**K2 Data Science - Loan Prediction Dataset**

**Goals:**

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.

Company wants to automate the loan eligibility process 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.

**Approach**

Determine the customer eligible for loan based on

* Education : Tells if the applicate is a graduate/non-graduate
* Income level: The applicant’s income is measured in dollars. There is another measure of income I.e the co applicant’s income
* Credit history : The value can either be 1 or 0. It is 1 if the credit history meets guidelines else 0

**Cleaning the data**

For determining if a customer is eligible for loan or not , I decided to consider “Applicant Income” and “Co-applicant Income” as a factor rather than just applicant income since some applicants have lower income but strong support Co-applicants. So it might be a good idea to combine both incomes as total income . This is represented by the column “TotalIncome”. It can be observed that Applicant Income and Co Applicant columns do not have any missing fields. However based on the box - plots it can be observed that both the attributes have outliers.



Fig 1: Outliers in Total Income



Fig 2: Applicant Income grouped by Education

These can be due to income disparity in society. Since we are also considering education as a key factor, I segregated total income based on education. Education attribute also does not have any missing values. The plots do not show any substantial difference between the mean income of graduate and non-graduates. But there are a higher number of graduates with very high incomes, which are appearing to be the outliers. Therefore I did a log transformation to nullify their effect.

Another key factor to determine the eligibility of a customs for loan is credit history. It can take a value of 0 or 1. A customer who has fulfilled the requirements of credit history is given 1 else 0. We can see that 84% of the customers have credit history. However this attribute has 50 missing values. 84% of customers have a valid credit history, so imputed the missing values with the mean.

Final Recommendations:

Fig 3: Probability of being approved loan based on credit history



Fig 4: Probability of being approved loan based on Education

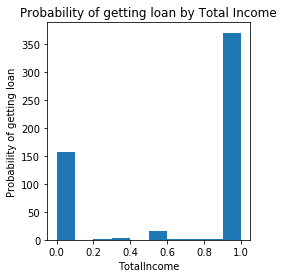


Fig 5: Probability of being approved loan based on Total Income

Fig 3,4 and 5 show that credit history and TotalIncome are important factors for a loan approval, however education does not.

Further Analysis

We can use other imputation techniques to impute missing values in the attributes such as gender and loan amount. Prediction models can be used to predict what the missing values could be in each attribute. This data can be used to predict if a customer will/will not be approved loan.