Capstone Project

Data Analysis of Lending Club Loan Data (2007 – 2014)

Foundations of Data Science Workshop on SpringBoard

Author: Sathish Kumar V

Mentor: Neal Fultz

Introduction:

Lending Club is an online marketplace connecting borrowers and investors and aims to transform the banking system to make credit more affordable and investing more rewarding. Lending Club, via its peer-to-peer lending model, operates at a lower cost than traditional bank lending programs and claims to pass the savings on to borrowers in the form of lower rates and to investors in the form of solid returns.

How peer-to-peer lending works -

- Customers interested in a loan complete a simple application at lendingClub.com

- The website leverages online data and technology to quickly assess risk, determine a credit rating and assign appropriate interest rates.

- Qualified applicants receive offers and can evaluate loan options with no impact to their credit score

- Investors ranging from individuals to institutions select loans in which to invest and can earn monthly returns

Lending club publishes the portfolio of loans issued via its website and this will be used as the data source for this capstone project. The data can be downloaded from https://www.lendingclub.com/info/download-data.action

From the URL listed above, the following data sets have been downloaded

1. LoanStats3a.csv – Loan Statistics from 2007 – 2011 containing 42542 observations of 112 variables
2. LoanStats3b.csv - Loan Statistics from 2012 – 2013 containing 188128 observations of 112 variables
3. LoanStats3c.csv - Loan Statistics for 2014 containing 67502 observations of 112 variables

In this capstone project, we will use the datasets listed in (1) & (2) and build logistic regression and random forest models. Using these models, we will predict the possibility of bad loans on the dataset in (3), i.e., (1) & (2) will be our training datasets & (3) will be our testing dataset.

Structure of the data:

The data dictionary describing the columns is published at https://resources.lendingclub.com/LCDataDictionary.xlsx

The loan data for 2007-2011 is imported to RStudio and the structure looks thus:

id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term
-6.77	Min. : 149512	Min. : 1000	Min. : 1000	Min. : 950	-6.736111111
10004577: 1	1st Qu.: 2169504	1st Qu.: 8000	1st Qu.: 8000	1st Qu.: 8000	36 months:14385(
10004609: 1	Median : 6047542	Median :12175	Median :12125	Median :12100	60 months: 44273
10004610: 1	Mean : 5910733	Mean :14355	Mean :14352	Mean :14340	NA
10004615: 1	3rd Qu.: 8721082	3rd Qu.:20000	3rd Qu.:20000	3rd Qu.:19975	NA
10004622: 1	Max. :12096968	Max. :35000	Max. :35000	Max. :35000	NA
(Other) :188119	NA's :4	NA's :5	NA's :5	NA's :5	NA

int_rate	installment	grade	sub_grade	emp_title	emp_length
12.12% : 9409	Min. : 21.62	B :62605	B3 : 15614	: 11720	10+ years:61175
13.11% : 8257	1st Qu.: 269.98	C :49988	B4 : 14440	Teacher: 830	2 years :16143
8.90% : 7638	Median : 398.21	A :28576	B2 : 13058	Manager: 666	5 years :14265
14.33% : 7142	Mean : 443.75	D :27881	C1 : 11191	US Army: 491	3 years :13965
7.90% : 6612	3rd Qu.: 578.34	E :12242	C2 : 10618	RN : 388	< 1 year :13222
11.14% : 6428	Max. :1408.13	F : 5706	B1 : 10357	(Other):174031	6 years :11916
(Other):142642	NA's :5	(Other): 1130	(Other):112850	NA's : 2	(Other) :57442

home_ownershi	annual_inc	verification_sta	issue_d	loan_status	pymnt_plan
################	Min. : 4800	################	Dec-13 : 15020	Fully Paid :120	################
MORTGAGE:96979	1st Qu.: 45000	Not Verified :588	Nov-13 : 14676	Current : 399	n:188122
NONE : 42	Median : 62000	Source Verified:4	Oct-13 : 14115	Charged Off : 2	y: 1
OTHER : 46	Mean : 72239	Verified :8735	Sep-13 : 12987	Late (31-120 days)	NA
OWN :15447	3rd Qu.: 87000	NA	Aug-13 : 12674	In Grace Period :	NA
RENT :75609	Max. :7141778	NA	Jul-13 : 11910	Late (16-30 days)	NA
NA	NA's :5	NA	(Other):106746	(Other) : 49	NA

