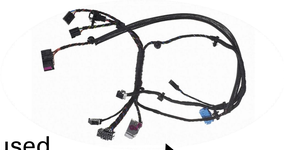
Using algorithmic feature selection to classify wiring harness data

Nick Melzner in cooperation with eXXcellent solutions

Problem and Background

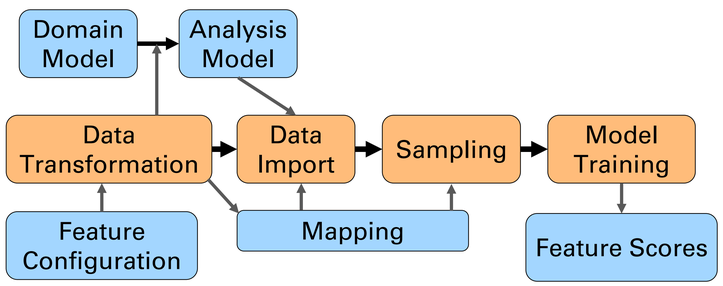
- EXXcellent manages Vehicle **Wiring Harness data** for a client in the German automotive industry
- Vehicles are evaluated using "Static Checks"
- Static Checks can not be performed on every vehicle because of their high cost
- Thus, Classification models shall be used to find data inconsistencies and flag vehicles, that should be further examined using Static Checks



exXcellent solutions

Data Preparation Pipeline

- The existing product stores Wiring harness data in a relational database model, Vehicles refer to contained Wiring Harnesses, these contain Cables, Connectors and other parts
- Data in this form is unsuitable for training models in ML libraries, it had to be transformed into a **non-relational model**
- This model contains information about features like the total wire length or numbers of certain parts for 111489 vehicle configurations
- Every step of data processing is implemented within an automated pipeline:

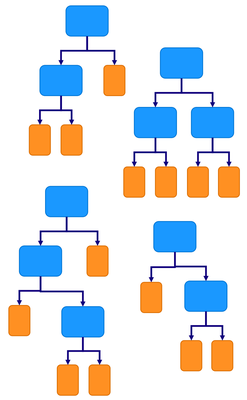


Comparing Different Models

- Goal of this thesis is to find correlations between vehicle properties like "vehicle series" or "sales country"
- Goal is to determine which features are most important to classify vehicles
- Results of different classification models will be compared for finding a suitable method to detect relevant features

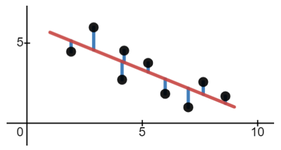
Classification - Random Forest

- Decision trees classify data following criteria like "wire length > 1000"
- 200 trees vote for the classification of each vehicle
- Random Forest reduces the decision tree's risk of overfitting
- Scikit-Learn implements methods to identify the most important features in a forest



Regression – Lasso and ElasticNet

- Advanced form of Linear Regression
- Examines linear relationships between features
- Can be used to filter out non-important features



- ElasticNet combines two regression approaches:

Lasso

Results are highly interpretable, flags features as non-relevant

Ridge Regression

Performs better than Lasso when features are strongly correlated

Results

- Random Forest outperformed ElasticNet in training speed
- Accuracies between both models were similar after tuning
- Using ElasticNet to pre-select features improved the performance of RandomForest
- Scores were label-dependent as expected

Accuracy of ElasticNet, Random Forest and Random Forest tuned using ElasticNet:

	ElasticNet	Random Forest (RF)	ElasticNet tuned RF
Best	0.99	0.98	**0.99**
Worst	0.77	0.62	**0.82**
Median	0.87	0.89	**0.92**

Results for different labels, performance was stable within each label

Conclusion and Next steps

- ElasticNet is more suitable for collecting detailed information about feature importance (selection per label value is possible)
- Random Forest and ElasticNet in combination perform well in classifying vehicles using wiring harness data (99% best, 82% worst, 92% average)
- Random Forest is a good quantifier to evaluate the quality of a feature set provided by ElasticNet

Next steps:

- Finish ElasticNet tuning of RandomForest to find the optimal feature-set
- Wrap up training pipeline implementation