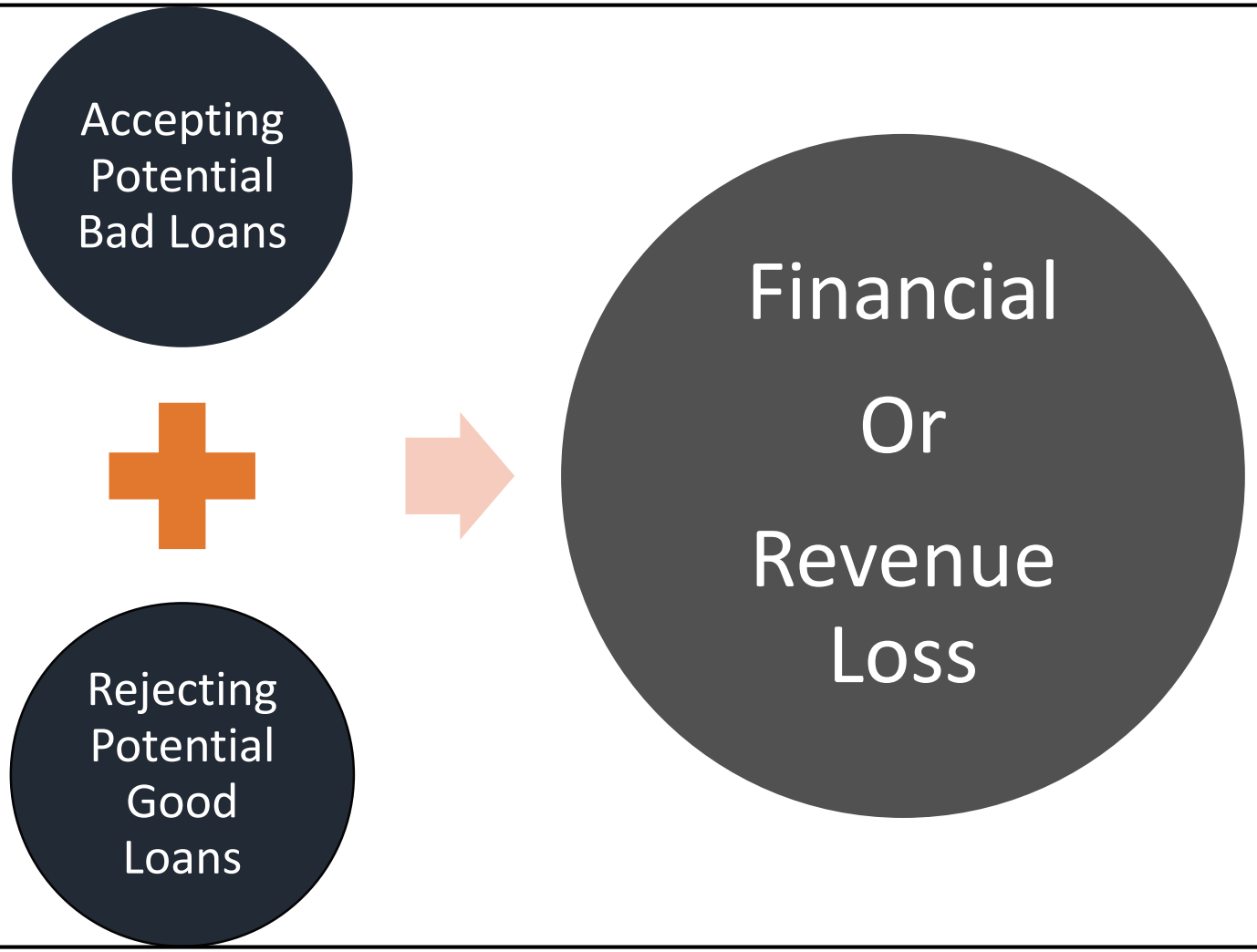
EDA - Gramener Case Study

AGENDA

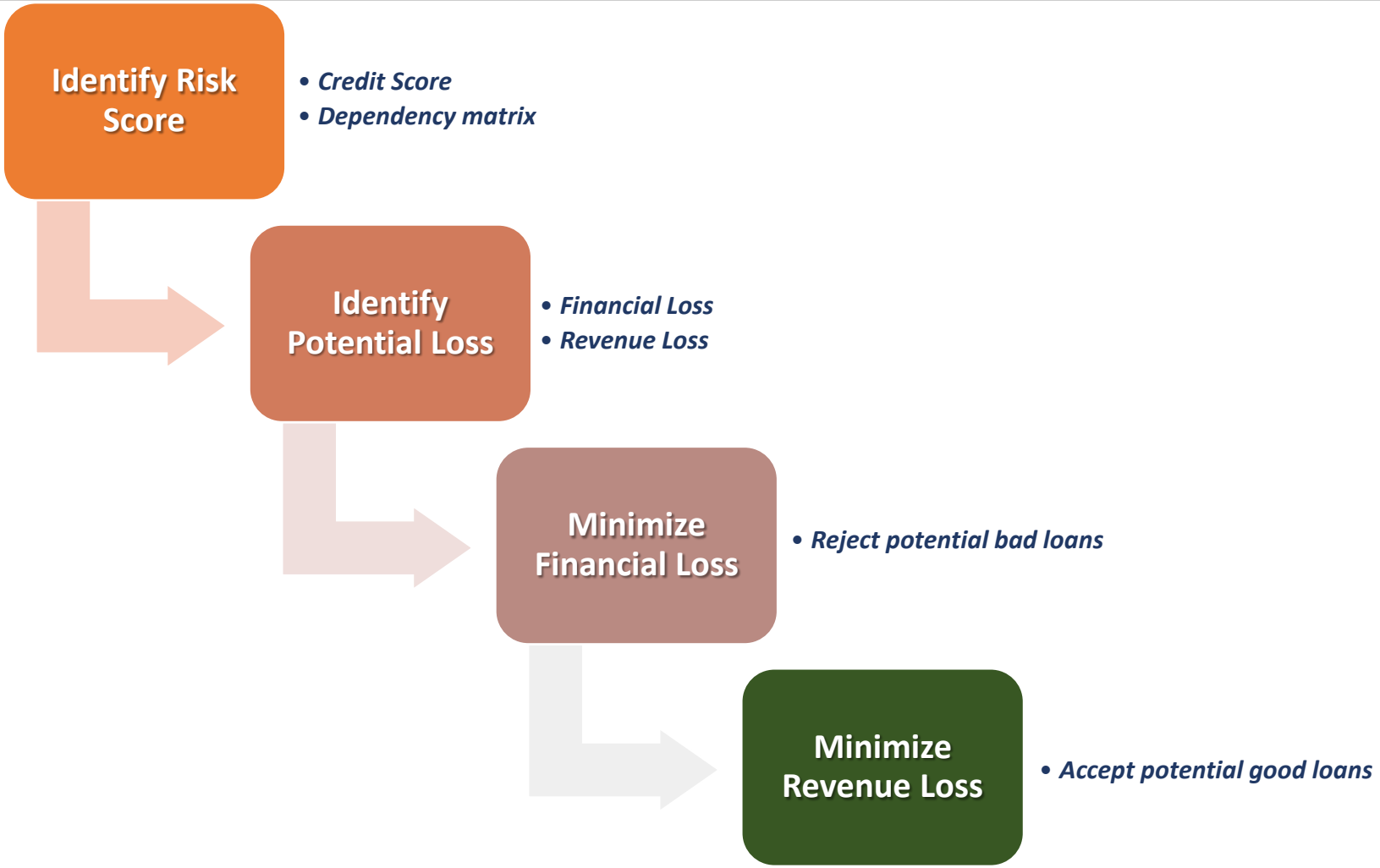
- BUSINESS PROBLEM
- BUSINESS OBJECTIVES
- ANALYTICS MODEL DEVELOPMENT STEPS
- DEVELOPING MODEL
- EXPECTED OUTCOME AND INFERENCE
- FUTURE FINE TUNING – BASED ON FEEDBACK

