**PROJECT PROPOSAL**

## **Problem Statement Formation**

When we ask a home buyer to describe their dream house, they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. There are a number of features and criteria being looked at, based on which the pricing of a housing is estimated.

## **Context**

## The problem is connected to the sale of individual residential property in Ames, Iowa.

## from 2006 to 2010. We have to consider all features of each residential property; we have to provide a price of that particular property.

## **Criteria for Success**

* To predict the accuracy of current sales price of houses in Ames, Iowa.

## Scope of solution space

The scope of the solution illustrates various home features and includes sales price and use the given features of the home to estimate the value.

## **Data Sources**

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

## What is the problem you want to solve?

This is indicated by the Problem statement

## **What data are you using? How will you acquire the data?**

**File descriptions**

* **train.csv**
  + Questions contains the title, body, creation date, closed date (if applicable), score, and owner ID for all non-deleted Stack Overflow questions whose Id is a multiple of 10.
  + **Format** - CSV file that has 81 columns
* **test.csv** 
  + Answers contains the body, creation date, score, and owner ID for each of the answers to these questions. The ParentId column links back to the Questions table.
  + **Format** - CSV file that has 80 columns

**Data availability** – The data is available in the Kaggle community as indicated in the Data Sources section above

## **Method and Solution**

* Data Wrangling – Extraction and Cleaning. Drop the columns that have least effect on the result. Null values for integer and float features are filled with mean value of the particular feature, and, those in the categorical features are filled with the most frequently occurring category.
* Perform Exploratory Data analysis for understanding the data to find out correlations between numeric variables and target variable.
* Perform Feature Engineering to further filter the features and select those which have either positive or negative impact on the target feature SalePrice.
* Data Normalization to transform numeric variables to a common scale. This step transforms multiscaled data and after that, all variables have a similar influence on the model, improving the performance of the learning algorithm.
* Apply different regression models.

## **Deliverables**

* A GitHub repo containing the work completed for each step of the project, including:
  + A project report
  + A slide deck