포르투갈 고등학생 성적 예측 및 요소 관련 연구

2019020310 이승태

I. 서론

포르투갈인의 교육 수준은 올라갔지만, 유럽 내 포르투갈의 교육 수준은 낮은 것으로 보였다. 따라서, 교육 성과에 영향을 주는 요소들이 무엇인지 밝히고 예측 모델을 세우는 것은 향후 교육 정책 수립에 의미가 있다. Cortez, Silva(2008)[1]는 포르투갈 2개 공립학교(secondary school: 9년 간의 의무교육을 받고 나서의 기관. 고등학교 개념)에서 2005~2006년에 649개의 유효 데이터를 수집했다. 이 데이터를 이용하여 분석을 진행했다. 이후의 부분에서는 데이터의 EDA를 진행 후 여러 모델을 적합 후에 선택할 것이다. 그리고 어떤 요소가 성과에 영향을 줬는지 서술할 것이다. 데이터 분석은 파이썬으로 이뤄졌으며, 이 연구에 사용된 코드 원본[2]은 https://github.com/mars-cookie/Advanced-Statistical-Learning-에서 이용할 수 있다.

II. EDA 및 전처리

다음은 데이터의 변수 설명 표이다. 빨간색은 명목형이고 파란색은 연속형이다.

표 1 <변수 설명표>

변수	설명	변수	설명
sex	성별	age	나이
school	2개 학교	address	도시, 시골
Pstatus	부모와 거주 여부	Medu	어머니 교육수준 0~4((없음, 4th grade, 5~9th grade, 2차 교육, 그 이상 교육)
Mjob	어머니 직업(교사, 건강관련, 공무원, 주부, 기타)	Fedu	아버지 교육수준
Fjob	아버지 직업	guardian	보호자(어머니, 아버지, 기타)
famsize	가족 구성원 수(3 이하 또는 초과)	famrel	가족 관계(리커트: 매우 안좋음~

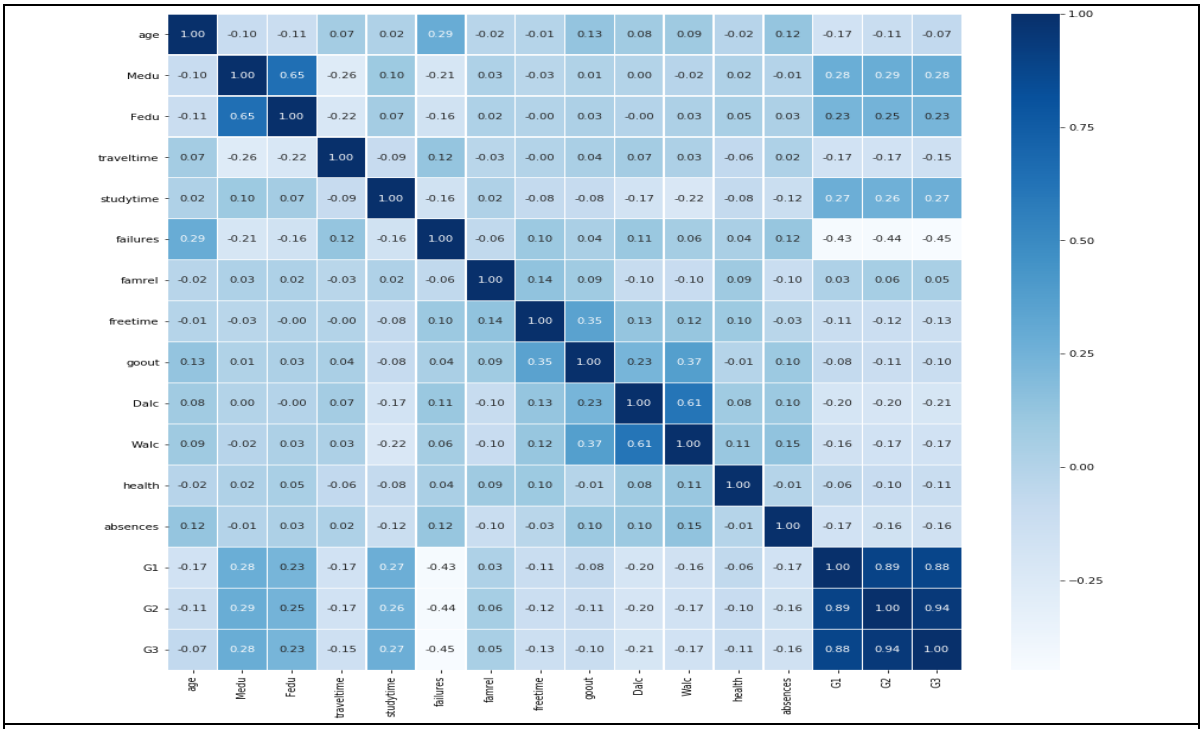
[1] P. Cortez and A. Silva(2008). "Using Data Mining to Predict Secondary School Student Performance." Proceedings of 5th FUture BUsiness TEChnology Conference, 5-12.

