

9월 8일 회의

1. 날씨 데이터 컬럼 정리한 거 같이 보기

2. 날씨 지표 생성 논의

- 어떤 지표로 사용하는 것이 좋을까?

→ NET 지수

활용방안)

a) 0%~100% 값을 -1~1 사이 값으로 변환

b) 5%/95% or 1%/99% 이하/이상의 값을 이상치로 판단, 스트레스 있는 날로 반영

c) 각 기간에 스트레스 가중치를 부여하여 예측에 활용

⇒ NET 지수를 어떻게 계산할 것인가

: 24시간 min, max로 NET 구하고 시각화

(max의 상위 5%가 이상치, min의 하위 5%가 이상치 ⇒ 이상치 내에서 1~5로 차등)

실외, 실내 공연 구분 필요성

: (내외여부) x (NET 이상치 여부)

⇒ 회귀분석을 진행해서 양의 잔차를 보인다는 것 (예측인원보다 실제인원이 더 많다), 해당 공통 점을 파악하기

(자외선 지수

: 주기적인 그래프를 보이기에 타지표나 자외선 지수 데이터를 12년으로 확장해서 사용해도 될 것 같음.)

3. 기존 모델에 날씨데이터 어떻게 적용할 것인지

- 행사인원 예측모델 (행사가 하루에 걸쳐 진행되지 않는다는 특성/ 어느 날을 기준으로/ 행사 기간 평균으로)

→ 기온 습도: NET 활용 (이상치 여부 1~5 차등)

→ 강수 대신 기상 특보 여부 (태풍, 폭염, 한파 등) → 태풍 데이터 O

→ 폭염 여부 (일 최고기온이 33°C 이상)

→ 한파여부 (일 최저기온이 -12°C 이하)

→ 폭우/폭설 뉴스 크롤링

→ 폭우 폭설 네이버 뉴스기사 일자

- 지하철 교통량 예측 모델 - 시간대별 날씨 데이터 적용, 코로나 이후로 시기 한정

4. 사업화 계획

1) B2B (행사인원 예측모델)

고객: 행사 주최사(기업)

경쟁사:

자사:

→ 웹사이트 구현

2) B2C (교통 이상치 탐지 모델)

-날짜, 시간 등 입력하면 대중교통 혼잡도 알려주는 앱

고객: 행사 참여자, 관객 + 정부부처

경쟁사: T맵, 카카오맵 등

차별화전략: 차량 혼잡도 제공하는 앱은 기존에 많지만 지하철의 경우 해당 앱이 존재하지 않는 것. 사실 있었음 t맵 대중교통

→ 날짜, 시간 입력하면 교통 혼잡도 알려주는 방식....?

→ 앱으로 구현

5. 해야할 일

1) 행사데이터 일자 정리/ 벡터 모델 생성

2) 날씨데이터

- NET 계산 (min, max 계산하고 이상치 여부 → 이후 1~5 차등 점수)

- 실내, 실외 여부와 NET과 상관이 있는가
- 폭우, 폭설 뉴스기사 일자 크롤링
- 태풍 데이터 (직접영향, 일자)
- 폭염, 한파 계산 컬럼 추가

3) 행사데이터 (요일, 계절 등등 컬럼 추가)

추가 변수

요일 - 월~일에 따른 더미화

계절 변수 - 봄~겨울에 따른 더미화

전후 휴일여부 - 휴일 전후에 따른 변수 더미화

연휴, 연말 여부 - 연휴, 연말에 따른 변수 더미화 (3일 연속 연휴)/ (12/21~31)

날씨 데이터 - 폭염, 비, 눈에 따른 더미화/ 체감온도, 강수량의 수치형 데이터

+ (요일별 평균인원/ 월별 평균인원/ 전주 인원)= 회귀분석으로 해당 변수의 설명력을 파악한 뒤 변수 포함여부 결정

4) 행사벡터화 모델 + 기존 이상치 탐지 모델 합치기

5) 행사인원 예측모델 예측률 높이기 - 예측방법 여러가지 적용

6. 일정

(~9/14)

권소영-행사데이터 일자 정리

김수민-행사데이터 (요일, 계절 등등 컬럼 추가)/ 태풍 데이터 (직접영향, 일자)/ 폭염, 한파 계산 컬럼 추가

정성희-행사데이터 일자 정리

오지민-벡터 모델 생성/ 폭우, 폭설 뉴스기사 일자 크롤링

이세은-NET 계산 (min, max 계산하고 이상치 여부 → 이후 1~5 차등 점수)/ 실내, 실외 여부와 NET과 상관이 있는가

(~9/19)

회귀분석으로 유의한 변수들 찾기

(이후)

이세은- 행사벡터화 모델 + 기존 이상치 탐지 모델 합치기

나머지- 행사인원 예측모델 예측률 높이기

(중간고사 이후)

ppt 완성 및 웹사이트 구현