# PROJECT SYNOPSIS

## PREDICTING HOUSE PRICES BY THEIR FEATURES

Step 1. Understand problem

1. Business objective and end goal : The company evaluates houses on various parameters and then gives it an actual value. If the house is undervalued the company buys it . Later the company sells it at higher price (actual predicted price) and makes a profit.
2. How is it done now? : by manual experts who judge it. However, the error rate is 25%

Step 2. Which model we will require

1. Supervised learning regression model because we have data with label and we want to find a value. It’s a batch learning because we have a fixed data to study which is not changing. In online learning the data keeps on changing eg- email spam

Step 3. Choosing a Performance measure

1. RMSE root mean square error- sqrt(1/n \* ( actual - predicted)^2 )

Why? Usually preferred in regression task

Step 4. Check my assumptions

1. We may have made some assumptions for the problem so cross check them before working on them. Eg I may assume that company needs value of price of the property however they require only 2 categories whether property is expensive or cheap

Step 5. Getting the data

1. Got housing data from uci (**University of California, Irvine**.)machine learning dataset

Converted the data to csv file  
506 data entries with no missing values

Boston housing data

Steps 6. Analysis of data

1. using histograms and dataframe.describe (mean std deviation percentile )to know about missing values or arbitrary large/small data (incorrect data )

Percentile in describe shows the value which is less than the given percentile

ie. 25% = 2.5 meaning top 25% of data is less than 2.5

<https://datascience.stackexchange.com/questions/52613/what-does-pandas-describe-percentiles-values-tell-about-our-data>

1. Using hist method we analysed histograms of all columns (value vs freq)

Steps 7. Data preprocessing

1. train test splitting
   * 1. we cannot use sklearn.preprocessing train test split because it may not divide CHAS values equally in train and test .
   1. splitting into test and training data using StratifiedShuffleSplit to preserve the percentage of CHAS coloumn

CHAS is **Charles River dummy variable** (= 1 if tract bounds **river ie house is beside river**;

0 otherwise

1. Looking for coorelatons
   1. Using df.corr(), we found pearson correlation index for each pairwise column- lies between [ -1 to 1]  
      if 1 then strongly corelated, ie increasing this would increase our column value
   2. Plotting scatter wrt each other and finding outliers. (house process are capped at 50k)
2. missing data can be managed by

1.deleting the row if missing data is less

2. deleting the entire column of missing data if the column is not that important

3. by filling 0 / mean/ median at its place we choose this

1. Scikit-learn Design

Primarily, three types of objects

1. Estimators - It estimates some parameter based on a dataset. Eg. imputer. It has a fit method and transform method. Fit method - Fits the dataset and calculates internal parameters
2. Transformers - transform method takes input and returns output based on the learnings from fit(). It also has a convenience function called fit\_transform() which fits and then transforms.
3. Predictors - LinearRegression model is an example of predictor. fit() and predict() are two common functions. It also gives score() function which will evaluate the predictions.
4. Feature Scaling

Primarily, two types of feature scaling methods:

1. Min-max scaling (Normalization) (value - min)/(max - min) Sklearn provides a class called MinMaxScaler for this
2. Standardization (value - mean)/std Sklearn provides a class called StandardScaler for this

This is better because a single value(min or max) does not affects to feature scaling

Also mean and standard deviation remains same

1. Creating a pipeline

Machine learning pipeline that remembers the complete set of preprocessing steps in the exact same order. So that whenever any new data point is introduced, the machine learning pipeline performs the steps as defined and uses the machine learning model to predict the target variable.

Pipeline is created only for data preprocessing

1. Choosing our model  
   Here we try with various models compare them using rmse. And choose the best after fine tuning them
2. Evaluating our model  
   RMSE – calculate the rmse of y pred with y label

cross validation - <https://www.analyticsvidhya.com/blog/2018/05/improve-model-performance-cross-validation-in-python-r/>

cross validation is done during training of data

1. Saving a machine learning Model
   1. In machine learning, while working with [scikit learn](https://www.geeksforgeeks.org/learning-model-building-scikit-learn-python-machine-learning-library/)library, we need to save the trained models in a file and restore them in order to reuse it to compare the model with other models, to test the model on a new data. The saving of data is called *Serializaion*, while restoring the data is called *Deserialization*.
   2. Also, we deal with different types and sizes of data. Some datasets are easily trained i.e- they take less time to train but the datasets whose size is large (more than 1GB) can take very large time to train on a local machine even with GPU. When we need the same trained data in some different project or later sometime, to avoid the wastage of the training time, store trained model so that it can be used anytime in the future.