2018 빅콘테스트
심사 의견

서희
Hee.seo@gmail.com

2018년 빅콘테스트 참가 분야

참가분야

Innovation분야		Analysis분야	
통신	금융	퓨처스리그	챔피언리그
통신서비스데이터 외 다양한 공공데이터를 활용한 "지역생활편의지수(Index)" 개발	금융 데이터를 활용한 "나의 금융생활 정보지수" 개발	개봉(예정) 영화 관객 수 예측	게임 유저 이탈 예측 모형 개발

분석(추론) Task

예측 Task

"어떤 Task에 참가를 해야 할까?"

http://tfcstar.dothome.co.kr/archives/262

2018년 빅콘테스트 참가 분야

참가분야	Innovation분야		Analysis분야	
	통신	금융	퓨처스리그	챔피언리그
	통신서비스데이터 외 다양한 공공데이터를 활용한 "지역생활편의지수(Index)" 개발	금융 데이터를 활용한 "나의 금융생활 정보지수" 개발	개봉(예정) 영화 관객 수 예측	게임 유저 이탈 예측 모형 개발

분석(추론) Task

"어떤 Task에 참가를 해야 할까?"

http://tfcstar.dothome.co.kr/archives/262

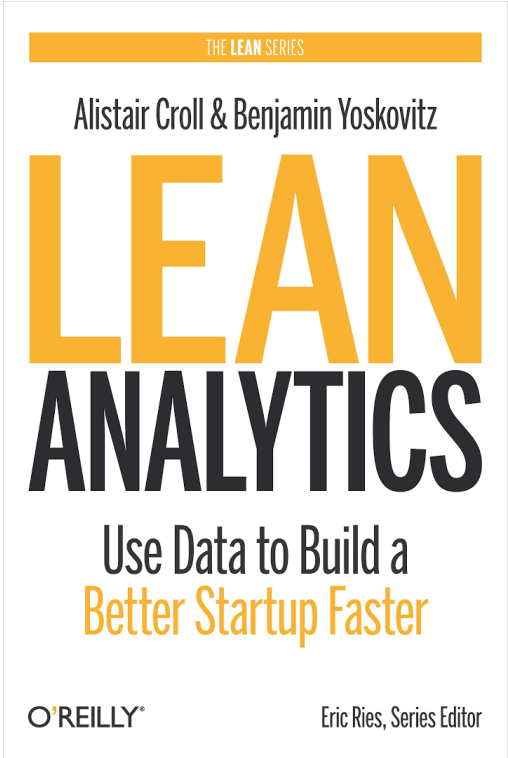
분석(추론) Task

"좋은 지표란 무엇인가?"

1) **상대적**이다
2) 이해하기 **쉬워야** 한다
3) **비율**로 표현된다 (**행동/비교**)
4) **행동 방식**을 바꾼다

참가분야

Innovation분야	
통신	금융
통신서비스데이터 외 다양한 공공데이터를 활용한 "지역생활편의지수(Index)" 개발	금융 데이터를 활용한 "나의 금융생활 정보지수" 개발



"분석의 본질은 사업에 매우 중요한 지표를 추적하는 것이다." – 린분석

http://tfcstar.dothome.co.kr/archives/262

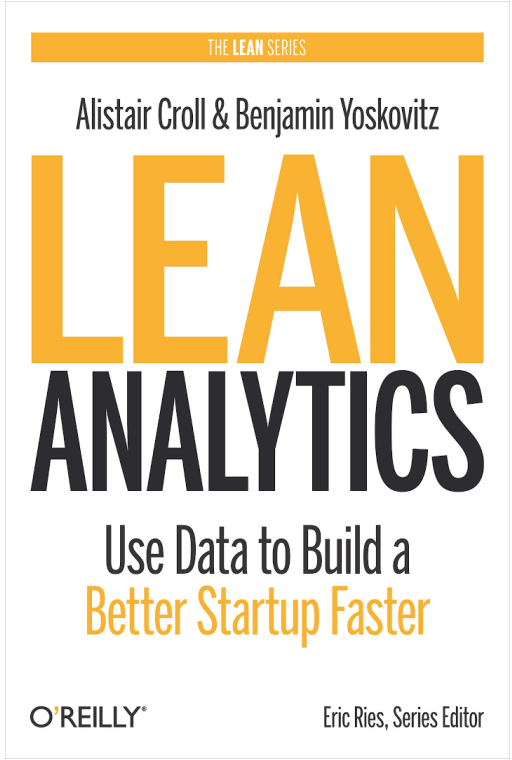
참가분야

Innovation분야	
통신	금융
통신서비스데이터 외 다양한 공공데이터를 활용한 "지역생활편의지수(Index)" 개발	금융 데이터를 활용한 "나의 금융생활 정보지수" 개발

