(1페이지)

안녕하세요 3팀 발표 시작하겠습니다. 발표를 맡은 변주섭입니다. 저희 팀이 선정한 주제는 사내식당 식수인원 예측입니다. 즉 식사시간에 회사 내 식당을 이용하는 수요인원의 수를 예측하는 것입니다. 지금까지는 단순한 시계열 추세와 식당 담당자의 경험과 직관에 의존하여 식사 수요 인원을 예측했습니다. 그러나 이제는 빅데이터 분석을 기반으로 식수 인원 예측의 정확도를 높여 과다한 잔반 발생 등의 문제를 예방하고자 합니다.

(2~4페이지)

저희 팀은 4명의 팀원으로 이루어져 있습니다. 3명의 팀원은 데이터 수집 및 전처리, SQL과 머신러닝을 활용하는 작업을 진행하였고, 팀장님은 DB구축과 R 통계분석을 비롯하여 각 팀원들이 작업한 내용을 깃허브를 통해 정리했습니다.

발표는 프로젝트 개요, 분석결과, 모듈, 향후 보완할 점을 순서대로 설명하겠습니다.

(5페이지)

앞서 말씀드렸듯, 집단급식소 이용 인원 예측은 담당자의 경험에 의해 대부분 이루어졌습니다. 다시 말해, 데이터를 기반으로 한 체계적인 식수 인원 예측이 이루어지지 않는다는 것입니다. 실패한 예측은 과다한 잔반 발생, 음식 및 서비스 품질 저하, 혹은 음식이 부족한 사태를 초래할 수 있습니다. 추가로 이에 따른 비용이나 이용객의 불만이 발생할 수도 있죠. 정확한 데이터기반 식수 예측 알고리즘이 필요함에도 충분한 연구는 이루어지지 않았습니다.

저희 3팀은 현장에 맞는 식수 알고리즘을 적용해 미배식 잔반의 감소를 통한 원가 절감, 조리 작업의 효율화 등의 효과적인 운영을 기대합니다.

또한 학교, 병원, 공공기관 등의 집단급식소에서 데이터를 기반으로 맞춤형 식수 예측 기법을 적용하는데 기여하고 맞춤 모델링을 통해 식수 예측과 관련된 실무적인 활용도를 높일 수 있는 틀을 마련하고자 본 프로젝트를 진행했습니다.

(6~7페이지)

저희는 L사의 구내식당 데이터를 중점으로 작업했습니다. 저희는 본 프로젝트를 통해 L사와 같은 회사의 구내식당 데이터를 csv로 받았을 때 식수 예측 서비스를 제공하는 급식 업체의 역할을 다음과 같이 하게 됩니다.

도식으로 보시면 먼저 분석에 필요한 데이터를 파이썬으로 병합하는 전처리를 진행합니다. 이후 SQLite에 데이터를 5개 테이블로 분할해 DB를 구성하고요. 이어서 R을 이용한 통계분석으로 변수들의 유의미 여부를 보고서로 작성합니다.

다시 파이썬의 Matplotlib과 Seaborn으로 시각화를 진행하였고, 다양한 회귀모델로 머신러닝을 진행하여 성능이 좋은 모델을 찾아 성능을 개선시킵니다.

(8-9페이지) 분석과정과 결과를 설명하겠습니다.

식당 데이터 중 음식메뉴는 분석과 머신러닝에 필요한 데이터만 추출했습니다. 조식메뉴의 경우 조식 인원에 대한 데이터가 없어서 삭제했습니다. 정원은 실제 회사 내에서 근무 중인 인원을 구했습니다. 저희는 날씨와 계절이 식수 인원에 영향을 줄 것으로 판단했고, 기상청에서 데이터를 수집하여 일자를 기준으로 결합하고자 했습니다.

(10페이지) 날씨 데이터입니다.

회사가 위치한 지역의 날씨 데이터에서 불필요한 칼럼은 제거했습니다. 기온과 풍속, 습도를 이용해 불쾌지수와, 체감온도 칼럼을 추가했습니다. 또한 비가 오지 않은 날의 강수량 null값은 0으로 바꿔 처리했습니다.

(11페이지)

가져온 두 데이터의 변수입니다. 기준시점 2016년 2월부터 2021년 1월까지 L사의 근무자 수, 식단 메뉴 및 식수 인원, 그리고 기준시점은 2016년 2월 1일부터 2021년 1월 26일까지의 요일과 날짜 약 5년간 데이터입니다. 그리고 기상청에서 해당날짜에 해당하는 평균기온, 강수량, 풍속과 습도 데이터가 있습니다.

(12페이지)

기존 데이터 변수를 통해 파생변수를 만들었습니다. 일자를 세분화하여 계절, 연도, 월, 일을 추가하였고 인원변수를 통해 실제 사내 근무자 수를 변수로 추가하였습니다. 메뉴는 그 가운데 밥, 국, 메인 메뉴로 나눴습니다.

