Technische Universität München

Data Mining Cup 2

Description

The topic of the second data mining cup is predictive maintenance, which is currently a very popular application area of predictive analytics.

The dataset is about predicting defects of a specific engine part in cars. There are at least two use cases, where a sufficient prediction model is of high value for both the car manufacture and the car driver: During a car inspection, the model can be used to indicate that the engine part should be replaced to prevent a future breakdown. While driving, the driver could be warned about an imminent engine defect and the chances of an accident are decreased when the driver brings the car to inspection.

Originally, the dataset is from a project between a leading German car manufacturer and Alexander Thamm Data Science Services GmbH. We highly appreciate that both have agreed to sponsor the dataset for academic purposes. The project is already finished and at the end of the data mining cup Alexander Thamm's data scientists will share insight of their solution. Furthermore, the three best teams will be awarded with an additional prize on top of the exam bonus points.

Classification Task

Within the scope of this dataset, a defect is defined as a failure of a specific engine component. Your task is to determine whether a system readout, including several different information about the car and engine, can be used to predict if the engine part will be defective.

Feature Overview

Feature	Description	Example
system_readout_id	The observation unit. This is unique identifier for records in the dataset. Predictions file should contain this feature.	"15" … "214459"
vehicle_identification_number	This is a unique identifier for each car.	"15" … "145302"
engine_type	There are five different car engine types.	"1", "2", "3", "4", "6"
vehicle_type	There are ten different car types.	"1", "2", "3", "4", "5", "6", "7", "8", "9", "20"
mileage	The amount of km the car has been driven	
engine_feature_2 to engine_feature_15	There are 14 features describing measure values from engine control units and other attributes such as e.g. fuel consumption. The car manufacturer is not providing further information regarding these features and their measurement.	
engine_feature_#_1 engine_feature_#_2 … engine_feature_#_n	One feature that is split into n columns	
engine_feature_11_1 to engine_feature_11_48	One feature that is split into 48 columns. This data represents points from a normalised histogram. All this features should approximately sum up to 100.	
engine_feature_14_1 to engine_feature_14_5	One feature that is split into 5 columns. This data represents points from a normalised histogram. All this features should approximately sum up to 100.	
defect	Dependent variable that should be predicted	"y":= engine part is defective "n":= engine part is not defective