Model to Predict Loan Eligibility Status Through Machine Learning

Solution Approach Document

Link to Dataset: https://www.kaggle.com/shadabhussain/credit-risk-loan-eliginility

About the Dataset:

Dataset has 63,999 rows and 45 columns. The following are the attributes of the dataset:

#	Column	Non-Null Count	Dtype
0	member_id	63999 non-null	int64
1	loan_amnt	63999 non-null	int64
2	funded_amnt	63999 non-null	int64
3	funded_amnt_inv	63999 non-null	float64
4	term	63999 non-null	object
5	batch_enrolled	53735 non-null	object
6	int_rate	63999 non-null	float64
7	grade	63999 non-null	object
8	sub_grade	63999 non-null	object
9	emp_title	60173 non-null	object
10	emp_length	60675 non-null	object
11	home_ownership	63999 non-null	object
12	annual_inc	63999 non-null	float64
13	verification_status	63999 non-null	object
14	pymnt_plan	63999 non-null	object
15	desc	9150 non-null	object
16	purpose	63999 non-null	object
17	title	63986 non-null	object
18	zip_code	63999 non-null	object
19	addr_state	63999 non-null	object
20	dti	63999 non-null	float64
21	delinq_2yrs	63999 non-null	int64
22	inq_last_6mths	63999 non-null	int64
23	mths_since_last_delinq	31168 non-null	float64
24	mths_since_last_record	9650 non-null	float64
25	open_acc	63999 non-null	int64
26	pub_rec	63999 non-null	int64
27	revol_bal	63999 non-null	int64
28	revol_util	63970 non-null	float64
29	total_acc	63999 non-null	int64
30	initial_list_status	63999 non-null	object
31	total_rec_int	63999 non-null	float64
32	total_rec_late_fee	63999 non-null	float64
33	recoveries	63999 non-null	float64
34	collection_recovery_fee	63999 non-null	float64
35	collections_12_mths_ex_med	63991 non-null	float64
36	mths_since_last_major_derog	15844 non-null	float64
37	application_type	63999 non-null	object
38	verification_status_joint	28 non-null	object
39	last_week_pay	63999 non-null	object
40	acc_now_delinq	63999 non-null	int64
41	tot_coll_amt	58875 non-null	float64
42	tot_cur_bal	58875 non-null	float64
43	total_rev_hi_lim	58875 non-null	float64
44	loan_status	63999 non-null	int64

Note:

- After taking a look at the dataset and going through data preprocessing and exploratory data analysis, it was determined that not all of these columns would be significant in determining the loan eligibility status either due to insignificance or too many null values.
- All the code to all the analyses will be provided in the final solution rather than this solution approach document

Exploratory Data Analysis:

Here is a preview of the dataset after dropping columns:

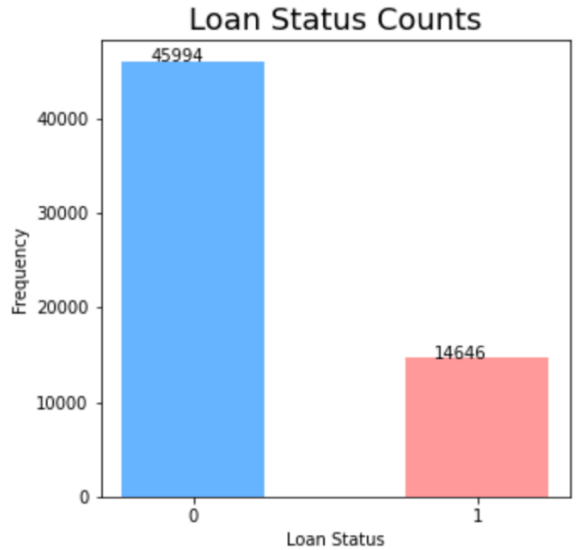
```
df.head()
```

	loan_amnt	term	int_rate	grade	emp_length	home_ownership	annual_inc	verification_status	pymnt_plan	purpose	...	pub_rec	revol_bal	revol_util	total_acc	initial_list_status
0	14350	36 months	19.19	5	9	OWN	28700.0	Source Verified	n	debt_consolidation	...	1	22515	73.1	28	f
1	4800	36 months	10.99	2	0	MORTGAGE	65000.0	Source Verified	n	home_improvement	...	0	7624	23.2	13	w
2	10000	36 months	7.26	1	2	OWN	45000.0	Not Verified	n	debt_consolidation	...	0	10877	31.2	19	w
3	15000	36 months	19.72	4	10	RENT	105000.0	Not Verified	n	debt_consolidation	...	0	13712	55.5	21	f
4	16000	36 months	10.64	2	10	RENT	52000.0	Verified	n	credit_card	...	0	35835	76.2	27	w

collections_12_mths_ex_med	application_type	last_week_pay	acc_now_delinq	loan_status
0.0	INDIVIDUAL	26th week	0	0
0.0	INDIVIDUAL	9th week	0	0
0.0	INDIVIDUAL	9th week	0	0
0.0	INDIVIDUAL	135th week	0	0
0.0	INDIVIDUAL	96th week	0	0

Bar Plot of Loan Status Counts:

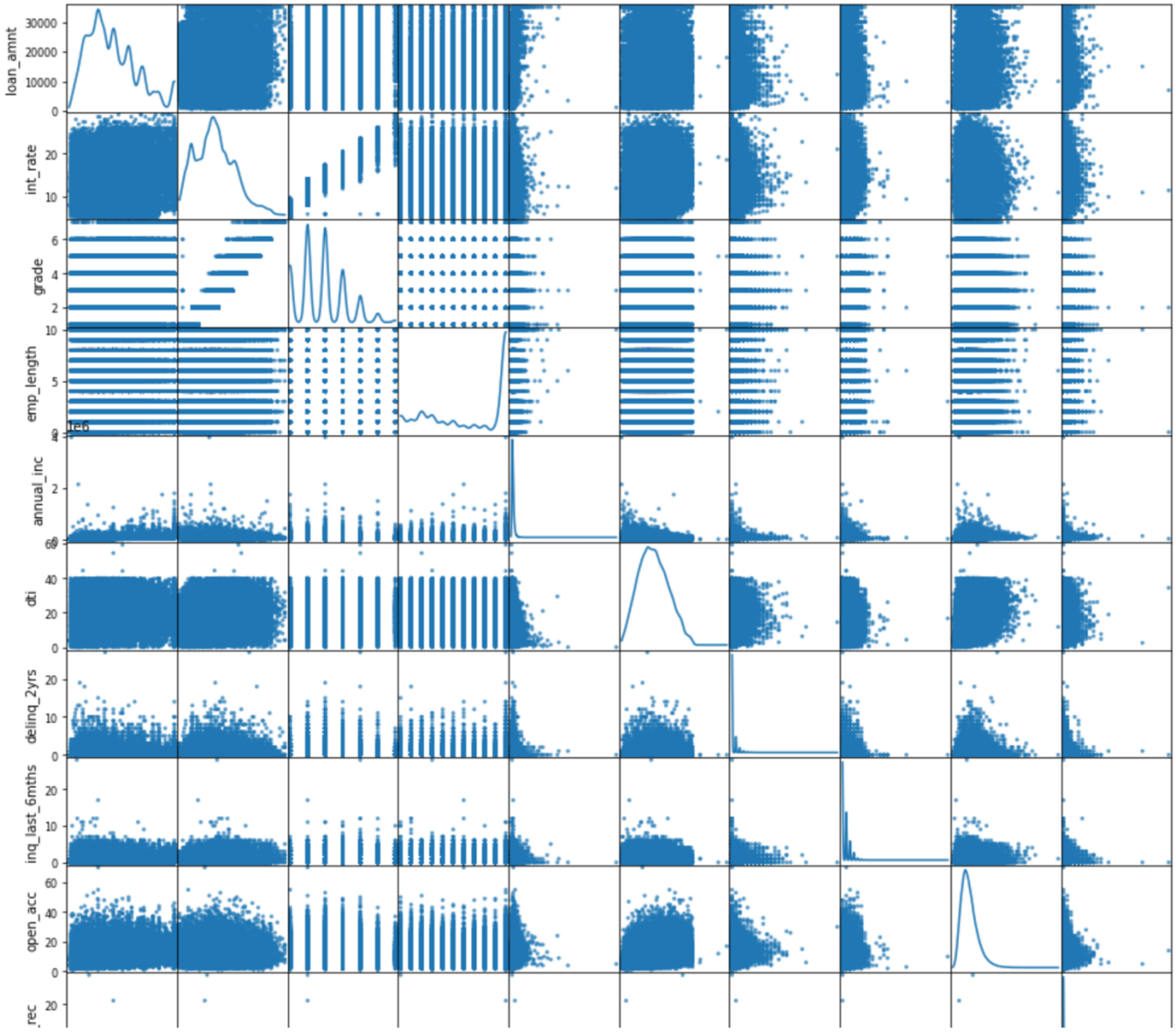
This bar plot is charted in terms of 1s and 0s, just as the dataset, and serves to provide an idea of how many loans were considered to be approved or not based on the conditions provided.



