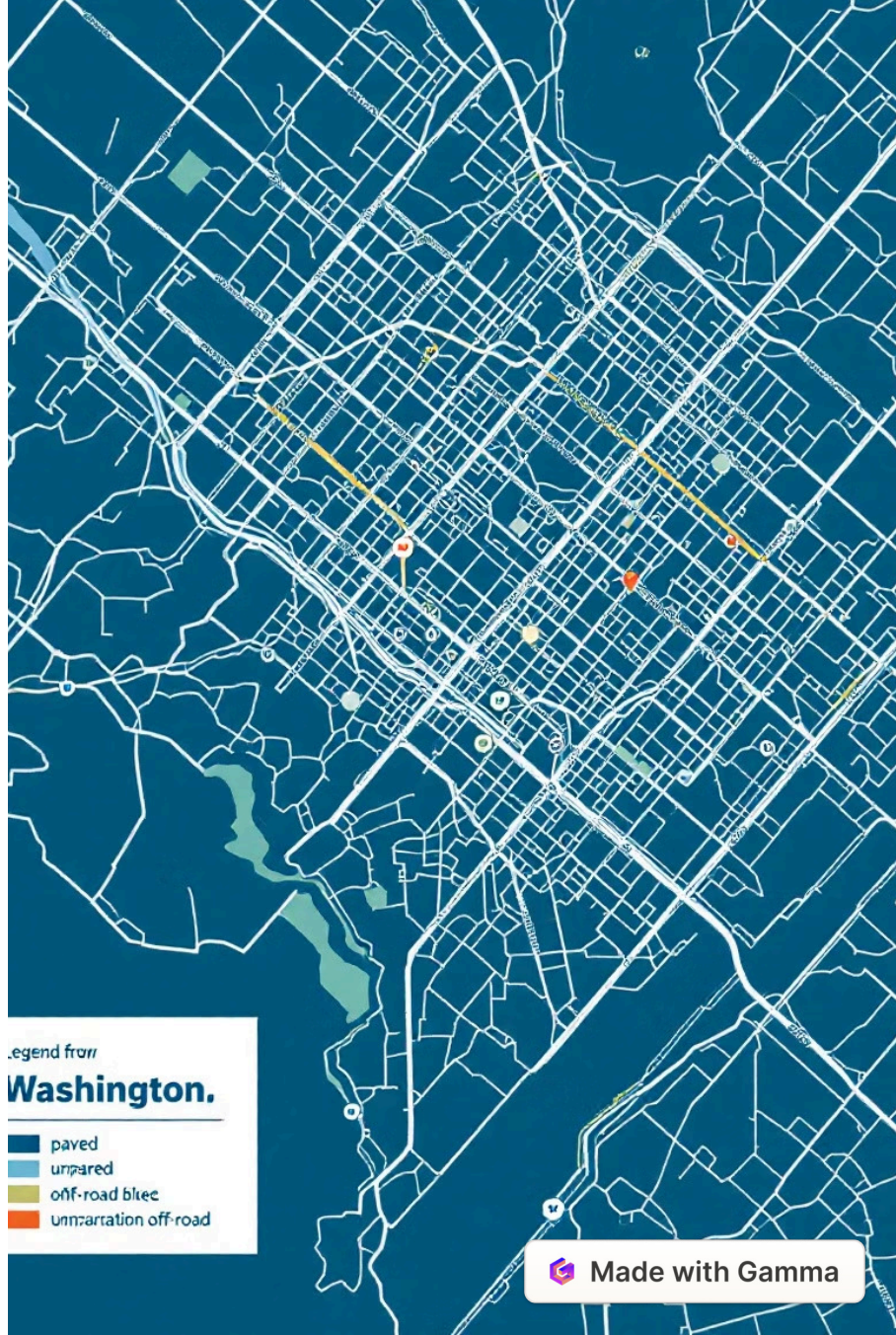
