Data Mining: Final Team Project

- Team Report -

2023.03.20

인하대학교 컴퓨터공학과

데이터마이닝 - 001 분반

팀 원 : 곽수연(12191549),

허건혁(12181898)

목차

1. 프로젝트 개요

본 프로젝트에서는 학생들의 수학, 읽기, 쓰기 성적과 개인적 사회 경제적 요인들을 포함한 데이터 세트를 이용하여 학생들의 성적을 예측 및 분류한다. 반응 변수로는 수학 점수를 선택하였고 나머지 변수들에 대하여 예측변수로 사용했다.

다양한 모델을 비교하여 모델의 예측 성능을 평가하여 최적의 모델을 찾고, 예측 결과를 해석한다. 유의미한 변수들을 식별하여 개인의 학습 수준과 능력에 영향을 주는 요인들을 이해한다.

이 프로젝트를 통해 점수에 유의미한 변수를 찾는다면 학생들의 맞춤형 학습 지도전략을 세우고 이를 토대로 교육기관의 교육방침 및 프로그램을 개선할 수 있을 것으로 기대한다.

1.1 데이터 소개

컬럼명	설명	타입
Gender	Gender of the student	object
EthnicGroup	Ethnic group of the student	object
ParentEduc	Parent(s) education background	object
LunchType	School lunch type	object
TestPrep	Test preparation course followed	object
ParentMaritalStatus	Parent(s) marital status	object
PracticeSport	How often the student parctice sport	object
IsFirstChild	If the child is first child in the family or not	object
NrSiblings	Number of siblings the student has	int
TransportMeans	Means of transport to school	object
WklyStudyHours	Weekly self-study hours	object
MathScore	math test score	int
ReadingScore	reading test score	int
WritingScore	writing test score	int

- 분류 모델의 경우 MathScore 값을 대한민국 수능 점수 등급의 백분율을 이용해 1~9 등급 라벨을 생성했다.

2. 프로젝트 수행 절차 및 방법

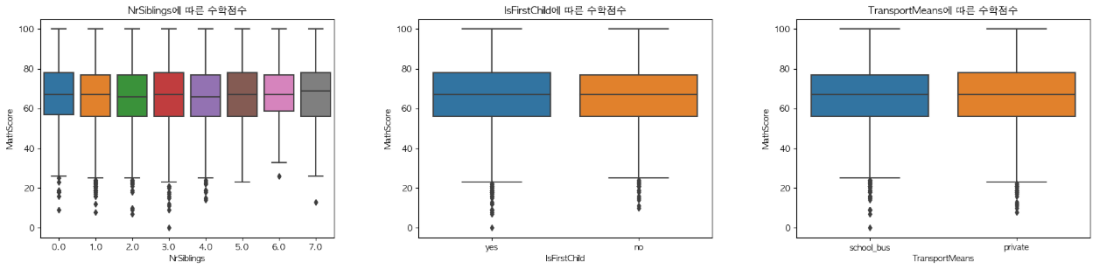
2.1 Data Processing

- **결측치 제거**

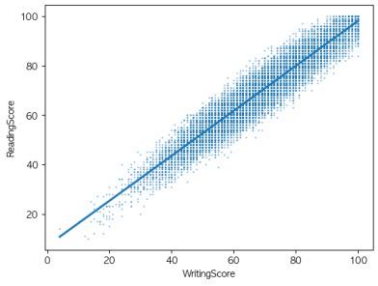
  주어진 데이터는 30,641 개의 행과 14 개의 열을 갖는다. 데이터의 수가 충분히 많다고 판단하여 결측치가 존재하는 데이터의 행은 모두 삭제하기로 했다. 따라서 총 11,398 개의 데이터를 제거하였다.

- **변수 제거**

  결측치를 제거한 데이터에 대하여 범주별 수학점수에 대한 EDA 를 진행하였다.



위 그래프를 보았을 때, NrSblings, IsFirstChild, TransportMeans 변수들에 대하여 시각적으로 그룹별 유의미한 차이가 존재하지 않는다고 판단하였다. 통계적으로 유의성을 확인하기 위해, t-test 를 통하여 그룹별 평균을 비교하였고 유의 수준 0.05 하에서 각 그룹별 평균의 차이가 없다는 결론을 내렸다. 따라서, 위 변수들을 예측변수에서 제외하기로 하였다.



