으로 삼는것)

대회 이름과 공모분야

공모분야	문제	제공데이터 및 내용	참가대상
퓨처스리그	개봉(예정) 영화에 대한 관객 수 예측 -킹스맨: 골든 서클 (9월 27일 개봉) - 남한산성 (10월 3일 개봉) - 넛잡2 (10월 3일 개봉)	▪ 약 2년 정도의 개봉 영화 관련 데이터 ▪ 매출액, 점유율, 관객수, 증감율, 스크린수 상영횟수 등	고등학생, 대학(원)생

김범준과 잘생긴 미니언들 팀

퓨처스 리그 (관객수예측)

김범준: 김범준
잘생긴 미니언들: 곽병우
김현우
조규형

목차

- 사용한 변수
  - 사용한 데이터 소개
  - 특정 독립변수 측정방법
- 사용한 모델 소개
  - 모델 세부 파라메터
- 전체적 모델 학습 과정
- 성능

사용한 변수 1(킹스맨2 예측 모델)

- 종속 변수

개봉 2주 후까지의 누적관객수

- 독립변수

네이버 영화 개봉 전 기대지수 '보고싶어요' 수
네이버 영화 개봉 전 리뷰 수
개봉 전 30일간 특정 영화 관련 인터넷 기사 수
배우점수
감독점수
배급사점수
개봉 2일 후 누적관객수

사용한 변수 2(남한산성, 넛잡2 예측 모델)

- 종속 변수

개봉 1주 후까지의 누적관객수

- 독립변수

네이버 영화 개봉 4일전 기대지수 '보고싶어요' 수
네이버 영화 개봉 4일전 리뷰 수
개봉 4일전 30일간 관련 인터넷 기사 수
배우점수
감독점수
배급사점수
개봉 4일전 누적 관객수

사용한 데이터

- 14~17년 개봉영화별 관련정보
  (모델에 직접 학습 시키기 위해 사용)
  (출처: 영화진흥위원회 통합전산망)

- 역대 박스오피스 탑 500 영화별 관련정보
  (배우, 감독, 배급사 점수 측정을 위해 사용)
  (출처: 영화진흥위원회 통합전산망)

- 그 외 모두 웹 크롤링으로 데이터 제작

특정 독립변수 측정 방법

- 배우 점수, 감독 점수, 배급사 점수 측정법

역대 탑 500 영화에 참여한 배우, 감독, 배급사 별로 특정 점수를 측정

$$\text{배우점수} = \sum_{1}^{\text{참여한 모든영화}} (\text{누적관객수} * \text{개봉일자가중치})$$

$$\text{개봉일자가중치} = \frac{100}{\text{현재날짜} - \text{개봉일자}} \quad (\text{최신일수록 가중치가 상승})$$

감독점수, 배급사점수도 똑같이 측정

(이후 모델 학습용 데이터인 14~17년도 개봉 영화마다 참여한 배우, 감독, 배급사에 따라 점수 부여)

사용한 모델

부스트 트리(Gradient Boosting Regressor)

결정트리 기반 모델

간단한 결정노드를 단계적으로 조합해 모델 구축
(각 단계마다 이전 단계의 결정노드 오차를 최소화하는
방식으로 새로운 결정노드를 생성)

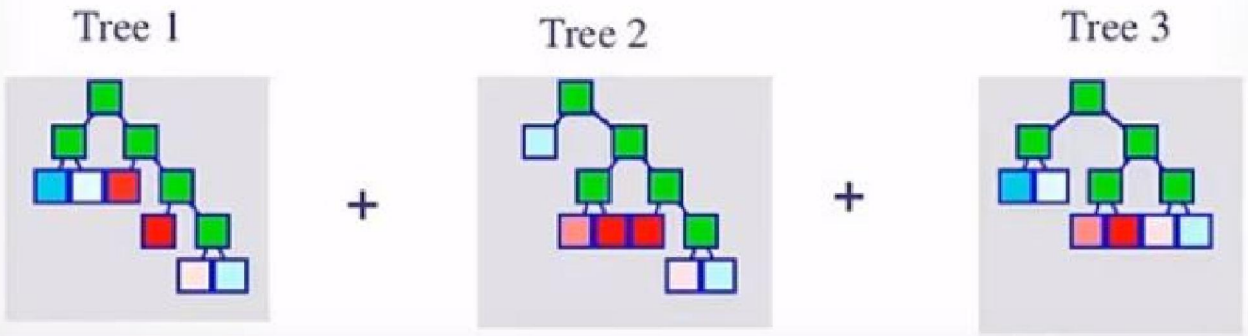
기존 회귀 모델의 문제

- Missing values
  - Results in record deletion OR
  - Requires imputation
- Nonlinearities
  - Ignores local effects
  - Requires manual transformations
- Interactions
  - Requires manual detection
- Variable selection
  - Could be thousands available
- Overfitting to the learn sample
  - Uses all available data just to build the model
  - No use of test sample to monitor performance