Presented By:
Subhra Sinha
Sathish Karthik Ramu
Saathvik K V
Amitava Mukherjee

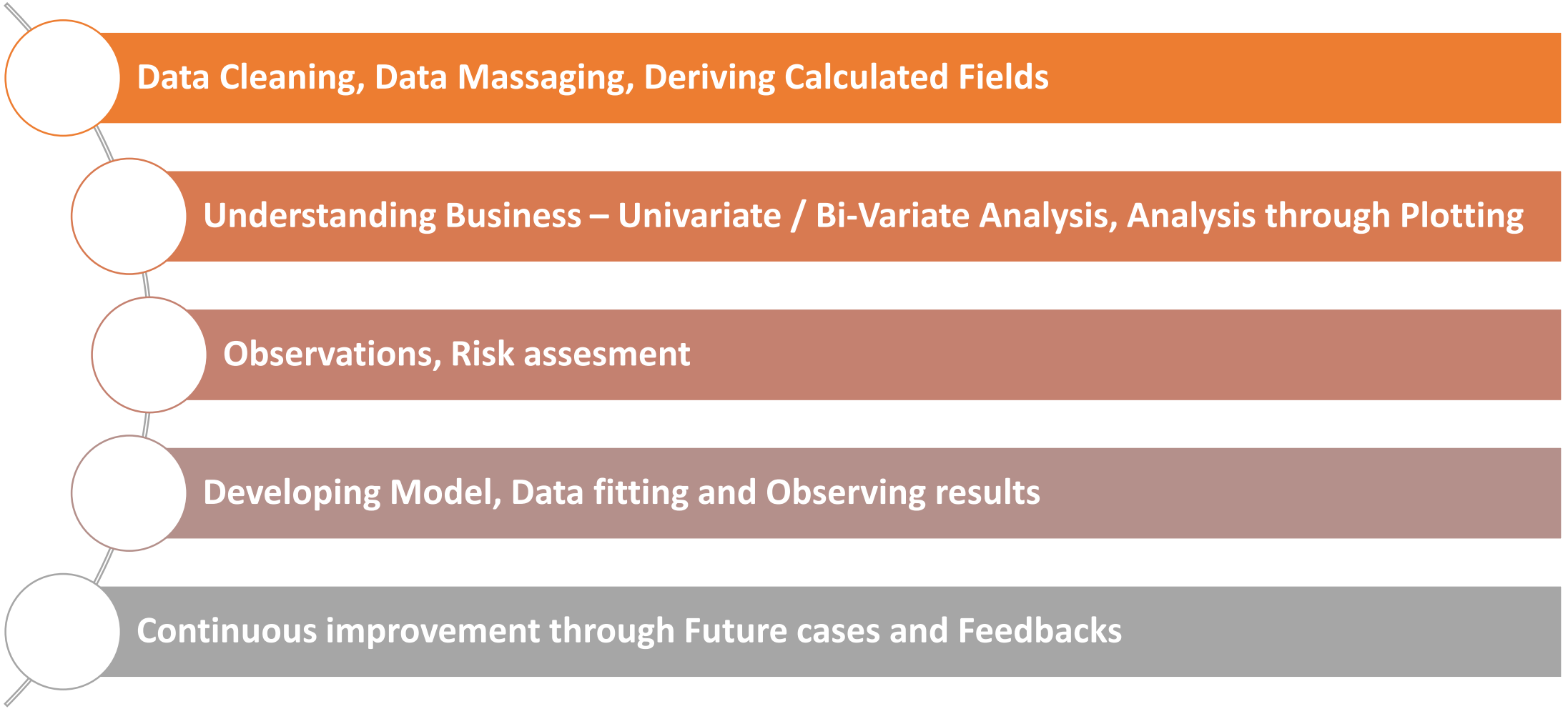
Gramener Case Study – Business Problem



Gramener Case Study – BUSINESS OBJECTIVES



Gramener Case Study – Analytics Model Development Steps



Developing Model – Data cleaning, Derived fields

DATA CLEANING

- Removed the columns having only NULL values
- Removed the columns having only 0
- Converted the date field in date format , added day as 01 for all the dates
- Filled the empty value of last_pymnt_d for the defaiulter's with old date 2000-01-01
- Removed the % from term to only numeric value
- cleaned the employee experience field by removing text
- Replaced NA with zero for mths_since_last_record

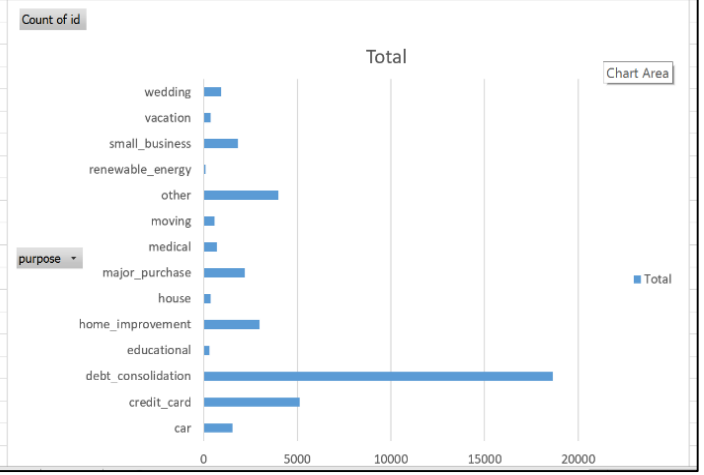
DERIVED FIELDS

- Create new column tenure by removing the ' months' string to use it in calculated field
- Create new field DTP_ratio = ((total_rec_prncp + total_rec_int)/debt) *100
- Create new column debt = Monthly installment * tenure
- Creat new emp_exp column for employee experience by removing the text
- New field created for the customer score
