Lending Club Case Study--info

Problem statement:

You work for a consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. The data given contains information about past loan applicants and whether they 'defaulted' or not.

Client is looking to deep dive into the data and get insights to help them reduce the overall default cases.

Expected results:

The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

Data sets used for analysis:

- Loan Data set with 39k records
- Data dictionary explaining fields present in the loan data set
- No third party data has been used during analysis

Data received

Below table represents various fields received in the data set. Fields highlighted in Yellow were considered to perform the analysis

Column Name	DataType	Column Name	DataType	Column Name	DataType	Column Name	DataType
id	int64	mths_since_last_record	object	tot_cur_bal	object	mths_since_recent_inq	object
member_id	int64	open_acc	int64	open_acc_6m	object	mths_since_recent_revol_delinq	object
loan_amnt	int64	pub_rec	int64	open_il_6m	object	num_accts_ever_120_pd	object
funded_amnt	int64	revol_bal	int64	open_il_12m	object	num_actv_bc_tl	object
funded_amnt_inv	float64	revol_util	object	open_il_24m	object	num_actv_rev_tl	object
term	int32	total_acc	int64	mths_since_rcnt_il	object	num_bc_sats	object
int_rate	float64	initial_list_status	object	total_bal_il	object	num_bc_tl	object
installment	float64	out_prncp	float64	il_util	object	num_il_tl	object
grade	object	out_prncp_inv	float64	open_rv_12m	object	num_op_rev_tl	object
sub_grade	object	total_pymnt	float64	open_rv_24m	object	num_rev_accts	object
emp_title	object	total_pymnt_inv	float64	max_bal_bc	object	num_rev_tl_bal_gt_0	object
emp_length	object	total_rec_prncp	float64	all_util	object	num_sats	object
home_ownership	object	total_rec_int	float64	total_rev_hi_lim	object	num_tl_120dpd_2m	object
annual_inc	float64	total_rec_late_fee	float64	inq_fi	object	num_tl_30dpd	object
verification_status	object	recoveries	float64	total_cu_tl	object	num_tl_90g_dpd_24m	object
issue_d	datetime64[ns]	collection_recovery_fee	float64	inq_last_12m	object	num_tl_op_past_12m	object
loan_status	object	last_pymnt_d	datetime64[ns]	acc_open_past_24mths	object	pct_tl_nvr_dlq	object
pymnt_plan	object	last_pymnt_amnt	float64	avg_cur_bal	object	percent_bc_gt_75	object
url	object	next_pymnt_d	datetime64[ns]	bc_open_to_buy	object	pub_rec_bankruptcies	object
desc	object	last_credit_pull_d	datetime64[ns]	bc_util	object	tax_liens	object
purpose	object	collections_12_mths_ex_med	object	chargeoff_within_12_mths	object	tot_hi_cred_lim	object
title	object	mths_since_last_major_derog	object	delinq_amnt	int64	total_bal_ex_mort	object
zip_code	object	policy_code	int64	mo_sin_old_il_acct	object	total_bc_limit	object
addr_state	object	application_type	object	mo_sin_old_rev_tl_op	object	total_il_high_credit_limit	object
dti	float64	annual_inc_joint	object	mo_sin_rcnt_rev_tl_op	object	Isdefault	int32
delinq_2yrs	int64	dti_joint	object	mo_sin_rcnt_tl	object	SalaryRange	category
earliest_cr_line	datetime64[ns]	verification_status_joint	object	mort_acc	object	Loanrange	category
inq_last_6mths	int64	acc_now_delinq	int64	mths_since_recent_bc	object		
mths_since_last_delinq	object	tot_coll_amt	object	mths_since_recent_bc_dlq	object		

Total no of records : 39,717
Data time frame : 06-Jan-2007 – 12-Jan-2011
Total Disbursed Loan amount: 445,602,650
Total defaulted loans : 4,312

Data preparation

Below are some of the data preparation/cleansing performed

- Converted several fields to appropriate data types. Notable conversions include:
  - Dates were converted from object type to datetime for fields such as "issue_d", "earliest_cr_line", "last_pymnt_d", "last_credit_pull_d", and "next_pymnt_d".
  - Cleansed the "term" field by removing the keyword "months" and then converted its data type from object to integer.
  - Removed default values across all fields. Any missing values (NA) were replaced with blank values.
  - Cleansed the "int_rate" field by removing the "%" symbol and then converted its data type from object to float.
- Additionally, a calculated field named "Isdefault" was created. This field was populated with the value "1" where the "delinq_2yrs" field is greater than or equal to 1.