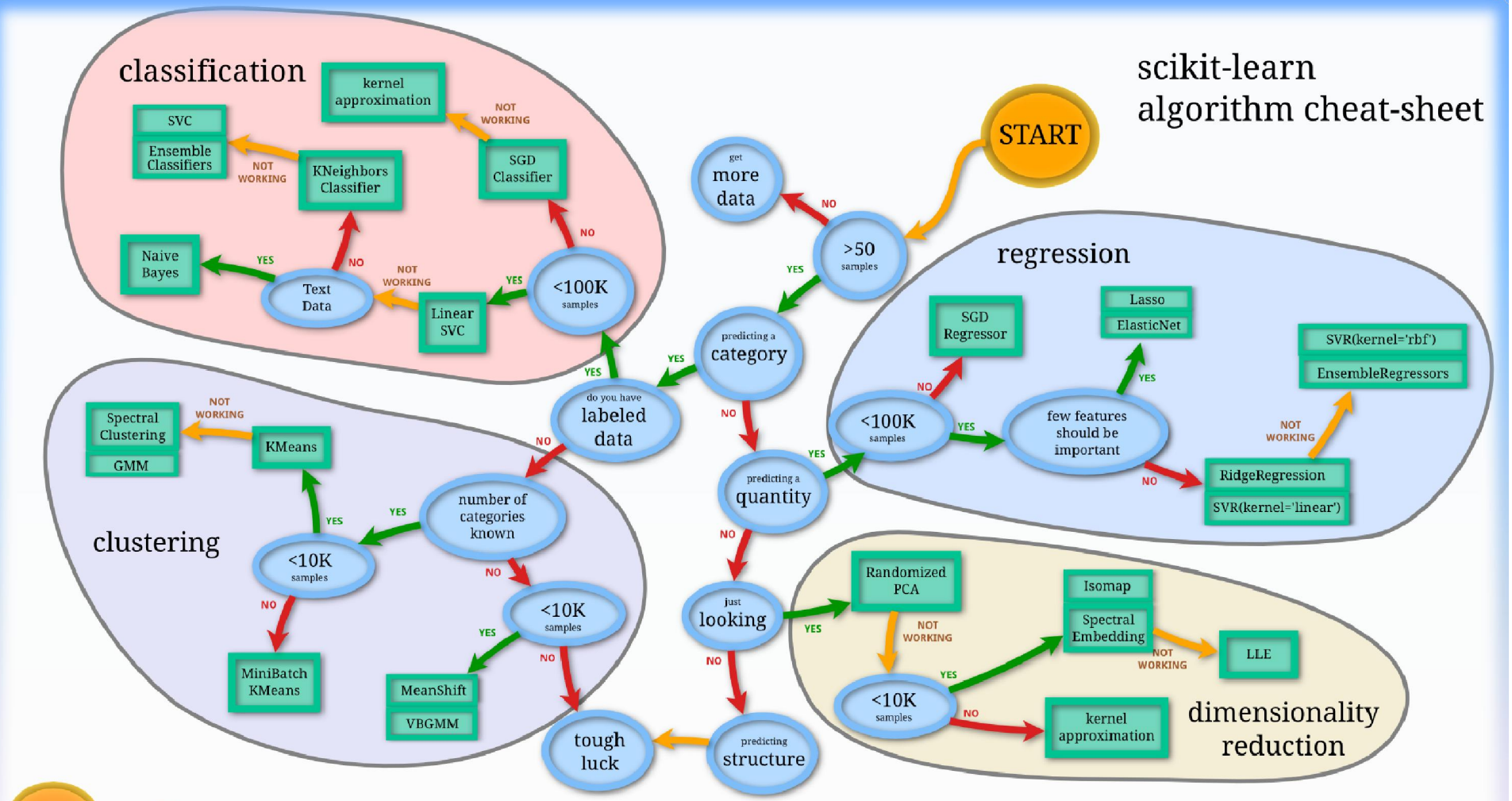
부스트 트리의 강점

Stochastic Gradient Boosting

- Small decision trees built in an error-correcting sequence
  1. Begin with small tree as initial model
  2. Compute residuals from this model for all records
  3. Grow a second small tree to predict these residuals
  4. And so on…
- Fast and efficient
- Data driven
- Immune to outliers
- Invariant to monotone transformations of variables



모델 선택시 참고할만한것



모델 세부 파라메터

오차계산은 최소절대편차(LAD)를 사용

(독립변수 개별 분산도가 높아 최소제곱법 대신 LAD 사용)

그 외 최대 결정노드 개수, 각 트리 학습도 등은 예측 영화 별로 상이하게 설정

(킹스맨2, 남한산성, 넛잡2 모두 장르 및 등급이 달라 영화별 특성에 따라 학습 데이터 및 모델 세부 파라메터를 다르게 설정.)

전체적 모델 학습과정

탑 500 영화 리스트를 통해 배우,감독,배급사 별 점수 측정

↓

그 외 데이터는 모두 크롤링

↓

14~17년도 개봉영화 관련정보에 영화별로 배우, 감독, 배급사 점수, 네이버 영화 보고싶어요 지수, 리뷰 수, 관련 기사 수, 일자별 누적관객수 추가 입력해 학습용 데이터세트 완성

↓

부스트 트리 모델에 학습 후 예측

최종 성능

- 킹스맨 2 예측용 모델 평균 오차율 14%
- 남한산성 예측용 모델 평균 오차율 18%
- 넛잡 2 예측용 모델 평균 오차율 17%

$$\text{오차율} = \left| \frac{\text{실측값} - \text{예측값}}{\text{실측값}} \right| * 100$$

최종예측값

- 킹스맨: 골든 서클 : 2,690,619명
- 남한산성 : 4,370,983 명
- 넛잡 2 : 364,289명