날씨 예측 기반 자전거 대여 활성화 전략

20212603 이가빈

목차

서론
1.1 연구 배경
1.2 연구 목적
1.3 데이터 수집 방법

문제 분석
2.1 문제의 중요성 및 영향

연구 분석
3.1 변수 설명
3.1 데이터 분석 방법
3.2 데이터 분석 결과

결론
4.1 연구 결과 요약

참고 문헌
5.1 데이터 출처

연구 배경

자전거 대여는 환경친화적이고 건강에 좋은 대중교통수단으로 인기를 얻고 있다. 그러나 자전거 대여량은 날씨 조건에 따라 변동성이 크며, 날씨가 좋을 때 자전거 대여 수요가 증가하는 경향이 있다. 이에 따라 날씨 예측을 통한 자전거 대여 수요 예측과 함께, 이를 기반으로 자전거 대여의 활성화를 위한 전략 수립이 중요한 과제로 부각되고 있다. 기존의 자전거 대여 시스템은 주로 정적인 요소를 고려하여 운영되었다. 하지만 날씨는 자전거 대여 수요에 민감한 요소로 작용하며, 날씨 조건에 따라 사용자들의 자전거 대여 패턴이 변화한다. 따라서, 날씨 예측 기반의 자전거 대여 활성화 전략을 개발하여 정확한 수요 예측과 효율적인 자전거 운영을 가능하게 하는 것이 필요하다. 날씨 예측 기술의 발전과 기상 데이터의 접근성 향상은 자전거 대여 시스템에 새로운 가능성을 제시하고 있다. 고객들은 자전거 대여를 결정할 때 현재 날씨 조건을 고려하며, 미리 예측된 날씨 정보를 활용하여 이용 계획을 세우는 경우가 많다. 따라서, 정확한 날씨 예측을 통해 불확실성을 감소시키고, 예상 수요에 맞게 자전거의 공급과 수요를 조절함으로써 자전거 대여 시스템의 효율성과 만족도를 향상시킬 수 있다. 본 연구에서는 날씨 예측 기반 자전거 대여 활성화 전략을 개발하여 기존의 정적인 운영 방식을 보완하고, 자전거 대여량을 증대시키는 방안을 제시한다.

연구 목적

본 연구의 목적은 날씨 예측 기반 자전거 대여 활성화를 위한 전략 수립이다. 첫 번째로 날씨와 자전거 대여량 간의 상관관계 분석한다. 분석을 통해 다양한 날씨 변수와 자전거 대여량 간의 관계를 탐색하고, 특정 날씨 조건이 자전거 대여 수요에 미치는 영향을 확인한다. 두 번째로는 자전거 대여 수요 예측 모델 개발하는 것이다. 분석을 통해 구한 날씨와 자전거 대여량 간의 관계를 이용하여 자전거 대여 수요를 예측하는 모델을 개발한다. 이를 통해 향후 날씨 상황에 따른 자전거 대여량을 예측할 수 있다. 이렇게 목적을 통해 날씨 예측 기반 자전거 대여 활성화를 위한 실용적인 전략을 제시하고, 자전거 대여 시스템의 효율성과 이용자 만족도를 향상시키는 데 기여하고자 한다.

데이터 수집 방법

본 연구에서 필요한 데이터를 수집하기 위해 다음과 같은 절차를 수행한다. 먼저 자전거 대여, 날씨 데이터를 획득한다. 그리고 수집한 데이터를 전처리하여 연구에 활용 가능한 형태로 가공한다. 이 과정에는 데이터 정제, 결측치 처리, 이상치 처리, 변수 변환 등의 작업이 포함될 수 있다. 이렇게 절차를 통해 적절하고 신뢰성 있는 데이터를 수집하여 연구에 활용할 수 있다. 데이터 수집 과정에서는 데이터의 정확성에 유의하여야 하고, 연구 목적에 맞는 데이터를 효율적으로 수집하는 것이 중요하다.

문제의 중요성 및 영향

자전거는 환경친화적이고 건강에 이롭다는 점에서 사회적으로 중요한 교통수단이다. 그러나 자전거 대여의 활성화와 이용자 만족을 위해서는 대여 시스템의 효율적인 운영이 필요하다. 이에 따라 자전거 대여 수요에 영향을 미치는 요인을 이해하고, 이를 기반으로 한 운영 전략

을 개발하는 것이 중요하다. 그리고 자전거 대여 시스템의 효율적인 운영은 이용자의 만족도 향상에 직결된다. 날씨 예측 기반의 자전거 대여 활성화 전략은 이용자들에게 더 나은 이용 경험을 제공할 수 있다. 이를 통해 이용자의 편의성과 접근성을 개선하고, 자전거 대여 가능성이 최적화되어 만족도를 향상시킬 수 있다. 이는 자전거 대여 시스템의 이용률 증가와 이용자들의 재방문율 상승에 긍정적인 영향을 미칠 것이다. 따라서 연구를 통해 날씨와 자전거 대여량 사이의 상관관계를 탐구하고, 이를 바탕으로 자전거 대여 활성화를 위한 전략 수립에 기여할 수 있다.

