GRAMENER Case Study

Submitted by:

Group Name: Team Insight

1. Kaushal Kashyap
2. Manjiri Paranjape
3. Snehil Gupta
4. Ravinder Gill

Business Understanding and Introduction

- A largest Consumer finance company which specializes in lending various types of loans to urban customers online.

- Company facilitates personal loans, business loans, and financing of medical procedures.

- Borrowers can easily access lower interest rate loans through a fast online interface.

- Major products are credit card loans, debt consolidation loans, house loans, cars loans etc.

Decision Variable

When a person applies for a loan, there are two types of decisions that could be taken by the company:

- **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:

    1. **Fully paid**: Applicant has fully paid the loan.

    2. **Current**: Applicant is in the process of paying the instalments.

    3. **Charged-off**: Applicant has defaulted on the loan.

- **Loan rejected:** The company had rejected the loan application.

Problem and Goal Analysis

Problem:

There are two types of risks associated with bank’s decision.

1. **Loss of business to the Company** → Not approving a loan when applicant is likely to repay the loan
2. **Financial loss for the Company** → Approving a loan when applicant is not likely to repay the loan

Goals:

1. Risk analysis to find out how consumer attributes and loan attributes influence the tendency of default.
2. To identify **driving factors** which are strong indicators of default
3. Potentially use the insights in **approval/ rejection** decision making.
4. The company can utilize this knowledge for its portfolio and risk assessment.

Data Understanding

Data Source: Data is provided in ***.csv file*** format in the form of two files:

1. Loan Data Set
2. Data dictionary

Data Understanding: The data contains the information about past loan applicants and whether they ‘defaulted’ or not.

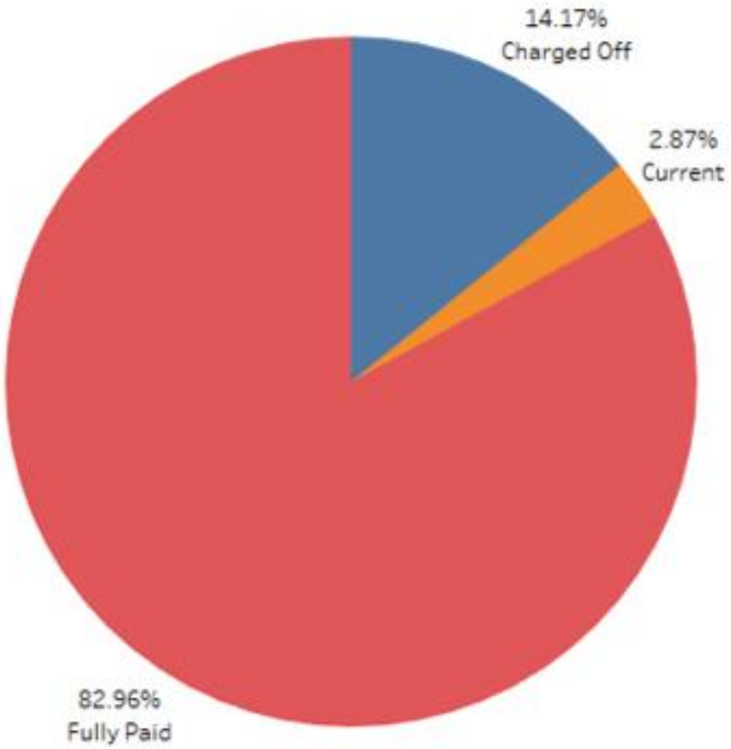
Data contains the information about past loan applications from 2007 – 2011.

Methodology

Analysis Type	Operations Performed	Methodology/ Tools used
Data Cleaning	• Data Preparation: This includes Data Cleansing and Formatting. • Remove columns that are not required.	▪ CRISP DM – Framework ▪ Univariate analysis, ▪ Segmented Univariate analysis. ▪ Univariate analysis, ▪ Segmented Univariate analysis. ▪ Derived Matrices. Tools and Packages Used: ▪ Lubridate Package, ▪ Tidyrverse Package, ▪ Stringr Package, ▪ Dplyr Package, ▪ ggplot Package ▪ Based on plots derived from the above analysis. ▪ R- studio ▪ Tableau
Univariate Analysis	• Visually identify effect of Verification status, Issue date, Loan status on the number of loans applied.	
Segmented Univariate Analysis	• Filtered data with respect to Charged off customers and • Visualize the effect the various parameter like home owned, Employment period on loan applications. • State wise analysis of Customers with charged off loan status.	
Derived Metrics Analysis	• Visualizing the effect of Principal loss calculated using Income slot and Loan status	

Data distribution w.r.t. Loan status

Plot No :- 1



% of Total Count of Number of Records and Loan Status. Color shows details about Loan Status. Size shows sum of Number of Records. The marks are labeled by % of Total Count of Number of Records and Loan Status.

