Lending Club Analysis

Case Study

Shakeeb Pasha and Vishal Khera

Agenda

1. Objectives
2. Insights & Assumptions
3. Recommendations

4. Appendix
   - Approach and Findings

Section

Objective

Object

We have been given the complete loan data for all loans issued through 2007 to 2011 for Lending Club

We need to surface insights based on the available information with the objective to

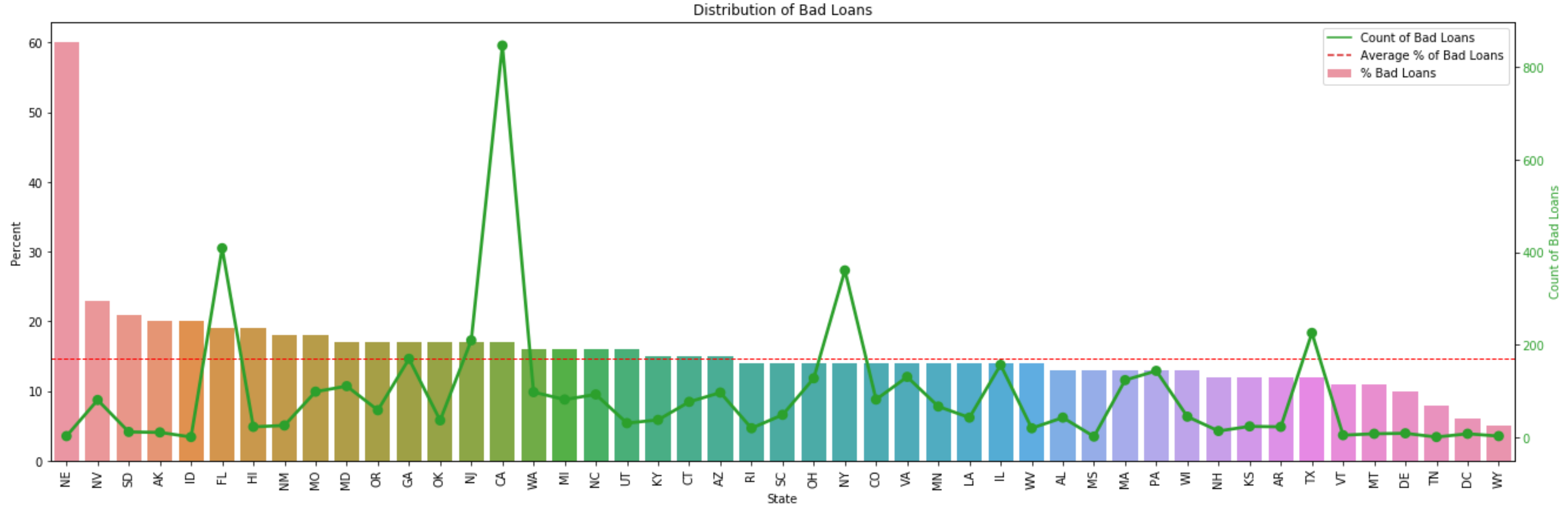
- Reduce losses for the business by improving the probability of identifying
  - Applicants that may not be able to pay back loans
  - Applicants that have a strong possibility of paying back loans

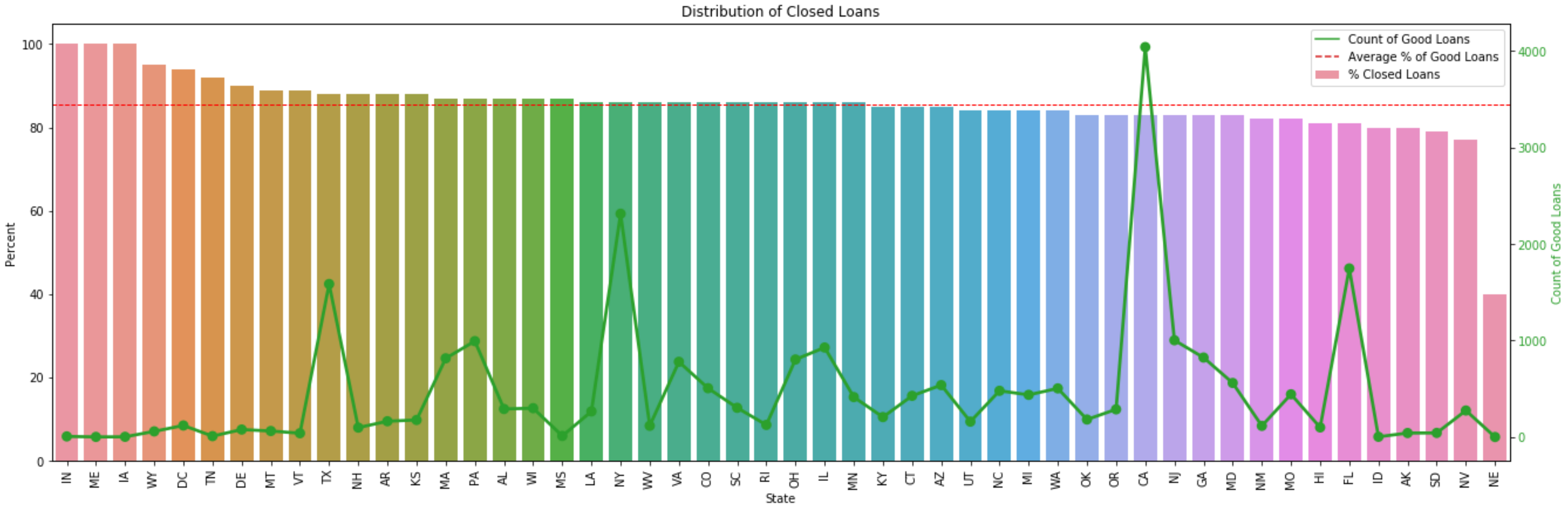
Section

Insights and Recommendations Cheat Sheet

The insights provided in the following slides can be used as a cheat sheet to identify the risk potential of loan applications.