Performed analysis using describe on the entire dataset to identify any outliers

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	annual_inc	issue_d	dti	delinq_2yrs	earliest_cr_line	inq_last_6mths	mths_since_last_delinq	mths_since_last_record	collection_recovery_fee
count	39717	39717	39717	39717	39717	39717	39717	39717	39717	39717	39717	39717	39717	14035	2786	39717
mean	683131.9131	850463.5594	11219.44381	10947.7132	10397.44887	12.02117657	324.5619221	68968.92638	2010-05-07 04:13:04	13.31512954	0.146511569	1996-07-30 03:59:11	0.869199587	35.90096188	69.69813352	12.40611189
min	54734	70699	500	500	0	5.42	15.69	4000	2007-01-06 00:00:00	0	0	1946-01-01 00:00:00	0	0	0	0
25%	516221	666780	5500	5400	5000	9.25	167.02	40404	2010-01-05 00:00:00	8.17	0	1993-01-11 00:00:00	0	18	22	0
50%	665665	850812	10000	9600	8975	11.86	280.22	59000	2011-01-02 00:00:00	13.4	0	1998-01-05 00:00:00	1	34	90	0
75%	837755	1047339	15000	15000	14400	14.59	430.78	82300	2011-01-08 00:00:00	18.6	0	2001-01-09 00:00:00	1	52	104	0
max	1077501	1314167	35000	35000	35000	24.59	1305.19	6000000	2011-01-12 00:00:00	29.99	11	2008-01-11 00:00:00	8	120	129	7002.19
std	210694.1329	265678.3074	7456.670694	7187.23867	7128.450439	3.724825435	208.8748735	63793.76579		6.678593595	0.491811516		1.070219332	22.02005955	43.82252903	148.6715935

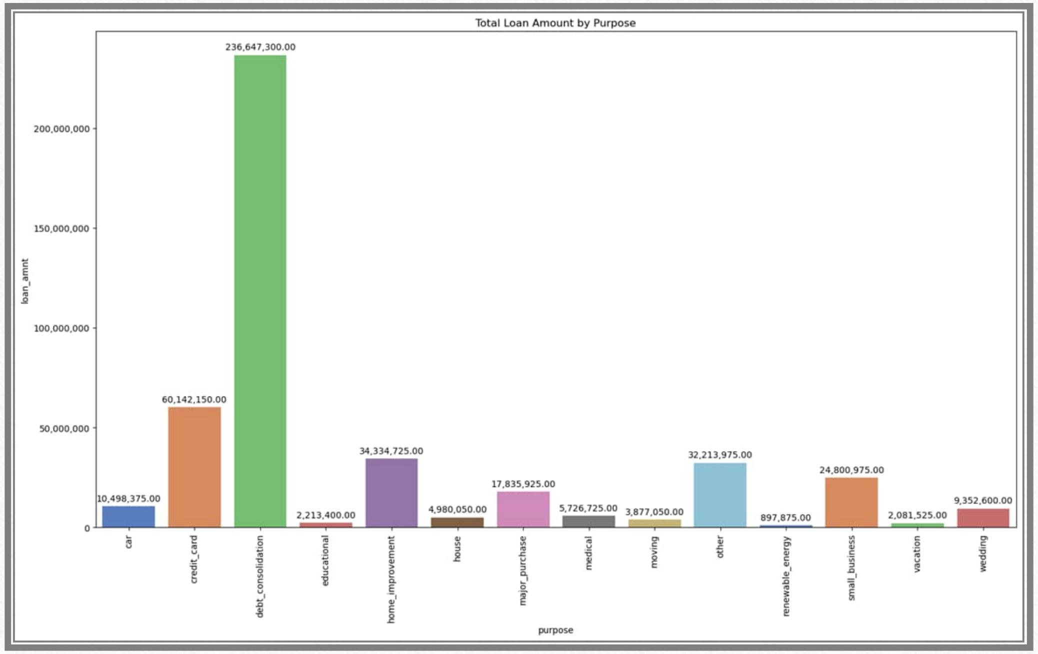
Technologies used

- Python 3.0
- Excel
- Various Python libraries
  - Pandas
  - Numpy
  - Matplotlib
  - Seaborn