	WritingScore	MathScore
WritingScore	1.000000	0.808533
MathScore	0.808533	1.000000

	ReadingScore	MathScore
ReadingScore	1.000000	0.818986
MathScore	0.818986	1.000000

두 수치형 자료 ReadingScore 와 WritingScore 간 선형관계가 있음을 시각적으로 확인하였다. 선형계수가 약 0.95 로 매우 강한 선형관계를 띄고 있었고, 다중공선성 문제를 해결하기 위하여 예측하고자 하는 MathScore 와 더 적은 선형관계를 갖는 변수인 WritingScore 를 예측변수에서 제외하기로 하였다.

- **데이터 변형**
  - 수치형 변수: 데이터의 평균을 빼고 표준편차로 나누는 Z-score 기반의 정규화를 진행하였다.
  - 범주형 변수 : 모델학습에 사용하기 위해 더미화를 진행하였다.

- **데이터 분할**

학습에 사용할 Train Data 와 Metric 을 평가할 Test Data 에 대하여 8:2 비율로 분할하여 프로젝트를 진행하였다. 분할된 Train Data 에 대하여 5-fold 를 통하여 hyperparameter 를 Grid search 하여 모델을 최적화 하는 hyperparameter 를 찾아 주었다.

2.2 사용 모델 소개

- **Classification Model**
  - **선형 판별 분석(Linear Discriminant Analysis, LDA)**

    : LDA 는 클래스 내의 분산이 최소가 되게 하고, 클래스 간의 거리가 최대가 되게 하는 벡터를 찾아 데이터 포인터들을 투영(projection)시키는 차원 축소 기법이다. 클래스별 같은 공분산 구조를 가진다고 가정한다.

solver	최적화 문제를 푸는 데 사용되는 알고리즘 지정	**svd**, lsqr, eigen

  - **이차 판별 분석(Quadratic Discriminant Analysis, QDA)**

    : QDA 는 LDA 와 달리 클래스 간 분산의 공분산 행렬이 다른 경우에 클래스 간 분산과 클래스 내 분산을 추정하여 비선형 분류를 수행하는 기법이다. 결정 경계선이 비선형 구조를 이룬다.

reg_param	공분산 행렬의 조정에 사용되는 정규화 매개 변수 지정	0.0, **0.1**, 0.2, 0.3, … , 1.0

  - **로지스틱 분류(Logistic Classification)**

    : logistic regression 은 일종의 확률 모델로서 독립 변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는 데 사용되는 기법이다. 종속 변수가 범주형 데이터를 대상으로 하며 입력 데이터가 주어졌을 때 해당 데이터의 결과가 로지스틱 함수에 의해 특정 분류로 나뉘며, 이진 분류를 위해 설계되었지만, 다중 클래스 분류에서도 사용할 수 있다.

penalty	규제를 적용할 유형 지정	l1, l2, elasticet, **none**
C	규제 강도를 조절하는 역수(작을수록 규제 강화)	10e-4 ~ 10e-4, **4.28**

solver	최적화 문제를 푸는 데 사용되는 알고리즘 지정	**newton-cg**

- **의사 결정 트리(Decision Tree)**

  : Decision Tree 는 트리구조로 일련 질문과 특정 조건에 따라 데이터를 분할하는 기법이다. Decision Tree 는 중요한 feature 일수록 해당 feature 를 기준으로 한 분할이 불순도를 크게 감소시킨다는 것을 바탕으로 변수 중요도를 구한다. 분류와 회귀에서 모두 사용 가능하며 직관적인 결정 경계를 생성하기 때문에 결과를 이해하기 쉽다는 장점이다. 반면 파라미터를 적절히 조절하지 않으면 과적합이 발생하기 쉽다는 단점이 있다.

max_depth	트리의 최대 깊이 제한	3, 5, **7**, 10
min_samples_spilt	노드를 분할하기 위해 필요한 최소 샘플 수 지정	**8**, 10, 12, 16, 18, 20
min_samples_leaf	리프 노드가 되기 위해 필요한 최소 샘플 수 지정	3, 5, 10, 15 ,**20**

- **Regression Model**
  - **선형 회귀(Linear Regression)**

    : Linear Regression 은 종속 변수와 하나 이상의 독립 변수가 간의 선형 관계를 모델링하는 가장 간단한 회귀 분석 방법으로, 독립 변수가 종속 변수에 미치는 영향력의 크기를 파악하고 이를 통해 독립 변수의 일정한 값에 대응하는 종속 변수 값을 예측하는 데이터 분석 기법이다.