- New column created for transactional score
- Create new column revol_util_1 for revol_util by removing the %
- New column for the behavioural score of the applicant
- New column for credit score of the loan applicant
- New column for the eligibility based on the credit score of the applican

Developing Model - Understanding Business in Details

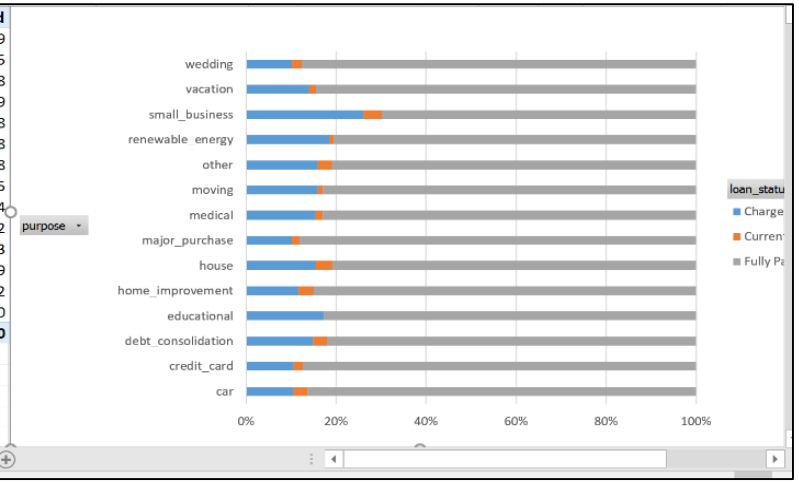
Where the most loan requests are coming from

Row Labels	Count of id
car	1549
credit_card	5130
debt_consolidation	18641
educational	325
home_improvement	2976
house	381
major_purchase	2187
medical	693
moving	583
other	3993
renewable_energy	103
small_business	1828
vacation	381
wedding	947
Grand Total	**39717**



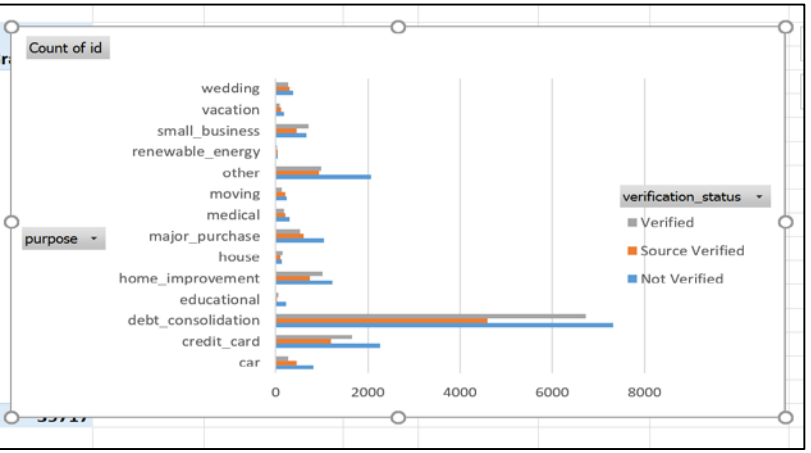
Status of loan per segment

Row Labels	Charged Off	Current	Fully Paid
car	160	50	1339
credit_card	542	103	4485
debt_consolidation	2767	586	15288
educational	56		269
home_improvement	347	101	2528
house	59	14	308
major_purchase	222	37	1928
medical	106	12	575
moving	92	7	484
other	633	128	3232
renewable_energy	19	1	83
small_business	475	74	1279
vacation	53	6	322
wedding	96	21	830
Grand Total	**5627**	**1140**	**32950**