Plotting and visualizing the whole data:

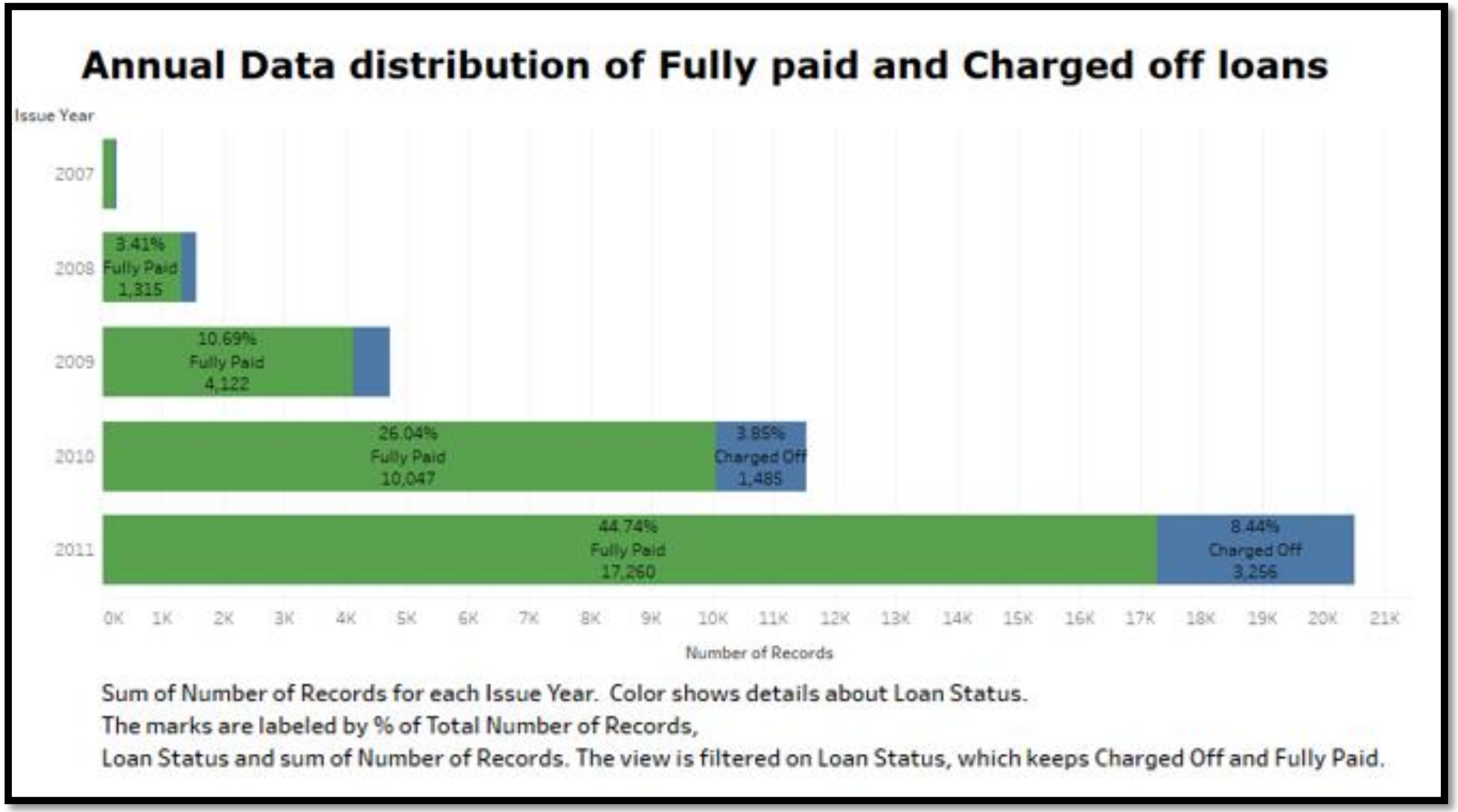
- Since 2007 till 2011, 82.9% of total number of loans have been paid fully by the customers
- 2.87% of the customers are currently paying the instalments.
- 14.17% of the customers have ‘Charged off’ status that is considered as defaulted.

Analysis:

- 14.17% ‘Charged off’ customers are the cause of business loss and there by it is expected to reduce this percentage.

Univariate Analysis

Plot No :- 2



Annual Data distribution of Fully paid and Charged off loans

Sum of Number of Records for each Issue Year. Color shows details about Loan Status.
The marks are labeled by % of Total Number of Records,
Loan Status and sum of Number of Records. The view is filtered on Loan Status, which keeps Charged Off and Fully Paid.

Observation:

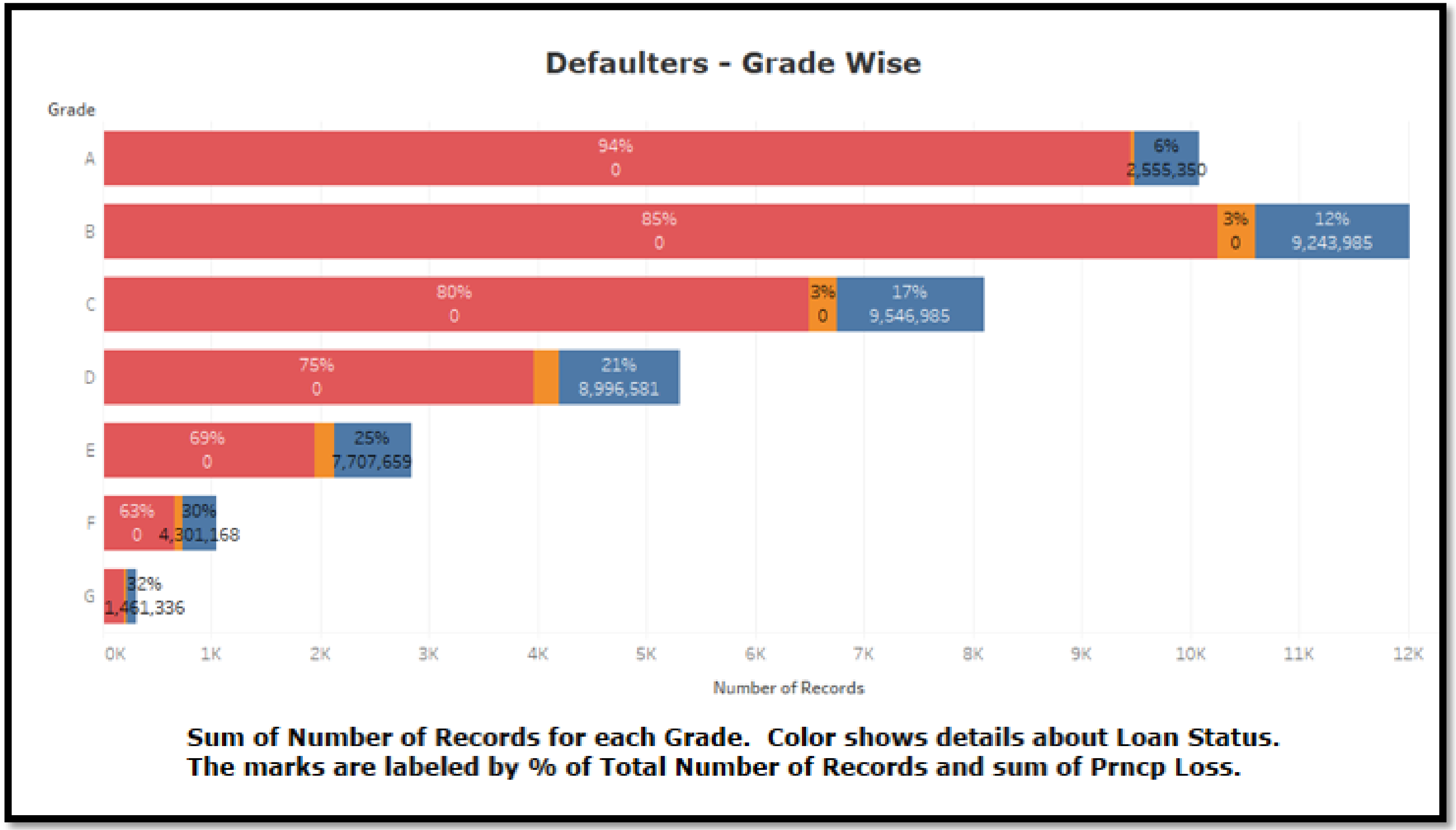
- In comparison to year 2010, in year 2011 total number of fully paid applicants increased by 71%.
- As a result, 'Charged off' customer increased by 120%.

Analysis:

- Same trend is expected in coming years.
- With increase in business, there will be increase in the percentage of charged off customers as well.

Univariate Analysis

Plot No :- 4

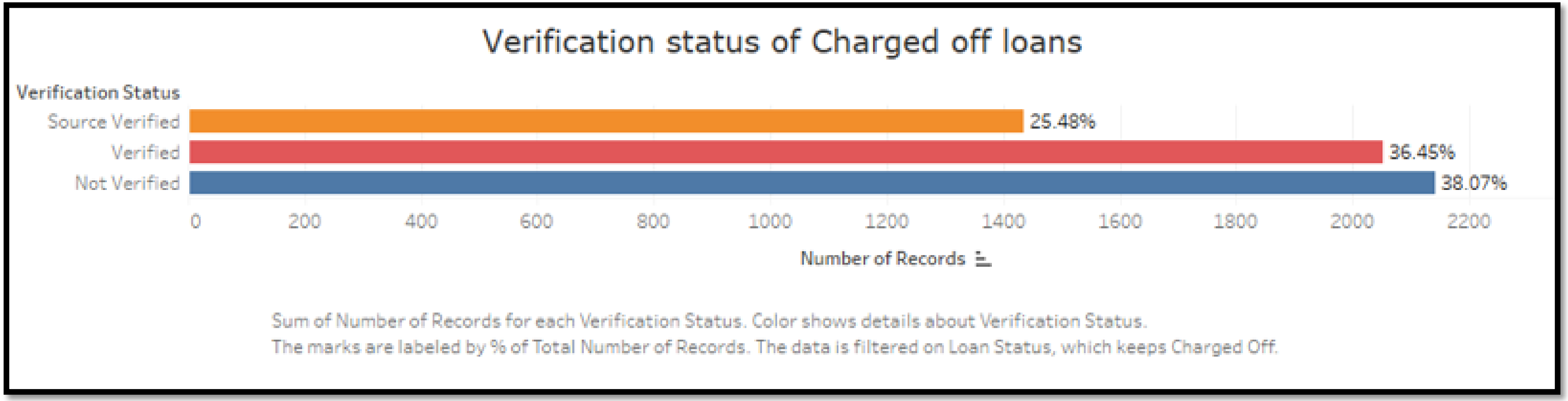


Sum of Number of Records for each Grade. Color shows details about Loan Status. The marks are labeled by % of Total Number of Records and sum of Prncp Loss.

