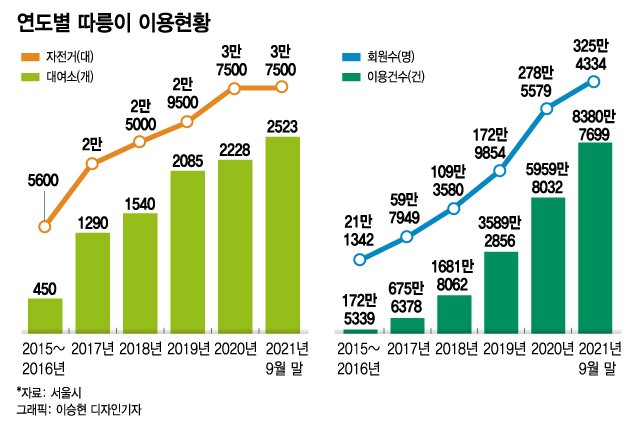
|  |
| --- |
| 서울시 공공자전거 이용자 수 예측을 통한 쏠림 현상 해결  컴퓨터공학전공 2022208026 장욱 |

**1. 서론**

현대 사회는 기술의 급속한 발전과 함께 환경 오염의 심각성이 증가하는 추세를 보이고 있습니다. Earth.Org의 ‘2022년 지구 위기 보고서’[1]에 따르면, 대기오염, 농업, 음식물 쓰레기 등 다양한 환경 문제가 심각한 수준에 이르렀습니다. 이러한 문제에 대응하기 위해 여러 국가에서는 공공 자전거 서비스를 도입하여 환경 개선, 교통 체증 완화, 시민들의 편리성 제공 및 건강한 생활 방식 촉진 등의 긍정적인 효과를 경험하고 있습니다. 특히, 네덜란드, 독일, 영국 등 유럽 국가들은 공공 자전거 서비스를 통해 도시 교통의 지속 가능성을 높이고 환경 친화적인 대안을 제공하고 있습니다.

대한민국 서울시도 이러한 글로벌 추세에 발맞추어 2015년 10월부터 '따릉이'라는 공공자전거 대여 서비스를 시작했습니다. '따릉이'는 서비스 시작 당시 5,600대의 자전거와 450개의 대여소로 시작하여, 2022년 현재 약 43,500대의 자전거와 2,719개의 대여소로 확장되었습니다. 이는 서비스 시작 이후 규모가 약 8배 증가한 것으로, 서울시 공공자전거 서비스의 성장과 시민들의 높은 수요를 반영합니다.



[2] 연도별 따릉이 이용현황

그러나 따릉이 서비스의 높은 수요로 인해 발생하는 '따릉이 쏠림 현상'과 '노동자 분배 작업'에 대한 문제가 대두되었습니다. 따릉이 쏠림 현상은 공공자전거의 특성상 발생하는 문제로, 대여소 간 이용 가능한 자전거 수의 편차가 크게 나타나는 현상입니다. 이로 인해 이용자들은 원하는 대여소에 자전거가 없을 경우 불편을 겪거나 다른 대여소로 이동해야 하는 상황에 직면하고 있습니다.

노동자 분배 작업 문제는 쏠림 현상을 해결하기 위해 서울시에서 노동자들을 고용하여 분배 작업을 진행하고 있지만, 이러한 노력에도 불구하고 쏠림 현상은 지속되고 있으며, 인건비 증가로 인한 재정적 부담이 커지고 있습니다.

이러한 문제를 해결하기 위해 본 프로젝트에서는 서울시 공공자전거 데이터를 활용하여 자전거 이용자 수를 예측할 수 있는 시스템을 개발하였습니다. 이 시스템은 자전거의 수요와 공급을 효율적으로 관리하고, 쏠림 현상을 완화하는 데 기여할 것으로 기대됩니다. 또한, 이 시스템을 통해 서울시 공공자전거 서비스의 운영 효율성을 높이고, 시민들의 이용 편의성을 개선하는 데 중요한 역할을 할 것입니다.

