



*Amity Institute of Integrative Sciences & Health (Manesar)*

# PROJECT REPORT

## Predicting Loan Using Logistic Regression

Postgraduate Diploma in Data Science

Year 2020/2021

WALID FARAH DJAMA

Enrollment No : A52513450006

# **Summary**

## **1.Executive Summary**

## **2.Introduction**

## **3.Preparing and Cleaning the Data**

## **4.Exploring and Transforming the Data**

## **5.The Logistic Model**

## **6.Optimizing the Threshold for Accuracy**

## **7.Optimizing the Threshold for Profit**

## **8.Results Summary**

## **9.Conclusion**

## **Executive Summary**

### **Project objective**

The main objective of the project, is to set out a good predictive Model, from the collected existing information on quality of the loan and improving the profit level with status of the loans, to set out the various variables of these standards and benchmarks to qualify for a good loans, and to extract from it a lists of minimum variables to predict a quality standards loans consideration for approval. The result is intended to improve a better revenue from the loans and be able to make profits while reducing the default loans in the process. The various elements of the project are summarised as follows:

- **Preparing and Cleaning the Data**
- **Exploring and Transforming the Data**
- **The Logistic Model**
- **Optimizing the Model for Accuracy**
- **Optimizing the Threshold for Profit**

## **Introduction**

This project is about predicting loan and reducing defaults while improving profits using statistical analysis, and the evaluation of probable logistic regression model. I have dataset 50000 loans which is composed of 32 variables. And I have used R programming Languages for the analysis.

In my analysis have used several comprehensive aspect of generalized linear model in order to optimize and predict the loans status as well as predict profit that based on my model. I have analysed it in terms of correct prediction percent of fully paid and default loan's status and from there. I was able to predict a higher percentage of the profit and reduced the defaulted loans.

## Preparing and Cleaning the Data

I have loaded all the library required for my analysis at first the loaded the dataset into the system.

Right after that I have started preparing/cleaning the data. I replaced original “n/a”, missing data, with the default value of NA for the purposes of my analysis. I usually impute missing data if its more than 5% but, in this case, the missing data is less than 5% so I had to drop some unnecessary variables from dataset as it will not have of any effect on my analysis. Our final dataset contains 28 variables instead of 32 variables.

### Data Descriptions:

<https://datascienceuwl.github.io/Project2018/TheData.html>

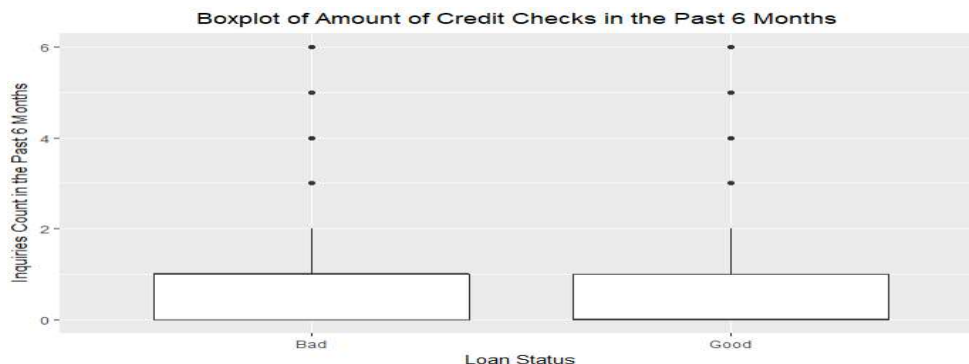
```
19 - ## 1. Preparing and Cleaning the Data----
20 Loans = read.csv("data.csv")
21 view(Loans)
22 Loans[Loans == "NA"] = NA
23 sum(is.na(Loans)) #sum of the na value
24 colMeans(is.na(Loans))
25 rowMeans(is.na(Loans))
26 Loans = na.omit(Loans) #removing or dropping the Na as it's less than 5% of the Data
27 Loans$loanID = NULL #dropping the id
28 Loans$state = NULL #dropping states
29 Loans$verified = NULL #dropping verification Income
30 Loans$employment = NULL #dropping Employment
31 sum(is.na(Loans))
32
```

```
33:1 1.Preparing and Cleaning the Data -
Console Terminal Jobs
~/Cluster Study in R/Section A PR mid term paper/Project DS/Endyear-Project/ #
[870] 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
[881] 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
[892] 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
[903] 0.06250 0.06250 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
[914] 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
[925] 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
[936] 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
[947] 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
[958] 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
[969] 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
[980] 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
[991] 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
[ reached getOption("max.print") -- omitted 49000 entries ]
> Loans = na.omit(Loans)#removing or dropping the Na as it's less than 5% of the Data
> Loans$loanID = NULL #dropping the id
> Loans$state = NULL #dropping states
> Loans$verified = NULL #dropping verification Income
> Loans$employment = NULL #dropping employment
```

