[서식1] 제출양식

「2022 공공데이터 활용 아이디어 경진대회」
참가 신청 제출양식

1 신청자 정보

이름	정재현	참가경로	숭실대학교 편시스템
학과(부)	산업정보시스템공학과	학번	
휴대폰		이메일	

2 아이디어 기획서

아이디어명 : 머신러닝/딥러닝을 이용한 공공임대주택 입주의향 예측

■ 아이디어 제안 배경 및 필요성

주거환경과 주택가격은 국민의 주거에 막대한 영향을 미치며, 특히 주택의 가격은 다양한 요인들에 의하여 결정됩니다. 그 요인 중 하나는 정부의 부동산 정책입니다. 대표적으로 주거복지정책 중 하나인 공공임대주택 공급이 있습니다. 공공임대주택 정책은 저소득층 주거 불안정 문제를 해결하고, 국민의 주거비 부담을 줄이며 주택시장 안정화 등의 목표를 가지고 있습니다. 과거에는 해당 정책이 서민층에서 주로 각광을 받았으나, 시간이 지남에 따라 중산층에서도 임대주택 거주에 대한 높은 의향을 보여왔습니다. 2020년 10월 28일, 매거진한경, 김영은 기자에 의하면 한경비즈니스가 시장 조사 마크로밀엠브레인을 통해 서울·경기와 5대 광역시에 거주하는 만 20~69세 국민 1,000명을 대상으로 조사한 결과 '중산층을 위한 20~30평대 공공임대주택이 필요하다'는 데는 73.4%가 동의하였으며, '중산층을 위한 공공임대주택 거주 의향'을 묻는 질문에는 71%가 거주 의향이 있다고 답했습니다. 실제로 결측치와 이상치를 제거한 '2020년 주거실태조사'의 데이터 분석 결과 적지 않은 수의 중산층에서도 공공임대주택에 입주할 의향이 있음을 확인할 수 있었습니다. 상대적으로 집값이 비싼 수도권(서울, 경

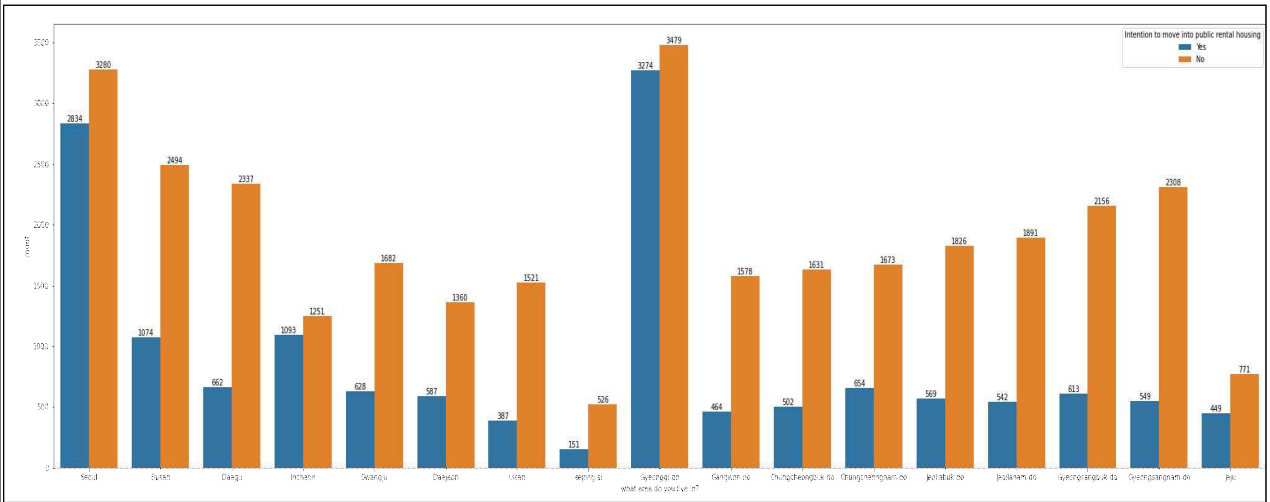


Fig. 1. 시도별 공공임대주택 입주의향 시각화

기, 인천)의 경우 Fig. 1과 같이 다른 지역에 비해 입주의향이 높게 나타났으며, 2020년 보건복지부에서 발표한 가구원 수에 따른 기준 중위소득을 바탕으로 OECD에서 정의하고 있는 계층별 중위소득 기

준에 따라 기준 중위소득 75%에서 200%까지의 계층을 중산층, 75% 미만을 서민층, 200%를 초과한 계층을 상류층으로 분류했을 때 서울시의 경우 중산층의 약 44%, 경기도의 경우 중산층의 약 52%, 인천시의 경우 약 41% 정도가 공공임대주택 입주의향이 있음을 Fig. 2 와 Fig. 3, Fig. 4에서 확인할 수 있습니다.

