Automated Repair Code Labelling Utilizing Machine Learning Algorithms

Ali El-Khatib

**Introduction**

Similar to the healthcare industry, the transportation industry requires extensive documentation of procedures and events. Upon initial receipt of a vehicle for maintenance and repair activities, technicians collect a variety of information about its current state. This information is comprised of numerical data (e.g., coolant levels, total mileage, oil pressure) as well as descriptive (e.g., driver complaints, technician’s initial impressions). After compiling the necessary information, a technician creates a repair order (RO) detailing all detected abnormalities as well as the repairs to be applied. Simultaneously, the identified repairs are applied to the vehicle, with the respective supervisor reviews the RO. Ultimately, the RO and all corresponding information of this event is stored within a data warehouse for future business analytics.

With the aforementioned process, the gathered data poses a challenge for business intelligence (BI) and analytics activities. Namely, the ROs are unable to be grouped by similar repairs without prior extensive, time-consuming labelling. This is typically done by sending the dataset to a subject-matter expert (SME), who analyses the various fields and applies the appropriate repair code label. This process results in a significant time delay in developing BI reports, in addition to a significant time commitment from the already-hectic schedules of the SMEs. By utilizing machine-learning algorithms, a model can be used to predict the appropriate repair code labeling for each repair job, ultimately resulting in a decrease processing time for generating accurate business intelligence reports.

**Background**

The purpose of this project is to predict the appropriate component repair code for different repairs as documented by various technicians. The data provided consists of a mixture of human-inputted text fields, categorical data, and a numerical field.

The dataset is comprised of the following fields:

* **Complaint** (text) – Description of problem with vehicle obtained from driver
* **Cause** (text) – Technician’s diagnosis of defective components leading to driver’s complaint
* **Correction** (text) – Solution applied by technician
* **Job Notes** (text) – Additional job-specific information included by technician
* **Vehicle Category** (category) – Category of vehicle (Truck, Trailer, Tractor)
* **Vehicle Model** (category) – Vehicle model (numerous different models, see dataset)
* **Cost** (float) – Total dollar amount needed to apply solution including labor and parts

**Note**: Due to the sensitive nature of the dataset, some or all of the fields will be anonymized and/or redacted. This will be in compliance with company protocols to ensure the protection of any vital information regarding business operations.

**Methods**

The purpose of this project is to predict the appropriate component repair code for different repairs as documented by various technicians. This will be done by utilizing supervised machine learning techniques on a labelled and verified dataset (i.e., a correctly labelled dataset).

Before applying the machine learning algorithms, the text fields will need to be converted to an appropriate datatype, specifically . This will be done by utilizing the term frequency-inverse document frequency (TFIDF) method. In TFIDF, each text value is converted to a matrix of weight factors indicating the importance of each word within that text field. Stop words, or words that do not contribute any relevant information (e.g., and, is, was, the, etc.), will be eliminated to optimize the model and reduce potential noise. Likewise, accents and other extraneous characters will be removed.

After preprocessing, the dataset will be fitted to a number of different algorithms. For each of the algorithms, various parameter settings will be applied to identify the best settings within each model. At this time, the identified algorithms are the following:

* Ridge Regression Classifier
* Support Vector Classifier
* Random Forest Classifier
* K-Nearest Neighbors Classifier

To analyze the effectiveness of the models, each model will predict the labels for the test dataset. The resulting labels will then be compared to the true label values and an accuracy score will be generated representing the effectiveness of each model.

**References**

All information is first-hand from my personal experience within the transportation / trucking industry. That is to say, this information primarily from the operations within my current business.