  - **Ridge Regression**

    : Ridge Regression 은 선형 회귀의 일반화된 형태로, 회귀 계수의 크기를 제어하여 과적합(overfitting)을 방지하는 방법이다. Ridge Regression 은 L2 규제를 추가하여 회귀 계수의 크기를 제한한다. L2 규제는 손실 함수에 회귀 계수의 제곱을 더한 항을 추가하는 방식으로 동작한다.

alpha	규제의 강도 조절(높은 값일수록 더 강한 규제를 의미)	0.0, 0.05, … **8.75**, …, 10.0

  - **LASSO Regression**

    : LASSO Regression 은 Ridge Regression 과 유사하지만, L1 Norm 을 사용하여 회귀 계수를 제한하는 차이가 있다. L1 Norm 은 손실 함수에 회귀 계수의 절댓값을 더한 항을 추가하는 방식으로 동작한다. LASSO Regression 은 예측에 영향력이 적은 변수의 회귀 계수를 0 으로 만들어 변수 선택(Feature Selection)의 효과를 가진다.

alpha	규제의 강도 조절(높은 값일수록 더 강한 규제를 의미)	0.0, 0.05, …, 10.0

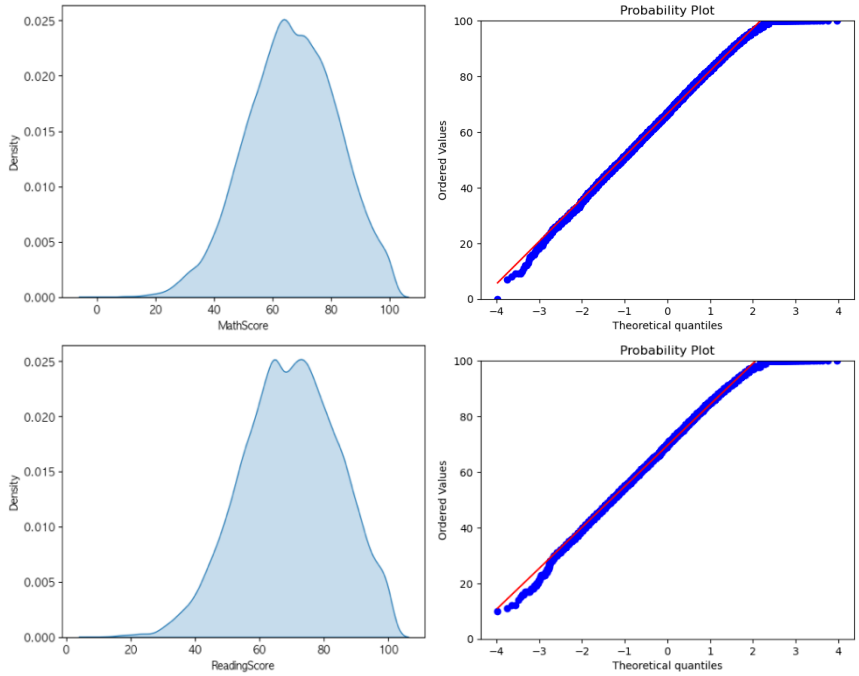
- **부분 최소 제곱(Partial Least Squares, PLS)**

  : Partial Least Square 는 종속 변수와 독립 변수 간의 관계를 추출하기 위해 사용되는 방법이다. PLS 는 독립 변수들의 선형 조합인 잠재 변수(latent variables)를 생성하여 이들과 종속 변수 간의 최대 공분산을 찾는다. PLS 는 변수 간의 다중 공선성(multicollinearity) 문제를 완화하고 예측 성능을 향상시킬 수 있다는 장점이 있다.

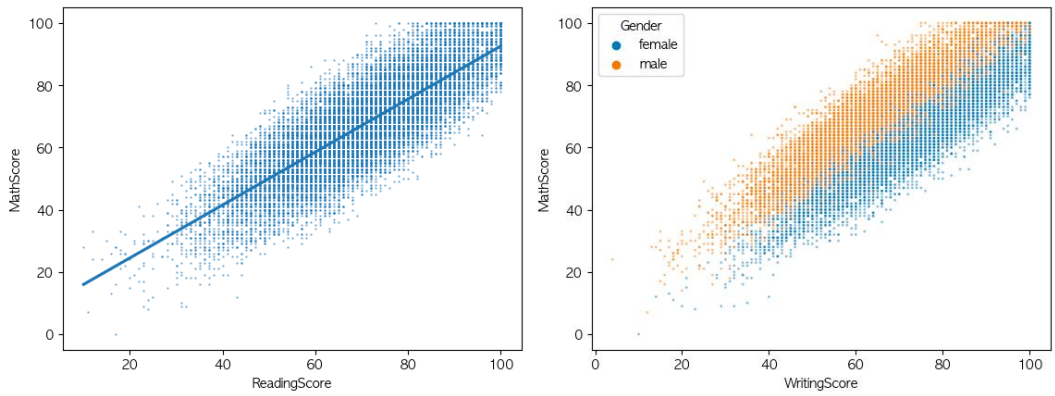
n_components	생성할 잠재 변수의 개수를 지정	1, 2, 3, 4, 5, **6**, 7, 8, 9

※ Grid Search 를 통해 찾은 Best Parameter 를 붉은 글씨로 표시하였다.