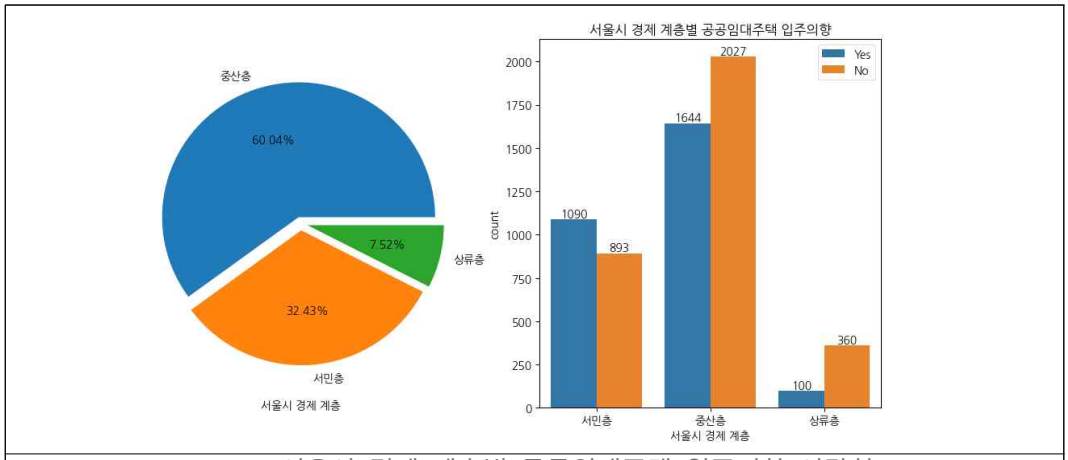


Fig. 2. 서울시 경제 계층별 공공임대주택 입주의향 시각화

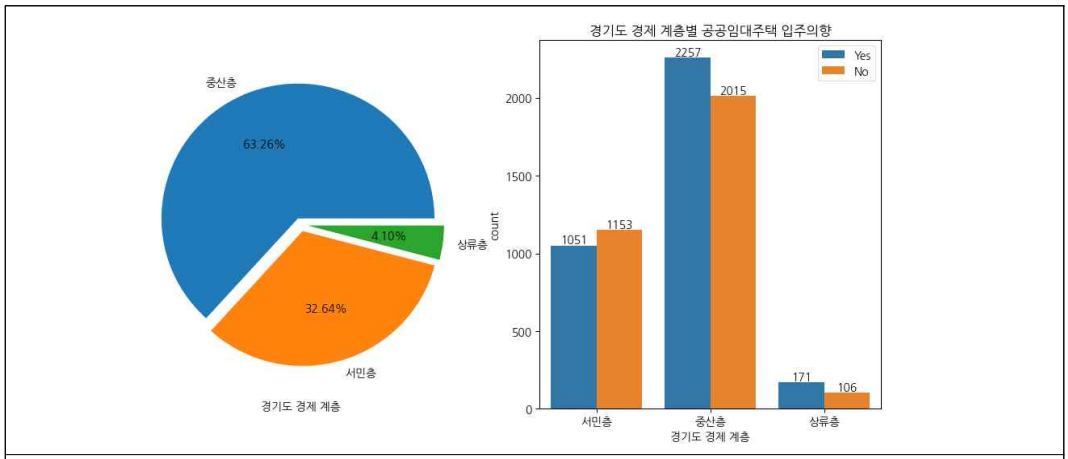


Fig. 3. 경기도 경제 계층별 공공임대주택 입주의향 시각화

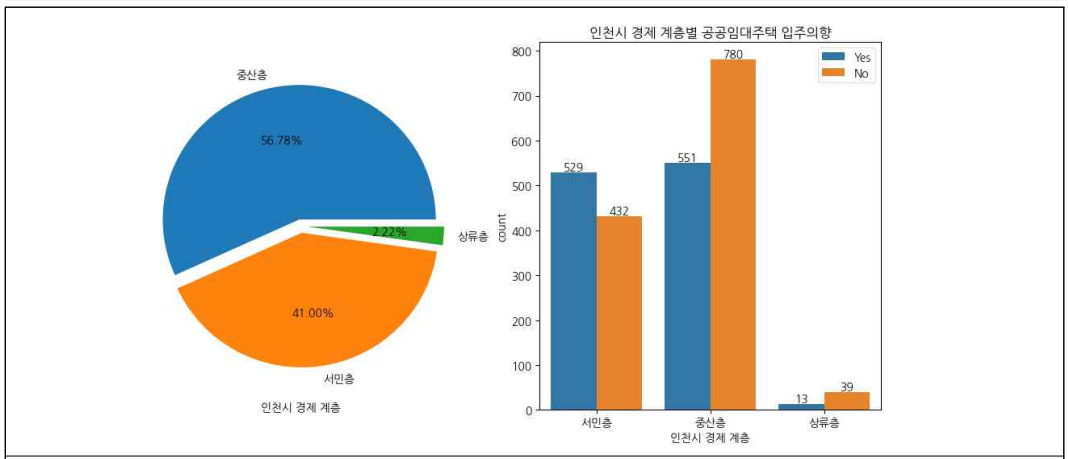


Fig. 4. 인천시 경제 계층별 공공임대주택 입주의향 시각화

반면 비수도권(서울, 경기, 인천을 제외한 14개 시·도)의 경우 중산층에서 약 76%가 공공임대주택 입주의향이 있음을 Fig. 5에서 확인할 수 있습니다. 그리고 자산을 기준으로 수도권과 비수도권으로 나누어 공공임대주택 입주의향을 확인한 결과 상위 소득 30% 미만의 그룹에서 입주의향이 높음을 확인할 수 있었습니다. 2021년 12월 6일, 동아일보, 신지환 기자에 의하면 우리금융경영연구소가 내놓은 '팬데믹 시대의 대중부유층 보고서'에서 2020년 중산층과 부유층 사이에 해당하는 소득 상위 10~30%

인 '대중부유층'의 총자산(부동산 자산, 금융자산, 기타자산)은 7억 6473만 원이라고 합니다. 이 금액을 기준으로 분류했을 때 수도권의 경우 소득 상위 30% 미만의 그룹에서 약 51% 정도가 Fig. 6과 같이 공공임대주택 입주의향이 있으며, 비수도권의 경우 소득 상위 30% 미만의 그룹에서 약 74% 정도가 입주의향이 있음을 Fig. 7에서 확인할 수 있습니다.

