**비즈니스 모델링 2**

**Assignment 2**

**Due: 12/17, 2020 (목요일), 11:59PM**

**(Total Points: 12)**

**Instructions:** 다음 문제들을 잘 읽고, 요구되는 코드를 작성하고 설명하시오. 코드와 설명을 포함한 모든 답안을 ‘HW4\_김일주(자기이름).r’ 하나의 파일에 저장하여 제출하시오. 설명을 요하는 문제의 경우 주석을 사용하도록 하고, 코드와 설명이 몇번 문제에 대한 답인지를 반드시 표시하시오. 늦은 제출은 받지 않으니, 제출시한에 맞추도록 유의하시오. 반드시 자신들의 이해를 충분히 보이는 방법으로, 자신들의 Words를 사용하여 답안을 작성하시오. 질문있을 시 언제든지 [ijkim@mju.ac.kr](mailto:ijkim@mju.ac.kr)로 연락하시오.

1. **수업시간에 활용한 ToyotaCorolla.csv 데이터를 사용하여 아래와 같이 실행하고 답하시오.**

**Predicting Prices of Used Cars (중고차 가격 예측). ToyotaCorolla.csv파일은 2004년 여름 후반 동안 네덜란드에서 판매된 중고차(도요타 코롤라)에 관한 데이터를 포함하고 있다. 이 파일은 1,436개의 레코드와 가격 (Price), 사용기간 (Age), 주행거리 (Kilometer), 마력 (HP) 등 38개의 상세 속성들을 포함하고 있다. 이 특성들을 기반으로 중고 도요타 코롤라의 가격을 예측하고자 한다 (앞선 수업에서의 예제는 이 데이터의 일부분에 속함).**

1. **(6pts) 데이터를 training (50%), validation (30%), 그리고 test (20%) 세개의 데이터셋으로 나누시오. 그리고 가격(Price)을 출력변수로 하고 입력변수로는 Age\_08\_04, KM, Fuel\_Type, HP, Automatic, Doors, Quarterly\_Tax, Mfg\_Guarantee, Guarantee\_Period, Airco, Automatic\_Airco, CD\_Player, Powered\_Windows, Sport\_Model, Tow\_Bar 를 설정한 후 multiple linear regression (다중선형모델)을 실행하시오.**
2. **(3pts)자동차 가격을 예측하는데 있어 가장 중요한 3~4개의 자동차 속성(car specification)은 무엇인지 쓰고 그 근거를 설명해 보시오.**

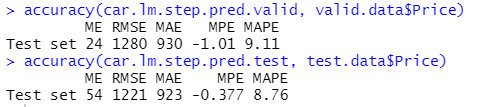
**Age\_08\_04, Automatic\_airco, KM, Quarterly\_Tax이다. 왜냐하면 R코드의**

**#1-(2)부분을 보았을 때 forward방식을 사용하여 predict 변수가 추가되면 AIC가 가장 많이 줄어드는 속성 4개이기 때문이다.**

**이 4개의 변수를 추가하였을 때 모델의 정확도가 올라가고 예측값을 계산하는 데에 도움이 되기 때문에 가장 중요한 자동차 속성은 Age\_08\_04, Automatic\_airco, KM, Quarterly\_Tax이다.**

1. **(3pts) 유용하다고 생각하는 평가지표(metrics)를 사용하여, 가격 예측 모델의 성능(performance)을 평가하시오. Validation data 와 test data에 각각 테스트를 해 보고 prediction accuracy를 평가해 보시오.**

**Validation data 와 test data에 forward방식으로 각각 테스트를 해보고 accuracy를 확인했을 때 결과는 다음과 같다.**



**전체 데이터의 에러 평균의 절댓값을 나타내는 MAE는 valid.data가 더 크다.**

**퍼센트 에러의 평균을 나타내는 MPE는 valid.data가 더 크다.**

**퍼센트 에러 평균의 절댓값을 나타내는 MAPE는 valid.data가 더 크다.**

**(에러의 제곱 / 개수) 의 제곱근을 나타내는RMSE는 valid.data가 더 크다.**

**에러의 평균을 나타내는 ME는 test.data가 더 크다.**

**이렇게 확인해보았을 때 ME를 제외한 모든 값이 valid.data가 크다.**

**Accuracy의 값이 작을수록 데이터의 예측이 정확하다는 의미이므로 그 값이 상대적으로 적은 test.data의 값이 predict accuracy가 더 정확하다고 할 수 있다.**