- **모델 선택 이유**



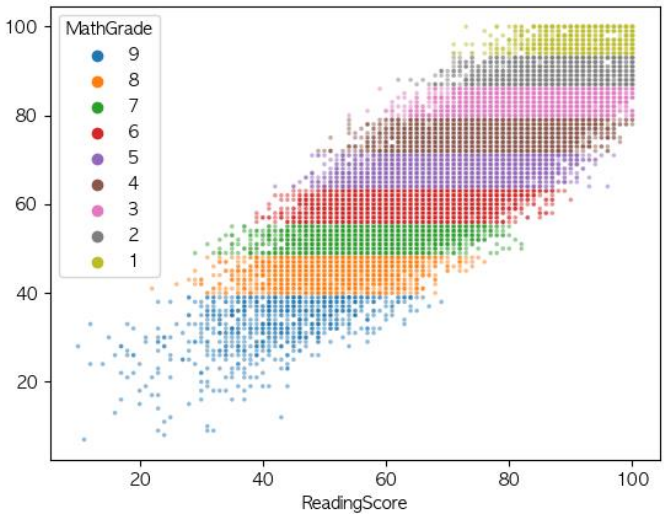
위 그래프에서는 예측하고자 하는 수학점수 와 예측변수인 읽기점수가 모두 정규분포 형태를 보이는 것을 시각적으로 확인할 수 있다.



위 그래프에서 확인할 수 있는 것처럼 읽기 점수와 수학점수 간 선형관계를 파악할 수 있다. 또한 성별과 같은 특정 범주에 의해서 두 집단으로 명확하게 구분되어 짐을 확인할 수 있었다. 따라서 선형모델을 통하여 점수를 예측하고 분류하고자 하였다.

위에서 확인한 선형관계를 바탕으로 예측모델로 선형회귀, Ridge, Lasso, PLS 를 사용하기로 결정했고 분류 모델로 Logistic Classification 을 사용하기로 결정했다.

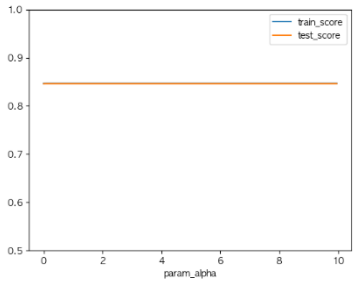
예측변수와 반응변수 간의 상관계수가 약 0.81 로 높은 값을 보여준 것을 고려하여, 차원 축소를 통한 회귀모형 중 주성분 회귀(PCR)을 통한 회귀모형은 사용하지 않았다. 대신 이러한 상관관계를 반영할 수 있는 편중 최소 제곱(PLS)회귀를 선택하였다.

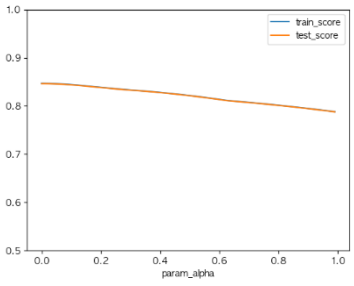


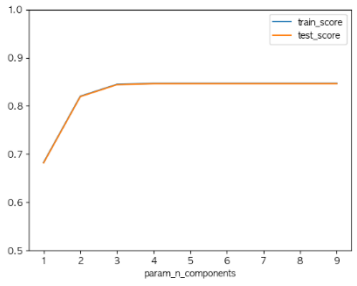
등급별 읽기 점수와 수학 점수간 산점도를 시각화해 보았다. 시각화 결과 결정경계가 비선형 구조보다 선형구조일 때 분류가 더 잘 이루어질 것으로 판단되어 LDA 를 분류모델로 사용하고 비교를 위하여 QDA 를 함께 분류모델로 사용하기로 결정했다.

