안녕하세요, 지금부터 5공대 발표 시작하겠습니다.

저희 프로젝트 목표는 시계열 데이터를 통해 전력사용량을 예측하는 것입니다. 따라서 시계열 데이터란 무엇인지 먼저 설명한 뒤, 데이터 소개를 하고자 합니다. 그 다음 EDA, 모델링, 평가 및 발전 방향 순서로 진행될 예정입니다.

그럼, 시계열 데이터에 대해 알아봅시다. 시계열 데이터란 무엇일까요? 일정한 시간 동안 수집된 일련의 순차적으로 정해진 데이터 셋의 집합을 말합니다. 데이터에서 법칙성을 발견하고, 이를 모형화, 추정된 모형을 통하여 미래의 값을 예측할 수 있습니다. 시계열 데이터는 보시는 바와 같이 추세변동, 계절변동, 순환변동, 불규칙변동의 특성을 가지고 있습니다. 불규칙변동을 제외하면 장기간에 주기적이거나 반복적으로 나타나는 특징을 가지고 있음을 확인할 수 있습니다.

이제, 이와 같은 특징을 가진 데이터를 소개하고자 합니다. 데이터는 122400개의 행과 10개의 열로 이루어져 있습니다. 처음 말씀 드렸다시피, 데이터 열 중 하나인 전력사용량을 예측하는 것이 저희 팀의 목표입니다. 근데 왜 전력사용량을 예측해야 할까요? 전력사용량을 예측하는 이유는 다음과 같습니다. 공급의 안정화 및 건물 운영 예산을 측정하는 데 도움이 되기 때문입니다. 그렇기에 전력사용량은 종속 변수가 되고, 다른 9가지의 변수들이 독립 변수가 됩니다. 그럼 어떤 독립변수가 있을까요? 독립변수에는 우선 건물 번호와 날짜와 시간의 정보를 가진 변수가 있습니다. 뿐만 아니라 기온, 풍속, 습도, 일조, 강수량, 비전기냉방설비운영, 태양광보유를 나타내는 변수도 존재합니다. 여기서 조금은 낯선 변수명도 있는데요, 바로 비전기냉방설비 운영입니다. 우리나라는 2011년부터 일정 규모 이상의 공공기관과 민간 건축물에 대해 비전기식 냉방 설치가 의무화되었습니다. 이는 주로 여름철 전력 피크 수요를 완화하고, 전력 의존도를 줄여 에너지 효율을 높이기 위해 시행하게 되었습니다. 마지막으로, 종속 변수인 전력사용량입니다. 전력사용량은 kWh(킬로와트시)라는 단위를 쓰는데요, 1kWh는 1시간 동안 1kW의 전력을 소비했을 때 사용되는 전력을 뜻합니다.

변수를 알았으니 그 다음엔 무엇을 해야 할까요? 변수 간의 관계를 알아볼 차례입니다. 오른쪽 상관 행렬 표를 보시면, 기후 변수들과 전력량 사이의 상관관계가 대체로 낮은 편임을 파악할 수 있습니다. 그래서 저희는 전력 사용량과 다른 독립변수의 Feature Importance를 확인해보기로 했습니다. 그 결과, 건물 번호가 가장 중요한 변수임을 알 수 있었습니다. 이렇게 건물 번호가 가장 중요한 변수임을 알았으니, 건물마다 어떠한 특징이 있는지 확인하기 위해 평일 평균 그래프를 시각화 해 보았습니다. 예를 들어, 6번과 14번처럼 비슷한 양상을 보이는 그래프가 눈에 들어옵니다. 따라서, 저희는 수동으로 그룹화를 진행하여 비교 및 분석을 진행하기로 결정하였습니다. 그 결과, 그룹은 총 5개로 나뉘게 됩니다. 그룹 1,2,3은 오전 6시부터 오후 6시까지 전력 사용이 집중되는 그래프입니다. 여기서 그룹 2는 오후 12시에 잠깐 전력 사용량이 줄어들고, 그룹 3는 새벽 시간에 전력 사용량이 다시 증가하는 특징을 가지고 있습니다. 그룹 4는 오전 12시부터 오전 6시까지는 감소, 오전 6시부터 오후 6시까지는 유지되고, 오후 6시 이후에는 증가했다가 다시 감소하는 그래프의 형태를 가지고 있습니다. 마지막으로 그룹 5는 분류하기 어려운, 특징을 가지지 못한 패턴을 가진 그래프로 이루어져 있습니다. 이렇게 건물의 평일 평균 전력사용량을 시각화 및 군집화를 해보았는데, 과연 이 건물들은 어떤 종류의 건물일까요? 이 건물이 어떤 건물인지 알아내기 위해, 75개 산업군의 시간대별 전력소비계수를 시각화해 비교해보기로 했습니다. 그룹 1은 도매업, 소매업, 숙박업, 영사 기록물 제작 및 배급업, 통신업, 금융 및 보험 관련 서비스업, 부동산업, 연구개발업의 전력사용량과 비슷한 패턴을 가지고 있습니다. 그룹 2는 어떨까요? 농업, 임업, 광업, 제조업의 전력사용량과 비슷한 패턴을 찾을 수 있었습니다. 이러한 산업군에서 작업 및 기계 가동이 점심 시간 동안 일시 중단되기 때문에 12시에 전력사용량이 잠시 하락이 됩니다. 마찬가지로, 그룹 3는 방송업, 정보 서비스업, 사회복지 서비스업의 전력사용량과 비슷한 패턴을 가지고 있습니다. 여기서 정보 서비스업의 종류에는 IT 서비스 및 소프트웨어 개발, 인터넷 서비스, 클라우드 및 데이터 센터 서비스, 통신 서비스, 디지털 콘텐츠 제공 등이 있습니다. 네 번째로, 그룹 4는 주택용 건물임을 확인할 수 있었습니다. 주택용 건물이기에 오전 12시부터 6시까지는 감소, 오전 6시부터 오후 6시까지는 유지하고, 오후 6시에 증가하는 특징이 나타나게 되는 것입니다. 그룹 5는 그룹 1~4의 그래프 형태와 유사성이 보이지 않아 산업군을 찾아내기 어려운 부분이 있었습니다. 이제 분류도 마치게 되었으니 그룹별 특징을 시각화해보겠습니다. 제조업, 광업, 정보 서비스업에서 전력 사용량이 가장 높고, 그룹2와 그룹3에 비해 비교적 주택용 건물이 전력사용량이 낮은 편임을 알 수 있었습니다. 이를 제외한 기온, 풍속, 일조, 습도, 강수량은 모두 시간대별 비슷한 추이를 보여 상관 행렬에서 확인했던 것처럼 상관 관계가 약한 편임을 확인이 가능합니다. 비전기냉방설비운영 보유 비율은 그룹 1이 83%로 가장 많고, 태양광 보유는 그룹 3이 71%로 가장 높다는 것을 확인할 수 있습니다.