Analysis:

- B, C, D: These loan grades seem popular as well contributing to max principal loss.
- E - This is a tricky grade where loans approved are moderate in count but max %defaulters in this category contributing to substantial principal loss.
- A - Popular but less risky (A4, A5 least risky).
- F, G - Safe but not many apply or get approved for this loan.

Segmented Univariate Analysis



Sum of Number of Records for each Verification Status. Color shows details about Verification Status.
The marks are labeled by % of Total Number of Records. The data is filtered on Loan Status, which keeps Charged Off.

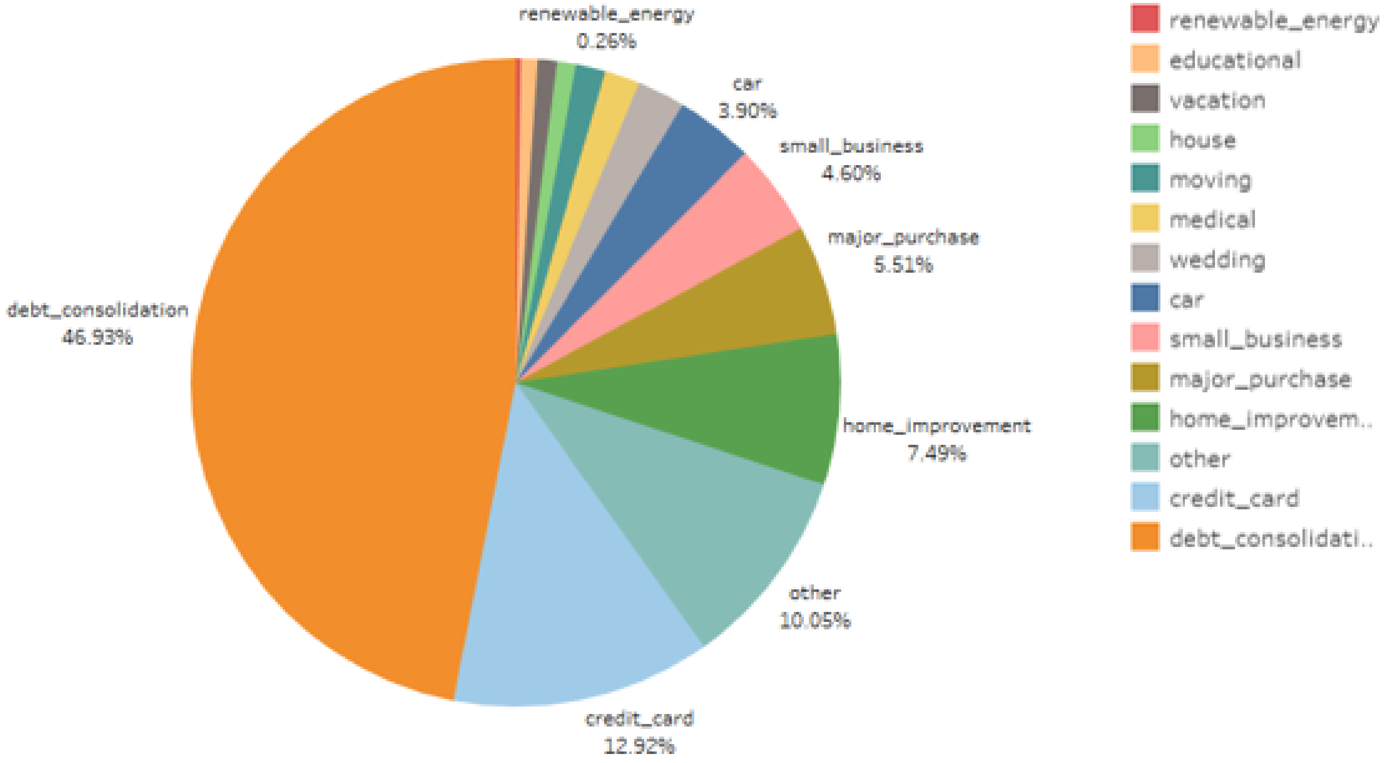
Analysis:

- Among all the 'Charged off' loan applicants, 38.07% has verification status as 'Not verified'.
- The above plot implies that verification process must be streamlined in order to increase the customer's credibility.
- Thereby, reducing the number of 'Charged off' loan applicants.

Segmented Univariate Analysis

Plot No :- 6

Product wise distribution of Charged off customers



Purpose and % of Total Number of Records. Color shows details about Purpose.
Size shows sum of Number of Records. The marks are labeled by Purpose and % of Total Number of Records.

