

# RandomForest와 Support Vector Machine을 활용한 차량 운전자 분류 모델 구축

차량주행 데이터 기반 도난탐지 트랙  
발표일자\_2018.11.30  
팀명\_차도둑들  
팀원\_박규리, 장철아, 황이은

# CONTENTS

1

연구 설명  
-데이터 설명

2

예선  
-Feature Selection  
-모델선정  
-분석결과

3

세션1  
-Feature Selection  
-모델선정  
-분석결과

4

세션2  
-Feature Selection  
-모델선정  
-분석결과

# 1. 연구 설명

2. 예선

3. 세션1

4. 세션2

# 1. 연구 설명



예선: 54개 변수 56700개 관측치  
세션1: 54개 변수 34569개 관측치  
세션2: 54개 변수 108224개 관측치




운전자 분류에 영향을 미치는 중요 변수를 도출하고 도출된 변수들을  
토대로 운전자 분류의 정확성을 계산



통계 기반 feature selection을 통해 총 29개의 중요변수를 도출했으며,  
RandomForest와 SVM을 실시

>> 목표\_ 운전 행동 데이터를 바탕으로 9명/5명의 운전자를 분류





1. 연구설명

2. 예선

3. 세션1

4. 세션2

## 2. 데이터 전처리

예선 변수 선택(feature selection) - 통계 기반



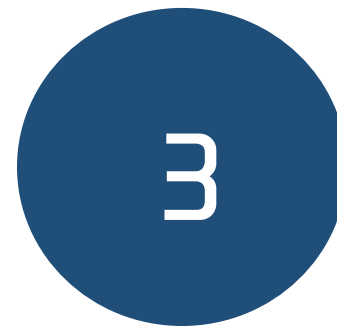
분산이 0에 가까운 변수 제거

nearZeroVar 함수를  
이용하여 분산이 0에 가까운  
12개의 변수들을 제거



모든 관측치가 0인 변수 제거

0값 만을 갖는 4개의  
Field 제거



상관성이 높은 변수를 제거

FindCorrelation 함수를 이용하여  
상관성이 높은  
9개의 변수 제거

## 2. 데이터 전처리

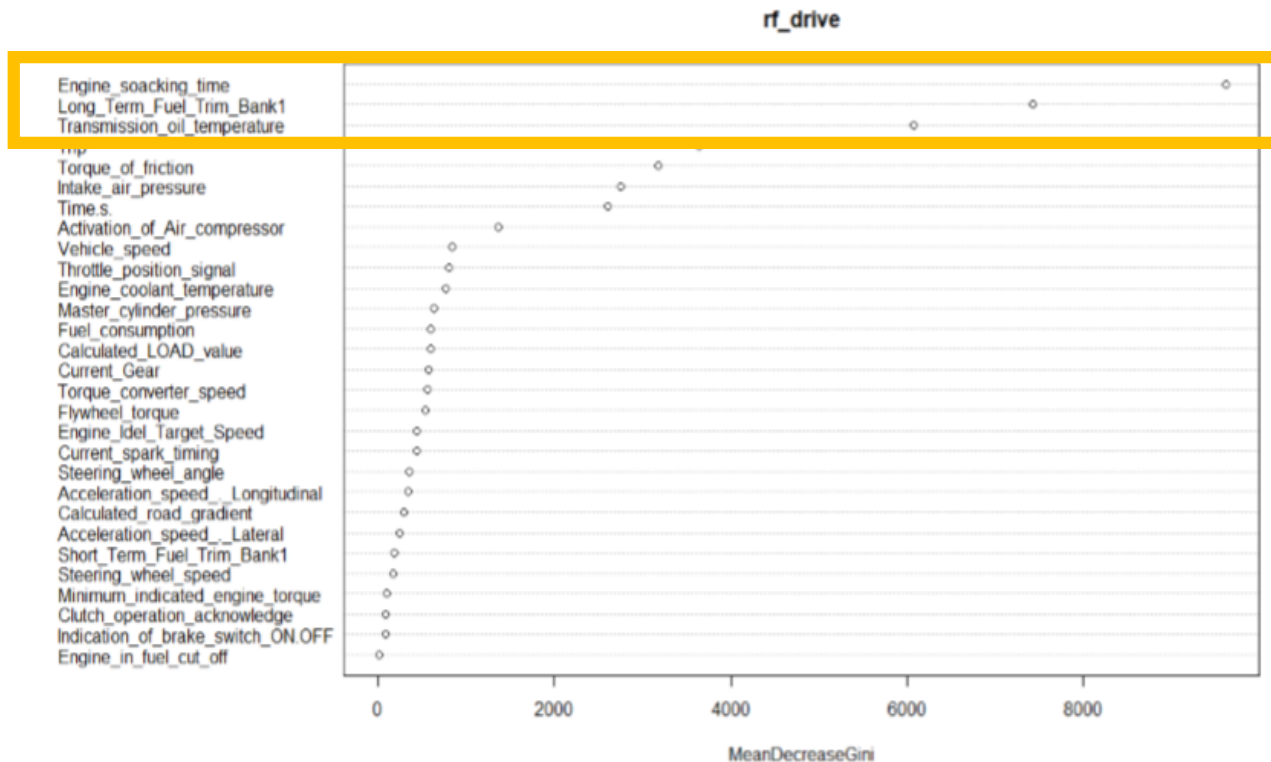