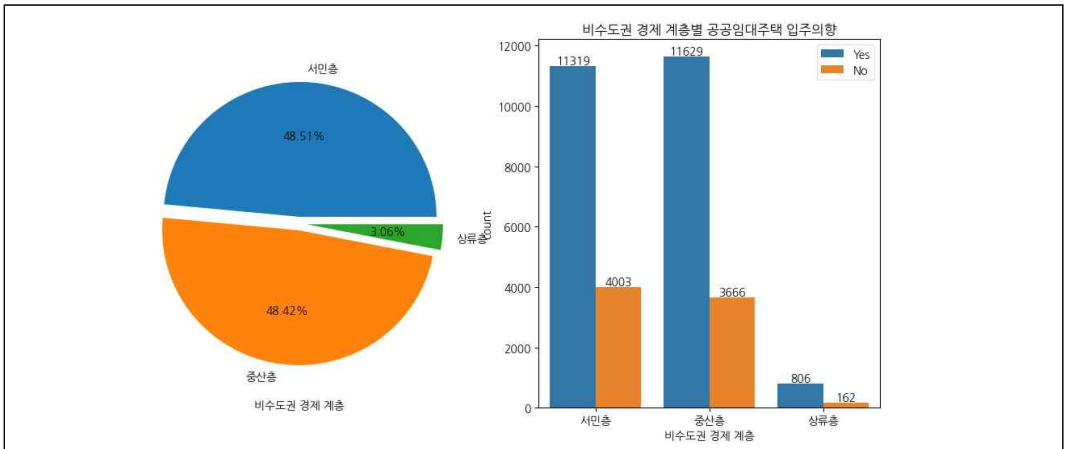


Fig. 5. 비수도권 경제 계층별 공공임대주택 입주의향 시각화

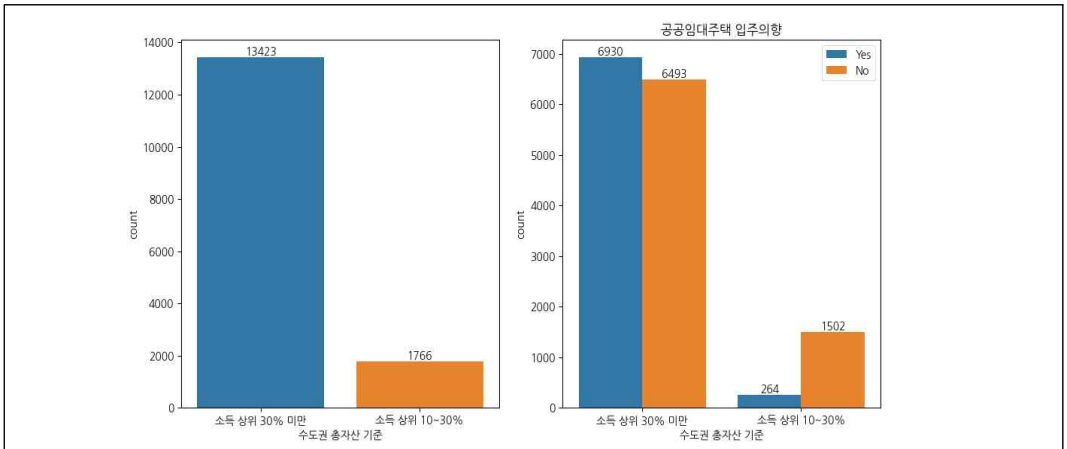


Fig. 6. 2020년 소득 상위 30% 총자산 기준
수도권 공공임대주택 입주의향 시각화

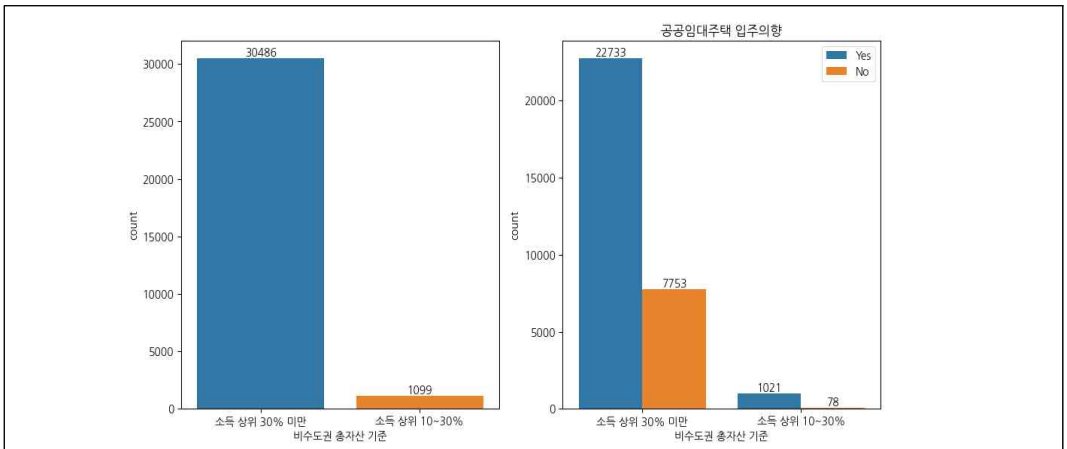


Fig. 7. 2020년 소득 상위 30% 총자산 기준
비수도권 공공임대주택 입주의향 시각화

이처럼 중산층에서도 공공임대주택 입주를 희망하고 있는 것을 확인할 수 있습니다. 또한 현재 중산층까지 정책을 확대하고 있는 시점이므로 공공임대주택 입주의향을 파악할 필요가 있습니다. 그러므로 국민의 인구통계학적 정보와 주택과 관련된 정보를 반영하여 다양한 집단에서 공공임대주택 입주

를 희망하는지 예측할 수 있다면 공공임대주택 공급 확대와 관련하여 정책적인 방향 제시에 도움이 될 것입니다. 예측을 위해 사용하는 데이터는 결측치와 이상치를 제거한 총 46,796가구의 '2020년 주거실태조사' 결과입니다. 주거실태조사는 우리나라 주택 및 주거정책을 수립하기 위한 기초자료를 작성하고자 주거기본법 제20조에 의거하여 2006년부터 국토교통부와 국토연구원이 공동으로 실시하고 있으므로 데이터를 지속해서 확보할 수 있습니다. 국토교통부와 국토연구원에서는 '2020년도 주거실태조사(통계 보고서)', '국토 이슈리포트' 등 주거실태조사를 통해 축적된 데이터를 기반으로 주거 이동 및 주택 보유 의식, 주거지원 정책 수요 및 평가, 맞춤형 주거지원 등 주거 관련 현황 및 주요 지표의 시계열적 변화를 분석하고 정책 효과 파악 및 새로운 정책방향 도출에 활용하고 있습니다.

■ 아이디어 기획 핵심내용

아이디어는 46,796개의 설문 응답자의 데이터를 기반으로 공공임대주택 입주의향 관련 연구에서 사용한 독립변수를 기반으로 월 소득, 자산 그리고 응답자의 심리 등을 물어보는 정보를 추가해 총 38개를 독립변수로 사용하고, 공공임대주택 입주의향(0,1)을 종속변수로 설정합니다. 먼저 extra tree classifier, XGBoost, Gradient Boosting Method(GBM)와 같은 머신러닝 기법을 통해 공공임대주택 입주의향 예측 모델을 구축하고 모델의 성능을 비교·분석하여 가장 적합한 모델을 찾습니다. Fig. 8은 average accuracy(%), average f1-score(%), average recall(%), average precision(%)을 비교한 것입니다.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.8000	0.8501	0.6092	0.7205	0.6600	0.5198	0.5236	0.5880
