# Problem Statement

Anomaly detection: Basis business goals and challenges, identify, track and action upon anomalies, which if ignored can have dire impact on businesses and users

**NOTE: anomalies are those column values which is having less importance percentage.**

# Proposed Solution

Here we are considering a scenario where we required data of past 10 or 20 year from an organisation and that data should be of annual turnover. This data will help us to understand well, which features or columns that contains values impacting the annual turnover.

**Objective**: Our main motive is to calculate the **percentage of importance** of columns that is impacting the business growth, after that we will consider only those columns which have higher **percentage of importance** because those columns will impact more to business growth and those columns which have less **percentage of importance** will impact less hence these columns will be consider as anomalies which we will handle separately.

**Now the next few steps we will perform to get more details**

1. Loading the data for analysis
2. Calculating the importance of every column which is related to annual turnover value.
3. Generating the report which shows how much importance does a particular column contains.

After performing above mentioned steps we will get one figure **(below)** stating out how much **percentage** **of importance** it plays in business and that **percentage** will help us to understand what are anomalies (**column values with less percentage**) which is impacting business growth. Along with this it will help us to get more clarity about which values are more important for generating more turnover for upcoming year.

# Detailed Steps

We have taken **Boston house pricing** dataset to give a small proof about how our idea is **practical and feasible**.

So, for calculating feature importance we need below mentioned things:

1. Boston house pricing dataset
2. **RandomForestRegressor**: It a ML Model which is required to calculate the feature importance of provided data.
3. **Matplotlib**:It is required for generating the plot/figure stating out importance of every column.
4. **Bayesian model Optimization using HyperOpt: it is required for hyper-parameter tuning of RandomForestRegressor Model.**

# This dataset contains below mentioned attributes details

These are the columns which is impacting the housing price, so if we are able to get **most important columns** from it then, we will able to see which values are

impacting most and least to house price.

**Attribute Information (in order):**

- CRIM per capita crime rate by town

- ZN proportion of residential land zoned for lots over 25,000 sq.ft.

- INDUS proportion of non-retail business acres per town

- CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

- NOX nitric oxides concentration (parts per 10 million)

- RM average number of rooms per dwelling

- AGE proportion of owner-occupied units built prior to 1940

- DIS weighted distances to five Boston employment centres

- RAD index of accessibility to radial highways

- TAX full-value property-tax rate per $10,000

- PTRATIO pupil-teacher ratio by town

- B 1000(Bk - 0.63)^2 where Bk is the proportion of black people by town

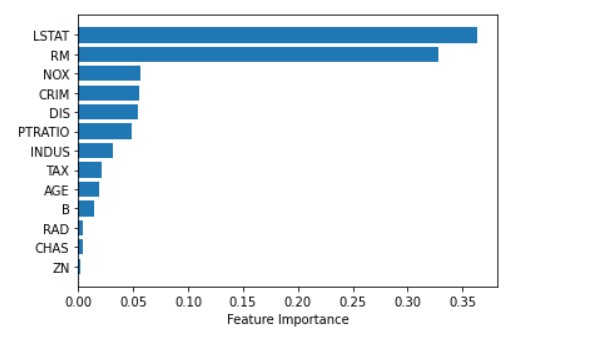
- LSTAT % lower status of the population

- MEDV Median value of owner-occupied homes in $1000's

# Steps performed for generating report

1. Load the data.
2. Split data into training and testing for calculating the importance of columns.
3. Hyper-parameter tuning of model.
4. Training the model so that **it will tell which columns** are more important.
5. Plotting the feature importance calculated by model.

# Final plot after performing above mentioned steps



From above we can see this plot is stating out how much percentage of impact it is making in deciding house price in an area. Higher the **percentage** more is the chance of making impact on pricing and lesser the **percentage** less is the chance of making impact on pricing.

# Few Questions may arise after going through document:

1. **Why we need data of past 10 or 20 years of an organization?**

This is needed because it will help to get detailed information about anomalies which is impacting the business.

1. **What if we ignore those columns have higher percentage of importance?**

If we ignore those columns then we will have loss in business which is equal to the percentage of importance that column contains because those with higher percentage will impact more than those columns which is having less percentage.

# Team Details

Team Name: Deep thinkers

Team Member:

1. Gaurav Kumar Daharia | [gauravdaharia80@gmail.com |](mailto:gauravdaharia80@gmail.com%20%20|) +91 7973071348
2. Dhananjay Kumar | [dhananjaykumar1598@gmail.com](mailto:dhananjaykumar1598@gmail.com) | +91 9131021841

Project Link: https://github.com/GauravDaharia20/Rakathon-22-Anomaly-detection