As seen in the bar chart, most of the Loans received a loan status of ‘0’, specifically 76.25%, is more dominant in that data column.

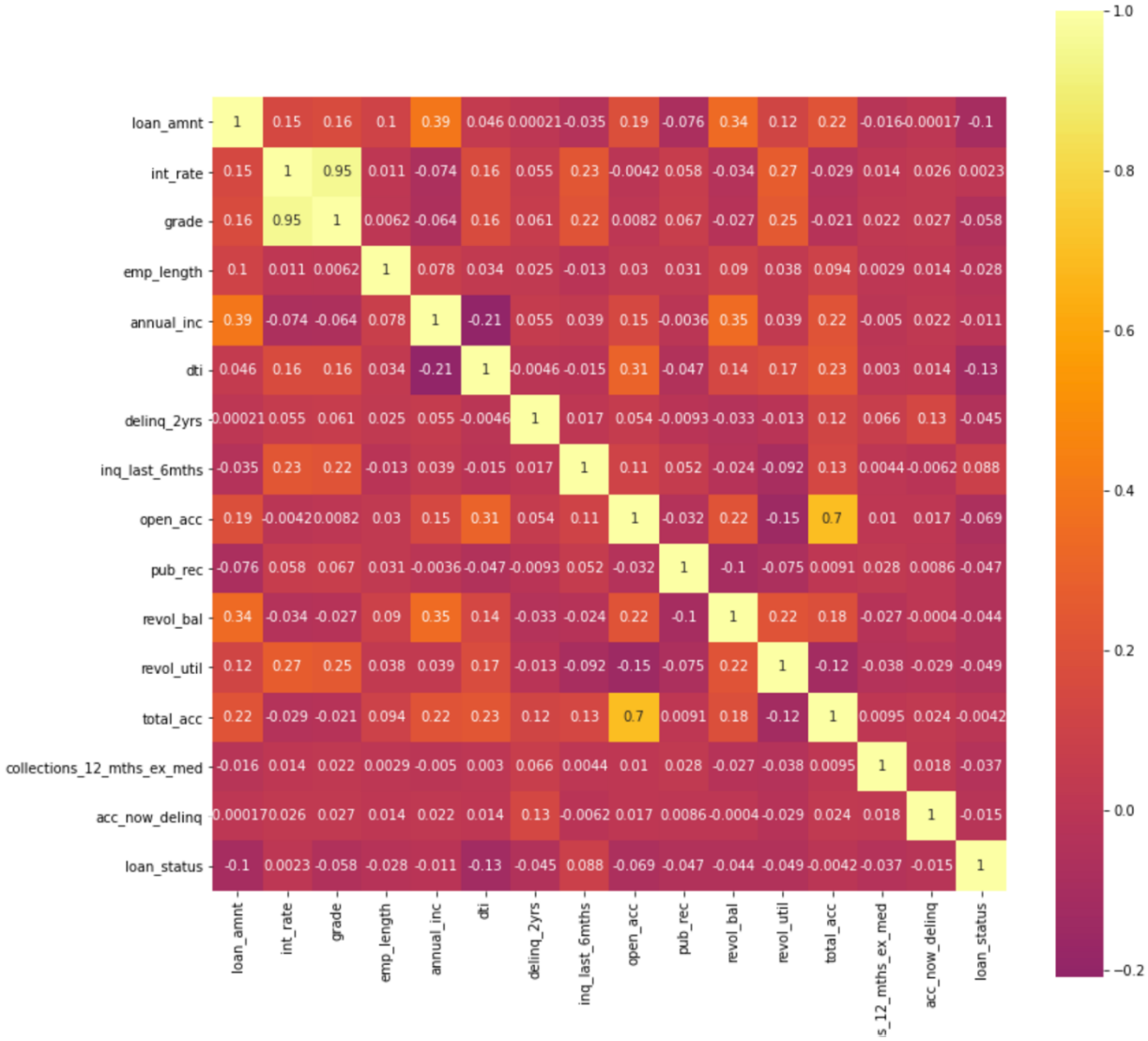
Pairplot:

The following shows the pairplot of this dataset which allows us to see various correlations between different variables in a compact manner.



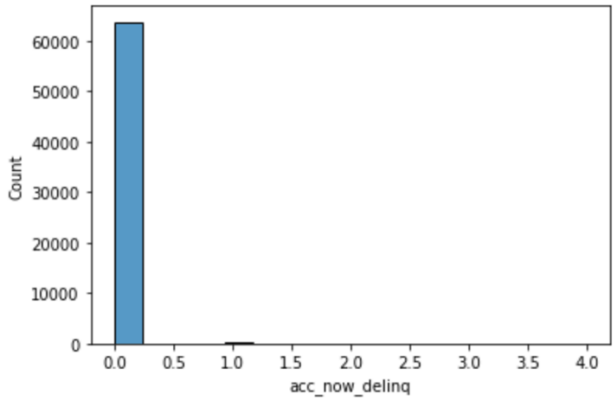
Correlation Heatmap:

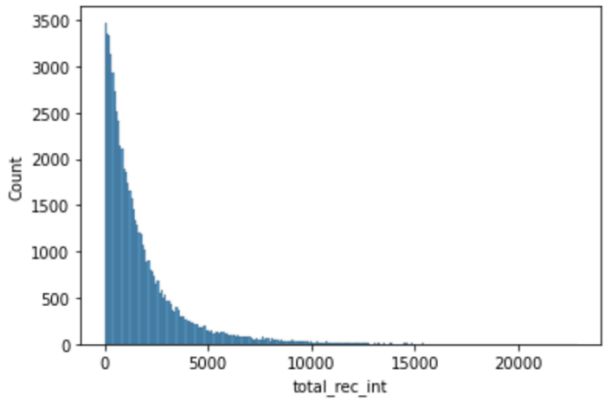
This heatmap displays the correlation between different features of the dataset in order to help us understand what the dataset and its attributes are like.