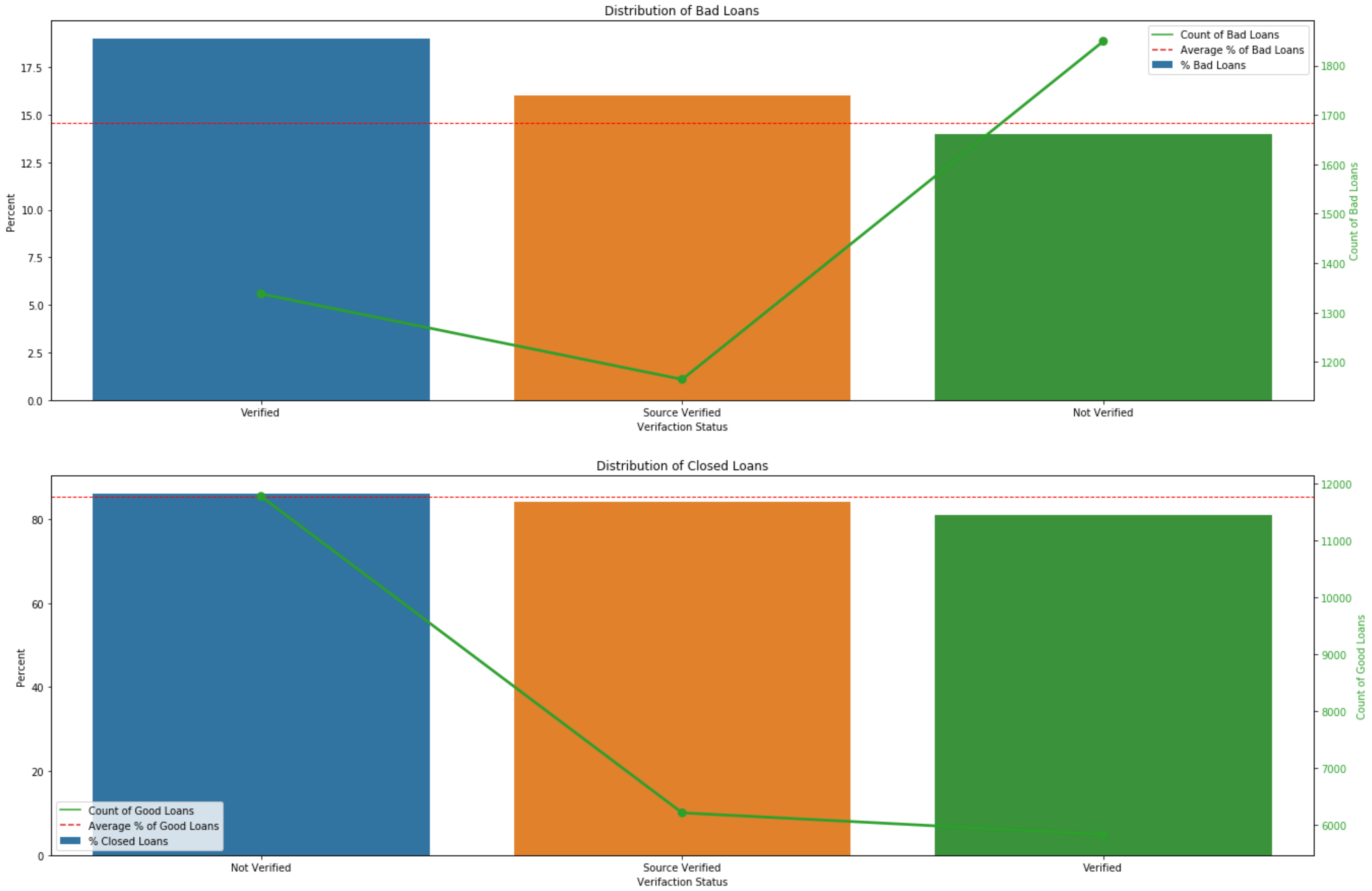
Segmented Univariate | Insights





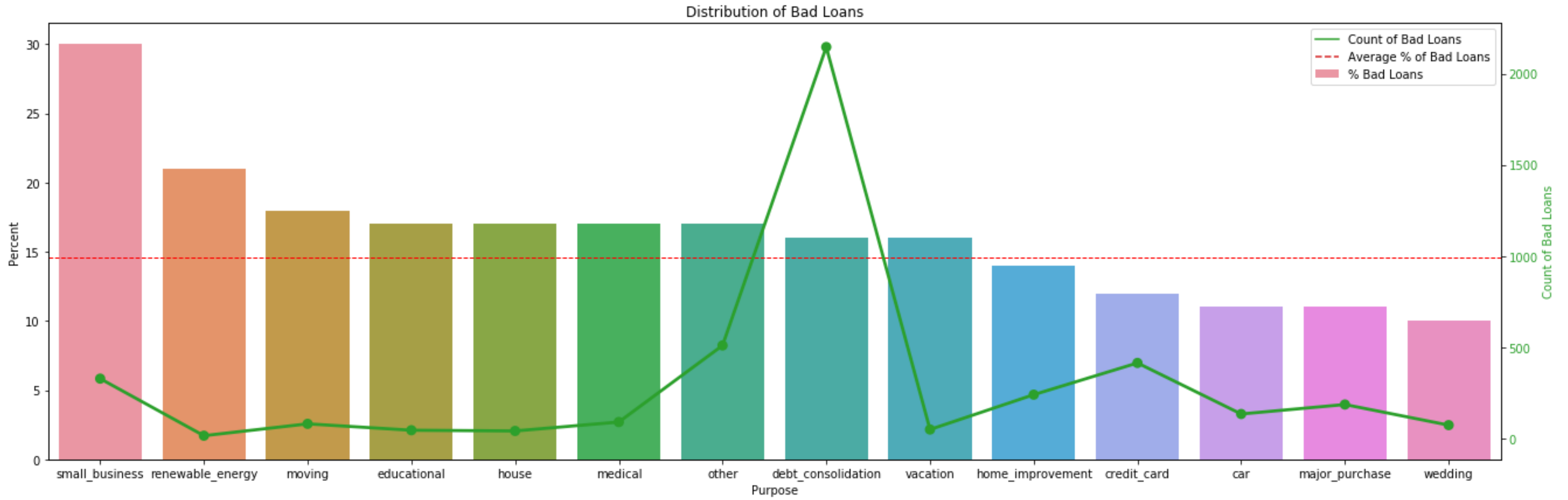
- High volume and higher than average bad loans are coming from FL, NJ, CA.
- NE, NV, AK, SD, are also high in their default rate even though the numbers are less
- NY, TX are good for Lending Club where the volumes are there and the default rates are low

