Q. 공채와 관련하여 포트폴리오는 어떻게 준비해야 하는지 궁금합니다.

A. 신입공채는 특별한 커리어 보다는 자세(Attitude)를 보려는 경향이 있습니다. 열정과 하려는 자세가최대한 드러나는 것을 어필하는 것이 좋습니다.

<빅데이터 분야 석박사 학위자 입사시험문제>

주어진 가상의 데이터셋(테이블 형태) 을 주어주고 어떻게 중요 feature를 도출하고 분석결과를 도출할 것인지 가정을 설명하라는 문제가 주어짐(데이터 해석, 전처리방법, 스코어링 방법, 가설 수립 등에 제한이 없고, 자유롭게 논리적으로 설명할 것이라고 명시)

<빅데이터 프로세스 분석 과정>

* 분석 프로세스

주제 선정 -> 데이터 수집 -> 전처리 및 저장 -> 데이터분석/모델링 -> 모델검증

최적화 ↖ ← |

주제 범위 선정 -> 세부 요구사항 수립 -> 필요 데이터 탐색 -> 데이터 수집 방법 결정 ->

데이터 수집 / 저장 -> 데이터 품질 확인 -> 데이터 마트 생성 -> 성능검증(A/B Test) ->

모델 검증 -> 데이터분석/ 모델링 -> 전처리 -> 데이터 탐색 -> 운영 시스템 적용 ->

모니터링 -> 유지보수

☞주제를 선정한 프로젝트의 예시

주제 선정: 주제(고양이 맞추기)

세부 요구사항

1. Input : 고양이 그림 Output: 고양이
2. 데이터 수집: Crawling & Download

데이터 수집:

전처리

분석/모델링

모델 검증

< 인물로 알아보는 NN 역사>

-----|---------------------|------------------------|-----------------------|---------------------|-------------------|--------🡪

1943 1957 1969 1974 1986 2006

제프리 힌튼

Deep Belief

Network

(심층신뢰신경망)

이라는

계층간에는 연결이 있지만

걔층 내의 유닛

간에는 연결이

없는 신경망 제시

↓

Vanishing

Gradiant

문제 해결

제프리 힌튼,

역전파를 재발견하여

세상에 내놓음

Multi-layer

Network

(Deep neural

Network)

를 제시함

힌튼,

데이비드

럼멜하트,

로널드

윌리엄스가

오류 역전파

알고리즘으로

다중 퍼셉트론을

학습시키는데

성공함

폴 워보즈가

박사과정 논문에서

‘역전파

(Backpropagation)’

를 제안

앞의 진행방향에서

고쳐가는 것이 아닌

결과를 보고 뒤로

가면서 weight와

Bias를 조정하는

알고리즘

마빈 민스키,

시모어 패퍼트가

퍼셉트론의 한계를 수학적으로 증명

(XOR연산이 불가능)

헵 의 학습이론

(두 뉴런 사이의 연결강도를 조정할 수 있는 학습이론)

“Cell A가 B를 자극시키는데 충분하고, 지속적인 원인으로 작용한다면, cell B에 대한 cell A 작용력이 내부적으로 변화를 일으켜 증가하게 된다”

Ex/ 파블로프의 개 실험(신경강화)

워터필츠,

워렌맥컬린

Eletronic

Brain

이라는

개념 제시

프랭크 로젠블럿의

‘퍼셉트론’

최초의 신경망

모델을 발표

선형분류를

수행할 수 있는

‘피드포워드

뉴럴네트워크’

제안

<Backpropagation 의 이해>

forward propagation을 통해 필요한 값들을 미리 저장해두고, backward propagation이 진행되면서 위에서부터 loss에 대한 derivative를 하나하나 계산해나가면서 다음 layer에서 바로 전 layer에서 계산한 값들과 각 neuron 별로 추가적으로 필요한 derivative들을 곱해나가면서 weight의 derivative를 계산하는 알고리즘이다.

이렇게 한 번 전체 gradient를 계산한 다음에는 learning rate를 곱하여 전체 parameter의 값을 update한 다음, 다시 처음부터 이 과정을 반복한다. 보통 에러가 감소하는 속도를 관측하면서 ‘이 정도면 converge한 것 같다’ 하는 수준까지 돌린다.

익숙해지려면 다소 시간이 걸리지만, 개념적으로 먼저 ‘error를 먼저 계산하고, 그 값을 아래로 전달해나가면서 바로 전 layer에서 계산한 미분값들을 사용해 현재 layer의 미분값을 계산한 다음, 그 값을 사용해 다음 layer의 미분값을 계산한다.’ 라고 개념만 이해해두고 다시 차근차근 chain rule을 계산해나가면서 계산하면 조금 편하게 익숙해 질 수 있을 것이다.

