뉴욕 택시 수요 예측 프로젝트

newyork

코드스테이츠
AI 07 한다운

Contents

프로젝트 소개

뉴욕의 택시는
언제, 어디서 가장 많은 수요가 있을까?

Google BigQuery에서 제공하는 데이터셋을
활용하여 뉴욕 택시 수요 예측 모델 구현.

데이터셋 소개 | 01. 프로젝트 소개

날짜별/지역별 수요

시간대에 따른 수요

Google BigQuery

SQL 활용하여 데이터 추출

데이터 로드 | 02. 데이터 전처리

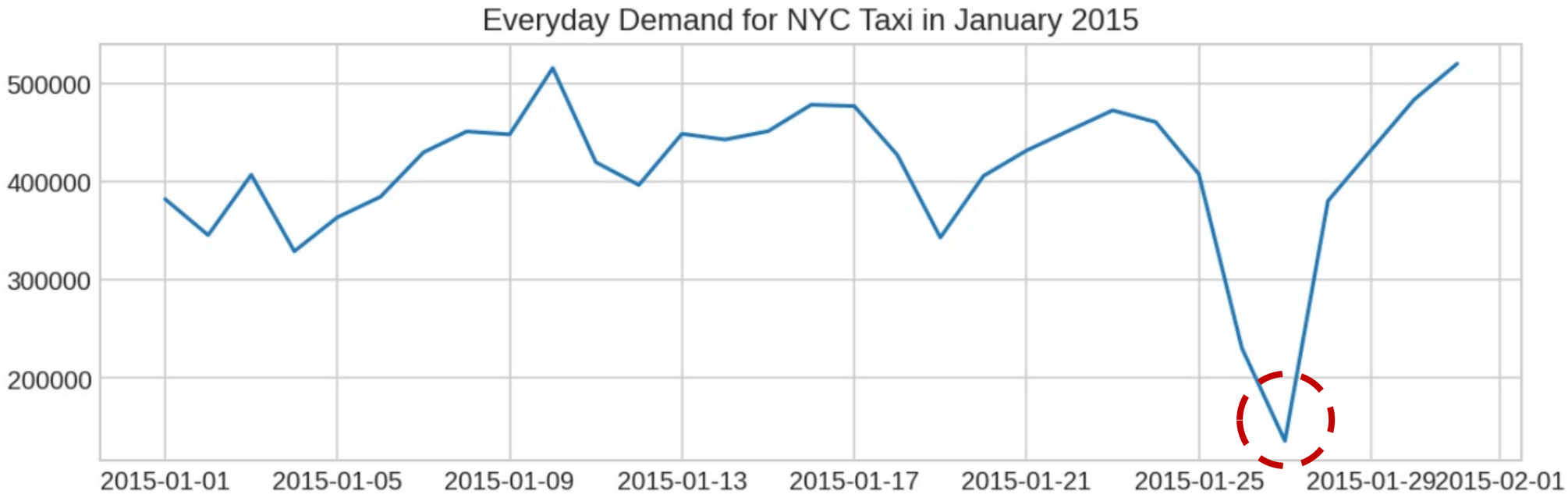
```
%%time
extract_query = """
SELECT
    *,
    EXTRACT(MONTH FROM pickup_hour) AS month,
    EXTRACT(DAY FROM pickup_hour) AS day,
    CAST(format_datetime('%u', pickup_hour) AS INT64) -1 AS weekday,
    EXTRACT(HOUR FROM pickup_hour) AS hour,
    CASE WHEN CAST(FORMAT_DATETIME('%u', pickup_hour) AS INT64) IN (6, 7) THEN 1 ELSE 0 END AS is_weekend
FROM (
    SELECT
        DATETIME_TRUNC(pickup_datetime, hour) AS pickup_hour,
        count(*) AS cnt
    FROM `bigquery-public-data.new_york_taxi_trips.tlc_yellow_trips_2015`
    WHERE EXTRACT(MONTH from pickup_datetime) = 1
    GROUP BY pickup_hour
)
ORDER BY pickup_hour
"""
PROJECT_ID = 'newyorktaxi-340407'

df = pd.read_gbq(query=extract_query, dialect='standard', project_id=PROJECT_ID)
```

df.tail()