3. 프로젝트 결과 분석

3.1 Cross Validation 결과







위 세 가지 그래프는 Ridge, LASSO, PLS 에 대한 각각의 파리마터 별 결정계수 (R-squared)를 Score 로 표시한 것이다. 관찰 결과, 결정계수를 score 로 사용했을 때 train score 와 test score 간의 차이가 없다. 이러한 현상은 데이터의 양이 많아 일반화 성능이 향상되었고, 회귀 모델을 훈련할 때 훈련에 사용된 수치형 변수가 하나뿐이라는 한계로 인해 모델의 단순함 때문으로 판단된다.

3.2 Regression Model 결과 분석

- Coefficient 에 대한 t-test

< Ridge >	coef	standard errors	t-value	p-value
Const	66.0273	0.0098	6769.081	0
ReadingScore	13.091	0.0187	700.1793	0
Gen_female	-5.6053	0.0176	-319.309	0
Gen_male	5.6053	0.0175	321.1834	0
Eth_group A	-1.413	0.049	-28.8239	0
Eth_group B	-1.2267	0.0326	-37.598	0
Eth_group C	-1.4008	0.0285	-49.1354	0
Eth_group D	-0.1015	0.0309	-3.2847	0.001
Eth_group E	4.1421	0.0376	110.1063	0
Par_associate's degree	0.2626	0.0342	7.6682	0
Par_bachelor's degree	0.4549	0.0417	10.9138	0
Par_high school	-0.2687	0.0346	-7.7784	0
Par_master's degree	0.1136	0.0524	2.1692	0.0301
Par_some college	0.0381	0.0326	1.1697	0.2422
Par_some high school	-0.6005	0.0359	-16.7401	0
Lun_free/reduced	-2.4394	0.0191	-128.032	0
Lun_standard	2.4394	0.0178	137.206	0
Tes_completed	-0.5752	0.0187	-30.6786	0
Tes_none	0.5752	0.0173	33.1717	0
Par_divorced	-0.1026	0.0413	-2.4846	0.013
Par_married	-0.0878	0.0339	-2.5893	0.0097
Par_single	-0.144	0.039	-3.6961	0.0002
Par_widowed	0.3343	0.0901	3.7094	0.0002
Pra_never	-1.1773	0.0328	-35.936	0
Pra_regularly	0.9886	0.0243	40.6103	0
Pra_sometimes	0.1887	0.0225	8.4008	0
Wkl_5 - 10	0.0596	0.0215	2.7732	0.0056
Wkl_< 5	-1.1199	0.0254	-44.0083	0
Wkl_> 10	1.0603	0.0296	35.8742	0

< LASSO >	coef	standard errors	t-value	p-value
Const	73.5813	0.0098	7543.513	0
ReadingScore	13.1004	0.0187	700.6822	0
Gen_female	-11.1783	0.0176	-636.772	0
Gen_male	0	0.0175	0	1
Eth_group A	-0.0502	0.049	-1.0243	0.3058
Eth_group B	0	0.0326	0	1
Eth_group C	-0.1421	0.0285	-4.9834	0
Eth_group D	1.0872	0.0309	35.172	0
Eth_group E	5.2987	0.0376	140.8532	0
Par_associate's degree	0.1541	0.0342	4.4991	0
Par_bachelor's degree	0.3126	0.0417	7.4986	0
Par_high school	-0.2738	0.0346	-7.9242	0
Par_master's degree	0	0.0524	0	1
Par_some college	0	0.0326	0	1
Par_some high school	-0.602	0.0359	-16.7819	0
Lun_free/reduced	-4.8288	0.0191	-253.436	0
Lun_standard	0	0.0178	0	1
Tes_completed	-1.1099	0.0187	-59.1986	0
Tes_none	0	0.0173	0	1
Par_divorced	0	0.0413	0	1
Par_married	0	0.0339	0	1
Par_single	-0.0097	0.039	-0.2489	0.8035
Par_widowed	0	0.0901	0	1
Pra_never	-1.288	0.0328	-39.3152	0
Pra_regularly	0.7723	0.0243	31.7276	0
Pra_sometimes	0	0.0225	0	1
Wkl_5 - 10	0	0.0215	0	1
Wkl_< 5	-1.1432	0.0254	-44.9244	0
Wkl_> 10	0.9447	0.0296	31.9619	0

