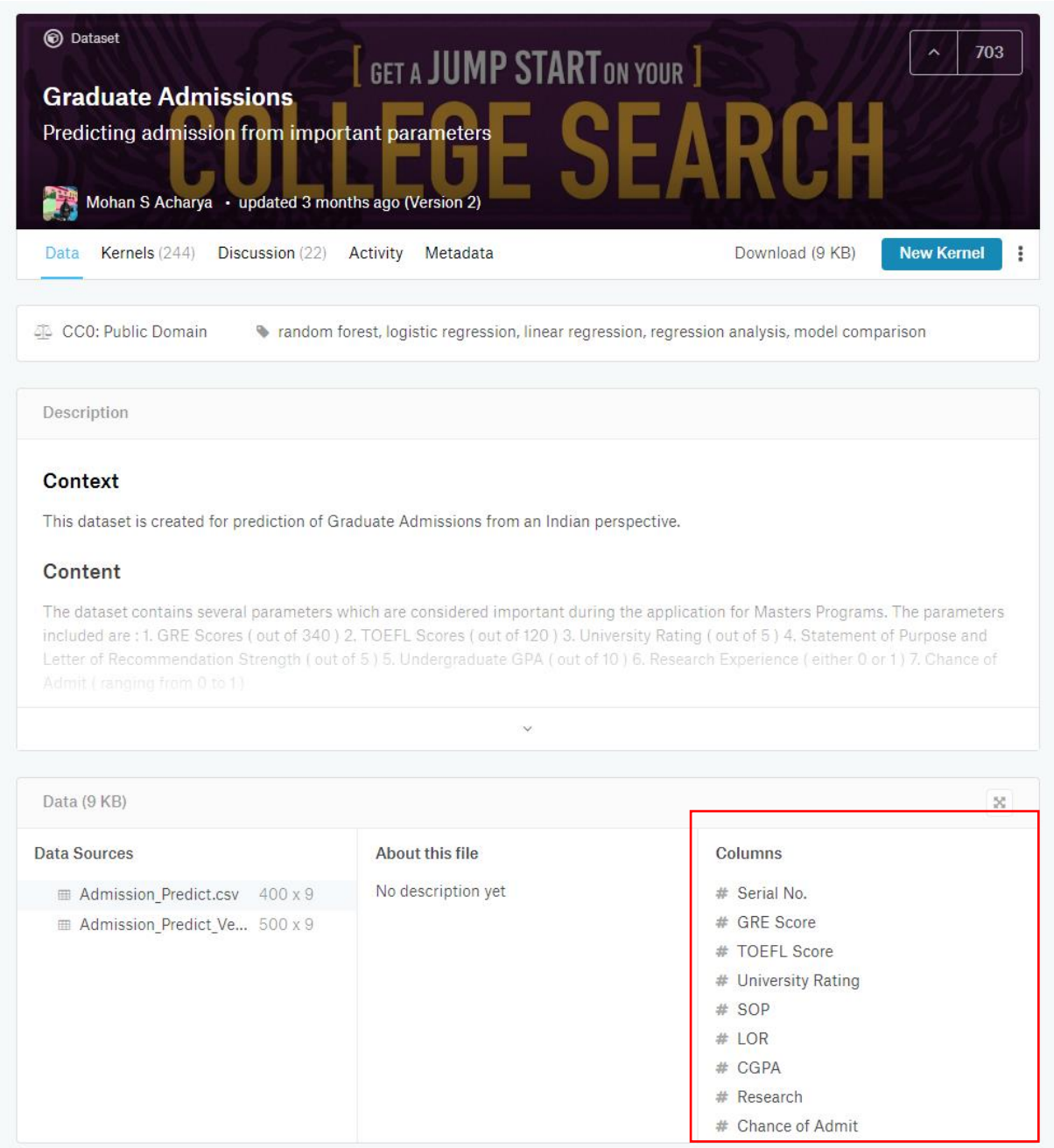
Multivariate Data Analysis Assignment #2

Logistic Regression

Dataset: Graduate Admissions (file name: Admission_Predict_Ver1.1.csv)

(https://www.kaggle.com/mohansacharya/graduate-admissions)

해당 데이터셋은 미국 대학원의 지원자들에 대한 여러 가지 점수(GRE, TOELF 등)와 대학의 등급에 따라 각 지원자들이 합격할 확률(Chance of Admit)을 기록한 데이터이다.



[Q1] Logistic Regression 모형 구축을 위해 필요하지 않은 변수는 어떤 것들이 있는가? 왜 그렇게 생각하는가?

다음 물음에 대해서는 [Q1]에서 선택한 변수들은 제외하고 답변하시오.

[Q2] 개별 입력 변수들에 대하여 각각 다음과 같은 단변량 통계량을 계산하고 Box plot을 도시하시오: Mean, Standard deviation, Skewness, Kurtosis. 전체 변수 중에서 정규분포를 따른다고 할 수 있는 변수들은 몇 개인가? 정규분포를 따른다고 가정한 근거는 무엇인가?

[Q3] [Q2]의 Box plot을 근거로 각 변수들에 대한 이상치(너무 값이 크거나 작은 객체) 조건을 정의하고, 해당하는 객체들을 데이터셋에서 제거해 보시오.

다음 각 물음에 대해서는 [Q3]에서 제거된 객체들을 제외시킨 데이터프레임을 사용하여 답하시오.

[Q4] 가능한 모든 두 쌍의 입력 변수 조합에 대한 산점도(scatter plot) 및 Correlation plot (hint: "corrplot" 패키지의 corrplot( ) 함수 사용) 상관관계를 계산해 보시오. 어떤 두 조합의 변수들이 서로 강한 상관관계가 있다고 할 수 있는가?

[Q5] 종속변수인 Change of Admit은 원래 데이터에서는 0부터 1사이의 확률 값으로 표현되어 있다. 이를 0.8을 기준으로 하여 0.8을 초과하는 경우 1 (positive class), 0.8 이하인 경우 0 (negative class)의 값을 갖는 binary target variable로 변환하시오. 이후 전체 데이터셋을 70%의 학습 데이터와 30%의 테스트 데이터로 무작위(random)로 분할한 후 모든 변수를 사용하여 Logistic Regression 모델을 학습해 보시오. 유의수준(Significance level) 0.1에서 Change of Admit에 유의미하게 영향을 주는 변수들은 어떤 것들이 있는가?

[Q6] Test 데이터셋에 대하여 예측을 수행하고 Confusion Matrix를 생성한 뒤, True Positive Rate, True Negative Rate, False Positive Rate, False Negative Rate, Simple Accuracy, Balanced Correction Rate, F1-Measure를 각각 구하고 그 의미를 해석하시오.

[Q7] Test 데이터셋에 대한 AUROC를 산출하는 함수를 직접 작성하고, random seed를 변경해가면서 학습-테스트를 5회 반복하여 산출된 AUROC값의 변화를 확인해보시오.

[Q8] 이 외 웹이나 기타 자료들을 통해 재미있는 데이터셋(fun dataset)을 찾아 나름대로의 로지스틱 회귀분석 모형 구축 및 결과 해석을 수행하시오.