Verification Status of loans per segment

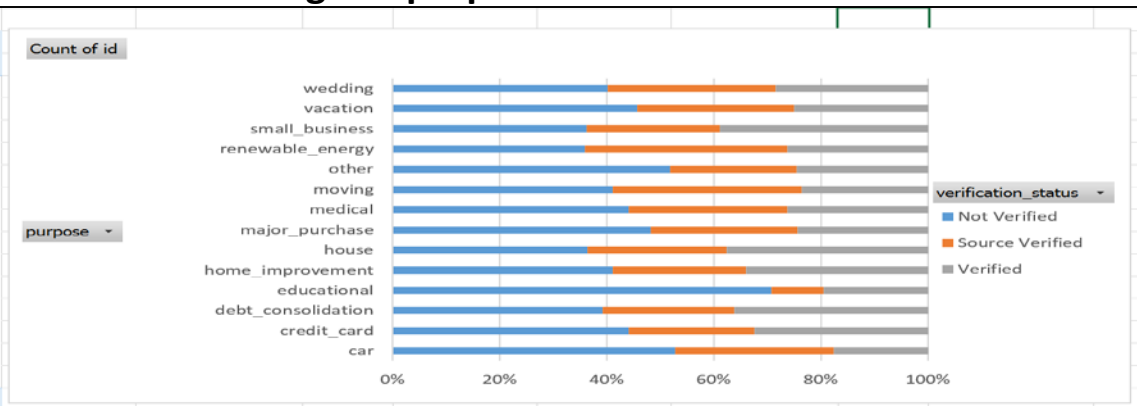
Count of id	Column Labels		
Row Labels	Not Verified	Source Verified	Verified
car	819	458	272
credit_card	2267	1202	1661
debt_consolidation	7315	4596	6730
educational	230	32	63
home_improvement	1224	742	1010
house	139	99	143
major_purchase	1054	601	532
medical	306	205	182
moving	240	206	137
other	2073	941	979
renewable_energy	37	39	27
small_business	663	456	709
vacation	174	112	95
wedding	380	298	269
Grand Total	**16921**	**9987**	**12809**



Developing Model - Understanding Business in Details (Continued)

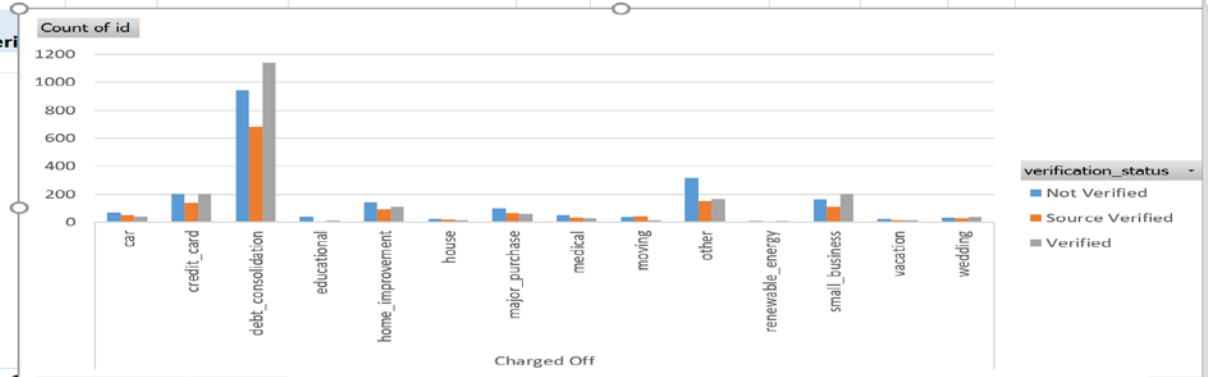
Verification Status of the loans per segment in % understanding the proportions

Count of id	Column Labels			
Row Labels	Not Verified	Source Verified	Verified	Grand Total
car	819	458	272	1549
credit_card	2267	1202	1661	5130
debt_consolidation	7315	4596	6730	18641
educational	230	32	63	325
home_improvement	1224	742	1010	2976
house	139	99	143	381
major_purchase	1054	601	532	2187
medical	306	205	182	693
moving	240	206	137	583
other	2073	941	979	3993
renewable_energy	37	39	27	103
small_business	663	456	709	1828
vacation	174	112	95	381
wedding	380	298	269	947
Grand Total	16921	9987	12809	39717



Which Segment is contributing to charged off/current and what's the verification status

Count of id	Column Labels	
Row Labels	Not Verified	Source Veri
Charged Off		
car	68	
credit_card	202	
debt_consolidation	941	
educational	39	
home_improvement	144	
house	24	
major_purchase	98	
medical	50	
moving	34	
other	317	
renewable_energy	9	
small_business	161	
vacation	24	
wedding	31	
Charged Off Total	2142	1
Grand Total	2142	1

