# Summary Report

## Understand the business problem

It is important to fully understand business problem before working on the project. If not, the project will be easily turn to wrong way and all of our effort will be wasted.

### Identify the business problem:

As is:

- The typical lead conversion rate at X education is around 30%.

To be:

- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot.
- The model should be able to adjust to if the company's requirement changes in the future

## Build Data Model

### Cleaning data

There are a lot of variables that have null values. We have many methods to handle missing data: fill with mod, fill with mean, median, … But, I don't understand the reason behind why these variables have many missing values. It's best to ask the business to know the root cause before trying to fill null data otherwise, the model is unreliable. Thus, I chose to delete all null data.

### EDA

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. Data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.)

These are analysis steps in EDA:

- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis

### Model Building

I used Recursive Feature Elimination (RFE) to build the model because we have a lot of variables (more than 100). First, we get top 15 important features from RFE and then we manually remove variables that have large P value one by one because they are insignificant.

I did not realize of the importance of choosing evaluation metrics before coming to the cutting point when building the model. Thanks to this project, I can now understand which metrics we need to choose to identify the cutting point.

In this project, precision – recall is the correct metric for model evaluation because we care about the true positive, false negative and false positive.

## Presentation and Recommendations

I learn from many sources in tableau (viz of the day) the way they present to the audience. The slides them self should give the audience understand 80% the content you want to share for the time 5-10 seconds. I tried to make the presentation based on that concept: simple and informative.