upGrad

Executive PG Programme in Machine Learning & AI - December2021

Lending Club Case Study

Ankit Kumar

Business Objective : To assess the risk associated with lending through the identification of risky applicants. This is achieved by leveraging the data available during loan application and ascertaining driving factors for loan default.

Business Constraint : Apart from the readily available data during loan application, there are other key parameters which determines if an applicant will default or not. An objective assessment with caution should taken while considering driving factor for loan default.

Null Value Analysis

feature_name	feature_type	null_count	null_percentage
verification_status_joint	float64	39717	100.00
annual_inc_joint	float64	39717	100.00
mo_sin_old_rev_tl_op	float64	39717	100.00
mo_sin_old_il_acct	float64	39717	100.00
bc_util	float64	39717	100.00
bc_open_to_buy	float64	39717	100.00
avg_cur_bal	float64	39717	100.00
acc_open_past_24mths	float64	39717	100.00
inq_last_12m	float64	39717	100.00
total_cu_tl	float64	39717	100.00
inq_fi	float64	39717	100.00
total_rev_hi_lim	float64	39717	100.00
all_util	float64	39717	100.00
max_bal_bc	float64	39717	100.00
open_rv_24m	float64	39717	100.00
open_rv_12m	float64	39717	100.00
il_util	float64	39717	100.00
total_bal_il	float64	39717	100.00
mths_since_rcnt_il	float64	39717	100.00
open_il_24m	float64	39717	100.00
open_il_12m	float64	39717	100.00
open_il_6m	float64	39717	100.00
open_acc_6m	float64	39717	100.00
tot_cur_bal	float64	39717	100.00
tot_coll_amt	float64	39717	100.00
mo_sin_rcnt_rev_tl_op	float64	39717	100.00
mo_sin_rcnt_tl	float64	39717	100.00
mort_acc	float64	39717	100.00
num_rev_tl_bal_gt_0	float64	39717	100.00
total_bc_limit	float64	39717	100.00
total_bal_ex_mort	float64	39717	100.00
tot_hi_cred_lim	float64	39717	100.00
percent_bc_gt_75	float64	39717	100.00
pct_tl_nvr_dlq	float64	39717	100.00
num_tl_op_past_12m	float64	39717	100.00
num_tl_90g_dpd_24m	float64	39717	100.00
num_tl_30dpd	float64	39717	100.00
num_tl_120dpd_2m	float64	39717	100.00
num_sats	float64	39717	100.00
num_rev_accts	float64	39717	100.00
mths_since_recent_bc	float64	39717	100.00
num_op_rev_tl	float64	39717	100.00
num_il_tl	float64	39717	100.00
num_bc_tl	float64	39717	100.00
num_bc_sats	float64	39717	100.00
num_actv_rev_tl	float64	39717	100.00
num_actv_bc_tl	float64	39717	100.00
num_accts_ever_120_pd	float64	39717	100.00
mths_since_recent_revol_delinq	float64	39717	100.00
mths_since_recent_inq	float64	39717	100.00
mths_since_recent_bc_dlq	float64	39717	100.00
dti_joint	float64	39717	100.00
total_il_high_credit_limit	float64	39717	100.00
mths_since_last_major_derog	float64	39717	100.00
next_pymnt_d	object	38577	97.13
mths_since_last_record	float64	36931	92.99
mths_since_last_delinq	float64	25682	64.66
desc	object	12940	32.58
emp_title	object	2459	6.19
emp_length	object	1075	2.71
pub_rec_bankruptcies	float64	697	1.75
last_pymnt_d	object	71	0.18
collections_12_mths_ex_med	float64	56	0.14
chargeoff_within_12_mths	float64	56	0.14
revol_util	object	50	0.13
tax_liens	float64	39	0.10
title	object	11	0.03
last_credit_pull_d	object	2	0.01

All the feature which had 100% of their values as null were dropped. Since they will not contribute any insight to the analysis.

