Data Mining Team Project

팀명: 왕경율

Team member:
2014147576 박건우
2014147519 손장현
2014147565 WANGGENGYU

Subject

기상 조건에 따른 음식물 쓰레기 배출량 예측

목록

<Data Pre-processing>

1. 데이터 결합
2. Domain knowledge를 통한 데이터 COLUMN 정리
3. Missing Value 처리
4. 카테고리 변수 처리
5. 데이터 시각화
6. Outlier 정리

<Data Mining execution>

1. Data partitioning
2. XGBOOST 실행

<Schedule>

<Data Pre-processing>

1. 데이터 결합

- Input data

  2015년 지자체별 일별 날씨 데이터(csv), 2015년 지자체별 일별 음식물 쓰레기 배출 데이터(csv), 2016년 서울 일별 날씨 데이터(csv), 2016년 서울 일별 음식물 쓰레기 배출 데이터(csv)

- 결합 규칙

  지차체, 일자를 기준으로 데이터를 결합. 지자체는 시, 군 단위로 하였으며 서대문구 등 구 단위는 시 데이터에 합산하였다.
  L'Python pandas'를 이용하여 같은 날짜이며, 같은 지자체인 날씨 데이터와 음식물 쓰레기 배출 데이터를 하나의 데이터 row로 만들었다.

```
result = pd.merge(waste_df, weather_df, how='inner', on=['city', 'year','month','day'])
```

- Output data

  67개의 속성, 10146개의 Record를 가지는 데이터

2. Domain knowledge를 통한 데이터 속성 정리

- 음식물 배출 습관에 영향을 미치지 않을 것이라 예상되는 데이터 속성 삭제

  최대 순간 풍속 풍향(16방위), 평균 중하층운량(1/10) 등 사람의 음식물 배출 습관에 크게 영향을 미치지 않을 것 같은 데이터 속성은 팀의 판단 하에 삭제한다.

- 지역적인 데이터 속성 삭제

  1시간 최다 강수량 시각(hhmi), 1시간 최다일사 시각(hhmi), 평균 5cm 지중온도(℃), 평균 10cm 지중온도(℃), 평균 20cm 지중온도(℃), 평균 30cm 지중온도(℃), 0.5m 지중온도(℃), 1.0m 지중온도(℃), 1.5m 지중온도(℃), 3.0m 지중온도(℃),5.0m 지중온도(℃) 와 같이 일별 기준보다 작은 단위의 지역적 정보를 가지는 데이터 속성은 삭제한다.

- 음식물 배출량과 관련 없는 데이터 속성 삭제

  quantity_rate: 월 기준 해당 날짜 배출 비율을 나타내는 속성으로, 2015년 일별 배출량 예측을 진행하는 경우에는 필요 없기에 제거한다.
  Year, time_rate(월 기준 해당 날짜 배출 횟수 비율) 또한 같은 이유로 더이상 필요없기에 제거한다.

- 최종데이터

  month(10146,) day(10146,) week(10146) quantity(10146), times(10146), city(10146), 평균기온(℃)(10146), 최저기온(℃)(10146), 최고기온(℃)(10146), 일강수량(mm)(10146), 최대 순간 풍속(m/s)(10146), 최대 풍속(m/s)(10145), 평균 풍속(m/s)(10140), 풍정합(100m)(10138), 평균 이슬점온도(℃)(10141), 최소 상대습도(%)(10143), 평균 상대습도(%)(10141), 평균 증기압(hPa)(10139), 평균 현지기압(hPa)(10144), 최고 해면기압(hPa)(10146), 최저 해면기압(hPa)(10144), 평균 해면기압(hPa)(10144), 가조시간(hr)(10144), 합계 일조 시간(hr)(10131), 합계 일사(MJ/m2)(6088), 일 최심신적설(cm)(86), 일 최심적설(cm)(112), 평균 지면온도(℃)(10134), 최저 초상온도(℃)(10144), 안개 계속시간(hr)(241)
  최종적으로 30개의 데이터 속성이 남았으며, 괄호 안의 숫자는 해당 속성이 총 데이터 record 10146개 중 몇개에 기록되어 있는지를 나타낸다.