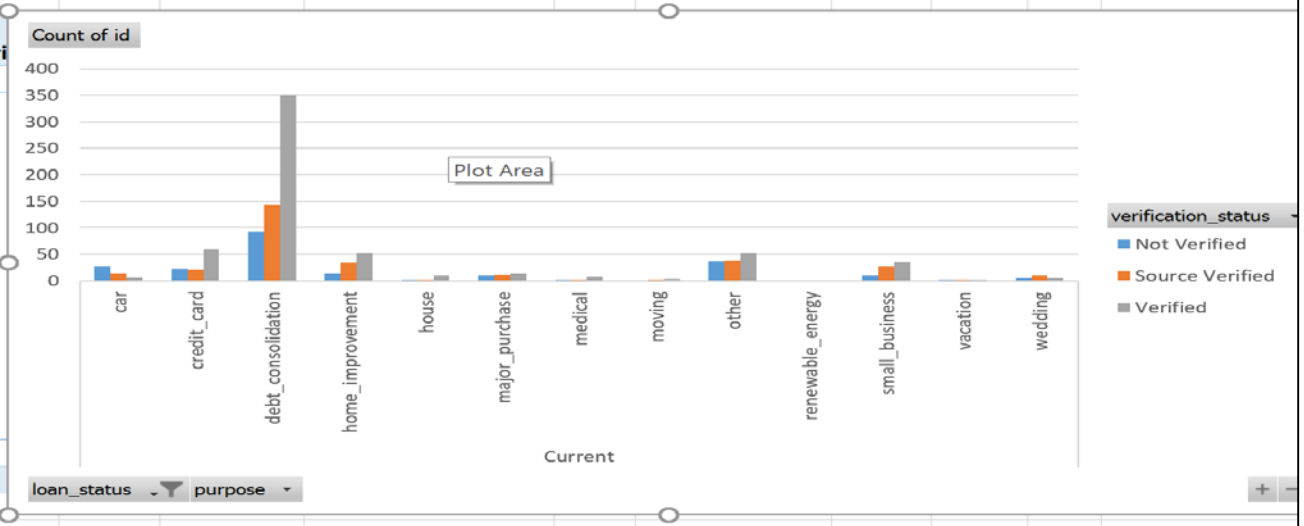


Sheet2 | loan_modified | Sheet3 | Sheet1

Developing Model - Understanding Business in Details (Continued)

Which Segment is contributing to charged off/current and what's the verification status

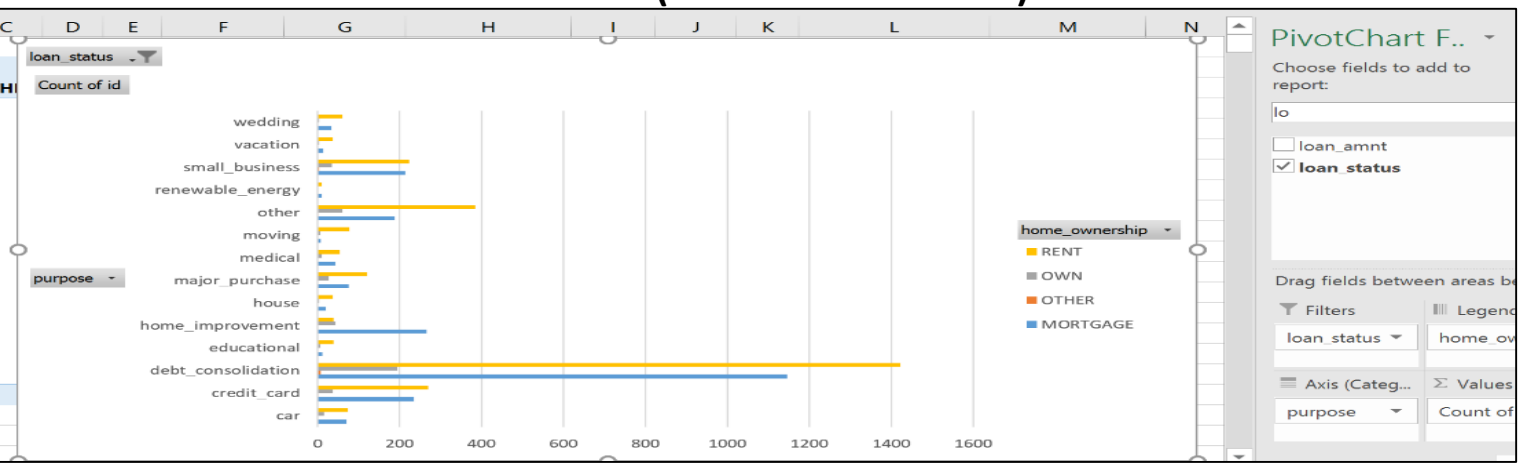
Count of id	Column Labels	
Row Labels	Not Verified	Source Veri
Current		
car	28	
credit_card	22	
debt_consolidation	93	
home_improvement	14	
house	2	
major_purchase	11	
medical	2	
moving	1	
other	37	
renewable_energy		
small_business	10	
vacation	2	
wedding	5	
Current Total	227	
Grand Total	227	



Developing Model - Understanding Business in Details (Continued)