Out of the remaining features with null values, features were dropped at the moment:
next_pymnt_d : more than 90% data is missing
mths_since_last_record: more than 90% data is missing
desc : purpose feature provides similar information
emp_title : Doesn't enrich our dataset
last_pymnt_d : customer behavior variable
collections_12_mths_ex_med : Zero variance, i.e. all the observations are same
chargeoff_within_12_mths : Zero variance, i.e. all the observations are same
tax_liens : Zero variance, i.e. all the observations are same
title : purpose feature provides similar information
last_credit_pull_d : customer behavior variable

For remaining features with null values, feature creation or missing value imputation was performed:

mths_since_last_delinq : create a column named delinquency_history with null values of mths_since_last_delinq as "0" in delinquency_history, and remaining as "1".
emp_length : mode imputation was performed and the existing categories were used.
pub_rec_bankruptcies : mode imputation was performed and the existing categories were used.
revol_util : mode imputation then object type converted to float.

Outlier Analysis

upGrad

	Feature	IQR	Lower_Limit	Upper_Limit	Outlier_Count	Outlier_Percent
0	loan_amnt	9700.000000	-9250.0000	29550.000000	1088	2.82
1	funded_amnt	9800.000000	-9500.0000	29700.000000	920	2.38
2	funded_amnt_inv	9000.000000	-8500.0000	27500.000000	1034	2.68
3	installment	259.810000	-223.9750	815.265000	1373	3.56
4	annual_inc	42000.000000	-23000.0000	145000.000000	1762	4.57
5	dti	10.430000	-7.5150	34.205000	0	0.00
6	delinq_2yrs	0.000000	0.0000	0.000000	4191	10.86
7	inq_last_6mths	1.000000	-1.5000	2.500000	3554	9.21
8	open_acc	6.000000	-3.0000	21.000000	495	1.28
9	pub_rec	0.000000	0.0000	0.000000	2070	5.37
10	revol_bal	13262.000000	-16243.0000	36805.000000	2423	6.28
11	revol_util	47.000000	-45.5000	142.500000	0	0.00
12	total_acc	16.000000	-11.0000	53.000000	513	1.33
13	out_prncp	0.000000	0.0000	0.000000	0	0.00
14	out_prncp_inv	0.000000	0.0000	0.000000	0	0.00
15	total_pymnt	10623.455092	-10421.6853	32072.135068	1276	3.31
16	total_pymnt_inv	10310.190000	-10445.6950	30795.065000	1450	3.76
17	total_rec_prncp	8675.000000	-8512.5000	26187.500000	943	2.44
18	total_rec_int	2027.990000	-2397.9950	5713.965000	2901	7.52
19	total_rec_late_fee	0.000000	0.0000	0.000000	1995	5.17
20	recoveries	0.000000	0.0000	0.000000	4218	10.93
21	collection_recovery_fee	0.000000	0.0000	0.000000	3782	9.80
22	last_pymnt_amnt	3229.830000	-4627.3250	8291.995000	3990	10.34
23	policy_code	0.000000	1.0000	1.000000	0	0.00
24	acc_now_delinq	0.000000	0.0000	0.000000	0	0.00
25	delinq_amnt	0.000000	0.0000	0.000000	0	0.00
26	pub_rec_bankruptcies	0.000000	0.0000	0.000000	1642	4.26
27	delinquency_history	1.000000	-1.5000	2.500000	0	0.00