rf	Random Forest Classifier	0.7992	0.8493	0.5969	0.7248	0.6545	0.5149	0.5198	1.1090
et	Extra Trees Classifier	0.7987	0.8470	0.6016	0.7209	0.6557	0.5152	0.5195	1.4910
xgboost	Extreme Gradient Boosting	0.7984	0.8460	0.6163	0.7124	0.6607	0.5184	0.5213	0.3420
gbc	Gradient Boosting Classifier	0.7979	0.8413	0.6154	0.7117	0.6599	0.5173	0.5202	1.6710
lr	Logistic Regression	0.7920	0.8335	0.5866	0.7105	0.6425	0.4977	0.5024	2.6150
ridge	Ridge Classifier	0.7911	0.8319	0.5937	0.7047	0.6444	0.4981	0.5018	0.0970
ada	Ada Boost Classifier	0.7911	0.8322	0.5964	0.7034	0.6453	0.4987	0.5023	0.6860
lda	Linear Discriminant Analysis	0.7908	0.8319	0.6072	0.6974	0.6490	0.5010	0.5035	0.2330
knn	K Neighbors Classifier	0.7700	0.7750	0.5193	0.6833	0.5900	0.4343	0.4423	0.3590
svm	SVM - Linear Kernel	0.7671	0.7637	0.5459	0.6732	0.5983	0.4376	0.4455	0.3390
nb	Naive Bayes	0.7650	0.7957	0.6751	0.6208	0.6468	0.4711	0.4721	0.1070
dt	Decision Tree Classifier	0.7196	0.6811	0.5747	0.5585	0.5665	0.3593	0.3595	0.2290
dummy	Dummy Classifier	0.6812	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0750
catboost	CatBoost Classifier	0.6453	0.6852	0.4989	0.5842	0.5381	0.4293	0.4321	3.9670
qda	Quadratic Discriminant Analysis	0.5509	0.7960	0.9193	0.4091	0.5662	0.2237	0.3079	0.1790