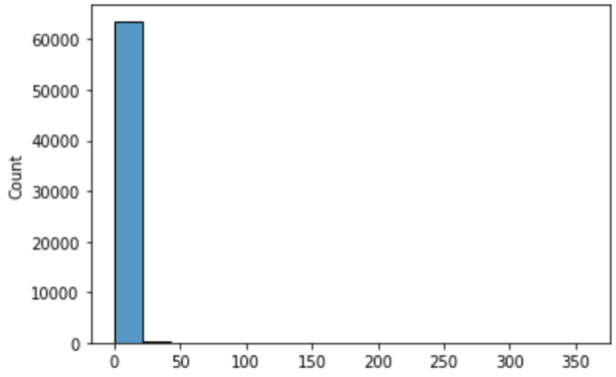


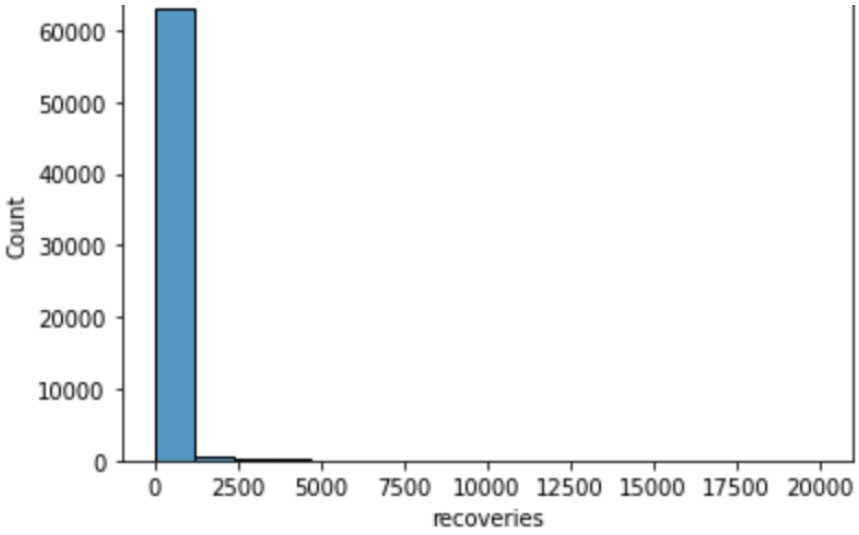
Data Values Check:

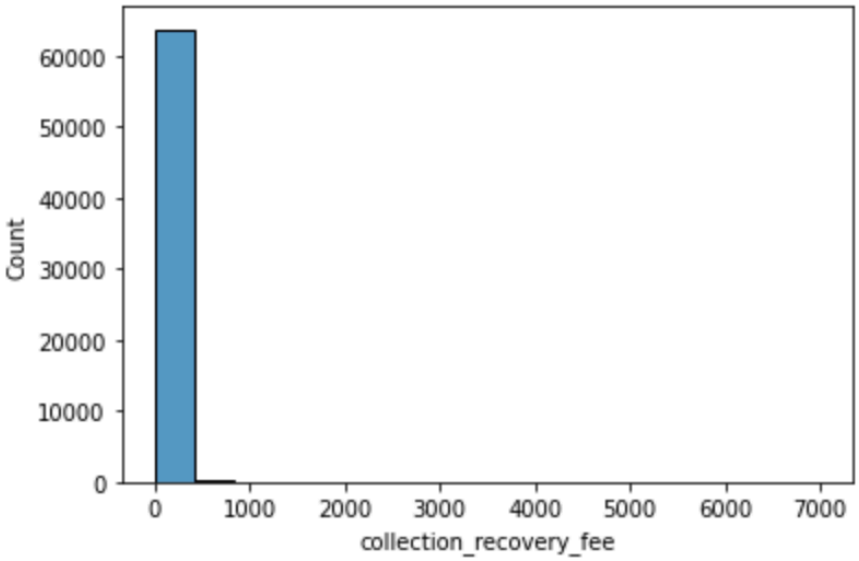
By taking a look at the following histograms, we can get an idea of the distribution of the values in each data column and see how biased the data is. After looking at all these plots, it can be seen that all of them are biased and the data is lobsided. Hence, these all of these were dropped and grouped as not being useful in building the model.