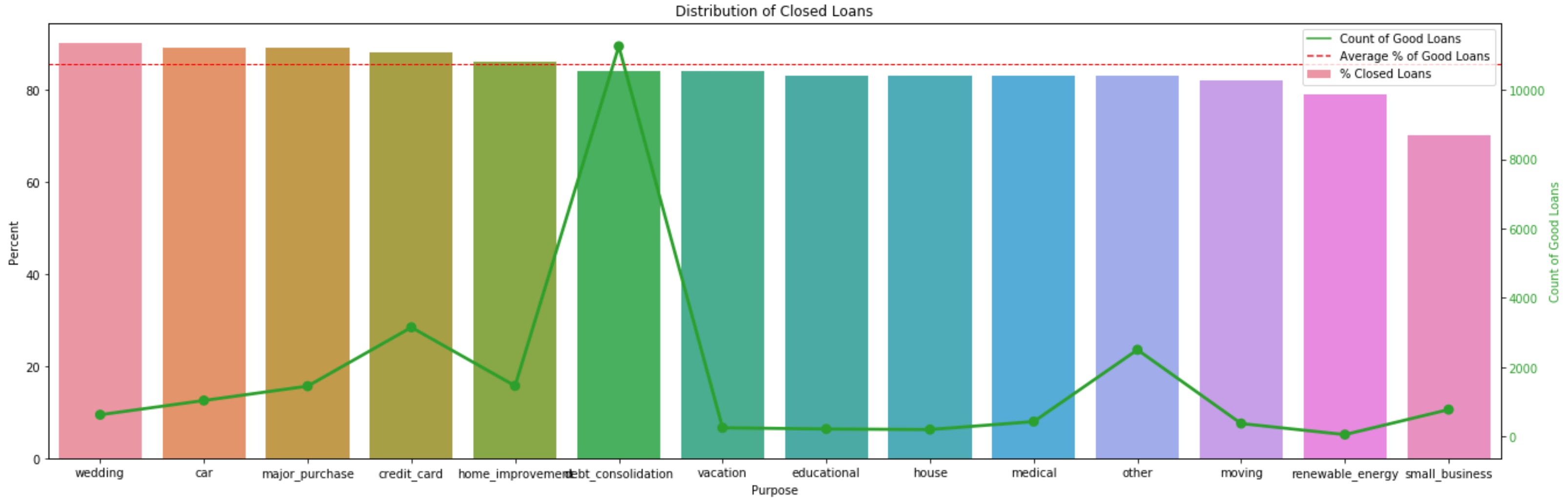
Segmented Univariate | Insights



- Verified applicants seem to have a have higher default rate! than not verified ones.

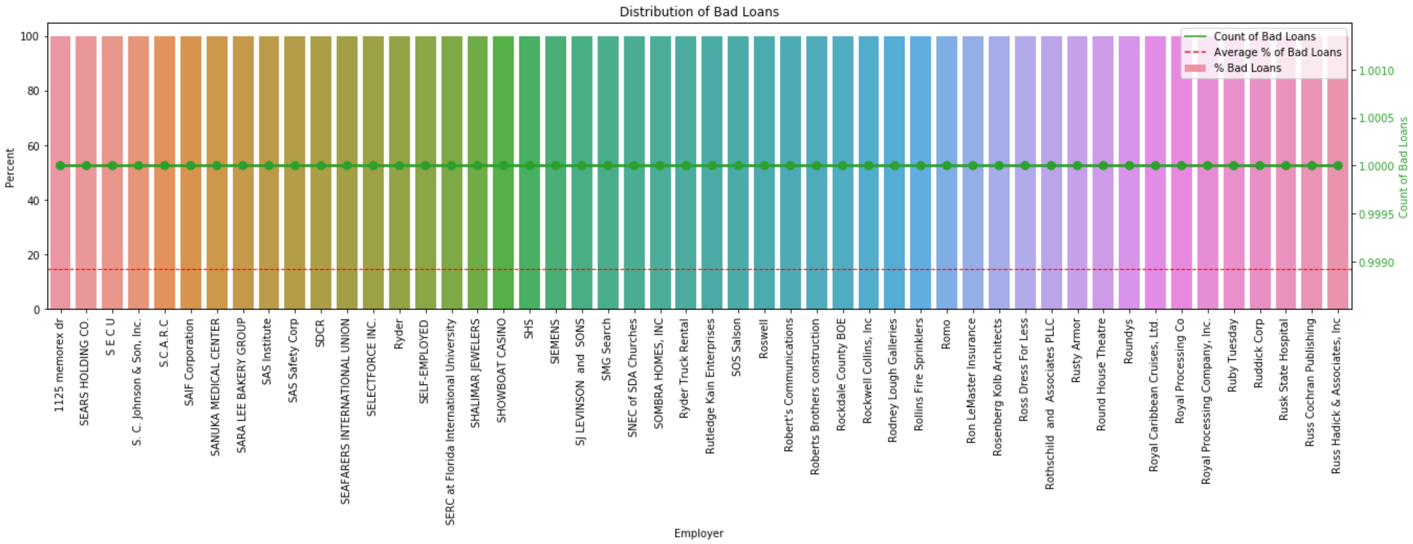
Segmented Univariate | Insights

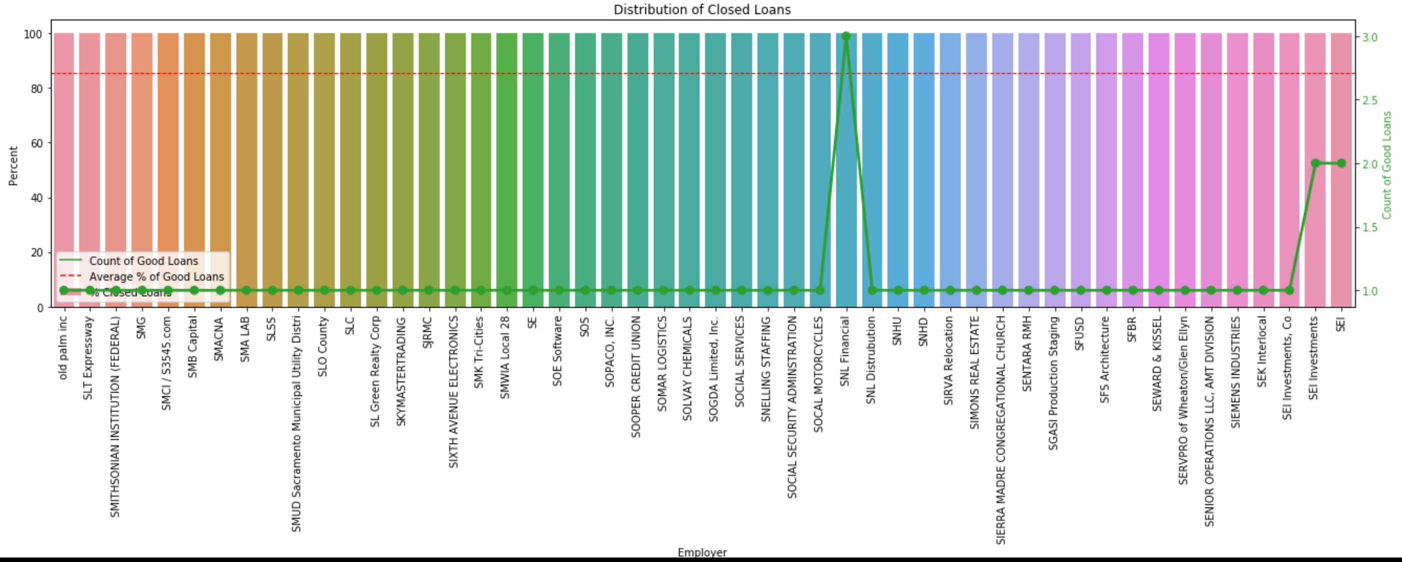




- Small business loans have the higher % of loans going bad and debt consolidation is the highest in terms of volumes but highest in terms of number of bad loans

