〈Script〉

안녕하세요, 10조 Synergy의 김형림입니다. 10조의 데이터 시각화 프로젝트, 〈기후 데이터를 활용한 모기 발생 예보〉에 대한 발표를 시작하겠습니다.

(next - 02)

1. 프로젝트 개요

먼저 연구 배경을 비롯한 프로젝트 개요에 대해 말씀드리겠습니다.

(next - 03)

저희 조는 다양한 전공 및 관심을 가진 3명의 인원이 머리를 맞대고 시너지를 내어, 여름철 지긋지긋한 '모기'에 대한 분석을 진행했습니다. 팀원들 간의 역할은 다음과 같이 나누어 맡았습니다.

(next - 04)

최근 기후 변화로 인한 영향을 주제로 다양한 연구가 진행되고 있습니다. 오존층 파괴, 해수면 상승 등 큼직큼직한 연구도 있지만, 저희가 주목한 것은 바로 기후 변화가 모기 발생에 미치는 영향입니다. 보시는 바와 같이 선행연구에 의하면 기온이 상승함에 따라 모기 성체의 발생량이 증가하고, 상대습도와 강수량 또한 기온과 마찬가지로 양의 상관관계를 가집니다. 모기 발생량이 증가하는 것은 ~~단순히 강가나 공원 등에서 모기에 많이 물리기 때문이 아니라,~~ 이로 인해 최근 들어 말라리아, 일본뇌염, 지카바이러스 등의 모기매개 감염병 환자의 수가 늘어나고 있기 때문에 가볍게 볼 문제가 아닙니다. 게다가 감염병의 종류 또한 해를 거듭할수록 늘어나고 있습니다. 실제로 질병관리청에서는 몇 년 전부터 각 시도 보건환경연구원으로부터 전달 받은 유문등, 그리고 디지털 모기 측정기(DMS)를 통해 채집한 모기 데이터로 모기매개 질병에 대해 연구하고 있습니다.

(next - 05)

모기는 알에서 부화해서 1~2주 정도 후에 유충에서 성충으로 성장하는데요, 이 부분이 저희 Synergy의 분석에서 핵심이라고 할 수 있습니다. 이 내용과 관련해서는 뒤에서 자세히 살펴보겠습니다. / 앞서 말씀드린 대로 모기채집 장치에는 유문등과 DMS가 있는데요, 유문등 데이터는 모기의 종류별로 포집량을 보여주지만, 그 수가 매우 적고 주 단위로 집계하기 때문에 매일매일 변하는 기후 데이터를 적용시키기에는 한계가 있습니다. 따라서 저희는 자동으로 일일 모기 포집량을 집계하는 DMS 데이터를 분석에 사용하기로 했습니다. / 서울특별시에서는 전국에서 거의 유일하게 모기예보제를 실시하고 있는데요, 매일매일 수변부, 공원, 주거지의 모기 활동지수를 계산해서 홈페이지를 통해 공개하고 있습니다. 이 모기 활동지수를 정확히 어떻게 계산하는지는 관련 논문이 현재 비공개 상태라 확인이 어렵기 때문에, 저희 나름의 모기지수를 만들어보기로 했습니다.

(next - 06)

분석에 앞서 다음과 같이 모기 성장 기간을 반영한 기후 데이터와 모기 개체수 증가량 사이에 상관관계가 있는지 여부로 귀무가설과 대립가설을 설정했습니다. 그리고 오른쪽 화면에서 (기후 데이터가 정규성을 가지고, 독립적인 집단을 비교하기 때문에 독립비교) t 검정을 수행해보니, 모든 변수에 대해서 p-value가 유의수준 0.05보다 작게 나오기 때문에 귀무가설을 기각하고, '모기 성장 기간을 반영한 기후 데이터와 모기 개체수 사이의 상관관계'에 대한 분석을 시작했습니다.

(next - 07)

2. 프로세싱

지금부터는 분석 과정에 대해서 말씀드리겠습니다.

