# WRITE UP

## OBJECTIVE

This study aims to identify the risky loan applicants for our bank using EDA and machine learning algorithms so that such loans can be reduced thereby cutting down the amount of credit loss. In other words, we want to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

## ANALYSIS

### RAW DATA

There are 39717 rows and 110 columns/variables in the data set covering various details about the loan applications & repayment.
The Variables can be broadly classified into 3 categories:

1) Applicant Details: Applicant's information such as "Annual Income", "Employment Length', "Address", "Home Owned Status" etc.

2) Loan Details: Loan application information such as "Loan Amount", "Interest Rate", "Grade", "Term" etc.

3) Customer Behavior Details: Variables describing customer behavior on repayment after the loan is sanctioned such as "Pub-rec", "Last Payment Date", "Next Payment Date", "Recoveries" etc.

   As per the Objective of the study we need to focus on variables that help us identify loan can be defaulted before sanction thus we would be focusing more on "Applicant Details" & "Loan Details" Variables

### EDA

After all null values are treated and insignificant columns are dropped, there were 35367 rows and 34 columns left.

### UNIVARIATE ANALYSIS

We performed Univariate Analysis on major columns to check which variables help to identify Charged off loans. Below are the Major Findings:

1) The percentage of charged off loans increases as the **loan amount** increases.

2) The number of 3-year loans (**loan term**) is very high compared to 5 years. 5 years loan get more prone to getting charged-off.

3) The **interest rate** for charged off seems higher than for fully paid. And as the interest rate increases the percentage charged off also increases.

4) Loans **grades** A, B have the lowest % of charged-off cases. Other grades like E, F,G are facing large charge offs.

5) 21.6 % loans taken are from people with 10 + years of **emp_length**.24.9 % of out these are charged off.

6) Charged off cases are higher for **Homeownership** type of Rent & Mortgage.

7) Majority (99 %) of the **annual income** lies below 2.5 lacs, Annual Income range is 4k to 39 lacs, since 95% of annual income is below 1.4 lacs, we removed 192 outliers from top income

8) Amongst charged off records, majority of the loans are from 2011 and 2010.

9) Amongst all Charged off records most loans are for same **purpose** debt consolidation. Small business and renewable energy have the greatest number of charged-off loans percentages amongst themselves.

10) Average value of **dti** is slightly higher for charged off, as the dti increases above 15, the % of defaulter also increases.

11) Applicants with **delinq_2yrs** value greater than 0 have slightly high percentage of defaults on average.

12) Charged off loans are affected by **revol_utils** as the revol % increases risk of loan being charged off increases with highest risk lying when revol_util is under the range of 60 -90 %

**BIVARIATE ANALYSIS**

To further narrow down the key indicators of loan default we performed bivariate & segmented univariate analysis. In this analysis we checked whether the univariate analysis holds true when variable is checked with others.

Below are the major findings:

1) **Annual Income** for charged off cases is lower than fully paid, whereas the annual income varies in a different pattern against different variables. Annual income is highest for mortgage home ownership applicants & for applicants having emp_length greater than 20 yrs. Univariate analysis showed that Highest defaulters are also from mortgage & over 10years employed.
2) **Dti** however stands true showing that people having high DTI have lower annual income, and as we know dti is higher for charged off cases, dti also holds true as it

increases with emp length and average dti is greater for charged off cases. Thus, it seems to be a strong indicator.

3) Dti for **home ownership** as Mortgage & Rent is on the higher side, indicating that these can be strong indicators for defaulters. Home Ownership vs delinq_2yrs does not relate heavily with defaults as its high for others case in both kind of loan status.
4) Loan Amount, Term, Grade & Interest rate all are directly proportional to each other.