Product opted by the most of

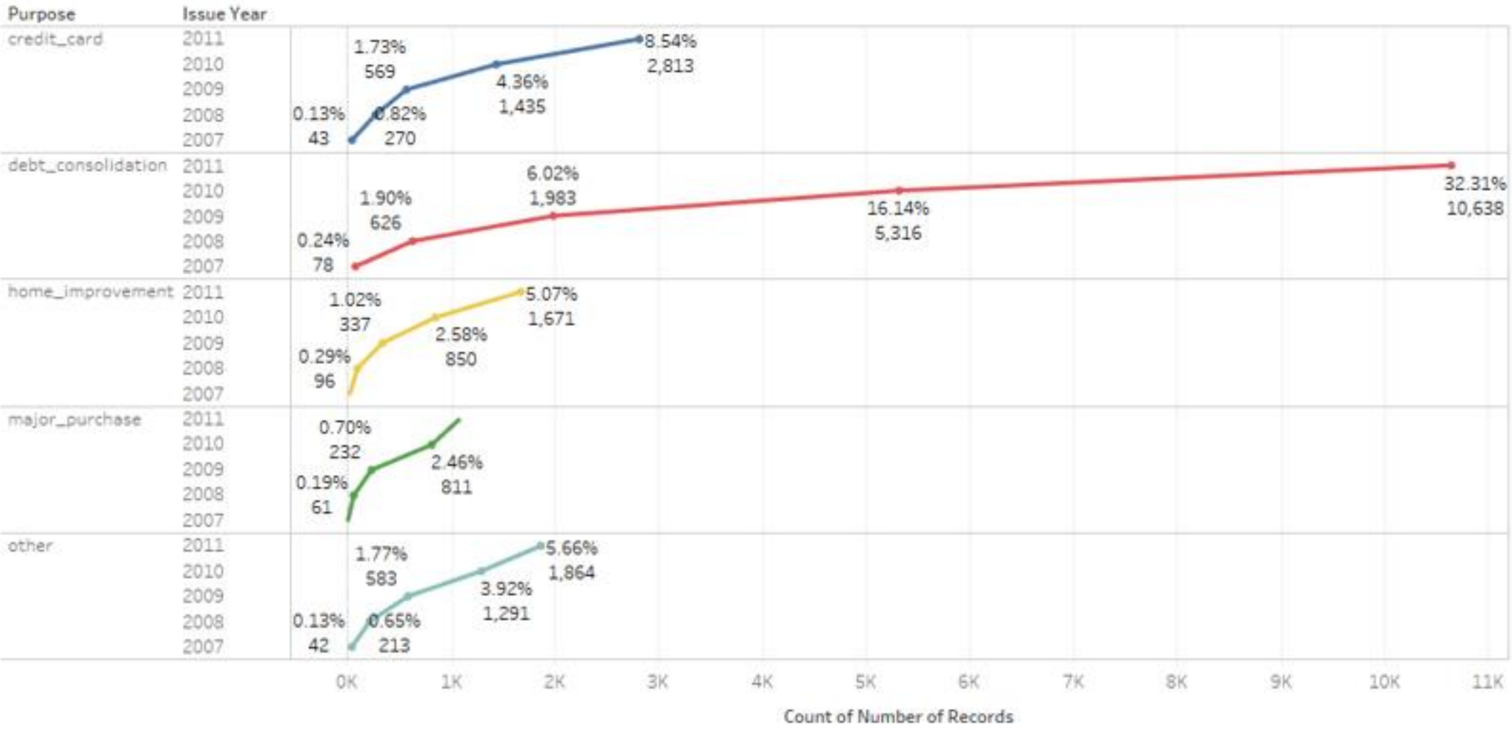
'Charged off' applicants:

- Debt Consolidation tops the list with 46.93%
- Credit card at second place with 12.92%
- Others at third place with 10.05%
- Home improvement is at fourth place with 7.49%
- Then followed by major purchase, small business, car, etc.