분석(추론) Task

"올바른 지표를 선택하기 위한 고려사항"

1) 정성적 지표와 정량적 지표
2) 허상 지표와 실질적 지표
3) 탐색 지표와 보고 지표
4) 선행 지표와 후행 지표
5) 상관 지표와 인과 지표



"분석의 본질은 사업에 매우 중요한 지표를 추적하는 것이다." – 린분석

http://tfcstar.dothome.co.kr/archives/262

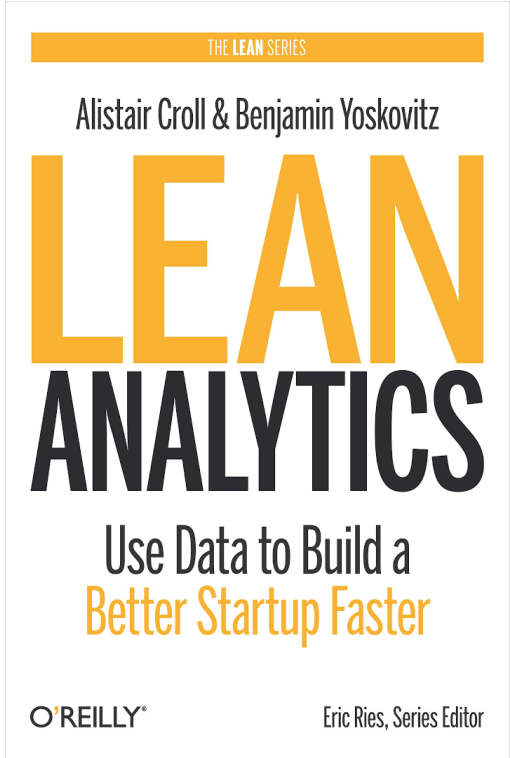
분석(추론) Task

“지표는 사업에 따라 다르다”

1) 전자 상거래
2) SaaS
3) 모바일앱
4) 미디어 사이트
5) 사용자 제작 콘테텐츠

참가분야

Innovation분야	
통신	금융
통신서비스데이터 외 다양한 공공데이터를 활용한 “지역생활편의지수(Index)” 개발	금융 데이터를 활용한 “나의 금융생활 정보지수” 개발



“분석의 본질은 사업에 매우 중요한 지표를 추적하는 것이다.” – 린분석

http://tfcstar.dothome.co.kr/archives/262

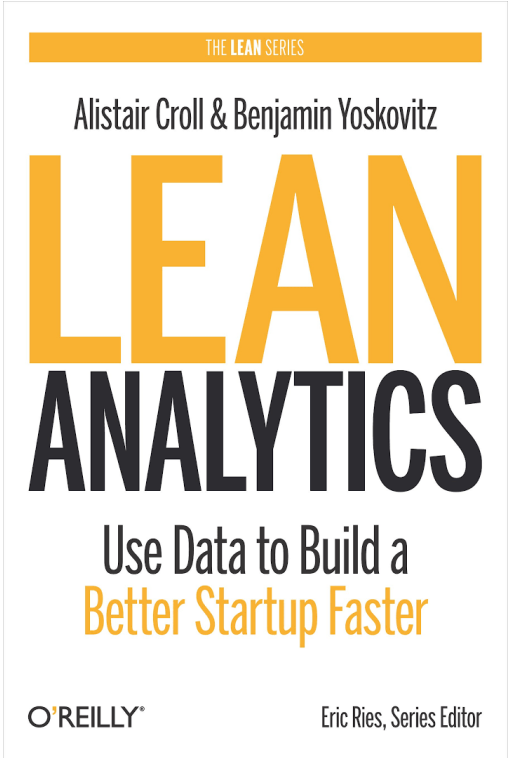
참가분야

Innovation분야	
통신	금융
통신서비스데이터 외 다양한 공공데이터를 활용한 "지역생활편의지수(Index)" 개발	금융 데이터를 활용한 "나의 금융생활 정보지수" 개발