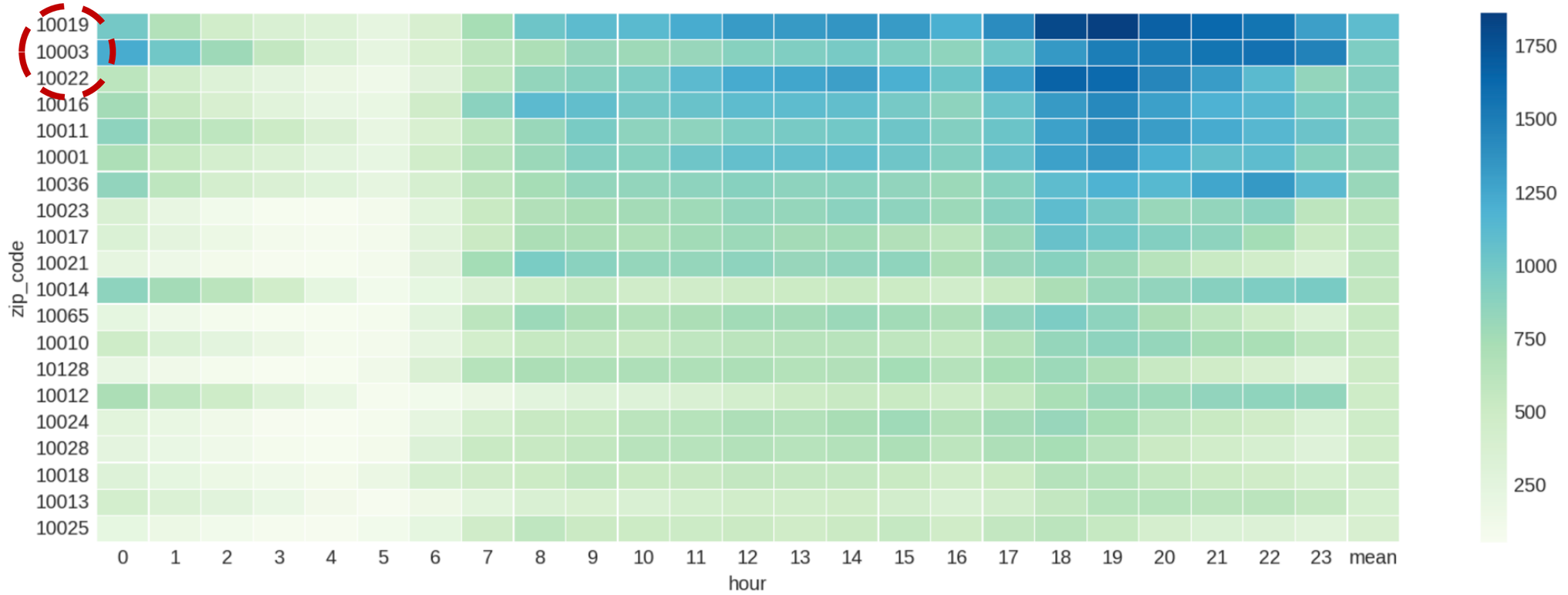
	pickup_hour	cnt	month	day	weekday	hour	is_weekend
739	2015-01-31 19:00:00	32436	1	31	5	19	1
740	2015-01-31 20:00:00	27555	1	31	5	20	1
741	2015-01-31 21:00:00	27477	1	31	5	21	1
742	2015-01-31 22:00:00	29862	1	31	5	22	1
743	2015-01-31 23:00:00	29856	1	31	5	23	1