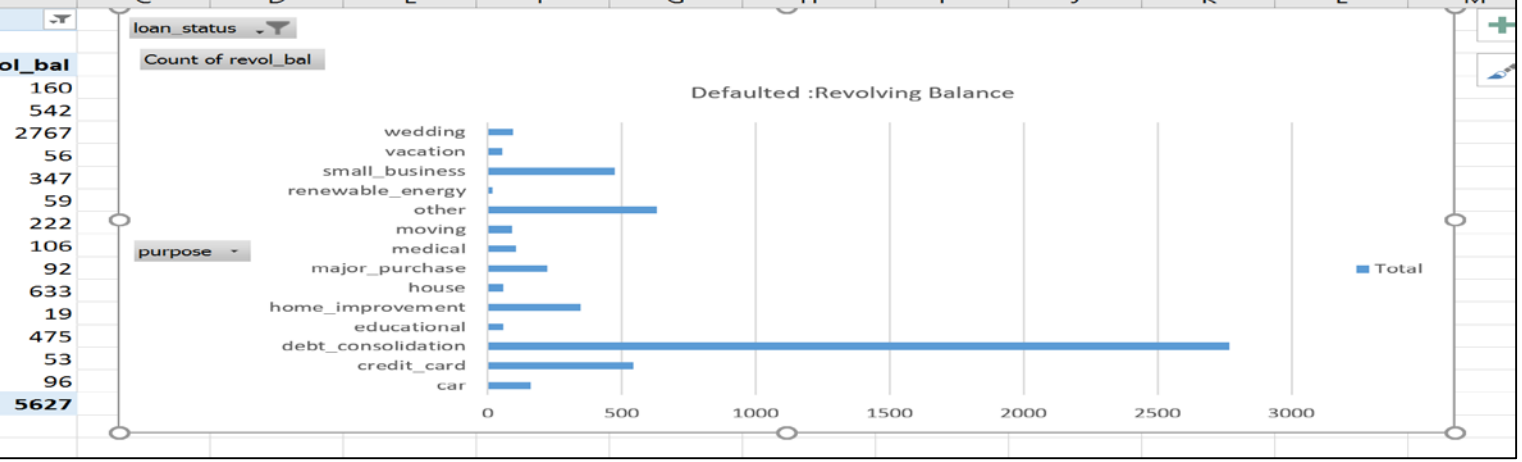
Understanding the Collaterals of Customers – Risk assessment (Status of Defaulters)

Count of id	Column Labels	
Row Labels	MORTGAGE	OTH
car	70	
credit_card	234	
debt_consolidation	1145	
educational	12	
home_improvement	266	
house	19	
major_purchase	75	
medical	42	
moving	8	
other	187	
renewable_energy	9	
small_business	214	
vacation	13	
wedding	33	
Grand Total	2327	



SECTORS at Risk: Based on the learnings from Defaulted, 4 sectors are in the risk zone

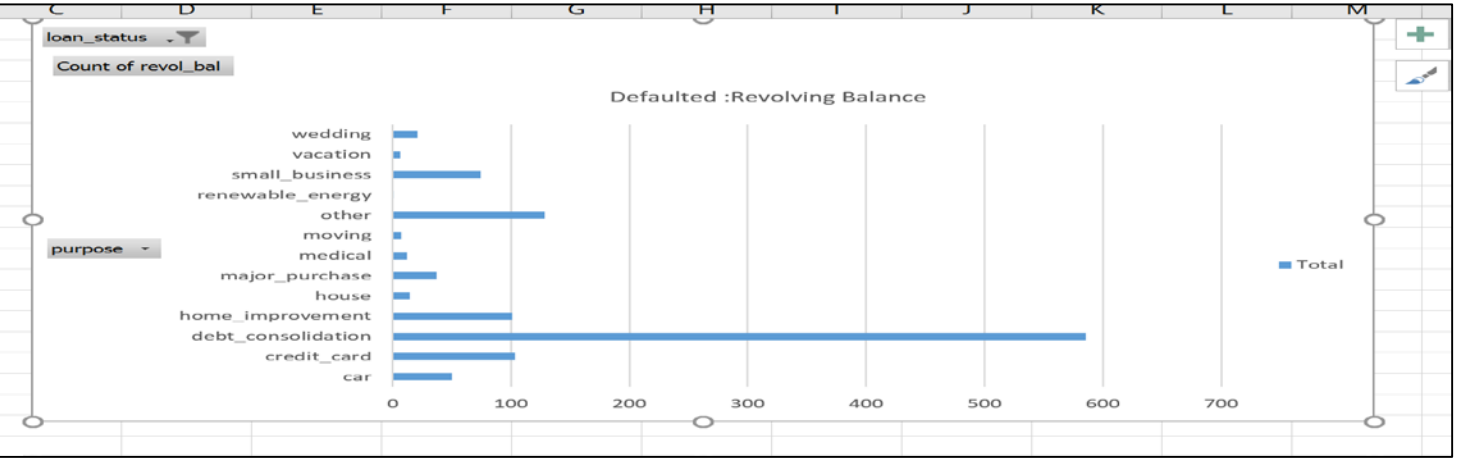
loan_status	Charged Off
Row Labels	Count of revol_bal
car	160
credit_card	542
debt_consolidation	2767
educational	56
home_improvement	347
house	59
major_purchase	222
medical	106
moving	92
other	633
renewable_energy	19
small_business	475
vacation	53
wedding	96
Grand Total	5627



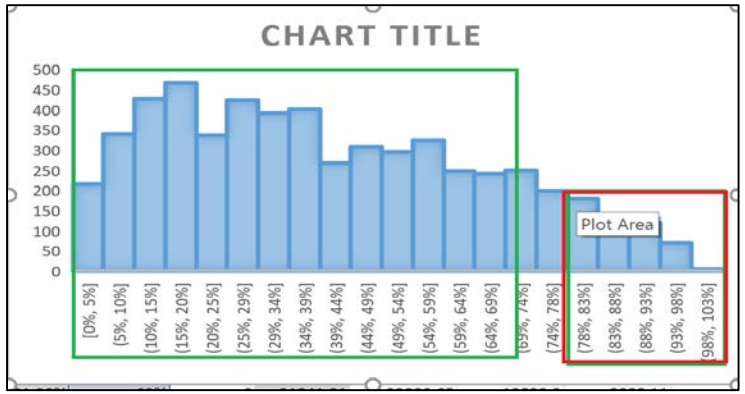
Developing Model - Understanding Business in Details (Continued)