Data Analysis

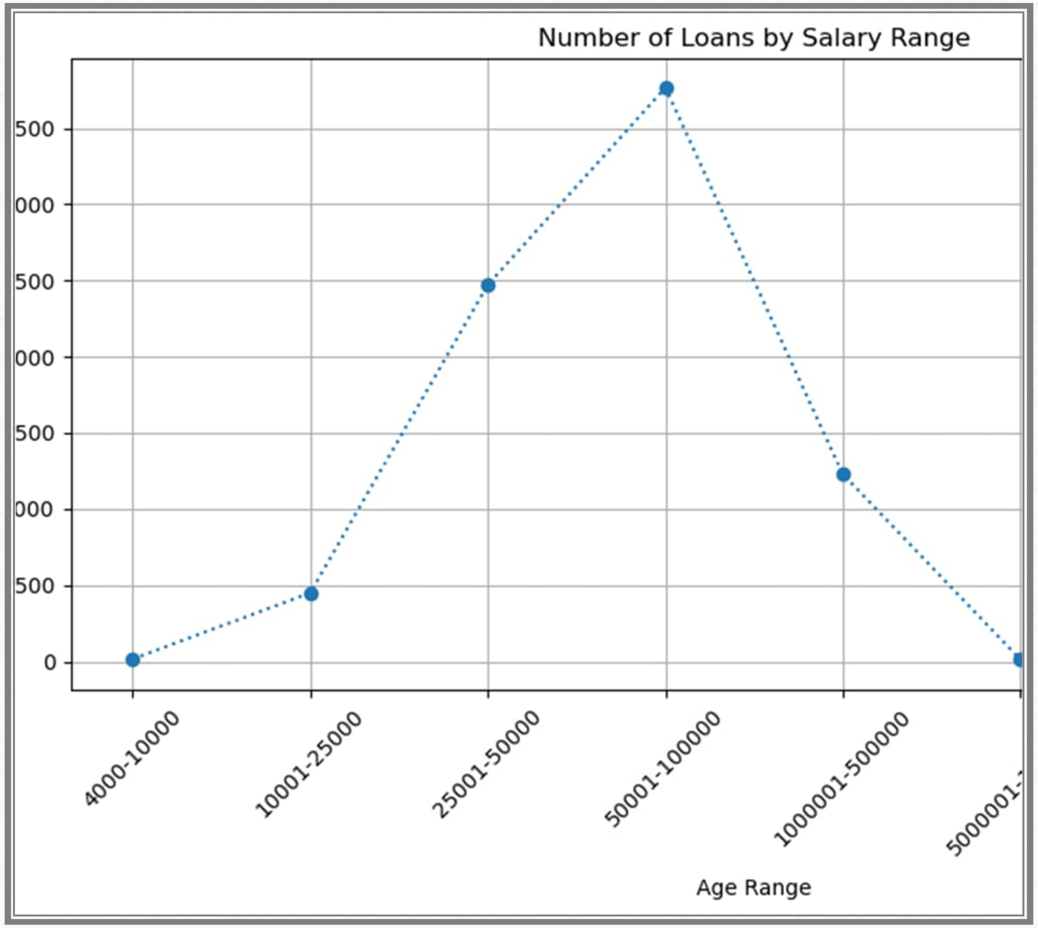
"Debt consolidation" is the common purpose used by 53% of the customers

- 53% of the loans approved(based on amount) have a generic purpose mentioned
- Second in the list is to pay off credit card bill which is something to review as these loans are given to people from Urban areas
- Agents should be instructed to provide proper details under purpose field which can lead to meaningful insights



