IEOR 135/290: Data-X - Final Report

**Predicting service operations from connected car telematics**

*Julian Kudszus, Nicholas Hirons, Spencer Lee, Soham Kudtarkar*

**1. The pitch**

Our project harnesses vehicle data on every level, from make and mileage to the error codes from engine control units, in order to improve the experience of car ownership and repair for drivers, dealerships, and repair shops alike. By being able to predict future services needed on a vehicle, we can preemptively save customers from failures that may arise if a vehicle is left unmaintained. Tailoring recommendations to each specific vehicle offers an invaluable tool for mechanics, who will be able to more quickly identify issues in a vehicle by knowing the most likely problems and required services.

**2. The problem**

Our project aims to solve a number of problems for both drivers and car dealerships/repair shops. 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. 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 journey**

Initially, we had developed approaches, potential questions and avenues to pursue without a complete picture of our data (as we did not receive it until late in the semester). Some of the questions and ideas for our project (such as using error codes to predict future vehicle behavior) became unfeasible once we received the data. Tools that we used included Python, Jupyter notebooks, a variety of the libraries that we’ve used in class (more details visible on our GitHub), and Shiny in R.

**4. Our solution**

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 below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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 with tire inflation** |
| https://lh3.googleusercontent.com/wj4aQXao_5GdH2jjo00rtlNUKNGAZQDjTyqrrSnBvkErotqkrYdShZmF0CrBafyax73M6SSWZs5Rz4Ic3JYymFbjpCCOX-7VhNYD59vyxu9_-BapqRT4WdGbfdfFgZOv8RJTpGmk | https://lh5.googleusercontent.com/xc7lP-aoFHaa9YMkSbtybGGwlB8sSAWfvzGp3scwR8Zp3MxZ77yaJ6M6EEkb3wjk6X_9q5Ucl0GP4Ts35x_Ok538sMtiSAvF34GqCF-x_G1tYRNmKn6IOo9o42NSO3BfeRnbkvOg |

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.

**5. Future improvements**

A key future improvement we could make to our project is creating models for each make of vehicle (e.g. Ford, Acura) in our database. We believe this would significantly increase precision and recall, as our models would be identifying predictive information only within a specific subset for each make, rather than over the entire population. It would also allow us to capture more information in the ‘tail end’ of the make distribution; in the current format, most makes were mapped to ‘other’ during one-hot-encoding, potentially eliminating valuable signals in the data. This would mean analyzing makes (and potentially even models of vehicles) independently to gain a deeper understanding of the issues faced by those cars in particular, rather than generalizing our solution unnecessarily.

There are a number of key limitations to this approach: (1) to do this for a large number of makes and models, we would need significantly more data, (2) training, inferring and updating models in live software could be prohibitively expensive and (3) the models may overfit to smaller sets of historical data and not be able to infer anything valuable from unseen cases.

**6. Team contributions**

Chronologically: Julian made contact with our mentor at the student mentor mixer held before semester began. Julian formed a team with Spencer, which Nick and Soham soon joined. As a team, we discussed directions for the project with Jessika Lora (our mentor) and brainstormed potential approaches and ways to make our data and analysis useful. For our low tech demo, Soham focused on reporting what our data looked like and spearheaded UI design and initial concepts, Julian developed examples of ways we envisioned the project working, and Nick and Spencer identified algorithms and techniques that could be useful to us. Nick further contributed to the design of the slides. Spencer and Julian worked on outlining our architecture.

Once we received our data, Nick led the analysis efforts and handled data cleaning/parsing as well as model formulation and evaluation. Spencer contributed to data analysis as well and explored an extension by prototyping separate models for each make. Soham built a working mockup of an application UI to allow end users to leverage our algorithms. Julian led the efforts on the 1-2 page and 2-3 page write ups with results/discussion from Nick. Nick led the slide design and all team members contributed to the final slides and presentation, focusing on the areas where they had contributed in the low tech demo and throughout the project.

**7. Mentor experience**

Our mentor experience was not ideal. While engaged with the project early on, contact became unreliable and we did not receive our data until mid November after repeated delays. However, the context for our project and how it is useful to end users was provided in part by Jessika Lora of CarForce. We unfortunately did not receive any replies after we were sent the data, so our understanding of the full data set was somewhat limited to obvious, well-labelled columns, and the final direction of the project based on the data was decided upon independently of our mentor. *Update 12/6/17: Jessika has offered to set up the call to assist with the project, unfortunately we have completed the project at this point.*