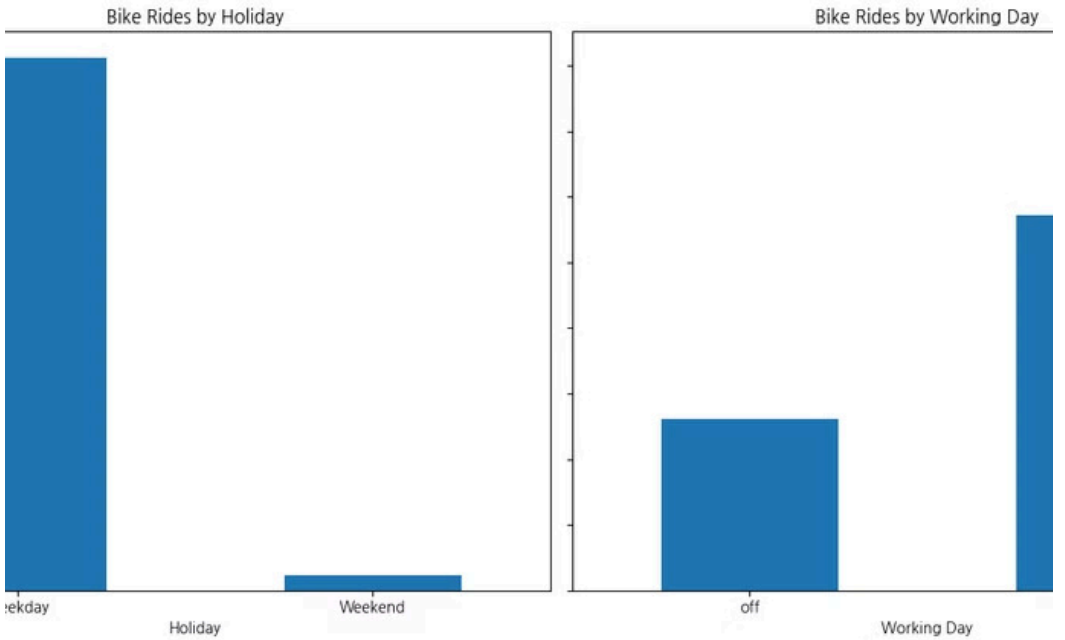
자전거 공유 서비스 데이터 분석 및 수요 예측 모델링