[2] https://github.com/mars-cookie/Advanced-Statistical-Learning-/blob/master/%ED%8F%AC%EB%A5%B4%ED%88%AC%EA%B0%88%20%EA%B3%A0%EB%93%B1%ED%95%99%EC%83%9D%20%EC%84%B1%EC%A0%81%20%EC%98%88%EC%B8%A1%20%EB%B0%8F%20%EC%9A%94%EC%86%8C%20%EA%B4%80%EB%A0%A8%20%EC%97%B0%EA%B5%AC.ipynb

			매우 좋음. 1~5)
reason	이 학교 온 이유(통학거리, 평판, 코스 선호, 기타)	traveltime	통학거리(15분 미만, 15~30분, 30~60분, 60분 초과. 7.5,22.5,45,60)
studytime	주 공부시간(2시간 미만, 2~5, 5~10, 10 초과. 1,3.5,7.5,12)	failures	이전 수업 F 개수(4개 이상이면 4로 처리)
schoolsup	추가 교육관련 학교 지원	famsup	가족의 교육적 지원
activities	방과후 활동	paid	방과후 수업
internet	가정에서 인터넷 가능 여부	nursery	nursery 다녔는지
higher	다음 단계 학교 진학 여부	romantic	로맨틱한 관계 있는지
freetime	방과후 자유시간(리커트)	goout	친구와 노는 정도(리커트)
Walc	주말 음주 정도(리커트)	Dalc	주간 음주 정도(리커트)
health	최근 건강 상태(리커트)	absences	결석 수(리커트)
G1	1학년 점수(0~20)	G2	2학년 점수(0~20)
G3	3학년 점수(0~20)		

결석 수와 점수는 학교 데이터에서, 나머지 변수는 설문지를 통해 작성됐다. 이 데이터로 점수를 맞추는 것이 목적이다. 하지만, 설문이 몇 학년에 이뤄졌는지 모르기에 설문 내용이 1~3학년 내내 동일했을 것이라는 가정을 하겠다. 그리고 위의 문제로 인해 G1,G2,G3의 평균(변수 G)을 예측하고자 한다. 여기서 음주 정도가 학생이 솔직하게 답하기 어려운 문항이라고 생각할 수도 있지만, 포르투갈에서는 2005년 당시에는 만 16세부터 음주가 가능하기에[3] 측정이 잘 됐을 것이라 가정한다.

[3]. Francesco Montanari (2015). "Portugal: Portugal — Alcohol Law: New Age Limit for Sale and Consumption of All Alcoholic Drinks and Related Enforcement" European Food and Feed Law Review, 10(6), 459-461.