Segmented Univariate Analysis

Plot No :- 10

Top 5 product Analysis over the past 4 years

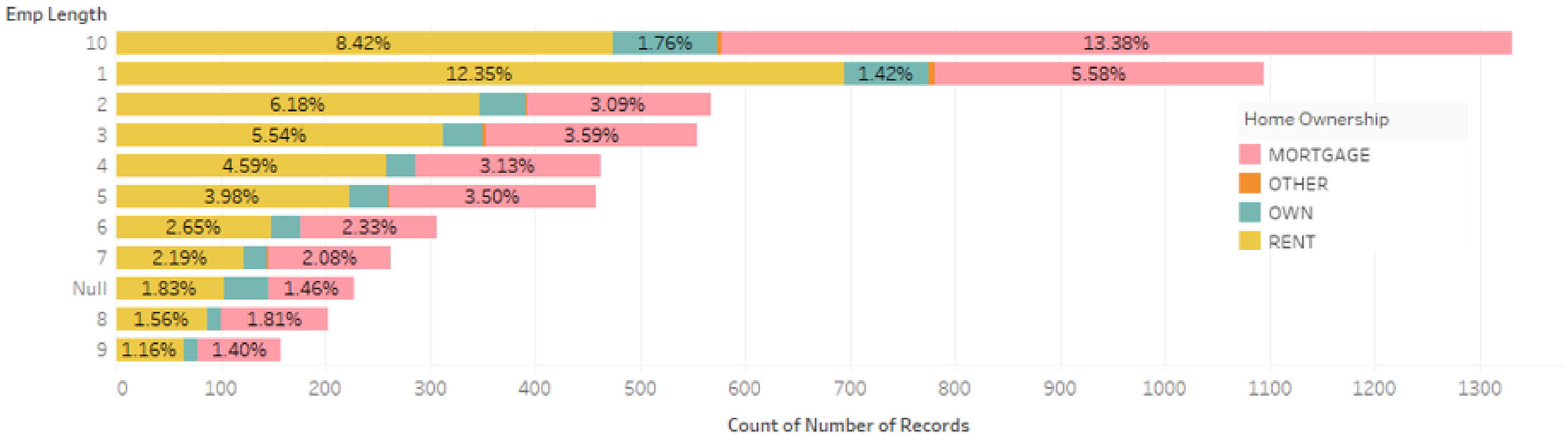


The trend of count of Number of Records for Issue Year broken down by Purpose. Color shows details about Purpose. The marks are labeled by % of Total Number of Records and sum of Number of Records. The view is filtered on Purpose, which keeps credit_card, debt_consolidation, home_improvement, major_purchase and other.

- This plot depicts the changing demand of top five products over the years.
- Debt consolidation shows an exponential increase in demand among the 'Charged off' customers.
- Followed by Credit card product.
- This trend can be seen as a predictor for the definite growth in demand of these products in future.

Segmented Univariate Analysis

Plot No :- 7



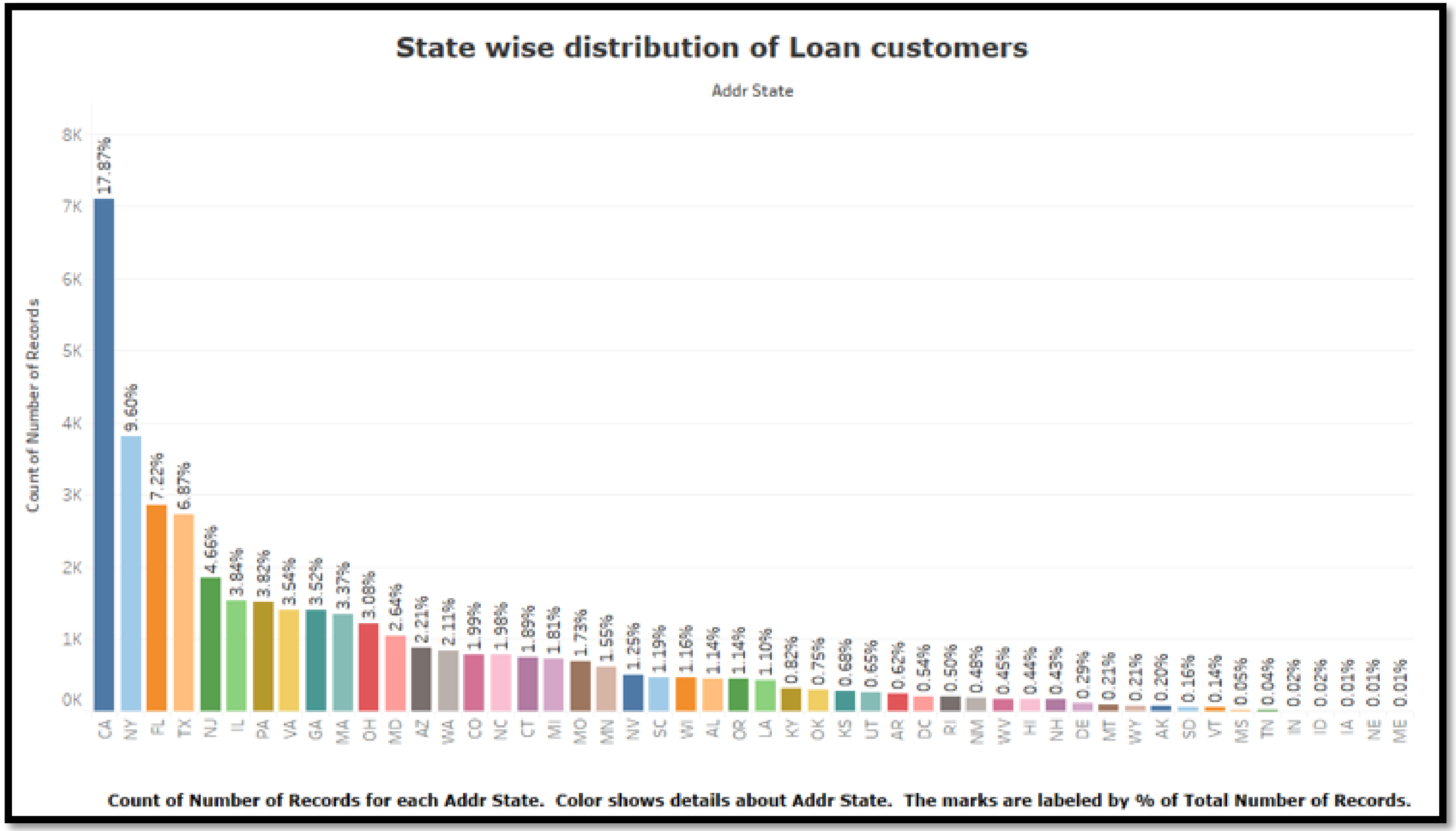
Count of Number of Records for each Emp Length. Color shows details about Home Ownership.
The marks are labeled by % of Total Number of Records. The data is filtered on Loan Status, which keeps Charged Off.