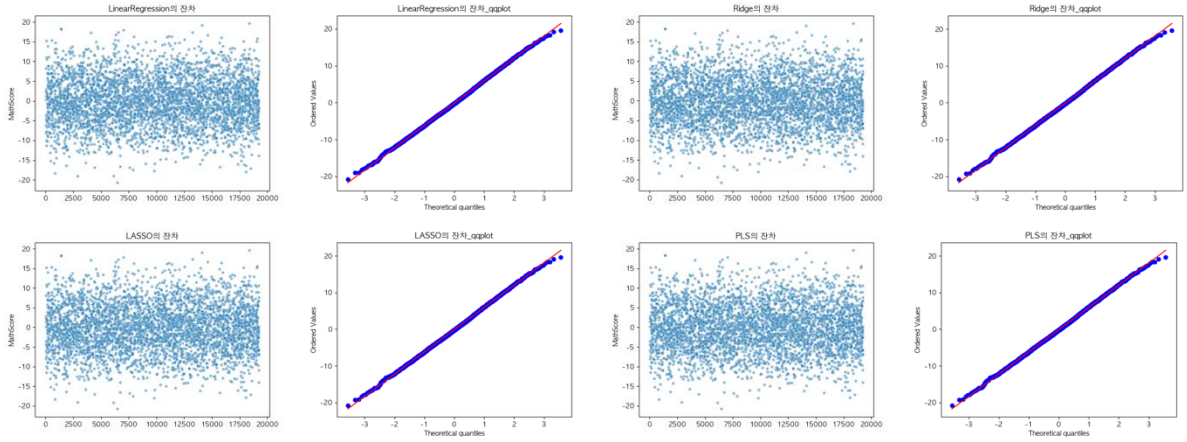
< Linear Regression >	coef	standard errors	t-value	p-value
Const	4.75E+13	0.0098	4.87E+15	0
ReadingScore	13.0942	0.0187	700.3507	0
Gen_female	9.28E+12	0.0176	5.29E+14	0
Gen_male	9.28E+12	0.0175	5.32E+14	0
Eth_group A	3.55E+13	0.049	7.24E+14	0
Eth_group B	3.55E+13	0.0326	1.09E+15	0
Eth_group C	3.55E+13	0.0285	1.25E+15	0
Eth_group D	3.55E+13	0.0309	1.15E+15	0
Eth_group E	3.55E+13	0.0376	9.44E+14	0
Par_associate's degree	2.41E+12	0.0342	7.03E+13	0
Par_bachelor's degree	2.41E+12	0.0417	5.78E+13	0
Par_high school	2.41E+12	0.0346	6.97E+13	0
Par_master's degree	2.41E+12	0.0524	4.6E+13	0
Par_some college	2.41E+12	0.0326	7.39E+13	0
Par_some high school	2.41E+12	0.0359	6.72E+13	0
Lun_free/reduced	1.18E+13	0.0191	6.19E+14	0
Lun_standard	1.18E+13	0.0178	6.64E+14	0
Tes_completed	-6.1E+13	0.0187	-3.2E+15	0
Tes_none	-6.1E+13	0.0173	-3.5E+15	0
Par_divorced	-7.5E+12	0.0413	-1.8E+14	0
Par_married	-7.5E+12	0.0339	-2.2E+14	0
Par_single	-7.5E+12	0.039	-1.9E+14	0
Par_widowed	-7.5E+12	0.0901	-8.3E+13	0
Pra_never	-7.1E+13	0.0328	-2.2E+15	0
Pra_regularly	-7.1E+13	0.0243	-2.9E+15	0
Pra_sometimes	-7.1E+13	0.0225	-3.2E+15	0
Wkl_5 - 10	3.24E+13	0.0215	1.51E+15	0
Wkl_< 5	3.24E+13	0.0254	1.27E+15	0
Wkl_> 10	3.24E+13	0.0296	1.1E+15	0

< PLS >	coef	standard errors	t-value	p-value
Const	66.612	0.0098	6829	0
ReadingScore	13.092	0.0187	700.24	0
Gen_female	-2.8026	0.0176	-159.65	0
Gen_male	2.8026	0.0175	160.59	0
Eth_group A	-0.3309	0.049	-6.7494	0
Eth_group B	-0.4252	0.0326	-13.033	0
Eth_group C	-0.5783	0.0285	-20.283	0
Eth_group D	0.0289	0.0309	0.9333	0.3507
Eth_group E	1.4905	0.0376	39.622	0
Par_associate's degree	0.1198	0.0342	3.4985	0.0005
Par_bachelor's degree	0.1596	0.0417	3.8277	0.0001
Par_high school	-0.0906	0.0346	-2.6221	0.0088
Par_master's degree	0.0394	0.0524	0.7517	0.4523
Par_some college	0.0333	0.0326	1.0221	0.3068
Par_some high school	-0.2206	0.0359	-6.1497	0
Lun_free/reduced	-1.1659	0.0191	-61.191	0
Lun_standard	1.1659	0.0178	65.576	0
Tes_completed	-0.2737	0.0187	-14.601	0
Tes_none	0.2737	0.0173	15.788	0
Par_divorced	-0.0027	0.0413	-0.0643	0.9487
Par_married	0.0038	0.0339	0.1125	0.9104
Par_single	-0.0208	0.039	-0.5332	0.594
Par_widowed	0.0583	0.0901	0.6463	0.5181
Pra_never	-0.4807	0.0328	-14.673	0
Pra_regularly	0.3633	0.0243	14.925	0
Pra_sometimes	-0.0216	0.0225	-0.9604	0.3369
Wkl_5 - 10	0.0777	0.0215	3.6132	0.0003
Wkl_< 5	-0.4592	0.0254	-18.045	0
Wkl_> 10	0.4404	0.0296	14.899	0

