Predict Revolving Balance

Group 4
Mentors: Sri Vinod
Munmun Bhagat

Team:
Amit Sharma
Hymavathi Samsani
Mandar Malekar
Vijay Sonawane

10-10-2020

Business Problem:

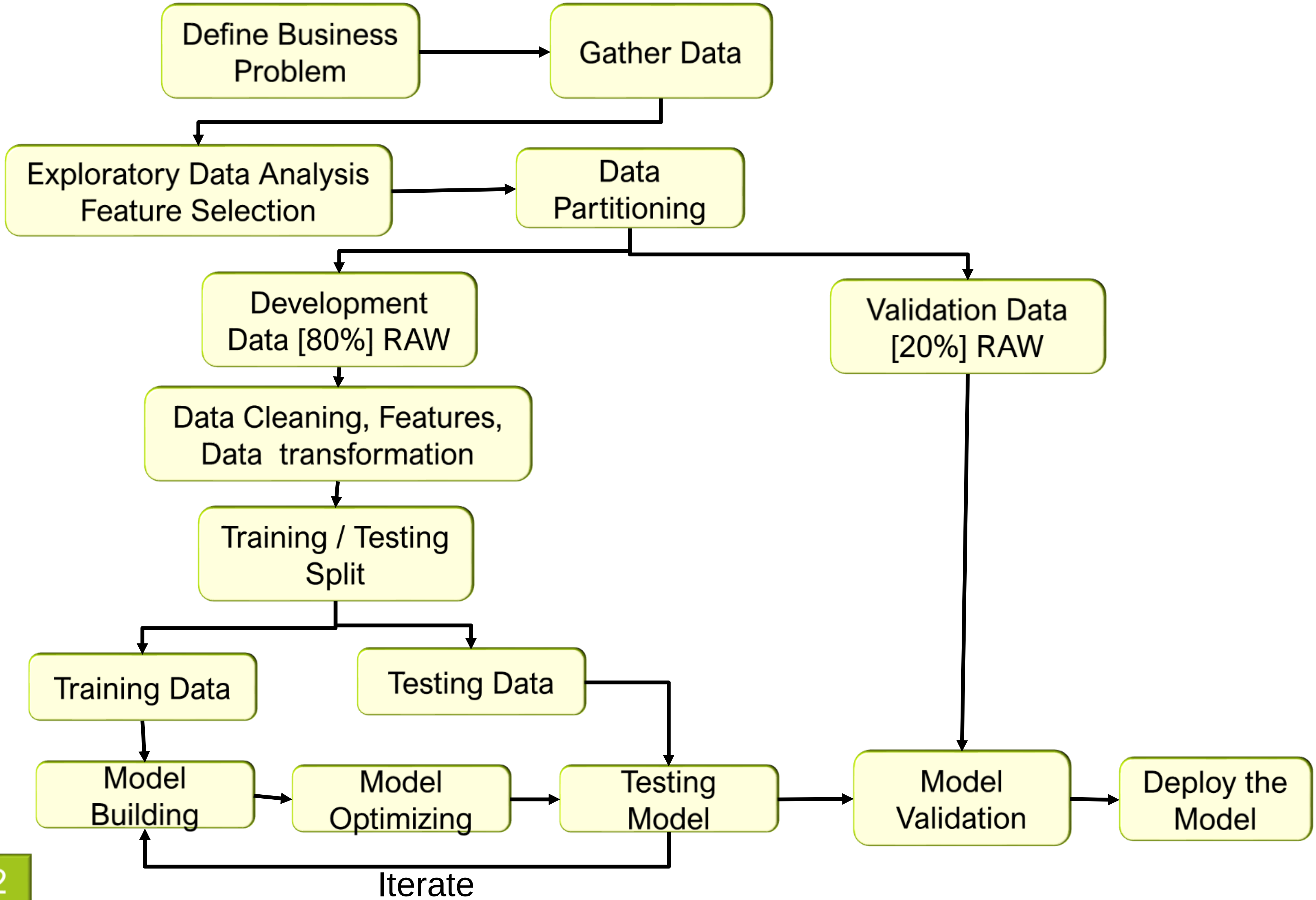
The investors who are into the Credit revolving balance, want to decide Marketing Strategies for revolving balance. The purpose is to maximise profitability by charging higher Interest rates

Objective:

The objective of the analysis is to predict the revolving balance maintained by the customer so that they can derive marketing strategies individually

Revolving Credit is similar to a credit card. Only difference being lower interest rate and secured by business assets. Revolving balance amount is balance payable

Project Architecture / Project Flow



Exploratory Data Analysis (EDA) and Feature Selection

Variables in the Data set

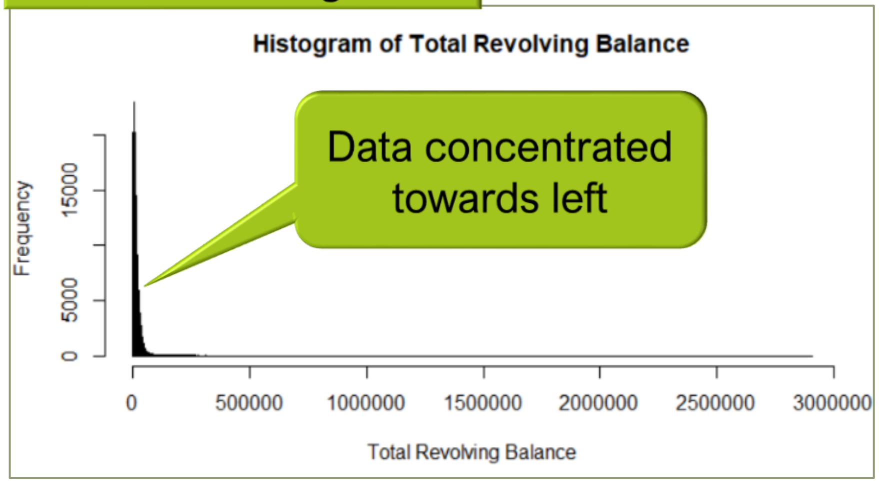
- Number of Records: 887379
- Variables: 36
- Target Variable: Total Revolving Balance

Categorical [Discrete] Factor	Integer Categorical / Numeric	Ccontinuous Numeric
terms	member_id	Rate_of_intrst
batch_ID	loan_amnt	annual_inc
grade	delinq_2yrs	debt_income_ratio
sub_grade	inq_last_6mths	total_rec_int
Emp_designation	mths_since_last_delinq	total_rec_late_fee
Experience	mths_since_last_record	recoveries
home_ownership	numb_credit	collection_recovery_fee
verification_status	pub_rec	tot_curr_bal
purpose	total_credits	
State	collections_12_mths_ex_m	
initial_list_status	mths_since_last_major_de	
application_type	acc_now_delinq	
verification_status_joint	tot_colle_amt	
last_week_pay		

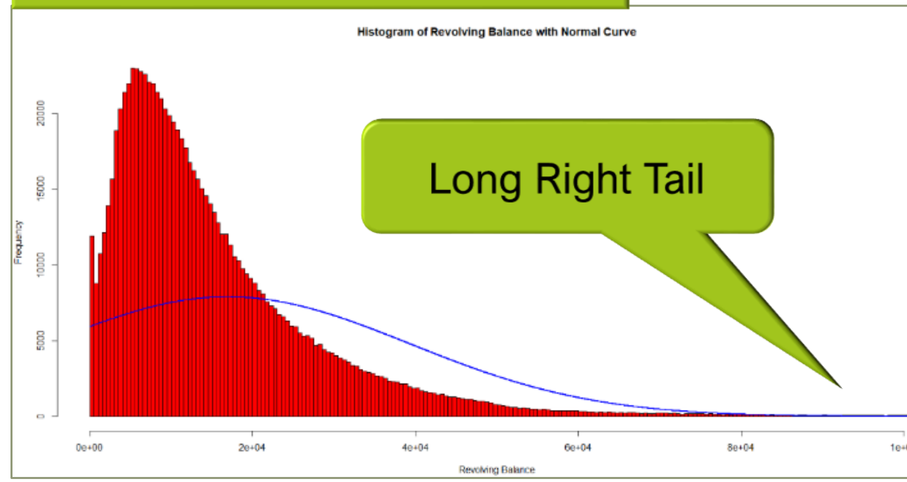
member_id	loan_amnt	terms	batch_ID	Rate_of_intrst	grade	sub_grade	Emp_designation	Experience	home_ownership	annual_inc	verification_status	purpose	State	debt_income_ratio	delinq_2yrs	inq_last_6mths	mths_since_last_delinq	mths_since_last_record	numb_credit	pub_rec	total revol_bal	total_credits	initial_list_status	total_rec_int	total_rec_late_fee	recoveries	collection_recovery_fee	collections_12_mths_ex_med	mths_since_last_major_derog	application_type	verification_status_joint	last_week_pay	acc_now_delinq	tot_colle_amt	tot_curr_bal
58189336	14350	36 months		19.19	E	E3	clerk	9 years	OWN	28700	Source Verified	debt_consolidation	FL	33.88	0	1	50	75	14	1	22515	28	f	1173.84	0	0	0	0	74	INDIVIDUAL		26th week	0	0	28699
70011223	4800	36 months	BAT1586599	10.99	B	B4	Human Resources Specialist	< 1 year	MORTGAGE	65000	Source Verified	home_improvement	MD	3.64	0	1			6	0	7624	13	w	83.95	0	0	0	0		INDIVIDUAL		9th week	0	0	9974
70255675	10000	36 months	BAT1586599	7.26	A	A4	Driver	2 years	OWN	45000	Not Verified	debt_consolidation	OH	18.42	0	0			5	0	10877	19	w	56.47	0	0	0	0		INDIVIDUAL		9th week	0	65	38295
1893936	15000	36 months	BAT4808022	19.72	D	D5	Us office of Personnel Management	10+ years	RENT	105000	Not Verified	debt_consolidation	VA	14.97	0	2	46		10	0	13712	21	f	4858.62	0	0	0	0		INDIVIDUAL		135th week	0	0	55564
7652106	16000	36 months	BAT2833642	10.64	B	B2	LAUSD-HOLLYWOOD HIGH SCHOOL	10+ years	RENT	52000	Verified	credit_card	CA	20.16	0	0			11	0	35835	27	w	2296.41	0	0	0	0		INDIVIDUAL		96th week	0	0	47159
10247268	15000	36 months	BAT2575549	8.9	A	A5	Design Consultant	2 years	MORTGAGE	120000	Not Verified	debt_consolidation	IN	12.3	0	0	56		18	0	19040	30	f	1957.24	0	0	0	0		INDIVIDUAL		113th week	0	0	350619
8089625	5000	36 months		7.9	A	A4	TOYOTA OF NORTH HOLLYWOOD	5 years	RENT	75000	Source Verified	debt_consolidation	CA	5.7	0	0		105	13	2	13272	23	f	578.36	0	0	0	0		INDIVIDUAL		117th week	0	1023	13272
23043116	6000	36 months		9.17	B	B1	Banker	8 years	MORTGAGE	54000	Not Verified	credit_card	AL	11.63	0	1	46		13	0	3484	49	f	637.51	0	0	0	0	54	INDIVIDUAL		78th week	0	0	272579
45900933	6000	36 months	BAT4136152	13.99	C	C4	LVN	7 years	MORTGAGE	92000	Not Verified	home_improvement	CA	30.85	0	0	77		16	0	47567	27	w	621.72	0	0	0	0		INDIVIDUAL		44th week	0	0	281521

