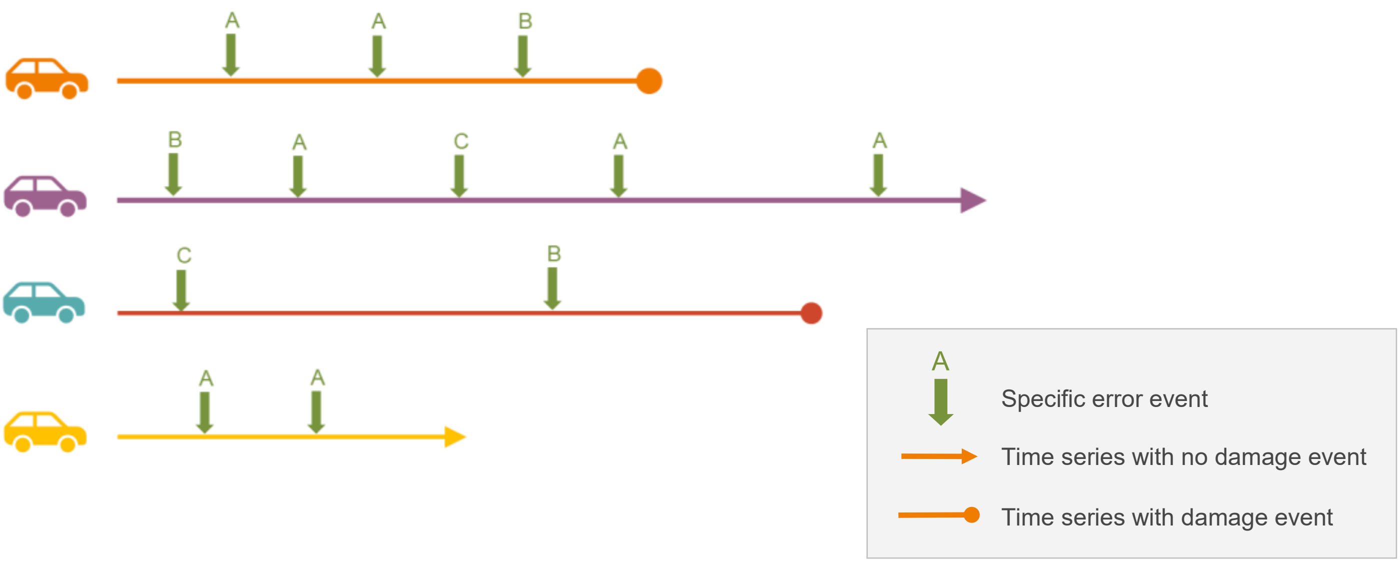
Failure prediction in innovative vehicle projects - how bio statistical approaches can be transferred to automotive Projects

Project Goal

The development and the first market launch of very innovative vehicles with new technologies involves particular challenges, especially for the development of a quality management framework. There is no historic information and the process cannot be based on past experiences. For the quality management it is important to identify potential damage types and their drivers or patterns. For newly developed cars there is no previous field information where patterns of car configurations and error patterns lead to specific car failures. The goal of this project was to develop an algorithm, that is able to identify clusters of error patterns and car configurations, that lead to certain car failures. Importantly, the clustering needs to take into account that the project started in the development phase of the vehicle so that the data provided is provided only by a small amount of test vehicles in different development stages and that the cars are in varying states of their life cycle and that the algorithm needs to re-learn as soon as new data comes in. The identified clusters were evaluated and labeled by experts, which allows for the training of a failure type classifier. To solve this challenge a biostatistical approach was transferred it into the automotive industry.

Provided Data

Each sample of the data consisted of information on the configuration of the individual car, data from the production process, the normalized log of the error messages from different control modules, text mined data from customer problem descriptions, repair reports and the current runtime of the car.



Challenges

There were a couple of major challenges to approach for the development of the clustering algorithm.

First of all, most of the provided features were nominal data with categories of labels. A feature engineering was needed here, since most of the clustering algorithms only work with numerical data.

The provided data consisted of car records taken at different time points of their life cycle and many data points were censored. In effect, for most entries the actual failure has not yet taken place.

There were a couple of error sources, which might have biased the precision. This included false repairs and inherited error messages, which both might cause misleading conclusions about the real failure.

The clusters needed to be self-adjusting to the incoming data-stream and interpretable according to their major drivers, that an expert can evaluate the clusters according to their failure pattern.

Applied Methods

To deal with the problem of censored data, we implemented a feature engineering approach, which accounts for the different stages of the life cycle. The error event variables were re-engineered, so that they relate the relevance of a certain error event for an individual car with the corresponding relevance of the error event for all cars. These variables were then weighted by time. In that way, series of error logs could be grouped according to their relative relevance and compared across time. This allowed to build clusters of failure types for cars that are in different stages of their life cycle.

To address the challenge with the nominal nature of the data, the provided features were projected into the numerical space by applying Multiple Correspondence Analysis, which at the same time also reduced dimensionality and saved computational power.

The re-engineered and projected features were then used for the clustering algorithms. The best algorithm was the DBSCAN, which was selected through the silhouette coefficient.

Based on the identified clusters of the principal components, the major drivers of the clusters were determined. An individual survival analysis for each cluster was performed to determine the time to a damage event. A co-occurrence score was further calculated to identify error log series, that lead to the same damage patterns.

After an expert revision of the identified clusters according to their prevalent failure type, a classifier was trained, that can classify new cars into their most likely failure pattern.

Project Outcome

The developed dynamic algorithm can be used to identify clusters of common damage patterns, that were spread across different car configurations.

Cars in different stages of their life cycle can now be incorporated in the clustering.

New car variations can be classified into the dominating damage pattern. This allows for the development of a suitable quality management strategy.