분석(추론) Task

"지표는 사업 목적에 따라 정의된다"

1) 전자 상거래 – 구매전환율 /장바구니 /검색효과
2) SaaS – 유로 서비스 가입 /상향판매 /이탈율
3) 모바일앱 – 다운로드/ 앱크기 /매출 /생애가치
4) 미디어 사이트 – 클릭율 /세션대 클릭 비율/리퍼러
5) 사용자 제작 콘텐츠 – 업로드 성공율/체류시간



"분석의 본질은 사업에 매우 중요한 지표를 추적하는 것이다." – 린분석

http://tfcstar.dothome.co.kr/archives/262

분석(추론) Task

문제
빅데이터를 활용한 “미세먼지의 사회적 영향 분석 및 비즈니스 아이디어 제시” ▪ 유동인구데이터(SK텔레콤), 카드매출데이터(신한카드), SNS데이터(와이즈넛), 환경기상데이터(케이웨더), 유통데이터(GS리테일), 공공데이터 등 다양한 데이터를 활용하여 미세먼지로 인한 소비/경제/행동변화에 따른 사회적 영향 분석 및 예측 모델링을 통한 비즈니스 아이디어 제시

“당신은 어떤 분석을 하시겠습니까?”

2018년 빅콘테스트 참가 분야

참가분야	Innovation분야		Analysis분야	
	통신	금융	퓨처스리그	챔피언리그
	통신서비스데이터 외 다양한 공공데이터를 활용한 "지역생활편의지수(Index)" 개발	금융 데이터를 활용한 "나의 금융생활 정보지수" 개발	개봉(예정) 영화 관객 수 예측	게임 유저 이탈 예측 모형 개발

예측 Task

"어떤 Task에 참가를 해야 할까?"

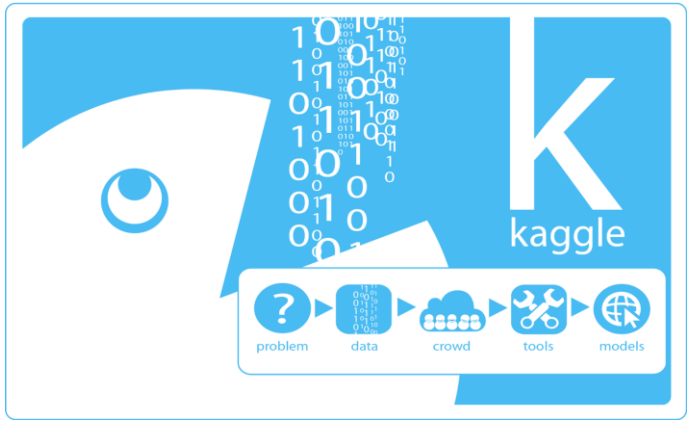
http://tfcstar.dothome.co.kr/archives/262

Analysis분야	
퓨처스리그	챔피언리그
개봉(예정) 영화 관객 수 예측	게임 유저 이탈 예측 모형 개발

예측 Task

2018년 예측 문제 경향

1) 예측 모델 단일화 (불균형/정확도)
2) 탐색적 데이터 분석
3) Feature Engineering & Data Imputation
4) 최적화 (Model Parameter)
5) ~~모델 운영~~
6) ~~재현성~~



"10 Tips to Get Started with Kaggle"

https://medium.com/@ODSC/10-tips-to-get-started-with-kaggle-fc7cb9316d27

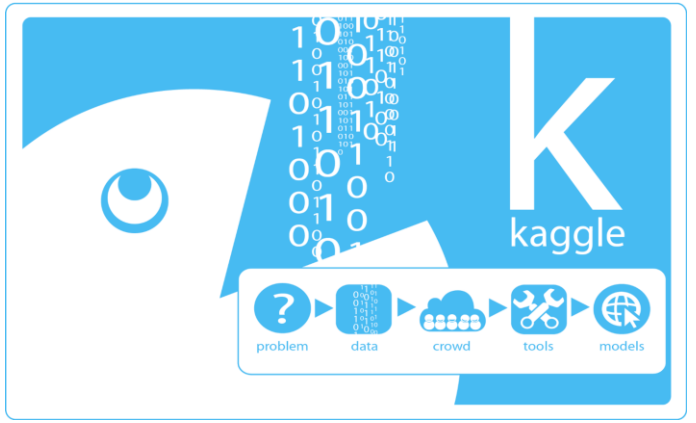
Analysis분야	
퓨처스리그	챔피언리그
개봉(예정) 영화 관객 수 예측	게임 유저 이탈 예측 모형 개발