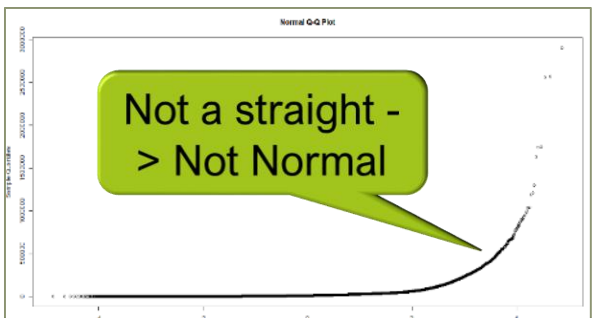
Target Variable [Total Revolving Balance]

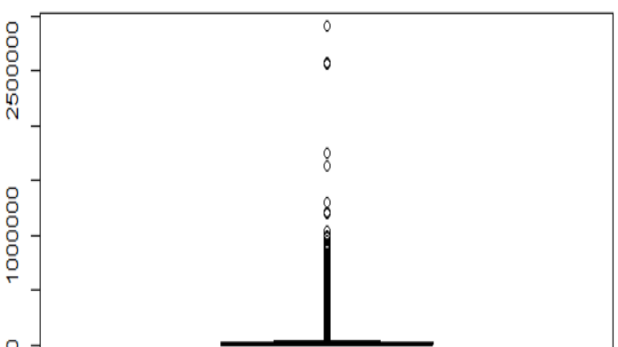
‘AS IS’ Histogram



X Axis limited to 100000







- Data is Right (Positive) Skewed – Long tail at Right
- X Axis limited to 100000 to show the shape of data

```
total.revol_bal
Min.    :      0
1st Qu.:   6443
Median :  11875
Mean   :  16921
3rd Qu.:  20829
Max.   :2904836
```

Mode < Median < Mean

- Data Not Normal
- Outliers present

Irrelevant Predictor Features

EXCELR Raising Excellence

Irrelevant Features for Target Variable

Member ID and Batch ID are unique Identification number for the member and batch. Logically not relevant data impacting the Revolving Balance

Member ID and Batch ID will be dropped prior to Data Analysis

Missing Values

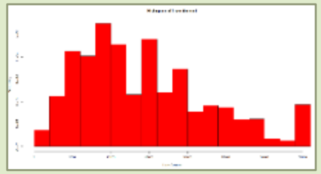
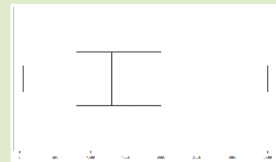
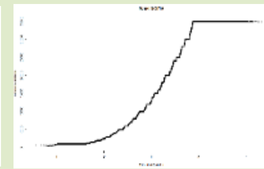
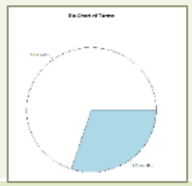
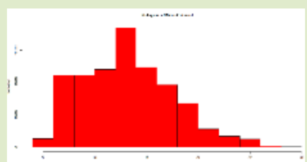
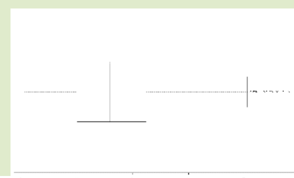
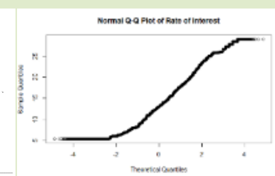
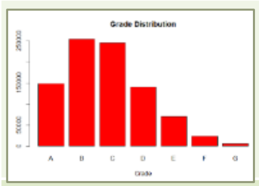
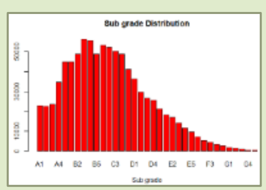
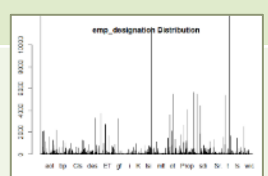
loan_amnt	terms	Rate_of_intrst	grade	sub_grade	Emp_designation
0	0	0	0	0	51457
Experience	home_ownership	annual_inc	verification_status	purpose	State
0	0	4	0	0	0
debt_income_ratio	delinq_2yrs	inq_last_6mths	mths_since_last_delinq	mths_since_last_record	numb_credit
0	29	29	454312 [51%]	750326 [81%]	29
pub_rec	total.revol_bal	total_credits	initial_list_status	total_rec_int	total_rec_late_fee
29	0	29	0	0	0
recoveries	collection_recovery_fee	collections_12_mths_ex_med	mths_since_last_major_derog	application_type	verification_status_joint
0	0	145	665676 [75%]	0	886868 [99.9%]
last_week_pay	acc_now_delinq	tot_colle_amt	tot_curr_bal		
0	29	70276	70276		

Features with very large missing Values [>75%] .To be dropped from Analysis because these features may skew the analysis prediction

Predictor Variables

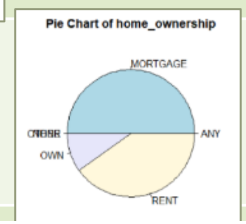
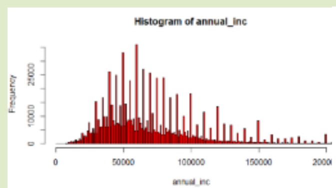
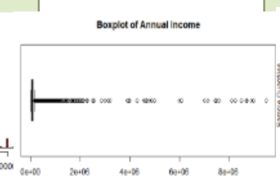
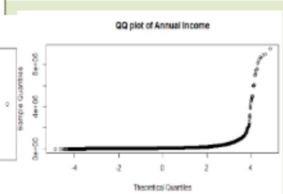
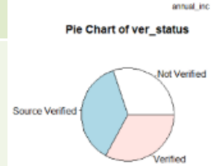
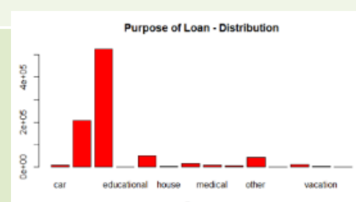
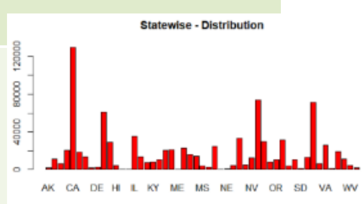
Numeric [Continuous and Integer]
Categorical [Factor]

EXCELR
Raising Excellence

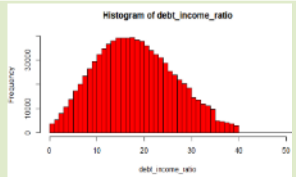
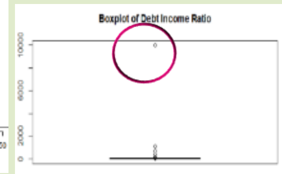
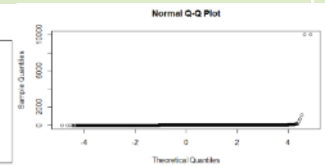
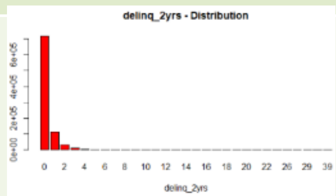
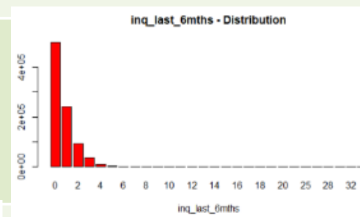
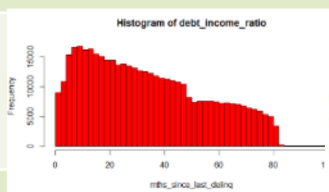
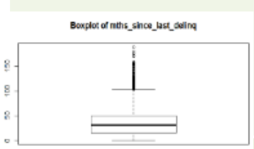
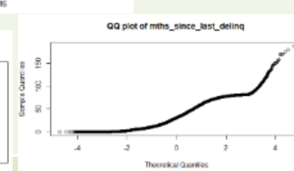
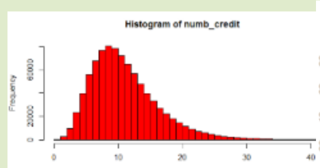
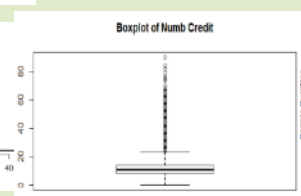
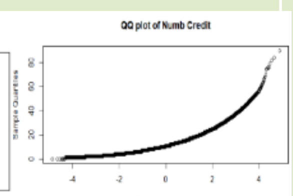
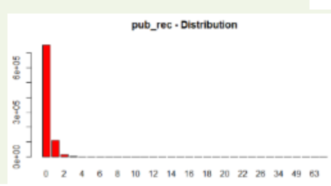
Predictor	Data Type	Graphs	Inference
Loan Amount	Integer	  	• Positively [Right] Skewed. • Not Normal. Limits at 35000
Terms	Factor		• Two levels 36, 60 months • Unequal proportions
Rate of Interests	Numeric	  	• Positively [Right] Skewed. • Not Normal
Grade	Factor		• A to G Levels • Not uniform
Subgrade	Factor		• Each grade sub grouped into 5 [A1,A2, A3,A4,A5] • use one feature out of Grade / subgrade
Employee Designation	Factor		• No inference