		purpose	title	zip_code	addr_state
################		debt_consolidati	Debt consolidatio	945xx : 2327	CA :30734
https://lendingclu	Borrower added	credit_card : 4	Debt Consolidatio	112xx : 2182	NY :16254
https://lendingclu	Borrower added	home_improvem	Credit card refina	750xx : 2075	TX :14557
https://lendingclu	Borrower added	other : 889	Consolidation	606xx : 1875	FL :12842
https://lendingclu	Borrower added	major_purchase	debt consolidatio	100xx : 1830	IL : 7312
https://lendingclu	Borrower added	small_business	Other : 2	900xx : 1804	NJ : 7210
(Other)	(Other)	(Other) : 796	(Other) :1	(Other):176035	(Other):99219

dti	delinq_2yrs	earliest_cr_line	inq_last_6mths	mths_since_last_	mths_since_last_
Min. : 0.00	Min. : 0.0000	Oct-00 : 1646	Min. :0.0000	Min. : 0.00	Min. : 1.00
1st Qu.:11.34	1st Qu.: 0.0000	Oct-01 : 1476	1st Qu.:0.0000	1st Qu.: 17.00	1st Qu.: 69.00
Median :16.78	Median : 0.0000	Oct-99 : 1470	Median :0.0000	Median : 32.00	Median : 95.00
Mean :17.06	Mean : 0.2396	Nov-99 : 1461	Mean :0.8036	Mean : 34.98	Mean : 86.02
3rd Qu.:22.58	3rd Qu.: 0.0000	Nov-00 : 1459	3rd Qu.:1.0000	3rd Qu.: 50.00	3rd Qu.:106.00
Max. :34.99	Max. :29.0000	Dec-00 : 1381	Max. :8.0000	Max. :156.00	Max. :121.00
NA's :5	NA's :5	(Other):179235	NA's :5	NA's :107549	NA's :170665

open_acc	pub_rec	revol_bal	revol_util	total_acc	initial_list_status
Min. : 0	Min. : 0.0000	Min. : 0	0% : 624	Min. : 2.00	################
1st Qu.: 8	1st Qu.: 0.0000	1st Qu.: 7132	61.50% : 342	1st Qu.: 16.00	f:148316
Median :10	Median : 0.0000	Median : 12437	64.60% : 340	Median : 23.00	w: 39807
Mean :11	Mean : 0.1062	Mean : 16319	66.50% : 337	Mean : 24.54	NA
3rd Qu.:14	3rd Qu.: 0.0000	3rd Qu.: 20670	67.40% : 334	3rd Qu.: 31.00	NA
Max. :62	Max. :54.0000	Max. :2568995	61.60% : 332	Max. :105.00	NA
NA's :5	NA's :5	NA's :5	(Other):185819	NA's :5	NA

out_prncp	out_prncp_inv	total_pymnt	total_pymnt_inv	total_rec_prncp	total_rec_int
Min. : 0	Min. : 0	Min. : 0	Min. : 0	Min. : 0	Min. : 0
1st Qu.: 0	1st Qu.: 0	1st Qu.: 8039	1st Qu.: 8032	1st Qu.: 6000	1st Qu.: 1148
Median : 0	Median : 0	Median :12934	Median :12922	Median :10000	Median : 2111
Mean : 1244	Mean : 1243	Mean :15019	Mean :15006	Mean :11705	Mean : 3163
3rd Qu.: 0	3rd Qu.: 0	3rd Qu.:20255	3rd Qu.:20239	3rd Qu.:15953	3rd Qu.: 3941
Max. :24973	Max. :24902	Max. :57898	Max. :57778	Max. :35000	Max. :25595
NA's :5	NA's :5	NA's :5	NA's :5	NA's :5	NA's :5