예측 Task

2018년 예측 문제 경향

1) 예측 모델 단일화 (불균형/정확도)
2) 탐색적 데이터 분석
3) Feature Engineering & Data Imputation
4) 최적화 (Model Parameter)
5) ~~모델 운영~~
6) ~~재현성~~

"상향 평준화"



"10 Tips to Get Started with Kaggle"

https://medium.com/@ODSC/10-tips-to-get-started-with-kaggle-fc7cb9316d27

예측 Task

Analysis분야	
퓨처스리그	챔피언리그
개봉(예정) 영화 관객 수 예측	게임 유저 이탈 예측 모형 개발

예측 모델 단일화

1) 클래스 불균형
2) 정확도
3) 기타 (앙상블/스태킹)

Main GBM libraries:

spark MLlib / scikit-learn / R CRAN	H2O.ai	dmlc XGBoost	Microsoft LightGBM
•Vanilla •Some tree implementations are plain bad •As extensible as one wants	•Vanilla + TONS of tweaks •Histogram-based optimisation •Feature parallel split search •Common tasks, some extensions	•Regularized tree structure •(new) histogram-based trees •Feature parallel split search •Common tasks + full customization	•Leaf-wise tree growth •Histogram-based trees •Feature & data parallel split search •Common tasks

문제에 가장 적합한 모델을 선택

http://blog.ladonneeintelligente.com/2018/09/introduction-to-ms-lightgbm-algorithm.html

예측 Task

Analysis분야	
퓨처스리그	챔피언리그
개봉(예정) 영화 관객 수 예측	게임 유저 이탈 예측 모형 개발

탐색적 데이터 분석

1) EDA는 중요
2) EDA의 주화입마는 회피

EDA를 통해 모든 feature를 뽑을 수 없다

Analysis분야	
퓨처스리그	챔피언리그
개봉(예정) 영화 관객 수 예측	게임 유저 이탈 예측 모형 개발

예측 Task

~~Feature Engineering &~~

Data Imputation

1) Bias
2) Package
3) Domain Knowledge

過猶不及 : 과유불급

지나친 것은 미치지 못한 것과 같다는 뜻.

평균와 **Package**가 답이 아닐 때가 많다

Analysis분야	
퓨처스리그	챔피언리그
개봉(예정) 영화 관객 수 예측	게임 유저 이탈 예측 모형 개발

예측 Task

최적화

1) Package
2) 과적합
3) 시간

Approach	*ML*	*DL*	*Manual/ Auto*	*Cost*	*Space expl.*	*History*	*Parallel/ Distributed*
Babysitting	👍	👎	🐼	💰	Low	✅	No
Grid	👍	👎	💻	💰 💰 💰 💰	High	🙈	Yes
Random	👍	👍	💻	💰 💰 💰	Medium	🙈	Yes
Bayes SMBO	👍	👍	💻	💰 💰	Medium - Driven	✅	*

* The SMBO by definition is sequential, however, *it's possible, but not trivial,* to build parallel/distributed solution upon these optimizations.