3. Missing Value 처리

- 평균 이슬점온도(℃)','최소 상대습도(%)','평균 상대습도(%)','평균 증기압(hPa)','평균 현지기압(hPa)','최저 해면기압(hPa)','평균 해면기압(hPa)','가조시간(hr)','합계 일조 시간(hr)'

  Missing value를 해당 날짜 전의 날씨 데이터로 replacement한다.

- 안개 지속시간, 일 최심적설, 일 최심신적설

  Missing value를 0으로 replacement한다. 적설량의 경우, 눈이 오지 않을 경우, 기록에 남지 않았기에 0으로 대체하는 것이 타당하다. 안개 또한 같은 이유로 0으로 대체하였다.

4. 카테고리 변수 처리

- One - hot encoding 방식을 통해서 처리하였다. Python과 Sklearn을 사용하였다.

```
from sklearn.preprocessing import OneHotEncoder
```

- 요일

  월, 화, 수, 목, 금, 토, 일은 1주일 주기로 반복 되기에 순서가 있다고 판단하지 않고, One-hot encoding 방식으로 처리하였다.

- 도시

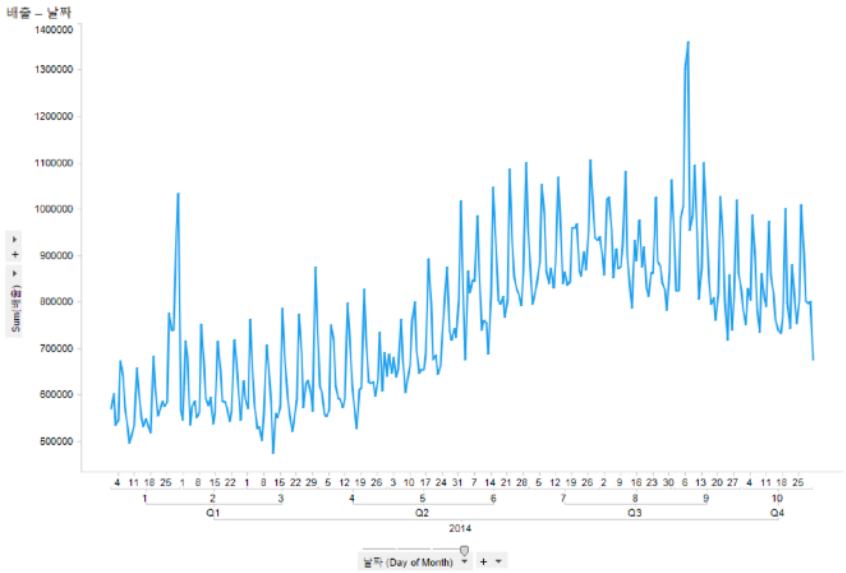
  각 도시명에 One-hot encoding 방식을 적용하였다.

- 월

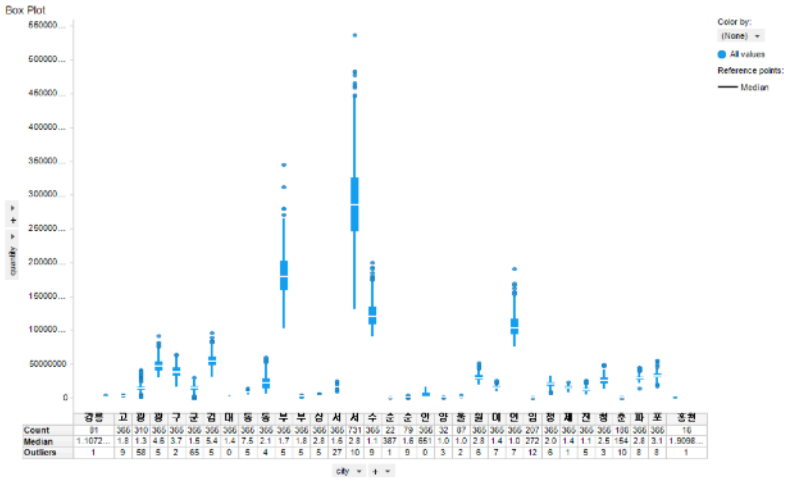
  2015년도내에서 순서가 있다 판단하여, 1월부터 12월을 해당 월을 나타내는 숫자를 그대로 사용하여 표현하였다.