변수설명
독립변수: Visibility(미세먼지), Solar.Radiation(태양복사량), Rainfall(강수량), Snowfall(강설량), Seasons(계절), Holiday(휴일), Functioning.Day(일하는 날), Date(날짜), Temperature(온도), Wind.speed(풍속)
종속변수: Rented.Bike.Count(자전거 대여량)

데이터 분석 방법
시간 데이터를 일별 데이터로 바꾸는 과정: 데이터 셋에서 시간 데이터를 추출하여 일별로 집계하고, 일별로 집계된 데이터 셋을 새로 생성한다.

이상치 및 결측치 확인: 일별 데이터 셋에서 이상치, 결측치를 확인하였지만 이상치, 결측치 둘 다 없었다.

다중 선형회귀 모델링: 독립 변수와 종속 변수 간의 다중 선형 관계를 분석하기 위해 회귀 분석을 수행한다.

정규 분포 확인: 회귀 모델의 잔차(residuals)를 분석하여 정규성을 확인한다. shapiro.test() 함수를 사용하여 잔차의 정규성을 검정한다. 결과에서 Shapiro-Wilk 검정의 통계량인 W 값은 0.98866이고, p-value 값은 0.006154이다. 일반적으로 유의수준 0.05를 기준으로 p-value가 이보다 작으면 정규성 가정을 한다. 따라서, 해당 결과에서 p-value가 0.006154로 유의수준보다 작으므로, 잔차는 정규 분포를 따르지 않는다고 할 수 있다. 이는 모델이 잔차의 정규성 가정을 만족하지 못한다는 것을 의미한다.

잔차의 자기상관 확인: 회귀 모델의 잔차 간에 자기상관이 있는지 확인한다.
durbinWatsonTest() 함수를 사용하여 잔차의 자기상관을 검정한다. Durbin-Watson 통계량은 0에서 4 사이의 값을 가지며, 2에 가까울수록 자기상관이 없다는 가정을 나타낸다. 해당 결과에서 Durbin-Watson 통계량은 1.124925이다. Durbin-Watson 검정은 귀무가설이 잔차들 사이에 자기상관이 없다는 것을 가정하며, 대립가설은 잔차들 사이에 자기상관이 존재한다는 것이다. p-value 값이 0으로 나왔으므로, 유의수준 0.05보다 작으므로 귀무가설을 기각하고, 대립가설을 채택한다. 따라서, 해당 회귀 모델의 잔차들 사이에는 자기상관이 존재한다고 할 수 있다. 이는 회귀 모델이 시계열 데이터에 적합하지 않을 수 있다는 것을 의미한다.

로그 변환 후 다중 선형회귀 모델 재구축: 변수 변환을 하는 이유는 모델의 성능 향상을 위해, 종속 변수와 독립 변수가 정규분포를 따르지 않아서 가깝게 만들려고 하는 것이다. 앞에서 정규분포를 따르지 않으므로 로그 변환을 수행한다. 로그 변환한 변수를 사용하여 다중 선형회귀 모델을 다시 구축한다.

정규 분포 확인: 로그 변환한 모델의 잔차를 분석하여 정규성을 확인한다. Shapiro-Wilk 정규성 검정은 데이터가 정규 분포를 따르는지를 확인하는 통계적인 방법이다. 검정 결과는 Shapiro-Wilk 통계량인 W 값과 p-value 값을 제공한다. 해당 결과에서 Shapiro-Wilk 통계량인 W는 0.76888이다. p-value 값은 2.2e-16보다 작다. Shapiro-Wilk 검정은 귀무가설이 데이터가 정규 분포를 따른다는 것을 가정하며, 대립가설은 데이터가 정규 분포를 따르지 않는다는 것이다. p-value 값이 매우 작기 때문에 유의수준 0.05보다 작다. 따라서, 귀무가설을 기각하고 대립가설을 채택한다. 이는 회귀 모델의 잔차들이 정규 분포를 따르지 않는다는 것을 의미한다.