**2. 이론적 배경**

**2.1 선행연구**

본 프로젝트는 서울시 공공자전거 서비스인 '따릉이'의 사용 패턴과 이용자 만족도에 초점을 맞추고 있습니다. 이는 서울시 공공자전거 서비스의 현황을 정확히 이해하고, 서비스의 효율성과 만족도를 향상시키기 위한 구체적인 방안을 제시하는 데 중요한 기초가 됩니다. 선행 연구를 통해 '따릉이' 서비스의 이용 증가 추세와 이에 따른 쏠림 현상, 그리고 이용자들의 만족도와 관련된 문제점을 파악하였습니다.

텍스트, 라인, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

* 공공자전거 이용 추이 변화

'코로나 19 전후 서울시 공공자전거 이용 패턴의 변화(2021)'[3] 논문에 따르면 서울시 공공자전거 대여 서비스인 '따릉이'의 이용자 수가 지속적으로 증가하고 있음을 보여줍니다. 특히 코로나19 팬데믹 이후 공공자전거에 대한 수요가 크게 증가한 것으로 나타났습니다. 이는 공공자전거 서비스가 서울시민의 생활교통 수단으로 자리 잡고 있음을 시사하며, 동시에 서비스의 확장과 개선의 필요성을 강조합니다**.**

**'**공공자전거 이용특성이 만족도에 미치는 영향 분석(2021)[4]' 논문은 서울시민들이 공공자전거 배차 대수에 대해 44.4%의 불만족도를 나타내고 있다고 지적합니다. 이는 늘어나는 수요에도 불구하고 시민들의 배차 대수에 대한 만족도가 낮다는 점을 드러내며, '따릉이' 서비스의 쏠림 현상이 존재함을 시사합니다. 이러한 쏠림 현상은 서비스의 효율성과 이용자 만족도에 부정적인 영향을 미치며, 이를 해결하기 위한 적극적인 대책 마련이 필요함을 나타냅니다.

텍스트, 스크린샷, 도표, 폰트이(가) 표시된 사진

자동 생성된 설명

- 자전거 배차대수 만족도

**2.2 분석 방법 및 모델 선정 기준**

선행 연구를 바탕으로, 본 프로젝트에서는 서울시 공공자전거 이용자 수 예측을 위한 분석 방법과 모델 선택의 기준을 설정하였습니다. 분석 방법으로는 데이터 수집, 정제, 탐색적 데이터 분석(EDA), 모델링 및 검증의 과정을 포함하며, 모델은 가장 대표적인 예측 모델인 3가지 회귀 모델을 선정하였습니다.

1. 선형회귀 (Linear Regression)

선형회귀는 가장 기본적이면서 널리 사용되는 예측 모델 중 하나입니다. 이 모델은 독립변수와 종속변수 간의 선형 관계를 찾아내는 방식으로 작동합니다. 선형회귀의 주요 장점은 그 구현의 단순성과 해석의 용이성에 있습니다. 모델이 생성하는 선형 관계는 이해하기 쉽고, 데이터의 특성과 결과 사이의 관계를 명확하게 설명할 수 있습니다. 그러나 선형회귀는 비선형 관계를 모델링하는 데 한계가 있으며, 이상치에 매우 민감하게 반응할 수 있는 단점이 있습니다.

1. 릿지회귀 (Ridge Regression)\*\*

릿지회귀는 선형회귀의 확장으로, 과적합을 방지하기 위해 정규화를 추가한 모델입니다. 이 모델은 회귀 계수의 크기를 제한함으로써 모델의 복잡성을 줄이고, 더 일반화된 모델을 생성합니다. 릿지회귀의 장점은 과적합을 방지하고, 다중공선성 문제를 완화하는 데 있습니다. 그러나 정규화 강도를 결정하는 하이퍼파라미터를 적절히 조정해야 하는 필요성이 있으며, 이는 추가적인 튜닝 과정을 요구합니다.