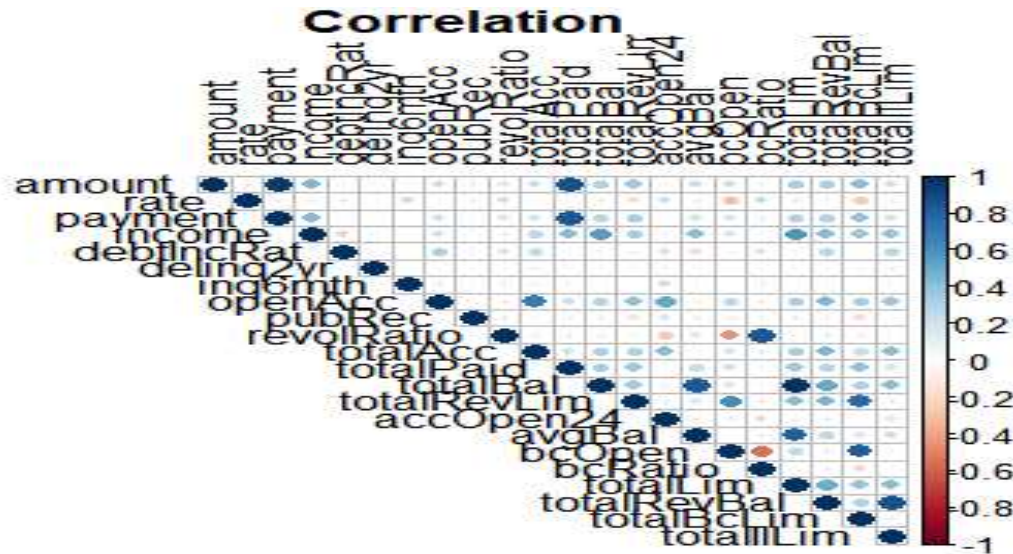
## Exploring and Transforming the Data

I have started exploring and transforming the data as factor and characters and integer properly. The loan status contains multiple categories, for the purpose of my analysis I will retain only two categories of loan status: 'Fully Paid' a good loan, 'Charged Off' and 'Default' as bad loan status. We will also replace original "n/a", missing data, with the default value of NA for the purposes of my analysis.

After data preparation, we can now explore our dataset; first let's look at how many credit inquiries have been done in the past 6 months for both good and bad loan status. As we can note on our boxplot, the median for Good loans is laying on 0 value meaning median inquiry for the past 6 months is no inquiries have been done, however for Bad loans status we see one inquiry; both boxplots are showing similar shape.



And then I checked the correlation between the variables and change the Loans status into categorical factor of Good and Bad Loans. And check the correlation using the Pearson method to see how they are correlated.



```
# 2.Exploring and Transforming the Data----
summary(Loans)
Profit = sum(Loans$amount-Loans$totalPaid)
Profit
Loans$status = as.factor(Loans$status)
levels(Loans$status) = list("Good"=c("Fully Paid"),"Bad"=c("Charged Off","In Grace Period","Late (16-30 days)","Late (31-120 days)"))
table(Loans$status)
Loans$status = as.character(Loans$status)
Loans$length = as.character(Loans$length)
Loans$reason = as.character(Loans$reason)
Loans$amount = as.numeric(LoansInt$amount)
Loans$rate = as.numeric(LoansInt$rate)
Loans$term = as.character.Date(Loans$term)
Loans$grade = as.character(Loans$grade)
Loans$length = as.character.Date(Loans$length)
```

R Script

## The Logistic Model

The logistic model that I used is generalized Linear model of a binomial family. The generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. So I used a glm algorithm where I used the factor of the as status against all other variable and the result was not satisfactory it was unbalanced as they were more good than bad in a greater value.

```
105
106 LrgModel = glm(status ~ ., data = TrainLoans, family = 'binomial')
107 summary(LrgModel)
108 prLrg = predict(LrgModel, newdata = TestLoans, type = "response") # to do prediction
109 trs = 0.5 # threshold of 0.5
110 prediction = cut(prLrg, breaks=c(-Inf, trs, Inf), labels=c("Bad", "Good"), header = TRUE)
111 mtab <- table(prediction, TestLoans$status)
112 addmargins(mtab)
113 PLevel = round(sum(diag(mtab)) / sum(mtab)*100, 2) # Percentage correctly predicted
114 bad = round(diag(mtab)[1]*100/(diag(mtab)[1]+(mtab)[3]), 2) # to calculate correctly predicted bad loans
115 good = round((mtab)[4]*100/((mtab)[2]+(mtab)[4]), 2) # to calculate correctly predicted good loans
116 incbad = round((mtab)[3]*100/(diag(mtab)[1]+(mtab)[3]), 2) # to calculate incorrectly predicted bad loans
117 incgood = round((mtab)[2]*100/((mtab)[2]+(mtab)[4]), 2) # to calculate incorrectly predicted good loans
118 print(paste('Percent correctly predicted = ', PLevel, '%'))
119 print(paste('Percent of loans correctly predicted as being bad is', bad, '%', 'and good is', good, '%'))
120 print(paste('Percent of loans incorrectly predicted as being bad is', incbad, '%', 'and good is', incgood, '%'))
121
122
123
124
125
126
127
```