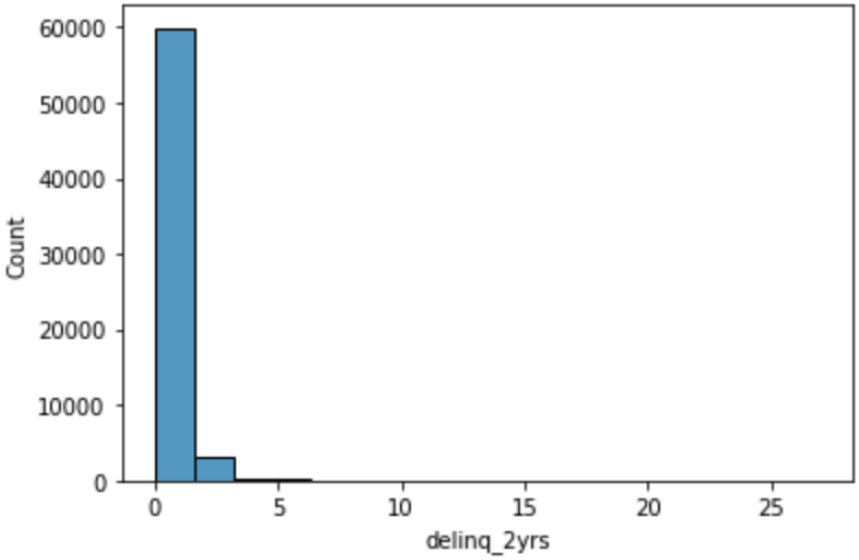


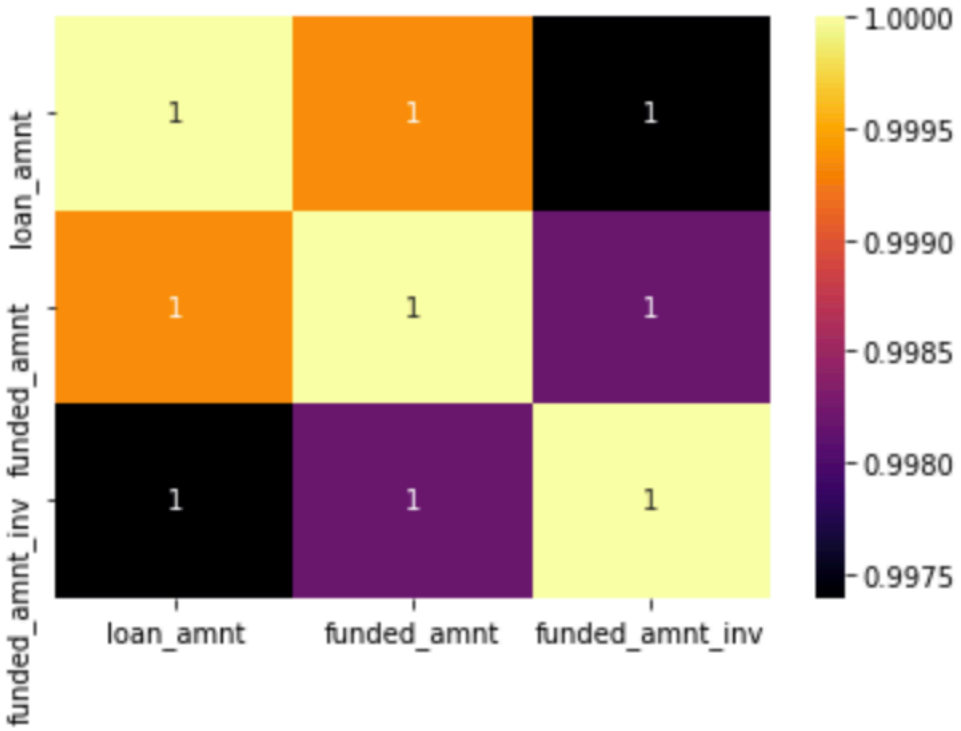





60000
50000
40000
30000
20000
10000
0
Count
0
2500
5000
7500
10000
12500
15000
17500
20000
recoveries


60000
50000
40000
30000
20000
10000
0
Count
0
1000
2000
3000
4000
5000
6000
7000
collection_recovery_fee


60000
50000
40000
30000
20000
10000
0
Count
0
5
10
15
20
25
delinq_2yrs



The above heatmap displays the correlation between the three columns “loan_amnt”, “funded_amnt”, and “funded_amnt_inv.” Since the correlation between these three columns is 1 in all of the squares in the heatmap, two of these three columns can be dropped.

Building the model:

After EDA and Data preprocessing on the dataset, the edited dataset would be exported from jupyter notebooks and then inputted into Vertex AI which would be used to build a model to predict the “loan status” or the loan eligibility based on certain features. After the model is built and validated, it can be exported from Vertex AI and deployed on the internet for usage.