DA 2기 최승연



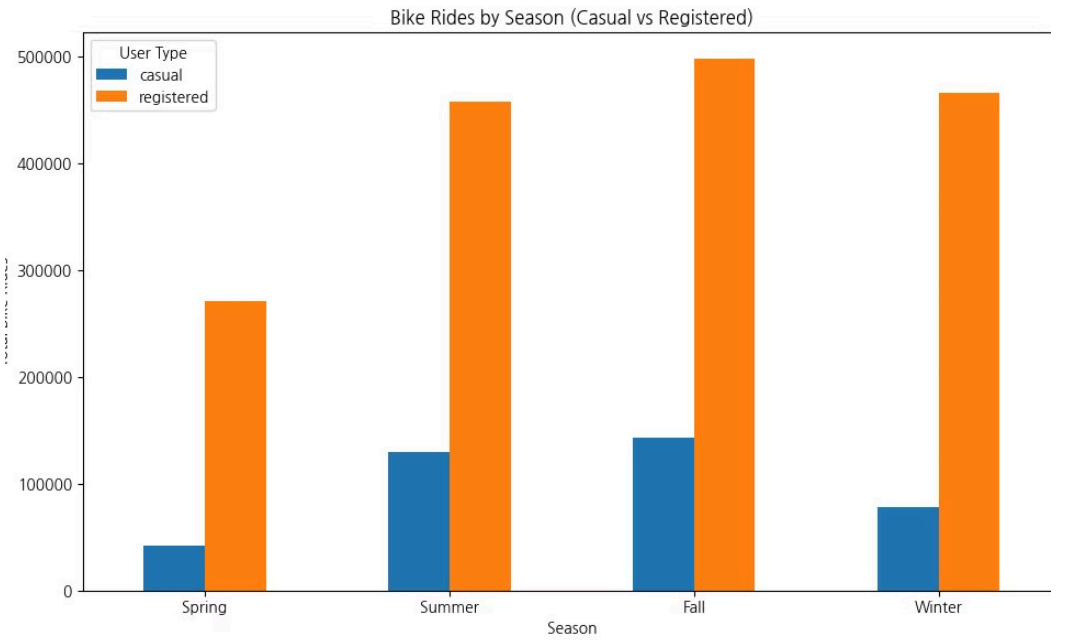
탐색적 데이터 분석 (EDA) - 데이터 시각화



공휴일/근무일별 자전거 사용량

평일에 훨씬 많이 이용하며 워킹데이의 이용자 수가 휴일의 이용자 수보다 많음.

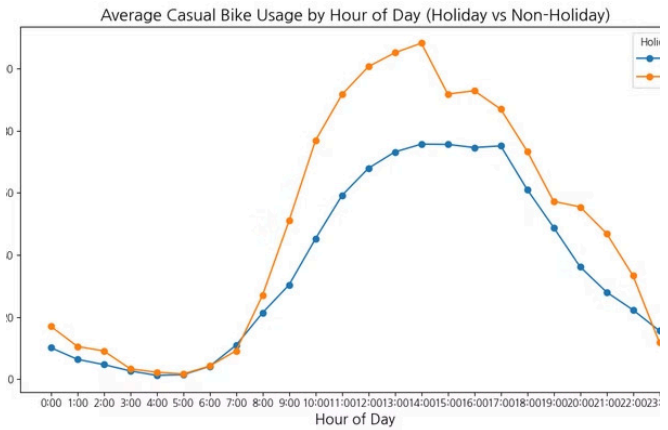
→ 통근용으로 주로 사용되는 것으로 보여짐

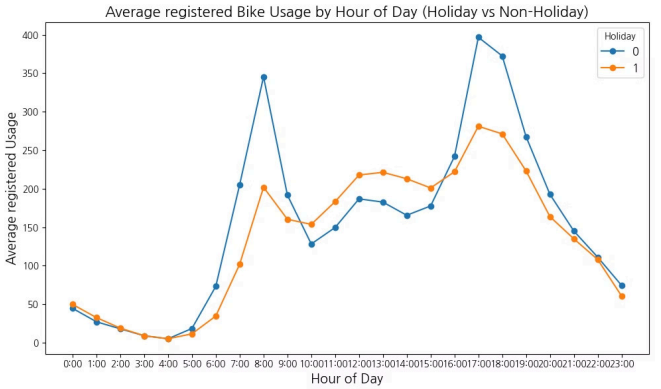


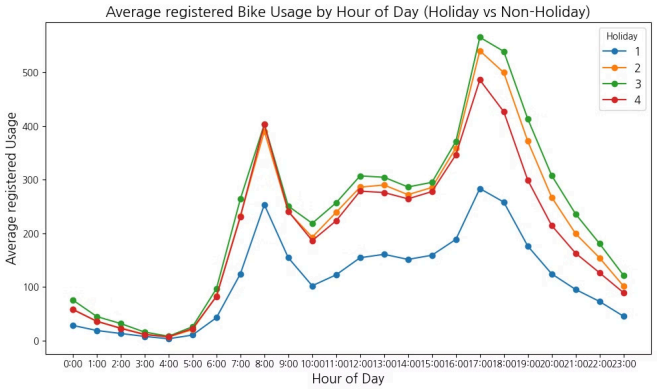
계절별 사용자 유형 비교

가을의 이용자 수가 제일 높고 봄의 이용자 수가 제일 낮았음. 봄을 제외한 계절의 이용자 수는 비슷함.