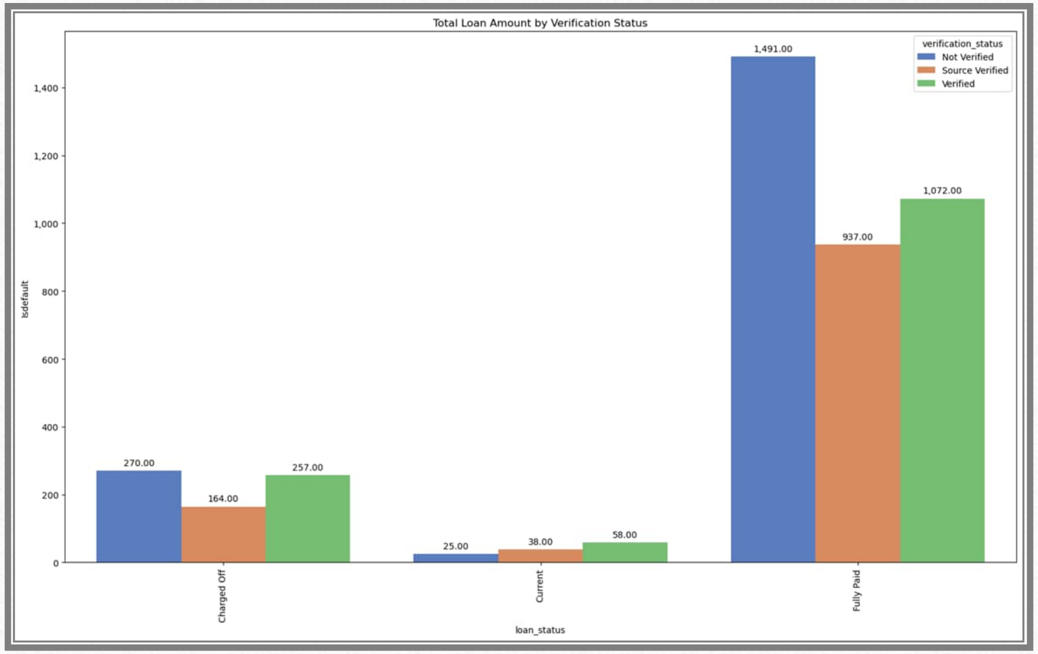
Majority of the loans are issued to customers between salary range of 50k-1 Lakh

- ~50% of the loans approved(based on amount) are to customers whose salary ranges between 50k to 1 Lakh per annum



Analysis confirms loans to verified customers have less default rates

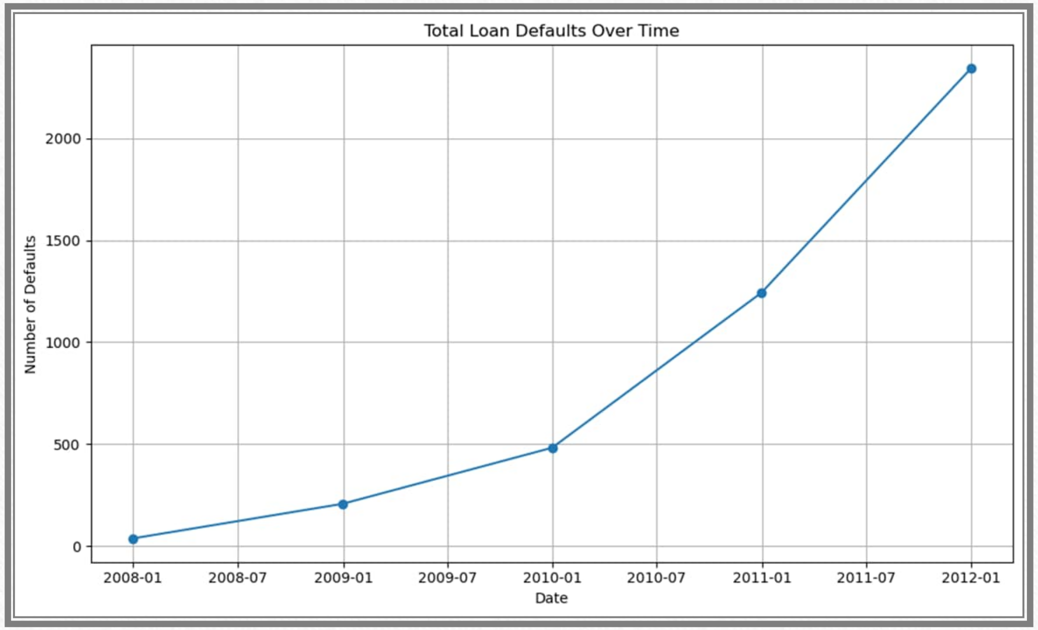
- More than 40 % of the loans defaulted are from Non verified customers