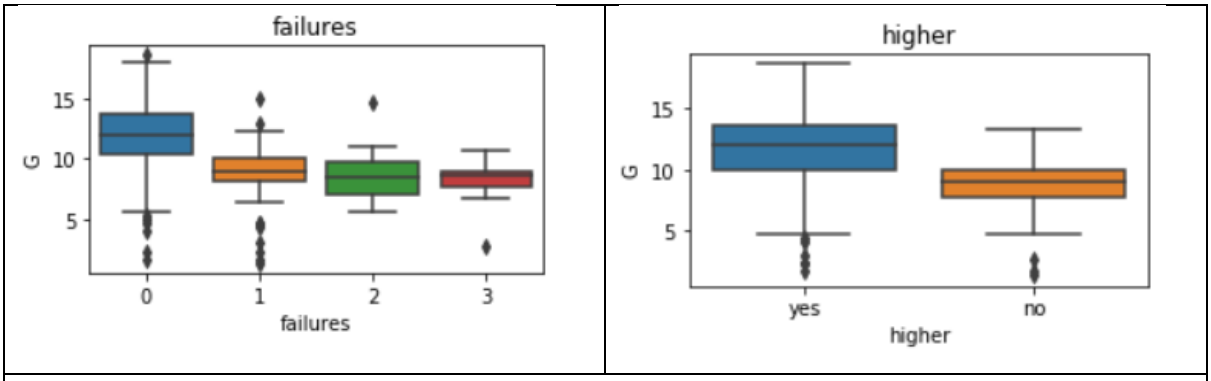



```
%matplotlib inline;
import matplotlib.pyplot as plt; import seaborn as sns;plt.figure(figsize=(15,15))
sns.heatmap(data=ds.corr(method='spearman'),annot=True,fmt='.2f',linewidths=.5,cmap='Blues')
```

그림 1 <연속형 변수들 간 상관계수>

먼저, 스피어만 상관계수 플롯을 통해 서로 비슷한 변수가 있는지 파악해보겠다. 피어슨 상관계수는 변수들이 정규분포일 때 써야하므로, 여기서는 스피어만 상관계수를 이용했다. 지면 절약을 위해 핵심 코드만 수록하였다. 부모 교육수준을 뜻하는 Medu와 Fedu는 계수가 0.65로 상당히 비슷해서 Medu만 쓰기로 했다. 참고로, Medu는 범주형으로 모델링했지만, 원데이터는 순서형이기 때문에 상관계수를 봐도 무방하다. 그리고 알코올 소비량을 뜻하는 Dalc와 Walc가 비슷하기에 Dalc만 남겼다. 그리고 G1,G2,G3 간의 상관계수가 약 0.9로 매우 강한 상관관계를 띄므로 3개 점수를 평균낸 앞의 방법이 적절함을 알 수 있다. 이후 나머지 범주형 변수들에 대한 일변수 막대 그래프를 그려보았다.