total_rec_late_fe	recoveries	collection_recove	last_pymnt_d	last_pymnt_amnt	next_pymnt_d
Min. : 0.0000	Min. : 0.0	Min. : 0.00	Jun-16 : 44296	Min. : 0.0	:145680
1st Qu.: 0.0000	1st Qu.: 0.0	1st Qu.: 0.00	Jul-15 : 5564	1st Qu.: 362.2	Aug-16: 1
Median : 0.0000	Median : 0.0	Median : 0.00	Oct-15 : 5360	Median : 757.8	Jul-16: 42427
Mean : 0.8196	Mean : 149.9	Mean : 14.73	Mar-16 : 5117	Mean : 3899.1	Jun-16: 19
3rd Qu.: 0.0000	3rd Qu.: 0.0	3rd Qu.: 0.00	Mar-15 : 5086	3rd Qu.: 5469.9	May-16: 1
Max. :358.6800	Max. :39443.6	Max. :6124.94	Apr-15 : 4971	Max. :35760.2	NA
NA's :5	NA's :5	NA's :5	(Other):117734	NA's :5	NA

last_credit_pull_d	collections_12_m	mths_since_last_	policy_code	application_type	annual_inc_joint
Jun-16 :102366	Min. :0.000000	Min. : 0.00	Min. :1	################	Mode:logical
Mar-16 : 6244	1st Qu.:0.000000	1st Qu.: 25.00	1st Qu.:1	INDIVIDUAL:1881	NA's:188128
Apr-16 : 6015	Median :0.000000	Median : 41.00	Median :1	NA	NA
May-16 : 5108	Mean :0.003173	Mean : 41.79	Mean :1	NA	NA
Feb-16 : 4738	3rd Qu.:0.000000	3rd Qu.: 58.00	3rd Qu.:1	NA	NA
Jan-16 : 4479	Max. :4.000000	Max. :165.00	Max. :1	NA	NA
(Other): 59178	NA's :5	NA's :155631	NA's :5	NA	NA

dti_joint	verification_statu	acc_now_delinq	tot_coll_amt	tot_cur_bal	open_acc_6m
Mode:logical	Mode:logical	Min. :0.000000	Min. : 0.00	Min. : 0	Mode:logical
NA's:188128	NA's:188128	1st Qu.:0.000000	1st Qu.: 0.00	1st Qu.: 27471	NA's:188128
NA	NA	Median :0.000000	Median : 0.00	Median : 80764	NA
NA	NA	Mean :0.002716	Mean : 76.75	Mean : 137331	NA
NA	NA	3rd Qu.:0.000000	3rd Qu.: 0.00	3rd Qu.: 208185	NA
NA	NA	Max. :5.000000	Max. :88303.00	Max. :8000078	NA
NA	NA	NA's :5	NA's :27746	NA's :27746	NA

open_il_6m	open_il_12m	open_il_24m	mths_since_rcnt_	total_bal_il	il_util
Mode:logical	Mode:logical	Mode:logical	Mode:logical	Mode:logical	Mode:logical
NA's:188128	NA's:188128	NA's:188128	NA's:188128	NA's:188128	NA's:188128
NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA

open_rv_12m	open_rv_24m	max_bal_bc	all_util	total_rev_hi_lim	inq_fi
Mode:logical	Mode:logical	Mode:logical	Mode:logical	Min. : 0	Mode:logical
NA's:188128	NA's:188128	NA's:188128	NA's:188128	1st Qu.: 13900	NA's:188128
NA	NA	NA	NA	Median : 23000	NA
NA	NA	NA	NA	Mean : 29894	NA
NA	NA	NA	NA	3rd Qu.: 37200	NA
NA	NA	NA	NA	Max. :9999999	NA
NA	NA	NA	NA	NA's :27746	NA

total_cu_tl	inq_last_12m	acc_open_past_24	avg_cur_bal	bc_open_to_buy	bc_util
Mode:logical	Mode:logical	Min. : 0.00	Min. : 0	Min. : 0	Min. : 0.00
NA's:188128	NA's:188128	1st Qu.: 2.00	1st Qu.: 3013	1st Qu.: 1062	1st Qu.: 49.50
NA	NA	Median : 4.00	Median : 7776	Median : 3523	Median : 72.20
NA	NA	Mean : 3.93	Mean : 13797	Mean : 8264	Mean : 66.83
NA	NA	3rd Qu.: 5.00	3rd Qu.: 19669	3rd Qu.: 9704	3rd Qu.: 89.00
NA	NA	Max. :40.00	Max. :958084	Max. :497445	Max. :339.60
NA	NA	NA's :7500	NA's :27752	NA's :9029	NA's :9116