제곱근 변환 후 다중 선형회귀 모델 재구축: 앞에 분석에서 정규분포를 따르지 않으므로 제곱근 변환을 수행한다. 제곱근 변환한 변수를 사용하여 다중 선형회귀 모델을 다시 구축한다.

정규 분포 확인: 제곱근 변환한 모델의 잔차를 분석하여 정규성을 확인한다. Shapiro-Wilk 정규성 검정은 데이터가 정규 분포를 따르는지를 확인하는 통계적인 방법이다. 검정 결과는 Shapiro-Wilk 통계량인 W 값과 p-value 값을 제공한다. 해당 결과에서 Shapiro-Wilk 통계량인 W는 0.75556이다. p-value 값은 2.2e-16보다 작다. Shapiro-Wilk 검정은 귀무가설이 데이터가 정규 분포를 따른다는 것을 가정하며, 대립가설은 데이터가 정규 분포를 따르지 않는다는 것이다. p-value 값이 매우 작기 때문에 유의수준 0.05보다 작다. 따라서, 귀무가설을 기각하고 대립가설을 채택한다. 이는 제곱근 변환된 회귀 모델의 잔차들이 정규 분포를 따르지 않는다는 것을 의미한다.

다중공선성 확인: 앞에서 정규 분포를 따르는지 확인하는 분석을 하였는데 계속 정규 분포를 따르지 않아서 정규성은 무시하고 독립 변수들 간의 다중공선성을 확인하여 모델의 안정성을 평가한다. vif() 함수를 사용하여 다중공선성을 계산하고, 문제가 있는 독립 변수를 식별한다. VIF 값이 1에 가까울수록 다중공선성이 적고, VIF 값이 크면 다중공선성이 높다고 판단된다. 해당 결과에서 VIF 값은 Seasons에 대해 5.885121, Temperature에 대해 5.369084로 나타난다. 나머지 변수들은 모두 1보다 작은 값을 가지므로 다중공선성의 문제가 없다고 판단할 수 있다. Seasons와 Temperature의 VIF 값이 상대적으로 크므로, 이 변수들은 다른 독립 변수들에 의해 상당한 설명력을 받고 있을 가능성이 있다. 이러한 다중공선성이 존재할 경우 회귀 계수의 추정이 불안정해질 수 있으므로, 다중공선성이 높은 변수들을 제거한다.

다중공선성이 높은 독립 변수 제거: 다중공선성이 높은 Seasons 변수를 제거하여 모델을 개선한다.

단계적 선택법 수행: 변수 선택을 위해 단계적 선택법(Stepwise Selection)을 수행한다.

stepAIC() 함수를 사용하여 최적의 변수 조합을 선택하고, 모델을 최적화한다. Multiple R-squared 값은 모델이 종속 변수의 변동을 얼마나 설명할 수 있는지를 나타내는 지표이다. 0부터 1 사이의 값을 가지며, 값이 클수록 모델이 데이터를 잘 설명한다는 의미이다. 예를 들어, Multiple R-squared 값이 0.8093인 경우, 모델이 종속 변수의 약 80.93%의 변동을 설명할 수 있다고 할 수 있다. 따라서, 이 회귀 모델은 다양한 독립 변수들이 종속 변수인 Rented.Bike.Count를 통계적으로 유의하게 예측하는 데 유의미한 영향을 가지고 있으며, 모델이 데이터의 약 80.93%를 설명할 수 있다는 결론을 내릴 수 있다.

회귀식:
Rented.Bike.Count = -4.719e+05 + 2.623e+01 * Date - 5.884e+01 * Humidity - 1.077e+03 * Wind.speed + 3.075e+02 * Dew.point.temperature + 1.306e+04 * Solar.Radiation - 1.371e+02 * Rainfall - 4.124e+01 * Snowfall + 2.829e+03 * HolidayNo Holiday + 2.023e+04 * Functioning.DayYes