데이터 분석 | 02. 데이터 전처리



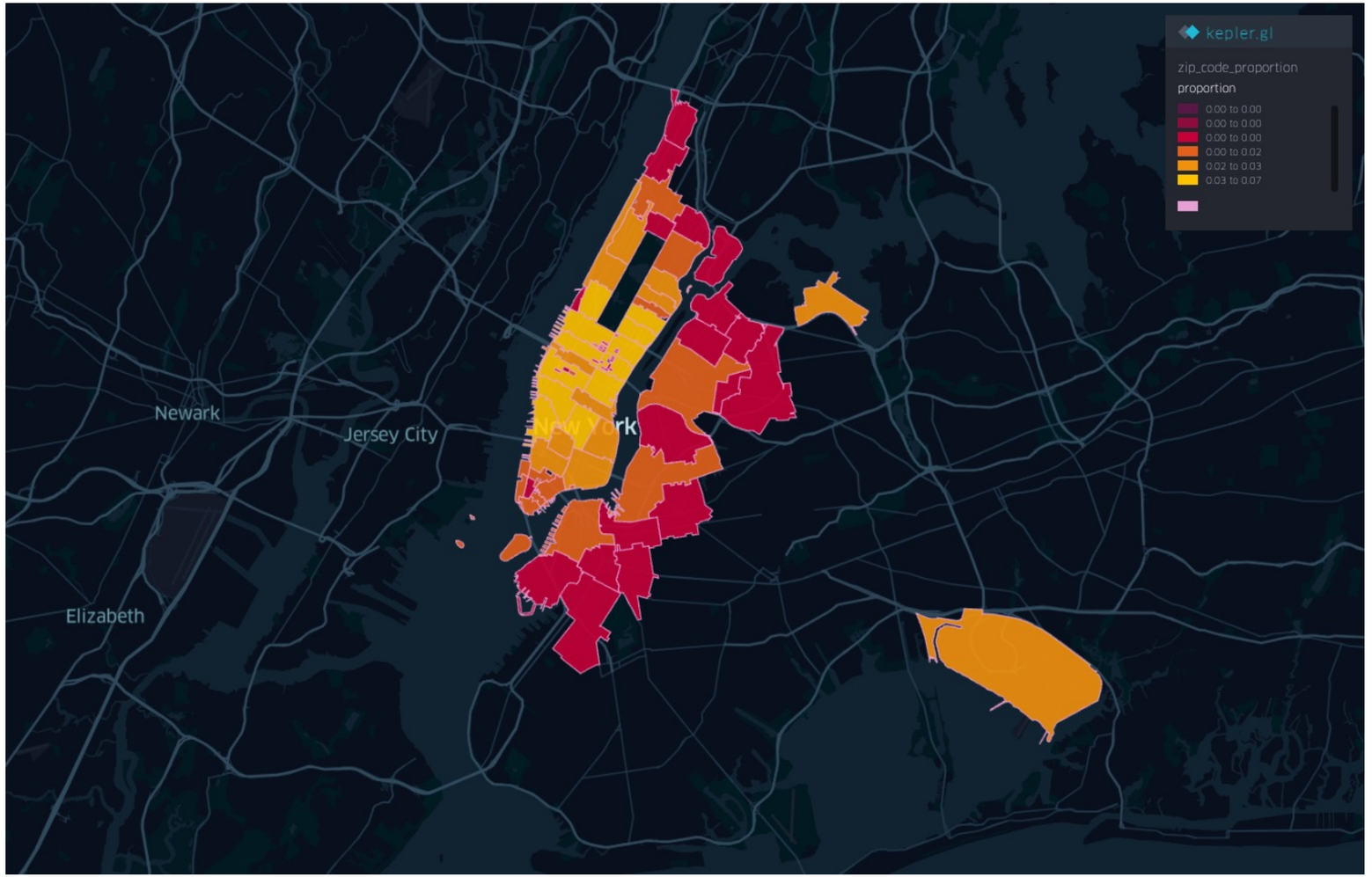
2015년 1월 동안의 뉴욕 주 택시 수요 데이터입니다.
2015년 1월 27일에 수요량이 급격히 감소 하는 형태입니다.
눈보라 사태로 인한 외출 감소가 원인으로 예상됩니다.

데이터 분석 | 02. 데이터 전처리



각 시간대별 지역에 따른 수요 데이터 시각화
우편번호가 10019, 10222인 지역의 택시 수요가 많은 편이며,
시간대가 17-20시 사이의 수요량이 높아집니다.

데이터 분석 | 02. 데이터 전처리



각 지역별로
택시 수요량을 직관적으로
시각화 하였습니다.