5. 데이터 시각화

- 2014년도 전국 일별 배출량(날짜 - 배출량 그래프)



- 도시별 배출량 Box plot



6. Outlier 정리

- 일별 배출량을 나타낸 Line chart를 보면 설과 추석 전후로 음식물 배출량이 급증하는 것을 알 수 있다. 따라서 2015년, 2016년 또한 설, 추석 기간 음식물 쓰레기 배출량 데이터를 제거하였다.
  2015년 설날 2월 18, 19, 20, 21일, 2015년 추석 9월 26, 27, 28, 29일, 2016년 설날 2월 7, 8, 9, 10일(서울), 2016년 추석 9월 14, 15, 16일 데이터(서울)를 제거하였다.

- 지자체별 음식물 쓰레기 배출량을 Box plot으로 나타낸 시각화를 통해 outlier 후보를 찾아내었다. 이를 통해 outlier들을 제거하였다.

<Data Mining execution>

1. Data Partitioning

- 'Python sklearn' 을 통하여 data를 partitioning하였다.

```
from sklearn.model_selection import train_test_split
seed = 7
test_size = 0.1
X_train, X_test, y_train, y_test = train_test_split(encoded_x, Y,
test_size=test_size, random_state=seed)
```

- 데이터는 Randomly partitioning되었으며, Training data가 전체의 90%, Test data가 전체의 10%가 되도록 하였다.

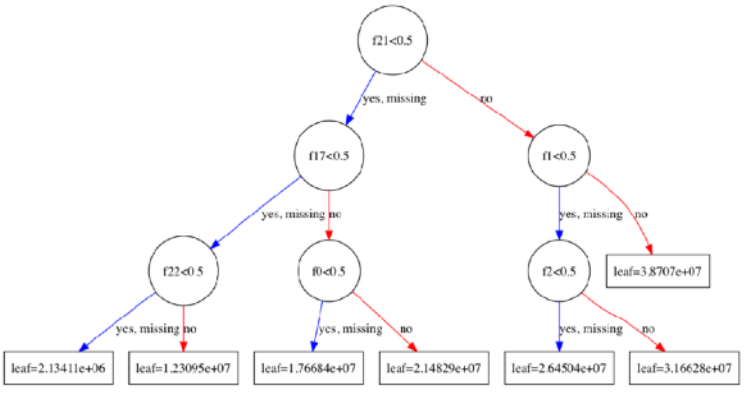
2. XGBOOST 실행

- 30개의 데이터 속성을 다 사용하여 XGBOOST를 실행한 결과
  Mean absolute percentage error: 2674.78035119%

- 사용된 XGBOOST model의 parameter value

```
XGBRegressor(base_score=0.5, colsample_bylevel=1, colsample_bytree=1,
       gamma=0, learning_rate=0.1, max_delta_step=0, max_depth=3,
       min_child_weight=1, missing=None, n_estimators=100, nthread=-1,
       objective='reg:linear', reg_alpha=0, reg_lambda=1,
       scale_pos_weight=1, seed=0, silent=True, subsample=1)
```

- XGBOOST model



<Schedule>

일정 및 역할	4월 30일	~5월 4일	~5월 15일	~5월 27일	~5월 29일
왕경율	제안서 작성	주제 정리 추가 데이터 마련	데이터 전처리	반복적인 실행 결과에 따른 모델 tuning 결과 분석	결과 분석 결과 활용 방안 마련 보고서 작성
박건우	데이터 획득	데이터 mining 기법 마련	XGBOOST 실행		
손장현	데이터 획득	데이터 pre-processing	중간 보고서 작성 데이터 전처리		

- 추가적인 Dimension Reduction 및 XGBOOST model Tuning 진행 필요
- Data Mining Architecture