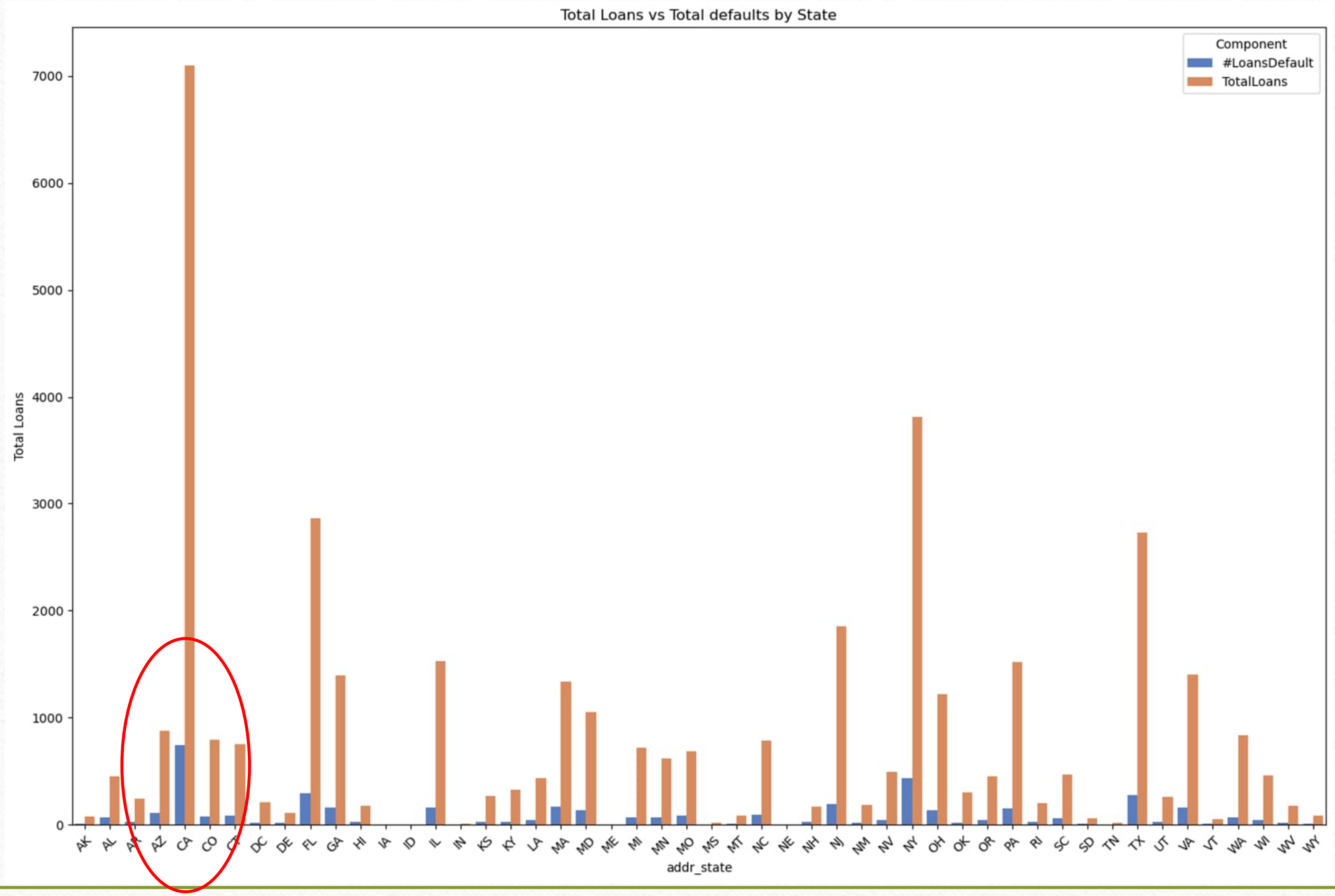
There is an exponential growth in the defaulted loans based on the issue date