Predictor Variables

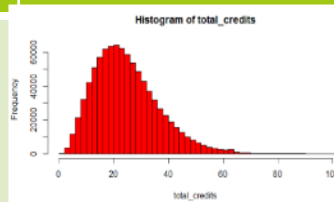
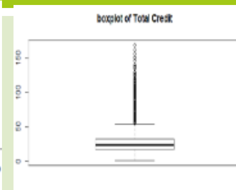
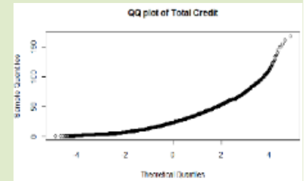
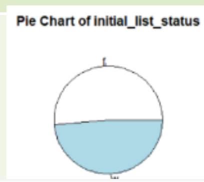
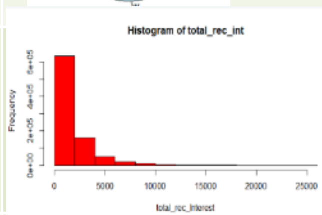
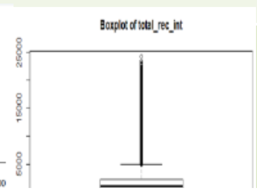
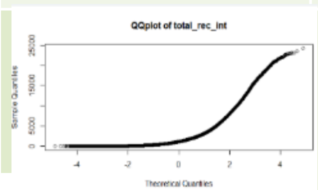
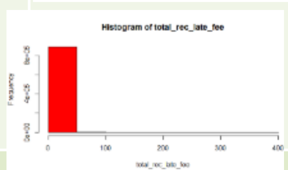
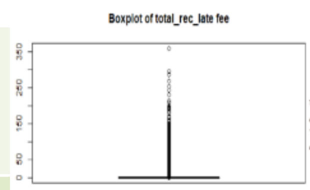
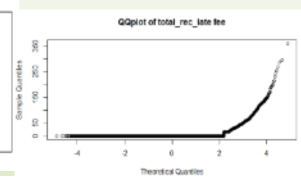
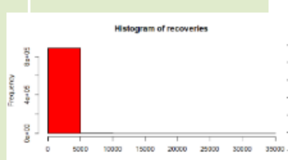
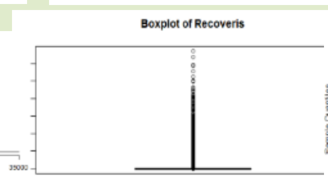
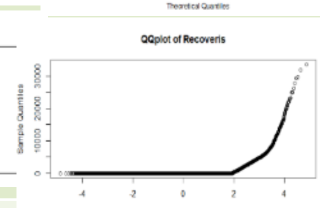
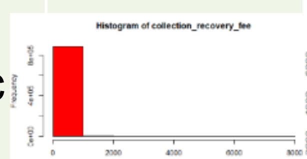
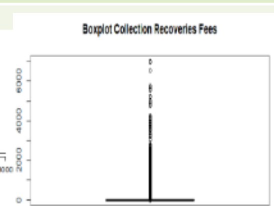
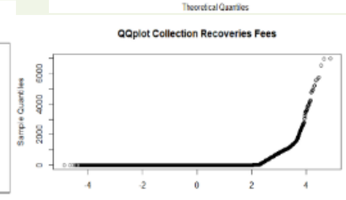
EXCELR
Raising Excellence

Predictor	Data Type	Graphs	Inference
Experience	Factor		• +10 years - peak
Home Ownership	Factor		• @ 50% Mortgage
Annual Income	Numeric	  	• Right Skewed – more people with lower Annual Income • Outliers • Not Normal
Verification Status	Factor		• 3 Levels • Equal Distribution
Purpose	Factor		• 14 levels • Credit Card and Debit card consolidation significant
State	Factor		• States from the USA • California has highest people in Revolving balance

Predictor Variables

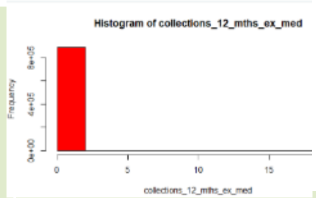
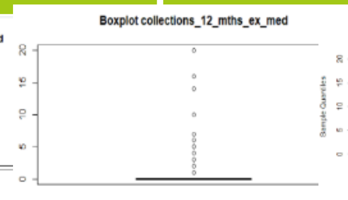
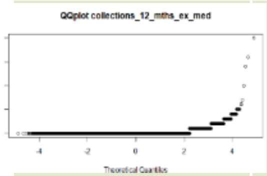
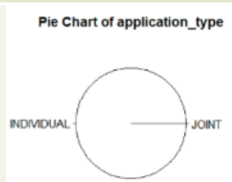
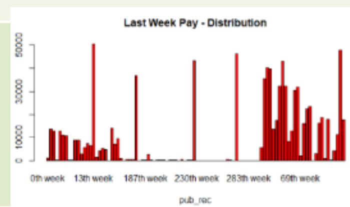
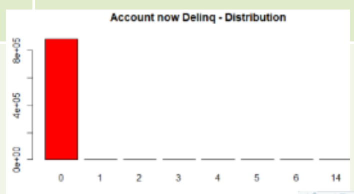
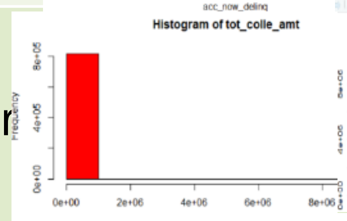
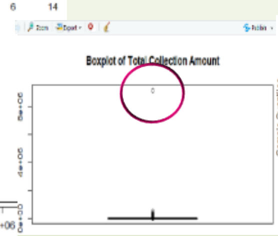
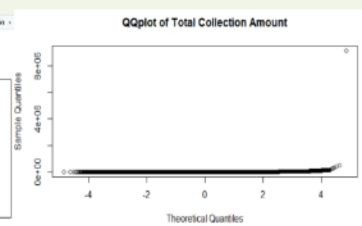
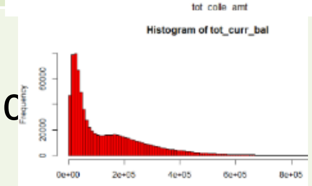
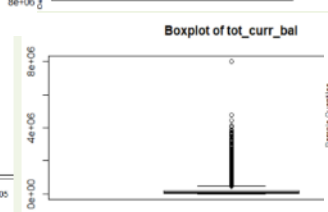
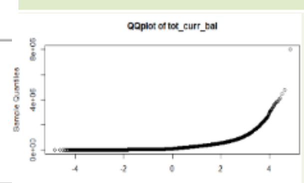
Predictor	Data Type	Graphs	Inference
Debt Income Ratio	Numeric	  	• Distinct outliers • May become Normal after removing outliers
Delinq 2 yrs	Integer		• Right Skewed Data
Inq. last 6 months	Integer		• Right Skewed
Months since last delinq	Numeric	  	• Not uniform
Numb Credit	Numeric	  	• Right skewed
Pub Record	Integer		• Right Skewed

Predictor Variables

Predictor	Data Type	Graphs	Inference
Total Credit	Integer	  	• Right Skewed • Not Normal
Initial List Status	Factor		• 2 Levels • Equal Distribution
Total Rec Interest	Numeric	  	• Right Skewed • Not Normal
Total Rec Late Fees	Numeric	  	
Recoveries	Numeric	  	• These 3 Distributions look similar • Need to check Collinearity
Collection Recovery fee	Numeric	  	

Predictor Variables

EXCELR
Raising Excellence

Predictor	Data Type	Graphs	Inference
Collection 12 months med	Numeric	  	• Distinct outliers • Not Normal
Application Type	Factor	 INDIVIDUAL JOINT 886868 [99.9%] 511[0.1%]	• Unequal Distribution • Need further analysis for Feature selection
Last Week Pay	Factor		• No inference
Acc now Delinq	Integer		• 8 Levels • Skewed to right • 99.5% data at 0 level
Total Collection Amount	Integer	  	• Collection Amount • Right Skewed • Distinct Outliers
Total Current Balance	Numeric	  	• Right Skewed • Distinct Outliers

Correlation Matrix to Reduce Features

	loan_amnt	terms	Rate_of_intrst	grade	sub_grade	Experience	home_ownership	verification_status	purpose	State	debt_income_ratio	total.revol_bal	initial_list_status	total_rec_int	total_rec_late_fee	recoveries	collection_recovery_fee	application_type
loan_amnt	1.000	0.412	0.145	0.151	0.157	-0.050	-0.196	0.281	-0.157	0.016	0.021	0.334	0.086	0.534	0.031	0.073	0.052	0.013
terms	0.412	1.000	0.428	0.443	0.452	-0.029	-0.111	0.168	-0.055	0.025	0.051	0.091	0.132	0.383	0.005	0.057	0.036	0.012
Rate_of_intrst	0.145	0.428	1.000	0.954	0.977	0.010	0.063	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0.115	0.446	0.057	0.107	0.071	0.011
grade	0.151	0.443	0.954	1.000	0.977	0.005	0.062	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0.073	0.377	0.053	0.091	0.064	0.014
sub_grade	0.157	0.452	0.977	0.977	1.000	0.005	0.065	0.242	0.155	0.007	0.086	-0.029	-0.068	0.388	0.054	0.094	0.066	0.014
Experience	-0.050	-0.029	0.010	0.005	0.005	1.000	-0.009	0.084	0.005	-0.010	0.016	-0.037	-0.011	-0.018	-0.005	0.002	-0.001	0.003
home_ownership	-0.196	-0.111	0.063	0.062	0.065	-0.009	1.000	-0.029	0.033	-0.068	0.001	-0.160	-0.032	-0.098	0.003	-0.004	-0.005	-0.009
verification_status	0.281	0.168	0.252	0.229	0.242	0.084	-0.029	1.000	0.011	-0.004	0.044	0.091	-0.031	0.273	0.017	0.052	0.033	0.008
purpose	-0.157	-0.055	0.150	0.151	0.155	0.005	0.033	0.011	1.000	-0.006	-0.046	-0.074	-0.074	-0.032	0.023	0.016	0.011	-0.002
State	0.016	0.025	0.006	0.007	0.007	-0.010	-0.068	-0.004	-0.006	1.000	0.022	-0.001	0.010	0.012	0.001	-0.003	-0.002	0.001
debt_income_ratio	0.021	0.051	0.080	0.084	0.086	0.016	0.001	0.044	-0.046	0.022	1.000	0.067	0.024	0.008	-0.006	0.001	0.002	0.074
total.revol_bal	0.334	0.091	-0.036	-0.030	-0.029	-0.037	-0.160	0.091	-0.074	-0.001	0.067	1.000	0.039	0.137	0.003	0.011	0.008	-0.001
initial_list_status	0.086	0.132	-0.115	-0.073	-0.068	-0.011	-0.032	-0.031	[illegible]	0.010	0.024	0.039	1.000	-0.157	-0.039	-0.059	-0.042	0.011
total_rec_int	0.534	0.383	0.446	0.377	0.388	-0.018	-0.098	0.273	[illegible]	0.012	0.008	0.137	-0.157	1.000	0.090	0.068	0.052	-0.017
total_rec_late_fee	0.031	0.005	0.057	0.053	0.054	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0.006	0.003	-0.039	0.090	1.000	0.074	0.068	-0.002
recoveries	0.073	0.057	0.107	0.091	0.094	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.074	1.000	0.802	-0.003
collection_recovery_fee	0.052	0.036	0.071	0.064	0.066	-0.001	-0.005	0.033	0.011	-0.002	[illegible]	[illegible]	[illegible]	[illegible]	0.068	0.802	1.000	-0.002
application_type	0.013	0.012	0.011	0.014	0.014	0.003	-0.009	0.008	-0.002	0.001	0.074	-0.001	0.011	-0.017	-0.002	-0.003	-0.002	1.000

