RandomForest와 Support Vector Machine을 활용한 차량 운전자 분류 모델 구축

차도둑들

*경희대학교 경영학과(박규리), 서울시립대학교 경제학부(황이은),
동국대학교 식품산업관리학과(장청아)

I. 서론

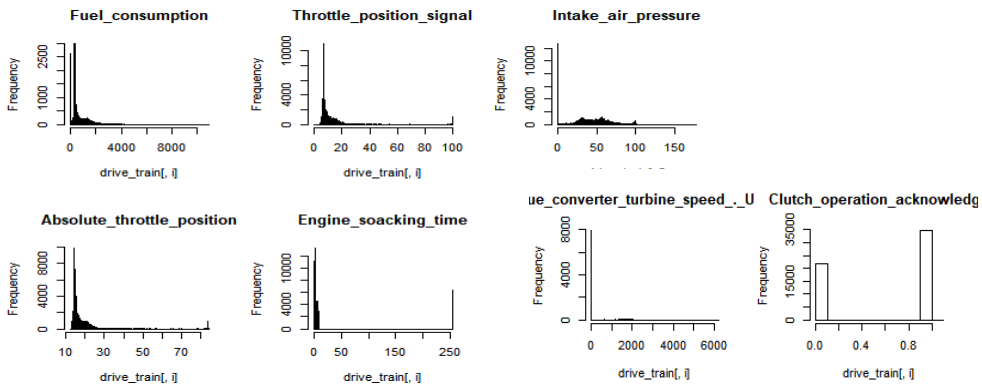
본 연구에서는 차량운전데이터를 통해 각각의 운전자를 얼마나 정확하게 분류할 수 있는가에 대한 연구를 진행하였습니다. 본 연구는 운전자별로 운전 행동이 다르며, 조작 불가능한 변수가 있다고 있다는 가정하에 실행하였습니다. 이 연구의 목적은 운전자 분류에 영향을 미치는 중요 변수를 도출하고 도출된 변수들을 토대로 운전자 분류의 정확성을 계산하는 것입니다. 통계 기반 feature selection을 통해 총 29개의 중요변수를 도출했으며, RandomForest와 SVM을 실시해 운전자를 분류한 결과, 각 99.91711%, 94.448%의 정확성을 보였습니다.