- More than 50 % of the loans defaulted are of loans which are issued in the year 2012

- It is evident from the graph that whatever checks that were previously performed are no longer effective and are not reducing the defaulted cases



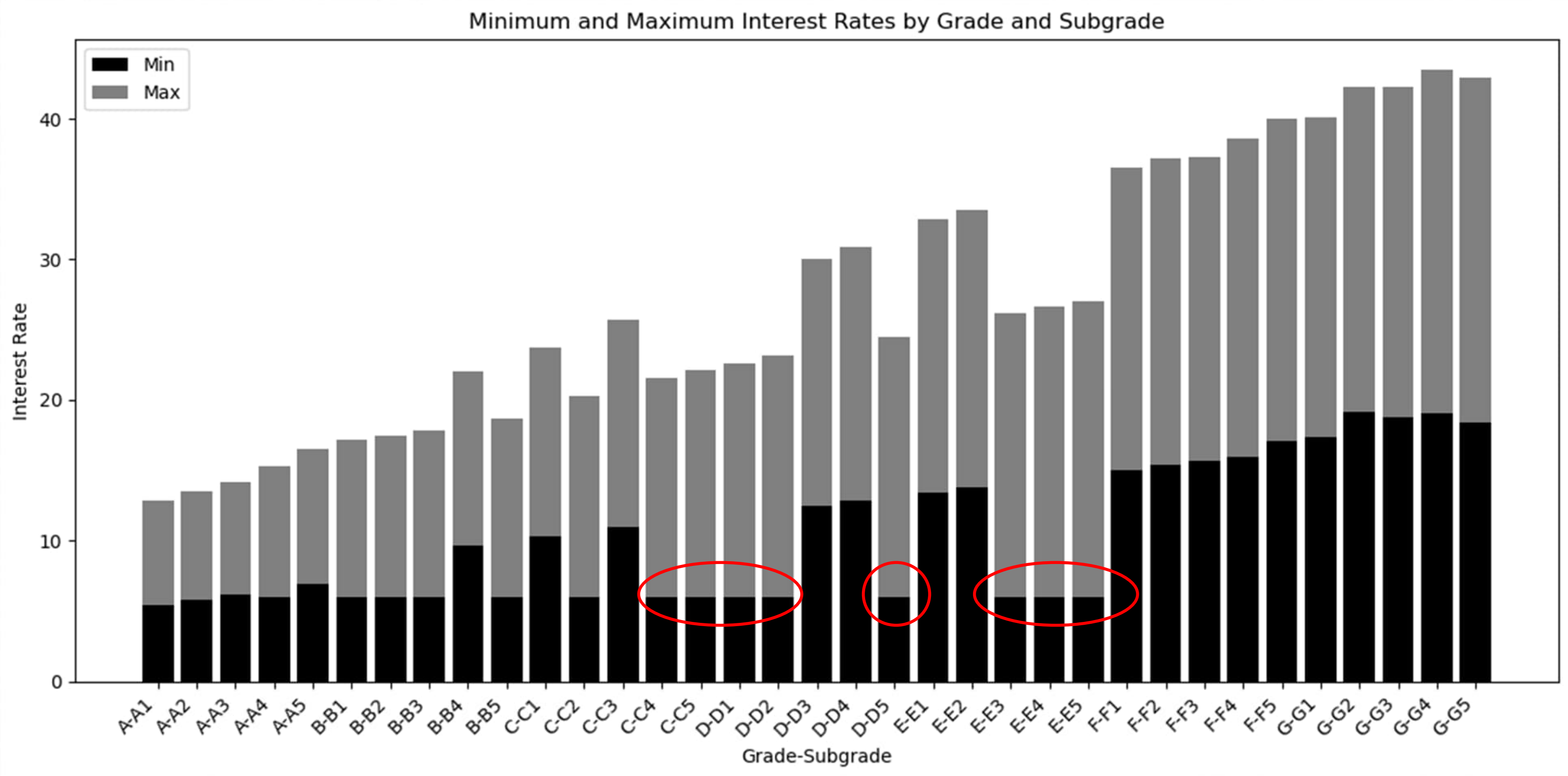
California(CA) has the max no of loans disbursed and in line defaults



