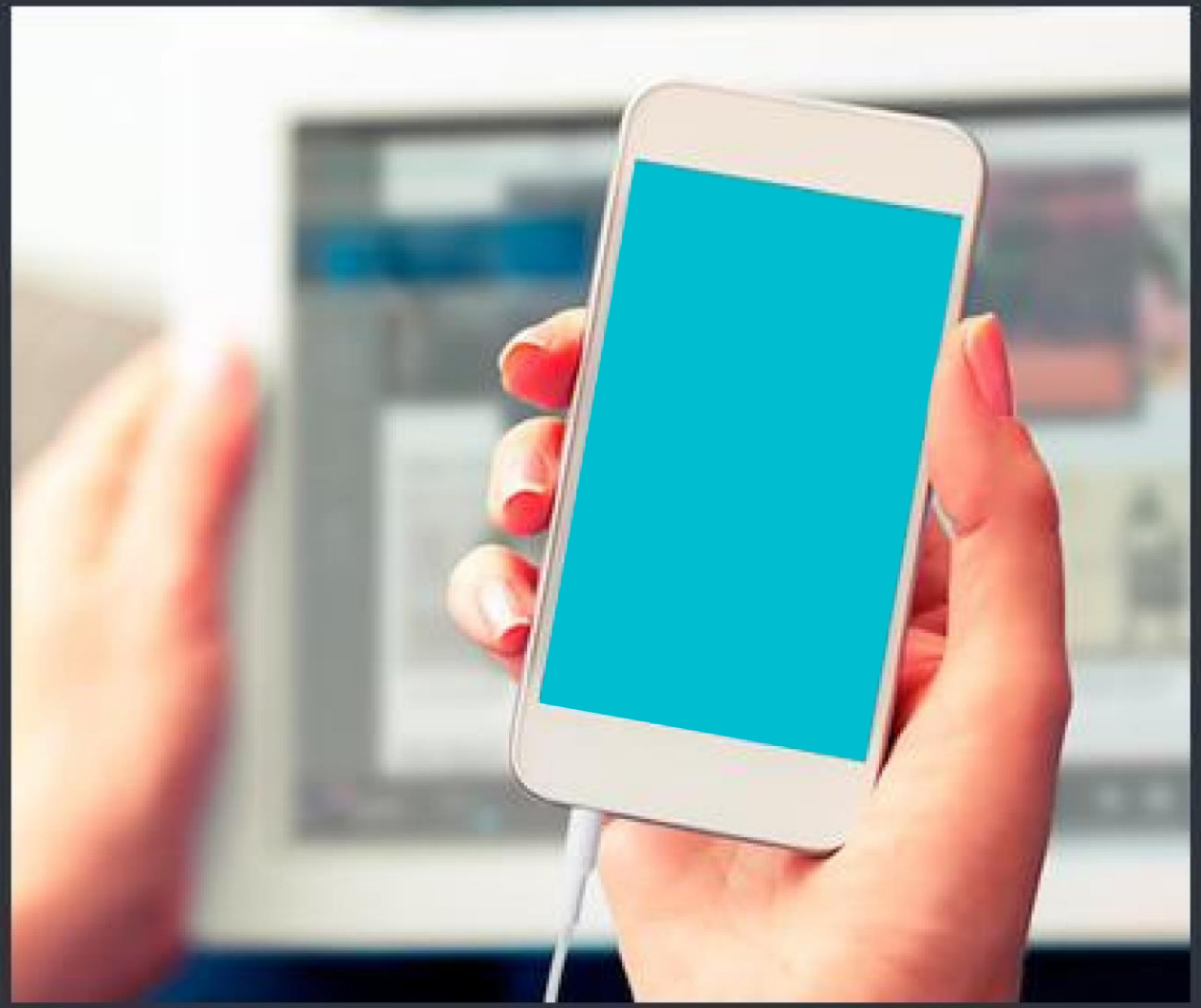


Presented by-

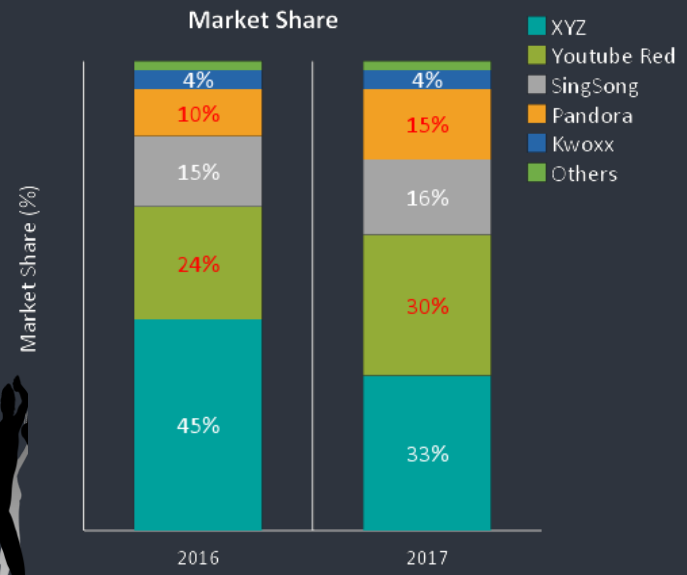
Puneeta Chaturvedi

Contents

Business Problem

Current State:

- XYZ used to follow “Freemium” model for providing music to its subscribers
- Market share of XYZ has seen a dip of 12% YoY in 2017
- “Business Strategies” division of “XYZ” in collaboration of “Analytics” division is responsible for providing actionable recommendations to sales and marketing team and enable them to acquire new customers and retain through active engagement



Desired State:

- XYZ is able to retain their customer base with their existing model and effectively monetize it

Objective:

- Classify customers on the basis of likelihood to leave system
- Identify factors affecting customer behavior significantly
- Analyze customer behavior of most engaged customer
- Identify Potential customers and pitch relevant services to them

Trigger:

- In 2016, “On-Demand” services model was introduced with the arrival of Pandora and Youtube Red
- This has made “XYZ” to reconsider their existing business model

Note:
- Light shade in map is showing market share of year 2016
- Dark shade in map is showing reduced market share in 2017

Problem Analysis

Identify target customer:

- Identify potential customers and prepare targeting strategy
- Pitch in same services to potential customers that was opted by customers with similar behaviour

Analyse behaviour of customer:

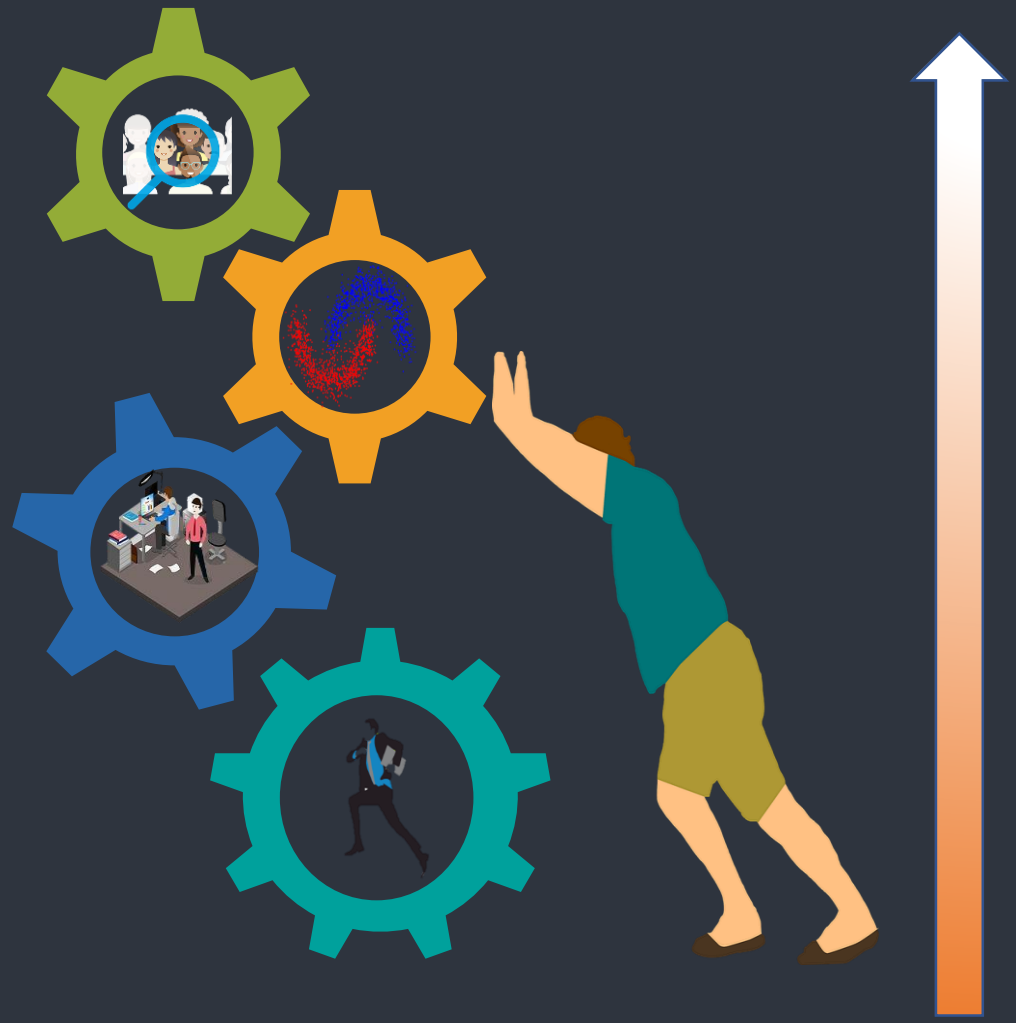
- Analyse behaviour of customer who are least likely to leave
- Compare behaviour across all the segments