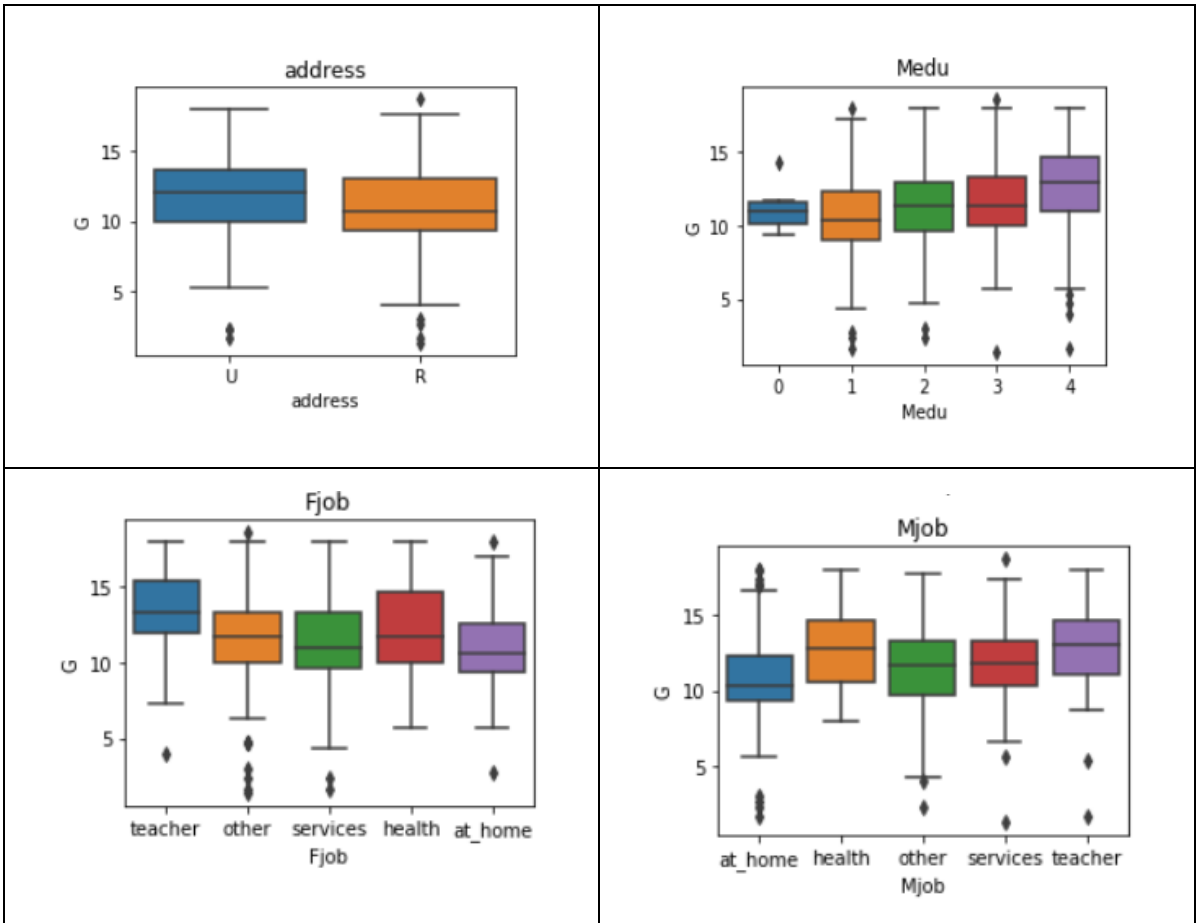
그리고 변수 G와 나머지 변수들과의 관계를 봄으로써, 어떤 변수가 유의할지 추측할 수 있다. 예측변수들을 Cortez, Silva(2008)처럼 학교 관련, 인구통계학 관련, 사회 및 교류 관련 변수로 나누어서 살펴보겠다. 학교 관련 변수 중 G와 크게 관련있는 것은 failures와 higher이었다. 수업에서 F를 받은 횟수가 많을수록 점수는 낮아진다. 만점이 20점이기에 여기서 3점 정도의 차이는 100점 만점에 15점 정도의 차이여서 상당히 큰 차이다. 그리고 다음 단계의 학교로 진학하고자 하는 학생들의 성적이 아닌 학생들보다 좋았다. 이 두 변수가 점수에 영향을 줄 것이라는 것은 전혀 이상하지 않다.



sns.boxplot(x='failures',y='G', data=ds1); plt.title('failures') #그래프 예시코드

그림 2 <학교 관련 변수와 점수와의 관계>

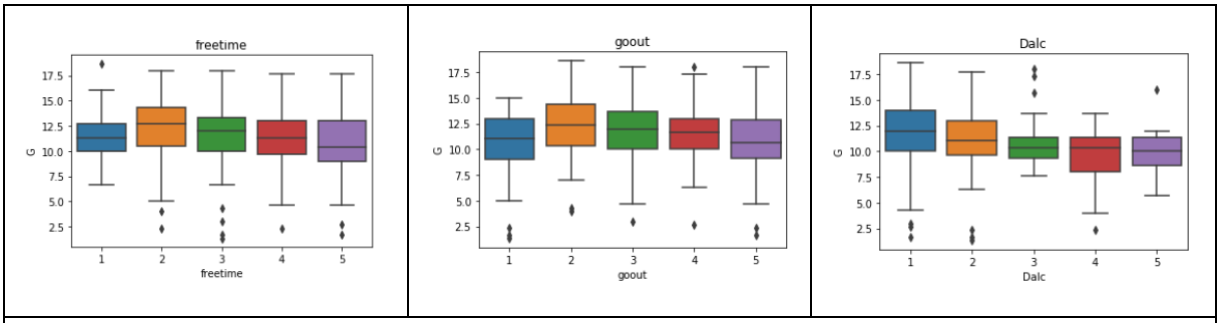
이제 인구통계학적 변수를 보겠다. address를 보면, 학생들이 도시에 사는 경우 점수가 더 높았다. Medu의 경우, 어머니의 교육수준이 높을수록 점수가 올라갔다. 그리고 Fjob의 경우, 아버지가 교사일 때, Mjob도 비슷하게 교사거나 건강 관련 직업을 가졌을 때 점수가 높았다. 여기에 나온 결과 모두 상식적으로 부합하는 것이다.