<u>Analysis</u>:

- Applicants with employment period over 10 years and 1 year or less, account for major contribution to the 'Charged off' customers.
- It is also observed that these applicants have either mortgaged their homes or living in rented houses.

Segmented Univariate Analysis

Plot No :- 8



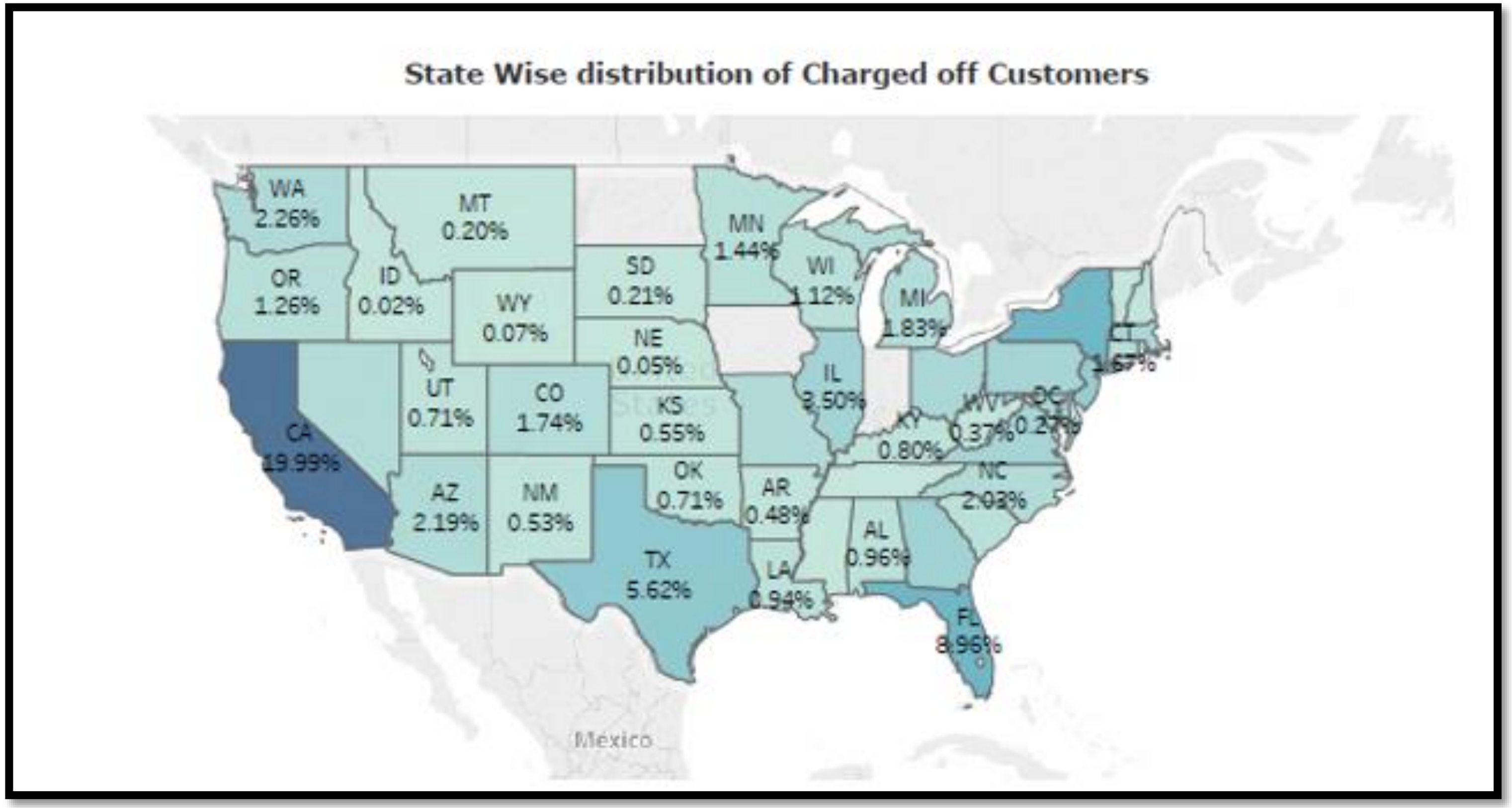
State wise distribution of Loan customers

Count of Number of Records for each Addr State. Color shows details about Addr State. The marks are labeled by % of Total Number of Records.

- Among the all types of loan applications from year 2007 – 2011, California accounts for the most number of loans applicants with almost 18%.
- Followed by New York with 9.60%.
- Then Florida with 6.87%.
- New Jersey at fourth place with 4.66% of total loan applied.