121:1 The Logistic Model

Console Terminal Jobs

```
~/Cluster Study in R/Section A PR mid term paper/Project DS/Endyear-Project/
> incgood = round((mtab)[2]*100/((mtab)[2]+(mtab)[4]), 2) # to calculate incorrectly predicted good loans
> print(paste('Percent correctly predicted = ', PLevel, '%'))
[1] "Percent correctly predicted = 97.82 %"
> print(paste('Percent of loans correctly predicted as being bad is', bad, '%', 'and good is', good, '%'))
[1] "Percent of loans correctly predicted as being bad is 98.59 % and good is 97.61 %"
> print(paste('Percent of loans incorrectly predicted as being bad is', incbad, '%', 'and good is', incgood, '%'))
[1] "Percent of loans incorrectly predicted as being bad is 1.41 % and good is 2.39 %"
>
```



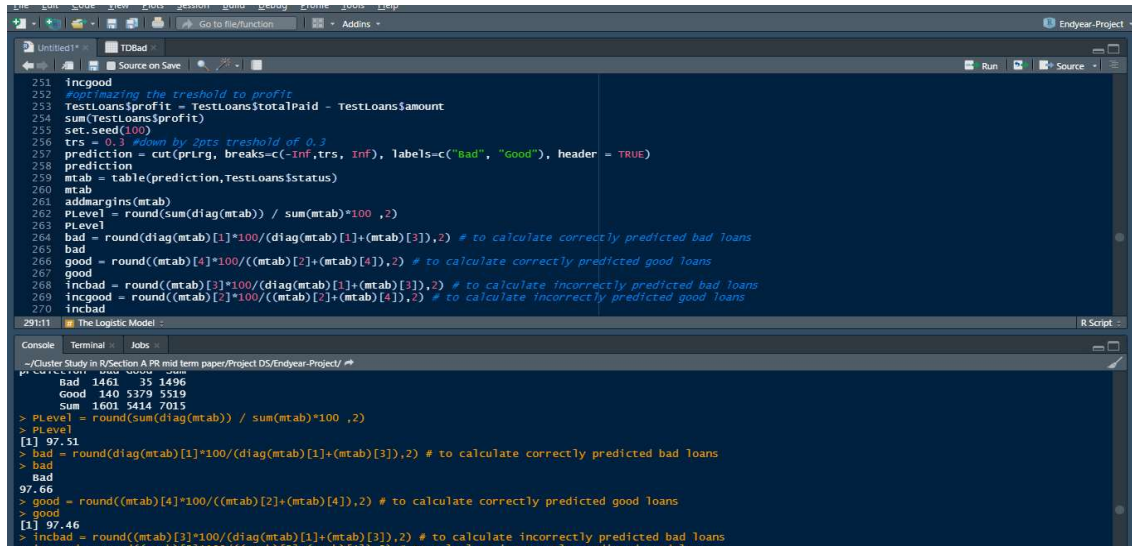
## Optimizing the Threshold for Accuracy (Balancing the data)

Based on the results produced, we can see that the full model produced correctly predicted results of 97.82% and the percent of bad loans were correctly predicted as being bad is 98.59% were correctly predicted). As for good loans we have had 97.61 %"predicted as good of correctly. I Can note that with current model the correct prediction of the bad loan is very higher than the good loan and needs improvement. So I used a new balanced and data the prediction as fallen a bit of Percent correctly predicted to 96.92 % but Percent of loans correctly predicted as being bad is 92.83 % and good is 98.15 %“ as good was good.

```
47 bad
48 good = round((mtab)[4]*100/((mtab)[2]+(mtab)[4]),2) # to calculate correctly predicted good loans
49 good
50 incbad = round((mtab)[3]*100/(diag(mtab)[1]+(mtab)[3]),2) # to calculate incorrectly predicted bad loans
51 incgood = round((mtab)[2]*100/((mtab)[2]+(mtab)[4]),2) # to calculate incorrectly predicted good loans
52 incbad
53 incgood
54
55 print(paste('Percent correctly predicted = ', PLevel,'%'))
56 print(paste('Percent of loans correctly predicted as being bad is', bad,'% ', 'and good is',good,'%'))
57 print(paste('Percent of loans incorrectly predicted as being bad is', incbad,'% ', 'and good is',incgood,'%'))
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81 The Logistic Model :
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
```