(next - 08)

분석에 사용한 데이터는 세 가지입니다. '서울특별시 DMS 설치 현황', '서울특별시 DMS 포집내역', 그리고 '서울특별시 기후' 데이터입니다. DMS 설치 현황은 2021년 자료로, 2020년과 동일하고, DMS 포집내역과 기후 데이터는 2015년~2020년 사이에 일자별로 구축되어 있습니다. DMS 포집내역은 구 단위가 아닌 서울 전체의 합계입니다. 세밀한 분석을 위해 관계부처에 구별 DMS 포집 데이터를 요청했지만, 제공을 거부했습니다. / 아래에 기후 데이터에서 분석에 사용한 변수들에 대한 설명을 적어놓았는데요, '온도'와 관련된 변수들이 많음을 확인할 수 있습니다. 그리고 accum_rain은 누적강수량으로, 일강수량(rain_per_day)로부터 누적 합산해서 새롭게 도출한 변수입니다.

(next - 09)

분석에 사용한 데이터로 이렇게 데이터 마트를 만들었습니다.

(next - 10)

먼저 서울의 DMS 설치 위치를 보시면, 25개 구별로 2개씩, 총 50개의 DMS를 운영 중이고, 지리타입을 '공원, 수변부, 주거지' 3개로 나누어서 설치 구역이 정해졌습니다. 이처럼 DMS가 서울 전역에 골고루 설치되어 있기 때문에 서울 전역의 모기 발생량을 분석하기에 적합하다고 판단했습니다.

(next - 11)

분석 목표는 모기 개체수를 목표변수로 해서 기온, 일강수량, 누적 강수량, 풍속, 습도, 일조량을 통해 목표변수를 예측하고, 저희 나름의 모기지수를 개발하는 것입니다. 이때, 모기 유충의 성장 기간을 반영하여 기후 데이터를 변형했는데요,

(next - 12)

다음과 같이 당일을 포함한 일주일 데이터의 평균값을 도출해서 새로운 분석 데이터를 구축했습니다.

(next - 13)

앞서 확인했듯이 기후 데이터에는 '온도' 관련 변수들이 많습니다. 분석에 들어가기 전에 다중공선성이 우려되어 다음과 같이, 데이터가 모두 numerical 변수들로 구성되어 있기 때문에, Pearson 상관계수의 Heatmap을 찍어보았더니, 역시 평균 기온, 최저기온, 최고기온, 이슬점온도 등 온도 관련 변수들 간의 높은 상관관계를 발견했습니다. 화면 오른쪽 그림과 같이 Pair Plot에서는 이 변수들 간의 높은 양의 상관관계, 즉 우상향하는 선형관계를 확인할 수 있습니다.

(next - 14)

좀 더 객관적인 수치로 확인하기 위해서 VIF(분산팽창요인)도 확인해보았습니다. VIF는 독립변수들 간의 상관관계를 보여줍니다. 온도 관련 변수들로 분석한 회귀 결과를 보면, 모기 개체수와 양의 상관관계가 있을 가능성이 높은 temp(평균 기온)의 coef, 즉 회귀계수가 -926.0829로 음의 값을 가집니다.

이를 보아 변수들 간의 높은 상관관계가 정확한 예측을 방해했음을 알 수 있습니다. / 보통 VIF가 10이 넘으면 다중공선성이 있다고 판단하는데, 온도 관련 변수들은 모두 10이 넘는, 심지어는 100보다도 큰 값이 나오기도 합니다. / temp(평균 기온)을 제외한 나머지 변수들을 drop 하고 나니, 오른쪽과 같이 temp의 회귀계수도 양의 값으로 제대로 나오는 것을 볼 수 있습니다. 따라서 저희는 온도 관련 변수는 temp만 분석에 사용하기로 결정했습니다.