sns.boxplot(x='Mjob',y='G', data=ds1); plt.title('Mjob') # 그래프 예시

그림 3 <인구통계학적 변수와 점수와의 관계>

마지막으로 사회 및 교류 관련 변수를 보자. freetime은 학생들의 방과 후 자유시간인데, 적당량의 자유시간을 보낸 집단이 점수가 높았다. goout도 친구와 노는 정도로, freetime과 비슷한 특징을 보인다. Dalc를 보면, 역시 술을 많이 마실수록 점수는 떨어졌다.




```
sns.boxplot(x='Dalc',y='G', data=ds1); plt.title('Dalc')
```

그림 4 <사회 및 교류 관련 변수와 점수와의 관계>

EDA 후에 본격적으로 모델을 적합시키기 위해 traveltime과 studytime은 스코어링을 했고, 범주형 변수는 one-hot encoding을 했고, 연속형 변수는 표준화 시 모델 성능을 비교하기 위해 표준화된 데이터를 하나 더 만들었다. 표준화하기 전에 train과 test set을 7:3으로 나눴다. 그 결과 비표준화, 표준화 데이터의 train과 test셋을 만들었다.

```
<스코어링>
ds1['traveltime']=ds1['traveltime'].apply(lambda x: 7.5 if x == 1 else 22.5 if x == 2
                                                    else 45 if x == 3 else 60)
<범주형 one-hot encoding>
ds1['school']=ds1['school'].apply(lambda x: 0 if x == 'GP' else 1)
<표준화>
from sklearn.preprocessing import StandardScaler   ; sc=StandardScaler()
sc.fit(ds4_X_train_std_var) # test data 말고 train 데이터만 이용해서 tranform 모델 만들기
ds4_X_train_std = sc.transform(ds4_X_train_std_var) # train에 적용
ds4_X_test_std = sc.transform(ds4_X_test_std_var) # test에 적용
```

그림 5 <스코어링, 범주화, 표준화 코드>

III. 모델링 및 결과 해석

여기서 풀 문제는 G를 나머지 다른 변수로 얼마나 잘 적합하는지의 회귀 문제이고 여기에 적용할 모델은 다음과 같다. 초모수는 10-fold 교차검증을 통해서 조절했고 scoring 방법은 MSE이다.

그리고 트리 기반의 모델은 표준화가 필요없기에 Elastic Net과 SVM에 대해서만 표준화를 한 데이터에 적합을 더 해봤다. 코드는 박유성(2019)[4]을 참조하였다.

본격적으로 결과 해석을 하기 전에 사용한 모델들에 대한 간단한 설명을 하겠다. Elastic Net은 Lasso와 Ridge의 혼합 모델이라고 보면 되는데 이는 규제화에 L1과 L2 패널티를 모두 썼기 때문이다. SVM 회귀는 점들이 회귀 직선과 특정 밴드 사이에 있도록 회귀 직선들을 적합하는 방식이다. 특정 밴드를 구성하는 점들을 서포트 벡터라 부른다. Decision Tree는 특정 변수를 어느 지점에서 나누면 데이터의 순도가 올라갈지에 대한 문제를 푸는 모델이다. Random Forest와 XGBoost는 일종의 많은 모델들을 결합하는 방식이다. Random Forest는 기존 배깅이 영향력이 큰 한 변수가 나무들을 지배할 수 있다는 단점을 보완했는데, 나무를 만들 때 모든 변수가 아닌 일부 변수만 사용한 것이 그 방법이다. XGBoost는 우선 나무에서 잘못 예측한 부분에 가중치를 두어 다시 나무를 만들다. 그렇게 만든 나무들을 성능에 가중치를 두어 합한다.