chargeoff_within_	delinq_amnt	mo_sin_old_il_ac	mo_sin_old_rev_	mo_sin_rcnt_rev_	mo_sin_rcnt_tl
Min. :0.000000	Min. : 0.00	Min. : 0.0	Min. : 5.0	Min. : 0.00	Min. : 0.000
1st Qu.:0.000000	1st Qu.: 0.00	1st Qu.: 95.0	1st Qu.:116.0	1st Qu.: 4.00	1st Qu.: 3.000
Median :0.000000	Median : 0.00	Median :128.0	Median :161.0	Median : 9.00	Median : 6.000
Mean :0.005268	Mean : 8.37	Mean :125.1	Mean :178.5	Mean : 14.11	Mean : 8.939
3rd Qu.:0.000000	3rd Qu.: 0.00	3rd Qu.:151.0	3rd Qu.:222.0	3rd Qu.: 17.00	3rd Qu.: 11.000
Max. :5.000000	Max. :65000.00	Max. :649.0	Max. :760.0	Max. :264.00	Max. :211.000
NA's :5	NA's :5	NA's :33876	NA's :27747	NA's :27747	NA's :27746

mort_acc	mths_since_recer	mths_since_recer	mths_since_recer	mths_since_recer	num_accts_ever_
Min. : 0.000	Min. : 0.00	Min. : 0.00	Min. : 0.000	Min. : 0.00	Min. : 0.000
1st Qu.: 0.000	1st Qu.: 7.00	1st Qu.: 23.00	1st Qu.: 2.000	1st Qu.: 18.00	1st Qu.: 0.000
Median : 1.000	Median : 15.00	Median : 40.00	Median : 6.000	Median : 34.00	Median : 0.000
Mean : 1.811	Mean : 25.66	Mean : 40.82	Mean : 6.992	Mean : 36.61	Mean : 0.329
3rd Qu.: 3.000	3rd Qu.: 33.00	3rd Qu.: 58.00	3rd Qu.:11.000	3rd Qu.: 52.00	3rd Qu.: 0.000
Max. :31.000	Max. :554.00	Max. :152.00	Max. :24.000	Max. :165.00	Max. :29.000
NA's :7500	NA's :8832	NA's :151393	NA's :27865	NA's :133702	NA's :27746

num_actv_bc_tl	num_actv_rev_tl	num_bc_sats	num_bc_tl	num_il_tl	num_op_rev_tl
Min. : 0.000	Min. : 0.000	Min. : 0.000	Min. : 0.000	Min. : 0.000	Min. : 0.000
1st Qu.: 2.000	1st Qu.: 4.000	1st Qu.: 3.000	1st Qu.: 6.000	1st Qu.: 3.000	1st Qu.: 5.000
Median : 3.000	Median : 5.000	Median : 4.000	Median : 8.000	Median : 6.000	Median : 7.000
Mean : 3.755	Mean : 5.675	Mean : 4.666	Mean : 9.019	Mean : 7.726	Mean : 8.095
3rd Qu.: 5.000	3rd Qu.: 7.000	3rd Qu.: 6.000	3rd Qu.:12.000	3rd Qu.:10.000	3rd Qu.:10.000
Max. :30.000	Max. :37.000	Max. :35.000	Max. :65.000	Max. :66.000	Max. :58.000
NA's :27746	NA's :27746	NA's :16060	NA's :27746	NA's :27746	NA's :27746

