**□ 모듈프로젝트 산출물**

클라우드 AI 융합 전문가 양성 과정

|  |
| --- |
| 모듈 프로젝트 3   * **주차수요 예측 AI 경진대회 -**   4조 : 배상현, 박준원, 민병석, 김서연 |

2021년 9월 27일

클라우드 AI 융합전문가

이름 : 박준원

|  |
| --- |
| **목 차**  1. 전처리 주요 코드  1) 임대건물구분  2) 주차규정  3) 임대료, 임대보증금  4) 대중교통  5) 지역 예상 보유수  2. 적용 알고리즘 및 결과 |

**1. 전처리 주요 코드**

|  |
| --- |
| https://lh6.googleusercontent.com/xPrvWuWY5JnNn9SCmeCV0topYVzdKwRE62VrgzGRz5AwLizEwjF_JgD2DJvi-ko9zTaBb1OEPkn0kIncrO5P_Ia2kMMKo_pYUrwH7IzcZ4P3dqaiDQv5c-3cdhsdfRsL6KkfhHwVr-Q=s0  [표1] 주어진 데이터셋 데이터프레임   1. **임대건물구분**   **C:\Users\user\Desktop\1_임대건물구분.PNG**  [주요코드1] 임대건물구분 컬럼 수치화  동일 단지코드 내에서 상가는 항상 아파트와 함께 존재하므로 이를 주상복합으로 보고 아파트=0, 주상복합=1로 수치화를 진행했습니다.   1. **주차 규정**   https://lh3.googleusercontent.com/pGmrAgPe2ALhRArNiK-5rLlaWCvlYHXMkOthfJ1U4tEws-FuXD6fbyJd9ATszuFD4ApF94y9TzXv92Pk5AvJk1TbC1wjWRN46jhotGtKnPIc2fQp2vB3m9zwxZIteOq3N6mxRuHD2kg=s0  [표2] 주차장 설치기준 자료  C:\Users\user\Desktop\2_주차규정.PNG  [주요코드2] 주차규정에 따른 예상등록수 컬럼 생성  주차 규정에 따라 지역과 전용면적 별로 기준치를 따로 적용하고 세대수를 이용해 단지코드 마다 예상되는 예상등록수를 계산하여 컬럼을 생성하였습니다.   1. **임대보증금, 임대료**   C:\Users\user\Desktop\3_임대보증금.PNG  [주요코드3] 임대보증금, 임대료 컬럼 전처리  임대보증금과 임대료 컬럼을 처리하는데 문제가 되었던 것은 세가지 입니다. 결측치 처리와 데이터 타입 변경은 수월하게 처리할 수 있었지만, 가장 문제가 되는 것은 동마다 다른 금액 수치를 단지 하나로 통합할 때 어떻게 처리할 것인가 였습니다. 이 두 값을 행별 전용면적으로 나누어 기준을 맞추고 평균값을 구하여 처리했습니다.   1. **대중교통**   C:\Users\user\Desktop\4_대중교통.PNG  [주요코드4] 세대수/대중교통 컬럼 생성  기존의 지하철역 수, 버스역 수 를 대중교통이란 하나의 컬럼으로 합치고 실세대수에서 대중교통을 나누었습니다. 이는 대중교통 역당 세대수 비중을 나타내는 새로운 지표가 됩니다.   1. **지역 예상 보유수**   C:\Users\user\Desktop\5_지역예상보유수.PNG  [주요코드5] 지역\_예상보유수 컬럼 생성  https://lh6.googleusercontent.com/pV4qHFFJx3V9T137WG7IMHKF1y4n67_q5544YTrm0ieeBtuTKQjNIUfO2Rl5NaxoMbVqeEZO8opJMeS7TTbUQ9dQ6nPQ9vl0gQkqImnAkQxGTeOAsFUXNYPdpyS7P_RQkGxmLzuTcXA=s0  [그래프1] 지역별 연령별 차보유비율 시각화  Age\_gender\_info의 데이터셋을 통해 지역별 연령, 성별 비율을 알 수 있었고 area\_age\_car 데이터셋을 통해 지역별 연령에 따른 자동차 보유율을 알 수 있었습니다. 이에 두 데이터의 내적곱으로 자동차 보유 확률을 구하고 여기에 실세대수를 곱하여 ‘지역 예상보유수’ 컬럼을 생성하였습니다.   1. **전처리 결과**     [표3] 전처리 결과 데이터셋 예시  C:\Users\user\Desktop\상관관계.png  [그래프2] 생성 컬럼들의 관계지수 시각화 |

**2. 적용 알고리즘 및 결과**

|  |
| --- |
| 1. **적용한 회귀 알고리즘**  * **Pycaret** * AutoML을 하게 해주는 파이썬 라이브러리 * scikit-learn 패키지를 기반, Classification, Regression 등 다양한 모델을 지원 * **Random Forest Regressor** * 다중 의사결정 트리를 결합하여 tree forest를 생성 * 군집 분석, 범주형 숫자형에도 잘 작동, 결측치가 있어도 잘 작동 * **XGB Regressor** * 성능이 낮은 예측 모형들의 학습 에러에 가중치를 두고 순차적으로 다음 학습에 반영 * **CatBoost Regressor** * target leakage 문제와 범주형 변수 처리 문제를 Ordered boosting과, 새로운 범주형 변수 처리 방법으로 해결 * 데이터의 대부분이 수치형인 경우에는 적합하지 못하다.  1. **적용 및 결과**  * **Pycaret** : MAE  =  113.2383 * **RF Regressor** : MAE  =  164.8272 * **XGB Regressor**  : MAE  =  154.7512 * **CatBoost**  : MAE  =  97.4411   CatBoost모델에서 earlystopping을 적용해 과적합을 방지하고 돌렸을 때, 다른 모델에 비해 가장 낮은 수치로 좋은 결과를 냈습니다. 전반적으로 만족스러운 MAE가 나오지 못했는데, 이는 제공된 데이터에 오류가 많고 신뢰성이 떨어지는 문제가 어느정도 반영된 결과라고 추측됩니다.   1. **Feature Importance**   C:\Users\user\Desktop\catboost_피처임포턴스.png  [그래프3] 회귀모델 적용 이후 컬럼 중요도 시각화 |

|  |  |
| --- | --- |
| 성명 | 프로젝트 후 소감 |
| 박준원 | 주차수요 예측 모델을 만들기란 쉽지 않았습니다. 도메인이 없는 상태로 시작하는 프로젝트라서 더 어렵게 다가왔었던 것 같습니다. 데이콘 측에서 제공한 데이터는 몇 가지 오류가 있어 이것을 수정했어야 했고 특정 컬럼의 의미를 숨겨놔서 처리해야 할 점도 많았습니다. 팀원들과 각 feature에 대한 다양한 가설을 수립하고 타당한지 확인 작업을 통해 결측치와 이상치를 수정하는 과정은 시간은 오래 걸렸지만, 마지막에 전치리가 완성된 데이터를 만들었을 땐 성취감을 느낄 수 있었습니다. 이후 다양한 모델을 만들고 parameter를 수정하며 다양한 모델의 특징을 파악할 수 있었고, 이들을 선택하는 과정 역시 중요하다는 것을 배울 수 있었습니다. |