1. 결정 트리 회귀   
   (Decision Tree Regression)

결정 트리 회귀는 데이터를 여러 하위 집합으로 분할하고, 각 분할에서 예측 값을 결정하는 방식으로 작동합니다. 이 모델은 비선형 관계와 복잡한 데이터 구조를 잘 포착할 수 있으며, 결과의 해석이 비교적 쉽다는 장점이 있습니다. 또한, 데이터의 스케일링에 민감하지 않고, 이상치에 강한 특성을 가집니다. 그러나 과적합의 위험이 있으며, 특히 데이터가 많은 경우에는 모델이 복잡해져 예측 성능이 저하될 수 있습니다.

위 3가지 회귀 모델을 학습시킨 후 가장성능이 뛰어난 모델을 선정하기 위해 다음과 같은 네가지 성능 지표를 비교하였습니다.

1. MSPE (Mean Squared Percentage Error, 평균 제곱 백분율 오차)

MSPE는 예측값과 실제값의 차이를 백분율로 나타낸 값입니다. 이 지표는 예측 오차의 크기를 백분율로 표현하여, 모델의 예측 정확도를 직관적으로 이해할 수 있게 해줍니다. 낮은 MSPE 값은 모델의 예측 정확도가 높음을 의미합니다.

2. MAE (Mean Absolute Error, 평균 절대 오차)

MAE는 예측값과 실제값의 차이의 절대값의 평균을 나타냅니다. 이 지표는 예측 오차의 평균 크기를 나타내며, 모델의 예측 오차가 얼마나 큰지를 평가하는 데 유용합니다. 낮은 MAE 값은 더 정확한 예측을 의미합니다.

3. R Square (결정 계수)

R Square는 모델이 데이터의 변동성을 얼마나 잘 설명하는지를 나타내는 지표입니다. 0에서 1 사이의 값을 가지며, 값이 높을수록 모델이 데이터를 잘 설명한다고 볼 수 있습니다. 이 지표는 모델의 설명력을 평가하는 데 중요합니다.

4. Adjusted R Square (조정된 결정 계수)

Adjusted R Square는 독립 변수의 수가 모델의 성능에 미치는 영향을 고려한 결정 계수입니다. 이 지표는 독립 변수의 수가 많을 때 과적합을 방지하는 데 도움을 줍니다. 높은 Adjusted R Square 값은 독립 변수의 수를 고려한 상태에서도 모델이 데이터를 잘 설명한다는 것을 의미합니다.

위 네가지 성능 지표들을 사용하여 모델의 성능을 비교한 후 가장 성능이 좋은 모델을 이용하여 서울시 공공자전거 이용자 수 예측 서비스의 모델로 선정하였습니다.

**3. 데이터 분석**

**3.1 데이터 설명**

본 프로젝트를 진행하기 위해 데이터 분석 플랫폼인 Kaggle에서 ‘Seoul Bike Sharing Demand’ 데이터 셋을 이용하였습니다. 이 데이터 셋은 총 8,760개의 행과 14개의 열로 구성되어 있으며, 서울시 공공자전거 이용에 관한 다양한 정보를 담고 있습니다. 각 열은 날짜, 자전거 이용량, 시간, 온도, 습도, 풍속, 가시거리, 이슬점, 태양 복사량, 강수량, 강설량, 계절, 공휴일 여부, 영업 여부 등을 포함합니다. 이 데이터 셋의 중요성은 서울시 공공자전거 서비스의 이용 패턴과 수요를 이해하는 데 있습니다. 이 데이터를 통해 서울시 공공자전거 이용량에 영향을 미치는 다양한 요인들을 분석할 수 있으며, 이는 공공자전거 서비스의 효율성을 높이고, 쏠림 현상을 해결하는 데 중요한 기초 자료가 됩니다.

텍스트, 메뉴, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

* Seoul Bike Sharing Demand DataSet

데이터 분석 과정에서 가장 중요한 단계 중 하나는 데이터의 품질을 검증하고 필요한 전처리를 수행하는 것입니다. 이를 위해 우선적으로 ‘Seoul Bike Sharing Demand’ 데이터 셋의 결측치 여부를 철저히 분석하였습니다. 분석 결과, 이 데이터 셋에는 결측치가 존재하지 않는 것으로 확인되었습니다.