num_rev_accts	num_rev_tl_bal_g	num_sats	num_tl_120dpd_2	num_tl_30dpd	num_tl_90g_dpd_
Min. : 0.00	Min. : 0.000	Min. : 0.00	Min. :0	Min. :0.000	Min. : 0.000
1st Qu.:10.00	1st Qu.: 4.000	1st Qu.: 8.00	1st Qu.:0	1st Qu.:0.000	1st Qu.: 0.000
Median :14.00	Median : 5.000	Median :10.00	Median :0	Median :0.000	Median : 0.000
Mean :14.95	Mean : 5.693	Mean :11.09	Mean :0	Mean :0.002	Mean : 0.064
3rd Qu.:19.00	3rd Qu.: 7.000	3rd Qu.:14.00	3rd Qu.:0	3rd Qu.:0.000	3rd Qu.: 0.000
Max. :94.00	Max. :37.000	Max. :62.00	Max. :2	Max. :4.000	Max. :24.000
NA's :27746	NA's :27746	NA's :16060	NA's :28002	NA's :27746	NA's :27746

num_tl_op_past_	pct_tl_nvr_dlq	percent_bc_gt_75	pub_rec_bankrup	tax_liens	tot_hi_cred_lim
Min. : 0.000	Min. : 15.0	Min. : 0.00	Min. :0.00000	Min. : 0.00000	Min. : 0
1st Qu.: 1.000	1st Qu.: 93.0	1st Qu.: 25.00	1st Qu.:0.00000	1st Qu.: 0.00000	1st Qu.: 44800
Median : 2.000	Median :100.0	Median : 50.00	Median :0.00000	Median : 0.00000	Median : 108504
Mean : 1.788	Mean : 95.4	Mean : 53.56	Mean :0.08475	Mean : 0.01403	Mean : 165554
3rd Qu.: 3.000	3rd Qu.:100.0	3rd Qu.: 80.00	3rd Qu.:0.00000	3rd Qu.: 0.00000	3rd Qu.: 243747
Max. :25.000	Max. :100.0	Max. :100.00	Max. :8.00000	Max. :53.00000	Max. :9999999
NA's :27746	NA's :27899	NA's :9032	NA's :5	NA's :5	NA's :27746

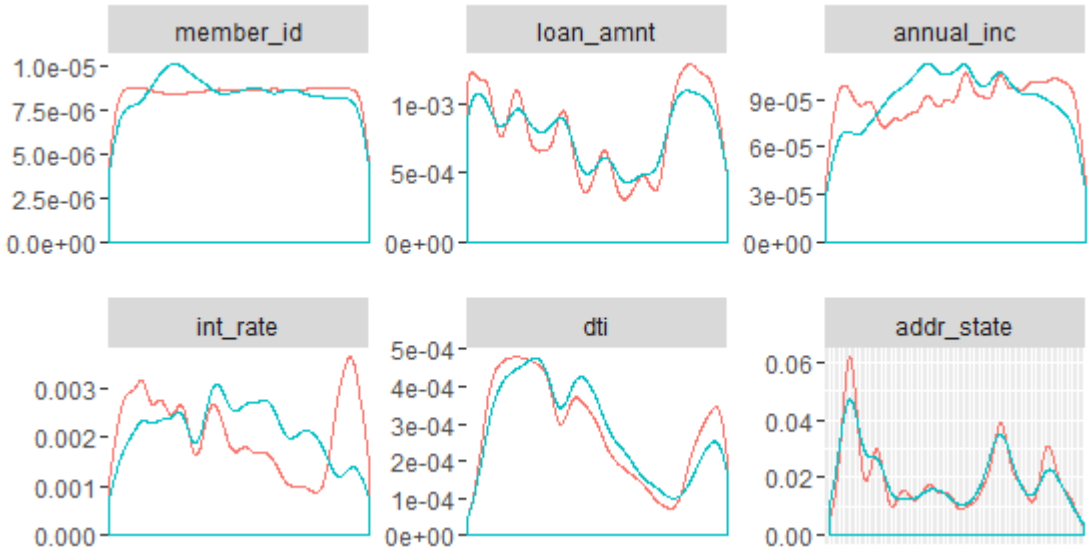
total_bal_ex_mor	total_bc_limit	total_il_high_credit_limit
Min. : 0	Min. : 0	Min. : 0
1st Qu.: 18905	1st Qu.: 7800	1st Qu.: 10047
Median : 32963	Median : 14700	Median : 25752
Mean : 42885	Mean : 20240	Mean : 34391
3rd Qu.: 53859	3rd Qu.: 26500	3rd Qu.: 46942
Max. :2644442	Max. :522210	Max. :1214546
NA's :7500	NA's :7500	NA's :27746

To classify a loan as bad, i.e., a higher risk of default, we will look at the loan_status column. Of the values in the loan_status, the following will be used to declare the loan as bad –

1. Charged Off
2. Late (31-120 days)
3. Late (16-30 days)
4. Default

A new binary column called bad_loan is added to reflect the value of the loan status and takes a value of 1 if the loan_status matches any of the strings listed above. Else the value will be set to 0.