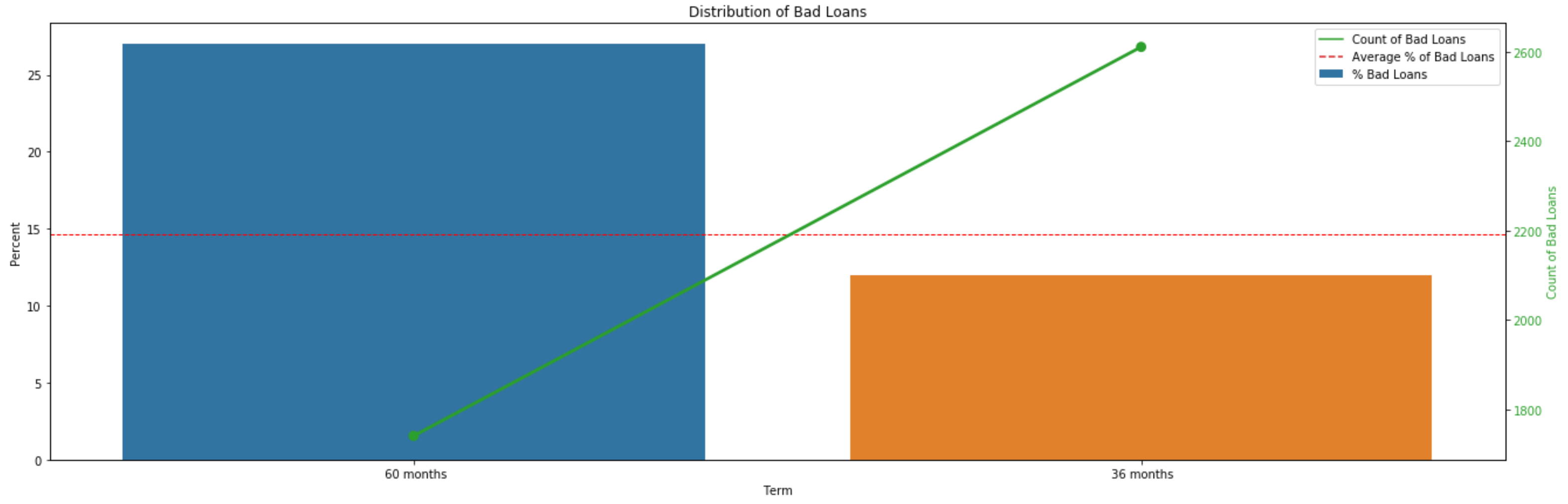
Segmented Univariate | Insights

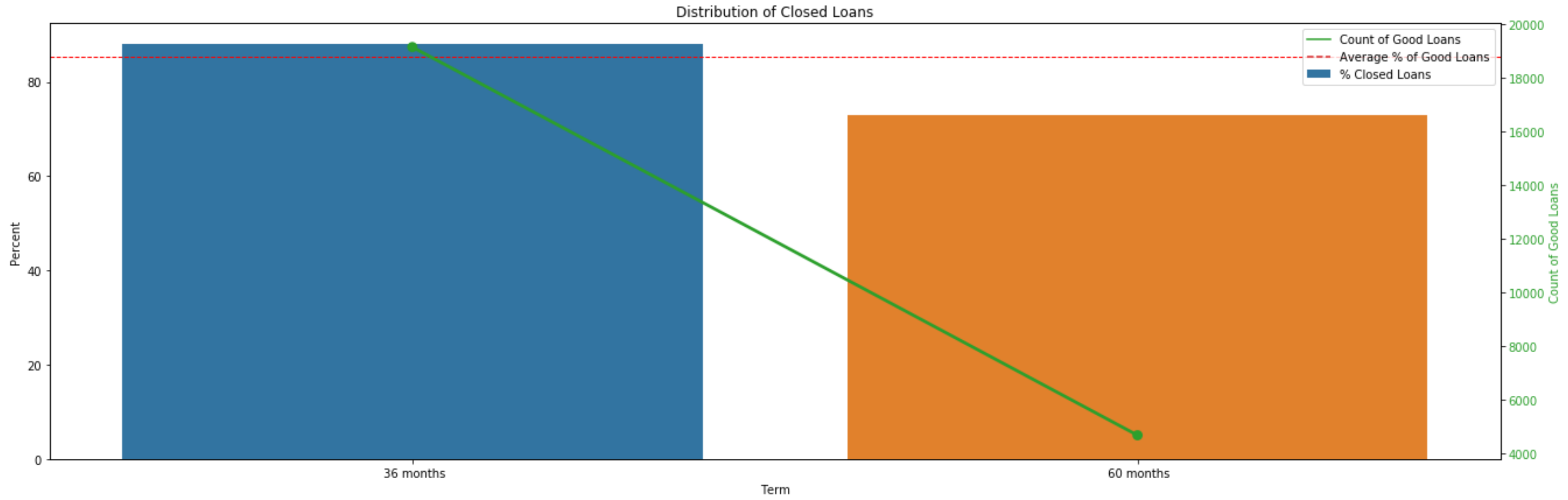




- On backgrounds of people, it is very scattered, however there are people working for companies and those companies have 100% default rate or 100% closed rate. Clearly company background though less volumes is a good predictor of bad loan
- Longer tenured loans have higher default rate
- %age wise “no available information” on current employee length seems to have the highest default rate. However the 10+ years seems to have a higher than average default rate with good volumes.
- Only A and B graded loans have the average default rates lower than average and the volumes of B is the highest in disbursement
- People with 0 Public records have the lowest default rates