Fig. 8. Model Performance 비교

Light Gradient Boosting Method(LGBM)에서 average accuracy와 f1-score의 값이 가장 높았으며, average precision의 값은 random forest, average recall의 값은 XGBoost에서 가장 높았습니다. 최종적으로 모델들의 성능을 비교했을 때 정확도가 높은 LGBM 모델을 입주의향 예측 모델로 선정하고 tune_model( ) 메서드를 이용해 튜닝된 LGBM 모델을 사용했습니다. 다음으로 딥러닝을 이용해서 공공임대주택 입주의향 가능성 예측을 위한 모델을 구현한 방법은 다음과 같습니다. 은닉층은 활성화 함수로 relu를 사용하고 마지막 층은 확률을 출력하기 위해서 시그모이드 함수를 사용합니다. 그리고 과적합을 방지하기 위해서 early stooping 함수를 사용합니다. 모델의 결과로 입력값에 대한 공공임대주택 입주의향에 대한 확률값이 나오게 되는데, 로지스틱 회귀분석의 ROC 곡선을 활용해서 임곗값을 구하여 입주의향 확률이 임곗값 이상이면 공공임대주택 입주의향이 1='있다', 임곗값보다 작으면 0='없다'로 분류할 수 있는 최종적인 모델을 만들었습니다. 최적의 컷오프 포인트는 참 양성률이 높고,

거짓 양성률이 낮은 곳입니다. 이 논리를 바탕으로 최적의 임곗값을 구하기 위해 다음의 함수를 사용합니다. $p^* = \arg_p \min |TPR(p) + FPR(p) - 1|$, 이 함수는 p가 임곗값일 때 'TPR(p)=1-FPR(p)'이라는 식을 이용한 것입니다. 최적의 컷오프 포인트는 Fig. 9과 같이 0.32500604297555484이므로 이보다 큰 값은 1, 낮은 값은 0으로 분류할 수 있습니다. 하지만 예측 모델을 만들기 위하여 사용한 데이터 세트가 다소 불균형하므로, 정확도를 최대화해서 임곗값을 찾는 방법으로 balanced_accuracy_score를 이용하였으며 결과는 Fig. 10과 같습니다. 최종적으로 두 번째로 구한 임곗값을 사용하여 공공임대주택 입주의향 분류에 사용했습니다. 예를 들어 2020년 주거실태조사 결과 입주의향이 있는 응답자의 데이터에 모델을 적용하면 다음과 같은 결과를 얻을 수 있었습니다. 먼저 응답자는 서울에 살며, 다세대주택에 26년 거주, 점유 형태는 자가, 자가주택을 마련한 적이 있으며 지금 사는 곳 이외에 가구주나 가구원의 이름으로 보유한 주택이 없고, 집 크기는 49.5㎡, 주택에 대한 만족도는 대체로 만족하며 주거환경에 대한 만족도는 상업시설과 문화시설, 도시공원 및 녹지, 접근 용이성은 약간 불만족, 의료시설과 공공기관, 대중교통 접근 용이성은 대체로 만족, 주차시설 이용 편의성은 약간 불만족, 주변 도로의 보행안전과 교육환경, 치안 및 범죄 등 방범 상태, 청소 및 쓰레기 처리 상태, 이웃과의 관계는 대체로 만족이며, 내 집을 보유해야 한다고 생각하고, 성별은 남성, 가구주의 나이는 90세, 현재 가구에서 함께 사는 가구원 수는 6명, 현재 주택의 주거 관리비는 약간 부담되며 국민기초생활보장 수급 가구가 아니고, 월평균 가구 근로/사업소득은 200만 원, 월평균 가구 재산소득은 0원, 월평균 가구 사회보험수혜금은 0원, 월평균 정부 보조금은 48만 원, 월평균 가구 사적이전소득은 0원, 월평균 가구 생활비는 230만 원, 부동산 자산은 2억 3천만 원, 금융자산은 0원, 기타자산은 0원, 부채는 없으며, 학력은 초졸 이하입니다. 해당 응답자의 데이터를 모델의 입력값으로 사용하였을 때 공공임대주택 입주의향이 있을 확률이 Fig. 11과 같이 약 80.8%로 나오며, 임곗값 이상이므로 1='공공임대주택 입주의향 있음'으로 분류됩니다. 그러므로 딥러닝을 이용한 예측 모델을 사용한다면 개개인의 공공임대주택 입주의향 가능성에 대한 예측 결과를 얻을 수 있습니다.

