
게임 산업에서
데이터
사이언스
netmarble

강사 소개

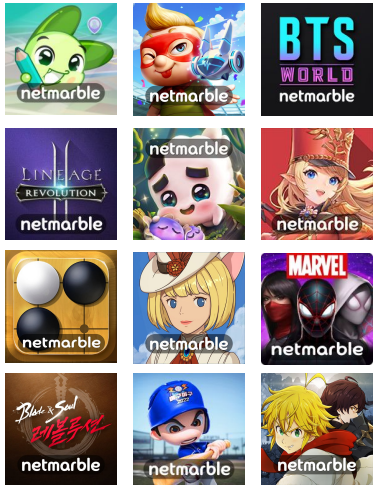
서창우, PhD/MBA

- 학력
  - KAIST 경영대학원 경영공학 박사, 테크노 MBA (-2017)
  - Research fellow at Brunel University London, UK (-2019)
- 경력
  - TmaxSoft 연구원 (-2005)
  - NCSoft Lineage2 (-2008)
  - 넷마블 AI 센터 (2020-)

게임 데이터는?

- 데이터계의 백화점

- netmarble 데이터는?



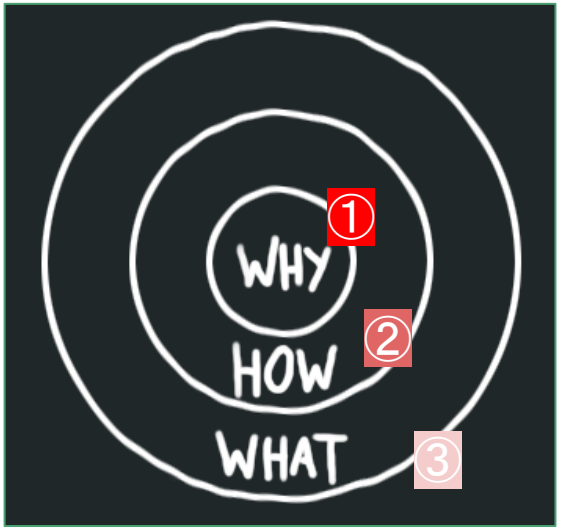
넷마블 데이터는?

- 게임에서 발생하는 모든 행동을 raw data로 트래킹
- 특정게임 1일 데이터 용량이 700GB가 되는 게임도 있음
- 클렌징 필수!

게임회사 직원들은 무슨 일을 하나요

- 광고 효과분석 (마케팅)
- 이상 탐지
- 소셜 미디어 분석 (SNS)
- 유저 플레이 패턴 분석
- 밸런스 분석 (재화, 경제, 허들)
- 상품 추천 (쇼핑)
- 그리고 게임...

문제 의식



WHY **(목적)**
현황 파악을 통한 문제 도출

HOW **(방법)**
why에 근거한 문제 해결 방법

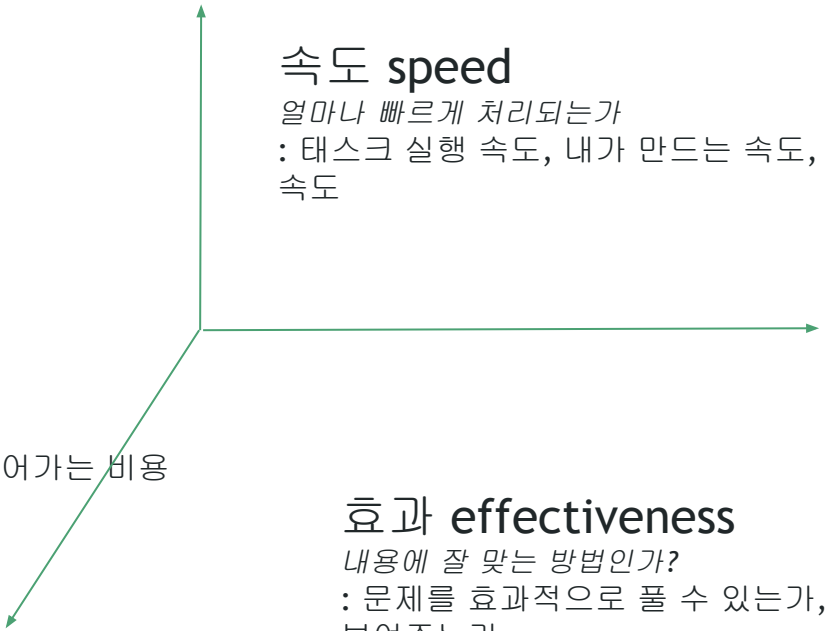
WHAT
why의 결과로 나오는 최종 프로덕

분석할 때 어떤 방법론을 많이 쓰나요

- 데이터 전처리
- 전통적인 통계 분석 방법론
- 머신 러닝
- 딥 러닝
- 데이터 시각화
- ...