The six columns listed below are chosen to build the model and the density plot of the bad_loan against each of the variables is plotted below.



Data Wrangling:

Once the datasets are imported, the following data cleanup activity is performed –

1. Interest Rate
   The interest rate column is imported as a character variable and has a % symbol at the end. The % value is removed using gsub & the column is converted to a numeric

2. State
   The factor type addr_state column has some missing values. These values are set to CA (the mode of the distribution)

3. Annual_inc
   The variable annual_inc is set to character. It is converted to numeric type

4. Handling NA values
   All the NA values in the data are set to the mean of the corresponding columns

Building the Logistic Regression Model:

The first logistic regression model is built using all the independent variables mentioned in the previous section.

logmodel1 <- glm(bad_loan ~ loan_amnt + annual_inc + grade + int_rate + dti + addr_state, data = df, family = "binomial")

- summary(logmodel1)

```
Call:

glm(formula = bad_loan ~ loan_amnt + annual_inc + grade + int_rate +

    dti + addr_state, family = "binomial", data = df)


Deviance Residuals:

    Min      1Q  Median      3Q     Max

-1.1072  -0.6036  -0.4887  -0.3480   4.3128
```

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.181e+01	3.449e+01	-0.343	0.731920	
loan_amnt	1.285e-05	8.813e-07	14.583	< 2e-16	***
annual_inc	-6.362e-06	2.125e-07	-29.940	< 2e-16	***
gradeA	8.333e+00	3.449e+01	0.242	0.809073	
gradeB	8.641e+00	3.449e+01	0.251	0.802154	
gradeC	8.774e+00	3.449e+01	0.254	0.799177	
gradeD	8.854e+00	3.449e+01	0.257	0.797389	
gradeE	8.870e+00	3.449e+01	0.257	0.797034	
gradeF	8.914e+00	3.449e+01	0.258	0.796053	
gradeG	8.865e+00	3.449e+01	0.257	0.797131	
int_rate	8.205e-02	4.440e-03	18.479	< 2e-16	***
dti	9.395e-03	8.464e-04	11.100	< 2e-16	***
addr_stateAL	2.445e-01	1.301e-01	1.879	0.060206	.
addr_stateAR	2.135e-01	1.379e-01	1.548	0.121535	
addr_stateAZ	1.915e-01	1.258e-01	1.523	0.127787	
addr_stateCA	2.342e-01	1.200e-01	1.952	0.050987	.
addr_stateCO	-3.270e-02	1.275e-01	-0.256	0.797619	
addr_stateCT	1.651e-01	1.289e-01	1.281	0.200198	
addr_stateDC	-2.952e-01	1.761e-01	-1.676	0.093757	.
addr_stateDE	1.283e-01	1.693e-01	0.758	0.448512	
addr_stateFL	3.568e-01	1.211e-01	2.947	0.003213	**
addr_stateGA	1.142e-01	1.241e-01	0.920	0.357388	
addr_stateHI	2.104e-01	1.420e-01	1.482	0.138419	
addr_stateIA	-4.380e-01	1.053e+00	-0.416	0.677383	
addr_stateID	-1.656e-01	1.062e+00	-0.156	0.876115	
addr_stateIL	4.310e-02	1.234e-01	0.349	0.726974	
addr_stateIN	1.330e-01	1.329e-01	1.001	0.316604	
addr_stateKS	-3.000e-02	1.372e-01	-0.219	0.826899	
addr_stateKY	9.867e-02	1.358e-01	0.726	0.467596	
addr_stateLA	2.190e-01	1.311e-01	1.670	0.094938	.
addr_stateMA	2.030e-01	1.254e-01	1.618	0.105573	

addr_stateMD 2.676e-01 1.253e-01 2.136 0.032670 *

addr_stateME -8.340e+00 6.861e+01 -0.122 0.903246

addr_stateMI 2.199e-01 1.253e-01 1.754 0.079355 .