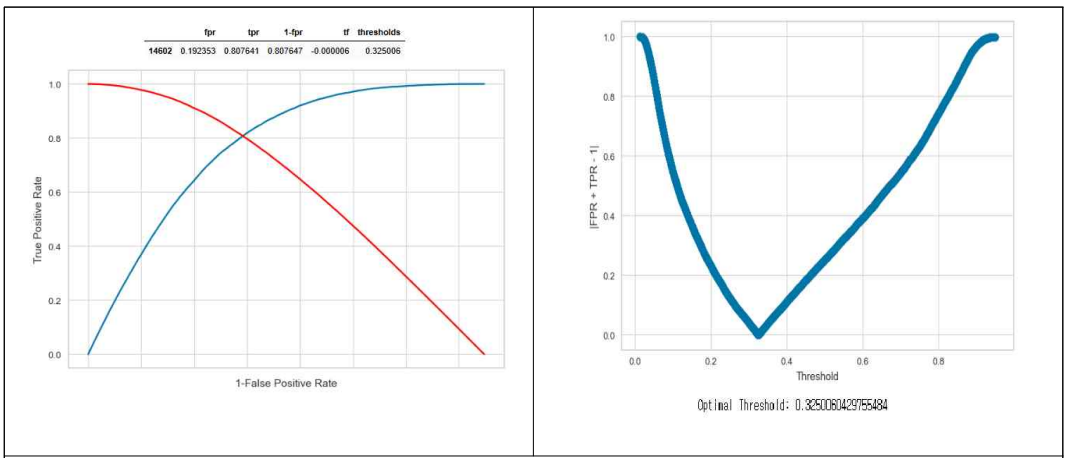
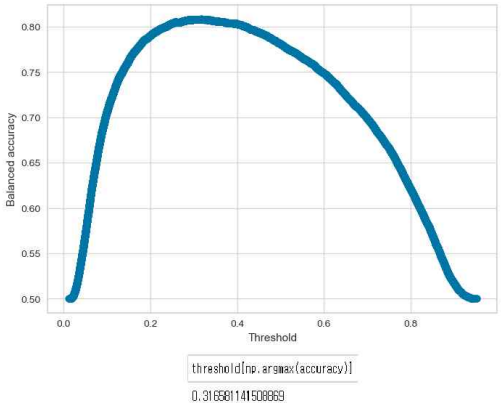


Fig. 9. 최적의 임곗값

```
from sklearn.linear_model import LogisticRegression
lrmodel=LogisticRegression()
lrmodel.fit(X_train,y_train)
threshold = []
accuracy = []
for p in np.unique(lrmodel.predict_proba(X_train)[:,1]):
    threshold.append(p)
    y_pred = (lrmodel.predict_proba(X_train)[:,1] >= p).astype(int)
    accuracy.append(balanced_accuracy_score(y_train,y_pred))
```




```
threshold[np.argmax(accuracy)]
```

```
0.31668114150869
```

Fig. 10. 불균형 데이터 세트를 고려한 최적의 임곗값

```
man=np.array([0.337662338, 0.15239726, 0.880952381, 0.555555556,
              0.1, 0, 0, 0.128, 0, 0.153, 0.021100917, 0., 0.,
              0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
              0., 0., 1., 0., 0.,0., 0.,0.,0.,0.,0.,0.,1.,
              0.,0.,0.,1.,0.,0.,0.,0., 0., 1.,1., 0.,
              0.,0.,0.,0.,1.,0.,1.,0.,0.,1.,0.,0.,0.,1.,
              0.,0.,0.,0.,1.,0.,0.,0.,1.,0.,0.,
              1.,0.,0.,0.,0.,1.,
              0.,0.,1.,0.,0.,0.,1.,0.,0.,0.,
              1.,0.,0.,0.,1.,0.,0.,
              0.,1.,0.,0.,0.,1.,0.,0.,0.,
              0., 1., 0.,1.,1.,0.,1.,0.,0.,
              0.,1.,0.,0.,1.,0.,0.,0.,]).reshape(1,125)
```

```
model.predict(man)
```

```
1/1 [==============================] - 0s 44ms/step
```

```
array([[0.8084758]], dtype=float32)
```

```
(model.predict(man)>=0.33349095754397905).astype("int32")
```

```
1/1 [==============================] - 0s 15ms/step
```

```
array([[1]])
```

