**주차수요 예측 AI 경진대회**

**목표분석**

임대주택 설계 시의 주변환경에 따른 장래주차수요 예측으로 단지 내 적정 주차구역의 산정을 필요로 함.

**데이터**

1. Train.csv

학습을 위한 데이터(2596 \* 15)

컬럼 구성 : 단지코드, 총세대수, 임대건물구분, 지역, 공급유형, 전용면적, 전용면적별 세대수, 공가수, 자격유형, 임대보증금, 임대료, 도보 10분거리 내 지하철, 도보 10분 거리내 버스, 단지내 주차면수, 등록차량수

1. Test.csv

모델 테스트를 위한 데이터(1022 \* 14)

컬럼 구성 : 단지코드, 총세대수, 임대건물구분, 지역, 공급유형, 전용면적, 전용면적별 세대수, 공가수, 자격유형, 임대보증금, 임대료, 도보 10분거리 내 지하철, 도보 10분 거리내 버스, 단지내 주차면수

1. Age\_gender\_info.csv

각 지역별 연령별 인구분포 데이터(16 \* 23)

1. Sample\_submission.csv

테스트 데이터 평가를 위한 제출용 데이터

**둘러보기**

1. **결측치**

(train)

임대보증금 569

임대료 569

도보 10분거리 내 지하철(환승노선 수 반영) 211

도보 10분거리 내 버스정류장 수 4

* 임대보증금과 임대료는 같은 위치에서 결측치가 발생

버스정류장에 결측치가 존재하는 행은 지하철에도 결측치가 존재

* 임대보증금과 임대료의 결측치에서 공백문자 외 “-“형태의 결측치가 추가적으로 발견된다. 이를 포함하면 임대보증금은 581개, 임대료는 590의 행에서 결측치가 나타나는 모습을 보이므로 같은 위치에서 결측치가 발생하는 것이 부정된다.

(test)

자격유형 2

임대보증금 180

임대료 180

도보 10분거리 내 지하철(환승노선 수 반영) 42

도보 10분거리 내 버스정류장 수 0

* 임대보증금과 임대료는 같은 위치에서 결측치가 발생

자격유형은 빼도 안빼도 크게 영향을 미치지 않을 것으로 예상

* 지하철과 버스정류장의 결측값은 0으로 대체하여 진행, 임대보증금과 임대료는 0으로 대체하는 것을 고려해봐야 함

1. **컬럼별 영향치 확인**
2. 단지코드별로 고정된 값을 가진 컬럼들(연속형)

* 총세대수, 공가수, 지하철역수, 버스정류장수, 단지내주차면수

선형회귀 p-value

공가수(5.19e-12)와 단지내주차면수(<2e-16)으로 유의미한 설명변수로 나타나는 모습을 볼 수 있다. 총세대수(0.0189)는 0.05보다 낮으나 크게 의미를 갖지 못하는 설명변수로 볼 수 있고 지하철역과 버스정류장 수는 의미없는 변수로 볼 수 있다.

상관관계차트

단지내주차면수(0.85)와 총세대수(0.60) 순으로 높은 상관관계를 보여주고 의미있는 설명변수로 적합하다고 볼 수 있다. 공가수(0.17)로 다소 낮은 상관관계를 보여주나 유의미한 수치로 볼 수 있다. 그리고 지하철역과 버스정류장 수는 무의미한 낮은 상관관계를 가진다.

**평**

단지내주차면수와 총세대수는 유효한 설명변수로 볼 수 있다. 다만, 매우 유효한 선형회귀의 모습을 보여준 공가수는 유의미하지만 매우 낮은 상관관계의 모습을 띄는데 이는 고려해봐야 할 사항이다. 마지막으로 지하철역과 버스정류장 수는 설명변수로 매우 부적합하다고 판단할 수 있다.

1. 단지코드별로 고정된 값을 가진 컬럼들(범주형)

* 지역

Train 데이터의 지역데이터는 등록차량수에 영향을 미치는지 확인할 수 없다. 따라서, age\_gender\_info데이터를 살펴보자. 지역별 연령데이터는 지역의 각 연령대별 비율을 나타내는 데이터로 하나의 행의 숫자데이터들을 모두 합치면 1이된다.