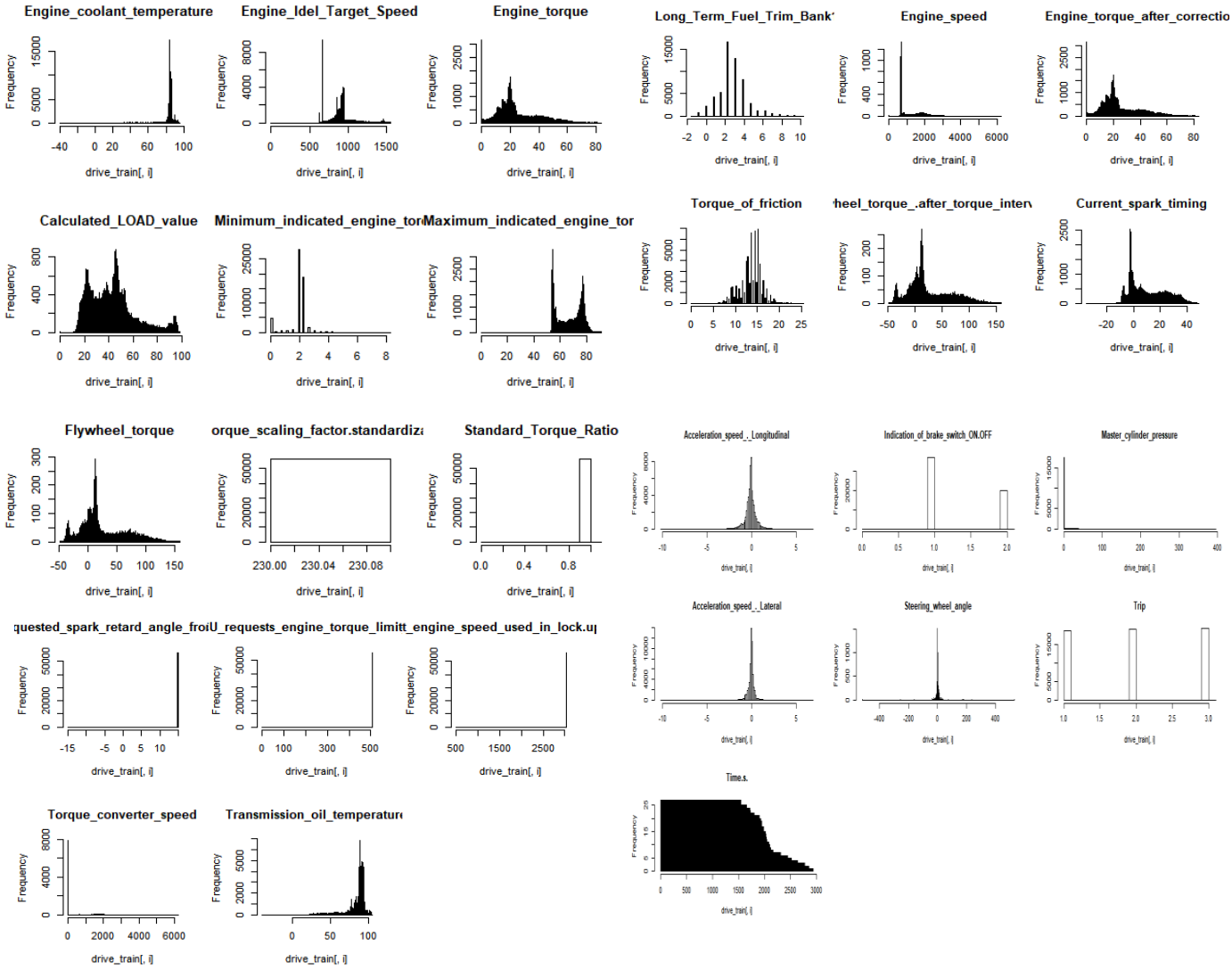
II. 방법론

2.1 데이터셋 분석결과

운전자 분류 모델 구축의 진행은 변수의 특성을 살펴보는 것으로 시작하였습니다. 54개의 변수들을 살펴보고자 각 변수들별로 히스토그램을 그려보았습니다.

[그림1. 변수별 히스토그램]