Winsorization Of Numerical Columns

	Feature	IQR	Lower_Limit	Upper_Limit	Outlier_Count	Outlier_Percent
0	loan_amnt	9700.000000	-9250.0000	29550.000000	0	0.0
1	funded_amnt	9800.000000	-9500.0000	29700.000000	0	0.0
2	funded_amnt_inv	9000.000000	-8500.0000	27500.000000	0	0.0
3	installment	259.810000	-223.9750	815.265000	0	0.0
4	annual_inc	42000.000000	-23000.0000	145000.000000	0	0.0
5	dti	10.430000	-7.5150	34.205000	0	0.0
6	delinq_2yrs	0.000000	0.0000	0.000000	0	0.0
7	inq_last_6mths	1.000000	-1.5000	2.500000	0	0.0
8	open_acc	6.000000	-3.0000	21.000000	0	0.0
9	pub_rec	0.000000	0.0000	0.000000	0	0.0
10	revol_bal	13262.000000	-16243.0000	36805.000000	0	0.0
11	revol_util	47.000000	-45.5000	142.500000	0	0.0
12	total_acc	16.000000	-11.0000	53.000000	0	0.0
13	out_prncp	0.000000	0.0000	0.000000	0	0.0
14	out_prncp_inv	0.000000	0.0000	0.000000	0	0.0
15	total_pymnt	10623.455092	-10421.6853	32072.135068	0	0.0
16	total_pymnt_inv	10310.190000	-10445.6950	30795.065000	0	0.0
17	total_rec_prncp	8675.000000	-8512.5000	26187.500000	0	0.0
18	total_rec_int	2027.990000	-2397.9950	5713.965000	0	0.0
19	total_rec_late_fee	0.000000	0.0000	0.000000	0	0.0
20	recoveries	0.000000	0.0000	0.000000	0	0.0
21	collection_recovery_fee	0.000000	0.0000	0.000000	0	0.0
22	last_pymnt_amnt	3229.830000	-4627.3250	8291.995000	0	0.0
23	policy_code	0.000000	1.0000	1.000000	0	0.0
24	acc_now_delinq	0.000000	0.0000	0.000000	0	0.0
25	delinq_amnt	0.000000	0.0000	0.000000	0	0.0
26	pub_rec_bankruptcies	0.000000	0.0000	0.000000	0	0.0
27	delinquency_history	1.000000	-1.5000	2.500000	0	0.0

Further Pre-Processing

upGrad

```
0.0    38577
Name: delinq_2yrs, dtype: int64
-------------------------
0.0    38577
Name: pub_rec, dtype: int64
-------------------------
0.0    38577
Name: out_prncp, dtype: int64
-------------------------
0.0    38577
Name: out_prncp_inv, dtype: int64
-------------------------
0.0    38577
Name: total_rec_late_fee, dtype: int64
-------------------------
0.0    38577
Name: recoveries, dtype: int64
-------------------------
0.0    38577
Name: collection_recovery_fee, dtype: int64
-------------------------
1    38577
Name: policy_code, dtype: int64
-------------------------
0    38577
Name: acc_now_delinq, dtype: int64
-------------------------
0    38577
Name: delinq_amnt, dtype: int64
-------------------------
0.0    38577
Name: pub_rec_bankruptcies, dtype: int64
-------------------------
```

Since 'delinq_2yrs', 'pub_rec', 'out_prncp', 'out_prncp_inv', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee', 'policy_code', 'acc_now_delinq', 'delinq_amnt', 'pub_rec_bankruptcies' have same values in all the rows. We will drop thee coumns.

Customer behaviour variables
delinq_2yrs
earliest_cr_line
inq_last_6mths
open_acc
pub_rec
revol_bal
revol_util
total_acc
out_prncp
out_prncp_inv
total_pymnt
total_pymnt_inv
total_rec_prncp
total_rec_int
total_rec_late_fee
recoveries
collection_recovery_fee
last_pymnt_d
last_pymnt_amnt
last_credit_pull_d
application_type

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

There are different categorical columns that needs either a type conversion or are not important for our analysis and will be dropped.

Columns needing type casting:

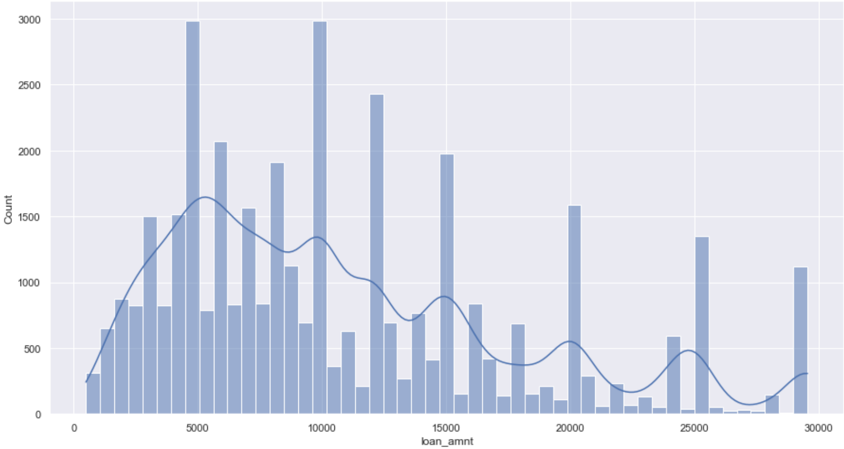
1. int_rate

Columns that needs to be dropped:

1. application_type (Reason : zero variance)
2. initial_list_status (Reason : zero variance)
3. earliest_cr_line (Reason : not relevant to credit decision)
4. zip_code (Reason : not relevant to credit decision)
5. url (Reason : not relevant to credit decision)
6. pymnt_plan (Reason : zero variance)
7. issue_d (Reason : not relevant to credit decision)
8. last_pymnt_amnt (Reason : not relevant to credit decision)

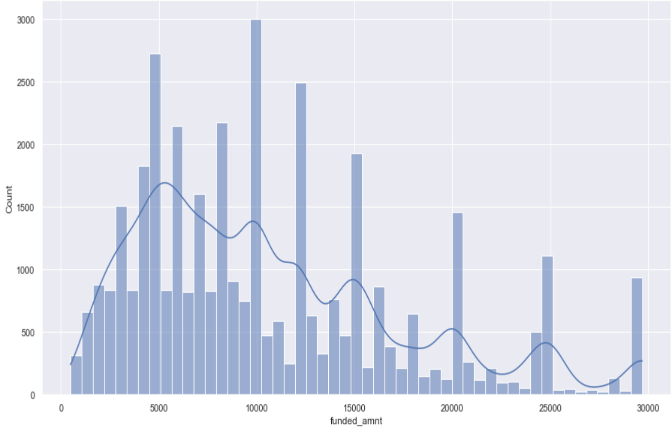
Loan Amount

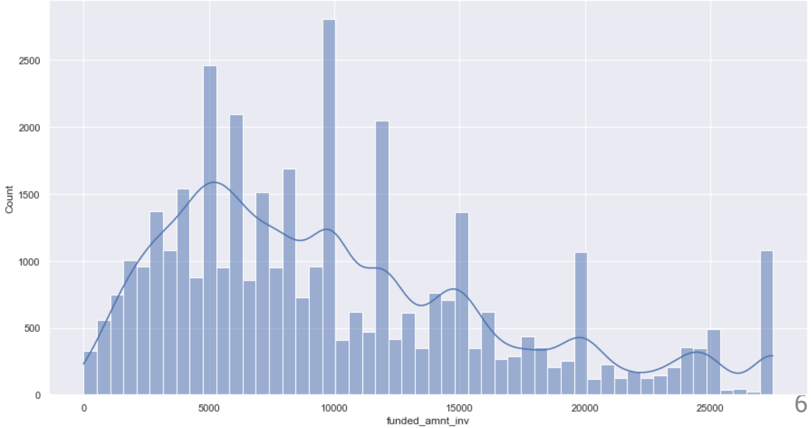
upGrad



The loan amount requested, funded and funded invested, all have their distribution spread around 8000.

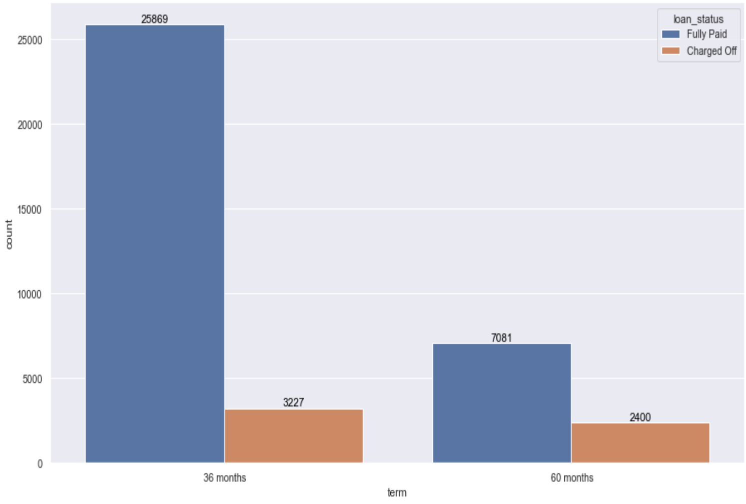
Most of the loan amount is <15000.