데이터 전처리

01

좌표를 zip_code로 변환하기

데이터셋 `bigquery-public-data.geo_us_boundaries.zip_codes`와 Join

02

One-hot Encoding

Zip_code를 원-핫 인코딩으로 변환

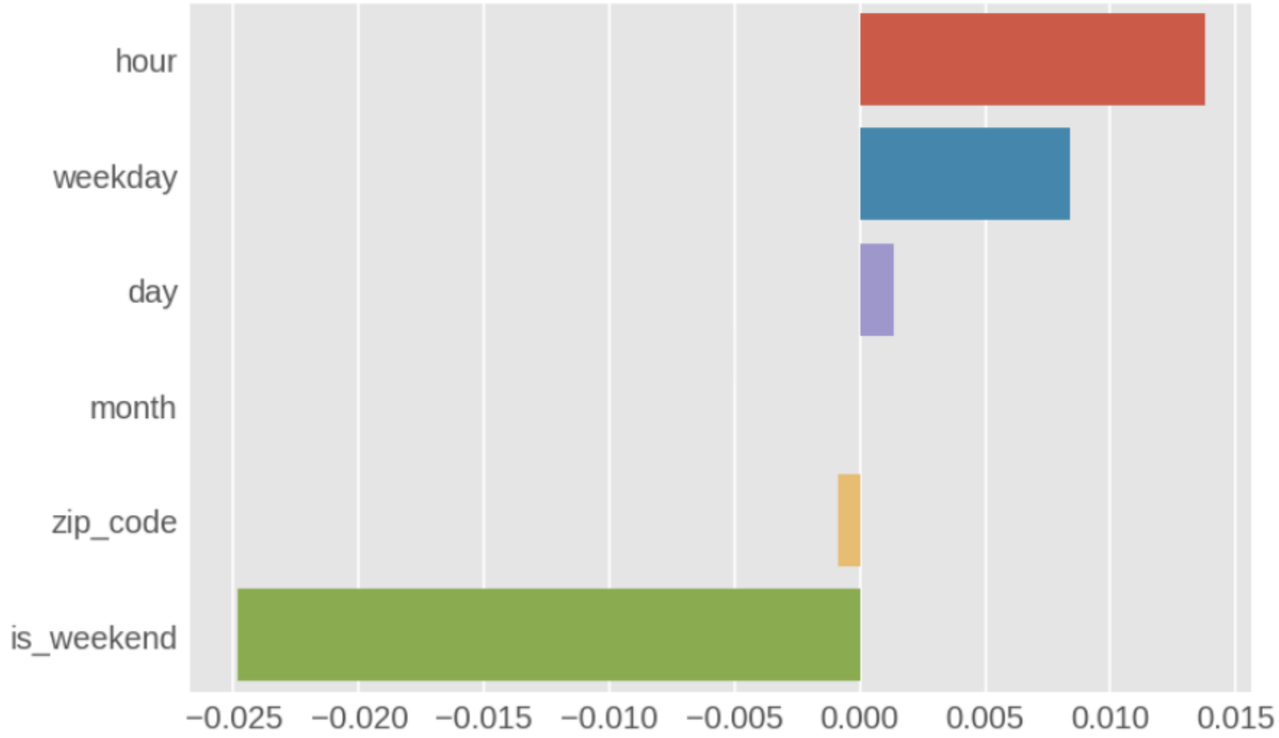
03

데이터셋 분리

시계열 데이터이므로 Random Sampling이 아닌, 과거-미래 순으로 분리

Simple Regression | 03. 모델 구축

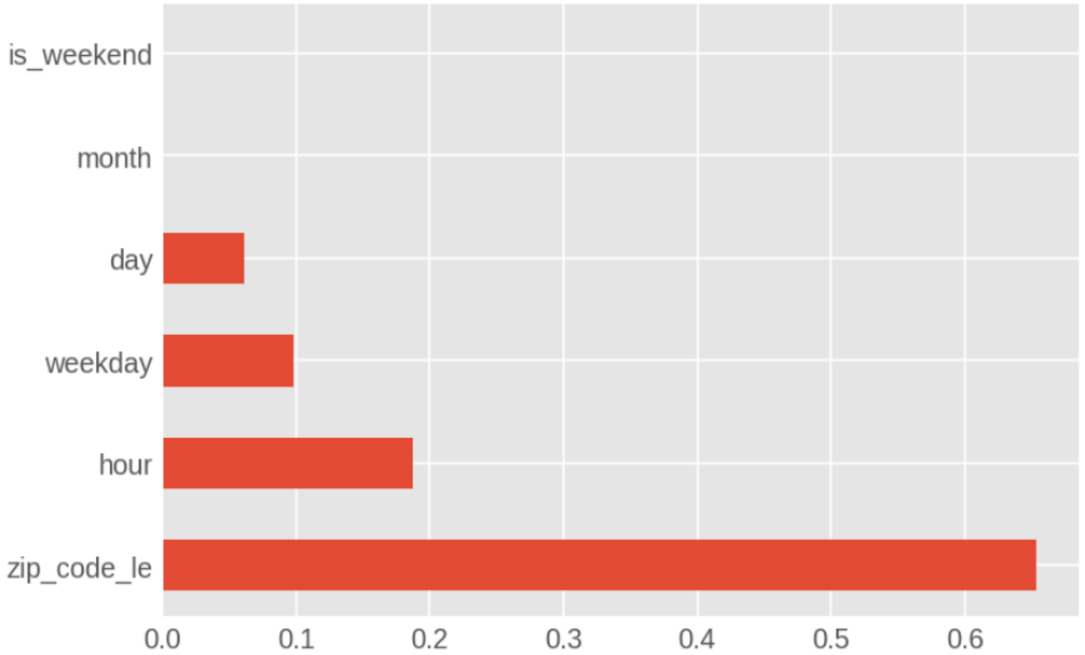
	MAE	MSE
Simple Regression	126.53	95916.68