분석 어려움 다음 챕터 진행

1. 단지코드별로 고정된 값을 가지지 않는 컬럼들(연속형)

* 전용면적, 전용면적별 세대수, 임대보증금, 임대료

같은 단지 코드를 가진 전용면적 중 같은 값을 가진 행들이 있다. 그런 경우는 세대수는 총합으로, 임대보증금과 임대료, 등록차량수는 평균값으로 구한다. (단, train데이터의 임대보증금과 임대료에는 결측치가 있으므로 이를 제거하고 진행한다.)

**평**

다중회귀방식으로 분석해보았을 때 전용면적별세대수, 임대보증금, 임대료가 유의미한 수치로 볼 수 있는데 이 모델의 설명률은 0.2로 낮은 모습을 보여주고 데이터의 구조로 보았을 때 전용면적과 전용면적별세대수는 단편적으로 보았을 때 등록차량수에 영향을 미칠만한 요소가 아니라고 판단할 수 있다. 다만, 선형회귀분석과 상관계수의 값을 보았을 때 임대료는 등록차량수에 영향을 주는 컬럼으로 볼 수 있다고 판단할 수 있다.

1. 단지코드별로 고정된 값을 가지지 않는 컬럼들(범주형)

* 임대건물구분, 공급유형, 신분(주차수요 예측 EDA & Catboost Baseline참고

URL : https://dacon.io/competitions/official/235745/codeshare/2851?page=1&dtype=recent)

상가의 전용면적별세대수는 모두 1, 상가와 아파트별 주차수요가 다르고 두개가 공존하는 주상복합인 경우 기존 수치에 비해 더 많은 수요가 발생할 수 있다.

세가지 변수에 대해서는 다른 자료 참고가 필요할 듯

이상 각 컬럼별 영향치분석을 끝내고 설명변수가 각각 가지고 있는 의미에 대해 생각해보자

1. **각 컬럼별 의미확인 (train.csv 기준)**
2. 단지코드

단지별 식별을 위한 정보

423개의 다른 단지코드

1. 총세대수(단지코드별)

단지별로 총세대수를 나타내는 정보

등록차량수와 회귀분석했을 때 크게 영향을 미치지는 않지만 유의미한 결과값이라고 판단할 수 있다. 또, 상관관계가 0.6의 다소 높은 수치가 나온 것을 확인할 수 있다.

* 총세대수가 많을수록 등록차량수가 많아진다고 판단할 수 있다.

1. 임대건물구분(단지코드별)

단지코드별로 아파트인지 상가인지 구분해주는 정보, 주상복합인 경우도 있다.

423개의 단지 중 상가로만 이루어져있는 단지는 없고 390개의 단지가 아파트만으로 이루어져있으며 33개의 단지는 주상복합이다. 이 때 주상복합과 그렇지 않은 단지의 등록차량수를 점 그래프로 비교했을 때 단지 내 상가비율이 높다고 해서 등록차량수가 많다고 판단할 수 없다.

1. 지역

지역에 관한 판단은 미루자.

1. 공급유형

각 행의 객체가 어떤 방식으로 임대되었는지 나타내는 정보

공공임대(분납), 공공임대(5년), 공공임대(10년), 공공임대(50년), 국민임대, 영구임대, 임대상가, 장기전세, 공공분양, 행복주택으로 나눠지는데 국민임대 1727건, 임대상가 562건, 행복주택 202건, 공공임대(10년) 175건, 영구임대 152건으로 상당 수가 국민임대와 임대상가로 치우친 모습을 나타낸다. 이는 test셋에서도 같은 모습을 보인다. 따라서, 국민임대와 임대상가, 그 외 세 계층정도로 나누어 구별하는 것이 좋다고 판단됨.

1. 전용면적, 전용면적별세대수

각 행의 객체의 전용면적과 세대수를 나타내는 정보

임대건물구분이 상가인 경우, 각 상가를 전용면적별세대수를 1로하여 전용면적에 상관없이 서로 다른 행으로 분리해 데이터를 정리해놓은 모습을 볼 수 있다. 전용면적과 전용면적별세대수를 곱해 단지내의 세대총면적을 이용해 데이터를 비교해보았을 때, 세대총면적은 총세대수, 단지내주차면수, 등록차량수와 모두 높은 상관관계와 비슷하게 수렴하는 선형방정식이 존재한다. 따라서, 세대총면적은 등록차량수를 유추하는데 중요한 변수로서 판단할 수 있다.