### 29개의 변수 선택

"Fuel_consumption"	"Throttle_position_signal"	"Short_Term_Fuel_Trim_Bank1"	"Intake_air_pressure"
"Engine_soaking_time"	"Engine_in_fuel_cut_off"	"Long_Term_Fuel_Trim_Bank1"	"Torque_of_friction"
"Current_spark_timing"	"Engine_coolant_temperature"	"Engine_Idel_Target_Speed"	"Calculated_LOAD_value"
"Minimum_indicated_engine_torque"	"Flywheel_torque"	"Activation_of_Air_compressor"	"Torque_converter_speed"
"Current_Gear"	"Transmission_oil_temperature"	"Clutch_operation_acknowledge"	"Vehicle_speed"
"Acceleration_speed_.Longitudinal"	"Indication_of_brake_switch_ON.OFF"	"Master_cylinder_pressure"	"Calculated_road_gradient"
"Acceleration_speed_.Lateral"	"Steering_wheel_speed"	"Steering_wheel_angle"	"Trip"
			"Time.s."

## 2. 데이터 전처리

### RandomForest 변수 중요도

29개의 변수들을 가지고 분류하는데 중요한 변수들을 분석



상위 3개 중요도가 높은 변수는

〰〰

Engine\_soaking\_time  
Long\_Term\_Fuel\_Time\_Bank1  
Transmission\_oil\_temperature

