## Project Proposal

**Project Title -** https://github.com/rgamerosl/capstone-project

**Title:** Fuel Consumption in the Ready-mix trucks

**Abstract:**
Roughly speaking the idea is to detect the main factors that contributes to the daily fuel consumption of the ready-mix trucks. In simple terms, if possible, the company would like to have like an equation that relates the fuel consumption (y-target: liters consumed per hour) in terms of the main external factors (predictors: manufacturer, year model, idle time, traveled kilometers, location, driver's driving habits, etc.). Afterwards, once the model is constructed it will be used to detect the ready-mix trucks whose fuel consumption is out of the ordinary in order to do a detail investigation and if required have a feedback talk with the corresponding driver looking for its improvement.

## Problem Statement & Business Case

The problem to be solved is to construct a common fuel consumption range for the different ready-mix trucks taking into account different external (not manipulable) factors as are:

- the different manufacturers for the trucks,
- the year model of the truck and
- some other internal factors intrinsic to the daily operation:
    - the city/location on which the truck operates,
    - if it supplies local or foreign deliveries,
    - duration of the idle times due to clients/costumers site conditions, etc.

Afterwards, the goal is to identify the manipulable factors (mainly driver's driving habits and maybe something else) that could be improved, and monetize these improvements (measure how these improvements could save fuel which in the end will translate in savings for the operation of CEMEX Mexico). Also, this model could help to identify unusual high fuel consumption which could be related to mechanical failures on the truck or in the worst-case scenario to some fuel stealing by the driver.

## Data Science Workflow

What Null/Alternative Hypothesis are you testing against?
We want to know how the X factors (not manipulable and manipulable factors) affects the target value Y (fuel consumption measure as liters of fuel consumed per hour).

What solutions would you like to deliver against?
The best-scenario is to deliver an equation model in which you input all the X factors and gave you a prediction for the target value Y.

What benchmarks are you looking to automate?
Identify usual range for the fuel consumption taking into account intrinsic factors to usual variations on the operation along the country.

What is the business case for your project?
Due to different conditions on the operation, one cannot expect the same fuel consumption for all the trucks. Therefore, the idea is to give a flexible goal taking into account those variations and with that in mind define a realistic goal for each group of trucks under that share the same conditions. Now with flexible baselines as the fuel consumption goal the company could really identify those trucks that are having unusual high fuel consumption and take actions to improve that, those actions will mainly be:

- Feedback talks with the driver regarding his driving habits,
- Mechanical checks for the trucks

How will your solution help generate revenue, reduce costs, or impact another Key Performance Indicator or Objective Key Result?

The model will help generate savings on fuel, which could add to important amounts of money taking into consideration that there are usually around 1000 – 1300 trucks operating around the country daily.

Who will be impacted (Executive Stakeholders/Sponsors) by your solution? Who is your ideal client/customer?

Cemex operations, specifically the batching plant managers or the cluster managers since it will help them improve their operations and generate important savings to the company.

## Data Collection

Already have a complete database for daily fuel consumption along the year 2020 for each ready-mix truck. Also, I have some other dataset regarding driving habits (measuring some relevant events) in a daily basis that is provided by the telemetry/GPS provider which could be crossed over with the original database to include relevant factors regarding the fuel consumption.

## Data Processing, Preparation, & Feature Engineering

First, I will perform some exploratory data analysis in order to understand main relations between the variables, identify some missing-values and decide how to handle those. Furthermore, outliers could play a relevant role in this model so they should not be discarded.

## Machine Learning: Model Selection

Mainly I am planning to apply a Linear Regression. Also, I think maybe Random Forest or Gradient Boosting could be useful, but not as easy to interpret as a Linear Regression Analysis.

## Model Persistence: Deployment, Training, & Data Pipelines

Train and test split will be relevant, specially trying to keep constant some proportions (truck manufacturer and year model mainly I will guess at the beginning, maybe something else could present as important in the EDA).