Identify factors affecting customer behaviour:

- Perform descriptive analysis to understand the impact of different variable on the customer behaviour across all the categories

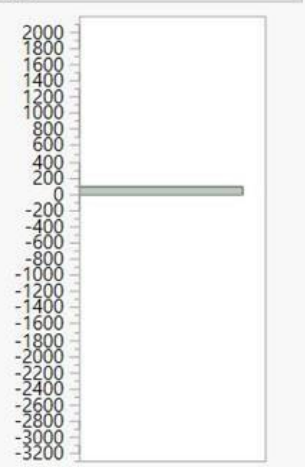
Prediction of customer attrition:

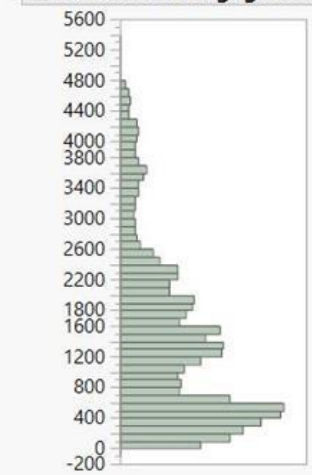
- Use predictive modelling techniques to classify customers based on likelihood to leave system

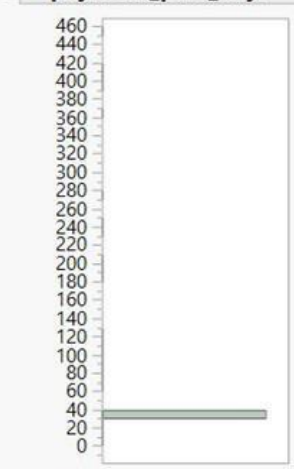


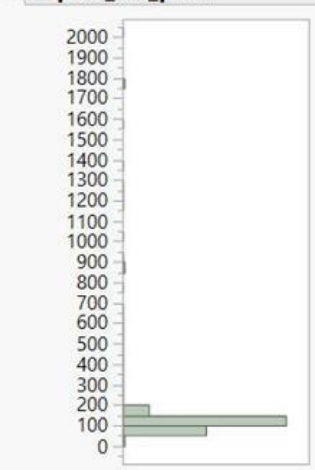
EDA - Outlier

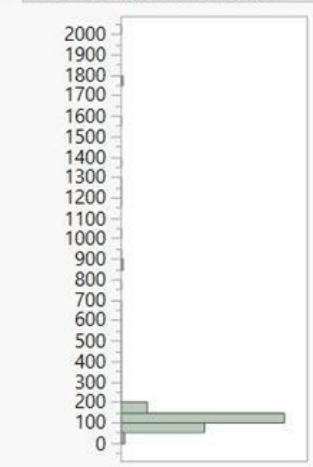
Distributions











bd

Quantiles		
100.0%	maximum	2016
99.5%		57
97.5%		47
90.0%		36
75.0%	quartile	28
50.0%	median	0
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	-3152

Summary Statistics	
Mean	14.258138
Std Dev	20.469914
Std Err Mean	0.0202983
Upper 95% Mean	14.297922
Lower 95% Mean	14.218354
N	1016983

Duration of Engagement

Quantiles		
100.0%	maximum	5393
99.5%		4634
97.5%		4188
90.0%		3143.6
75.0%	quartile	1996
50.0%	median	1238
25.0%	quartile	516
10.0%		243
2.5%		89
0.5%		58
0.0%	minimum	-15

Summary Statistics	
Mean	1420.7973
Std Dev	1105.3918
Std Err Mean	1.0961233
Upper 95% Mean	1422.9457
Lower 95% Mean	1418.649
N	1016983

payment_plan_days

Quantiles		
100.0%	maximum	450
99.5%		395
97.5%		30
90.0%		30
75.0%	quartile	30
50.0%	median	30
25.0%	quartile	30
10.0%		30
2.5%		30
0.5%		10
0.0%	minimum	0

