**Background**

통계청 2015년 자료에 의하면 ([https://bit.ly/2SFyzMA)](https://bit.ly/2SFyzMA)

일반적인 한국인의 절반은 48.1%는 아파트에 살고 있습니다.

그들은 아파트 주거 선호도가 매우 높습니다.

또한 부의 증식 수단으로 생각 하기 때문에 아파트 가격에 관심이 많습니다.

이번 대회의 데이터 제공자는 직방입니다.

직방은 부동산 정보의 비대칭성과 불투명성을 해소하기 위해 노력하며,

중개사와 구매자를 연결하여  부동산정보 서비스 시장의 신뢰도를 높이는데 기여합니다.

최근 매물 가격 정보는 직방, 다음부동산, 네이버부동산에서 볼 수 있습니다.

하지만 최근 매물 가격은 아직 거래되지 않아 정확하지 않은 정보 일 수 있습니다.

이에따라, 본 대회는 아파트 구매자들의 비대칭성 정보를 해결하기 위해 미래의 실 거래가 예측을 목표로 합니다.

**Description**

서울/부산 지역 아파트 실 거래가를 예측하는 모델을 만들어주세요.

실 거래가와 아파트, 학교, 지하철역 정보를 제공합니다.

**Validation**

약1,600,000여개의 실거래 데이터, 아파트 거래일, 지역, 전용면적, 공급면적 등의 정보가 제공됩니다.

\* 국토교통부 실거래가 공개시스템 (<http://rt.molit.go.kr/>)과 같은 법적인 제약이 없는 외부 데이터(공공 데이터) 사용이 가능합니다.

제출한 예측 결과물은 [Root Mean Squared Error(RMSE)](https://en.wikipedia.org/wiki/Root-mean-square_deviation)로 평가합니다(평가는 게시된 데이터셋만을 기준으로 합니다).

스코어가 상대적으로 낮은 참가자가 수상 후보가 되며 동점자~~는~~ 발생시 가장 최신 순으로 순위를 정합니다.

모델 검증 후 순위는 바뀔 수도 있습니다.

후보 선정 시 주석(한글, 영문)이 포함된 코드를 전달해야 합니다.

상위 5명에게서 코드를 받고 합격자가 없을 시 평가와 대회는 마감됩니다 .

코드 검증 절차(모델 검증)

1.        Pre-Processing for Data Cleaning

2.        Feature Engineering and Variable Selection

3.        Model Selection and Regularization

4.        Optimization Processing

준수 사항

1.        코드 제출시 확장자가 .R 이나 .py인 스크립트 파일만 제출

2.        전체 프로세스를 일목요연하게 정리하여 주석을 포함하여 하나의 파일로 제출(주석은 분석 과정 중 각 단계와 결과를 충분히 설명하여야 함)

3.        모든 코드는 에러 없이 실행되어야 함(패키지 또는 라이브러리 설치 코드는 제외 하여도 되나 라이브러리 로딩 코드는 포함되어야 함)

4.        코드와 주석의 인코딩은 모두 UTF-8을 사용하여야 함

\* 상금은 환율에 따라서 변동될 수 있습니다.

**[Files]**

1. train.csv : 서울/부산 지역의 1,600,000여개 거래 데이터, 아파트 거래일(약 10일 간격), 지역, 전용면적, 공급면적, 실 거래가 등의 정보

2. test.csv : 실 거래가를 제외하고 train.csv와 동일

3. schools.csv : 서울/부산 지역의 초,중,고교에 대한 정보

4. subways.csv : 서울/부산 지역의 지하철역에 대한 정보

**[Data fields]**

        train.csv

|  |  |
| --- | --- |
| key | Primary key |
| apartment\_id | 아파트 ID |
| city | Seoul: 1 Busan: 0 |
| transaction\_year\_month | 실거래가 발생 년월 : yyyymm |
| transaction\_date | 실거래가 발생일 dd ~ dd |
| year\_of\_completion | 아파트 준공 년도, 입주년 |
| exclusive\_use\_area | 전용면적 |
| floor | 층 |
| latitude | latitude |
| longitude | longitude |
| address\_ by\_law | 법정동코드 (앞2자리: 시/도, 앞5자리: 시/군/구, 앞8자리: 읍/면/동)  법정동 코드 (10자리) : 법적주소  시도(2)+시군구(3)+읍면동(3)+리(2)    보다 자세한 법정동 코드체계는 아래 행정표준코드관리시스템 > 법정동코드목록조회에서 찾아보실 수 있습니다.  <https://www.code.go.kr/stdcode/regCodeL.do> |
| total\_parking\_capacity\_in\_site | 아파트 단지내 총 주차대수 |
| total\_household\_count\_in\_sites | 아파트 단지 내 총세대수 |
| apartment\_building\_count\_in\_sites | 아파트 단지 내 총 동수 |
| tallest\_building\_in\_sites | 아파트 단지 내 최고층 |
| lowest\_building\_in\_sites | 아파트 단지 내 최저층 동의 층수 |
| heat\_type | 난방방식 : 개별난방(individual), 중앙난방(central), 지역난방(district) |
| heat\_fuel | 난방연료 : 도시가스(gas), 열병합(cogeneration) |
| room\_id | 평형 ID  (평형 =(전용면적㎡+ 주거공용면적㎡)\*0.3025))    (1평형=3.3058㎡) |
| supply\_area | 공급면적 (= 전용면적 + 주거공용면적) |
| total\_household\_count\_of\_area\_type | 평형의 총세대수 |
| room\_count | 방 수 |
| bathroom\_count | 욕실 수 |
| front\_door\_structure | 현관구조 : 복도식(corridor), 계단식(stairway), 복합식(mixed) |
| transaction\_real\_price | 실거래가 (단위: 만원) |

subways.csv

|  |  |
| --- | --- |
| station\_id | 지하철역 ID |
| latitude | latitude |
| longitude | longitude |
| subway\_line | 지하철 노선 서울 : 1 ~ 9 (1호선 ~ 9호선), KJ (경의중앙선), US (우의신설경전철), AP (공항철도), KC (경춘선), DL (분당선), ND (신분당선)    부산 : B1 ~ B4 (1호선 ~ 4호선), BD (동해선), BK (부산김해경전철)    복수일 경우 : 환승역 |
| address\_by\_law | 법정동코드 (앞2자리: 시/도, 앞5자리: 시/군/구, 앞8자리: 읍/면/동) |

schools.csv

|  |  |
| --- | --- |
| school\_code | 학교 코드 |
| latitude | latitude |
| longitude | longitude |
| school\_class | 초등학교(elementary school), 중학교( middle school), 고등학교(high school) |
| operation\_type | 운영주체 : 국립(national), 공립(public), 사립(private) |
| highschool\_type | 고등학교 구분 : 일반고등학교 (general), 자율고등학교 ( autonomous), 특성화고등학교 (specialized), 특수목적고등학교 (objective) |
| gender | 남학교(male), 여학교(female), 남녀공학(both = coeducation) |
| foundation\_date | 설립일 |
| address\_by\_law | 법정동코드 (앞2자리: 시/도, 앞5자리: 시/군/구, 앞8자리: 읍/면/동) |