Collinearity among Predictor Variables

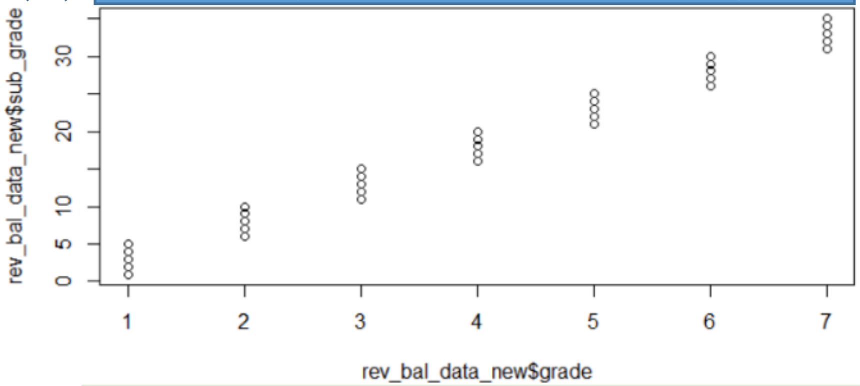
Target Variable

Collinearity among Predictor Variables

‘3’ Variables have similar correlation with rest of the variable

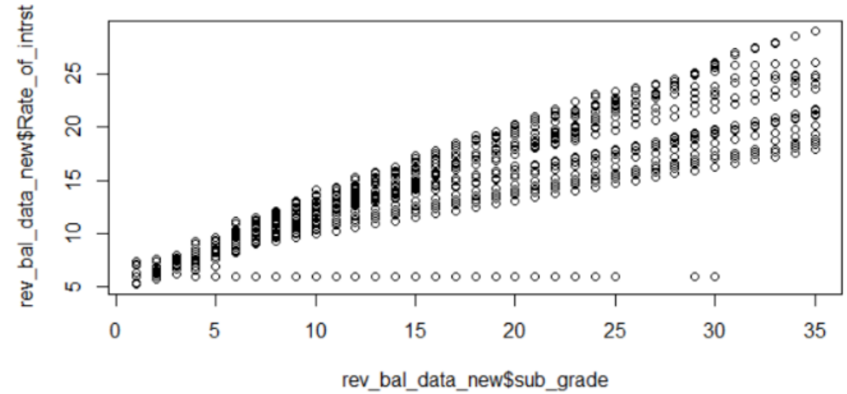
Correlation to check Collinearity

Collinearity between Grade and Subgrade



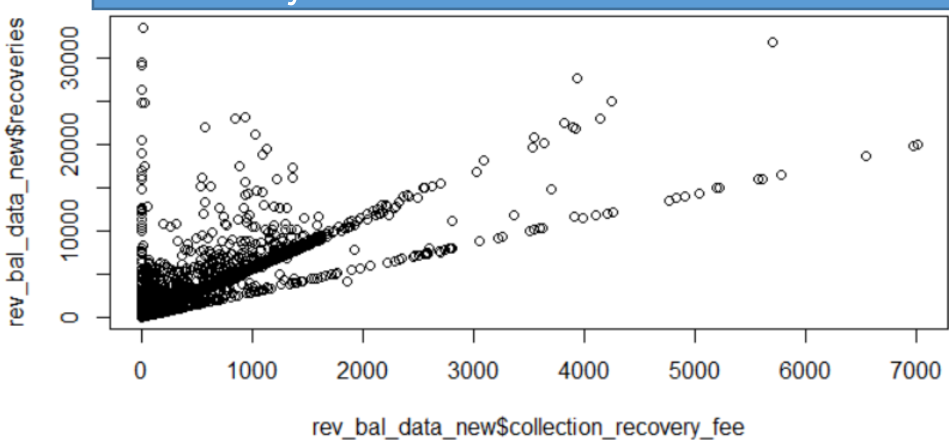
Correlation Coefficient r = 0.976
Collinearity between predictors
Will Drop Grade, as the subgrade has higher resolution

Collinearity between Subgrade and Rate of Interest



Correlation Coefficient r = 0.977
Collinearity
Will drop Subgrade, and keep Rate of Interest as Rate of Interest is numeric as Subgrade is Factor

Collinearity between Recoveries and Rec Late fee



Correlation Coefficient r = 0.8024196
From plot it is NOT evident correlation. Will decide later

Simple Linear Regression of each predictor

EXCELR Raising Excellence

Slide

Predictor Variable	R2 Sqaured on Target
tot_curr_bal	0.1954
loan_amnt	0.1113
annual_inc	0.0875
numb_credit	0.0504
total_credits	0.0358
home_ownership	0.0255
total_rec_int	0.0189
pub_rec	0.0101
terms	0.0083
verification_status	0.0082
purpose	0.0055
debt_income_ratio	0.0045
initial_list_status	0.0015
Experience	0.0013
Rate_of_intrst	0.0013
delinq_2yrs	0.0011

Also have Collinearity

Also have >75% Missing Values

Predictor doesn't have significant effect on Target

Predictor Variable	R2 Sqaured on Target
grade	0.0009
sub_grade	0.0009
mths_since_last_delinq	0.0007
mths_since_last_record	0.0007
collections_12_mths_ex_med	0.0005
Emp_designation	0.0004
mths_since_last_major_derog	0.0004
inq_last_6mths	0.0003
recoveries	0.0001
collection_recovery_fee	0.0001
tot_colle_amt	0.0000
total_rec_late_fee	0.0000
last_week_pay	0.0000
application_type	0.0000
State	0.0000
verification_status_joint	0.0000
acc_now_delinq	0.0000

Results of Simple Linear Regression of Target Variable with each individual predictor variable

Main Effect of predictor on to Target

May have Interactions effects alongwith other predictors

Insignificant Features based on Regression

EXCELR
Raising Excellence

Main Effect 1:1 Regression Lowest Effect

Predictor Variable	R2 Sqaured on Target
grade	0.0009
sub_grade	0.0009
mths_since_last_delinq	0.0007
mths_since_last_record	0.0007
collections_12_mths_ex_med	0.0005
Emp_designation	0.0004
mths_since_last_major_derog	0.0004
inq_last_6mths	0.0003
recoveries	0.0001
collection_recovery_fee	0.0001
tot_colle_amt	0.0000
total_rec_late_fee	0.0000
last_week_pay	0.0000
application_type	0.0000
State	0.0000
verification_status_joint	0.0000
acc_now_delinq	0.0000

Based on poor Main and Interactive effect on Target can be eliminated

MLR – Interactions

Coefficients:	t value	Pr(>|t|)
tot_curr_bal	39.999	2E-16
annual_inc	24.513	2E-16
debt_income_ratio	23.197	2E-16
numb_credit	19.513	2E-16
loan_amnt	16.919	2E-16
total_credits	-10.529	2E-16
home_ownership	7.47E+00	8.53E-14
mths_since_last_record	-6.72E+00	1.87E-11
Emp_designation	-3.408	0.000656
purpose	-3.06	0.002216
sub_grade	-2.877	0.004019
pub_rec	-2.824	0.004745
delinq_2yrs	-2.823	0.004766
tot_colle_amt	-2.493	0.012685
total_rec_int	2.076	0.037907
inq_last_6mths	-1.811	0.07011
mths_since_last_major_derog	1.758	0.078801
verification_status	1.404	0.160333
initial_list_status	-1.365	0.172123
verification_status_joint	-1.303	0.192607
Rate_of_intrst	1.278	0.201232
Experience	-1.259	0.207979
collection_recovery_fee	1.149	0.250704
mths_since_last_delinq	1.046	0.295346
last_week_pay	-1.019	0.308053
acc_now_delinq	-0.997	0.31879
total_rec_late_fee	0.995	0.319983
terms	-0.927	0.354114
State	0.919	0.357855
application_type	0.579	0.562534
(Intercept)	-0.502	0.615984
recoveries	-0.385	0.699895
collections_12_mths_ex_med	-0.242	0.809145
grade	-0.143	0.886478

P>0.05 No Interac tion

Important features based on Random Forest

EXCELR
Raising Excellence

Slide

Feature	%IncMSE	IncNodePurity
loan_amnt	4.958	5.33E+07
tot_curr_bal	4.081	6.89E+07
numb_credit	3.968	6.96E+07
Rate_of_intrst	3.586	6.86E+07
total_credits	2.802	6.09E+07
last_week_pay	2.474	3.09E+07
delinq_2yrs	1.803	4.62E+06
annual_inc	1.449	5.99E+07
purpose	1.386	1.67E+07
total_rec_int	0.973	3.94E+07
verification_status_joint	0.747	2.57E+07
sub_grade	0.652	6.19E+07
tot_colle_amt	0.596	6.50E+07
pub_rec	0.517	1.88E+07
grade	0.237	4.00E+07
total_rec_late_fee	0	0.00E+00
recoveries	0	0.00E+00
collection_recovery_fee	0	0.00E+00
collections_12_mths_ex_med	0	1.01E+06
application_type	0	0.00E+00
acc_now_delinq	0	0.00E+00
State	-0.267	3.71E+07
inq_last_6mths	-0.39	1.34E+07
Emp_designation	-0.464	8.05E+07
debt_income_ratio	-0.662	4.89E+07
mths_since_last_record	-8.86E-01	5.28E+07
mths_since_last_major_derog	-0.996	2.30E+07
initial_list_status	-1.319	7.42E+06
home_ownership	-1.40E+00	5.59E+06
verification_status	-2.225	1.79E+07
terms	-2.645	7.91E+06
Experience	-3.033	2.36E+07
mths_since_last_delinq	-3.507	3.27E+07