날씨 변수는 기상청 데이터를 활용하여 불쾌지수와 체감온도 변수를 추가하였습니다. 불쾌지수는 기온과 습도의 조합으로 더위를 표현한 수치이고, 체감온도는 기온과 풍속의 조합으로 추위를 표현한 수치입니다. 그 외에도 요일과 연휴 전날, 연말 여부를 변수로 사용했습니다.

(13페이지)

일별 중식과 석식 인원수의 변화 추이를 시각화했습니다. 매년 비슷한 증감 패턴을 보이며 석식의 경우 사내 제도로 있는 ‘자기 계발의 날’로 인해 석식 인원이 0명인 날이 존재함을 확인했습니다.

다음은 재택근무 인원 변화 추이입니다. 코로나 이후 재택근무자 수가 늘어났으나 재택근무자 수가 늘어남에 따라 식수 인원 증감 폭이 크지 않았습니다.

(14페이지)

SQL 쿼리문을 통해 각 변수들과 식수인원의 관계를 파악해봤습니다.

식단 데이터에서는 밥과 특식, 국과 찌개, 면의 유무, 신메뉴 여부가 중식 인원에 영향이 있는 편이었지만 석식 인원에는 영향이 적었습니다. 한편, 이용인원이 국 종류를 선호하는 것으로 유추했습니다.

날씨데이터에서는 기온에 따라 구내식당 이용 인원이 늘어나는 경향을 보이지는 않았습니다. 비가 오는 날은 중식 이용 인원이 늘어남을 확인했습니다. 단, 비가 오는 날에도 석식 이용 인원은 변화가 없는 편이었습니다.

요일의 경우 주초에서 주말로 갈수록 이용인원이 감소했습니다. 또한 연휴 전날 또는 연말이 되면 식수 인원이 줄어드는 경향을 보였습니다.

(15페이지) 다음은 머신러닝 과정을 소개합니다.

계절, 요일, 특식여부, 신메뉴여부, 휴일직전의 내용을 담은 데이터 칼럼을 컴퓨터가 학습할 수 있도록 판다스의 겟더미스를 통해 원핫인코딩 하여 데이터프레임을 가공했습니다. 연도, 월, 일의 정보는 정량적 의미를 가지지 않도록 카테고리화하여 머신러닝용 데이터프레임을 완성했습니다.

(16페이지)

다음으로 그리드서치를 이용해 검증 점수를 최대화할 수 있는 머신러닝 모델의 최적의 파라미터 값을 찾고, kfold 교차검증을 이용해 오차가 적은 모델을 만들려고 하였습니다.

(17페이지)

먼저 하이퍼 파라미터를 조절할 필요없는LinearRegression모델을 이용했습니다. 중식 데이터에 대해선 훈련 정확도가 80%, 점검 정확도가 81%, 석식데이터에 대해선 훈련 정확도가 59%, 점검 정확도가 49%였습니다. 중식에 비해 석식 데이터의 예측 설명력이 많이 부족하다는 걸 알 수 있었습니다.

(18-20페이지)

선형회귀를 보완하며 알파 값을 조절할 수 있는 릿지와 라쏘, 엘라스틱넷 회귀모델도 이용했습니다. 모델 별로 중식의 훈련 정확도는 중식데이터가 80% 내외, 테스트 정확도는 81~82% 대 사이로 소폭 성능이 개선되었습니다. 석식의 훈련 정확도도 60% 미만, 테스트 정확도도 훈련 정확도보다 10%p 아래로 계속해서 나타났습니다.

(21-22 페이지)

많은 하이퍼 파라미터 조절이 가능한 XGBRegressor 모델과 LightGBMRegressor 모델도 활용했습니다. 그리드서치로 찾은 최적의 파라미터를 대입하여 중식 데이터에서는 훈련 정확도를 90%, 테스트 정확도를 86%까지 올릴 수 있었습니다. 석식 데이터에서는 훈련 정확도 79%, 테스트 정확도는 60%에 가깝게 올렸습니다. 여전히 부족한 감이 있지만 이전 모델에 비교하면 엄청난 성능개선이 되었다고 생각합니다.

(23페이지)

머신러닝 결과, XGB가 LGBM보다 정확도 측면에서 약간 더 높은 점수를 나타냈습니다. 그 이유는 MSE 지표가 좀 더 낮기 때문인 것 같습니다. 이러한 트리 구조의 알고리즘이 선형회귀 알고리즘보다 뛰어난 성능을 나타내는 경향을 통해 저희는 딥러닝을 적용하면 또 좋은 결과를 낼 수 있지 않을까 추측했습니다.

**24 페이지**

다음으로 저희가 현재 만들고 있는 모듈에 대해 설명드리겠습니다.