Segmented Univariate Analysis

Plot No :- 9

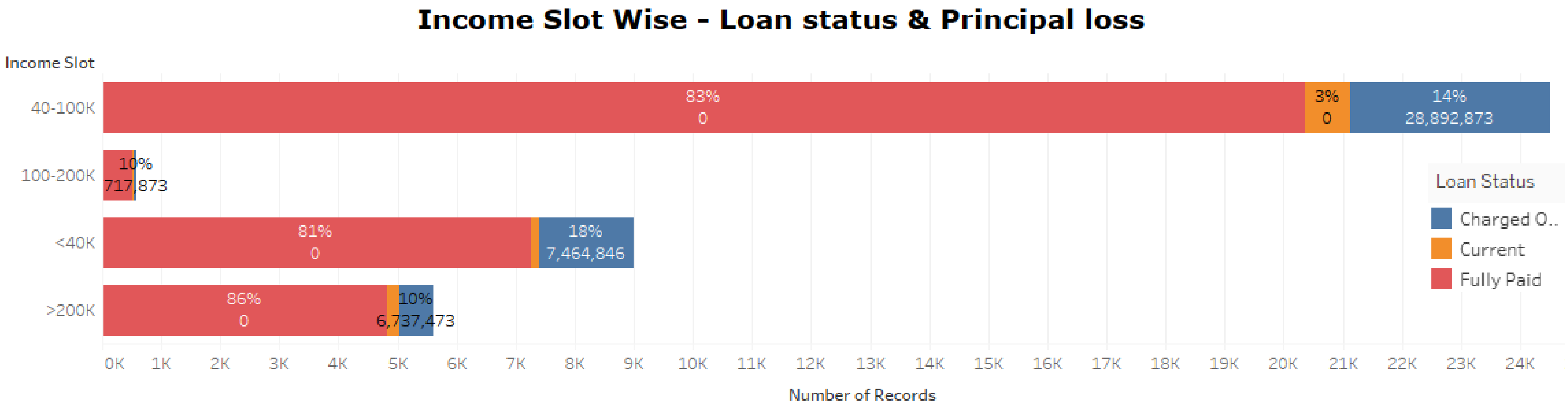


State Wise distribution of Charged off Customers

- California is the state with highest percentage of defaulters (19.99%) followed by Florida (8.96%), New York (8.80%) and Texas (5.62%).
- Increase number of loan applicants in California can be considered as a reason to lead to comparatively increased number of defaulters.
- Nevertheless, Florida outnumbers the number of defaulters as compared to New York.

Derived Metrics Analysis

Plot No :- 11



Analysis:

The marks are labeled by % of Total Number of Records and sum of Principal Loss.

Income Slot: Analysis

- < 40K: 18% defaulters contributing to 7M loss
- 40-100K: Almost 3400 defaulters - 29M loss - 14% loans defaulted in this category. Highest revenue loss in this category.
- 100K -200K: Safest category with least % wise defaulters & principal loss
- > 200K: 10% defaulters but still contributing to 7M loss. So this category also needs to be carefully scrutinized while applying for loans.

Analysis and Conclusion

- B, C, D these **loan grades** are most popular as they contribute to **maximum principal loss**.
- **Verification process** must be streamlined in order to increase the Customer credibility there by reducing the number of '**Charged off**' loan applicants.
- **Debt consolidation** has the major contribution as a product responsible for increased number of '**Charged off**' loans. **Credit card** being the second most common product leading same.
- The constant increase in the trend of 5 popular products can be seen as predictor for the definite growth demand of these product in future.
- This implies that the customer approaching for loans for products such as '**Debt consolidation**', '**Credit card**' must be very well scrutinized and mandatorily undergo verification process.
- **Employment length** of >10 years and <= 1 year account for increased number of '**Charged off**' applicants with **home ownership** status as **mortgaged/ staying on rent**. It can be believed that such applicants have compromised capacity for repayment of loan with the above mentioned home ownership status.

Model for Predictive Analysis

- The below observed 6 factors have a direct bearing on the defaulting tendency of a loan applier.
- If these characteristics are known at the time of loan application, hence can be crucial to avoid Business Loss / curb Principal Loss due to default.
- To prove this hypothesis, we apply a weighted average to the below 6 parameters.

Sr.	**Dimension**	**Observation**	**Weighted average for prediction**
1.	Income slot	Except 100-200K, all others cause loss in range of 7M~28M	APPLY WEIGHT 1 to all income groups other than 100-200K. The 100-200K bin has weight 0
2.	Grade	Grades C to G have more than 20% defaulters	APPLY WEIGHT 1 for grades C-G, for rest weight=0
3.	State	CA tops the list followed by big 4 states of NY, FL, TX, NJ these account for more than 50% of total defaulters	APPLY WEIGHT 2 to CA; WEIGHT 1 to NY, FL, TX, NJ. All others weight=0
4.	Home ownership	As expected, Rent and Mortgage top the list	APPLY WEIGHT 1 to RENT, MORTGAGE. For the rest, weight=0
5.	Employment length	Unemployed / Students default at 21% while others are in 13-14% defaulting rate	APPLY WEIGHT 1 to employment length not available. For the rest, weight =0
6.	Purpose	The Debt Consolidation is the topper in the list	APPLY WEIGHT 1 to debt consolidation. For the rest, weight=0

Model for Predictive Analysis

WEIGHTED AVERAGE METHOD:

Add up the weights based on the above parameters to get a defaulting score for a given application

If score > 5 , we conclude that the application is in High Risk Defaulting Zone.

Note: Weights were assigned based on the observations (% contribution, principal loss etc.) while threshold score for defaulting (>5) was obtained by trial.

PREDICTION ACCURACY:

The predicted values of an application are then compared with actual Loan Status to calculate accuracy of the hypothesis

Applying this method, we were able to successfully predict the outcome of **84.5%** of the applications (33581 out of 39718 correct predictions)