데이터 처리의 효율성을 높이기 위해, 데이터 셋의 열 이름에 포함된 공백과 특수문자를 언더바(\_)로 대체하였습니다. 이러한 작업은 데이터 처리 과정에서 발생할 수 있는 오류를 최소화하고, 데이터에 대한 접근과 조작을 용이하게 만듭니다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

* 결측치 분석과정

텍스트, 스크린샷, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

자동 생성된 설명

* 공백 및 특수문자 제거

**3.2 데이터 분석 과정**

데이터 분석 과정은 서울시 공공자전거 서비스의 이용 패턴을 이해하고 자전거 이용자 수를 예측하기 위해 진행하였습니다. 데이터 분석 과정은 총 5단계로 진행되었습니다.

1. 온도, 강수량과 이슬점의 관계

이슬점은 대기 중 수증기의 양을 나타내는 지표로, 온도와 습도를 통해 결정됩니다. 이슬점은 기본적으로 해당 온도에서 공기가 포화되어 이슬이 형성되기 시작하는 지점을 의미합니다. 이는 온도와 습도가 서로 밀접하게 연관되어 있음을 나타냅니다. 따라서, 예측 모델을 학습할 때 온도와 습도만을 사용하면 이슬점의 정보를 충분히 반영할 수 있습니다. 기온과 습도 데이터는 이슬점을 유추하는 데 필요한 모든 정보를 포함하고 있습니다. 이는 이슬점이 기온과 습도의 함수로 표현될 수 있기 때문입니다. 예측 모델에서 온도와 습도를 독립 변수로 사용하면, 이슬점을 별도로 고려하지 않아도 그 효과를 모델에 반영할 수 있습니다. 이는 데이터의 중복성을 줄이고, 모델의 복잡성을 감소시키는 동시에 효율성을 높이는 방법입니다. 이를 통해, 모델은 기온과 습도 데이터만으로도 이슬점이 미치는 영향을 효과적으로 포착하고 예측할 수 있습니다.

2. 날짜, 공휴일, 계절과 자전거 이용량의 관계

날짜, 공휴일, 계절과 자전거 이용량 간의 관계를 분석하기 위해 아래 코드를 사용하여 데이터를 시각화하였습니다. 이 시각화를 통해 몇 가지 중요한 패턴을 확인할 수 있었습니다. 특히, 5월부터 7월, 그리고 9월부터 10월 사이에 자전거 이용 수요가 높은 것으로 나타났습니다. 반면, 1월부터 3월 사이에는 수요가 상대적으로 낮았습니다. 이러한 데이터 분석 결과는 계절적 요인이 자전거 이용 패턴에 영향을 미치고 있음을 시사합니다. 그러나 여름철 이용자 수의 변화를 더 자세히 살펴보면, 계절 자체보다는 기온이 자전거 이용량에 더 큰 영향을 미치고 있음을 알 수 있습니다. 또한, 공휴일의 유무와 자전거 이용량 사이의 관계는 상대적으로 미미한 것으로 나타났습니다. 이는 공휴일이 자전거 이용 패턴에 큰 영향을 미치지 않음을 의미합니다. 이러한 분석은 자전거 이용량 예측 모델을 개발하는 데 있어 중요한 기준이 될 수 있으며, 특히 기온과 같은 기후 요인을 중요한 변수로 고려해야 함을 시사합니다.

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 도표, 지도이(가) 표시된 사진

자동 생성된 설명

3. 강수량, 강설량과 자전거 이용량의 관계

강수량과 강설량이 자전거 이용량에 미치는 영향을 분석하기 위해, 아래 코드를 사용하여 데이터를 시각화했습니다. 눈이 온 날은 파란색, 비가 온 날은 빨간색으로, 그리고 맑은 날은 초록색으로 나타냈습니다. 시각화된 그래프를 살펴보면, X축은 날짜를, Y축은 자전거 이용자 수를 나타냅니다. 특히 주목할 점은 비가 온 날에 자전거 이용자 수가 현저히 감소하는 경향이 있다는 것입니다. 이는 비가 오는 날씨가 자전거 이용에 부정적인 영향을 미친다는 것을 시사합니다. 반면, 눈이 온 날과 눈이 오지 않은 날의 자전거 이용자 수 사이에는 큰 차이가 없는 것으로 나타났습니다. 이는 눈이 내리는 날씨가 자전거 이용 패턴에 큰 영향을 미치지 않음을 의미할 수 있습니다. 이러한 분석은 날씨 조건, 특히 강수량이 자전거 이용량 예측 모델에 중요한 변수로 고려되어야 함을 나타냅니다.