- Based on the analysis 15% of the loans from CA have defaulted
- Overall the default rate of loans across locations is ~11%

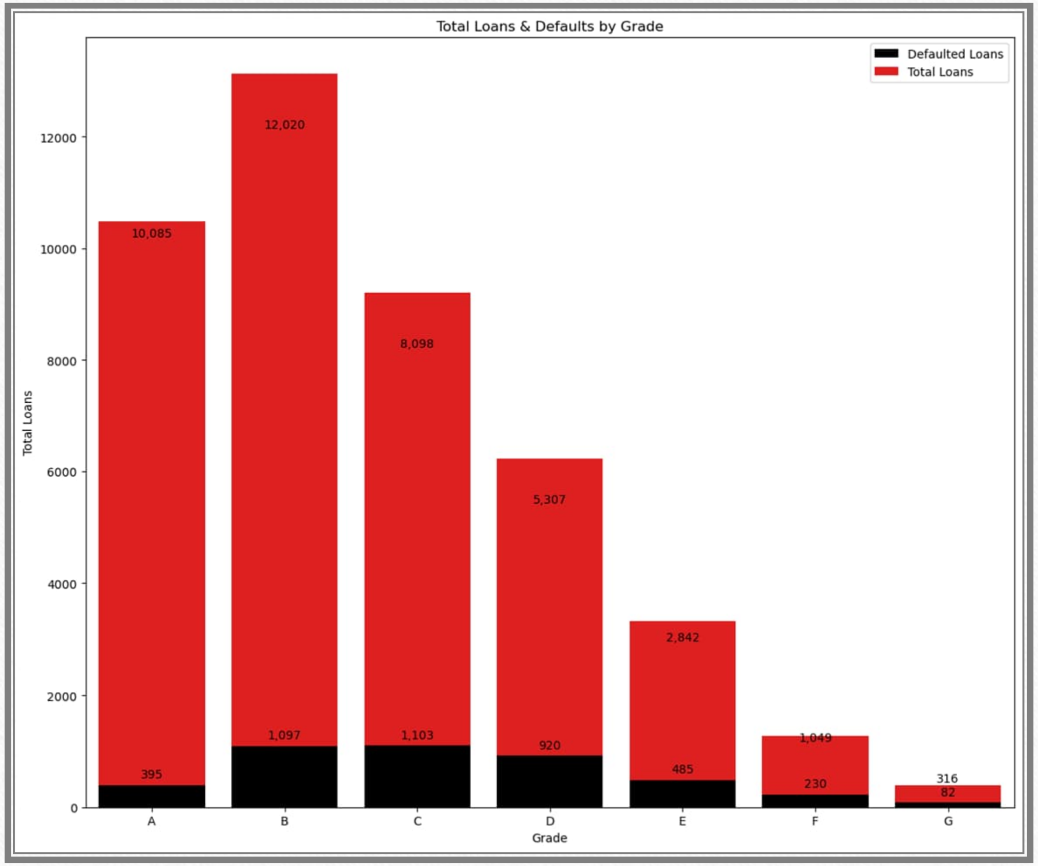
Loans to verified customers have less default rates

- Inconsistent interest rates offered to customers. For E.g. customers with grades E3/E4/E5 are offered less interest rates than customers with Grade D3/D4