표 2 <모델 및 초모수>

모델	초모수
Elastic Net(EN)	alpha(L1, L2 규제화 초모수 합), lambda1(L1 규제화 초모수)
SVM	C(오류 규제화), epsilon(밴드), gamma(가우시안 커널 시의 초모수)
Decision Tree	max_depth(최대 깊이)
Random Forest	k(나무 당 변수 개수), M(나무 개수)=3000
XGBoost	eta(학습률), max_depth, colsample(나무의 변수 선택 개수), M=1000

```
### Grid search에 의한 초모수 결정 (Elastic Net) ###
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import ElasticNet
elt=ElasticNet(random_state=1) # 기초 엘라스틱 모델
param_range=np.linspace(0, 1, 101) # range 정의
param_grid=[{'l1_ratio':param_range,'alpha': [0, 0.001,0.01,0.1,1,10]}] # grid 정의
gs = GridSearchCV(estimator=elt,param_grid=param_grid,
                  scoring='neg_mean_squared_error',cv=10) # nmse(크면 좋음).. 10-fold
gs = gs.fit(ds4_X_train_non_standard,ds4_y_train) # 10-fold 교차검증 시행
print(gs.best_score_) # 최적 스코어
print(gs.best_params_) # 최적 초모수

clf = gs.best_estimator_ # 최적 초모수를 이용하여 모델 조건 구성
```

[4] 2019학년도 2학기 고려대학교 일반대학원 통계학과 고급통계적머신러닝(박유성) 9장 코드.

```
clf.fit(ds4_X_train_non_standard,ds4_y_train) # 모델 적합시켜서 모수까지 구함
ds4_y_train_pred = clf.predict(ds4_X_train_non_standard) # train set 예측값 구하기
ds4_y_test_pred = clf.predict(ds4_X_test_non_standard) # test set 예측값 구하기

from sklearn.metrics import mean_squared_error
print('train mse: %s' % mean_squared_error(ds4_y_train,ds4_y_train_pred)) # train mse(높을수록 안좋다.)
print('test mse: %s' % mean_squared_error(ds4_y_test,ds4_y_test_pred)) # test mse
```

그림 6 <모델 적합 및 초모수 조절 예시>

결과는 다음과 같다. Test MSE를 살펴본 결과 EN 비표준화 모델이 가장 좋았다. 이는 처음 예상한 것과 다른 결과였다. 사실, 적합 측면에서 Random Forest와 XGBoost가 강력하다고 배웠기에 아래 두 모델이 선택될 것이라 예상됐다. 하지만, 데이터에 따라서 이전에 나온 모델이 더 좋을 수 있다는 점을 알 수 있었다.

표 3 <초모수 조절 및 Test MSE>

모델	초모수	Test MSE
EN 비표준화	alpha=0.1, lambda1=0.17	6.488
EN 표준화	alpha=0.1, lambda1=0.15	6.52
SVM 비표준화	C=0.1, epsilon=0.001, 커널=선형	6.6
SVM 표준화	C=10, epsilon=0.1, 커널=가우시안, gamma=0.01	6.74
Decision Tree	max_depth=2	7.124
Random Forest	max_depth=7	6.514
XGBoost	colsample=0.6, gamma=0.4, eta=0.0775, max_depth=4	6.852

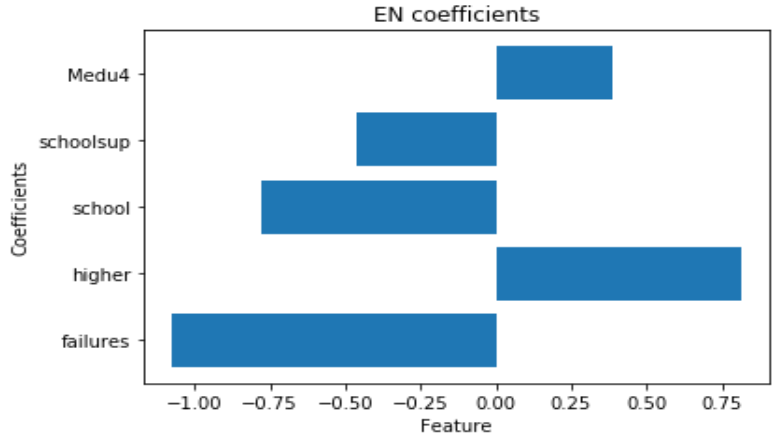
이제 <그림 7>을 통해 EN 비표준화 모델의 변수 중요도[5]를 살펴보자. 먼저, failures의 경우, 이전 과목의 Fail이 많을수록 점수가 떨어졌다. 그리고 higher의 경우, 더 높은 학교를 진학하고자 하는 학생들의 성적이 더 높았다. 이는 공부 동기가 강한 학생들이 그렇지 않은 학생들보다 더 성적이 좋았다는 것이다. 여기서 아쉬운 것은 공부 동기 중에서도 타인의 시선과 경쟁을 의식하여 공부하는 외적 동기와 학습 자체에 흥미를 갖고 자기 발전을 위해 공부하는 내적 동기가 있는데 이를 구분할 수 있는 설문 자료가 있었다면, 기존의 연구결과[6]처럼 내적 동기를 가진 학생들이 외적 동기를 가진 학생들보다 더 성적이 좋았는지를 확인할 수 있었을 것이다. school을 보면 학