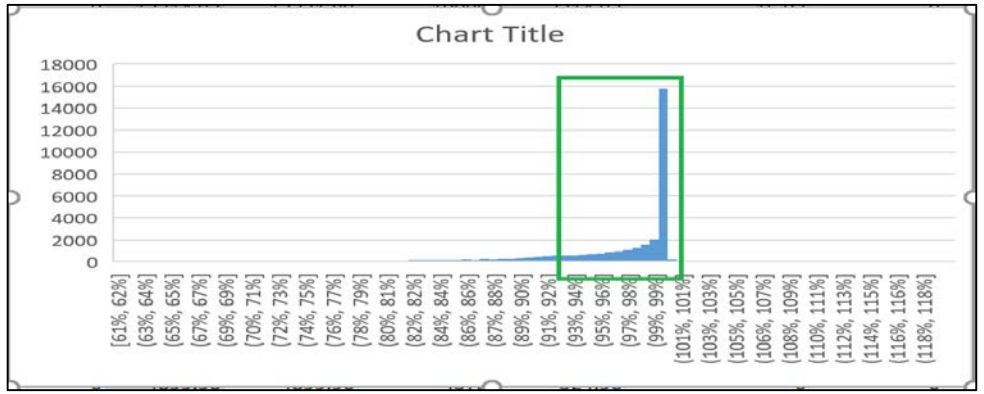
CURRENT Status also follows a similar trend

loan_status	Current
Row Labels	Count of revol_bal
car	50
credit_card	103
debt_consolidation	586
home_improvement	101
house	14
major_purchase	37
medical	12
moving	7
other	128
renewable_energy	1
small_business	74
vacation	6
wedding	21
Grand Total	1140



Histogram to understand the Risk pattern [based on the DTP ratio]

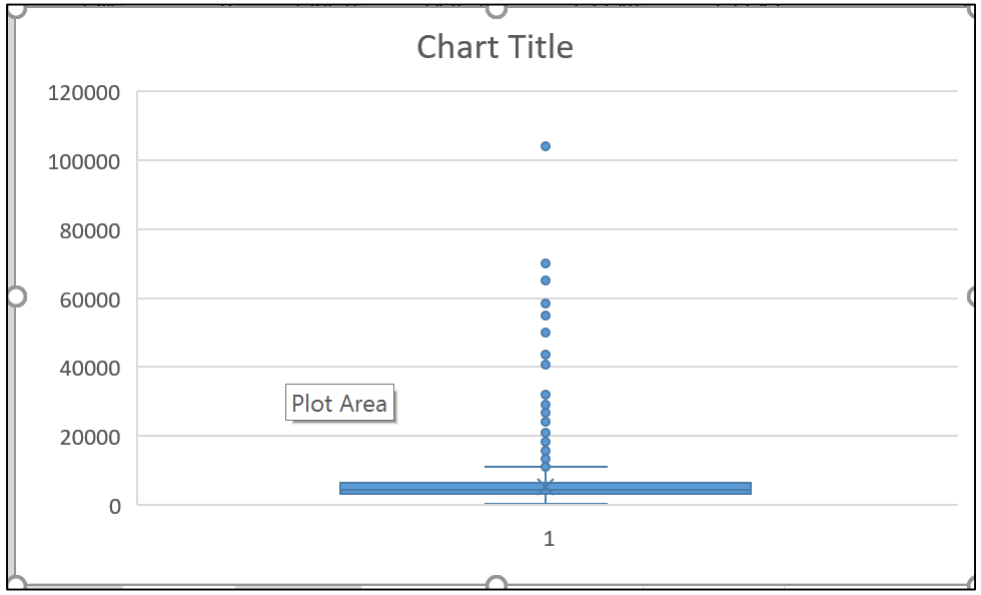




This is for defaulted, up to 70%; safe limit This clearly shows that DTP ratio < 70% would be considered as a risk Conclusion: DTP is a real factor to consider for Risk

Developing Model - Understanding Business in Details (Continued)

Understanding the data distribution of Monthly income of Salary for Credit calculation



	A	B	C	D	E	F	G	H
1	Monthly_inc		Monthly income Distribution					
2	2500.00				First Quartile	3083.33		
3	3333.33		Median		Second Quartile	4416.67		
4	1250.00		4416.67		Third Quartile	6250.00		
5	2500.00				Fourth Quartile	104166.67	Large Outlier	
6	8333.33						Affects Average	
7	8750.00							

This doesn't give any clear picture and can be omitted.

