# STRUCTURED PYRAMID ANALYSIS PLAN
## Classification & Prediction of Loan Defaulters
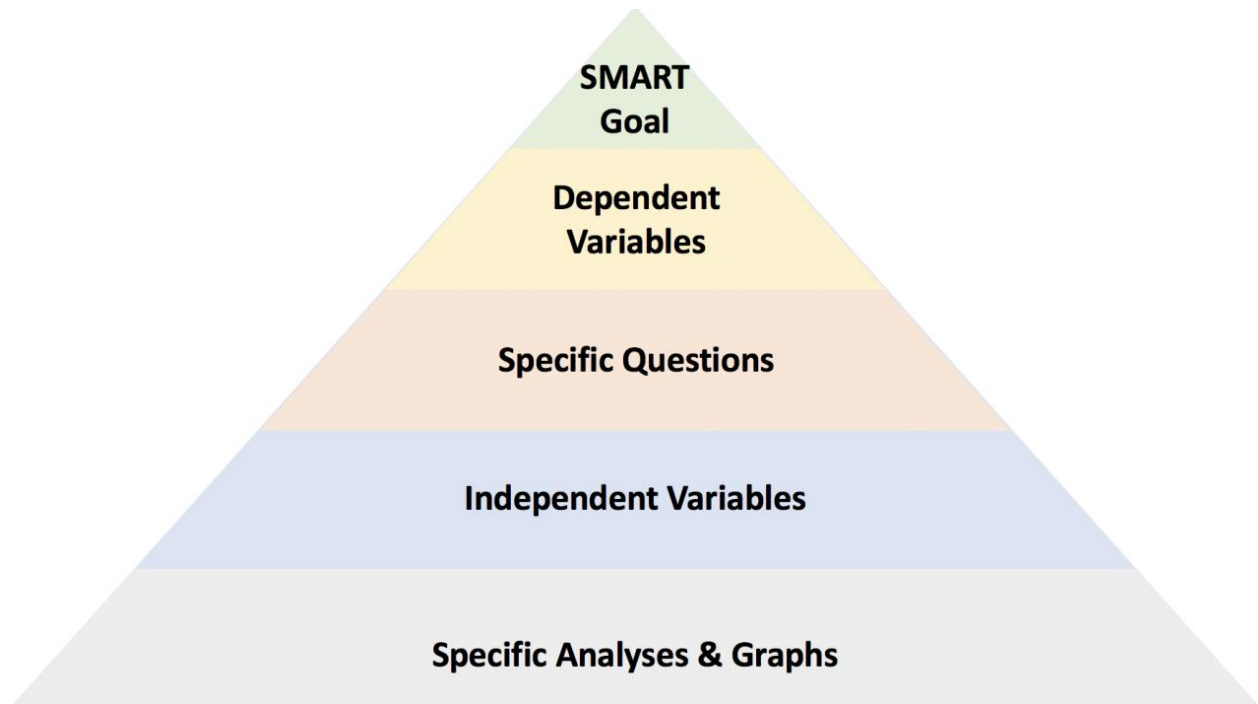
**Team 8:**

| | |
|---|---|
| Adarsh Lalchandani | GitHub - https://github.com/AdarshRL2109 |
| Karthik Grandhi | GitHub - https://github.com/karthii24 |
| Richard Britto | GitHub - https://github.com/richardbritto97 |

**DESCRIPTION:**

- **SMART Goal**: A goal that is Specific, Measurable, Attainable, Realistic, and Time-Bound.
- **Dependent Variables**: variables that directly pertain to the accomplishment of the S.M.A.R.T. goal.
- **Specific Questions to Investigate**: ideas and intuitions about what parameters may impact the independent variables.
- **Independent Variables**: variables that may help explain the possible connections in the questions, above, to the dependent variables.
- **Specific Analyses**: Statistical analyses and plots of independent versus dependent variables.

## GOAL:

To classify & predict whether the loan applicant is going to default or not (Boolean) based on the information we receive by the applicant at the time of application.

## DEPENDENT VARIABLES:

- default_flag taking values of 1 or 0 to determine if the borrower is going to default or not respectively

- loan_status taking values of Charged Off and Fully Paid

## SPECIFIC QUESTIONS:

What is the effect of loan amount to the percentage of loan defaults?

What effect does the borrowers loan enquiry in the last 6 months have on percentage of defaults?

How does the number of active credit lines of the borrower affect the percentage of defaults?

How does the DTI affect the number of loan defaults?

What is the effect of delinquency by the borrower in the last 12 months affect the loan default percentage?

## INDEPENDENT VARIABLES:

int_rate, loan_amt, Annual_inc, purpose, open_acc, No_of_Ac, emp_length dti, revol_util, total_pymt, total_rec_late_fee, inq_last_6mths, delinq_2yrs

## SPECIFIC ANALYSIS & GRAPHS:

- Univariate Analysis of specific variables vs number of defaults
- Bivariate Analysis on continuous variables vs default percentages
- Bivariate Analysis on categorical variables vs default percentages
- Multivariate Analysis to create the co-relation plot and find significant of each feature

## KEY PROCESSING INDEX:

Key Processing Index (KPI) can be achievable by determining what is your objectives, how you plan on achieving them, and who can act on this information. The following are the KPI's which will help us to evaluate and determine the final analysis.

| Key Processing Indexes |
| --- |
| Default Percentage |
| Interest Rate |
| Annual Income |
| Loan Amount |
| Loan inquiry in lat 6 months |
| Delinquency in last 12 months |
| Purpose for loan |
| Debt to Income ratio |
| Number of active credit lines |

KPIs need to be defined according to critical or core business objectives.
Following steps helps to define a KPI:

1. What is your desired outcome?
   Classify and predict loan defaulters based on information given at the time of request.

2. Why does this outcome matter?
   There is a particular business interest and risk prevention by the outcome of our analysis.

3. How are you going to measure progress?
   Progress will be measured by the accuracy and predictive power of the different models which we will be training and testing.

4. How can you influence the outcome?
   To find out the most affecting parameter that drives default rates, we shall achieve model parsimony so as to have specific influencing variables.

5. Who is responsible for the business outcome?
   The lenders are responsible for the business outcome.

6. How are borrowers benefiting from this?
   Borrowers will get a value assigned to their profile.

7. How often will you review progress towards the outcome?
   Will be reviewing on a quarterly basis to see the effects.

## HYPOTHESIS:

There exists some co-relation between the information we get from a user and the probability of that person to be defaulting with an outcome being either a defaulter or not a defaulter. The strength of co-relation will vary for each distinct variable and we will determine the effect of each variable over the outcome of this prediction.

## DATA DICTIONARY:

| Variable | Description |
|---|---|
| member_id | Unique identifier |
| loan_status | Current status of the loan |
| int_rate | Interest Rate on the loan |
| default_flag | A Boolean value where 0 means no default & 1 means default |
| inq_last_6mths | The number of inquiries in past 6 months (excluding auto and mortgage inquiries) |
| Annual_inc | The self-reported annual income provided by the borrower during registration. |
| Purpose | A category provided by the borrower for the loan request. |
| open_acc | The number of open credit lines in the borrower's credit file. |
| No_of_Ac | Categorical value for open_acc variable |
| emp_length | The job title supplied by the Borrower when applying for the loan. |
| loan_amnt | The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. |
| dti | A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income. |
| revol_util | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |
| total_pymnt | Payments received to date for total amount funded |
| total_rec_late_fee | Late fees received to date |
| delinq_2yrs | The past-due amount owed for the accounts on which the borrower is now delinquent. |