Fig. 11. 공공임대주택 입주의향 가능성 예측 결과

■ 기존 (또는 유사한) 사례와 차별성

유사한 사례로 한수정·전희정(2021)의 공공임대주택 입주의향 영향요인에 대한 연구가 있습니다. 해당 연구에서 설명변수는 크게 가구 특성, 주택 특성, 주거환경 특성으로 구분하여 공공임대주택 입주의향 유·무에 따른 집단별 가구, 주택 및 주거환경 특성 차이를 비교하기 위해 차이 검정을 실시해서 모든 변수들에서 통계적으로 유의미한 차이가 있는 것을 확인하고, 이항 로짓 회귀분석을 실시하여 오즈비를 통해 공공임대주택 입주의향에 영향을 주는 요인을 분석했습니다. 공공임대주택 입주의향 예측은 위의 연구와 달리 예측에 초점이 맞춰져 있습니다. 그래서 세분화된 예측을 하기 위해 위의 연구처럼 수도권 또는 비수도권, 아파트 또는 기타 주택같이 데이터를 분류해서 사용하지 않고 그대로 사용했습니다. 예를 들어 설문 응답자의 거주 지역인 전국 17개 시·도를, 주택 유형의 경우 일반 단독주택, 아파트, 오피스텔부터 기타(상가, 사무실, 공장, 농장 내 거주, 종교시설 등)까지 11개를 그대로 사용했습니다. 또한 위의 연구에서 사용한 독립변수 외에 공공임대주택 입주 기준 중 하나인 기준 중위소득과 자산을 고려하여 월평균 근로/사업 소득, 재산소득, 정부 보조금 등의 가구 소득과 부동산 자산, 금융자산, 기타자산 등 수치형 변수를 추가했으며, 국민기초생활보장 수급 가구인지, 부채가 있는지 등의 범주형 변수를 추가했습니다. 이렇게 선정된 총 38개의 독립변수를 기반으로 범주형 변수에는 원핫 인코딩을, 연속형 변수에는 min-max scaler를 적용했습니다. 전처리 결과, 종속변수를 포함해 총 126개의 속성을 얻었습니다. 이를 기반으로 딥러닝을 활용해 입주의향 가능성을 예측하고 분류할 수 있는 모델을 구축하고, 머신러닝 기법을 이용해 예측 모델을 구축했다는 점에서 차별성을 찾을 수 있으며, 중산층까지 공공임대 정책을 확대하고 있는 시점에서 기존에 연구되지 않았던 공공임대주택 입주의향 예측을 시도했다는 점에서 의미가 있다고 할 수 있습니다.

■ 기대효과 및 활용방안

계속해서 바뀌는 부동산 정책과 경제 상황에 따라 개별 가구의 주거문제가 점차 복잡해지고, 다양해지고 있습니다. 그러므로 가구 특성, 주택 특성, 주거환경 특성 데이터를 기반으로 만든 머신러닝/딥러닝 모델을 사용한다면 공공임대주택 입주의향 예측을 통해 입주의향에 미치는 중요 요인을 분석해서 기존의 정책을 개선하거나 데이터를 세분화해서 청년, 신혼부부, 고령자 등 입주자 특성에 맞춰 새로운 맞춤형 공공임대주택 개발 등 공급 계획의 방향성을 제시할 수 있을 것입니다. 또한 현재 공공임대주택 정책이 정책 대상을 중산층으로 확대하고 있는 시점에서 많은 예산이 사용되고 있습니다. 만약 공급에 비해 수요가 낮아지면 주택 공실이 발생하게 되므로 정부 예산이 낭비될 수 있습니다. 그러나 예측을 통해 공공임대주택 입주의향을 정량적으로 파악한다면 예측에 기반한 수요를 확인할 수 있을 것입니다. 혹은 해당 모델을 대한민국 행정권이 미치는 지역에 거주하는 모든 일반 가구를 대상으로 적용한다면 더 정확한 수요 파악이 가능해지므로 예산을 효율적으로 사용할 수 있을 것입니다. 이외에도 딥러닝으로 구현한 모델의 경우 머신러닝 모델과 달리 확률값으로 입주의향 가능성을 결과로 출력할 수 있으며, 임곗값을 기준으로 Fig. 11과 같이 개개인에 대한 입주의향까지 이진 분류가 가능합니다. 그러므로 이 점을 활용한다면 개개인이 월 소득, 자산, 가구원 수 등 자신의 데이터를 입력하여 본인의 입주의향을 확인할 수 있는 자가 진단 서비스를 제공할 수 있습니다.

