LOAN APPROVAL AND REJECTION BASED ON HISTORIC BORROWER DATA

Problem Statement:

- Predict loan Approval and rejection basses on different features

Business Value:

- Help to identify credibility of borrowers
- Will borrower able to pay the entire loan with interest

DataSet information:

Data Source:

- https://www.lendingclub.com/info/download-data.action
- https://www.dataquest.io/blog/machine-learning-preparing-data

Features available in dataset

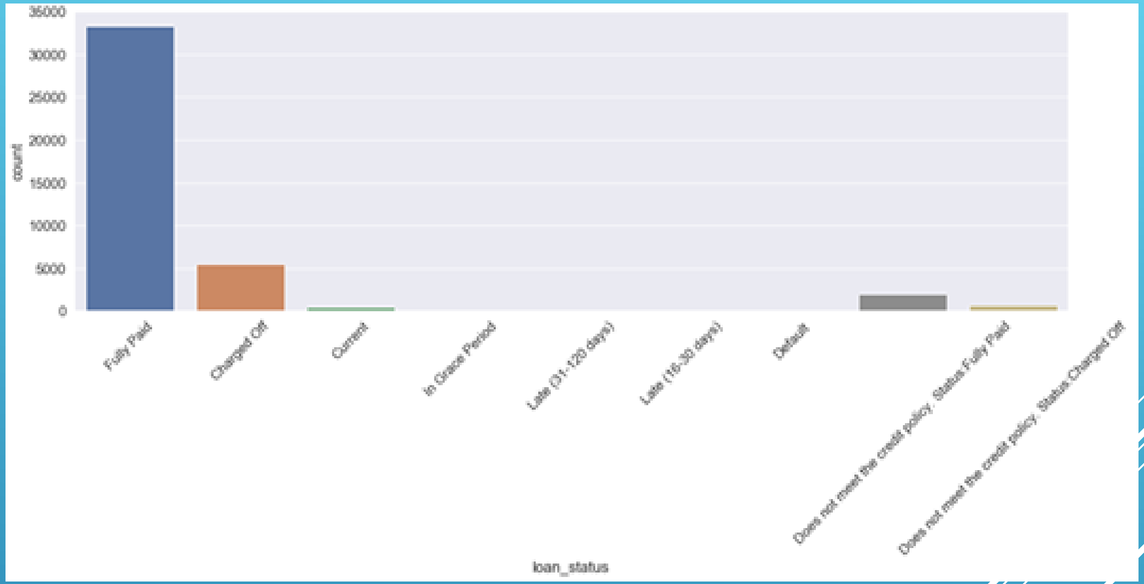
acc_now_delinq	emp_title	last_pymnt_d	num_actv_bc_tl	out_prncp	total_bal_ex_mort
acc_open_past_24mths	fico_range_high	loan_amnt	num_actv_rev_tl	out_prncp_inv	total_bal_il
addr_state	fico_range_low	loan_status	num_bc_sats	pct_tl_nvr_dlq	total_bc_limit
all_util	funded_amnt	max_bal_bc	num_bc_tl	percent_bc_gt_75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

