PROBLEM STATEMENT

The question every advertiser wants to answer is this: How can I spend my ad dollars to get the most sales at the lowest cost-per-sale? I have a 2017 dataset for television advertising placements, the resulting leads (someone calls the phone number listed on the ad) and sales (the lead converts to a sale - on average, 10 weeks later).

My approach will be to first construct a model to determine how the factors of a TV ad placement affect the conversion of a viewer from a lead to a sale. Then I would determine the average conversion rate from a lead to a sale, at the most granular level possible. Applying the conversion rate to the cost-per-lead, we could predict the cost-per-sale. This would give us to the desired end result: determine the optimal media placement to hit the most efficient cost-per-sale.

Hypothesis:

* Station will have the largest impact on the conversion from a lead to a sale.

Assumptions:

* All media buying and sales data is accurate
* All sales are created equal (ie. Doesn’t matter the age, gender, etc of the sale)
* All placements are being purchased at the lowest rate possible (ie, media buyers are negotiating and getting the lowest rates possible – the model cannot recommend a different rate of purchase)

Goals for the project:

* Determine the conversion rate from lead to sale
* Use the conversion rate to calculate the cost-per-sale from any cost-per-lead

Success metrics:

* Find a way to spend all ad dollars keeping cost-per-sale below $5000

Risks/Limitations:

* There are so many factors to conversion contained in the ad buy, but there are also many external factors as well. This project will have to be done inside a vacuum of sorts, assuming all external factors are essentially remaining equal, however, this is most likely not the case. This is a limitation of a project at this scale. Certain external factors could be:
  + Competitor ad spending increases
  + Competitor rates drop OR Client rates increase/are less desirable
  + Client increases spending in other ad channels (direct mail, digital)

Data Set Details:

* I selected this data set because it is a perfect real-world example of using data to produce tangible, actionable results in a business case.
* 2017 data stored in MySQL table, could be exported to a CSV file.
* Over 400,000 rows and 72 columns
* No index/primary key currently exists
* All the data is maintained in one file – the inputs, which are the details of the ad buy (columns such as Air Type, Station, Creative) and the response (calls, lead, sale) as well as details about the response (age group, AARP member, etc)
* Some columns are probably unnecessary for this analysis and could be removed. There are also columns with inaccuracies that should be excluded. There are also columns that show the same information 2 ways, such as Station Code and Station Name. One of these should be removed, to avoid a severe collinearity problem.
* Some columns have sparse data, such as Program Name. This will result in a significant amount of null values and would make the effect of that field difficult to read, so it should be excluded.
* There are 2 date fields, Call Date and Date Aired – one of these will always be null while the other is filled in, because the data is a merged document of both airings and responses. It would make sense to create a Date field that pulls from whichever is populated, and use this new Date for the analysis.