(next - 15)
이번에는 DMS 모기 포집내역 데이터를 살펴보겠습니다. Strip Plot과 Box Plot을 그려보니 2015년~2016년에 몇몇 이상치가 발생하는 것을 확인할 수 있었습니다. 2017년 이후로는 보건환경연구원에서 데이터를 공개하기 전에 이상치를 제거했을 것으로 추정됩니다. 하지만 이상치도 실제로 측정된 값이기 때문에 저희는 모기 포집량의 이상치를 제거하지는 않기로 결정했습니다. 또한 그림에서 보이듯이 여름철에 5,000~7,000마리 정도의 모기가 집계된 날이 가장 많음을 확인할 수 있었습니다.

(next - 16)
DMS 포집내역에서 모기를 제외한 기타 벌레 포집량은 분석에서 제외했습니다. 그리고 기후 데이터 중에서 일강수량(rain_per_day)은 기상청 문의 결과 결측치는 비가 전혀 오지 않았음을 의미하고, 0은 비가 조금이라고 온 날이라는 답변을 얻었습니다. 하지만 0.1mm 이하는 비가 전혀 오지 않은 것과 같다고 가정해서 0으로 변환하고, 나머지 결측치들도 이어서 0으로 변환했습니다. 또 반복해서 말씀드리지만, 분석의 핵심인 모기 유충의 성장 기간을 고려하여 기후 데이터의 7일 평균값을 계산했습니다. 결측치는 계절별 특성을 고려해서 결측치의 앞뒤 3일 평균값을 적용했습니다. 그리고 아래와 같이 최종적으로 분석에 사용한 데이터를 구축했습니다.

(next - 17)
전처리 전후의 히트맵을 비교해보겠습니다. 왼쪽 그림에서 상관관계가 높은 변수들을 drop 해서 오른쪽 히트맵과 같이 다중공선성을 제거했습니다.

~~(next)~~
~~목표변수인 mosquito(모기 개체수)와 남은 입력변수들 간의 Pair Plot은 다음과 같이 나타납니다.~~
~~음.....~~

(next - 18)
~~분석에 사용한 모델은 그리드 서치를 통해 최적의 하이퍼 파라미터 조합을 찾아서 모델에 적용하고, 각 모델별로 loss, 즉 RMSE(Route Mean Square Error) 값을 비교했습니다.~~ 분석을 진행할 때는, 모기 데이터와 기후 데이터 모두 연속형 변수로 구성되어있기 때문에 트리 기반의 Regressor 모델을 사용했고, LGBM, GBR, 그리고 랜덤 포레스트 모델에 적용한 결과 RMSE 값이 가장 작은 LGBM을 저희의 분석 모델로 채택했습니다.

(next - 19)
이 과정에서 분석 데이터 구축 방법이 적절했는지 검증해보았습니다. 선행연구를 참고해서 모기 유충의 성장 기간을 반영할 때 처음에는 기후 데이터를 단순히 7일씩 앞당긴 경우, 오히려 RMSE 값이 증가했습니다. 이는 모기 유충의 성장 기간이 제대로 반영되지 않았음을 반증하는 결과입니다. 이에 비해

오른쪽과 같이 누적된 기후 데이터의 영향을 반영해서 평균값을 적용하니 더 작은 RMSE 값, 즉 보다 정확한 예측력을 가진 모델을 만들 수 있었습니다. 이를 통해 저희가 생각한 데이터 구축 방법이 적절했음을 확인할 수 있었습니다.

(next - 20)
각 변수의 중요도를 살펴보니, temp(평균 기온)의 중요성이 크게 나타났습니다. 다른 모델들로 Feature Importance를 확인했을 때도 동일한 결과를 얻을 수 있었고, 다른 변수들에 비해 평균 기온이 모기 발생량에 매우 큰 영향을 미친다는 결론을 얻었습니다. 실제로 선행연구 중에서는 일본에서 기온만으로도 모기 발생량을 예측한 사례가 있었습니다. / 이에 비해 새롭게 도출한 누적 강수량의 경우에는 영향력이 미미한 수준임을 확인할 수 있었습니다...ㅠㅠ

(next - 21)
지금 보시는 게 Test Set을 예측한 결과입니다. 파란색 선이 실제 모기 발생량으로, 머신러닝으로 예측한 결과가 실제값과 비슷한 추이를 보입니다.