계절별 사용자 유형 비교







미등록 유저 평일/주말 이용패턴 차이가 존재

- 평일: 오전오후 5시까지 이용하고 이후부터 급격히 사용량이 줄어듬. 아마 아침오후까지 여행이나 단거리로 이용하는 사람들이 주류일 것임.
- 주말: 오전8시부터 사용량이 급증하고 오후1시 이후부터 급격히 줄어들고 저녁 7시부터 다시 급감함.

등록된 유저는 러시아워때 급격한 사용량을 띄고 있으며 평일보다 주말에 이러한 사용량이 더 두드러짐.

계절과 상관없이 많이 이용하는 시간대는 일정함.

가을-여름-겨울-봄 순으로 이용함.

모델링 - 데이터 전처리

1

시간별 칼럼 생성

- datetime 칼럼의 형변환 및 월별, 요일별, 일별, 시간대 등 4개의 파트로 나눔.
- 요일별 칼럼: 원핫 인코딩을 사용하여 변환한뒤 int 형변환
- 월별 칼럼: 숫자 매핑 후 주기를 확인하기 위해 sin/cos 칼럼으로 변환

2

결측치/이상치

결측치: 존재하지 않음

이상치: 수가 미미하여 제거하지 않음

3

중요 변수 선택

```
최적의 RMSLE: 0.4874641759305189
최적의 변수 조합: ('hour', 'month_sin', 'month_cos', 'holiday', 'workingday')
```

