**CAPSTONE 2 – PROJECT IDEAS**

1. **House Price Prediction**

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

#### **Problem**

To predict the final price of each home according to the market prices, taking into account different features ranging from, the basic amenities, to that of its proximity to public transport hubs.

#### **Problem Category**

* Advanced Regression techniques like random forest and gradient boosting.

#### **Tasks**

#### To predict the sales price of the house.

#### For each ID in test set, predict the value of SalesPrice variable.

#### **Data used**

This consists of the following data files:

* **train.csv:** Training data for house price prediction
  + 1460 data points for houses (rows) and 81 features of those houses (or columns).
* **test.csv:** Testing data for house price prediction.
  + 1459 data points for houses (rows) and 80 features of those houses (or columns).

1. **New York City Taxi Trip Duration**

### https://www.kaggle.com/c/nyc-taxi-trip-duration/data

#### **Problem**

To predict the total ride duration of taxi trips in New York City.

#### **Problem Category**

* Regression Techniques

#### **Tasks**

* Based on individual trip attributes, predict the duration of each trip.

#### **Data used**

The dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform. The data was originally published by the NYC Taxi and Limousine Commission (TLC).

It contains the following files:

* **train.csv:** Training data for
  + It contains 1458644 trip records (rows) and 11 features (columns).
* **test.csv:** Testing data for.
  + It contains 625134 trip records (rows) and 9 features (columns).

1. **Diabetes 130 US hospitals for years 1999-2008**

## https://www.kaggle.com/brandao/diabetes

#### **Problem**

To predict if a patient will be readmitted in some hospital.

#### **Problem Category**

* Advanced Regression Techniques like Logistic Regression.

#### **Tasks**

There are 3 outputs given below, based on which, the readmission of a patient has to be figured out.

1. No Readmission.
2. A readmission in less than 30 days (this situation is not good, because maybe your treatment was not appropriate).
3. A readmission in more than 30 days (this one is not so good as well the last one, however, the reason can be the state of the patient.

#### **Data used**

Original Source of data:

<https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008>

The data set represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria:

* It is an inpatient encounter (a hospital admission).
* It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
* The length of stay was at least 1 day and at most 14 days.
* Laboratory tests were performed during the encounter.
* Medications were administered during the encounter.

The dataset contains the following file:

* **diabetic\_data.csv**
  + 101766 records (rows) and 50 attributes (or columns) of patients.