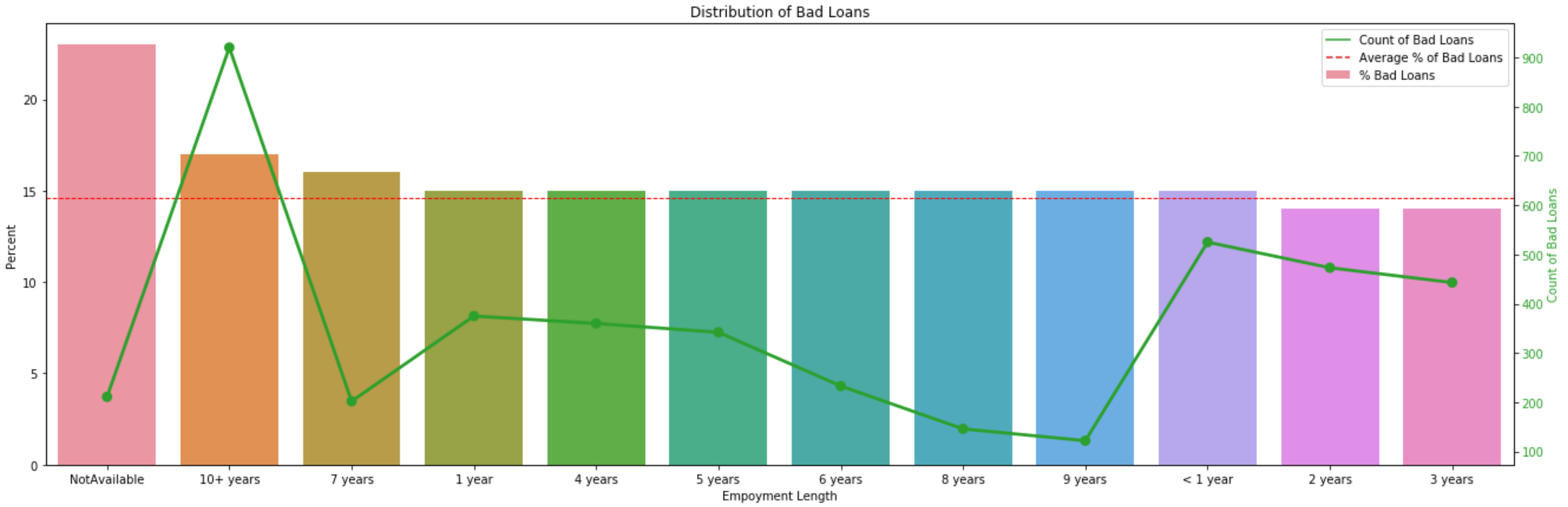
Segmented Univariate | Insights

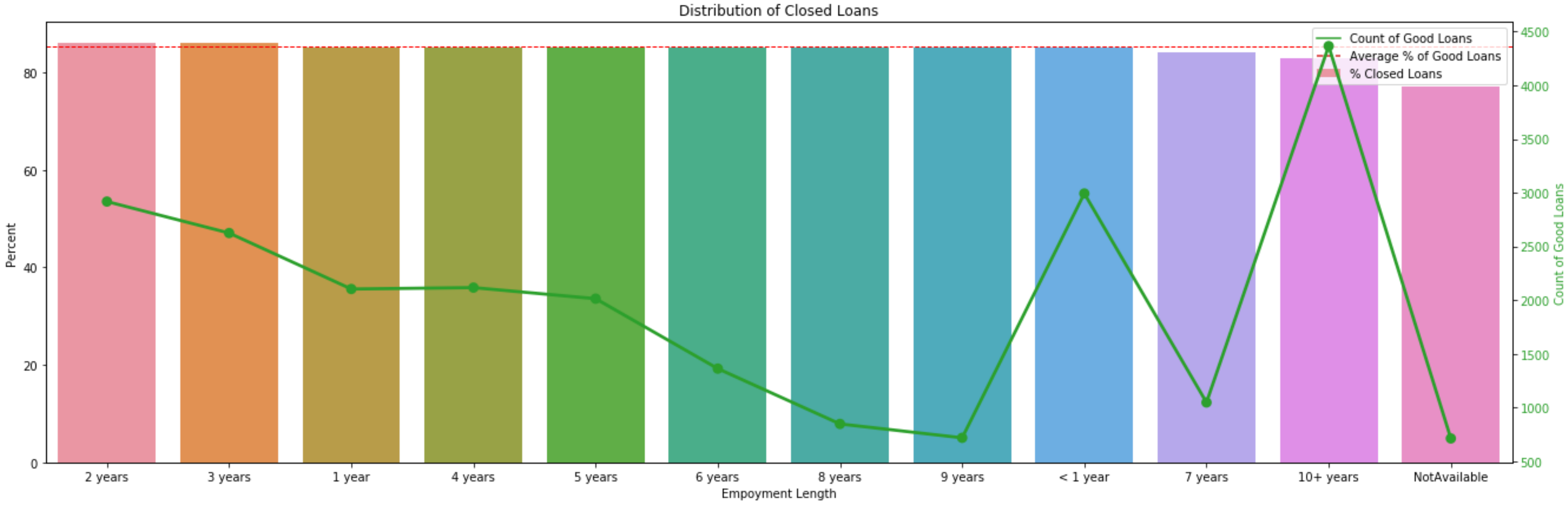




- Longer tenured loans have higher default rate

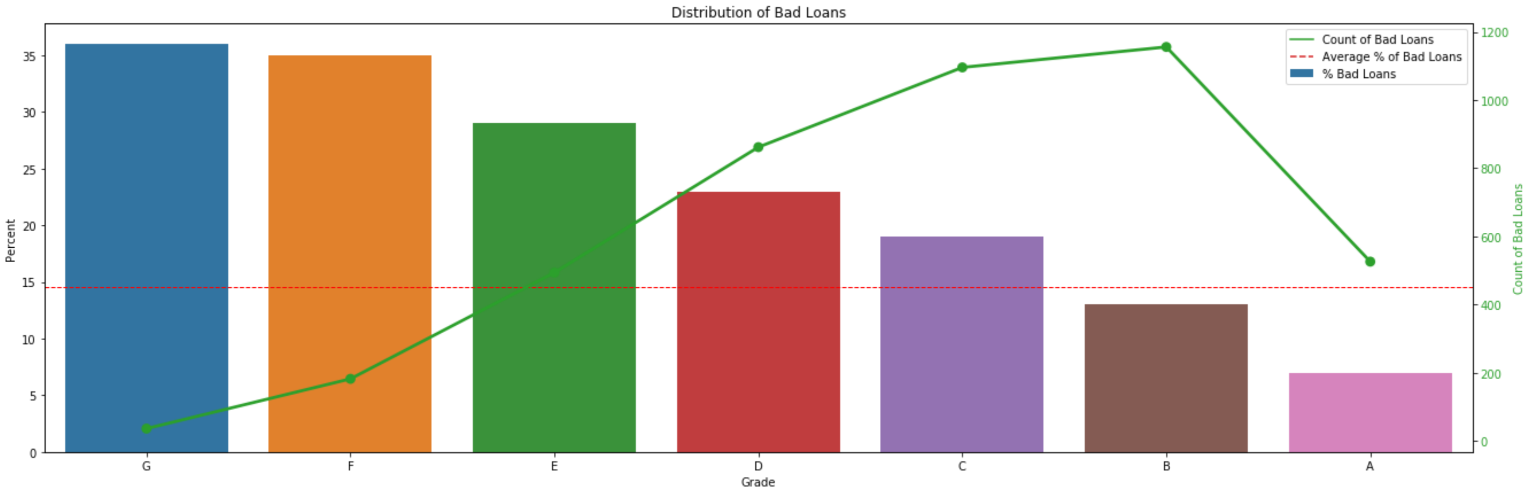
Segmented Univariate | Insights

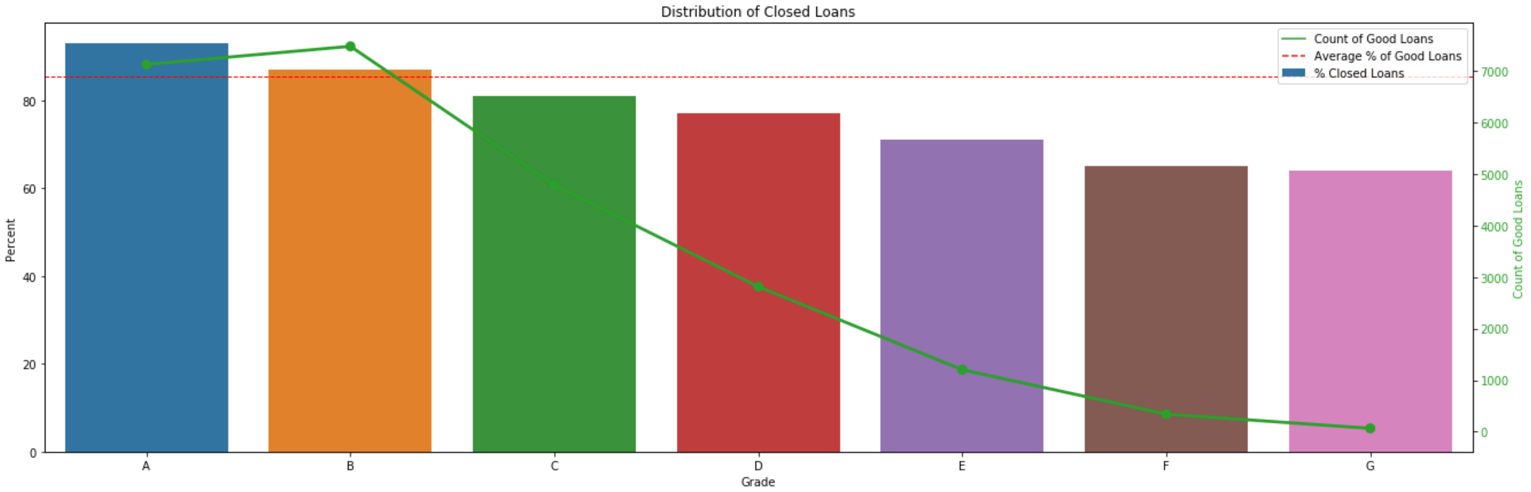




- %age wise “no available information” on current employee length seems to have the highest default rate.
- However the 10+ years seems to have a higher than average default rate with good volumes.

Segmented Univariate | Insights





- Only A and B graded loans have the average default rates lower than average and the volumes of B is the highest in disbursement

Bivariate Analysis | Cheat Sheet on States

State / House Ownership	AL	AK	AZ	AR	CA	CO	CT	DE	FL	GA	HI	ID	IL	IN	IA	KS	KY	LA	ME	MD	MA	MI	MN	MS	MO	MT	NE	NV	NH	NJ	NM	NY	NC	ND	OH	OK	OR	PA	RI	SC	SD	TN	TX	UT	VT	VA	WA	WV	WI	WY
OWN								100				100			100	91							91	100		90		29	100							91		90		91	100	100		25	100			28		
RENT												50		100	100				100							94	50	25	90												28							92		96
MORTGAGE	90			93							90	100		100	100												67															100								97
OTHERS			50	100		40	100		33	33			100			100				100	100	100	100	100	100	100		100		33		25	100		100					100			25		100				100	

Public Record / State	AL	AK	AZ	AR	CA	CO	CT	DE	FL	GA	HI	ID	IL	IN	IA	KS	KY	LA	ME	MD	MA	MI	MN	MS	MO	MT	NE	NV	NH	NJ	NM	NY	NC	ND	OH	OK	OR	PA	RI	SC	SD	TN	TX	UT	VT	VA	WA	WV	WI	WY
0	TRUE			90										100	100	90			100								60															92								95
1	29	100		30	25		38	100			50		27	100	100	40	31	39		32	30			100		100		33	40		29						26		100	92	100			35	100	28		91	90	100
2			100		100	100			100				100							100					100				100	67		50	50				100	100		100							100		25	
3																														100																				100
4																																														100				