〰〰

로 나타남



## 2. 데이터 전처리

### 변수 선택2

실제로 운전자가 조작할 수 있는 변수들만을 가지고 분류했을 때의 정확도를 비교

운전자가 조작할 수 있는 변수



기어



클러치



속도



가속페달



브레이크



핸들

## 2. 데이터 전처리

총 8개의 변수 선택

Current\_Gear

Clutch\_operation\_acknowledge

Vehicle\_speed

Acceleration\_speed\_.\_Longitudinal

Acceleration\_speed\_.\_Lateral

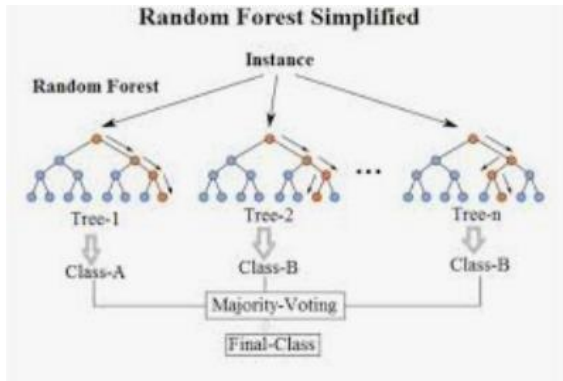
Indication\_of\_brake\_switch\_ON.OFF

Steering\_wheel\_speed

Steering\_wheel\_angle

### 3. 모델 선정

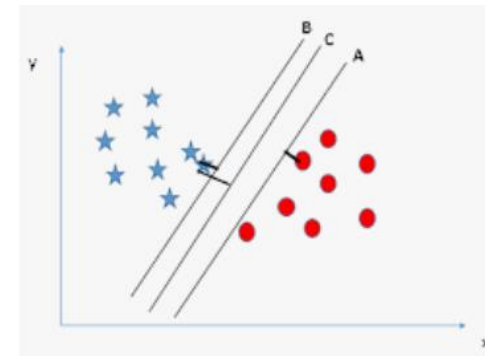
#### RandomForest



앙상블 학습방법의 일종으로 결정트리들을  
다수 생성하고 학습시켜 다수결의 결과를  
도출하는 원리로 작동됨.

장점 :분류에서 널리 사용되고 높은 정확도를 가짐

#### SVM



지도학습 방법 중 분류하는데 사용됨

장점: 정확도 측면에서 우수하다고 평가받고 있으며  
다양한 변수가 존재할 때 효과적인 알고리즘

## 4. 분석 결과

### 정확도

29개의 변수로 모델링 실시

10-fold cross validation로 검증

RandomForest

99.91711%

SVM

94.448%

->데이터가 대부분 연속형 변수라는 점에서 SVM이 유리

->RandomForest는 과적합을 막는다는 점에서 유리

## 4. 분석결과

### 정확도

8개 변수로 모델링

정확도 30%를 웃도는 굉장히 낮은 결과

RandomForest

34.43913%

SVM

25.6067%

->차원이 낮은 데이터

-> 운전자가 조작할 수 있는 변수가 다른 변수들에 비해 영향이 크지 않음

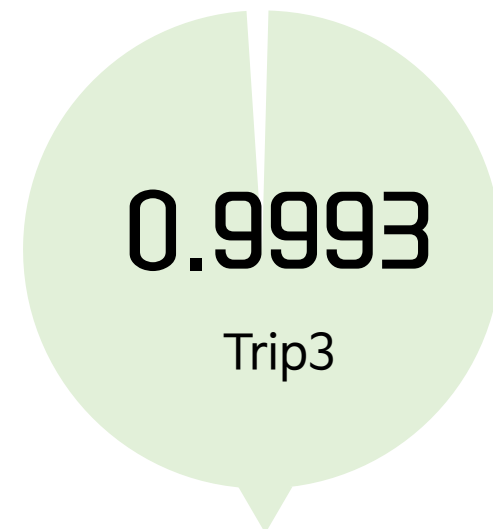
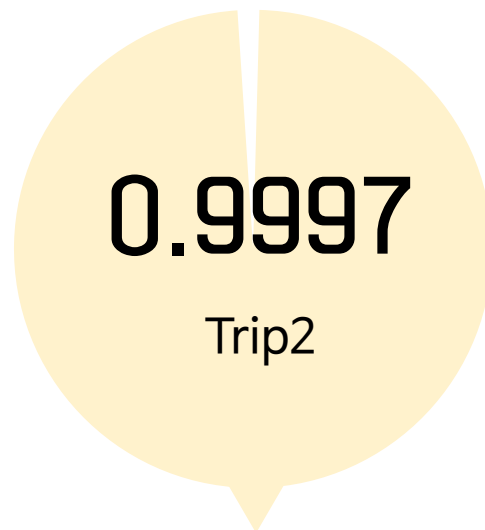
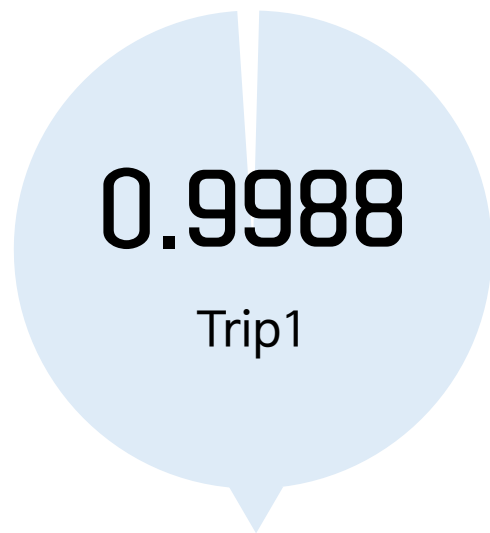
-> 9명이라는 적은 수의 운전자 데이터셋이 운전자의 특성을 반영하여 분류하기에 충분치 못하다 판단함