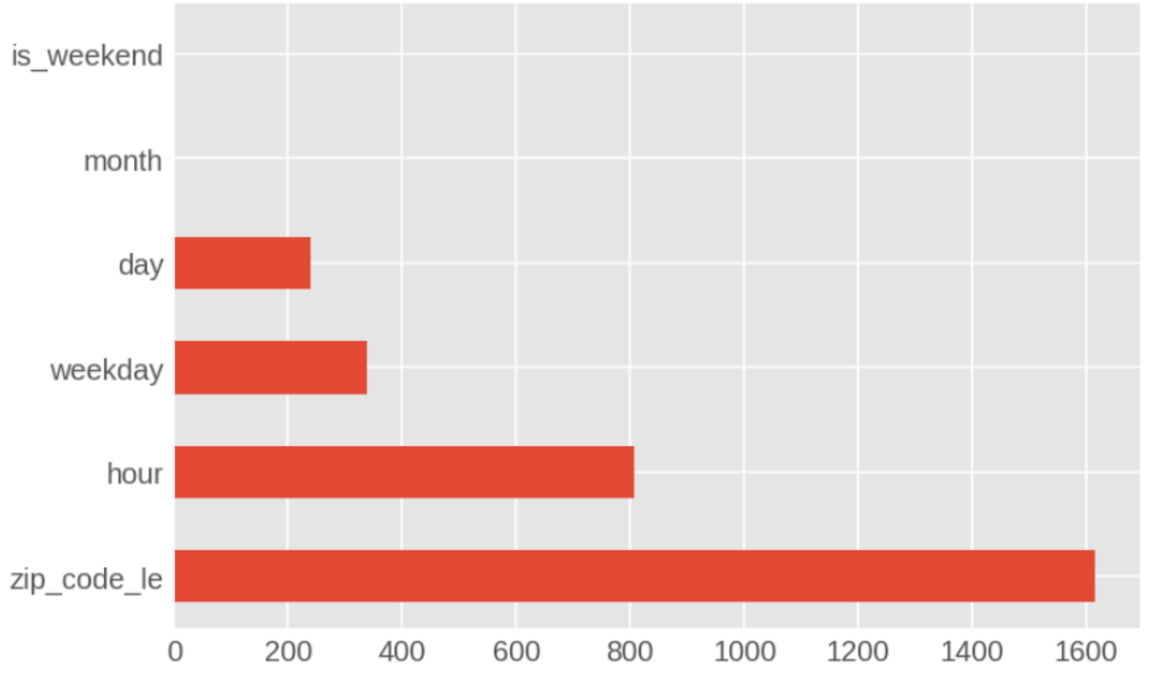


단순 회귀 모델을 구현하여
수요량을 예측하는 모델을 구축하였습니다.

Feature 중에서 타겟 값에 가장 큰 영향을
끼치는 것은 hour, weekday 였으며,
가장 관련 없는 것은 주말 여부 입니다.