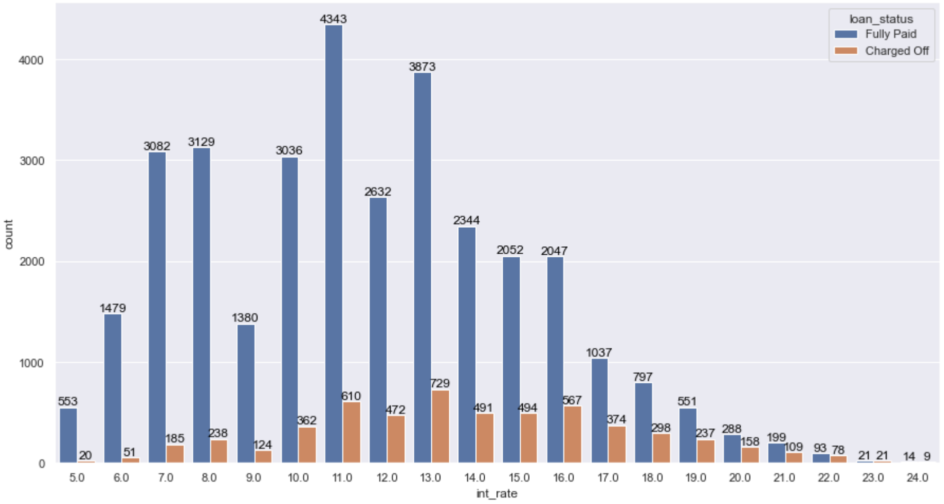




Loan Term , Interest Rate

upGrad



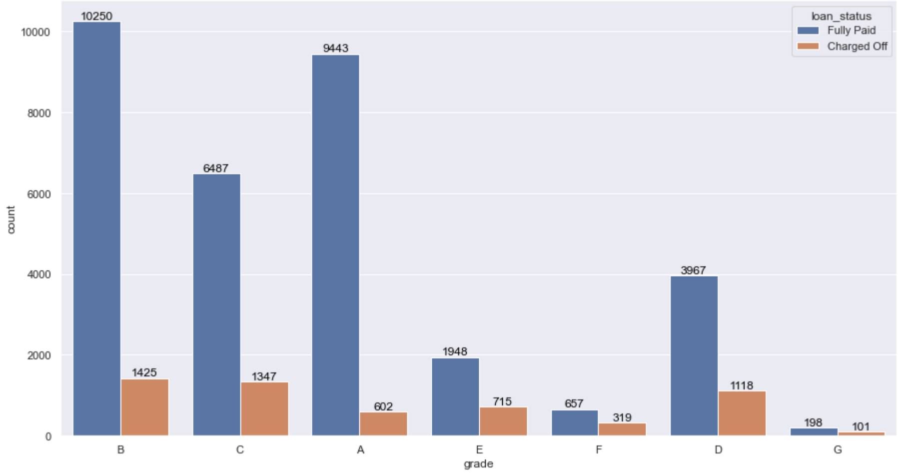


In terms of proportionality, loans whose terms were of 60 months are more likely to default.

Customer who are granted loans at higher rate of interest are more likely to default.