Missing Values > 75%
Collinear with Rate of Interest

Missing Values > 75%
Collinear with Rate of Interest

Poor Correlation with Target Variable

✓ Important Features from Correlation and Regression

Features to be dropped

	Feature	Justification
1	Member ID	Not relevant to Revolving Balance
2	Batch ID	
3	Month since last Record	>75% Missing Data
4	Month Since last Derog	
5	Verification Status Joint	
6	Acc now Delinq	Highly Imbalance Data [99.5% : 0.5%]
7	Grade	Collinearity with Rate of Interests Correlation Matrix [Heat Map]
8	Subgrade	

	Feature	Justification
9	collections_12_mths_ex_med	Poor correlation with Target variable as Main Effect And/or interaction effect
10	recoveries	
11	last_week_pay	
12	application_type	Imbalance Data and poor regression
13	State	Poor correlation with Target variable as Main Effect And/or interaction effect And Random Forest
14	total_rec_late_fee	
15	collection_recovery_fee	
16	Months Since Last Delinq	
17	terms	
18	Inq Last 6 months	

Features Selected for Model Building

	Feature	t Statistics	%Inc in MSE
1	tot_curr_bal	39.999	4.08
2	annual_inc	24.513	1.45
3	debt_income_ratio	23.197	-0.66
4	numb_credit	19.513	3.97
5	loan_amnt	16.919	4.96
6	total_credits	-10.529	2.80
7	home_ownership	7.470	--1.40
8	Emp_designation	-3.408	0.46
9	tot_colle_amt	-2.493	0.60
10	total_rec_int	2.076	0.97
11	verification_status	1.404	2.23
12	initial_list_status	-1.365	1.32
13	Rate_of_intrst	1.278	3.59
14	Delinq 2 Years	-2.82	1.80
15	Experience	1.26	-3.03
16	Purpose	-3.06	139
17	Pub Rec	-2.80	0.52

Data Preprocessing –

- Missing Value Imputation
- Outlier Treatment
- Data Transformation into Normal
- Data Scaling

80% RAW Data [titles as ‘Development’ Data used for Data Preprocessing

Missing Values Treatment

EXCELR
Raising Excellence
Slide

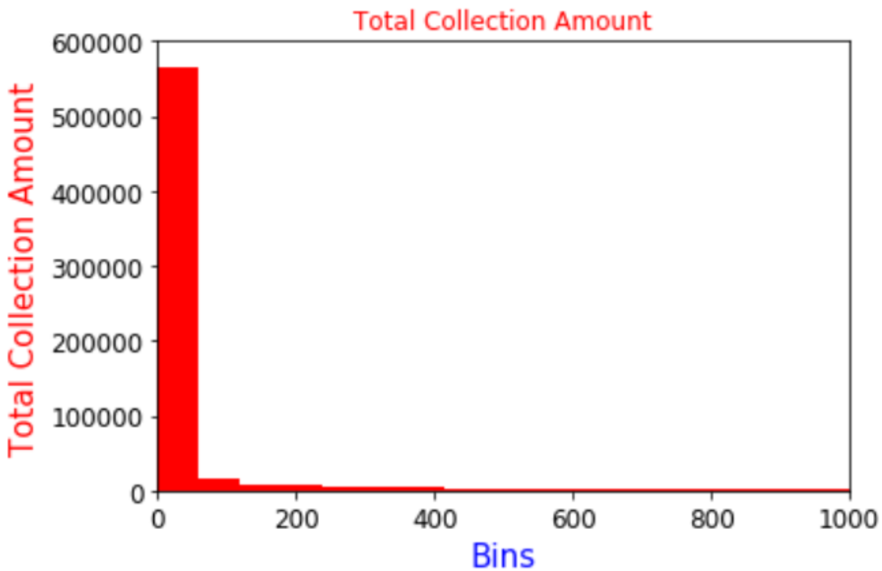
Feature	Data Type	Missing Values [Count]	% Missing Values
Loan_amt	int64	0	0.0%
rate_of_int	float64	0	0.0%
emp_designation	object	41092	4.6%
experience	object	35807	4.0%
home_ownership	object	0	0.0%
annual_inc	float64	2	0.0%
verification_status	object	0	0.0%
purpose	object	0	0.0%
debt_income_ratio	float64	0	0.0%
delinq_2_yrs	float64	22	0.0%
numb_credit	float64	22	0.0%
pub_rec	float64	22	0.0%
tot_revol_bal	float64	0	0.0%
tot_credits	float64	22	0%
initial_list_status	object	0	0.
tot_rec_int	float64	0	0.0%
tot_coll_amt	float64	56285	6.3%
tot_curr_bal	float64	56285	6.3%

Missing Values > 6%

- Used 80% - RAW Data
  [titled as ‘Development’ for Data Preprocessing
- Insignificant Features dropped
- Total Collection Amount and Total Current Balance have 6.3% Missing Values
- ‘4’ features have same missing values – for the same rows – These observations can be removed

Same observations
Will be deleted

Missing Values Treatment – Total Collection Amount



count	653911
mean	213.50
std	1849.06
min	0
25%	0
50%	0
75%	0
Max	296368
Missing Values	56000

Total Collection Amount :

- Highly Right skewed, >75% data have ‘0’ values
- If we impute for 6% missing values with a median, it further skew the data
- Decided to remove the feature

Missing Values Treatment – Actions Taken

Feature	Missing Values [Count]	% Missing Values	Actions on Missing Values
delinq_2_yrs	22	0.0%	Removed 22 observations
numb_credit	22	0.0%	
pub_rec	22	0.0%	
tot_credits	22	0.0%	
emp_designation	41083	4.6%	Imputation by Mode
experience	35807	4.0%	Imputation by Mode
tot_coll_amt	56263	6.3%	Removed feature
tot_curr_bal	56263	0.063403574	Imputation by Interpolation

Data Types change for Model Building

EXCELR
Raising Excellence
Slide

Feature	Feature
loan_amt	**int64**
rate_of_int	**float64**
emp_designation	**int64**
experience	**int32**
home_ownership	**int32**
annual_inc	**float64**
verification_status	**int32**
purpose	**int32**
debt_income_ratio	**float64**
delinq_2_yrs	**float64**
numb_credit	**int64**
pub_rec	**float64**
tot_revol_bal	**float64**
tot_credits	**float64**
initial_list_status	**int32**
tot_rec_int	**float64**
tot_curr_bal	**float64**

Transformed all data into Numeric –
Continuous variables retained as Float
Categorical variables transformed into Integers

Outlier Counts of Numeric Variables

EXCELR
Raising Excellence

Slide

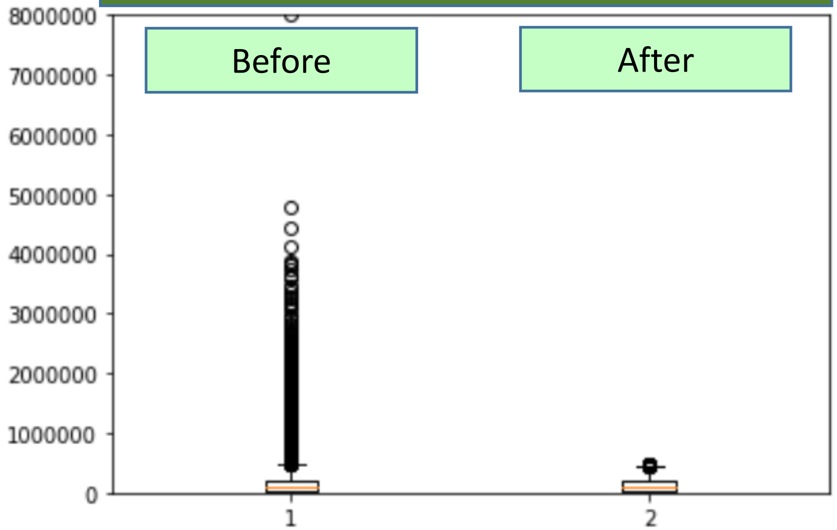
Numeric Feature	outliers count
loan_amt	0
rate_of_int	5025
annual_inc	31766
debt_income_ratio	65
delinq_2_yrs	136215
numb_credit	22018
pub_rec	108527
tot_credits	14720
tot_rec_int	51093
tot_curr_bal	23780

Used Interpolation to treat the outliers
If we use one single value – it may create bi-modal shape or skew the data

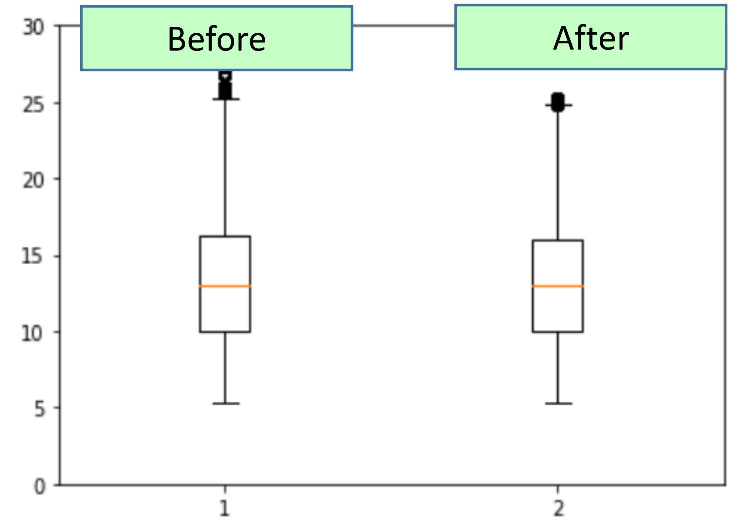
Question:
After imputation of outlier, another datapoints became outliers..
Is it an iterative process of Outliers Imputation ?