Summary Statistics	
Mean	33.960048
Std Dev	33.513789
Std Err Mean	0.0332328
Upper 95% Mean	34.025183
Lower 95% Mean	33.894912
N	1016983

plan_list_price

Quantiles		
100.0%	maximum	2000
99.5%		1599
97.5%		180
90.0%		180
75.0%	quartile	149
50.0%	median	149
25.0%	quartile	99
10.0%		99
2.5%		99
0.5%		0
0.0%	minimum	0

Summary Statistics	
Mean	149.31874
Std Dev	139.81589
Std Err Mean	0.1386436
Upper 95% Mean	149.59047
Lower 95% Mean	149.047
N	1016983

actual_amount_paid

Quantiles		
100.0%	maximum	2000
99.5%		1599
97.5%		180
90.0%		180
75.0%	quartile	149
50.0%	median	149
25.0%	quartile	99
10.0%		99
2.5%		99
0.5%		0
0.0%	minimum	0

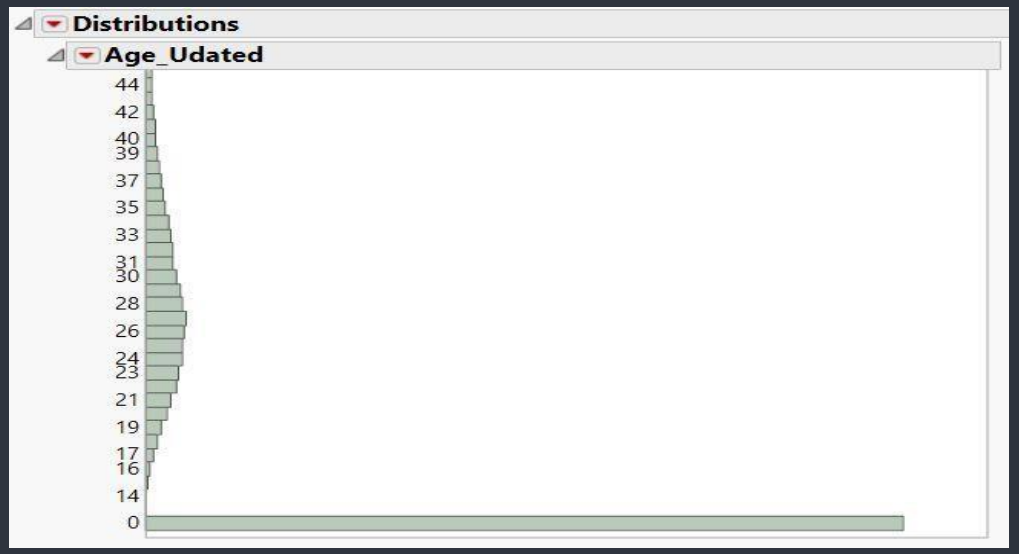
Summary Statistics	
Mean	148.76236
Std Dev	139.98476
Std Err Mean	0.138811
Upper 95% Mean	149.03442
Lower 95% Mean	148.49029
N	1016983

Recency

Quantiles		
100.0%	maximum	1066
99.5%		965
97.5%		834
90.0%		277
75.0%	quartile	273
50.0%	median	262
25.0%	quartile	252
10.0%		246
2.5%		246
0.5%		246
0.0%	minimum	246

Summary Statistics	
Mean	299.13356
Std Dev	140.28663
Std Err Mean	0.1391104
Upper 95% Mean	299.40621
Lower 95% Mean	298.86091
N	1016983

EDA



- Imputed age with “Multivariate Normal Imputation” wherever age is missing or less than 13
- Capped all the age above 57 with 57
- Derived columns : “Duration of Engagement”, “Recency” from
  - registration_init_time
  - transaction_date
  - membership_expire_date