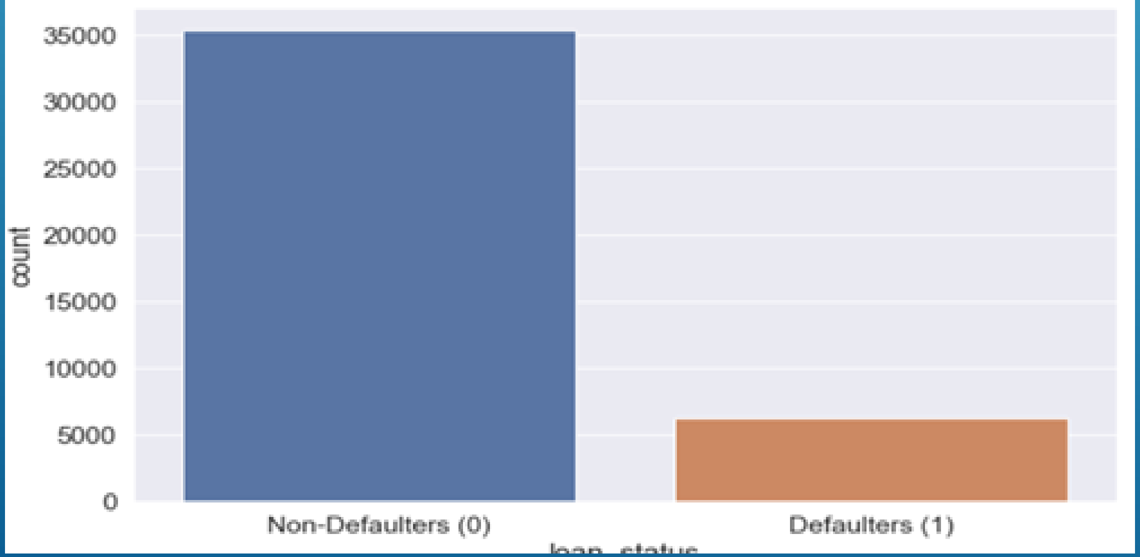
Data Wrangling

- Drop any columns with 50 % null row
- Handling missing values by dropping row if those columns have 3 or 4 missing row
- Remove columns (features) which not required in machine learning like ‘desc’,’url’ ‘zip   code’ etc.
- Select object columns and worked on data conversation required for machine learning like ‘emp_length’,’ grade’ etc.
- Converted target categorical value in binary for machine learning ‘loan_status’
- Finding outliers and removing from dataset because it can affect model prediction like annual income. So remove value above 99.5% of quartile