No silver bullet

가장 효과적이면서도 효율적인 방법은 없음



속도 speed

얼마나 빠르게 처리되는가

: 태스크 실행 속도, 내가 만드는 속도, 팀원과 코웍 속도

효율 efficiency

투자 비용 대비 효과

: 태스크 수행 비용, 만들때 들어가는 비용
유지보수 비용 등 고려

효과 effectiveness

내용에 잘 맞는 방법인가?

: 문제를 효과적으로 풀 수 있는가, 결과를 효과적으로 보여주는가

무엇을 배워야 하나?

데이터 전처리

- 데이터를 **column** 단위로 보고 분석할 수 있는가?

- **SQL, SQL, SQL**
  - 가능한 부분은 SQL에서 처리할 수 있도록 훈련 필요
    - SQL은 대규모 시스템으로 빠른 처리 가능 (예. BigQuery, SnowFlake, RedShift, ...)
- 익숙하지 않은 데이터라면 **Excel**로 데이터 디테일 확인하고 **prototyping**
  - column 내에 어떤 내용이 있는지 엑셀과 같은 스프레드시트로 확인
  - 데이터베이스 연결 기능을 이용하여 엑셀에서
- Python 이용한 전처리는 비용 (시간, 가격, 데이터 입출력)이 더 필요함
  - 그래도 어쩔 수 없이 사용해야 하는 경우 발생
  - Pandas는 필수
  - column 단위 처리를 위해 for-loop 돌리는 일은 없어야 함
    - 대신 apply, lambda 함수 사용

(전통적인) 통계 방법론

- 검정 (내 가설이 맞는지 틀린지 확인)
  - 두 집단이 같은지 다른지?
  - 상관관계

- 사후 효과 분석
  - 인과관계
  - 시계열 데이터 분석

- Python에서는?
  - 기본적인 통계량 추출은 Pandas, Pandas profiling으로 충분
  - 기본적인 검정을 넘어 회귀분석이라도 하려면 statsmodel, linearmodels, ...

머신러닝

- 예측
  - 이 다음 값이 어떻게 될까?
  - 이탈/유지
  - 구매

- 어떤 ML 알고리즘을 사용하는지도 중요하나...
  - 적절성
  - 데이터 유효성
  - 피처 선택

- Python에서는 scikit-learn이 표준이라 생각하면 됨

딥 러닝

- 예측을 좀 더 잘 하고 싶다면...
  - CNN, RNN, LSTM, Attention model, Transformers, ...
  - 각각을 사용하는 수준에서 끝나면 안되고, 각 알고리즘에 대한 "이해"가 필수.

- 비용, Pre-trained models
  - GPU는 언제나 필수 (안그래도 비싼데...)
    - 참고: Bert의 확장
  - Bert와 같이 큰 pre-trained model을 만들기 보다는 이해하고 잘 사용할 줄 알아야 함
    - 참고: 당근마켓 사례
  - Pre-trained model도 서비스로 사용해야 하는 환경으로 바뀌는 중
    - 참고: GPT-3 유료화 결정

- 무엇을 공부해야 하나?
  - Python/Tensorflow/Keras/Tensorflow lite
  - 다른 언어/패키지는 "추가" 옵션

무엇을 알아야 할까?

- 데이터 사이언티스트 모집 공고 공통사항
  - 필수 사항
    - **데이터 전처리**
    - **통계** 혹은 관련 전공 (경영학 포함)
    - **SQL** 사용 가능
    - **프로그래밍** (대체로 python) 가능
    - 다양한 직군의 사람들과 **원활한** 커뮤니케이션
    - **협업** 가능
  - 우대사항
    - 여러가지가 있지만 ...
    - "**클라우드 환경** 경험"
      - GCP, AWS, Azure

주요업무

- 게임 데이터 분석 (유저행동 분석)
  - 게임서비스 고도화를 위한 유저세그먼트 및 지표 개발
  - 게임데이터 기반의 통계분석 및 예측모델 개발,
    머신러닝, 데이터분석 관련 신기술 탐색 및 적용
  - 개인화 ID 개발

- 게임 데이터 분석 (시계열 이상탐지 데이터 분석 및 모델설계)
  - 어뷰징 행위 탐지를 위한 게임로그 탐사분석
    게임로그 기반 시계열 데이터 분석
    ML 모델 설계를 위한 Feature Engineering 수행
  - ML 기반 어뷰징 행위 탐지 모델 설계
    탐지 성능 평가 및 지속적 개선

- 마케팅 데이터 분석
  - 디지털 마케팅 최적화를 위한 분석 업무 수행
  - 디지털 마케팅 최적화를 위한 각종 예측 모델 제안 및 개발
  - 머신러닝 서비스 개발/운영

자격요건

- 관련 분야 경력 3년 이상
- 정형/비정형 데이터 분석 및 모델링 가능
- 통계 분석 도구 (R, Python, Tensorflow 등) 및 SQL 사용 가능
- 다양한 직군의 사람들과 원활한 커뮤니케이션 및 협업 가능