- The following table shows Good Loans with 90+% probability of being closed with no default (green) and Bad loans with 25% probability or more of going bad (amber) against
  - State and House ownership of the applicant
  - State and Public records of the applicant

- Clearly some of the States are good such as IN, IA that irrespective of other parameters they probabilities are good and other states like NE, NV which have higher than default rates.

Bivariate Analysis | Cheat Sheet on Loan Purpose

Purpose / Loan Amt	(472.5,3250.0]	(3250.0,6000.0]	(6000.0,8750.0]	(8750.0,11500.0]	(11500.0,14250.0]	(14250.0,17000.0]	(17000.0,19750.0]	(19750.0,22500.0]	(22500.0,25250.0]	(25250.0,28000.0]
car	90		90	92				25	100	
credit_card	92	91		90					25	
debt_consolidation									25	34
educational						33	100	100		
home_improvement							32	90		43
house						96	100			100
major_purchase	92	90							31	100
medical						30	38			100
moving					100	100	25	100	100	
other							30	26		50
renewable_energy				33	100			100	100	
small_business		25	25	28	38	34	30	38	50	45
vacation				90	30	44	100	100	50	
wedding		94		91		92				

Public Record / Employee Length	A	B	C	D	E	F	G
car	94					50	100
credit_card	94	90			26	28	100
debt_consolidation	93				29	39	35
educational	94			28	43	100	100
home_improvement	94						
house	96	90		25	45	60	100
major_purchase	95	91			26	94	50
medical	93					44	50
moving						25	100
other	92			28	30	30	25
renewable_energy			31	100	100	33	
small_business		30	35	34	41	41	48
vacation				92	44	100	
wedding	97	92				25	100

- The following table shows Good Loans with 90+% probability of being closed with no default (green) and Bad loans with 25% probability or more of going bad (amber) against
  - Loan Purpose and Loan Amount
  - Loan Purpose and Rating of the Loan applicant