회귀 모델들을 적합 시킨 결과, 각 변수에 대한 회귀 계수를 얻을 수 있었다. 이후, 각 변수의 유의성을 확인하기 위해 t-test 를 수행했다. 유의 수준을 0.05 로 설정하고, p-value 가 0.05 보다 큰 경우 해당 변수를 유의미하지 않다고 판단하였다. 위의 표에서는 유의미하지 않다고 판단된 변수를 표시했다.

결과적으로, LASSO Regression 과 PLS Regression 모델에서는 Parent Marital Status 에 해당하는 더미 변수가 수학 점수를 예측하는 데 영향을 주지 못한다고 판단할 수 있다. 각 범주형 데이터들에 대한 계수 값은 회귀모형의 절편을 결정한다. 따라서, 범주형 데이터로부터 파생된 더미 변수 간의 차이가 크다면 회귀모형이 예측하는 수학 점수에 대하여 영향을 준다고 판단할 수 있다. 이에 따라 성별과 점심 유형은 다른 범주형 변수들에 비하여 수학 점수 예측에 영향을 주었다고 판단된다.

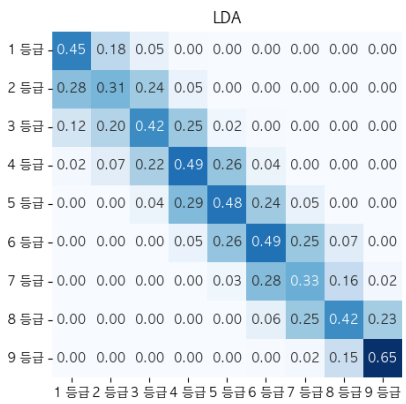
- 예측 결과의 잔차 분석

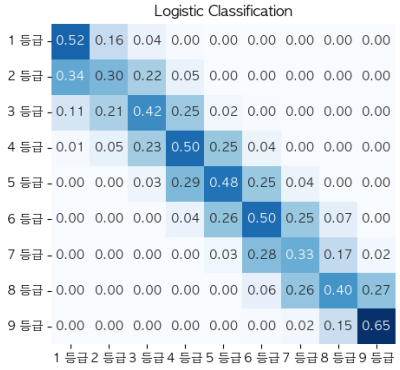


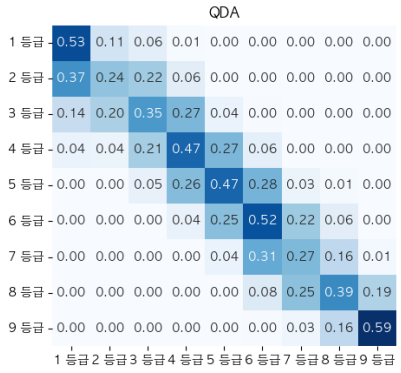
회귀 모델로 예측한 결과에 대해 잔차 및 잔차의 분포를 scatter plot 과 QQplot 으로 시각화하였다. 위 그래프에서 잔차가 정규분포를 따르고 등분산성을 만족함을 확인할 수 있었다. 이는 회귀 모형의 가정을 올바르게 따르면서 모델이 훈련되었다고 판단할 수 있다.