RMSLE 최소화하는 변수를 설정하는 함수를 만들어 도출된 변수들을 독립변수로 설정

4

다중공선성 확인

선정된 독립변수들 간의 다중공선성을 확인하여 모델의 정확도를 높임

5

데이터 분리

모델링

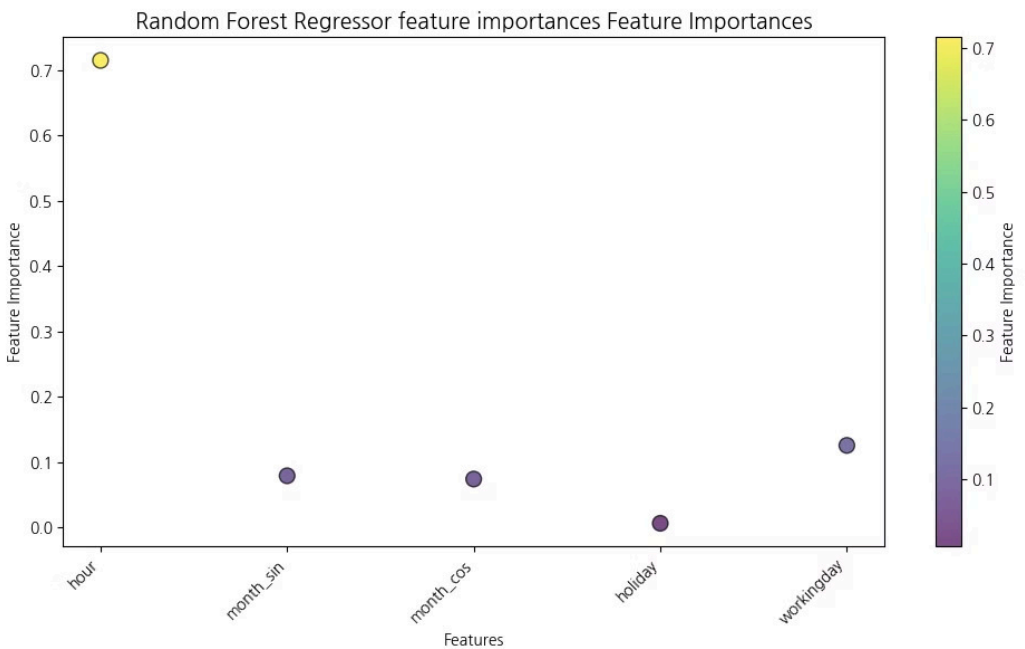
사용 모델

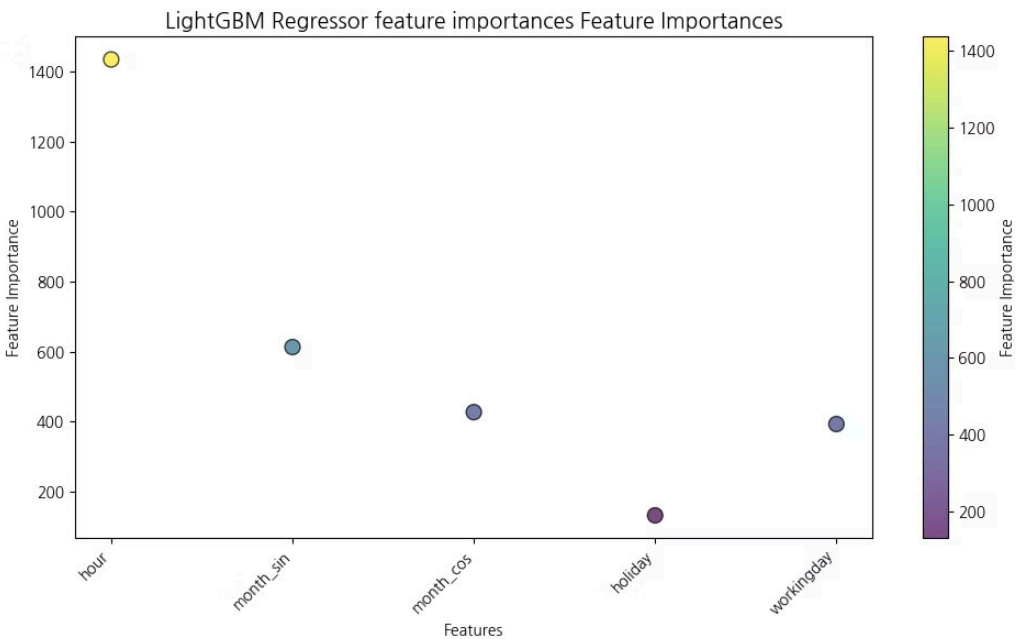
- LightGBM (LGBMClassifier)
- 선형 회귀 (Linear Regression, Ridge, Lasso)
- 랜덤 포레스트 (RandomForestRegressor)

