**AAS Project Report**

**(Team 2)**

**Problem statement:**

This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.  The company can utilise this knowledge for its portfolio and risk assessment.

To Achieve the model following methods were followed:

1. **Understanding the data :**

We just import files and by using different methods , we got to know the no.of rows and columns each file had……We also observed if there a re some null values are present or not.

1. **Cleaning The Data:**

The data is cleaned where null values and columns with single values were handled. The columns with more than 50% of null values were dropped, then still we had more no of rows….so we dropped columns having 15% null values. Many variables are having dataframes with single value which is leading to imbalance so we will drop the following column EXT\_SOURCE\_3,EXT\_SOURCE\_2….Some variables were affecting the model accuracy rate due to which these variables were removed from the data set. The variables are namely columns having names starting with word FLAG……..

We found some values were negative……..so we have converted them in positive values….like employed days and Registration days etc……..

1. **EDA:**

Exploratory Data Analysis was done to check categorical variables . We can find that some numerical variables consisted of very high values as compared to their respective means. That's why we have created charts using boxplot to understand the patterns. We have observed that the outliers are very high and we need to treat it.

1. **Model Building:**

Now all the data is numeric and our response variable (target) is BINARY, hence we can use Binary logistic regression model. As it was difficult to understand the correlation between the variables so we find the top 10 variables, from the train dataset this will help in training the model……..

**CONCLUSION :**

1. Banks should focus more on Age group 30-50 as they have highest no. of payment issues…….
2. Banks should focus more on income type ‘Working’ as they are having most number of payment issue.
3. Banks should focus more on MEDIUM income group people and as they again have payment issue
4. Customers owning House/apartment are most likely to make payments on time compared to those living in CO-OP apartment.
5. Get as much as clients from housing type ‘With parents’ as they are having least number of unsuccessful payments.