Outlier Treatment

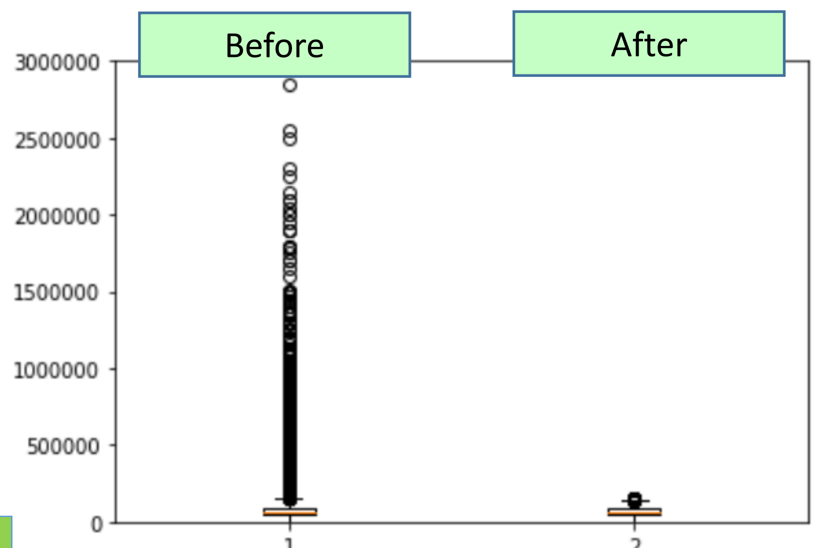
Total Current Balance



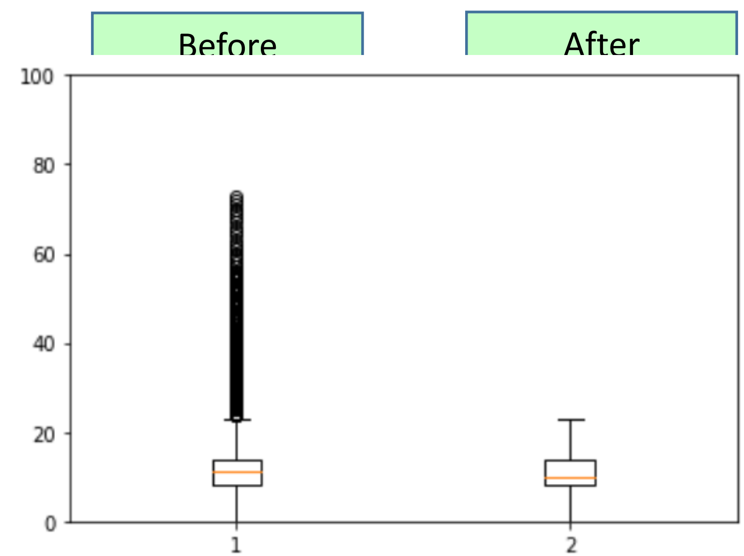
Rate of Interest



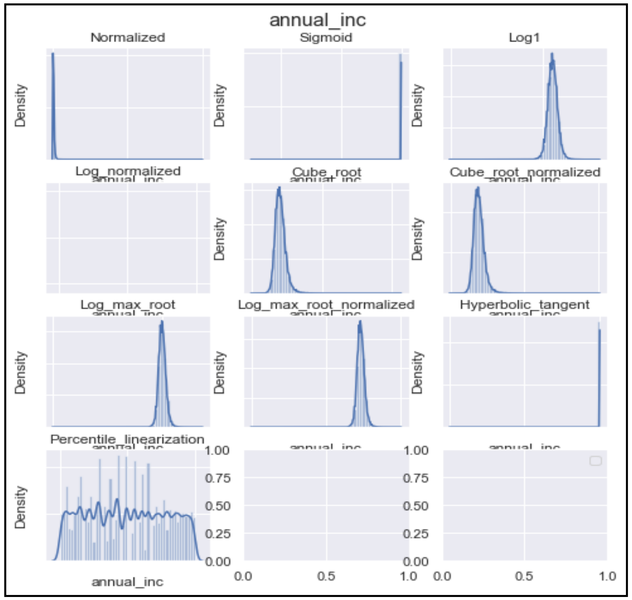
Annual Income

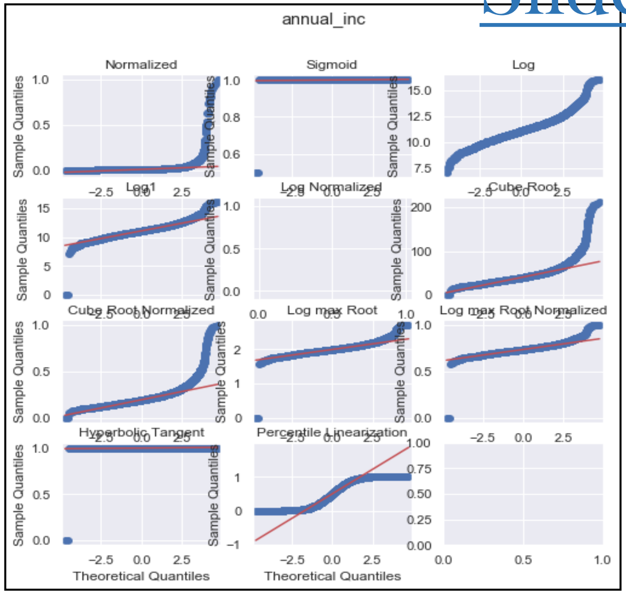


Numb Credit



Skewed to Normal - Annual Income





	Feature	Method	P-Value	Result
0	annual_inc	Normalized	0.0	reject H0
1	annual_inc	Sigmoid	0.0	reject H0
2	annual_inc	Log	1.0	fail to reject H0
3	annual_inc	Log+1	0.0	reject H0
4	annual_inc	Log Normalized	1.0	fail to reject H0
5	annual_inc	Cube Root	0.0	reject H0
6	annual_inc	Cube Root Normalized	0.0	reject H0
7	annual_inc	Log Max Root	0.0	reject H0
8	annual_inc	Log Max Root Normalized	0.0	reject H0
9	annual_inc	Hyperbolic Tangent	0.0	reject H0
10	annual_inc	Percentile Linearization	0.0	reject H0

For entire data set Shapiro test showed p value as 1 with a warning of accuracy.
Checked on a sample pf 5000, P value was close to 0.
Also checked with Kolmogorov-Smirnov test : Data transformation not feasible

Data Scaling - Normalization

loan_amnt	rate_of_int	emp_designation	experience	home_ownership	annual_inc	verification_status	purpose	debt_income_ratio	delinq_2_yrs	numb_credit	pub_rec	tot_credits	initial_list_status	tot_rec_int	tot_curr_bal
0.401	0.586	0.032	0.9	0.8	0.003	0.5	0.154	0.003	0	0.192	0.012	0.168	0	0.048	0.004
0.125	0.240	0.406	1	0.2	0.007	0.5	0.308	0.000	0	0.082	0.000	0.075	1	0.003	0.001
0.275	0.082	0.380	0.2	0.8	0.005	0	0.154	0.002	0	0.068	0.000	0.112	1	0.002	0.005
0.420	0.608	0.658	0.1	1	0.012	0	0.154	0.001	0	0.137	0.000	0.124	0	0.201	0.007
0.449	0.225	0.417	0.1	1	0.006	1	0.077	0.002	0	0.151	0.000	0.161	1	0.095	0.006
0.420	0.151	0.374	0.2	0.2	0.013	0	0.154	0.001	0	0.247	0.000	0.180	0	0.081	0.044
0.130	0.109	0.399	0.5	1	0.008	0.5	0.154	0.001	0	0.178	0.023	0.137	0	0.024	0.002
0.159	0.163	0.473	0.8	0.2	0.006	0	0.077	0.001	0	0.178	0.000	0.298	0	0.026	0.034
0.159	0.366	0.417	0.7	0.2	0.010	0	0.308	0.003	0	0.219	0.000	0.161	1	0.026	0.035
0.987	0.499	0.045	0.2	0.2	0.008	1	0.154	0.003	0	0.164	0.000	0.180	1	0.229	0.010

$$\text{Scaled Value [Normalisation]} = \frac{\text{Actual Value} - \text{Minimum Value}}{\text{Maximum Value} - \text{Minimum Value}}$$

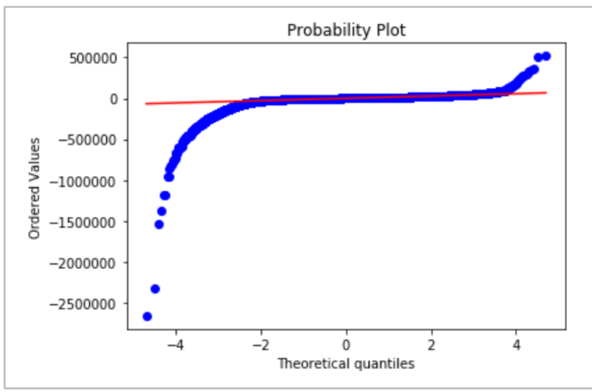
Standardisation should not be done since data does not follow Gaussian's distribution

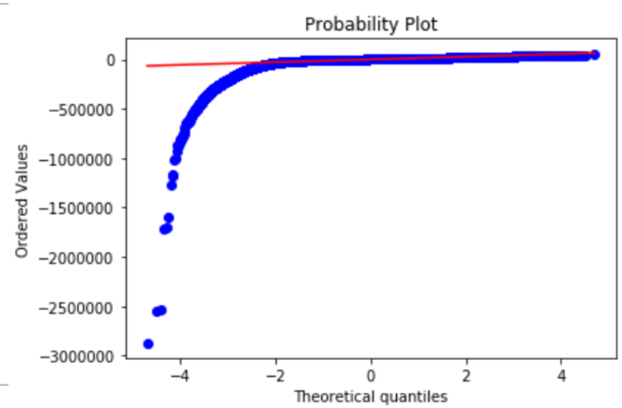
$$\text{Scaled Value [Standardised]} = \frac{\text{Actual Value} - \text{Mean}}{\text{Standard Deviation}}$$