(next - 22)
좀 더 명확하게 확인하기 위해서 LGBM을 사용한 예측 결과만을 시간순으로 그려본 결과는 다음과 같습니다. 완전히 동일하지는 않지만, 역시 실제값과 비슷한 추이를 보이며, 계절에 따라 일정한 주기성을 가지는 것을 확인할 수 있었습니다.

(next - 23)
분석 결과를 토대로 최종 목표인 지수도 개발해보았습니다. 선행연구의 구별 모기 활동지수 적용 기준을 참고해서 서울 전체에 적용하기 위해 25씩, 즉 서울시 구의 개수만큼 곱해주었습니다. 또 기존 10단계를 두 단계씩 묶어서 5단계로 만들어 오른쪽과 같이 '쾌적, 관심, 주의, 불쾌, 심각'으로 새로운 지수를 만들었습니다.

(next - 24)
이렇게 만든 지수를 Folium으로 지도에 시각화한 결과입니다. 보다 직관적으로 위험도를 판별할 수 있게 원색으로 표현했습니다. 이와 같은 지도 시각화를 모기예보제에 적용하면 좋을 것 같습니다.

(next - 25)
3. 기대효과 및 제한점
마지막으로 본 분석 결과의 기대효과와 한계점에 대해 말씀드리겠습니다.

(next - 26)
이러한 분석을 통해 좀 더 정확한 예측 모델이 구현된다면, 질병관리청에서 진행하는 모기매개 감염병 연구와 서울시 모기예보제의 발전에 도움이 될 것으로 기대합니다. 모기 발생량을 예측해서 선제적인 방역으로 모기매개 감염병 환자의 감소와, 방역 예산 감소에 긍정적인 효과를 불러올 수 있습니다.

(next - 27)
하지만 아쉬운 점은 생각보다 기후 데이터 외에 모기 발생에 영향을 미치는 요인이 많았다는 겁니다.

저희가 가진 데이터는 서울과 인천의 DMS 데이터인데요, 서울 데이터로 학습을 하고 인천에 적용했을 때는 썩 마음에 드는 결과를 얻을 수 없었습니다. 왼쪽 그림과 같이 해안가에 모기가 많이 발생하는 등, 지리적인 요인을 모델에 반영하지 못한 탓이라고 판단했는데요, 이는 DMS 데이터가 시도별이 아니라 구/군별로 구축은 되어있으나 관계부처에서 제공을 거부하고 있어서, 만약에 이러한 데이터를 추후에 얻을 수 있다면, 각 DMS별로 고도 등의 지리적 데이터를 추가해서 좀 더 정확한 예측력을 얻을 수 있을 것으로 기대하고 있습니다. / 그래서 당초 목표였던 모기지수의 전국 단위 적용은 어려워졌지만, 좀 전에 말씀드린 작은 단위의 DMS 데이터를 얻을 수 있다면, 보시는 그림과 같이 녹지 비율 등의 데이터를 추가해서 지역별로 비교 가능한 모델을 구축할 수 있을 거라고 생각합니다.

(next - 28)

4. 느낀점

네, 마지막으로 팀원들의 프로젝트 소감입니다.

(next - 29)

이번 프로젝트를 통해서 팀원들 모두 많은 경험을 해볼 수 있었고, Synergy라는 조 이름처럼 서로서로 엄청난 시너지 효과가 일어나서 많은 것을 얻어가는 것 같습니다.

(next - 30)

(next - 31)

네, 이상으로 10조 Synergy의 데이터 시각화 프로젝트 발표를 마치겠습니다. 감사합니다.