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

4. 습도, 햇빛량, 기온과 자전거 이용량의 관계

습도, 햇빛량, 기온과 자전거 이용량 간의 관계를 파악하기 위해 아래 코드를 사용하여 데이터를 시각화했습니다. 동그라미의 크기로 자전거 이용자 수를 나타내며, 색상은 날씨 조건을 표현합니다. 빨간색은 햇빛이 강한 날을, 파란색은 햇빛이 적은 날을 나타냅니다. 시각화 결과, 습도가 40%에서 60% 사이일 때와 기온이 10°C에서 25°C 사이일 때 자전거 이용자 수가 가장 높은 것으로 나타났습니다. 이는 이러한 기후 조건이 자전거 이용에 가장 적합하다는 것을 시사합니다. 따라서 습도와 기온은 자전거 이용량 예측 모델에서 중요한 변수로 고려될 수 있습니다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

스크린샷, 텍스트, 다채로움이(가) 표시된 사진

자동 생성된 설명

5. 시간, 날짜와 자전거 이용량의 관계

시간 및 날짜와 자전거 이용량 간의 상관관계를 분석하기 위해, 특정 코드를 사용하여 데이터를 시각화했습니다. 이 과정에서 시간 데이터는 정규화되어 그래프에 표현되었습니다. 시각화 결과 오전 8시와 오후 6시에 자전거 이용자 수가 가장 많은 것으로 나타났습니다. 이는 출퇴근 시간대에 자전거 이용이 증가한다는 것을 시사합니다. 또한, 그래프에 표시된 반투명한 부분은 날짜에 따른 이용량의 변동 폭을 나타내는데, 이 변동 폭이 상대적으로 작다는 점은 날짜보다는 시간이 자전거 이용량에 더 큰 영향을 미친다는 것을 보여줍니다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

그래프, 텍스트, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

데이터 분석을 통해 다음과 같은 중요한 사실들을 확인할 수 있었습니다

1. 이슬점: 이슬점은 온도와 강수량에 상대적인 데이터로 나타났습니다. 이는 이슬점이 온도와 습도의 함수로 결정되며, 이를 통해 온도와 습도 데이터만으로도 이슬점의 영향을 반영할 수 있다는 것을 의미합니다.

2. 계절 데이터: 계절은 날짜와 온도 데이터에 의해 결정되는 변수로 파악되었습니다. 이는 계절 데이터가 다른 변수들과 중복되는 정보를 담고 있음을 시사합니다.

3. 공휴일과 강설량: 공휴일과 강설량은 자전거 이용자 수와의 상관관계가 낮게 나타났습니다. 이는 이 두 변수가 자전거 이용 패턴 예측에 큰 영향을 미치지 않을 수 있음을 나타냅니다.

4. 시간과 자전거 이용자 수: 시간과 자전거 이용자 수 간에는 높은 상관관계가 관찰되었습니다. 이는 시간이 자전거 이용 패턴 예측에 중요한 요소임을 시사합니다.

텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

이러한 발견을 시각적으로 더 잘 이해하기 위해, 히트맵과 데이터들의 상관관계 막대그래프를 분석하였습니다.

텍스트, 스크린샷, 사각형, 패턴이(가) 표시된 사진

자동 생성된 설명

막대 그래프에서는 자전거 이용자 수와 높은 상관관계를 보이는 변수로 시간, 온도, 이슬점이 확인되었습니다. 그러나 이슬점은 이전 분석에서 온도와 습도에 상대적인 데이터로 밝혀졌기 때문에, 이슬점 대신 습도를 사용하기로 결정하였습니다. 따라서 최종적으로 독립변수로 시간, 온도, 습도를 선택하고, 종속변수로 자전거 이용자 수를 설정하였습니다.

