This project is about loan prediction,we have features like app income,credit history,dependents. we are solving a real life case study of Dream Housing Finance. The company deals in all home loans. They have a presence across all urban, semi-urban and rural areas. Customers first apply for a home loan after that company validates the customer's eligibility. The company wants to automate the loan eligibility process (real-time) based on customer detail provided while filling online application form.

By the end of the course, you will have a solid understanding of Classification problem and Various approaches to solve the probem

About Company

Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan.

Problem

Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers. Here they have provided a partial data set.

Note:

1. Evaluation Metric is accuracy i.e. percentage of loan approval you correctly predict.

-we divided columns into categorical columns and numerical columns based on their datatype.

-

### #After that we checked out type of values in categories,then we replaced string values with numerical values for dependents and loan\_status columns

We are validating the train data by using accuracy and predicting the values in test data and giving back to the company.

test data -it’s a new data

Data cleansing:

1)In the dataset we have null values in few columns,so we filled values in those columns.we divided columns into categorical columns and numerical columns.

**For numerical variables: imputation using mean or median**

**For categorical variables: imputation using mode(occuring max times)**

We observed outliers when we visualized data by using histogram.we applied log on those columns.

## Dropping unnecessary columns and splitting data

Dropped the column which have correlation with other column(since loanamount log and loan amount column are highlycorrelated with each other ,we dropped it) and also the column which doesn’t have any effect in predicting the model

We are using dummies to convert categorical features to numerical features.

## We are using accuracy and f1 score as metrics[¶](http://localhost:8888/notebooks/Downloads/DP_Project/Loan_prediction.ipynb#We-will-use-accuracy-as-metrics)

First we built a logistic regression model-we have binary classification in our dataset and logi regr goes well with bin class

Next we are calculating f1 score

Based on the domain knowledge, we can come up with new features that might affect the target variable. We will create the following three new features:

Total Income - As discussed during bivariate analysis we will combine the Applicant Income and Coapplicant Income. If the total income is high, chances of loan approval might also be high.

EMI - EMI is the monthly amount to be paid by the applicant to repay the loan. Idea behind making this variable is that people who have high EMI’s might find it difficult to pay back the loan. We can calculate the EMI by taking the ratio of loan amount with respect to loan amount term.

Balance Income - This is the income left after the EMI has been paid. Idea behind creating this variable is that if this value is high, the chances are high that a person will repay the loan and hence increasing the chances of loan approval.

In [26]:



total income values were left skewed that’s why we are logging the total income so that values will be distributed normally,and there will be no much variance.

Since the newly created features have high correlation with features from which these are created.We remove the

###################old columns. As, in Regression Analysis, we assume features are independent

Since, we have done data Cleaning and Feature Engineering. We can train and test some models.#################

## Planned Models are:

## Logistic Regression

## Decision Tree

## Random Forest

## KNN

Data analysis and finding pattern is missing