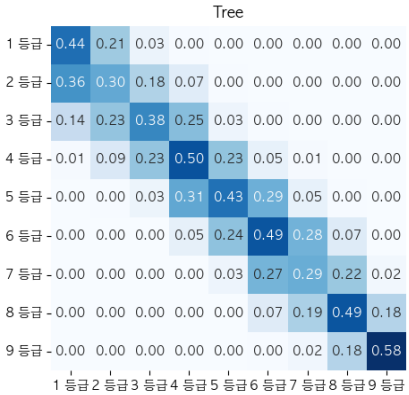
3.3 Classification Model 결과 분석

- Confusion Matrix









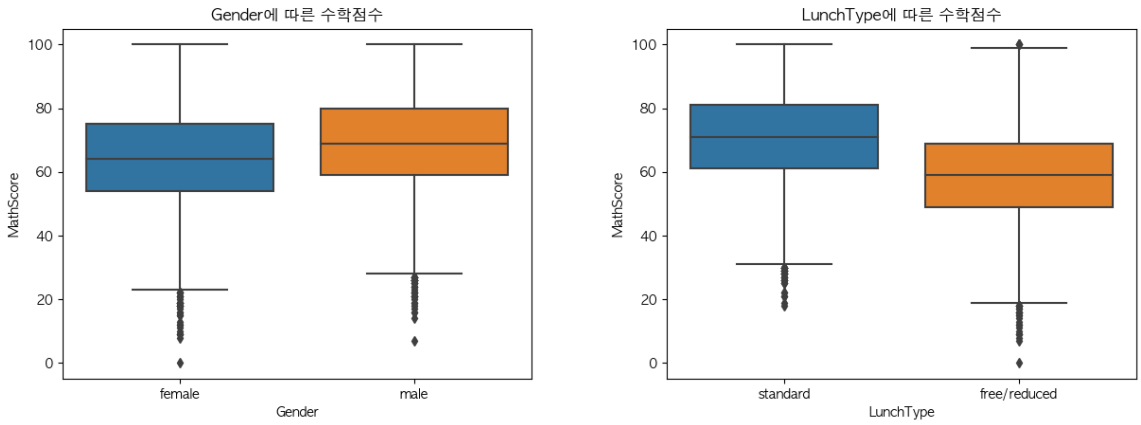
모델	분류율
Logistic Classification	0.4523
LDA	0.4515
Tree	0.4375
QDA	0.4315

4 가지 모델 중 Logistic Classification 과 LDA 가 보다 좋은 성능을 보여주고 있다.

Logistic Classification 모델의 좋은 성능은 앞서 EDA 를 통해 예측변수인 읽기 점수가 반응변수인 수학 점수와 선형관계를 갖기 때문으로 보인다. 또한 QDA 보다 LDA 가 더 좋은 성능을 보여줌으로써 앞서 모델 선정에 있어서 결정경계가 선형인 모델이 더 좋은 분류율을 보여줄 것이라는 판단이 옳았다고 생각된다.

또한, Decision Tree 의 feature_importances_를 확인해 보았을 때, ReadingScore 변수가 63%로 가장 높은 영향력을 가짐을 알 수 있었고, Gen_female 변수가 17%로 두번째로 높은 영향력을 보임을 알 수 있었다.

4. 프로젝트 고찰



위 분석 결과들을 바탕으로 다른 변수들에 비하여 수학 점수에 큰 영향을 미치는 변수들은 Gender 와 점심 유형으로 보인다. 이때 점심 유형의 경우 무료 급식을 받는 학생들에 대하여 경제적으로 낮은 수준이라는 것을 추측할 수 있다. 따라서 수학 점수에 영향을 미치는 변수는 성별과 학생들의 집안의 경제적 상황이라는 결론을 내렸다.

본 프로젝트를 진행하면서 아쉬웠던 점은 데이터의 다양성이 부족했다는 점이다. 하나의 수치형 변수만을 사용하여 모델을 훈련시켜야 하는 제약이 있었고, 이를 극복하기 위해 변수 변환을 통해 새로운 수치 자료를 생성해 보았지만 모델 성능은 오히려 저하되었다. 다양한 모델을 사용하여 최적의 모델을 선택하고자 하였으나 회귀 모델의 경우 모든 모델이 결정계수가 약 0.84 로 비슷한 성능을 보임에 그쳤다. 이를 통해 실제로 산업 현장에서 데이터를 분석하기에 앞서 다양한 데이터를 수집하기 위한 전략이 필요하다는 것을 느낄 수 있었다.

5. 참고문헌

- Project GIt : https://github.com/GeonHyeock/Capstone-Design-Score-Prediction
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning. : Springer, 2021
- scikit-learn Developers. “scikit-learningL Machine Learning in Python.: scikit-learning Documentation, Version 0.24.2, 2021. [Online]. Available: https://scikit-learn.org/stable/index.html
- 데이터 출처 : https://www.kaggle.com/datasets/desalegngeb/students-exam-scores