Loans to customers having good grades have less defaults

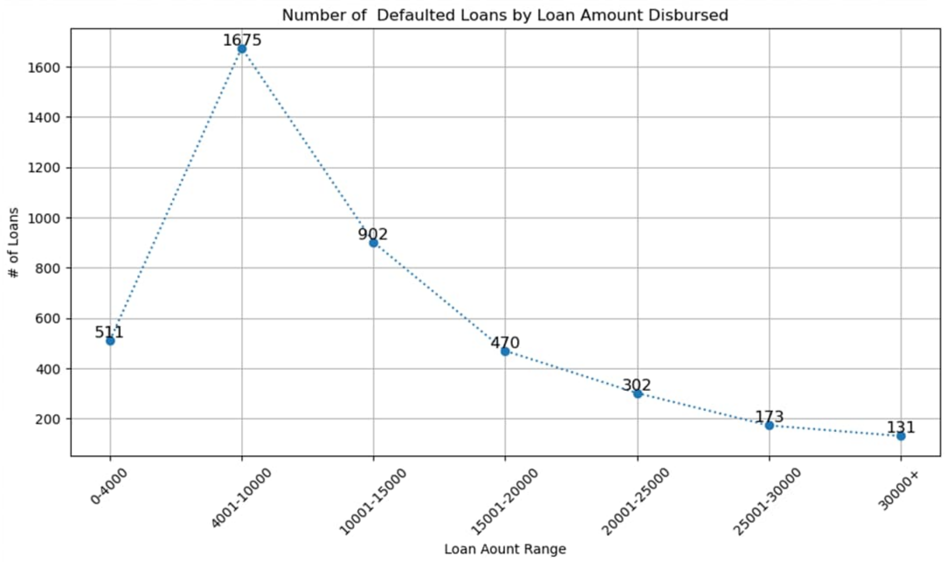
- Grade can be leveraged as a factor to reduce the overall defaulted loans
- Analysis confirms the default rates are higher in lower grades compared to others
- Based on the data provided the default rates as per the analysis are as below
  - A—3%
  - B—9%
  - C—13%
  - D—17%
  - E—17%
  - F—21%
  - G—26%
- Total loans given to customers with Grade "A","B","C" is ~30.2k out of which 2.5 K were defaulted i.e.~8% of loans
- Total loans given to customers with Grade "D","E","F","G" is ~9.5k out of which 1.8 K were defaulted i.e.~18% of loans

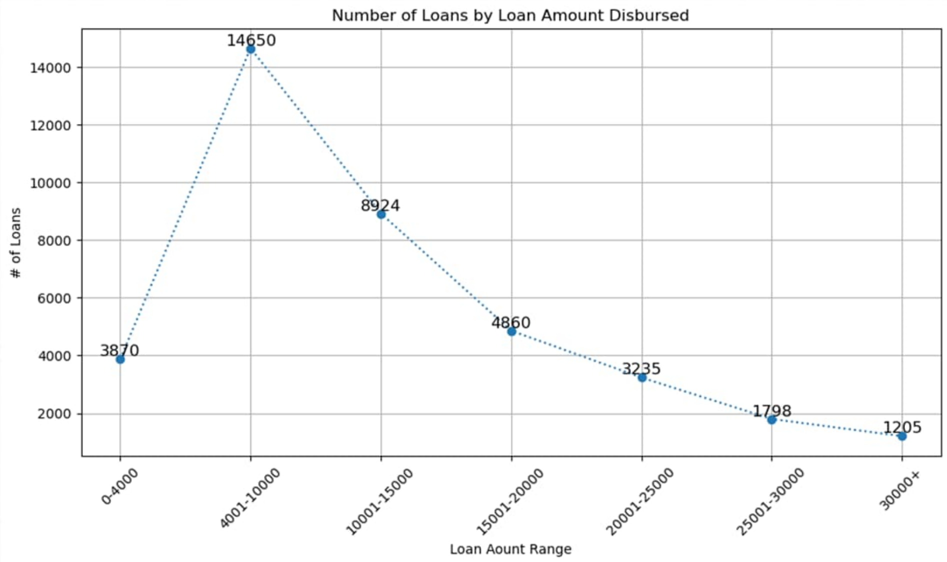


Higher loan amounts have less default rates

- Loans less than 15k have higher default rates compared to loans more than 15k
- So additional rules are required to risk score loans less than 15k

Loans Range	Total Loans	Defaulted Loans	% of default
0-4000	3,870	511	13.20%
4001-10000	14,650	1,675	11.43%
10001-15000	8,924	902	10.11%
15001-20000	4,860	470	9.67%
20001-25000	3,235	302	9.34%
25001-30000	1,798	173	9.62%
30001+	1,205	131	10.87%





Customers having more than 10 years of employment length have higher default rates

- This can be a feature considered to build the model to reduce the defaults