기후 데이터를 활용한 모기 발생 예보

1. 연구 배경 및 필요성

현재는 서울특별시와 인천광역시에서만 원격모기감시장비(Digital Mosquito Monitoring System, DMS)를 운영 중이지만, 2021년에는 보다 신속한 매개모기 감시결과를 획득하기 위하여 DMS를 통해 위험지역의 일일모기감시체계를 활용하여 매개체감시를 수행할 계획

2. 데이터

2-1. Data Collection

① 서울특별시 DMS 설치 현황
- 기간 : 2021년
- 출처 : 서울특별시 보건환경연구원
- DMS 지역명 / 세부 주소 / 지리 타입 / 위도 / 경도
- 구별로 2개씩 총 50개의 DMS 설치
- 지리 타입은 3개(공원, 수변부, 주거지)로 분류
- DMS 설치 현황 시각화 자료를 통해 구별, 지리 타입별로 골고루 설치되어 있어 서울 전역의 모기 발생량을 분석하기에 적합하다고 판단했다고 설명

② 서울특별시 DMS 포집내역
- 기간 : 2015년 ~ 2020년, 대부분 04월 초 ~ 11월 초
- 출처 : 서울특별시 보건환경연구원
- 일자별 total(전체 벌레 포집량) / mosquito(모기 포집량) / etc(모기 외 벌레 포집량)
- 분석에는 mosquito만 사용
- 1년 내내 DMS를 운영하지는 않기 때문에 겨울 데이터가 제외되어 추운 날씨에 모기 발생량을 알 수 없다는 한계

③ 서울특별시 기후
- 기간 : 2015.01 ~ 2021.06
- 출처 : 기상청 (기상자료개방포털) 종상기상관측(ASOS)
- 일자별 loc_num(지점) / loc_name(지점명) / temp(평균기온(°C)) / l_temp(최저기온(°C)) / h_temp(최고기온(°C)) / rain_per_day(일강수량(mm)) / accum_rain(누적 강수량(mm))(★파생변수) / wind(평균 풍속(m/s)) / dew(평균 이슬점온도(°C)) / humidity(평균 상대습도(%)) steam_pressure(평균 증기압(hPa)) / sunshine_time(합계 일조시간(hr)) / sunshine(합계 일사량(MJ/m2)) / ground_surface_temp(평균 지면온도(°C))

④ 인천광역시 DMS 계측정보
- 기간 : 2018년 4월 ~ 10월
- 출처 : 인천광역시 보건환경연구원
- 서울특별시 DMS 포집내역과 같은 내용으로, 이에 더해 군/구별 데이터까지 확인 가능

⑤ 서울특별시 DMS 포집내역
- 기간 : 2021년 05월 ~ 06월
- 출처 : 서울특별시 보건환경연구원
- 학습한 데이터를 test 하는 데 활용

2-2. Data Preprocessing

① 서울특별시 DMS 포집내역
- Boxplot을 통해 2015년 ~ 2017년에 몇몇 이상치가 발생하는 것을 확인
- 2018년 이후에는 데이터를 공개하기 전에 이상치를 제거했을 것으로 추정
- 정확한 분석을 위해서 이상치를 제거하지 않기로 결정

② 서울특별시 기후

②-1. 다중공선성
- Heatmap으로 각 변수 사이의 상관관계를 확인하고, 온도 관련된 변수들의 다중공선성이 우려되어 temp(평균기온(°C))만 분석에 사용하기로 결정

②-2. 분석 데이터 구축
- 각 변수는 모기 유충의 성충으로의 성장 기간을 고려하여 당일을 포함한 일주일 데이터의 평균값을 도출

②-3. 결측치
- 기상청 전화 문의 결과 rain_per_day(일강수량(mm))에서 확인되는 많은 결측치는 비가 전혀 오지 않은 날, 0으로 표시된 부분은 아주 조금이라도 비가 온 날 → 0.1 이하는 0으로 변환하고, 모든 결측치는 SimpleImputer()를 통해 0으로 처리
- 이외 변수들의 결측치는 계절별 날씨 특성을 고려하여 결측치 앞뒤 3일씩의 평균값으로 대체

②-4. 파생변수
- rain_per_day(일강수량(mm))으로부터 파생변수 accum_rain(누적 강수량(mm))을 생성
- 분석 결과 유의미한 영향력은 없는 것으로 판단

3. 분석

3-1. Comparing Models

① 데이터가 모두 연속적이기 때문에 트리 기반 Regressor 사용 결정

② Grid Search CV로 LGBM(Light Gradient Boosting Machine)과 Random Forest Regressor, GBR(Gradient Boosting Regressor)의 최적 하이퍼 파라미터를 확인하고 각각의 모델을 튜닝

③ 세 가지 모델의 RMSE 비교 → RMSE가 가장 작은 LGBM으로 최종 결정

3-2. Comprehensing Anlysis Result

① Test(Validation) Set을 통해 검증한 결과, 실제값과 완벽히 일치하지는 않지만 동일한 추이를 보여주는 그래프 도출

② 참고 문헌 〈기상자료와 GIS 활용 수도권 모기 활동지수 개발〉의 '구별 모기 활동지수 분류 기준'을 서울 전체로 확대 적용하여, '쾌적-관심-주의-불쾌-심각'의 5단계로 구분

③ 모기 활동지수를 지도에 색깔로 시각화

3-3. 보완점

① 모델 튜닝이 완벽하지 않기 때문에 머신러닝 이론에 대한 학습 이후 보완 필요

② 인천광역시의 DMS 데이터를 적용하면 정확하지 않은 값이 나오는데, 이는 고도, 해안가와의 거리, 녹지 비율 등 지리적인 특성을 고려하지 않았기 때문

③ 서울특별시는 서울 전체의 데이터만 공개하기 때문에 구/군별 지리적인 특성을 반영 불가능 → 데이터가 존재할 것으로 보이나 관계 부처에서 제공 거부

4. 기대효과

5. 연구 확장성