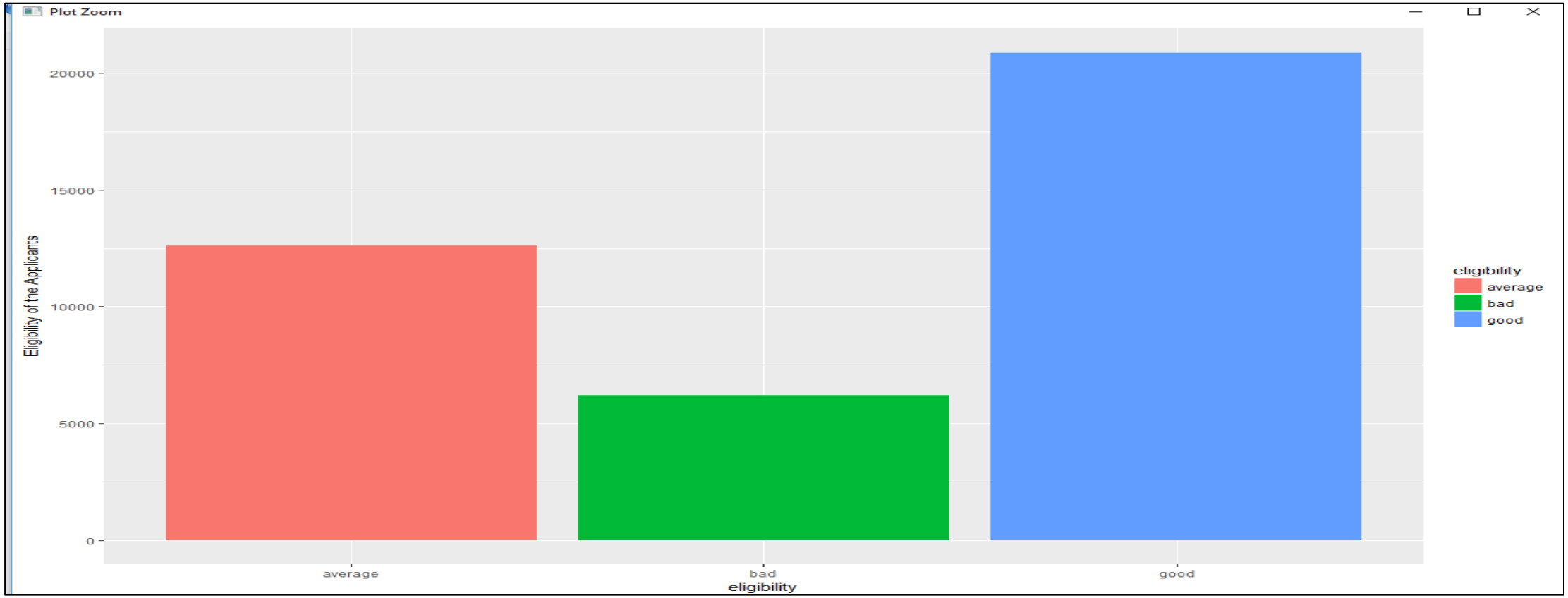
Developing Model - Understanding Business in Details (Continued)

Credit Score calculation mechanism & Risk assessment through weighted parameter values

Risk Assesment							
Category: Customer	**Factors to be considered**	**Attribute from Dataset**	**Value**	**Score ***	**Value**	**Score***	**Comments**
Assesment: Potential to handle debts/Failures	1.High income	Monthly_inc	>=4400	10	<4400	5	Box plot to understand customers income range
Weightage: 40%	2.Home Ownership	Home_ownership	Mortgage| Own	10	rent	5	Reduced risk with Mortgage and Own
	3.Verification Status	verification_status	Verified	10	not verified	5	Reduced risk once verified compared to not verified
	4.Debt2IncomeRatio	DTI	<10	10	>=10	5	Assumption: Keep the DTI below 10%; this can be changed
	5.Delta_payments_debts	calculated field : DTP ratio	>=70%	10	<70	5	Delta between debts and payments must be less
	6.Length of employment	emp_length	>=3years	10	<3y	5	Greater the employeement considered healthy
	7.Loan term	Months	Shorter ; reduced risk (36months)	10	60 mon	5	Assumption:Data set has only 36 and 60; so 60 gets 5
Category: Transaction	**Factors to be considered**	**Attribute**	**Value**	**Score**	**Value**	**Score**	**Comments**
Assesment: How disciplined in handling Finances	1.Date Opened	issue_d	>2y	10	<=2y	5	less years considered to be risky
Weightage: 40%	2.Date of last activity	last_pymt_d	< =6 months	10	>6 mon	5	less indicates healthy transactional activity
	3.Date Reported	use deliq_last_2_yrs	<2	10	>=2	5	Assumption is >=2 reported is considered risky;
	4.Inquiry Date	use inq_last_6_mon	<2	10	>=2	5	Large numbers will affect the score adversly
	5.Months since last deliquency	mths_since_last_delinq	<89	10	>89	5	Assumption:Give benefit of doubt for 2 payments; rest risky
Category: Behaviourial	**Factors to be considered**	**Attribute**	**Value**	**Score**	**Value**	**Score**	**Comments**
Assesment: Reflection of customer behaviour	1.length of credit history	issue_d - earliest_cr_line	>=2	10	<2	5	how long the oldest - recent opened
Ex:Does he utilize credit well or gets into bankrupcy	2.Public Records	mths_since_last_records	<=1	10	>1	5	Assumption:1-benift of doubt(economy) >1 consider risky
Weightage: 20%	Revolving balance Utilization	revol_util	<50%	10	>50%	5	Greater value is risky as customer is unable to pay min balan
	Recoveries	recoveries	<=10	10	>10	5	indicates followup and it's a cost to the lender (time/money)
	Collections	collection_recovery_fee	<=10	10	>10	5	indicates followup and it's a cost to the lender (time/money)
** Scores are Just Assumption and an illustration of logic; the actual numbers can vary in real time*							
*** Weightages are just assumptions that can vary in real time*							

Developing Model - Understanding Business in Details (Continued)

Model Outcome on Existing Data



EXPECTED OUTCOME AND INFERENCES

- CREDIT SCORE < =79 MOST LIKELY TO BE DEFAULTER
- CREDIT SCORE BETWEEN 80 to 84 MAY REPAY LOAN
- CREDIT SCORE >= 85 IS MOST LIKELY TO REPAY
- CREDIT SCORE VALUE RANGE IS 1 to 100
- CALCULATE CREDIT SCORE BASED ON INPUT IN THE MODEL
- ANALYZE NEW APPLICANT DATA

FUTURE FINE TUNING – BASED ON FEEDBACK

TEST THE MODEL WITH NEW INPUT

New applicants' data

Previously known data

FINE TUNE BASED ON OUTCOME

Match Model outcome Vs. Real outcome

Fine Tune Model to match outcome as close as real outcome

GOAL

- Less Rejection of applications in case of probability of repayment is high
- Less acceptance in case of probability of default is high

Continuous Fine Tuning

- Fine tune range and credit score calculation in case model outcome is not matching with real outcome
- Fine tune range and credit score calculation in case any new parameter is explored and seems to have major impact