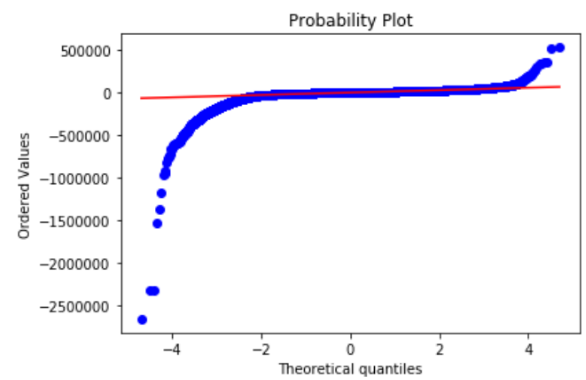
Regression Model Comparison of Data

Original Data Sets Missing Values imputed	Outliers Removed	Data Scaled to 0 to 1
Adj. R-squared: 0.257	**Adj. R-squared:** 0.167	**Adj. R-squared:** 0.254

Residual Plots







RMSE for the training data **19435**	RMSE for the training data **21020**	RMSE for the training data **19675**

Linear Model is not a GOOD fit since Residuals are not NORMAL →
Try non linear models
Data transformation look comparable

Model Building
Train model [60% Data]
Test Model [20% Data]
Validate [20% Data]
Kaggle Results

Multi Linear Regressionn Models
Logy ~ Log(cont. x) + (categorical x)

EXCELR
Raising Excellence
Slide

Log transformation was a better since data for all continuous parameters were Right Skewed

	Features used
1	tot_curr_bal
2	annual_inc
3	debt_income_ratio
4	numb_credit
5	loan_amnt
6	total_credits
7	home_ownership

	Features used
8	Emp_designation
9	total_rec_int
10	verification_status
11	initial_list_status
12	Rate_of_intrst
13	Experience
14	Purpose

RMSE for the training data 20550

RMSE for the testing data 20800

RMSE for the Validation data 20255

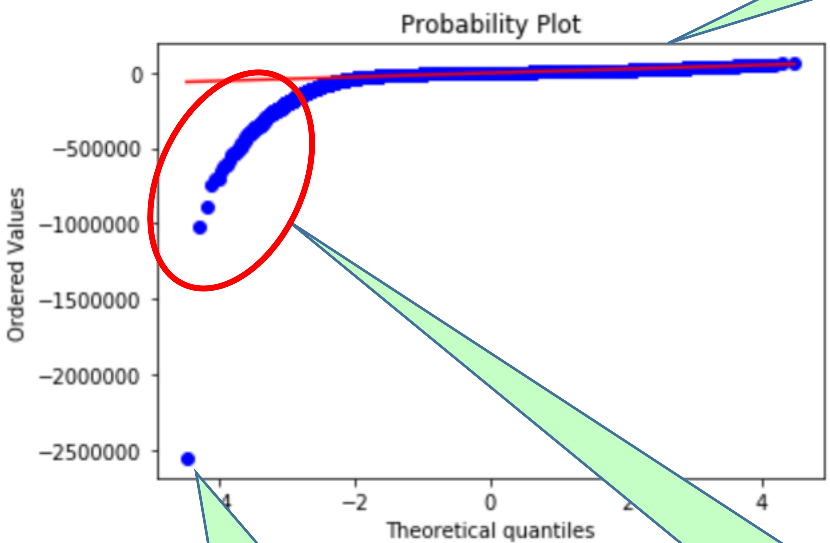
On Kaggle: 20540

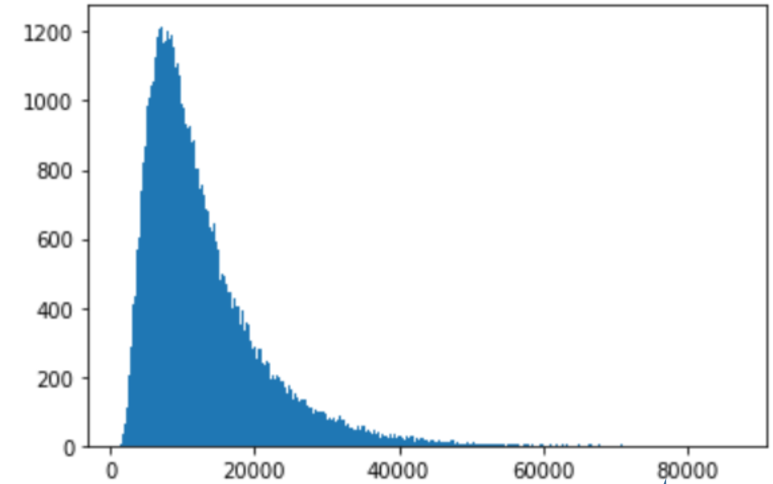
V

Prediction from MLR

Slide

Prediction good in this range where residuals are close to zero





This error probably aggravated error

Residuals start deviating from normality curve

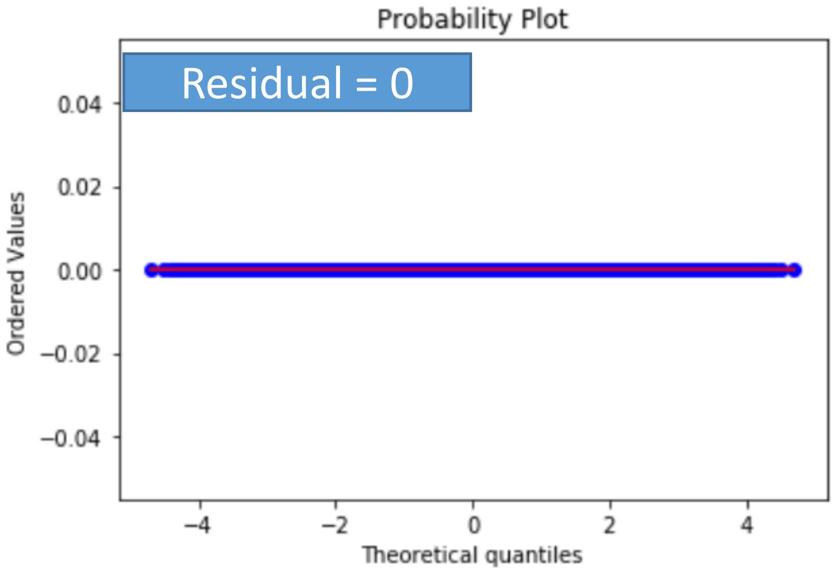
MLR Failed to predict higher credit revolving balances

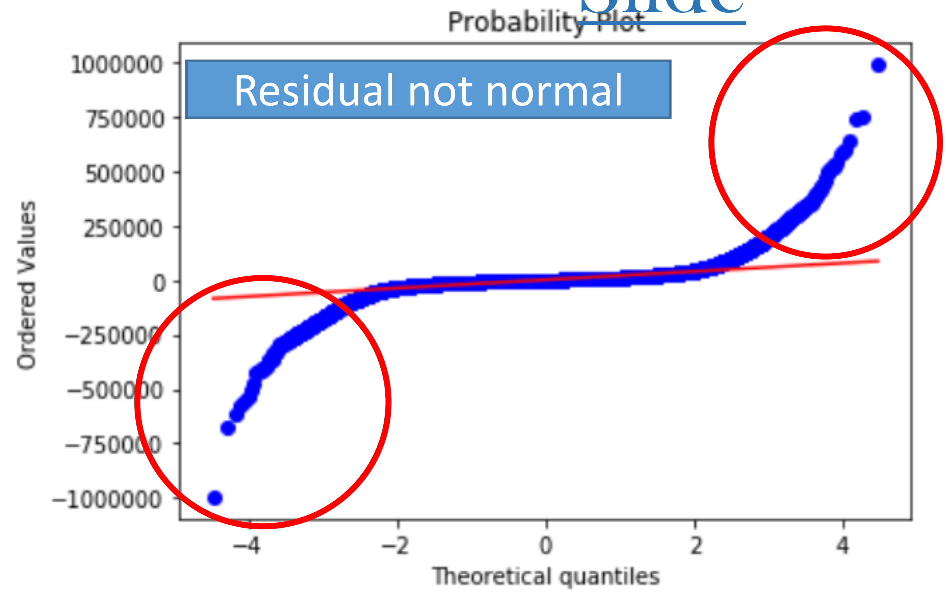
V

Decision Tree Models on Training Data

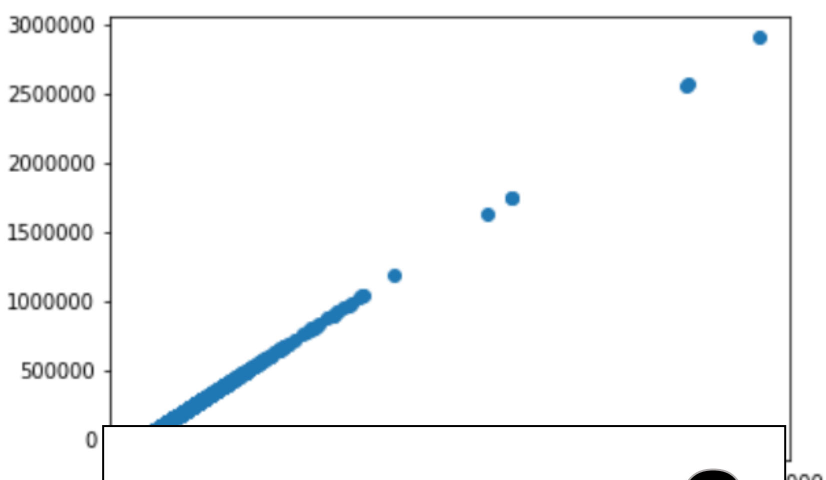
EXCELR
Raising Excellence

Slide



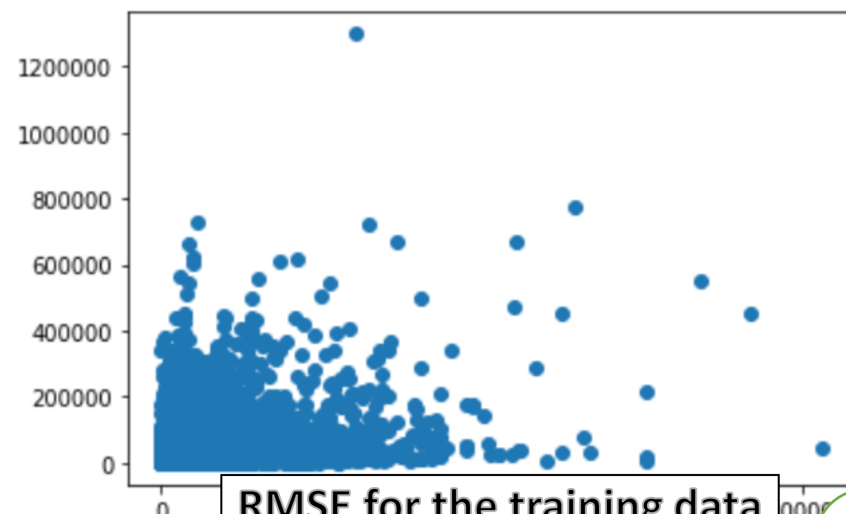


Predicted Y Vs Actual Y



RMSE for the training data **0**

Predicted Y Vs Actual Y



RMSE for the training data
24675

V

Random Forest and Ridge

Model	No. of features	Train RMSE	Kaggle test RMSE
Random Forest	15	2988	20634
Ridge	15	7601	20492

Numerical Features (9) : loan amount, rate of interest, annual income, total credits, total received interest, debt to income ratio, numb credits, last week pay, total current balance

Categorical features(6): terms, grade, verification status, experience, home, ownership , state,

M

What Next ???