최적화를 잘 하는 팀은 많지 않았음

https://blog.floydhub.com/guide-to-hyperparameters-search-for-deep-learning-models/

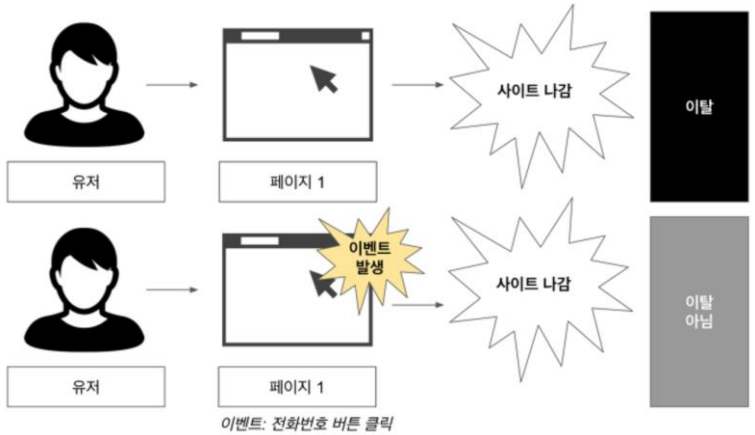
예측 Task

Analysis분야	
퓨처스리그	챔피언리그
개봉(예정) 영화 관객 수 예측	게임 유저 이탈 예측 모형 개발

~~모델 운영 & 재현성~~

1) 예측 결과가 좋으면 어떤 일을 할 수 있는가?

2) 코드와 데이터를 어떻게 전달 할 수 있을까?



예측 문제 이지만 활용과 재현성을 꼭!

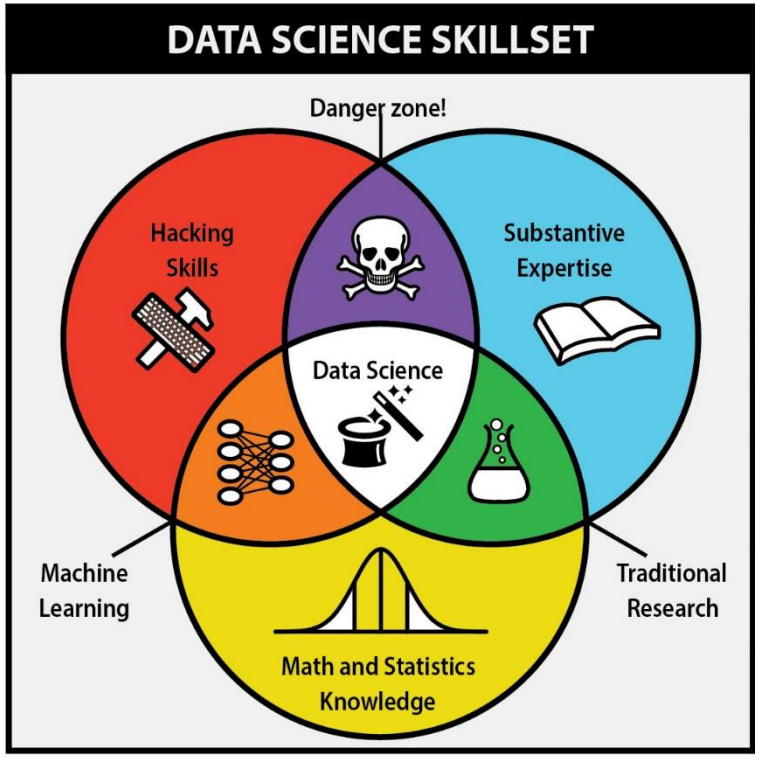
https://www.firstpythonnotebook.org/markdown/

예측 Task

구분	문제
퓨처스리그	항공 운항 데이터를 활용한 “항공 지연 예측” ▪ 항공 시즌 스케줄, 운항데이터 등 항공운항데이터(한국공항공사)와 항공기상데이터 등을 활용하여 항공지연 예측 모형 개발을 통하여 9월 16일부터 9월 30일까지의 항공편별 지연 여부 예측
챔피언리그	게임 활동 데이터를 활용하여 “게임유저 잔존가치를 고려한 고객 이탈 예측 모형” 개발 ▪ 엔씨소프트에서 제공하는 ‘리니지’ 고객 활동 데이터를 활용하여 향후 고객 이탈 방지를 위한 프로모션 수행 시 예상되는 잔존가치를 산정하는 예측 모형 개발

“당신은 어떻게 예측 모형을 만드시겠습니까?

데이터 분석가에게 필요한 SKILLSET 이란...



도메인 전문가

데이터 **핸들링**

문제 정의 & 모델링

가설**검증** & **최적화**

의사**결정** & **서비스**

Good
Storyteller

https://www.slideshare.net/HeeSeo/skillset-123722714

E.O.D