## 4. 분석 결과

### Trip Type별 정확도

Trip 1,2,3 데이터를 분리하여 모델링  
Trip type별 정확도의 차이가 근소함  
RandomForest Accuracy





1. 연구설명

2. 예선

3. 세션 1

4. 세션 2

## 2. 데이터 전처리

세션2 변수 선택(feature selection) - 모델기반



분산이 0에 가까운 변수 제거

nearZeroVar 함수를  
이용하여 분산이 0에 가까운  
12개의 변수들을 제거



모든 관측치가 0인 변수 제거

TRIP1만 선택



데이터 분포 확인

TRAIN TEST데이터를  
직접보면서  
변수 선택

## 2. 데이터 전처리

### 27개의 변수 선택

"Fuel_consumption"	"Throttle_position_signal"	"Short_Term_Fuel_Trim_Bank1"	"Intake_air_pressure"
"Engine_soaking_time"	"Engine_in_fuel_cut_off"	"Long_Term_Fuel_Trim_Bank1"	"Torque_of_friction"
"Current_spark_timing"	"Engine_coolant_temperature"	"Engine_Idel_Target_Speed"	"Calculated_LOAD_value"
"Minimum_indicated_engine_torque"	"Flywheel_torque"	"Activation_of_Air_compressor"	"Torque_converter_speed"
"Current_Gear"	"Transmission_oil_temperature"	"Clutch_operation_acknowledge"	"Vehicle_speed"
"Acceleration_speed_.Longitudinal"	"Indication_of_brake_switch_ON.OFF"	"Master_cylinder_pressure"	"Calculated_road_gradient"
"Acceleration_speed_.Lateral"	"Steering_wheel_speed"	"Steering_wheel_angle"	