우대사항

- 통계학 또는 관련 전공
- 시계열 데이터 분석 경험
- 모바일 마케팅 분석 경험
- 클라우드 환경 경험

데이터 사이언티스트가 **꼭 할 수 있어야** 하는 스킬들

- 데이터 전처리
  - 10만 줄 이하: 엑셀로도 OK
  - 1000만 줄 (10GB) 미만: 내 컴퓨터에 설치한 RDBMS, NoSQL로 OK
  - 그 이상: Cloud DW, Spark clusters, Hive clusters
- 데이터 다루기
  - 엑셀 함수 (vlookup, if, iferror, sumif, countif …)
  - SQL (select 문, 서브쿼리, 최적화 방법)
  - BigQuery (Standard SQL, BigQuery on jupyter notebook, Airflow 2, …)
- 데이터 분석
  - SQL (aggregation), BigQuery
  - Python (R, Java, … 데이터 사이언티스트 실무에서 거의 사용하지 않습니다.)
    - Pandas, scikit-learn, statsmodel, tensorflow, …
  - 데이터 시각화 on python
    - matplotlib, seaborn, pandas profiling, ...

Free credit/tier on GCP, AWS, Azure

- GCP
  - 한 계정당 / 한 번 / 일년간 사용할 수 있는 $300 의 free credit 제공됨
  - Google Drive의 **Google Colaboratory** (Colab)에서 Cloud Storage, **BigQuery** 연결 가능
    - Colab을 사용하면 Computing Engine/AI Notebook을 따로 사용할 필요 없음
    - Colab에서는 제한적이지만 **TPU/GPU** 무료 사용 가능
      - github 연결/배포 가능
- AWS/Azure
  - Free tier

- "회사에서" 가장 많이 사용되는 클라우드 플랫폼은 AWS
  - 현실적인 **장벽(?)**을 고려한다면 BigQuery/AI Notebook/Composer 등이 있는 **GCP** 사용해 보길 권함

질문 및 답변 (1/3)

- 예상 질문1) 무엇을 알아야 할까요?
  - 필수: 엑셀, SQL, Python, Pandas, scikit-learn
  - 옵션1: 컴퓨터 아키텍처, (python) numpy, lambda function, BigQuery, statsmodel
  - 옵션2: linux 기본 명령들, (python) 병렬 프로세싱, GCP command tools, tensorflow, keras

- 예상 질문2) R이나 NoSQL에 대해 잘 알고 있어도 도움이 될까요?
  - 답변: R을 사용하는 회사는 별로 없습니다. 취직이나 이직에 큰 도움이 되지 않습니다.
  - 답변: NoSQL은 필요할 때 배워서 사용하시면 됩니다.
  - 답변: Spark, Hive 등의 분산처리는 스킬은 필수가 아닙니다.

- 예상 질문3) pytorch를 쓰면 안되나요?
  - 회사에서는 일반적으로 tensorflow를 사용합니다. 써도 되지만 TF 다시 공부해야 할 가능성이 훨씬 큽니다.

- 예상 질문4) 말씀하신 내용들 다 알고 있는데, 하나만 더 공부해야 할 내용 알려주신다면?
  - 답변: 인과관계 분석에 대해 공부해 보세요

질문 및 답변 (2/3)

- 현직자 선배분이 현업에 대해서 강의를 해주실 때에 도메인에 대한 지식 또한 높이 평가하시는 걸 들을 수 있었습니다. 게임 산업의 데이터 사이언티스트 직무로 활동하기 위해서는 해당(게임) 도메인에 대한 지식을 어느 정도 갖추고 있어야 하는지 궁금합니다.

- 채용 사이트 "원티드"에서 넷마블 데이터 사이언티스트 포지션에 "GCP와 같은 클라우드 환경(AWS, Azure) 경험이 있으신 분"을 우대한다고 하시는 걸 볼 수 있었습니다. Google BootCamp 과정에서 GCP ML Engineering Certification / GCP Data Engineering Certification 등 GCP 관련 자격증을 취득하는 내용이 있는데, 이 자격증도 우대사항에 포함될 수 있을지 궁금합니다. 또한, 해당과정이 직무에 얼마나 도움이 될지 궁금합니다.

- 넷마블 데이터 사이언티스트 직무에서도 5년 이상 경력을 요구하고 있습니다. 하지만, 대학생 신분으로서 관련된 부분을 많이 경험할 수 없는게 현실인 것 같습니다. 따라서 추천을 받아 Dacon/ Kaggle 과 같은 곳에서 데이터를 다뤄보고 있는데요, 학생이 Kaggle/Dacon Competition에 참여하는 것을 어떻게 평가하시는지 개인적인 의견을 듣고 싶습니다!

질문 및 답변 (3/3)

- 전체적으로 직무에서 3년 이상의 경력을 요구하고있는데 인턴이나 신입에 대한 채용 일정을 문의드립니다.

- 강화학습 엔지니어 직무에 관심이 있습니다. 유니티로 환경을 만들고 ml-agent를 이용하여 학습시키는 등의 프로젝트가 입사에 도움이 되는지 궁금합니다. 실제 회사에서도 이런 방식의 시뮬레이션 관련일을 하는지 궁금합니다.