지금까지 목표를 설정하고, 데이터를 파악했으니 드디어 데이터를 설계해볼 차례가 되었습니다. 저희 팀은 시계열 모델 3개와 머신러닝 모델 3개를 각각 사용해보기로 했습니다. 또한, 모델에 넣을 데이터 방식도 두 가지로 나눴는데요. 첫 번째로는 전체 데이터를, 두 번째는 60개 건물 별로 각각 모델을 돌리기로 하였습니다. 첫 번째로, 시계열 모델입니다. 저희 팀이 시도한 시계열 모델은 ARIMA, LSTM, PROPHET 총 3가지입니다. 먼저, ARIMA는 자기회귀, 차분, 이동 평균의 조합으로 구성된 모델로, 추세가 명확한 시계열 데이터에 적합합니다. 만약 데이터가 비정상일 경우 차분을 통해 정상화가 가능합니다. LSTM은 장기 의존성을 잘 학습하는 RNN, 비선형 관계를 모델링하며 다양한 입력 길이에 대응할 수 있는 시계열 데이터 예측에 강력한 성능을 보입니다. 마지막으로 PROPHET은 수식에 보이는 것처럼 계절성과 추세 변화를 자동으로 감지하기 때문에 계절적 패턴 예측이 뛰어납니다. 두 번째로, 머신러닝 모델을 들어가기 전에 Bagging과 Boosting에 대해 간단하게 설명하고자 합니다. Bagging은 독립된 모델들의 결과를 합산하여 다수결 투표(Voting)을 통해 결과를 산출하는 방식이고, Boosting은 특정 모델의 결과를 다른 모델의 입력값으로 사용하는 방식으로, 모델 간 가중치를 부여하여 결과 산출합니다. 그 중 저희가 선택한 방식은 Boosting이고, Boosting 안에서도 Gradient Boosting을 골랐습니다. Gradient Boosting은 부스팅 기법을 이용한 앙상블 모델 중 하나로, 잔차(실제값과 예측값의 차이)를 이용하여 이전 모형의 약점을 보완하는 새로운 모형을 순차적으로 적합한 뒤 이들을 선형 결합하여 얻어진 모형을 생성하는 지도 학습 알고리즘입니다. 저희 팀이 시도한 머신러닝 모델은 Gradient Boosting에 포함된 XGBoost, LightGBM, CatBoost 총 3가지입니다. XGBoost는 레벨별 확장 방식으로 강력한 성능과 안정성을 제공합니다. 하지만 대용량 데이터에서는 느리고, 메모리 사용이 많기 때문에 LightGBM을 사용했습니다. LightGBM은 리프별 확장 방식으로 빠르고 메모리를 효율적으로 사용할 수 있습니다. 하지만 과적합 위험이 있기에 유의해야 합니다. 마지막으로, CatBoost입니다. 범주형 데이터를 직접 처리하며, XGBoost와 마찬가지로 데이터가 커지면 학습 속도가 느릴 수 있습니다.