addr_stateMN 9.278e-02 1.285e-01 0.722 0.470281

addr_stateMO 1.892e-01 1.282e-01 1.476 0.139925

addr_stateMS -3.454e-02 6.299e-01 -0.055 0.956267

addr_stateMT -1.558e-01 1.733e-01 -0.899 0.368752

addr_stateNC 1.911e-01 1.245e-01 1.535 0.124880

addr_stateNE 1.374e+00 5.855e-01 2.347 0.018922 *

addr_stateNH -1.205e-01 1.557e-01 -0.774 0.439212

addr_stateNJ 3.440e-01 1.227e-01 2.804 0.005042 **

addr_stateNM 1.767e-01 1.452e-01 1.217 0.223679

addr_stateNV 4.397e-01 1.279e-01 3.439 0.000585 ***

addr_stateNY 2.809e-01 1.208e-01 2.326 0.020005 *

addr_stateOH 1.627e-01 1.239e-01 1.313 0.189091

addr_stateOK 2.393e-01 1.350e-01 1.773 0.076231 .

addr_stateOR 7.879e-02 1.311e-01 0.601 0.547826

addr_statePA 1.159e-01 1.237e-01 0.937 0.348531

addr_stateRI 2.505e-01 1.499e-01 1.671 0.094722 .

addr_stateSC -3.095e-02 1.337e-01 -0.231 0.817012

addr_stateSD 1.326e-01 1.787e-01 0.742 0.457808

addr_stateTN 1.755e-01 1.339e-01 1.311 0.189983

addr_stateTX 3.725e-02 1.213e-01 0.307 0.758874

addr_stateUT 1.994e-01 1.378e-01 1.447 0.147931

addr_stateVA 2.159e-01 1.239e-01 1.743 0.081330 .

addr_stateVT -9.237e-04 2.011e-01 -0.005 0.996336

addr_stateWA 1.172e-01 1.258e-01 0.932 0.351242

addr_stateWI 7.247e-02 1.319e-01 0.549 0.582799

addr_stateWV -1.667e-01 1.544e-01 -1.080 0.280247

addr_stateWY -3.204e-01 1.889e-01 -1.696 0.089897 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 188410 on 230669 degrees of freedom

Residual deviance: 178859 on 230609 degrees of freedom

AIC: 178981

Number of Fisher Scoring iterations: 9

From the model above, we notice that the addr_state variable does not have much bearing on the data. Let us try another model by removing this variable.

logmodel <- glm(bad_loan ~ loan_amnt + annual_inc + grade + int_rate, data = df, family = "binomial")

- summary(logmodel)

Call:

glm(formula = bad_loan ~ loan_amnt + annual_inc + grade + int_rate,

family = "binomial", data = df)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.9964	-0.6060	-0.4920	-0.3499	4.4517

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.149e+01	3.449e+01	-0.333	0.739	
loan_amnt	1.382e-05	8.684e-07	15.911	<2e-16	***
annual_inc	-6.806e-06	2.072e-07	-32.841	<2e-16	***
gradeA	8.304e+00	3.449e+01	0.241	0.810	
gradeB	8.607e+00	3.449e+01	0.250	0.803	
gradeC	8.728e+00	3.449e+01	0.253	0.800	
gradeD	8.794e+00	3.449e+01	0.255	0.799	
gradeE	8.794e+00	3.449e+01	0.255	0.799	
gradeF	8.826e+00	3.449e+01	0.256	0.798	

```
gradeG      8.774e+00  3.449e+01   0.254    0.799
int_rate    8.765e-02  4.390e-03  19.967   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 188410  on 230669  degrees of freedom
Residual deviance: 179324  on 230659  degrees of freedom
AIC: 179346

Number of Fisher Scoring iterations: 9
```

Based on the above data, let us make our predictions on logmodel.

- predictlog <- predict(logmodel, newdata = dfTest, type = "response")

The confidence matrix can be realized using the model –

- table(round(predictlog), dfTest$bad_loan)

```
      0     1
0 59099  8360
1    31    12
```

We see that sensitivity is 12/43 = 0.28 & the specificity is 59099/67459 = 0.87.