## 4. 분석 결과

### 정확도

27개의 변수로 모델링 실시


10-fold cross validation로 검증

RandomForest

24%

-> RandomForest는 과적합을 막는다는 점에서 유리





1. 연구설명

2. 예선

3. 세션1

4. 세션 2

## 2. 데이터 전처리

세션1 변수 선택(feature selection) - 데이터 기반



모든 관측치가 0인 변수 제거

0값만을 갖는  
Field 제거

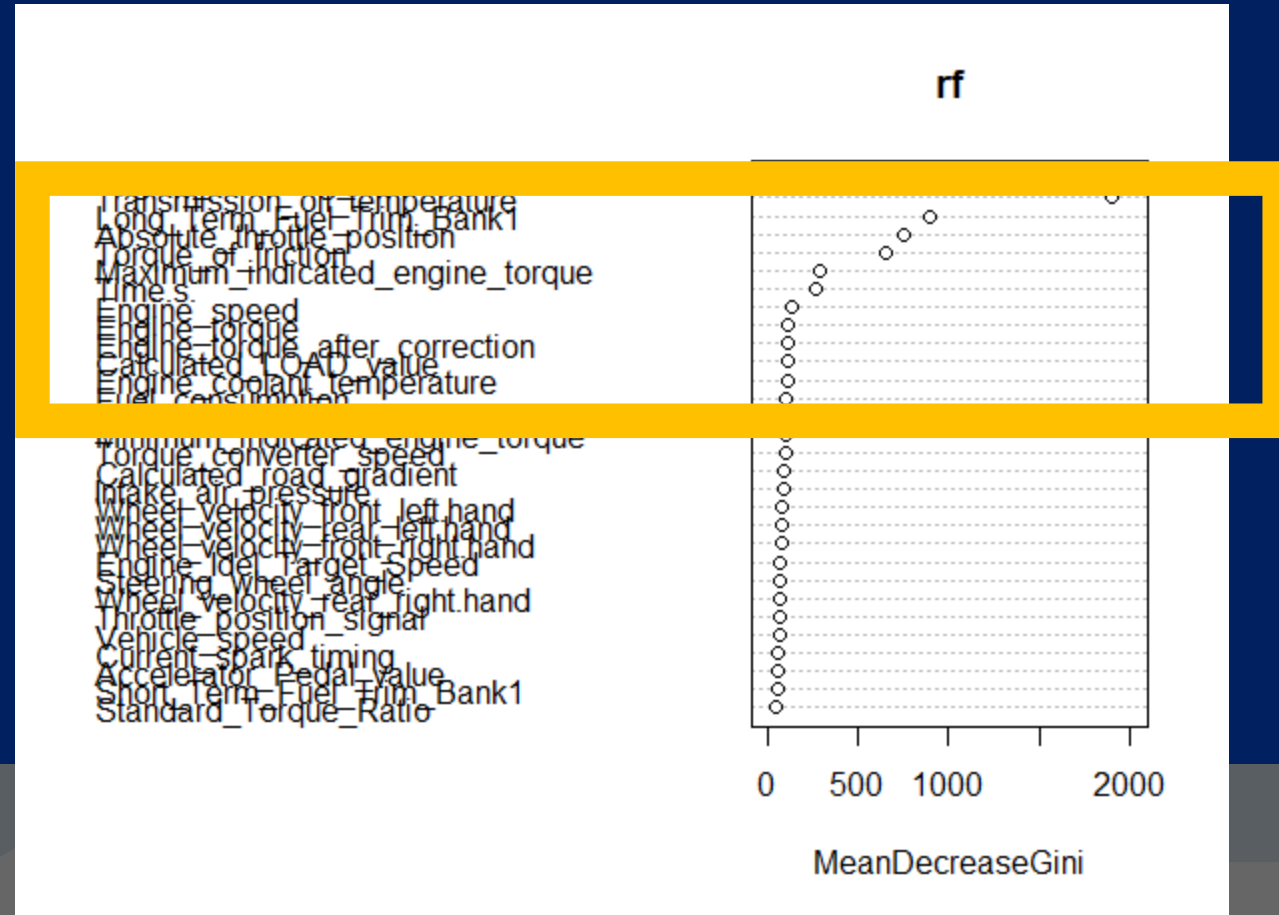


랜덤포레스트 변수 정확도 돌림

랜덤포레스트의 분석 요소인  
변수 정확도를 중심으로 돌림

## 2. 데이터 전처리

36개의 변수 선택



## 4. 분석 결과

### 정확도

29개의 변수로 모델링 실시

10-fold cross validation로 검증

RandomForest

48%

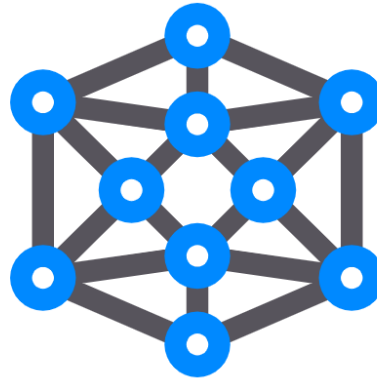
->RandomForest는 과적합을 막는다는 점에서 유리

## 4. 분석 결과

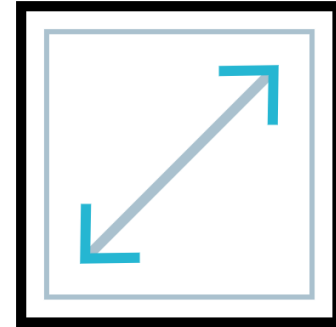
### 한계점



9 명밖에 되지 않는  
운전자 데이터셋의 한계



일반화 할 수 있는 모델링을 하지 못함



데이터로 볼륨을 늘리는 것이 필요함



A close-up, low-angle shot of a person's hands on a steering wheel at night. The driver is wearing a blue jacket with a red ribbed cuff and a silver watch with a blue face. The background is a blurred cityscape at night, with warm lights from buildings and streetlights creating a bokeh effect. The car's dashboard and a digital display are partially visible on the right side of the frame.

THANK YOU  
&  
Q & A