[5] 변수(계수): failures(-1.074), higher(0.812), school(-0.779), schoolsup(-0.465), Medu4(0.385), reason_reputation(0.369)

[6] 김희삼(2010), "학업성취도, 진학 및 노동시장 성과에 대한 사교육의 효과 분석", 한국개발연구원, 2010-05.

군에 따라 점수가 달랐다. schoolsup의 경우, 학교로부터 교육적 지원을 받는 학생들을 의미하는데, 이는 교육적 지원을 받아서 성적이 낮은 것이라고 해석하기 보다는 성적이 낮은 학생들에게 교육적 지원을 해줬을 것이라 추측하는 것이 더 합리적이다. Medu4의 경우, 어머니의 교육수준이 제일 높을 경우 다른 경우보다 학생의 성적이 높았다고 해석할 수 있다. 부모의 높은 학력이 부모의 높은 소득수준으로 연결이 되고, 이 사회경제적 환경이 아이에게 전달되고, 이 환경으로 인해 자녀의 교육수준과 성적이 좋아질 수 있다.[7] reason_reputation은 학교 평판을 보고 들어온 학생들이 다른 경우보다 성적이 높았다고 해석될 수 있다. 김희삼(2010)의 연구에 따르면, 내적 동기라 할 수 있는 reason의 course preference가 아닌 다른 변수가 선택되었다는 것이 의외이다. 이는 추후에 학교 평판이 어떤지 확인을 해볼 필요가 있겠다.

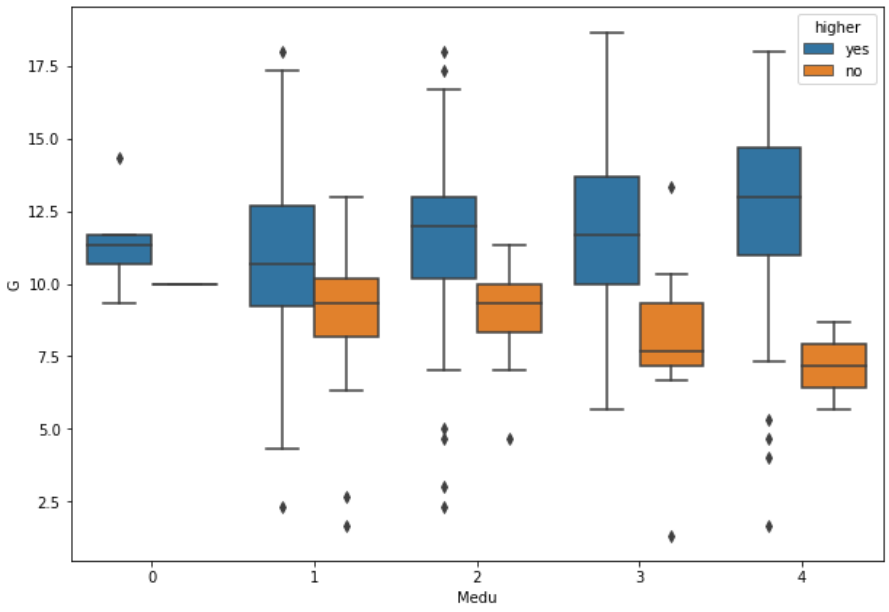
위의 모델 해석 결과를 바탕으로 포르투갈 현실에 정책을 제언하겠다. 포르투갈은 가정환경에 따라 학생의 성적이나 진학의지에 악영향을 미치는 것을 막아야 한다. 이를 위한 근본적인 방법은 공교육의 투자이다. 이렇게 판단하게 만든 변수는 Medu4와 higher이다. 부모의 학력수준으로 인해 학생의 성적이 결정되는 현실이 포르투갈에서도 강하게 의심된다. 여기서, higher를 보고 더 공부하려는 학생들이 높은 성적을 얻는 것이 무엇이 문제냐고 의문이 들 수 있다. 하지만 <그림 8>을 보면, 학생들이 다음 교육기관을 가려는 것과 부모의 학력수준은 강하게 연결되어 있다. 부모의 학력수준이 낮아 소득도 낮아서 학생들이 다음 교육기관을 가고 싶은데 못 가는 경우가 생길 수도 있다고 예측된다. 이는 추후 소득 데이터를 얻을 수 있다면, 좀 더 잘 알 수 있겠다.




```
# 중요변수 그림그리기
en_feature_import5 = en_feature_import.iloc[0:5,0:2]
x=en_feature_import5.iloc[:,0]; y=en_feature_import5.iloc[:,1]
plt.barh(x,y); plt.title('EN coefficients')
plt.xlabel('Feature'); plt.ylabel('Coefficients'); plt.show()
```

