Master Universitario di I Livello in "ANALISI DATI PER LA BUSINESS INTELLIGENCE E DATA SCIENCE", Università degli Studi di Torino

A.A. 2015-2016

Titolo della tesi: ETL and mining of on-board diagnostics data of construction equipment: the case of Tierra's telematics

Autore: Lucia Salvatori

Abstract

The aim of this study is to investigate a sample dataset from Tierra's telematics, applying ETL and data mining techniques in order to evaluate if a predictive analysis for vehicle failures could be implemented in mid-term to support preventive maintenance cycle of construction equipment. The project was developed during a four months internship c/o Tierra, a joint venture company established in 2008 by Topcon Positioning Systems and Divitech and based in Northern Turin area.

Tierra develops telematics solutions and produces devices that provide users vehicle parameters and diagnostic messages for working machines fleets (90.000 devices in 20 countries). Tierra aims to improve its offered services by using data collected by vehicle telematics, in order to provide: analytics for fleet management, predictive rules to improve maintenance cycle and support for precision agriculture.

The analysis was focused on on-board diagnostics data, compliant with SAE J1939, a family of standards used for car, highway and off-highway equipment, agricultural equipment, construction equipment and other vehicles. In particular Active Diagnostic Trouble Codes (Diagnostic Message 1 / DM1) at t_0 and Controller Area Networking messages (CAN messages) on vehicle parameters (fuel consumption, position, speed, diagnostic of mechanical components, etc.) at t_{-1} and t_{-2} and were analyzed.

Data are stored in a highly normalized structure designed using Hadoop Distributed File System (with several hundred millions measurements per month); the extraction process from was achieved using a direct SQL access via Cloudera Impala; it was preferred to Knime Big Data Connectors for performance and flexibility reasons.

As no data warehouse with metadata is yet available, a deep data cleaning and transformation process was needed for the following analysis. An “R” script ensures a robust reshaping of the dataset, highlighting also relationships between CANs and DM1s by the intersection of timestamps.

Furthermore, a detailed descriptive statistics was performed on the dataset, with a special in-depth analysis on more frequent failures (DM1 ‘652|5’, Injector Cylinder #2 - Current below normal; DM1 ‘652|5’: Engine Intake Manifold Temperature - Value above normal). All activities were developed in “R” again, allowing a fast and standardized reporting procedure, that uses a template with tables and charts (barplot, boxplot, Pareto’s chart).

In conclusion, a logit model was applied on the most frequent failures, using step-wise regressions on numerical parameters. As input, it was used a complete set and a restricted one, selected by expert judgment. The results show that the predictive power of model strongly relies on DM1 considered and on available CANs.

About DM1 ‘652|5’, very poor explanation power was found; it was concluded that further investigation of equipment operations and data collection are needed. Concerning DM1 ‘105|15’, the results provide some support for predictive analysis, as a mix of environmental condition, vehicles history and contingent conditions at t_{-1} and t_{-2} could significantly explain an important amount of failures; variables at t_{-2} a could play a role as a predictive warning of an incoming failure.

Additional investigation on the predictive power of missing and outliers have been developed, without getting an increased explanation power; for further analysis, the use of a longer series of t_{-n} on a longer-lasting observation together with the use categorical variables will require a significant increase of computation resources.