XGBoost / LightGBM Regression | 03. 모델 구축

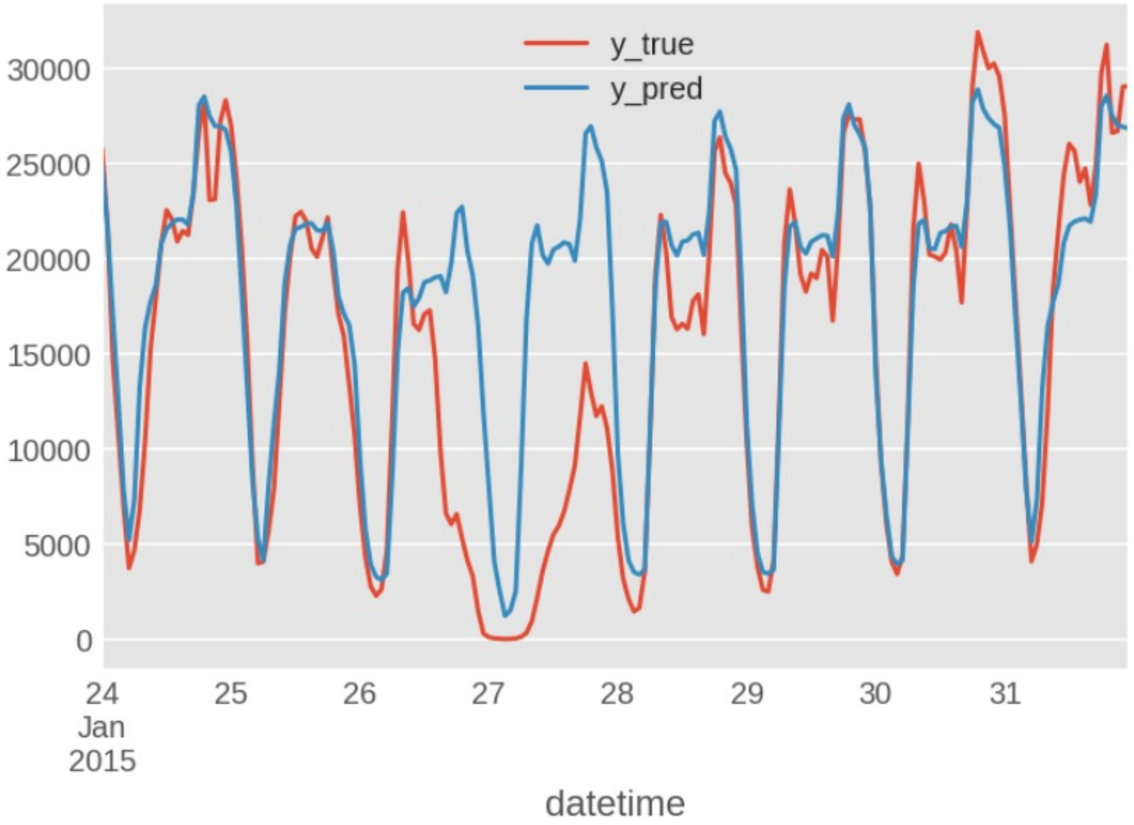


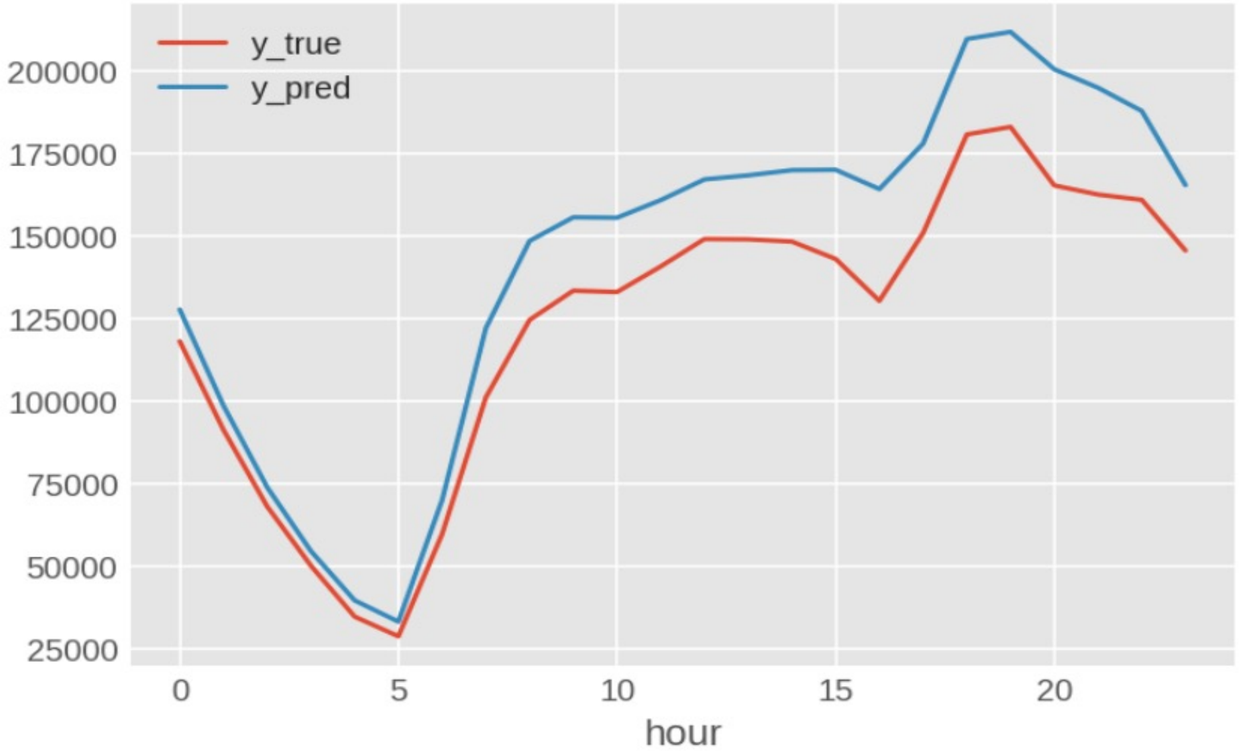


	MAE	MSE
XGBoost Regression	57.67	16512.33

	MAE	MSE
