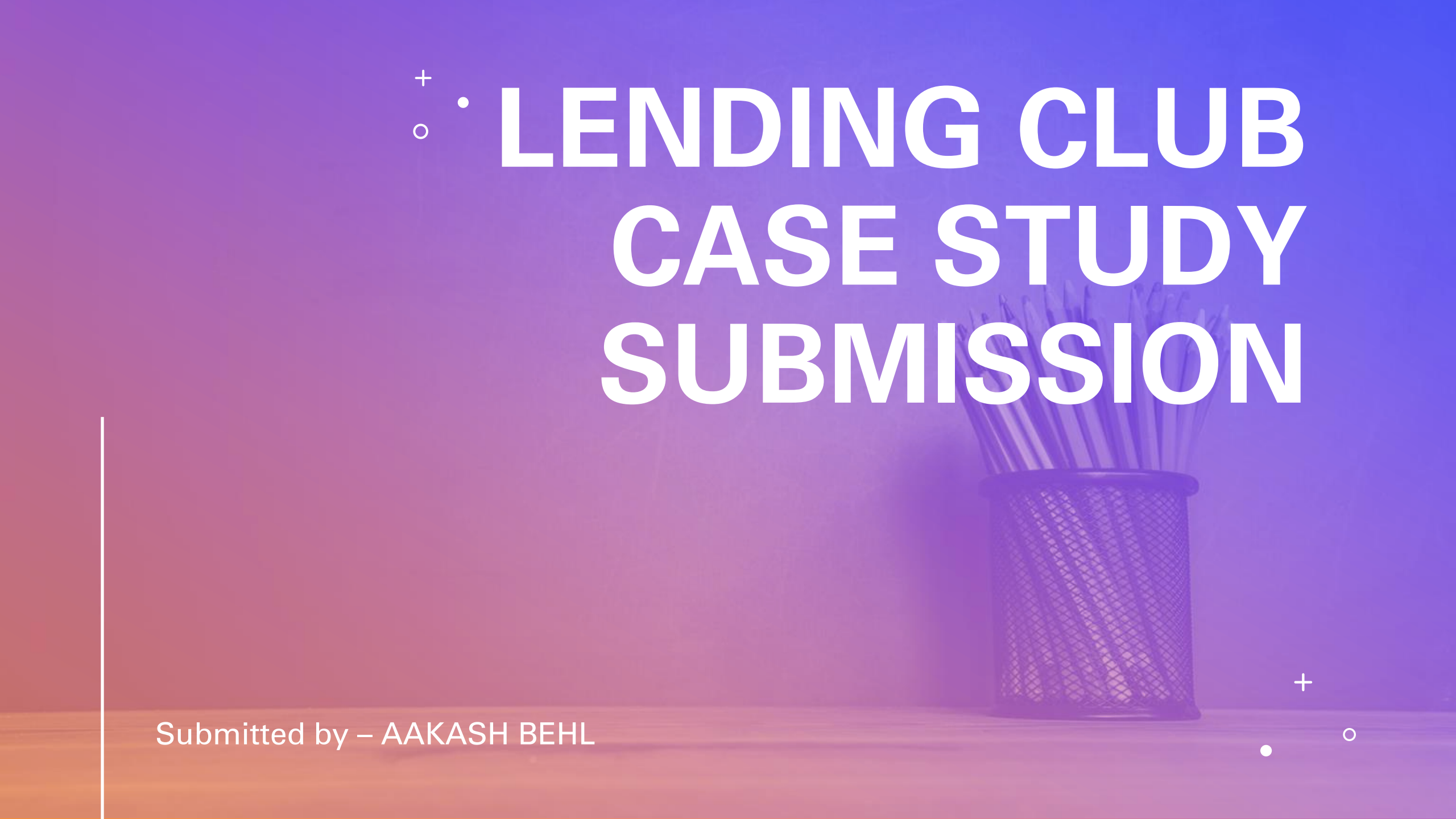
LENDING CLUB CASE STUDY SUBMISSION



Submitted by – AAKASH BEHL

Problems solving steps

1. Understanding data
2. Data cleaning
3. Data preparation
4. Exploratory data analysis – Performing univariate, segmented univariate and bivariate analysis.

Data Cleaning steps

1. Many columns in the dataset which are having lot of missing values, so we dropped the columns with more than 60% missing values, after dropping the columns the number of columns dropped from 111 to 54

2. Now removing the columns that are the customer behavior variables are not available at the time of loan application, and thus they cannot be used as predictors for credit approval

3. We removed the columns with only zero values, policy codes, single values or text columns which cannot be used for the analysis.

4. We also removed rows which are for current loan status, because for the guy who is paying the loan there is no way to say if it is a good or bad loan, so we will not use that current data in the analysis. Considering only fully paid and charged off loans

5. We can see that our data is imbalanced in favour of Fully paid, but that was expected we don't have to do much about it since we are not doing predictive modelling

6. I think it is good idea to remove people which are having income of more than 200,000 USD since they are few and can be considered outliers

Exploratory data analysis

Univariate Analysis

1. The debt-to-income ratio varies from 0 to 30% having mean of 13%
2. Loan amount varies from 0 to 35,000 and mean of 11,047 USD
3. Most people prefer to take loan for 36 months
4. Most of the loans are Fully Paid and We have a class imbalance here.
5. Most of the loans are of around 11% and they range between 6% and 23%
6. Most loans are of A, B and C grade
7. Only 1.4% of the people have income more than 200,000 USD
8. The mean income is around 64k and the median is 57K, after removing outliers of above 200k USD which are only 1.4%
9. Most of the loans are debt_consolidation
10. Many of the loans are taken by people with more than 10 years of experience
11. Most people with loans live in rented or mortgaged house
12. Many people whose status was not verified were given loans
13. Many loan are taken for debt consolidation and lot of them are becoming bad loans
14. Most people who took loan are in rented home or mortgaged home

Exploratory data analysis

Bivariate Analysis

1. We are making charged off proportion as the key indicator as that explains the proportion of defaults, which is most important for us

2. We can see that Grade A has the least defaults in all grades

3. We can see that as income goes higher the percentage of charged off loan decreases

4. Small business have the highest rate of default. Car, credit card and Wedding loans have the lowest rate of default

5. Debt to income of over 19% has the highest rate of default and as people are more in debt as a ratio of their income, the chances of default increases

6. Since small business owners had the maximum default rate, so looking down further into that data. We can see that grade D, E, F and G of small business owners have 30%- 45% default rate. Small business owners with D,E,F,G grade have the highest default rate of all groups