Let us proceed ahead and build the randomForest model for the same dataset.

Modeling using randomForest:

Due to the huge dataset, randomForest throws up a memory allocation error. This memory allocation error is fixed by setting the nodesize for the model to 6 (default is 5)

- X <- as.matrix(df$int_rate)
- rfmodel <- randomForest(X,df$bad_loan, nodesize = 6)

Let us build predictions on this model

- Y <- as.matrix(dfTest$int_rate)
- predictrf <- predict(rfmodel, Y, type = "response")

We now select various threshold levels and evaluate the sensitivity & specificity.

- table(predictrf <= 0.1, dfTest$bad_loan)

	0	1
FALSE	40034	7132
TRUE	19096	1240

For the threshold value of 0.1, Sensitivity = 0.061, Specificity = 0.85

- table(predictrf <= 0.2, dfTest$bad_loan)

	0	1
FALSE	12424	2980
TRUE	46706	5392

For the threshold value of 0.2, Sensitivity = 0.103, Specificity = 0.806

- table(predictrf <= 0.3, dfTest$bad_loan)

	0	1
FALSE	1422	411
TRUE	57708	7961

For the threshold value of 0.3, Sensitivity = 0.121, Specificity = 0.776

Plotting the ROC:

Finally, we will plot the ROC for each of these models and evaluate the AUC to gauge the effectiveness of the models.

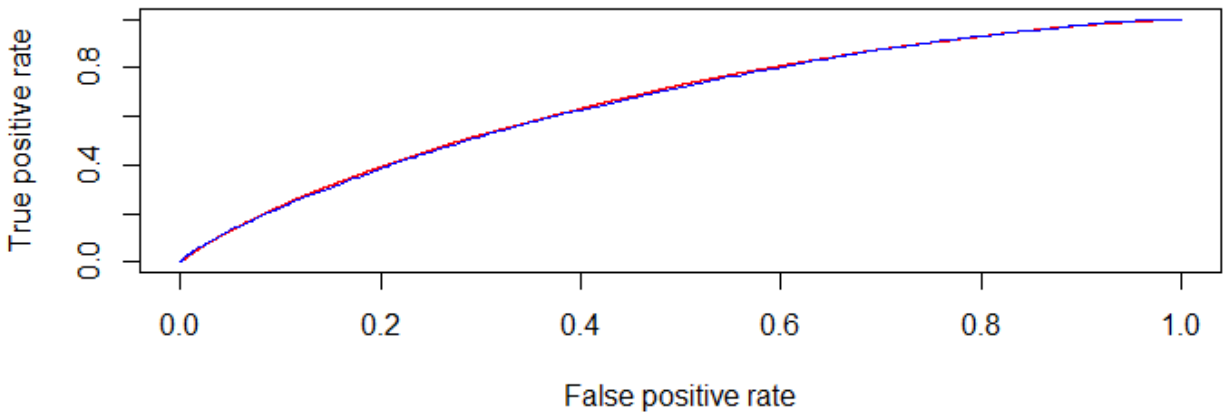
- predlog <- predict(logmodel, df, type = "response")
- ROCRlog <- prediction(predlog, df$bad_loan)
- ROCRperflog <- performance(ROCRlog, "tpr","fpr")

- Z <- as.matrix(df$int_rate)
- predrf <- predict(rfmodel, Z, type = "response")
- ROCRrf <- prediction(predrf, df$bad_loan)
- ROCRperfrf <- performance(ROCRrf,"tpr","fpr")

- plot(ROCRperflog, col = 'red')
- plot(ROCRperfrf, add = TRUE, col = 'blue')

- a <- rbinom(length(predlog), 1, 0.25)
- roc1 <- roc(a ~ predlog)
- roc2 <- roc(a ~ predrf)

- roc1$auc
- roc2$auc



Conclusion:

For the peer-to-peer lending model to flourish, the system should be able to offer healthy loans to the investors and therefore, must build a robust system to offer higher returns and flag potentially risky loans. This solution can be deployed to check new loan applications and validate their credentials before putting forth the proposal before the investors.