* DBN, RBM,CNN, RNN 등 신경망 용어별 개념 정리

° 영등포구 축제 홍보 시스템

°<https://www.youtube.com/watch?v=v3nK56kzA2I> KTaxi 외래관광객 택시투어 추천시스템(2018)

* 의미: 택시차량을 이용해서 접근성이 어려웠던 관광지투어 등 여행스케줄을 진행하는 투어
* 사례: 타이베이 근교 택시투어 탄생배경 및 성공비결

관광 집중도가 높았던 타이베이, 근교 관광지를 묶어 택시투어 형태로 1day 투어 서비스 제공

20대 젊은 층으로부터 높은 호흥, 고정 관광지 화(化)

(미디어에 등장한 스타 관광지, 젊은층의 호응으로 시작된 바이럴 마케팅 선순환 효과)

* 국내 외래관광의 문제점: 외래관광이 서울에 한정되어 있으며(78.8%) 이에 따라 서울 관광자원 고갈의 우려됨 (단순 소비성 관광에만 국한됨)
* 시사점: 외국인 관광객의 재방문을 위해서 새로운 관광지를 유도할 수 있는 방안을 모색해야 함
* 목적 : 외래관광객을 대상으로 서울 근교 다양한 관광지를 체험할 수 있는 택시투어 플랫폼 개발

페르소나 선정(프로젝트 타겟), 경기도 관광객 정의 및 현황분석

많은 관광객이 방문할 것, 오랫동안 체류할 것, 소비를 유발할 것

* 많은 관광객이 방문하는 곳은 어디인가, 오랫동안 체류하고 있는 곳은 어디인가, 소비를 유발하는 곳은 어디인가(질문 구체화)
* 경기도 지역 내 많은 관광객이 방문하는 곳에 대한 이해, 경기도 내 오랫동안 체류하는 곳에 대한 이해, 경기도 내 소비를 유발하는 곳에 대한 이해(insight 발견)
* 많은 관광객이 방문하는 곳은 어디인가?

국내 관광객 대비 외래관광객이 많이 찾는 지역, 지역별 최다 방문국적, 지역별 최다 방문 시간대

국가별/우러별/시간대별 선호하는 지역이 달라지며 이를 고려한 추천시스템이 구성되어야 함

중국/태국/미국 관광객 타겟 전략 또는 일본/대만 관광객 타겟 전략을 나눠서 접근할 수 있음(시사점)

* 오랫동안 체류하고 있는 곳은 어디인가?

지역별 외래관광객 숙박비율, 지역별 내국인관광개개 숙박비율, 지역별 외래관광객 평균숙박일수

도심지역에서 숙박하는 외래관광객을 감안하여 픽업을 위한 출발지, 도착지 선정 필요

외래관광객은 하루 단위 투어를 선호하기 때문에 하루 단위 투어 추천시스템이 구성되어야 함 (시사점)

* 소비를 유발하는 곳은 어디인가

전국 지역별 전체소비 대비 관광지주변 소비 비율, 경기도 지역별 전체소비 대비 관광지 주변 소비 비율경기도 관광지별 관광객 1명당 소비액

관광지별 관광객수와 관광지주변 소비액은 상관관계 X (시사점)

* SWOT 분석 : 관광전략 시사점 도출

ST전략 ‘서울에서 가깝다는 이점을 살려 서울에서 숙박하는 관광객을 위해 1day 투어로 구성하자’

WO전략 ‘정부주도 택시산업 발전 계획을 활용해 미흡했던 지자체 간 관광지 연계 프로그램 구성하자’

* 추천모델구현
  + Classification 모델용 데이터셋 준비

독립변수 선택 -> 변수 소각 과정 -> 변수 유의성 검증(독립변수/종속변수간 1:1검증) ->

다중공선성(독립변수간) 검증 -> 변수 EDA -> 변수 전처리 -> 파생변수 생성(모델의 설명력을 높이기 위해 독립변수 12개에서 파생된 변수 생성) -> 데이터 리샘플링 -> 파생변수 재평가

* + Classification 모델용 데이터셋 준비

분류모델 선택(XGBoost, RandomForest, SVM, LGBM, Catboost), 관광지 테마 예측

* + Clustering용 데이터셋 준비

관광지 위경도 좌표 수집(Bigsight, TourAPI 활용) -> 관광지 제거

* + Clustering 모델링

1단계: 군집 모델선택

<데이터 분석을 이용한 COVID-19 예측 모형>

: 지금까지 전염병과 같은 예측 불가의 현상은 유사 사례에 기반한 주관적인 경험적 예측 방법을 사용했지만, 이제는 빅데이터를 이용한 객관적인 데이터 기반의 예측이 가능.

전염병은 확신 시점과 발병 시점간 차이가 크기 때문에 ‘지역 폐쇄’ 와 ‘철저한 검진’ 이라는 선택지를 가질 수 있어 정책적 혼선이 나올 수 있음

중국을 포함한 다른 국가에서 코로나 감염과 관련한 데이터를 정확하게 공개하지 않거나, 감염자를 적극적으로 찾으려는 노력이 없어 데이터를 이용한 예측을 곤란하게 만듬.

Ex / 데이터의 신뢰성

2020/3/1 기준으로 한, 중, 일 등 주요 코로나 발생 국가의 확진자를 Log 스케일 반영

일본내 코로나 확진자는 한국과 중국을 포함한 다른 나라에 비하여 상대적으로 낮기 때문에 상대적 비율인 완치율은 높게 나타남 -> False negative의 중요성