   Higher Loan amount tends to have higher term

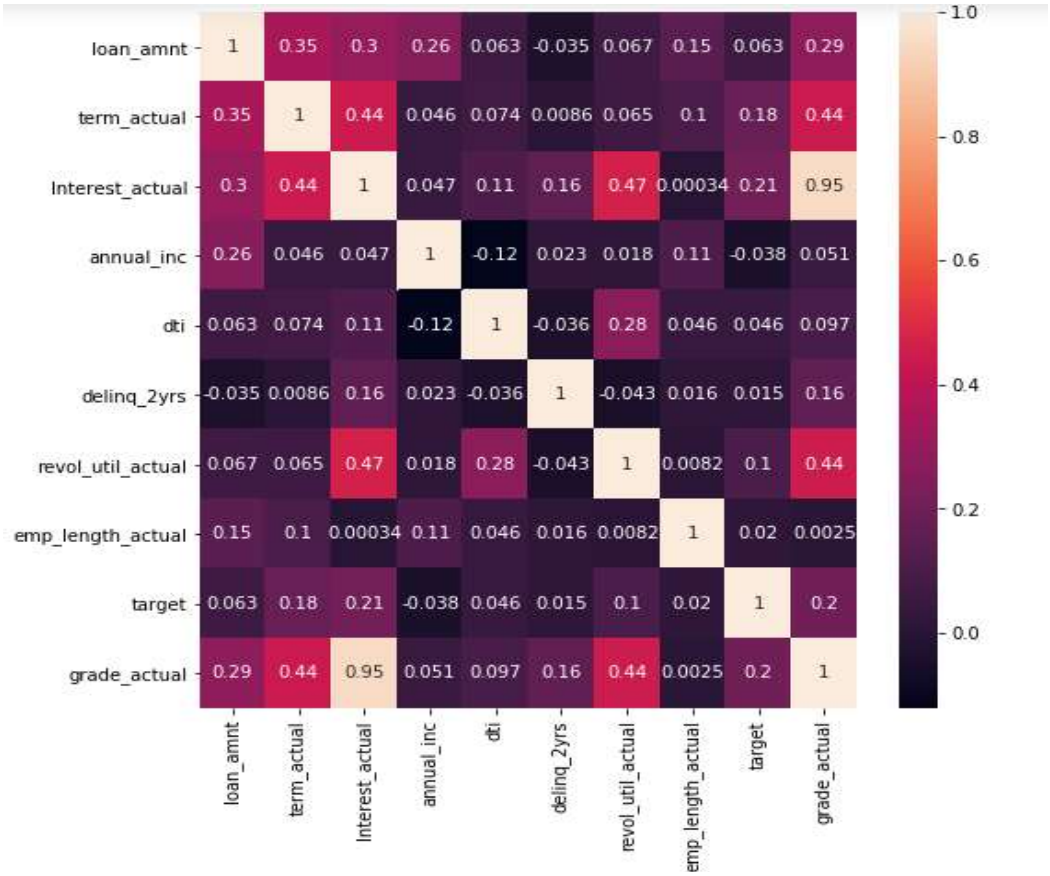
   Higher loan amount & higher term tends to have higher grade

   Higher grade will imply higher interest rate

   **Loans with higher amount, term, grade & interest rate tends to get charged off more often**

5) All kinds of **interest rates** are spread over various range of Annual Income, most dense being annual income lower than 10000 & interest rate 10~15 %
6) **Revol_util** is one of the stronger indicators of laon default, as revol_util increases risk of default increases.

On drawing the **correlation matrix** also, the variables which are the most correlated with the target ("Charged Off loans") are:



Grade, Term , interest rate, revol_util & Dti.

## MACHINE LEARNING ALGORITHMS

We applied three classification models:

1) Logistic Regression
2) Random Forest
3) Decision Tree

Below is the comparison matrix for models:

| | Logistic Regression | Random Forest | Decision Tree |
|---|---|---|---|
| **Accuracy** | 0.862501 | 0.788616 | 0.505796 |
| **Misclassification Rate** | 0.137499 | 0.211384 | 0.160871 |
| **True Positive Rate** | 0.999672 | 0.885515 | 0.845082 |
| **False Positive Rate** | 0.999314 | 0.820178 | 0.782341 |
| **Specificity** | 0.000686 | 0.179822 | 0.217659 |
| **Precision** | 0.862732 | 0.871519 | 0.871219 |
| **Prevalence** | 0.999623 | 0.876543 | 0.557629 |

The utilized logistic regression model has an accuracy and precision of 0.862 when evaluated on the validation dataset. The model has the lowest misclassification rate of 0.13.These results suggest that the validation dataset may be used to forecast new data using the logistic regression equation accurately.

> **Recommendation:** We would suggest using the Logistic Regression model for classification since it has the highest accuracy, lowest misclassification rate, and highest True Positive Rate. Additionally, it also has the lowest False Positive Rate, which indicates that it is better at identifying true negatives. However, it is important to consider other factors such as the problem domain, the dataset, and the available computational resources.

**CONCLUSION**

After the exploratory analysis and applying the machine learning algorithms, we found out that five key variables that drive the Charged Off loans are “the Term, Grade, Dti, Revol_Util and Interest rate” for the loan amount that can be defaulted.