- Clearly house loan seems to be the least risk followed by wedding and loans to small business carries the highest default rate.
- Best track record is in Grade A loans and the changes of bad loans worsens with fall in Applicant’s Grade

Bivariate Analysis | Cheat Sheet on Loan Purpose - 2

Purpose / House Ownership	OWN	RENT	MORTGAGE	OTHER	NONE
car				50	
credit_card					
debt_consolidation					100
educational				100	
home_improvement					
house	91			100	
major_purchase			90	100	
medical					
moving			90	100	
other					
renewable_energy					
small_business	32	31		40	
vacation					
wedding	93	90		100	

Loan Purpose / Employee Length	<1	1	2	3	4	5	6	7	8	9	10+	NotAvailable
car			92						93	91		90
credit_card	90	91	90	90								27
debt_consolidation												
educational				37			100		100	100	93	43
home_improvement							91	90				
house				90		91	40	31	25	100		100
major_purchase			90		91			97	94			
medical							92	26	26	100		
moving							90	100		25		36
other										90		
renewable_energy	100			38	60	33	100		50	100		
small_business	28	33	25	29	31	28	27		32	29	35	31
vacation	25	100	93	90				50				
wedding	91	96			94	90	93		91	92		25

Loan Purpose / Public Record	0	1	2	3	4
car			100		
credit_card			50	100	
debt_consolidation				100	
educational		27			
home_improvement			100		
house		35			
major_purchase			33		100
medical		26	50		
moving					
other		25		100	
renewable_energy					
small_business	29	37	50		
vacation					
wedding	90				

- The following table shows Good Loans with 90+% probability of being closed with no default (green) and Bad loans with 25% probability or more of going bad (amber) against
  - Loan Purpose and House Ownership
  - Loan Purpose and Length of Employment in last organization
  - Loan Purpose and N. of Public records

- Reinforces the previous finding that
  - small business is a problem irrespective of any other parameter and
  - credit card loans is safe for people working in organizations where we have details of employee length.

Bivariate Analysis | Cheat Sheet on Loan Purpose - 3

Purpose / House Ownership	OWN	RENT	MORTGAGE	OTHER	NONE
car				50	
credit_card					
debt_consolidation					100
educational				100	
home_improvement					
house	91			100	
major_purchase			90	100	
medical					
moving			90	100	
other					
renewable_energy					
small_business	32	31		40	
vacation					
wedding	93	90		100	

Loan Purpose / Employee Length	<1	1	2	3	4	5	6	7	8	9	10+	NotAvailable
car			92						93	91		90
credit_card	90	91	90	90								27
debt_consolidation												
educational				37			100		100	100	93	43
home_improvement							91	90				
house				90		91	40	31	25	100		100
major_purchase			90		91			97	94			
medical							92	26	26	100		
moving							90	100		25		36
other										90		
renewable_energy	100			38	60	33	100		50	100		
small_business	28	33	25	29	31	28	27		32	29	35	31
vacation	25	100	93	90				50				
wedding	91	96			94	90	93		91	92		25

Loan Purpose / Public Record	0	1	2	3	4
car			100		
credit_card			50	100	
debt_consolidation				100	
educational		27			
home_improvement			100		
house		35			
major_purchase			33		100
medical		26	50		
moving					
other		25		100	
renewable_energy					
small_business	29	37	50		
vacation					
wedding	90				

- The following table shows Good Loans with 90+% probability of being closed with no default (green) and Bad loans with 25% probability or more of going bad (amber) against
  - Loan Purpose and House Ownership
  - Loan Purpose and Length of Employment in last organization
  - Loan Purpose and No. of Public records
- Reinforces the previous finding that
  - small business is a problem irrespective of any other parameter and
  - credit card loans is safe for people working in organizations where we have details of employee length.

Bivariate Analysis | Cheat Sheet on Loan Amounts

Emp Len / Loan Amt	(472.5,3250.0]	(3250.0,6000.0]	(6000.0,8750.0]	(8750.0,11500.0]	(11500.0,14250.0]	(14250.0,17000.0]	(17000.0,19750.0]	(19750.0,22500.0]	(22500.0,25250.0]	(25250.0,28000.0]
<1										38
1							27		27	25
2										40
3										
4									25	
5									25	
6										
7							29			33
8							28		41	100
9	93								25	80
10+							25	26	31	38
NotAvailable					31				47	50

Loan Amt / Grade	(472.5,3250.0]	(3250.0,6000.0]	(6000.0,8750.0]	(8750.0,11500.0]	(11500.0,14250.0]	(14250.0,17000.0]	(17000.0,19750.0]	(19750.0,22500.0]	(22500.0,25250.0]	(25250.0,28000.0]
A	90	94	93	94	93	94	91	98	94	
B										
C										28
D						26	33		31	43
E	28	26	28	30	30	26	37	31	30	50
F		31	37	35	27	33	38	44	48	25
G		50	78			43		33	33	

- The following table shows Good Loans with 90+% probability of being closed with no default (green) and Bad loans with 25% probability or more of going bad (amber) against
  - Loan Amount Length of Employment in last organization
  - Loan Amount and Grading of the Loan
- Reinforces the fact that
  - Higher loans have higher default rates and
  - Lower grade applicants have higher rate of default.
  - Grade A applicants seem to be the best in terms of closure rate

Bivariate Analysis | Misc.

UpGrad

Delinq / Grade	A	B	C	D	E	F	G
0	93				30	34	34
1	90				25	38	50
2					32	43	25
3	100			30		29	100
4	50	100	33	36	25	50	
5		100	25	100	100	100	
6		100	100	100	100		100
7			100	100	100	100	
8				100		100	
9							
10							
11		100					

Veridication Status / Deliq	0	1	2	3	4	5	6	7	8	11
Not Verified				25	36	29	100	100	100	
Source Verified						100	100	100	100	100
Verified					40	100	25	100		

Tenure / Emp Len	<1	1	2	3	4	5	6	7	8	9	10+	NotAvailable
36					28	90	100	100	100			
60		27	26	29	29	50	33	50	100			

- The following table shows Good Loans with 90+% probability of being closed with no default (green) and Bad loans with 25% probability or more of going bad (amber) against
  - Grade of Applicant's loan and no. of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
  - Verification Status and no. of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
  - Employee Length and the duration of the loan

- Reinforces the fact that
  - Lower grade loans have higher default rates irrespective of the delinquency number
  - From 3 and above Delinquency rate we see it impacts the loan going bad irrespective of the Verification status
  - Higher tenured loan has a higher default rate than lower tenured loans.