**3.3 모델 성능 평가**

**텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명**

모델 성능 평가는 이론적 배경에서 언급된 세 가지 모델 - 선형회귀, 릿지회귀, 결정 트리 회귀 - 을 사용하여 모델들을 학습 시켰습니다. 이 모델들을 효과적으로 학습시키기 위해, 각 모델을 딕셔너리 구조에 할당하고 반복문을 사용하여 일괄적으로 학습시켰습니다. 이러한 접근 방식은 코드의 효율성을 높이고 모델 관리를 용이하게 만들었습니다. 모델의 성능 평가를 위해, 네 가지 주요 성능 지표를 사용하였습니다: 평균 제곱 백분율 오차(MSPE), 평균 절대 오차(MSAE), 결정 계수(R Square), 그리고 조정된 결정 계수(Adjusted R Square). 이 지표들은 모델의 예측 정확도와 일관성을 평가하는 데 중요한 역할을 하며, 각 모델의 성능을 객관적으로 비교하는 데 사용되었습니다.

**텍스트, 스크린샷이(가) 표시된 사진

자동 생성된 설명**

**4. 결론**

**4.1 분석 결과**

텍스트, 스크린샷, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명성능 지표 비교 결과, 결정 계수는 모든 모델에서 유사한 값을 보였으나, 평균 제곱 백분율 오차(MSPE)와 평균 절대 오차(MSAE)에서 결정 트리 회귀 모델이 가장 우수한 성능을 나타냈습니다. 이러한 결과를 바탕으로 결정 트리 회귀 모델을 서울시 공공자전거 이용자 수 예측에 가장 적합한 모델로 선정하였습니다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

결정 트리 회귀 모델을 사용하여 온도, 습도, 시간 등의 변수에 따른 자전거 이용량을 예측하는 모델을 구축하였습니다. 아래코드는 모델을 사용하여 입력 값을 온도가 18도, 습도가 40%, 시간이 오후 6시(18시)로 설정한 결과입니다. 그 결과 자전거 이용량이 약 3,245대로 추정되었습니다.

**4.2 기대효과 및 한계점**

서비스의 기대 효과는 다음과 같습니다. 첫째, 이용자 수 예측을 통해 사용자의 수요에 맞춰 자전거를 적절히 배치함으로써, 자전거 쏠림 현상을 완화할 수 있습니다. 이는 사용자 경험을 개선하고 서비스의 효율성을 높이는 데 기여할 것입니다. 둘째, 이러한 예측 시스템은 노동자들의 업무 효율을 증진시키고, 그들의 업무 부담을 경감시킬 것으로 기대됩니다.

현재 모델은 서울시 전체 자전거 이용자 수 데이터를 기반으로 하여 개발되었습니다. 이는 각 개별 대여소별 이용자 수 예측에는 직접 적용할 수 없는 한계를 가지고 있습니다. 그러나 본 프로젝트에서 사용된 방법론을 각 대여소의 자전거 이용 데이터에 적용한다면, 대여소별 이용자 수 예측 또한 가능할 것으로 보입니다.

**참고문헌**

1. [Earth.Org](http://Earth.Org) ‘2022의 가장 큰 환경문제 12가지’보고서 발표   
https://www.waterjournal.co.kr/news/articleView.html?idxno=62399

2. 연도별 따릉이 이용현황

https://news.mt.co.kr/mtview.php?no=2021103020523964746

3. 서일정 and 조재희. (2021). 코로나19 전후 서울시 공공 자전거 이용 패턴의 변화. 한국빅데이터학회 학회지, 6(2), 139-149.

4. 백지혜. (2021). 공공자전거 이용특성이 만족도에 미치는 영향분석. 한향대학교 석사학위논문 - https://hanyang.dcollection.net/public\_resource/pdf/200000485847\_20231126112245.pdf