LightGBM Regression	48.24	13755.68

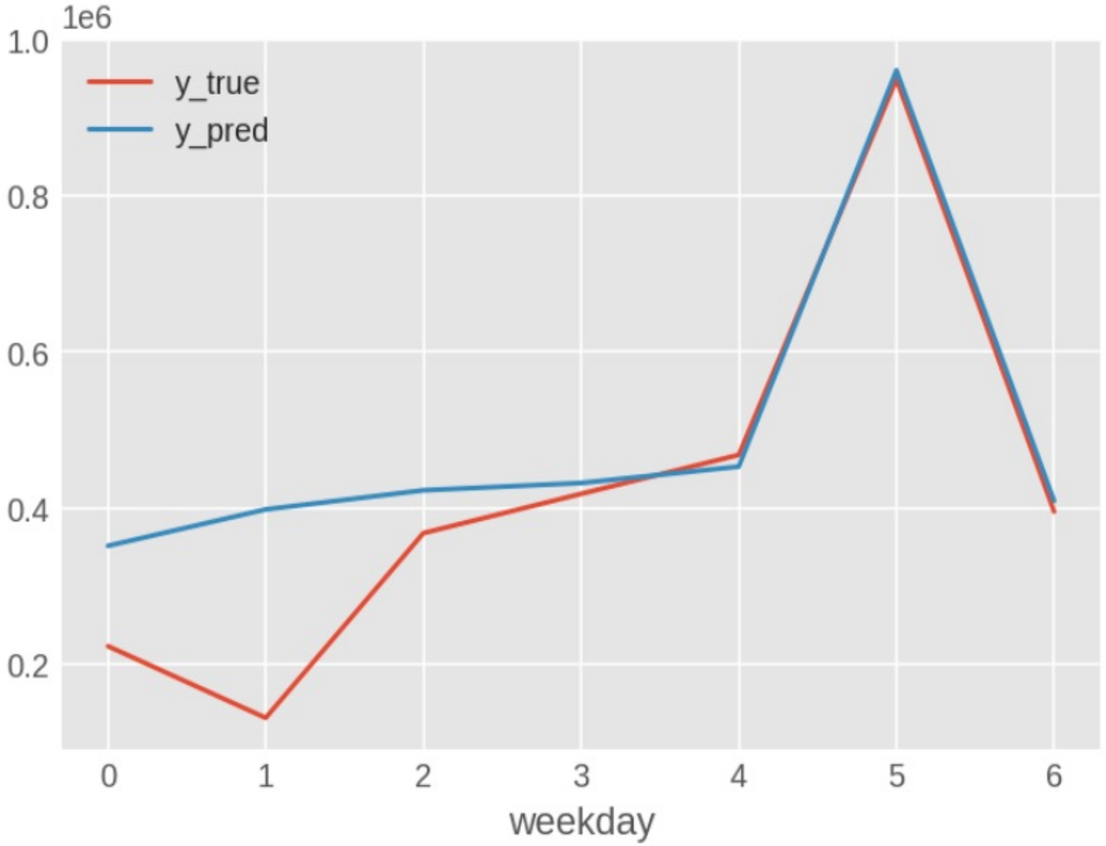
날짜 / 시간별 분석 | 04. 결과 분석

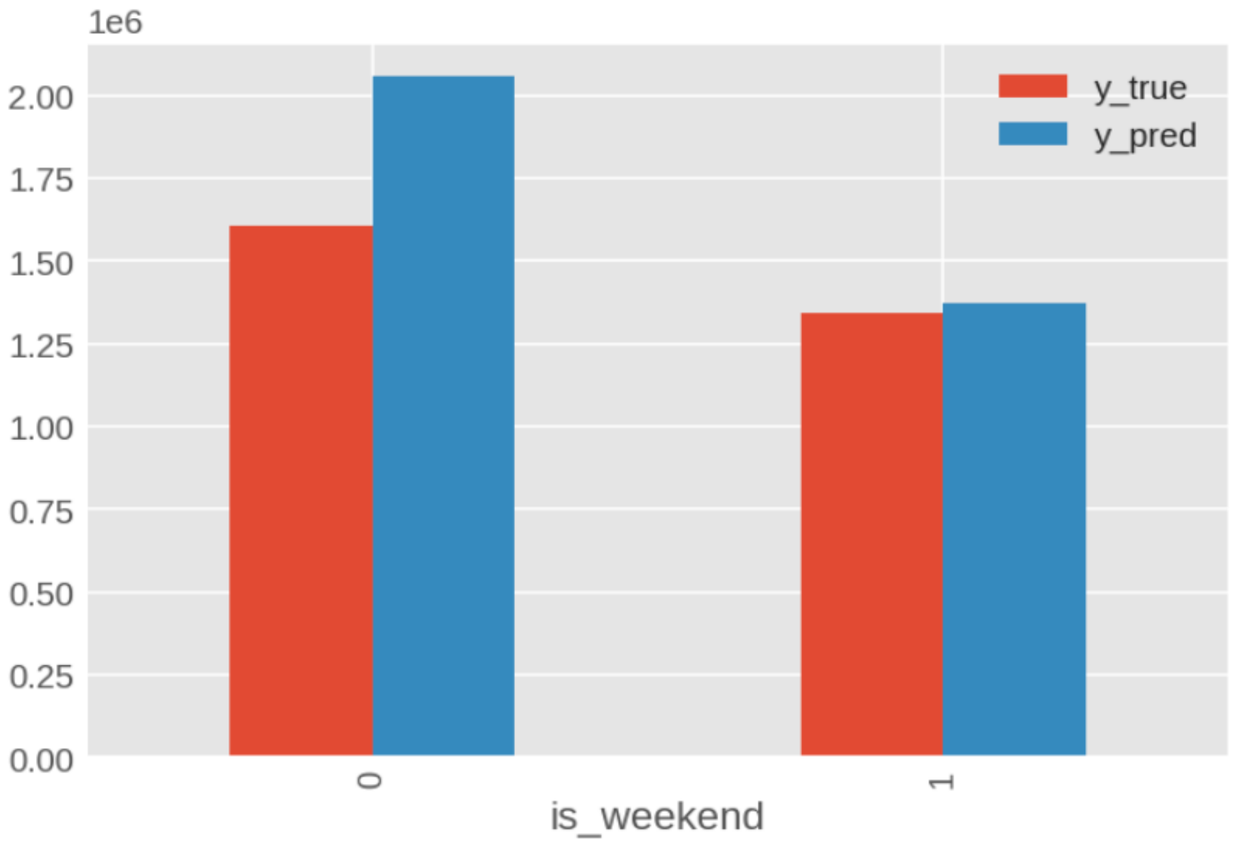




날짜별 예측 결과는 눈보라로 인한 비상 상황의 경우 정확도가 낮아집니다.
시간대별 예측 결과의 경우 실제 값과 큰 흐름이 같습니다.

요일 / 주말 여부 분석 | 04. 결과 분석





요일에 따른 예측 결과는 금, 토, 일의 경우 정확합니다.
주말의 경우 평일 보다 모델이 더 정확하게 예측합니다.

New York Taxi predict

THANK YOU.