# Introduction:

Project idea document suggested four projects. This document shall propose the Capstone Project required for Data Science Workshop – Intensive.

# Problem to be solved:

The project is to predict whether a mobile ad will be clicked or not. Click-Through-Rate (CTR) metric is used for evaluating ad performance and CTR systems are widely used by internet economy. The problem and solutions are applicable to current internet economy. The problem provides quite a good learning opportunity with a typical setup in real word with lots of data and lots of features.

# Project Client:

The project is completed Kaggle competition. There is already a solution available for the project and reward has been awarded to winning team.

The project client is myself to see if I am able apply advanced Machine Learning technologies and Data Science methodologies learned in this workshop to arrive to a solution that would match up to the top teams in the competition. Since the solution is available, there is a certainty to solution that one should arrive.

As mentioned before, the main aim here is to apply learning of this course to a real life problem to 1) refine the learnings further 2) demonstrate the learnings and 3) be ready to apply the learnings to the next projects.

# Data Set:

The data-set is available [here](https://www.kaggle.com/c/avazu-ctr-prediction/data). The data-set contains training and test data. The data-set is in csv format. The zipped training data, is about 1 gb in size and test data-set is about 118 Mb. The data-set is well defined with some of the categorical features anonymized.

# Approach:

## Dealing with data:

The data-set was used for data story exercise. The data-set in csv format about 6 GB. The first thing that has to happen is to find a small sample that can be used locally to run machine learning models.  
Once the model is ready. The models could be run on full dataset on servers in Amazon cloud.

## Models:

* Create a small sample that represents full data set for training purpose.
* Create a small sample for cross validation.
* Run multiple classification models to which one fits the best and check against the cross validation data set.
* Once the model receives reasonable accuracy, check its accuracy on test data-set.

# Deliverables:

The deliverables will be ipython notebook containing code and data story with code explanation and result analysis.