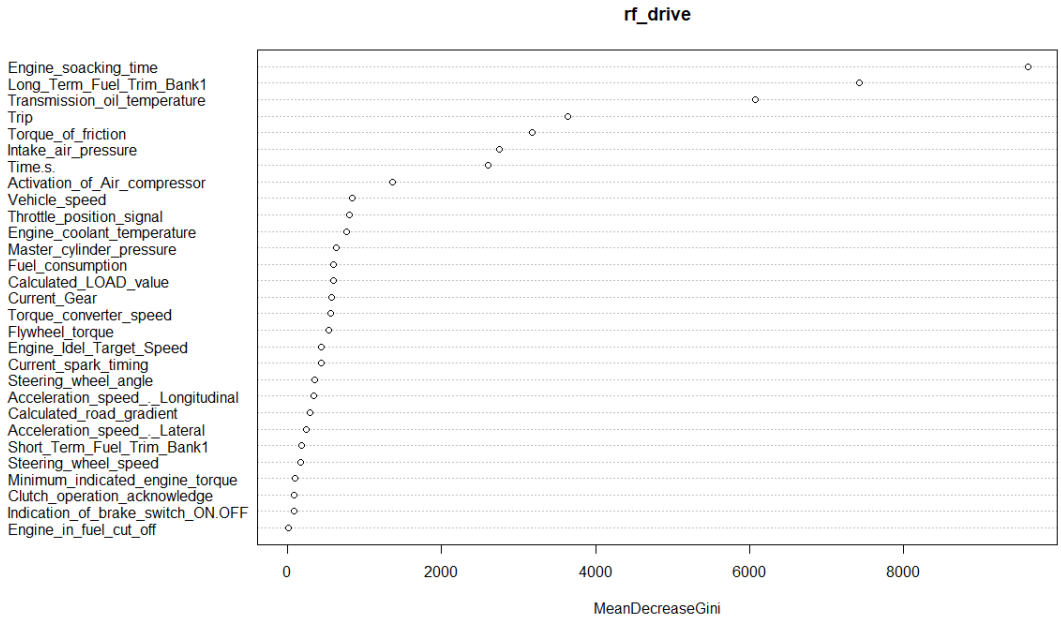


데이터를 탐색해본 결과 변수값이 존재하지 않거나 분산이 적은 변수들이 다수 발견되어 예측에 앞서 feature selection을 하였습니다. feature selection은 관측값이 모두 0인변수 제거, 분산이 0에 가까운 변수 제거, 변수들간 상관성이 높은 변수를 제거하는 식으로 하였습니다.

먼저 분산이 0에 가까운 변수를 제거했습니다. 분산이 0에 가깝다는 것은 분류하는데 큰 영향을 미치지 못하다는 것을 의미합니다. Caret package에 있는 nearZeroVar 함수를 이용하여 분산이 0에 가까운 12개의 변수들을 제거하였습니다. 또한 남은 변수중에서 Filtered_Accelerator_Pedal_value, Inhibition_of_engine_fuel_cut_off, Fuel_Pressure 변수는 모든 관측치의 값이 0이어서 삭제했습니다. 세번째로는 남은 변수들의 상관관계를 분석하여 상관성이 높은 변수를 제거하였습니다. 변수들 간 상관성이 높으면 모델의

신뢰성과 정확도가 떨어지기 때문에 제거해주는 것이 좋다고 생각했기 때문입니다. FindCorrelation 함수를 이용하여 상관성이 높은 변수들을 제거하고 나면 fuel_Consumption 변수를 포함하여 29개의 변수가 남게 됩니다(분류변수인 Driver 변수 제외). 29개의 변수들을 가지고 분류하는데 중요한 변수들을 분석해보았습니다. 위에서부터 중요도가 높은 순이며 상위 3개 중요도가 높은 변수는 Engine_soaking_time,Long_Term_Fuel_Time_Bank1,Transmission_oil_temperature 으로 나타납니다.

[그림2. RandomForest 변수 중요도 결과]



저희 팀은 분산과 상관성 처리를 해준 29개의 변수를 가지고 분류했을 때와 29개의 변수중에서 실제로 운전자가 조작할 수 있는 변수들만을 가지고 분류했을 때의 정확도를 비교해보았습니다. 우선 29개의 변수 중에서 운전자가 직접적으로 조작할 수 있다고 판단된 변수는 Current_Gear, Clutch_operation_acknowledge, Vehicle_speed, Acceleration_speed_._Longitudinal,Acceleration_speed_._Lateral,Indication_of_brake_switch_ON.OFF,Steering_wheel_speed,Steering_wheel_angle로 총 8개입니다. 이 변수들을 선택하게 된 이유는 참고논문과 배경지식을 바탕으로 기어,클러치,속도,가속페달,브레이크,핸들만을 운전자가 조작할 수 있다고 생각했기 때문입니다. 그 외의 변수들은 엔진이나 연료와 관련되어서 운전자가 직접적으로 제어할 수 있는 부분이 아니라고 판단하였습니다.

2.2. 알고리즘 설명

데이터 분류 알고리즘으로는 RandomForest와 SVM을 사용하였습니다. RandomForest는 앙상블 학습방법의 일종으로 결정트리들을 다수 생성하고 학습시켜 다수결의 결과를 도출하는 원리로 작동이 됩니다. 분류에서 널리 사용되고 높은 정확도를 가진다는 장점을 가지고 있습니다. SVM은 지도학습 방법 중 분류하는데 사용됩니다. 분류기법 중 정확도 측면에서 우수하다고 평가받고 있으며 다양한 변수가 존재할 때 효과적인 알고리즘입니다.

III. 프로그램 설명

모든 통계분석은 가장 보편적으로 쓰이는 통계 툴인 R을 이용하였습니다.