[7] 김진영, 전영준, 임병인(2014), "부모 학력에 따른 학업성취도 격차의 국제비교". 재정학연구, 7(2), 27-57.

그림 7 <EN 비표준화 중요 변수 시각화>




```
sns.boxplot(x='Medu',y='G',hue='higher',data=ds1)
```

그림 8 <어머니 교육수준, 진학의지에 따른 성적과의 관계>

IV. 결론

이상 포르투갈 고등학교 성적데이터를 통해서, 학업에 영향을 주는 요소를 밝히고 예측 모델을 적합해봤다. 그 결과 비표준화 EN 모델이 선택되었다. 그리고 이전의 학업 성과, 상위 교육기관으로 진학하려는 의지, 학군, 어머니의 교육수준, 학교 선택 이유 등이 점수 결정에 중요한 변수로 나타났다. 이 모델은 G1, G2, G3를 평균한 G를 예측했기에 Cortez, Silva(2008)의 모델과 직접적으로 비교하기는 무리지만, 그래도 수치적으로는 더 좋은 성능을 보인 것은 성과[8]라 할 수 있겠다. 그리고 어머니의 교육수준과 상위 교육기관으로 진학하려는 의지를 통해서 학생들이 나쁜 가정환경으로 인해 성적과 진학의지에 악영향을 끼치는 일을 막아야 함을 주장했다.

한계점으로는 설문 응답의 정확한 시점을 파악할 수 없어서 3개 학년의 점수를 평균한 점수를 예측했다는 것이 있다. 그리고 초모수 조절 시, 그리드 설정을 임의적으로 했다는 것이 또 다른 한계이다. 사실, 초모수 그리드를 정하기에 따라 모델의 성능이 쉽게 바뀔 수가 있다. 따라서, 초

[8] Cortez, Silva(2008)의 Table 5의 Portuguese C행 RF 열의 RMSE는 2.67, MSE는 7.13이다.

모수 그리드를 올바르게 선택하는 방법과 교차검증을 대체할 베이지안 최적화같은 다른 초모수 조절 방법을 연구할 필요가 있겠다. 또한, 소득수준 변수를 확보하지 못해 부모의 높은 교육수준이 높은 소득으로 이어지고 있다는 것을 선행연구를 통해 가정했다는 점도 한계점이다. 그리고 포르투갈 교육 현실을 좀 더 자세히 파악하고 결론을 냈어야 됐다는 점도 부족한 점으로 꼽힌다.

V. 참고문헌

P. Cortez and A. Silva(2008). "Using Data Mining to Predict Secondary School Student Performance." Proceedings of 5th FUture BUsiness TEChnology Conference, 5-12.

Francesco Montanari (2015). "Portugal: Portugal — Alcohol Law: New Age Limit for Sale and Consumption of All Alcoholic Drinks and Related Enforcement" European Food and Feed Law Review, 10(6), 459-461.

2019학년도 2학기 고려대학교 일반대학원 통계학과 고급통계적머신러닝(박유성) 9장 코드.

김희삼(2010), "학업성취도, 진학 및 노동시장 성과에 대한 사교육의 효과 분석", 한국개발연구원, 2010-05.

김진영, 전영준, 임병인(2014), "부모 학력에 따른 학업성취도 격차의 국제비교". 재정학연구, 7(2), 27-57.