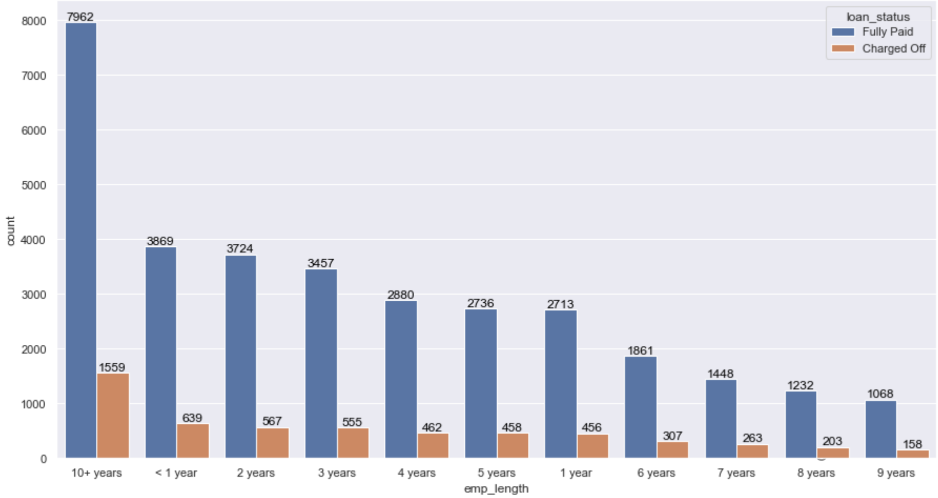
Grade , Employment Length

upGrad



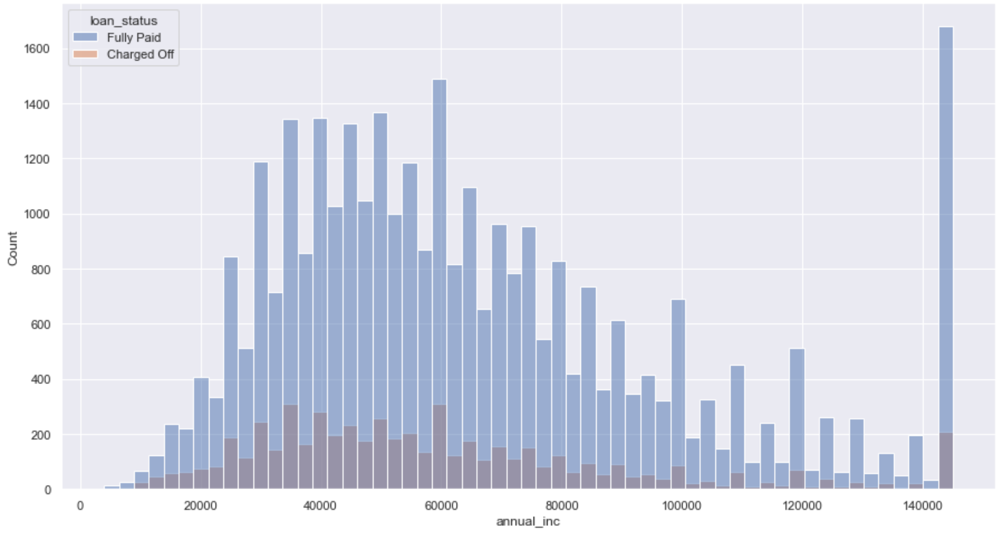
Grade D, E, F & G are more susceptible to defaulting.

There's a decreasing trend in loan default as the number of employment year increases .