1. 전용면적별세대수

-

1. 공가수(단지코드별)

단지 내 비어있는 세대의 총량을 나타내는 정보

선정된 단지내주차면수와 등록차량수에는 차이가 존재한다. 이는 단지내주차면수가 잘못선정된 것을 의미할 수도 있지만 공가수도 차이에 영향을 줄 수 있다. 그러나 단지내주차면수와 등록차량수간의 차이의 분산은 43202.66인 반면 공가수의 분산으 106.6143으로 오차가 공가수의 영향을 무시해도 된다고 판단할 수 있다.

1. 신분

각 행 객체의 임대자 신분을 나타내는 정보

공급유형과 비슷한 데이터 분포의 모습을 볼 수 있다. A유형 1756개, D유형 569개와 이외 다른 유형들에 분포되어있는 모습인데, 임대건물구분이 상가인 경우 모두 D유형이고 ‘C1350’단지만이 아파트임에도 D유형이다.

1. 임대보증금, 임대료

각 행 객체의 임대보증금과 임대료를 나타내는 정보

임대건물구분이 상가인 경우 임대보증금과 임대료가 모두 결측치이고 이외 ‘C1350’단지만 아파트임에도 임대보증금과 임대료가 존재하지 않는다. 임대보증금과 임대료는 “”와 “-“의 결측치가 발생한다. 이는 거의 같은 위치에서 결측치가 발생하는데 임대보증금이 “”인 행에서는 임대료도 “”이나 임대보증금이 “-“인 경우 임대료도 “-“지만 임대료만이 “-“인 경우가 9개가 더 존재한다. 임대보증금과 임대료 모두 등록차량수와 유의미한 p-value값을 가지지만 임대료가 설명력이 훨씬 높은 것을 볼 수 있다. 하지만, 등록차량수 예측에 주요한 변수로 자리잡긴 어렵다는 판단이다.

1. 임대료

-

1. 지하철역수, 버스정류장수(단지코드별)

단지마다 도보 10분 거리 내 지하철역 수(환승노선 수 반영)를 나타내는 정보

Train셋 지하철역수 211개, 버스정류장수 4개

Test셋 지하철역수 38개, 버스정류장수 0개의 결측치를 갖고 있다.

버스정류장수가 결측치 인경우 지하철역수도 결측치를 가지는 모습을 띈다.

지하철역수는 train셋 최대 3개, test셋 최대 2개이고 상당수가 0개인 경우이고 1개를 초과하는 경우가 매우 적다. 버스정류장수는 train셋 최대 20개, test셋 최대 50개(그 전 19개)로 지하철역수에 비해 다양한 스펙트럼의 수치를 갖고 있다. 또한, 중위값 근처의 값들에 골고루 분포되어있는 모습을 볼 수 있다.

선형회귀식을 이용해 지하철역수와 버스정류장수가 등록차량수에 영향을 미치는지 확인해보면 두 설명변수 모두 무의미한 값이 나온다. 따라서, 지하철역수와 버스정류장수는 참고사항으로 접근할 뿐 등록차량수에 영향을 미치는 변수로 판단할 수 없으므로 0으로 대체하는 것이 좋아보인다.

1. 버스정류장수

-

1. 단지내주차면수(단지코드별)

단지 내 이미 산정된 단지내주차면수를 나타내는 정보

기존의 방식으로 단지에 필요한 주차면수를 계산하여 실제 단지내에 존재하는 주차면수를 뜻한다. 이론적으로는 공가수가 없을 경우 단지내주차면수와 등록차량수가 근사한 모습을 띄어야한다. 하지만, (단지내주차면수-등록차량수)는 양수와 음수, 표준편차 200으로 크기도 크고 다양하게 오차가 발생한다. 따라서, 공가수는 (단지내주차면수-등록차량수)의 오차를 증명할 수 있는 수단으로 볼 수 없다. 그러나, 모든 설명변수들 중 단지내주차면수가 등록차량수를 가장 잘 설명하는 변수로 볼 수 있다.

1. 등록차량수

목적변수