IEOR 135/290: Data-X – 1-2 page writeup (pre-Final Report)

Predicting service operations from connected car telematics

Julian Kudszus, Nicholas Hirons, Spencer Lee, Soham Kudtarkar

1. Project overview

The goal of our project was to analyze the data that is captured by a vehicle's engine control unit to improve the process of car repairs, maintenance, and ownership for all those involved. This includes dealerships and repair shops who would be able to better diagnose and repair vehicle problems, as well as drivers who could use our platform to mitigate unexpected vehicle breakdowns, maintain their vehicles more reliably, and gain more insight into the inner workings of their vehicle. In the past, much of the data that is processed by the engine control unit was either obscured or not easily visible to the end user (whoever that may be). We are capturing that data and making it useful. The data is routed from the vehicle to our own servers, where we will apply our models and find any useful info.

2. The problem

The car repair process is subject to many uncertainties, and our project aims to solve a number of problems for both drivers and car dealerships/repair shops. It can be difficult for dealerships and repair shops to identify issues with a vehicle due to complications such as reproducibility of the problem and a lack of clarity on the vehicle's history. Currently, mechanics spend a lot of time diagnosing car issues and performing tests, while customers often feel that there is a lack of transparency and necessity for certain service operations. Our solution aims to change this: we are able to take standard information on a car, as well as any error codes from car telematics data, and infer what service operations are likely to be performed and by extension, what issues may need immediate addressing. These insights allow dealerships and auto repair shops to preemptively reach out to customers, before their vehicle experiences potentially very problematic issues that cause a breakdown. As CarForce's initial work has shown, car servicers have increased business by 2x by using this data and reaching out directly to customers. Our data and analysis also makes the service and repair process more efficient by suggesting likely causes and solutions for car issues, while also providing more transparency to consumers ahead of any required service.

3. Our solution

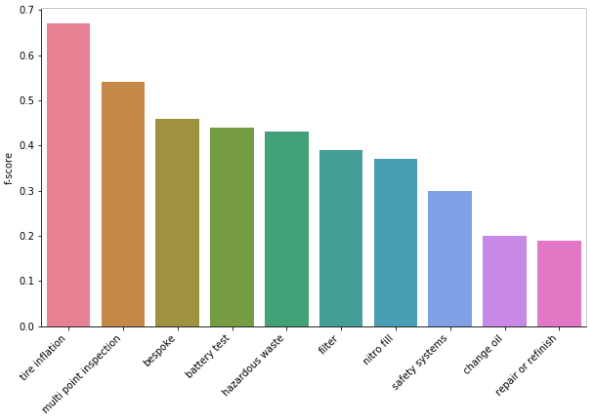
Our solution uses standard vehicle information and engine error codes in order to determine likely services that would need to be performed. This allows for dealerships to identify a range of issues that may not fall into their typical patterns, or take longer to find if they did not have our data to predict likely issues. Our project also allows customers to take better care of their vehicles by being able to look at their vehicle's history and likely future, so they can prepare for any otherwise unexpected issues.

We compared the performance of random forests, logistic regression, SVMs and k-Nearest Neighbors, with a parameter grid search for each algorithm using stratified 5-fold cross validation on 75% of the data. Given class imbalance in our target variables (each service operation appeared in no more than a third of services, with most occurring around 10% of the time), our key metric was the F1-score, which we evaluated on the 25% test set after selecting the best parameters from grid search CV. Using this approach, k-Nearest Neighbors significantly outperformed the other classification algorithms and we present a summary of the best parameters and key metrics on the following page.

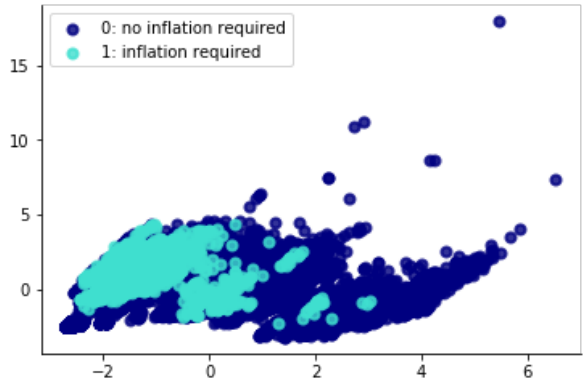
target variable / service	leaf size	n neighbors	accuracy	precision	recall	f-score
tire inflation	30	3	0.92	0.66	0.67	0.67
multi point inspection	30	3	0.68	0.55	0.53	0.54
bespoke	100	1	0.65	0.45	0.46	0.46
battery test	10	1	0.73	0.45	0.42	0.44
hazardous waste	100	1	0.91	0.41	0.45	0.43
filter	10	1	0.69	0.4	0.38	0.39
nitro fill	100	3	0.95	0.41	0.34	0.37
safety systems	10	1	0.88	0.31	0.29	0.30
change oil	10	1	0.85	0.20	0.20	0.20
repair or refinish	100	1	0.80	0.21	0.18	0.19

Examining the first two principal components of the data, classes did not appear to be easily separable, but local clusters of data could be seen (and it intuitively makes sense that similar cars in similar condition would need similar services). This may explain why an algorithm that is sensitive to the local structure of data, such as k-NN, outperformed the other separating/input-splitting algorithms. However, even k-NN was only effective for a few services, and in general, we did not get strong F-scores on test data.

F-score for predicted service operations



First two principal components vs. tire inflation



Our final product involves a clean, simple web app that uses these models to allow end users to infer what service operations are likely to be required for a car. A prototype can be found in our GitHub repo. The system that our solution could be integrated to is the one described by Jessika Lora of CarForce, in which data is streamed from vehicles to our own servers which will run our analysis and display useful conclusions back to the end user (typically dealership and service providers, but potentially consumers) via a web application. For our analysis, we primarily used Python scripts and Jupyter notebooks.