데이터 분석 결과
다중 선형회귀 모델을 통해 얻은 결과를 바탕으로 데이터 분석 결과이다.
날짜(Date) 변수는 자전거 대여량에 양의 영향을 미친다. 일별 날짜가 증가할수록 자전거 대여량이 증가하는 경향을 보인다. 날짜 변수가 2017-12-01 이런 식으로 표시된 것은 변수의 형식을 나타내는 것이며, 실제로는 해당 변수를 숫자형 변수로 변환하여 사용한다. 예를 들어, 2017-12-01을 1로, 2017-12-02를 2로 변환하는 방식이다. 이렇게 변환된 숫자 변수를 회귀식에 포함시킨다. 습도(Humidity) 변수는 자전거 대여량에 음의 영향을 미친다. 습도가 증가할수록 자전거 대여량이 감소하는 경향을 보인다. 풍속(Wind.speed) 변수는 자전거 대여량에 음의 영향을 미친다. 풍속이 증가할수록 자전거 대여량이 감소하는 경향을 보인다. 이슬점 온도(Dew.point.temperature) 변수는 자전거 대여량에 양의 영향을 미친다. 이슬점 온도가 증가할수록 자전거 대여량이 증가하는 경향을 보인다. 태양 복사량(Solar.Radiation) 변수는 자전거 대여량에 양의 영향을 미친다. 태양 복사량이 증가할수록 자전거 대여량이 증가하는 경향을 보인다. 강수량(Rainfall) 변수는 자전거 대여량에 음의 영향을 미친다. 강수량이 증가할수록 자전거 대여량이 감소하는 경향을 보인다. 강설량(Snowfall) 변수는 자전거 대여량에 음의 영향을 미친다. 강설량이 증가할수록 자전거 대여량이 감소하는 경향을 보인다. 하지만 통계적 유의성은 보이지 않는다. 휴일(Holiday) 변수는 자전거 대여량에 양의 영향을 미친다. 휴일이 아닌 날보다 휴일인 날에 자전거 대여량이 증가하는 경향을 보인다. 일하는 날(Functioning.Day) 변수는 자전거 대여량에 양의 영향을 미친다. 일하는 날에는 자전거 대여량이 증가하는 경향을 보인다. 위의 회귀식을 통해 각 독립 변수가 자전거 대여량에 미치는 영향을 분석할 수 있다. 이를 통해 날씨 조건과 근무일과 휴일에 따른 자전거 대여량 변화를 분석한 결과, 다음과 같은 결론을 도출할 수 있다. 휴일(Holiday)이 있는 날에는 자전거 대여량이 상대적으로 높았다. 휴일은 사람들의 여가 활동 및 외출 활동이 많아지는 시기이므로, 자전거 대여량도 증가하는 경향을 보였다. 일하는 날(Functioning.Day)에도 자전거 대여량이 높았다. 일하는 날에도 사람들은 자전거를 이용하여 출퇴근하거나 일상적인 이동에 사용할 가능성이 높기 때문이다. 따라서, 자전거 대여량을 증가시키기 위해서는 휴일에는 관련 이벤트나 프로모션 등을 통해 사용자들의 이용 동기를 높일 수 있고, 근무일에는 자전거 이용을 편

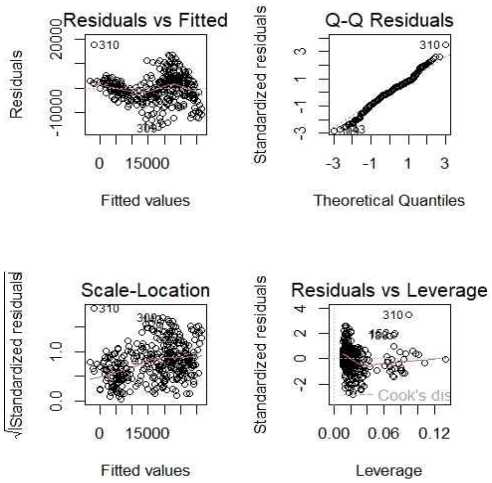
리하게 하는 인프라 개선 및 접근성 향상을 고려해야 한다. 이를 통해 자전거 대여량을 더욱 늘릴 수 있을 것이다.