평가 지표

- RMSLE (Root Mean Squared Log Error)
- R^2 Score

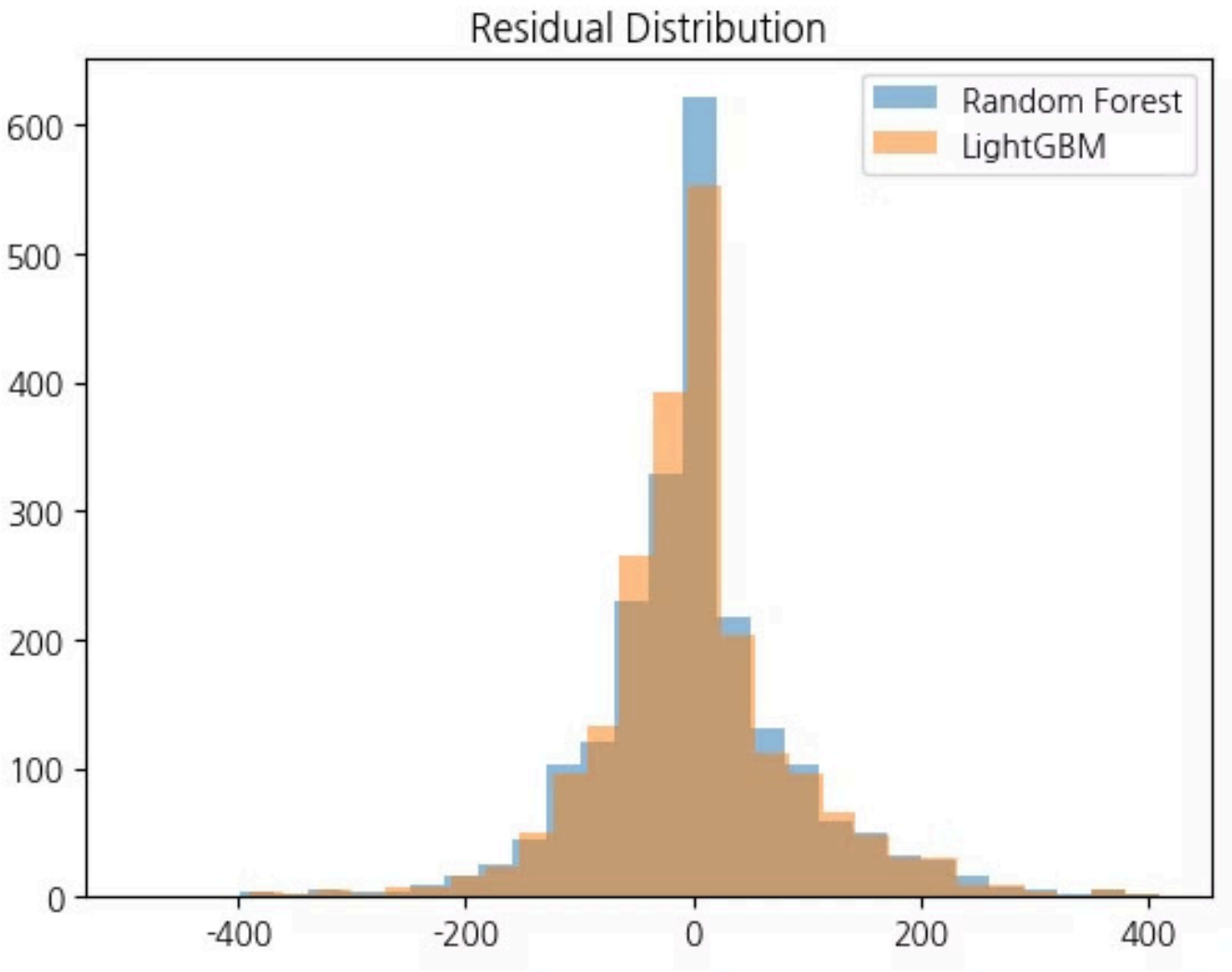
모델 선정





LightGBM이 랜덤 포레스트보다 month_sin과 month_cos의 중요도를 좀 더 반영하는 것을 알 수 있음.

모델 성능 비교 및 활용 방안



항목	Random Forest	LightGBM	비교 및 해석
R^2 (Train)	0.789	0.786	두 모델 모두 학습 데이터에서 비슷한 성능을 보임. Random Forest가 약간 더 높음.
R^2 (Test)	0.749	0.761	LightGBM이 테스트 데이터에서 더 높은 R^2로, 일반화 성능이 더 나음.
잔차 평균	0.296	0.540	Random Forest가 잔차 평균이 더 0에 가까워, 예측의 균형이 더 잘 맞음.
잔차 분산	8302.31	7905.79	LightGBM이 잔차 분산이 더 작아, 예측이 조금 더 안정적임.
잔차 분포 해석	잔차가 0 근처에 좁게 분포 (약간 뾰족)	잔차가 0을 중심으로 더 균일하게 분포	LightGBM이 잔차 분포에서 더 일반화된 예측을 보임.

LightGBM이 테스트 데이터에서의 결정 계수값이 더 높고 잔차 분산도 더 작아 안정적이고 일반적인 성능을 보여줌. Random Forest는 잔차 평균이 0에 더 가까워 train 데이터셋에 대해 예측 정확도가 조금 더 나음. 그러나 우리는 값이 정해지지 않은 미지의 테스트 데이터에서의 정확도를 중시해야 함.

→LightGBM을 선택