■ 활용데이터 및 참고문헌 출처 등

1. 활용데이터: 국토교통부. 2020년 주거실태조사(2020). 통계청 MDIS, 다운로드. (20221227 받음)
2. 김경은, 『전국 73.4% "공공임대주택 필요"…'저소득층 이미지'·'공급 부족'이 실패 원인』, 매거진한경, 2022. 10. 28. https://magazine.hankyung.com/business/article/202010287933b(접속일자 2022. 12. 30.)
3. 신지환, 『'중산층 부자' 총자산, 1년새 1.5억 늘어… 78% "금리 5%대면 부동산 구매 포기"』, 동아일보, 2021. 12. 06. https://www.donga.com/news/Economy/article/all/20211206/110622619/1(접속일자 2022. 12. 30.)
4. 국토연구원. (2021) <2020년도 주거실태조사 요약보고서>
5. Are you still using 0.5 as a threshold?[YOUR DATA TEACHER]. (2023.01.05.). URL: https://www.yourdatateacher.com/2021/06/14/are-you-still-using-0-5-as-a-threshold/
6. 한수정·전희정. (2021). 공공임대주택 입주의향 영향요인에 관한 연구. 대한국토·도시계획학회 2021 추계학술대회. https://kpa1959.or.kr/file/E104.pdf
7. 이길제·우지윤. 2022. 소득수준과 생애단계별 공공임대주택 필요가구 현황 및 시사점. 국토이슈리포트 54호. 세종: 국토연구원.

■ (자유롭게 작성)

Fig. 12는 머신러닝 모델로 LGBM을 사용했을 때 변수 중요도입니다.

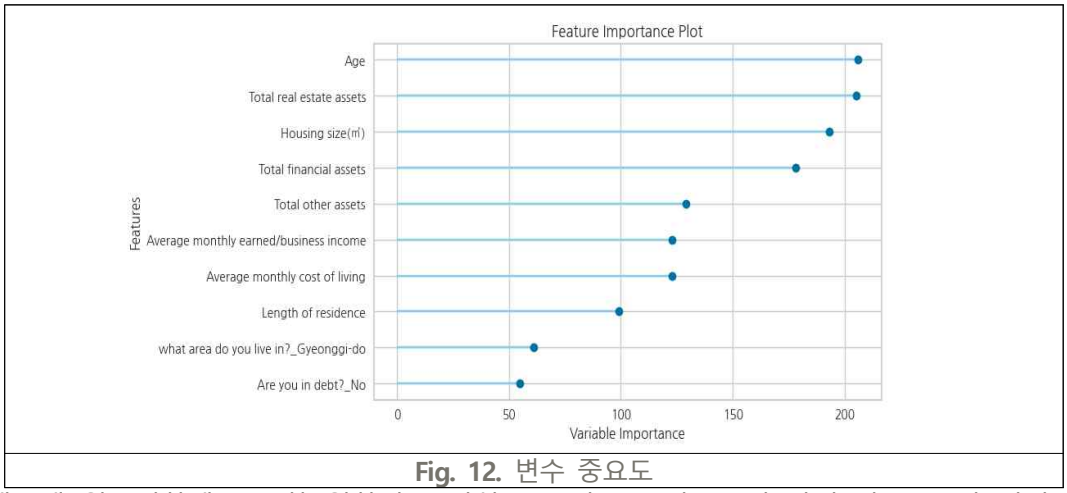


Fig. 12. 변수 중요도

공공임대주택 입주의향에 중요한 영향력을 미치는 요인으로 가구주의 나이, 총 부동산 자산, 주택 크기, 총 금융자산, 총 기타자산, 월평균 근로/사업 소득, 월평균 생활비, 현재 주택에서 거주하고 있는 기간, 경기도에 가주, 부채가 없는 가구 순으로 나타났습니다.

한계점은 다음과 같습니다. 먼저 MDIS에서 제공한 csv 파일에 이상치와 결측치가 존재하는 속성들을 제거하고 예측 모델을 구현해서 '대출을 얼마 정도 받았는지?', '맞벌이 가구인지?'와 같은 공공임대주택 입주와 관련된 속성이 반영되지 않았습니다. 또한 설문조사 기법의 한계를 반영하고 있습니다. 설문조사 응답을 기반으로 작성된 데이터이므로 데이터에 응답자의 거짓 답변과 무응답이 존재할 수 있다는 점입니다. 응답자의 재산을 묻는 과정에서 응답자가 이를 거짓으로 답변하거나 무응답으로 해당 질문을 넘겼을 수 있다는 점입니다. 그러므로 이와 같은 한계점을 극복하고 하이퍼파라미터 튜닝이나, PCA를 사용해 독립변수를 선정해서 예측 모델의 성능을 높일 수 있다면 정확도 높은 예측을 수행할 수 있을 것입니다.

※ 도표, 이미지, 영상 등 활용 가능하고, 10페이지 이내로 작성