# Optimizing the Threshold for Profit

I have optimized the threshold three model with 0.8 threshold the accuracy has dropped meanwhile at 0.3 it was far with 97.51 with considerable profit .



```
251 incgood
252 #optimizing the threshold to profit
253 TestLoans$profit = TestLoans$totalPaid - TestLoans$amount
254 sum(TestLoans$profit)
255 set.seed(100)
256 trs = 0.3 #down by 2pts threshold of 0.3
257 prediction = cut(prlg, breaks=c(-Inf, trs, Inf), labels=c("Bad", "Good"), header = TRUE)
258 prediction
259 mtab = table(prediction, TestLoans$status)
260 mtab
261 addmargins(mtab)
262 PLevel = round(sum(diag(mtab)) / sum(mtab)*100, 2)
263 PLevel
264 bad = round(diag(mtab)[1]*100/(diag(mtab)[1]+(mtab)[3]), 2) # to calculate correctly predicted bad loans
265 bad
266 good = round((mtab)[4]*100/((mtab)[2]+(mtab)[4]), 2) # to calculate correctly predicted good loans
267 good
268 incbad = round((mtab)[3]*100/(diag(mtab)[1]+(mtab)[3]), 2) # to calculate incorrectly predicted bad loans
269 incgood = round((mtab)[2]*100/((mtab)[2]+(mtab)[4]), 2) # to calculate incorrectly predicted good loans
270 incbad
```

```
291:11 The Logistic Model
Console Terminal Jobs
~/Cluster Study in R/Section A PR mid term paper/Project DS/Endyear-Project/ ➤
> Corrected Confusion Matrix
Bad 1461 35 1496
Good 140 5379 5519
Sum 1601 5414 7015
> PLevel = round(sum(diag(mtab)) / sum(mtab)*100, 2)
[1] 97.51
> bad = round(diag(mtab)[1]*100/(diag(mtab)[1]+(mtab)[3]), 2) # to calculate correctly predicted bad loans
[1] 97.66
> good = round((mtab)[4]*100/((mtab)[2]+(mtab)[4]), 2) # to calculate correctly predicted good loans
[1] 97.46
> incbad = round((mtab)[3]*100/(diag(mtab)[1]+(mtab)[3]), 2) # to calculate incorrectly predicted bad loans
[1] 97.46
> incgood = round((mtab)[2]*100/((mtab)[2]+(mtab)[4]), 2) # to calculate incorrectly predicted good loans
```

## **Results Summary**

The determined model provides an overall accuracy where predicted correctly fully paid loans are at 93.26% with a proposed threshold of 0.3 with our estimated predicted profit of \$ 12761340. However, the trade off of the profit comes with the price of denying some of the loans that actually would have been fully paid, the loan's status that was incorrectly predicted is 1.17%. The variation of the threshold can increase and decrease the percent of correctly predicted status, if we look at the graph of threshold level comparison to predicted profit on the right we will note that the profit is increasing up until threshold of about 0.3 and then decreasing as it goes up. Therefore, by implementing the proposed model the bank can increase their potential profit by \$ 10855971.

## Conclusion

So After performing data set cleaning and preparing as well as oversampling our training data set, then applying statistical model analysis we were able to build a plausible logistic regression model to predict good and bad loans and find a reasonable classification threshold in order to achieve the most profit for the bank. The model have shown the ability to predict the loans that will be fully paid off (Good) and charged off or default (Bad)

**Therefore, we suggest for the bank prediction for the loan status is the following:  $\text{status} \sim \text{totalPaid} + \text{amount} + \text{rate} + \text{payment} + \text{grade} + \text{bcRatio} + \text{debtIncRat} + \text{totalBal} + \text{length} + \text{term} + \text{totalAcc} + \text{openAcc} + \text{delinq2yr} + \text{totalBcLim} + \text{totalIllLim} + \text{totalRevBal} + \text{totalLim} + \text{totalRevLim} + \text{revolRatio} + \text{accOpen24} + \text{bcOpen}$  with 0.3 threshold applied.**

**I.e. : Though I dropped the employment variable when doing the preparation because of inconstancy , I do believe that a well classified employment variable would be helpful when running prediction.**

