Cipla Data Scientist

**HIRING CHALLENGE - June 2021**

#### Name: Nilesh Jaiswal

Most people avail home loans to buy their dream house. The bank only lends up 80% of the total amount based on person’s finances.

You work for XYZ bank. Predict the loan amount that can be sanctioned to customers who have applied for a home loan using the features provided in the dataset.

Problem Statement

We are provided with:

* Customer details (gender, age, dependents, etc.)
* Customer finance details(Income, Profession, Current loan expenses, etc.)
* Details of property (Property age, property location, price, etc.)

Feature Description

**Features:**

22 independent feature

1 Target variable

**Test Data:**

20,000 Unique customer

**Training Data:**

30,000 Unique customer

Data Distribution

# Approach

* EDA for ﬁnding key insights in data
* Key drivers are identiﬁed which
* Pre-processed Features
* Handling missing data
  + Build LightGBM Model on 36 encoded features and few other important features

affect bank’s intent to sanctioned loan to the customer



Extracted features from Raw Data

Stacking

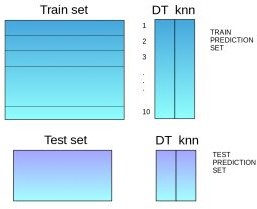
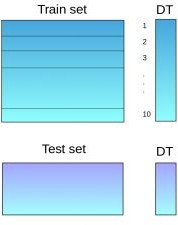
Build 2nd Level of Model

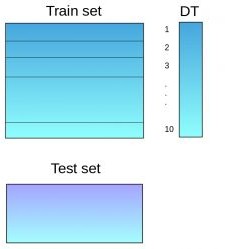
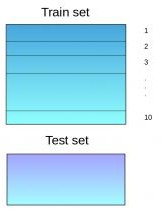
Tuned the Model parameters

Identiﬁed top variables

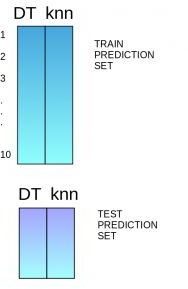
* + - Divided data into 10 K-Fold
    - Created random subsets of feature
    - Build: 1 LightGBM model
    - Hyper Parameter tuning

## Stacking

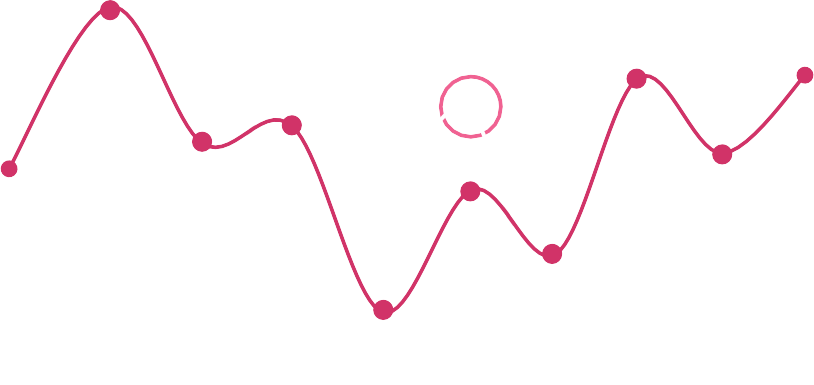




* 1. The train set is split into 10 parts
  2. A model is ﬁtted on 9 parts and predictions are made for the 10th part. This is done for each part of the train set.



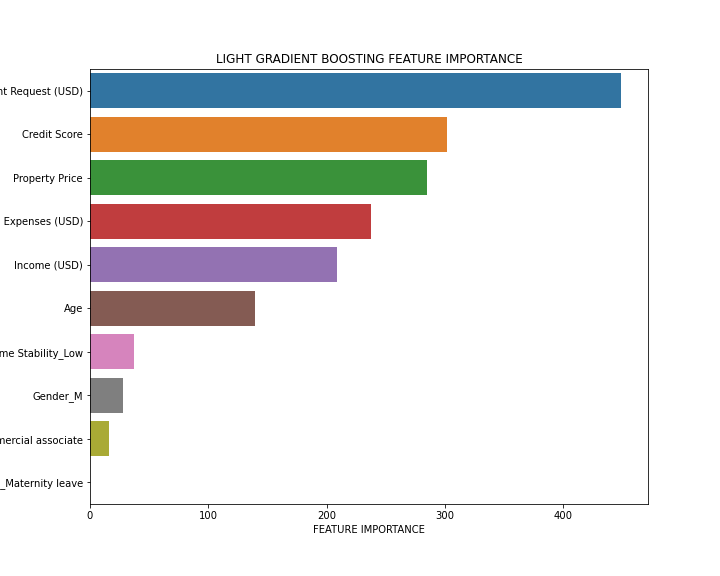
* 1. The model is then ﬁtted on the whole train dataset.
  2. Using this model, predictions are made on the test set
  3. The predictions from the train set are used as features to build a new model



# Result

80.9733 R2 achieved

## Feature Importance



## Thank You