**25페이지**

저희가 모듈을 만든 이유는 크게 두 가지입니다. 하나는 식수 인원 예측 모델을 프로그램으로 만들기 위해서이고, 다른 하나는 데이터가 업데이트됐을 때 매번 코드를 짜는 것보다 모듈로 간단히 분석하기 위해서입니다.

개발언어는 파이썬 3.9.7 버전이고 현재 DB 관련 소스를 담은 sql.py, 데이터 분석에 용이하게 데이터프레임 형태를 변환해주는 [train\_encoding.py와 forecast\_encoding] // XGBRegressor 훈련을 자동으로 돌리고 결과 및 시각화를 제공, 저장해주는 learnining.py 로 구성되어있습니다.

**26페이지**

모듈을 통한 저희가 구상하고 있는 프로세스는 이미지와 같습니다. 식수인원 CSV 파일과 날씨 CSV 파일을 입력하면 저희가 지금까지 한 모든 포트폴리오 작업, SQL로 DB 생성 및 데이터 저장, 데이터프레임 전처리, 인코딩, 그리고 분석이 자동으로 이뤄지는 것이죠. Sql.py는 db 생성 과정에서, 머신러닝을 통한 데이터 분석은 train\_encoding.py, forecast\_encoding.py, learnining.py에서 작동하게 할 예정입니다.

**27페이지**

현재 모듈 내 필요한 클래스, 함수는 약 85% 완성했으며 테스트 및 기타 기능 추가 부분을 담겨두고 있습니다. 모듈 구성을 보면 클래스 내 함수를 호출해 사용하는 방식이며 init을 설정하여 class 호출 때 데이터를 입력하면 클래스 내 함수에서 원하는 결과물을 바로 산출할 수 있도록 만들었습니다. 다만, 함수를 기능에 따라 여러 개 나눴는데 오류가 발생했을 때 빠르게 수정하기 위해, 즉 유지보수 측면의 용이성을 위해 이렇게 만들었습니다.

**28페이지**

다음은 저희가 프로젝트를 진행하며 느꼈던 아쉬운 점과 프로젝트 후 디벨롭해야할 점에 대해 말씀 드리겠습니다.

**29페이지**

먼저, 미완성된 모듈을 완성하는 것입니다. LSTM을 통한 딥러닝 분석모델, Pickle로 분석모델을 저장하는 기능을 추가로 만들고 모듈 test 및 오류 수정을 거쳐 모듈을 완료해야 하며 올 2월 안에 해당 작업 완료를 목표로 하고 있습니다.

(30페이지) 두번째입니다.

수요일이 자기 계발의 날이라 석식 인원이 없었음에도 이에 대한 데이터도 같이 훈련되었습니다. 그런데 석식 인원 데이터 예측에서 0에 가깝게 떨어지는 값, 이상치를 전혀 도출하지 못하고 있습니다. 그렇다면, 차라리 훈련 정확도에 지장을 주지 않게 0인 날의 데이터를 제거 후 다시 훈련 시켜볼 계획입니다.

(31페이지)

다른 해결 방안으로 EDA 도중 파악한 중식 인원 수와 석식 인원 수의 상관관계를 근거로 석식 예측에 중식 인원수를 넣어 훈련 시키는 방안입니다. 다만, 이를 위해선 매일 중식 인원 수가 바로 업데이트 되고 그 값을 포함해서 머신러닝을 돌려야하는 부담이 있습니다.

(32페이지) 식단에 관한 세번째 미션입니다.

메뉴가 너무 많고 카테고리화를 하기 어려워 머신러닝에서는 쌀밥과 특식으로 나눈 후 나머지 데이터는 다 제거했습니다. (밥은 1200여개의 데이터 중 대부분을 차지함)

국립농업과학원에서 농식품 분류 API(카테고리, 이미지)를 제공하고 있습니다. API데이터를 CNN으로 카테고리별로 분류한 다음 원핫인코딩을 하고 메뉴 데이터가 식수 인원에 영향을 주는지 파악하고 모델링을 하는 방법으로 보완할 계획입니다.

(33페이지)

또, 과제 논의 중 회사 근처 주변 음식점이나 먹자골목 데이터 등을 추가해 분석을 진행하면 좋을 것 같다는 의견이 있었습니다. 이는 비용이 많이 들지만, 이를 통해 정확한 분석이 가능하다면 추가 비용을 받고 사옥 주변 상권의 상황을 추가 입력해 예측하는 서비스를 제공하는 방향으로 진행 가능하다 판단 했습니다.

(34페이지) 마지막입니다.

날씨 데이터 중 기온, 풍속, 습도를 일일평균 데이터로 활용했습니다.

보다 정확한 방법으로 훈련시키기 위해 중식은 12시, 석식은 18시 데이터를 적용해서 다시 훈련시키는 방법을 고려하고 있습니다. 다만, 이 경우, 시간별 데이터를 확보하기가 어려워 결측치가 발생할 수도 있고, 모듈을 완전히 재수정해야 하는 단점이 있습니다.