Attribution modelling and Relative Importance

Mehdi Farhangian

The Importance of a factor depends on how importance is defined

By using multiple regression, we explore the extent to which each variable contributes to the prediction of the criterion. The word "importance" has a fuzzy nature open to different interpretations and the importance of a factor depends on how importance is defined. By variable importance, some people mean theoretical importance that refers to the change in the dependent variable based on a change in the predictor variable that can be measured using the regression coefficient. Other people take importance to mean the increase in the score of a dependent variable measured by the unstandardised regression coefficient. This interpretation is popular in economics. Finally, another group of people take importance to mean dispersion importance which is popular in behavioural science and refers to the amount of variance of the dependent variable explained by the regression equation that is attributable to each predictor variable.

Relative Importance and Digital marketing

Digital analysts are constantly asked to answer the question: What are the most important marketing channels our company should focus on? Determining the true importance of each marketing channel is called attribution modelling. You can read more about attribution modelling here, here and here. Attribution modelling has seen a lot of interest in measuring relative importance in recent years. Variable importance in regression refers to the quantification of an individual regressor's contribution to multiple regression models. In this article, I shed light on the black box model of Google 360 by investigating different regression models, such as the Shapley value regression and pmvd (a relatively new proposed model by Fledman) and several other models. In addition to multivariate linear regression, some other models such as random forest have received a lot of attention recently for assessment of variable importance. Please note I wrote this article initially here

zero-order correlation

In an associated study, relative importance is defined as percentage contribution. The simplest measure of importance is the zero-order correlation in which importance is defined as the direct predictive ability of the predictor's variable. In order to have a big picture of the importance of channels, you can start by visualising them in terms of their correlations to the conversion.

Multiple-analysis methods

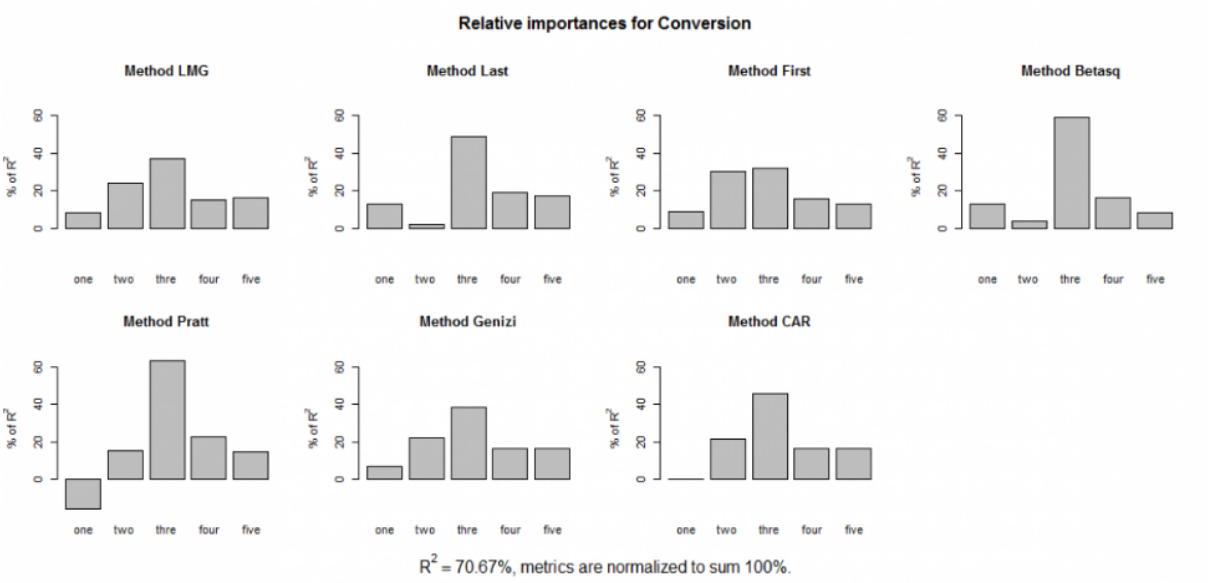
Correlation gives us the overall picture. However, if we define importance as the extent of the predictive ability of each factor in conjunction with other factors, the correlation is lacking and fails to consider the effect of each predictor in the context of other predictors. We need to measure the relative importance using regression to know what is changing in the expected value by changing the predicted variable.

R has a package for calculating relative importance. Ulrike Grömping has written an R package called relaimp that calculates relative importance. A description of the package is available in the Journal of Statistical Software.

Using this package, we can build a model that measures the relative importance of the factors with different metrics. In the following example, we measure the relative importance of five channels on conversions with 7 metrics.

Shapley value

Google's Data-Driven Attribution is based on a method known as the Shapley Value, but what is this?



The Shapley Value originated in 1953 by Nobel Prize winning mathematician Lloyd Shapley. It is a concept in Game Theory, specifically in systems where many factors need to 'cooperate' to achieve a given outcome. The Shapley Value is a way of allocating credit for the total outcome achieved amongst these many cooperating factors.

A simple analogy for building our intuition is that of a soccer game. If the striker scores most goals, he or she will traditionally get all of the credit (this is effectively Last Interaction attribution as the striker got the last touch before the ball went in the goal). But what if these goals were actually because of a brilliant pass from the midfield? A pass so perfectly timed that anyone could have simply tapped the ball in, surely the credit must be given to the mid-fielder? If this is the case, the striker is not adding much 'marginal-value'. Would the team score as many goals if the brilliant midfielder was not playing?

Clearly, it is not just what factors are involved, but how they interact, and in what order that is truly important to understand how they contribute to the overall goal.

Going back to a digital analytics example, we can see a comparison of two pathways below. Including 'Display' at this point in the sequence increases the likelihood of a purchase by 50%. Therefore we can attribute this increase to 'Display' despite it being only a link in the sequence. To get the complete credit given to 'Display', more comparisons need to be made with 'Display' occurring at different locations and working with different touch points.

I'll Add r codes and vidualisations here