연구 결과 요약

본 연구에서는 자전거 대여량과 날씨 요소 간의 관계를 분석하기 위해 회귀 분석을 수행하였다. 회귀 모델을 통해 자전거 대여량을 예측하는데 에는 유의미한 결과를 얻었다. 날짜, 습도, 풍속, 이슬점 온도, 태양 복사량, 강수량, 강설량, 휴일 여부, 일하는 날 여부 등의 변수들이 자전거 대여량에 영향을 미치는 것으로 나타났다. 날짜 변수(Date)는 양의 회귀 계수를 가지고 있으며, 이는 날짜가 자전거 대여 수에 양의 영향을 미친다는 것을 의미한다. 따라서, 날짜가 증가할수록 자전거 대여 수도 증가할 것으로 예측된다. 습도(Humidity) 변수는 음의 회귀 계수를 갖고 있다. 이는 습도가 증가할수록 자전거 대여 수가 감소한다는 것을 나타낸다. 이는 습도가 높을수록 사람들이 자전거를 대여하기를 꺼릴 가능성이 있거나, 기후 조건이 좋지 않아서 자전거를 이용하는 사람들이 줄어들 수 있다는 것을 시사한다. 풍속(Wind.speed) 변수도 음의 회귀 계수를 가지고 있다. 이는 풍속이 증가할수록 자전거 대여 수가 감소한다는 것을 의미한다. 강한 바람의 영향으로 자전거 이용이 불편해지거나, 안전에 대한 우려로 인해 자전거 대여 수가 감소할 수 있다. 이슬점 온도(Dew.point.temperature), 일사량(Solar.Radiation), 강수량(Rainfall), 적설량(Snowfall), 휴일 여부(HolidayNo Holiday), 정상 운영 여부(Functioning.DayYes) 변수들은 각각 자전거 대여 수에 유의한 영향을 미치는 것으로 나타났다. 따라서, 날짜, 습도, 풍속, 이슬점 온도, 일사량, 강수량, 적설량, 휴일 여부, 정상 운영 여부 등의 요인들이 자전거 대여 수에 영향을 미친다는 것을 확인할 수 있다. R-squared 값이 0.8093로 나타나며, 이는 모델이 자전거 대여량의 변동성의 약 80.93%를 설명할 수 있음을 의미한다. 이러한 연구 결과는 날씨 예측 기반 자전거 대여 활성화에 대한 중요한 정보를 제공한다. 날씨 변수들을 고려하여 자전거 대여량을 예측하는 모델을 개발함으로써, 도시 교통 문제 해결과 환경 보호에 기여할 수 있다. 또한, 향후 도시 계획 및 자전거 대여 서비스 제공 업체에게 참고 자료로 활용될 수 있다.

데이터 출처

data set source: Sathishkumar V E, Jangwoo Park, and Yongyun Cho. 'Using data mining techniques for bike sharing demand
prediction in metropolitan city.' Computer Communications, Vol.153, pp.353-366, March, 2020
Sathishkumar V E and Yongyun Cho. 'A rule-based model for Seoul Bike sharing demand prediction using weather
data' European Journal of Remote Sensing, pp. 1-18, Feb, 2020
Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of
California, School of Information and Computer Science.
URL: https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand



"Residuals vs Fitted" 그림은 회귀 모델의 잔차(residuals)와 예측값(fitted values) 사이의 관계를 시각화한 그래프이다. 이 그림은 회귀 모델의 적합도와 잔차의 패턴을 평가하는 데에 사용된다. x축에는 예측값(fitted values)이, y축에는 해당 예측값에 대한 잔차(residuals)가 나타난다. 잔차는 실제 관측 값과 모델이 예측한 값 간의 차이를 나타내는 오차이다. 이 그림을 통해 회귀 모델이 예측한 값과 실제 데이터 간의 차이를 시각적으로 확인할 수 있다.

"Q-Q Residuals" 그림은 회귀 모델의 잔차(residuals)가 정규 분포를 따르는지 확인하기 위해 사용되는 그래프이다. Q-Q (Quantile-Quantile) 플롯은 잔차의 분포와 정규 분포 간의 비교를 시각화한다. 그림에서 x축은 정규 분포의 분위수(quantiles)이고, y축은 해당 분위수에 해당하는 잔차의 값이다. Q-Q Residuals 그림은 회귀 모델이 가정하는 잔차의 정규성을 확인하는 데 유용하다.

"Scale-Location" 그림은 회귀 모델의 잔차(residuals)의 표준화된 값 또는 제곱근에 대한 예측값(fitted values)의 분산에 대한 그래프이다. 이 그림은 회귀 모델의 잔차들이 일정한 분산을 가지고 있는지 확인하기 위해 사용된다. 그림에서 x축은 예측값(fitted values)이고, y축은 잔차의 표준화된 값 또는 제곱근이다. 각 점은 해당 예측값에서의 잔차의 표준화된 값 또는 제곱근을 나타낸다. Scale-Location 그림은 회귀 모델에서 잔차들의 분산을 확인하고 모델의 가정을 평가하는 데 사용된다.

"Residuals vs Leverage" 그림은 회귀 모델에서 각 관측치의 잔차(residuals)와 해당 관측치의 영향도(Leverage)를 시각화하는 그래프이다. 이 그림은 회귀 분석에서 이상치나 영향력이 큰 관측치를 식별하는 데 사용된다. 그림에서 y축은 잔차의 표준화된 값 또는 제곱근이다. 잔차는 해당 관측치의 실제 값과 모델로 예측한 값 간의 차이를 나타낸다. 이 그림을 통해 모델에서 중요한 관측치를 확인하고, 모델의 안정성과 적합성을 평가할 수 있다.