IV. 실험

모든 통계분석은 *Know Your Master: Driver Profiling-based Anti-theft Method* 라는 논문에서 실험된 데이터를 바탕으로 하였습니다

V. 평가

변수 선택을 통해 남은 29개의 변수를 가지고 위의 상기된 이유로 RandomForest와 SVM을 실시하였습니다. 10-fold cross validation로 검증한 결과, RandomForest는 약 99.91711% 의 정확도를 보였습니다. 또한 SVM을 돌려본 결과 94.448%의 정확도출 도출할 수 있었습니다. 결과적으로 SVM과 RandomForest 모두 높은 정확도를 보였습니다. 이는, 데이터가 대부분 연속형 변수라는 점에서 SVM이 유리했으며, RandomForest는 많은 의사결정나무를 만들어 투표를 하여 과적합을 막는다는 점에서 유리했다고 생각합니다. 또한 29개 변수 중에서 운전자가 통제 가능하고, 운전자의 특성을 나타낸다고 판단되는 8개 변수[1] 가지고 모델링을 해본 결과 정확도 30%를 웃도는 굉장히 낮은 결과를 얻었습니다. 이는, 차원이 낮은 데이터를 가지고 모델링을 한 이유도 있지만, 운전자의 특성이 다른 변수에 비해 영향일 크지 않기 때문이라고 볼 수도 있습니다. 또한 trip별로 정확도 또한 확인하기 위해 trip 1,2,3 데이터를 분리하여 모델링을 해보았으나 이 3 타입이 굉장

[1]Current_Gear,Clutch_operation_acknowledge,Vehicle_speed,Acceleration_speed_._Longitudinal,Acceleration_speed_._Lateral,Indication_of_brake_switch_ON.OFF,Acceleration_speed_._Lateral,Steering_wheel_speed, Steering_wheel_angle,Drive

히 미미한 차이를 보였습니다. 이것으로 보아 trip별 차이는 그리 크다고 볼 수 없다고 볼 수 있을 것 같습니다.

[표1. Trip열 더미변수 처리 및 분리 후 정확도]

	모든 데이터	Trip1 분리	Trip2 분리	Trip3 분리
RandomForest 정확도	0.99886	0.9988	0.9997	0.9993

여러 모델을 만들어 본 결과 feature selection된 변수 29개를 가지고 RandomForest 모델링 한 것이 가장 정확도가 높았습니다. 그 모델을 가지고 test set을 예측해 볼 예정입니다. 운전자를 분류하는데 있어서 운전자가 통제할 수 있는 변수들이 중요성을 가질 것이라 생각하였으나 그렇지 않은 결과가 나왔습니다. 이에 대해 저희는 9명이라는 적은 수의 운전자 데이터셋은 운전자의 특성을 반영하여 분류해내기에는 충분치 못하다고 생각했습니다. 그래서 운전자가 통제할 수 없는 변수들이 영향을 더 미친 것이라고 추측하였습니다. 데이터 볼륨의 자체적인 한계로 인해 운전자 특성 차이가 다른 변수보다 미미한 차이로 나타났을 수도 있다고 생각합니다.

VI. 결론

현 프로젝트에서는 데이터 탐색,feature selection, 모델링을 하였고, 평가까지 마쳤습니다. 운전자를 분류한다는 새로운 컨셉의 분류는 다양한 시도가 가능했으며, 데이터를 만지면 만질수록 다양한 면모를 볼 수 있다는 점에서 신선한 연구였다고 보여집니다. 운전자 분류 모델은 테스트했을 때 100%에 가까운 높은 정확률을 보였습니다. 하지만 9명밖에 되지 않는 운전자 데이터의 한계로 일반화할 수 있는 모델링을 하지 못했다고 보여집니다. 앞으로 더 질 높은 데이터로 볼륨을 늘린다면 운전자를 정확하게 분류할 수 있는 일반적인 모델을 만들 수 있을 것이라 생각합니다.

[참고문헌]

[1] Byung Il Kwak, Jiyoung Woo, and Huy Kang Kim. “*Know Your Master: Driver Profiling-based Anti-theft Method.*” PST (Privacy, Security and Trust). IEEE, 2016