Section

Appendix

APPROACH

Approach

Data Understanding

- Import Data from csv
- Understand the Data Types and Shape
- Identify target column for analysis (loan status)

Data Clean up - 1

- Remove unused columns (e.g. URL, description)
- Convert % into numbers
- Identify and Remove missing columns (assumption 50% and above empty)
- Remove columns that have 0 as a value across all rows
- Remove columns with single value across all rows (5 columns identified)
- Remove columns with single value across all rows (5 columns identified)
- Drop duplicate rows if any
- Potentially 2 Imputes values based on mean, median or mode (not done)

Data Clean up - 2

- Drop "current" loans given we are not sure if these will go bad or not. Not relevant for study.
- Categorizing variables based on (see Note-1)
  - Applicant Related
  - Loan Characteristics
  - Customer Behaviour
- Drop all behaviour related columns as these are not used for loan processing, these come after loan has been issued.

Univariate Analysis

- Categorize all variables into (see Note-2)
  - Not Relevant
  - Unordered
  - Ordered and
  - Continuous
- Drop all the 'Not Relevant; columns
- Plot graphs of the ***unordered categorical variables*** and search for insights
- Plot graphs of the ***Ordered categorical variables*** and search for insights
- Plot Histograms for ***continuous variables*** and search for insights
- Carry out ***Segmented Univariate Analysis*** for Bad loans and Good loans and search for Insights

Bivariate Analysis

- Select subset of columns for Bivariate analysis
- Get Heatmap on the correlation b/w the columns and identify insights on ***correlation***
- Apply Category codes for unordered columns e.g. State, verification status, purpose, grade, month, home ownership, term, emp_length…
- Get Pair plot on the variables to see if there are any relationships
- Keep single column among the high correlated sets of data and drop the others in that set.
- Apply heatmaps on paired columns for bivariate analysis and search for insights

Note – 1 : Segregation of Columns

APPLICATION RELATED	LOAN CHARACTERISTICS	CUSTOMER BEHAVIOUR
id	loan_amnt	out_prncp
member_id	funded_amnt	out_prncp_inv
emp_title	funded_amnt_inv	total_pymnt
emp_length	term	total_pymnt_inv
home_ownership	int_rate	total_rec_prncp
annual_inc	installment	total_rec_int
title	grade	total_rec_late_fee
zip_code	sub_grade	recoveries
addr_state	verification_status	collection_recovery_fee
earliest_cr_line	issue_d	last_pymnt_d
delinq_2yrs	loan_status	last_pymnt_amnt
inq_last_6mths	purpose	last_credit_pull_d
open_acc	dti	revol_bal
pub_rec		revol_utl
total_acc		
pub_rec_bankruptcies		

Note – 2 : Segregation of Variables

UpGrad

NOT RELEVANT	UNORDERED	ORDERED	CONTINUOUS
id	title	term	int_rate
member_id	addr_state	emp_length	loan_amnt
earliest_cr_line	verification_status	Grade	installment
zip_code	total_acc	sub_grade	dti
	purpose	loan_status	funded_amnt
	home_ownership	pub_rec	funded_amnt_inv
	emp_title	issue_d	annual_inc
		inq_last_6mths	
		pub_rec_bankruptcies	
		delinq_2yrs	
		open_acc	

Univariate Analysis | Insights

Unordered Categorical

- Majority of the Club’s business is in 10 states
- A significant portion of the loans show as non-verified
- Majority of loans taken are for Debt Consolidation
- Majority of loans taken are by people with rented / mortgaged homes
- The number of credit lines on the Borrower's file follows a standard distribution

Ordered Categorical

- Almost 75% of the loans are of tenure 36 months
- Majority of Loans are given to people with jobs. There are close to 1000 people who have not disclosed or are not in jobs
- Majority of loans are with A,B and C gradings
- Further breakdown by sub grade shows D, E and F grades are not encouraged with the outlier of D2 which seems to be high in comparison
- Employee Title Contains too much text, no processing done on this column
- 85.4% of loans closed have been fully paid up, 14.6% of loans have gone bad
- We have about 2100 cases with derogatory public records which is a little over 5% of the population.
- No real abnormal patterns seen in the distribution of loans by month
- Not many loan applications with more than 3 inquiries were approved
- About 94% of the loans were given to applicants with 0 public record bankruptcies and about 4% with 1 record
- About 2.5% of the applicants given loan have had over 1 incident of delinquency in the last 2 years
- The number of credit lines on the Borrower's file follows a standard distribution

Univariate Analysis | Insights contd.

Segmented

- High volume and higher than average bad loans are coming from FL, NJ, CA.
- NE, NV, AK, SD, are also high in their default rate even though the numbers are less
- NY, TX are good for Lending Club where the volumes are there and the default rates are low
- Verified applicants seem to have a have higher default rate! than not verified ones.
- Small business loans have the higher % of loans going bad and debt consolidation is the highest in terms of volumes but highest in terms of number of bad loans
- People who have mortgaged houses seem to have lesser rate of default
- On backgrounds of people, it is very scattered, however there are people working for companies and those companies have 100% default rate or 100% closed rate. Clearly company background though less volumes is a good indicator of bad loan
- Longer tenured loans have higher default rate
- %age wise “no available information” on current employee length seems to have the highest default rate. However the 10+ years seems to have a higher than average default rate with good volumes.
- Only A and B graded loans have the average default rates lower than average and the volumes of B is the highest in disbursement
- People with 0 Public records have the lowest default rates

Bivariate Analysis | Insights

Correlation Analysis

- From the calculated Correlation coefficients we see strong correlation seen b/w
  - Number of derogatory public records (pub_rec) and Number of public record bankruptcies (pub_rec_bankruptcies)
  - Loan amount applied (loan_amnt) and monthly installment by the borrower (installment)
  - Loan amount applied (loan_amnt) and the committed Funding amount (funded_amnt)
  - Loan amount applied (loan_amnt) and committed funding amount by the investors (funded_amnt_inv)
  - Committed Funding amount (funded_amnt) and committed funding amount by the investors (funded_amnt_inv)
  - Monthly installment by the borrower (installment) and committed funding amount by the investors (funded_amnt_inv)

Paired Bivariate Analysis (90%+ confidence for Bad Loans)

- From the various heatmap plots we see
  - States of AR / Others have a 90% of going bad
  - States of CT / Others have a 100% of being good
  - States of DC / Others have a 100% of being good
  - States of DC / Mortgages have a 100% of being good
  - States of DC / Others have a 100% of being good
  - States of DC / Own have a 100% of being good
  - States of DC / Rent have a 92% of being good
  - States of DE / Own have a 100% of being good
  - States of HI / Mortgages have a 90% of being good
  - States of IA / Mortgages have a 100% of being good
  - States of IA / Own have a 100% of being good
  - States of IA / Rent have a 100% of being good
  - States of ID / Own have a 90% of being good