Slide

- **Regression failed to predict extremely skewed Credit Revolving Balance**
- **Decision Tree, Random Forest worked well on Training but failed to predict on Validation data**
- **Hence, we proceeded to XGBoost – an Ensemble Method** ***"Best of both worlds"***

V

XGBoost Model

EXCELR
Raising Excellence
Slide

- **We had 35 features, used 30% [0.3] as colsample [trees to build total model]**

- **Learning Rate 0.1 (typical range 0.1 – 0.3)**

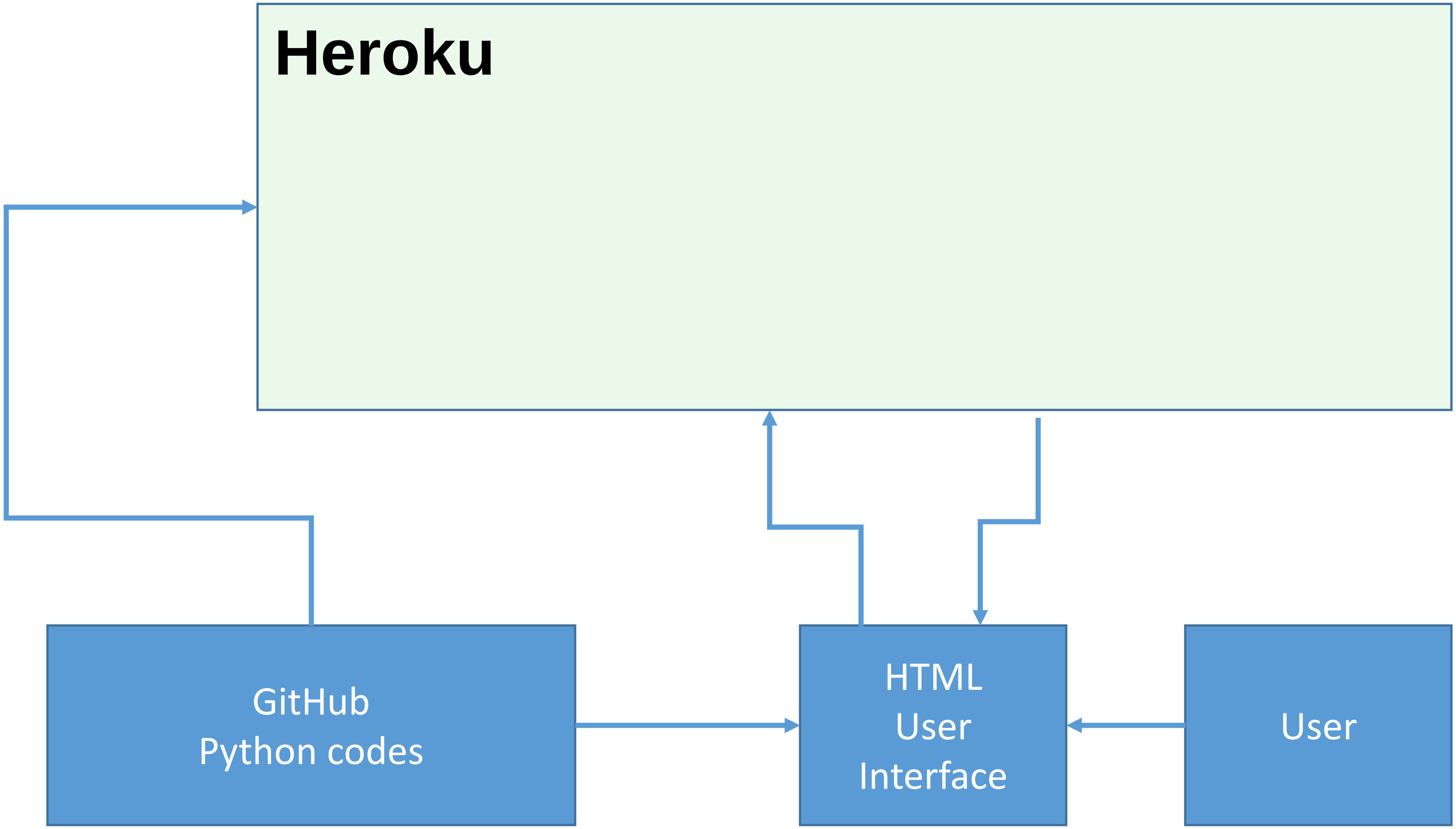
- **Max. Depth as 5**
- **Alpha as 10**

- *RMSE of Final Model : 18314*

H

Model Deployment

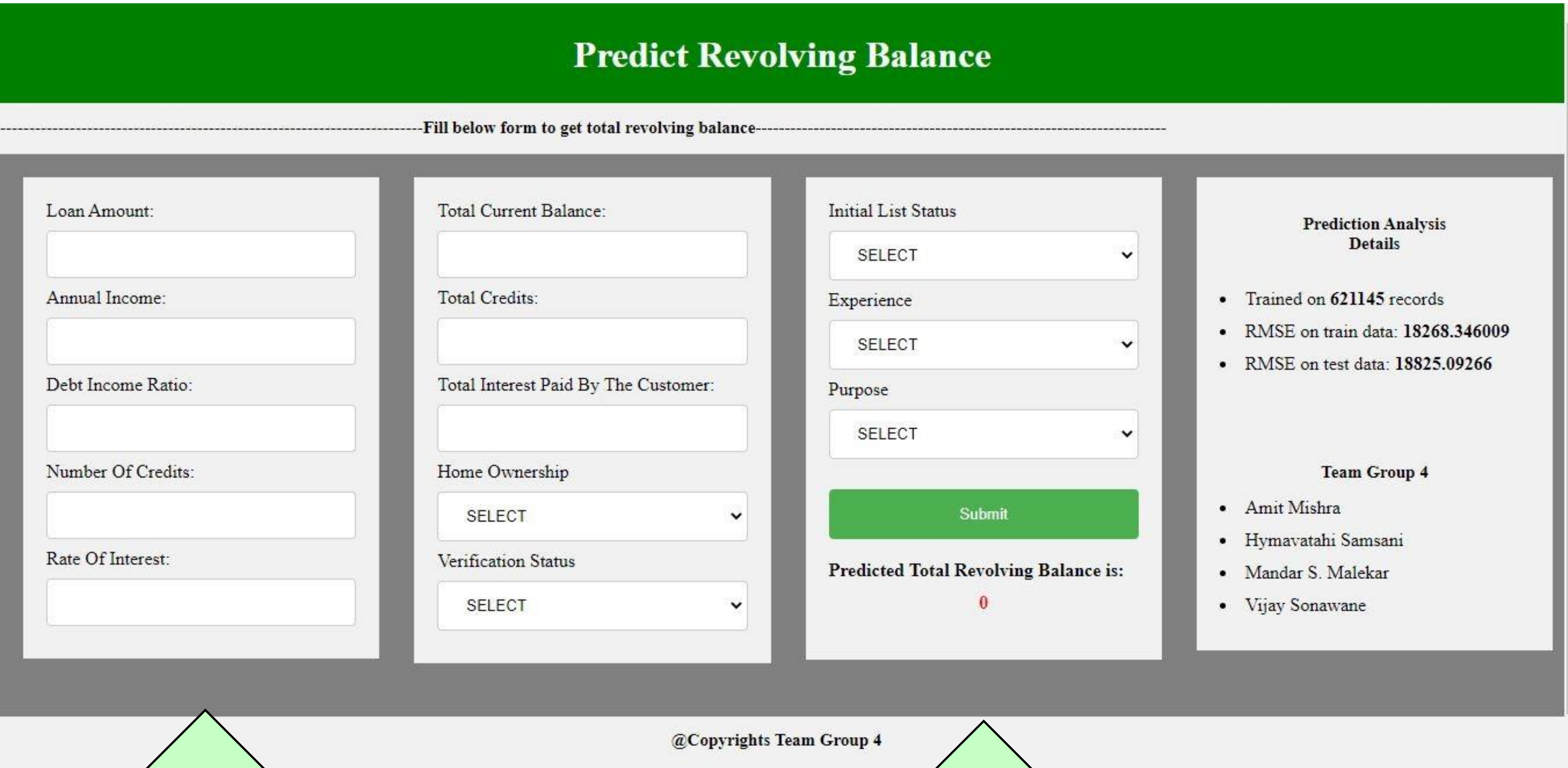
Approach for Model Deployment



H

GUI Interface for Prediction

Slide



Continuous Features tabs for entering values

Pull down selection menu for categorical features

H

Next Steps

EXCELR
Raising Excellence

Slide

Upload following on Kaggle :

- **Final Python Model**
- **Presentation as a Project documentation**
- **Deployment Model**

V

Lesson Learnt

EXCELR
Raising Excellence
Slide

- **EDA, Feature Engineering** are the crucial steps in defining the strategy for model building
- Challenging when the data is **right skewed**
- The difficulties aggravates when **correlation** of predictors with the target variable is very **poor**
- Advanced algorithms of Machine Learning such as **XGBoost, Neural Network** are helpful over traditional prediction models such as MLR or Decision tree

V

We take the opportunity to thank our Mentors Sri Vinod and Ms Munmun