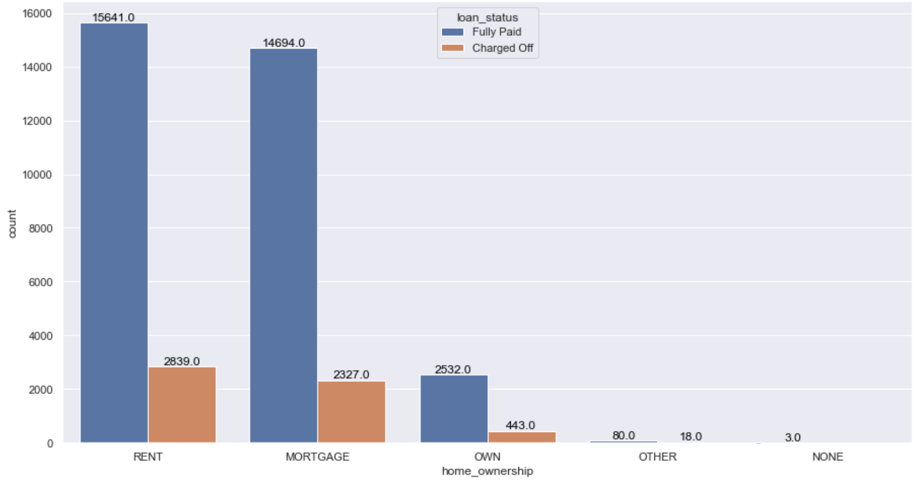


Annual Income, Home Ownership

upGrad



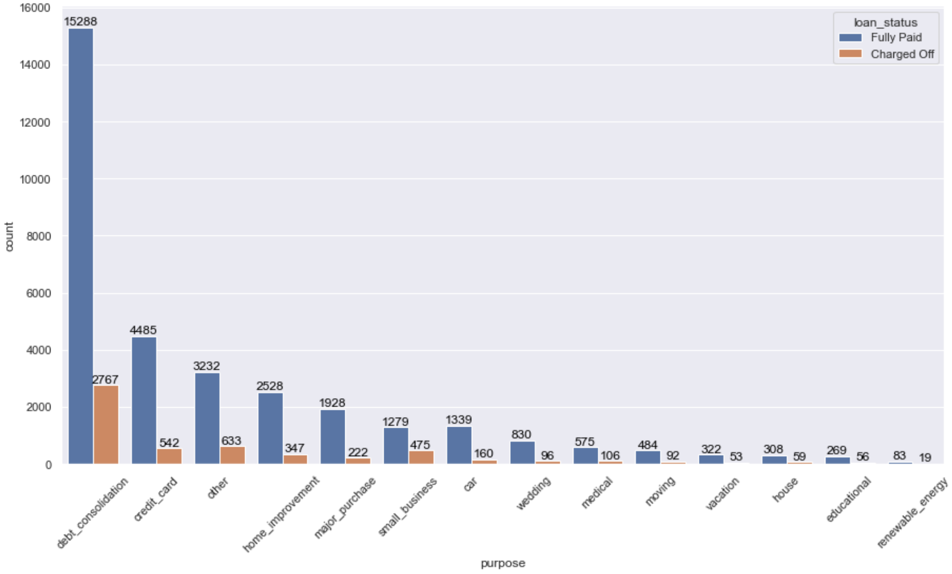
Most of the loan defaults are for customers who have annual income less than 10,000.



Customers living in rented place or who have mortgage to pay are more prone to defaulting.

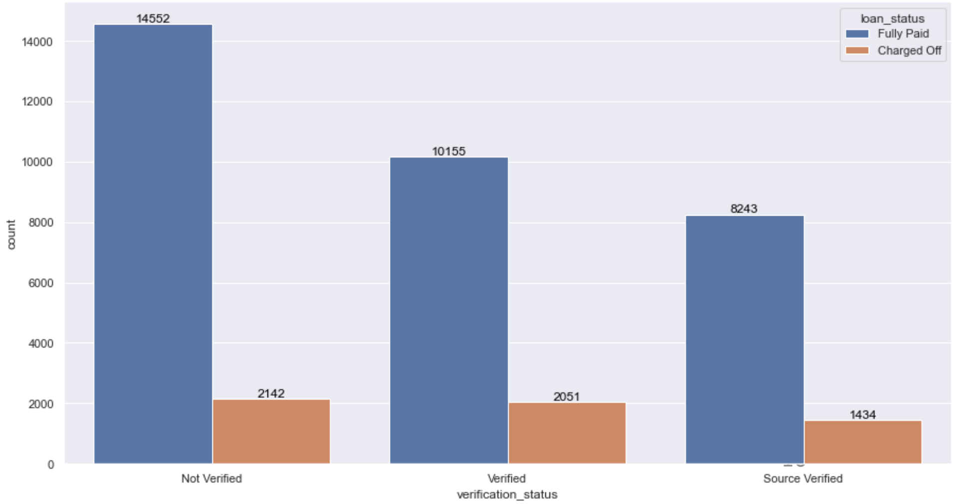
Loan Purpose, Verification Status

upGrad



Debt consolidation one of the leading cause for defaulting.

Comparatively, source verified loans are more less likely to be in the default list.



Summary

upGrad

- *The loan amount requested, funded and funded invested, all have their distribution spread around 8000.*
- *In terms of proportionality, loans whose terms were of 60 months are more likely to default.*
- *Customer who are granted loans at higher rate of interest are more likely to default.*
- *Grade D, E, F & G are more susceptible to defaulting.*
- *There's a decreasing trend in loan default as the number of employment year increases .*
- *Most of the loan defaults are for customers who have annual income less than 10,000.*
- *Customers living in rented place or who have mortgage to pay are more prone to defaulting.*
- *Debt consolidation one of the leading cause for defaulting.*
- *Comparatively, source verified loans are more less likely to be in the default list.*