Converted target feature to binary for Machine learning

Loan Status

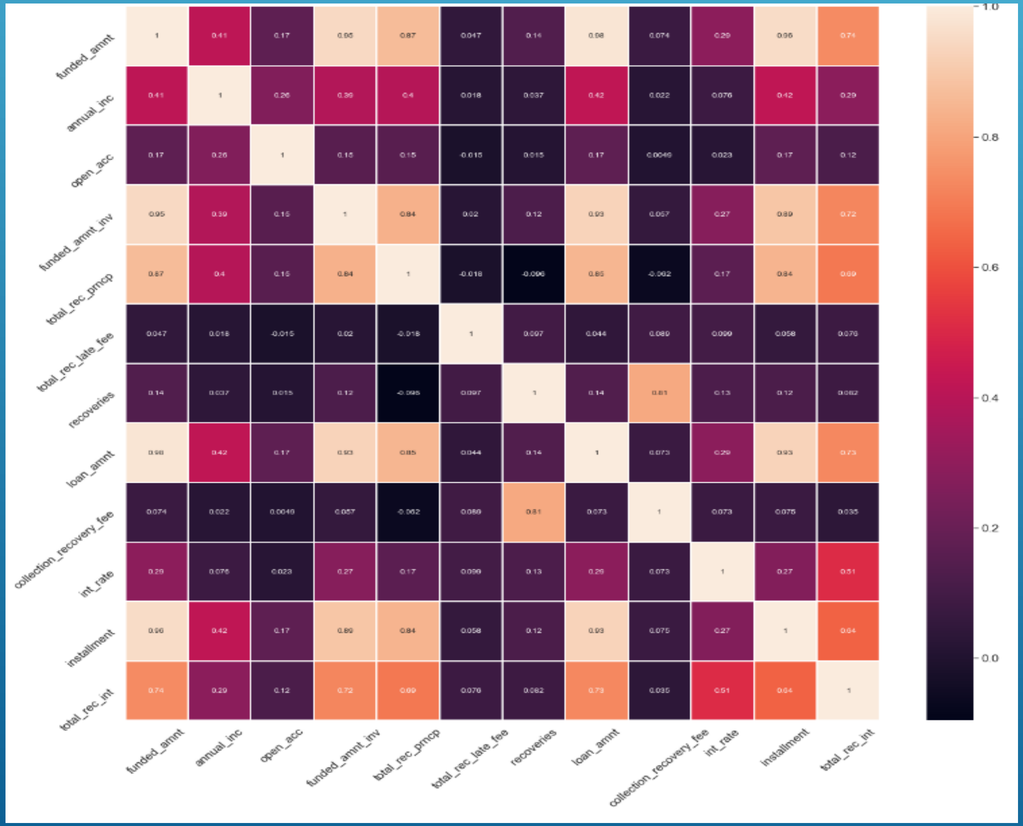


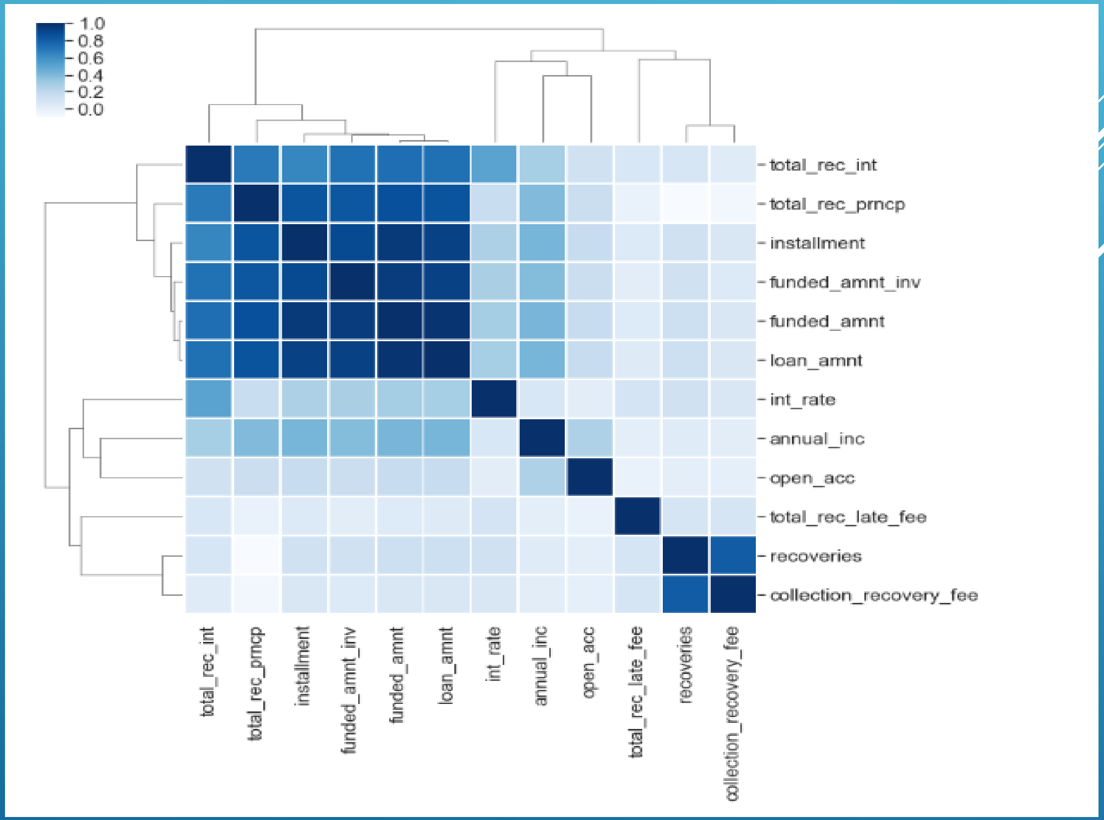


Loan Status after converting binary

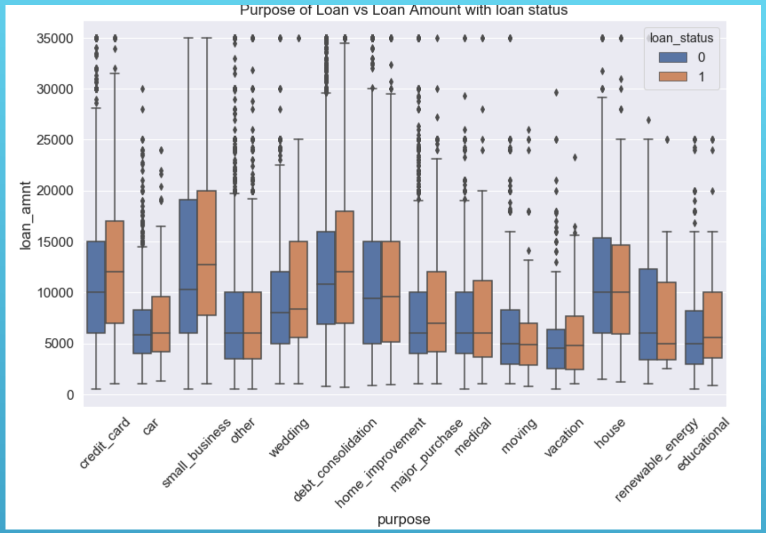
Exploratory Data analysis

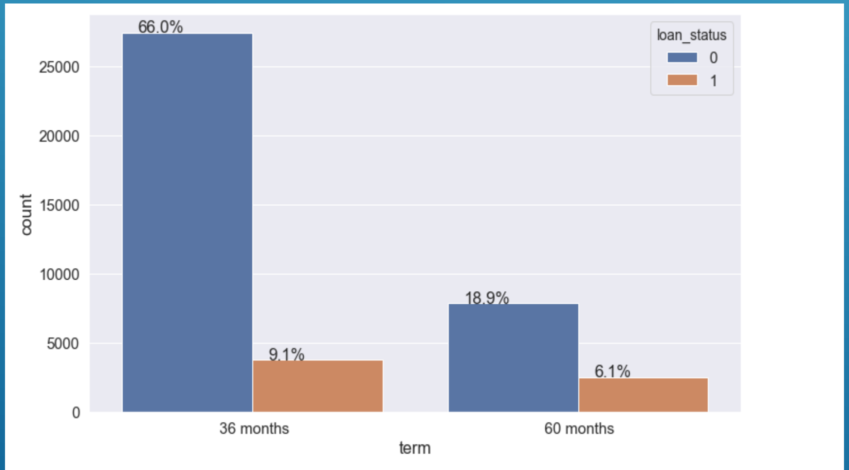
- ❖ Correlation heat map, pair plot and cluster map to identify correlation between different features with loan status





Wedding and Major Purchase are more towards Defaulter



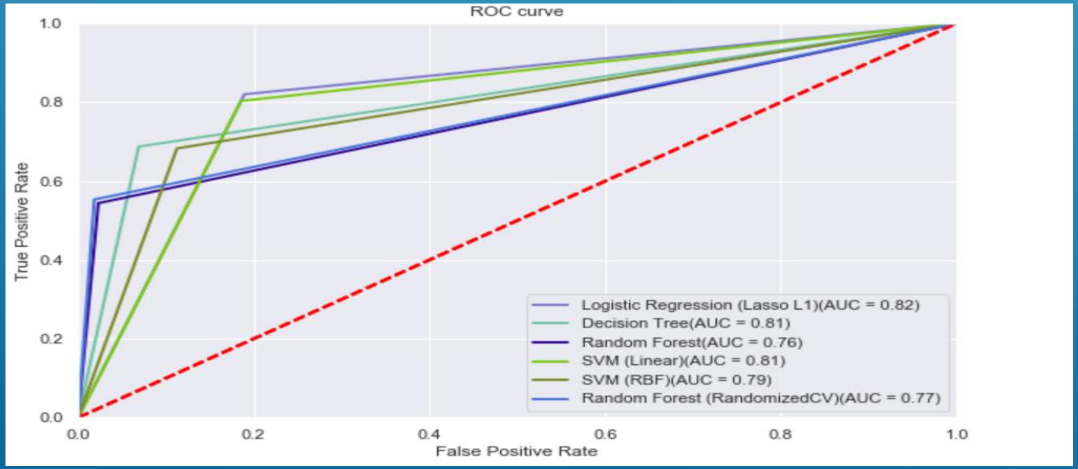


Defaulter percentage are more in long term payments

Machine Learning and observations

Score table from Machine leaning

	Model	Accuracy	Precision	Recall Score	F1 Score
0	Logistic Regression	0.812102	0.444398	0.820216	0.576464
1	Decision Tree Classifier	0.892939	0.647529	0.687500	0.666916
2	Random_Forest_n=100	0.909419	0.813149	0.543981	0.651872
3	SVM (Linear)	0.812703	0.444302	0.803241	0.572135
4	SVM (RBF)	0.855527	0.528358	0.682870	0.595759
5	Random_Forest_RandomCV	0.914712	0.846517	0.553241	0.669155



ROC Curve from Machine learning

Conclusion and Next Steps

- Random forest accuracy is around 90 % on test data.
- Test model with more data to check further accuracy and result
- Pick and choose more Features and try loan prediction for more accurate results