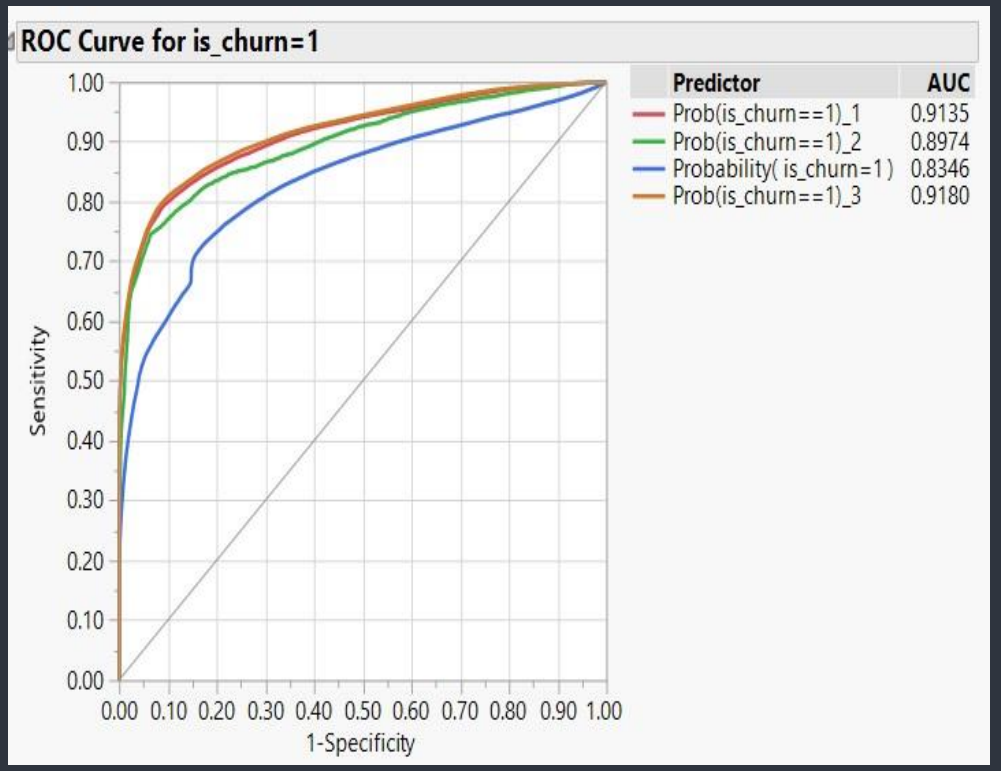
Correlations

	registered_via	Duration of Engagement	payment_method_id	payment_plan_days	plan_list_price	actual_amount_paid	is_auto_renew	Recency	Age_Updated
registered_via	1.0000	0.5621	0.0039	-0.0159	-0.0220	-0.0178	0.1782	-0.0055	0.2035
Duration of Engagement	0.5621	1.0000	-0.1111	-0.0029	0.0160	0.0197	0.1731	-0.0171	0.2652
payment_method_id	0.0039	-0.1111	1.0000	-0.2783	-0.2525	-0.2414	0.2617	0.1861	0.0230
payment_plan_days	-0.0159	-0.0029	-0.2783	1.0000	0.9591	0.9568	-0.3877	-0.0277	-0.0401
plan_list_price	-0.0220	0.0160	-0.2525	0.9591	1.0000	0.9965	-0.3700	-0.0328	-0.0333
actual_amount_paid	-0.0178	0.0197	-0.2414	0.9568	0.9965	1.0000	-0.3714	-0.0277	-0.0309
is_auto_renew	0.1782	0.1731	0.2617	-0.3877	-0.3700	-0.3714	1.0000	0.0470	0.2279
Recency	-0.0055	-0.0171	0.1861	-0.0277	-0.0328	-0.0277	0.0470	1.0000	0.0216
Age_Updated	0.2035	0.2652	0.0230	-0.0401	-0.0333	-0.0309	0.2279	0.0216	1.0000

Model Comparison

Measures of Fit for is_churn

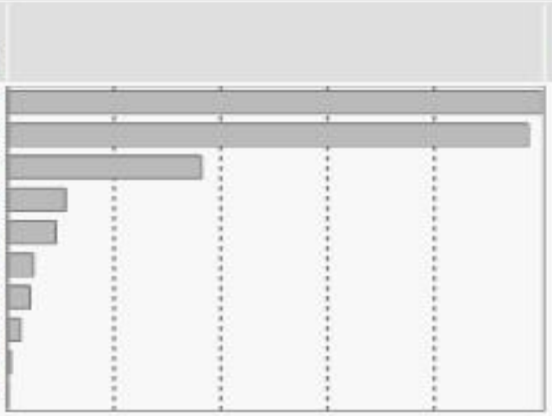
Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N	AUC
Boosted Tree		0.5105	0.6020	0.1816	0.2204	0.0973	0.0613	478113	0.9135
Bootstrap Forest		0.4375	0.5292	0.2087	0.2368	0.1269	0.0740	478113	0.8974
Fit Generalized Logistic Regression		0.3017	0.3829	0.2591	0.2679	0.1440	0.0892	478113	0.8346
Partition		0.5246	0.6156	0.1764	0.2173	0.0946	0.0596	478113	0.9180



Factors affecting Customer Behavior

Column Contributions

Term	Number of Splits	G^2	Portion
Recency	41	41314.0147	0.3709
payment_plan_days	8	40524.6672	0.3638
is_cancel	1	15029.5907	0.1349
actual_amount_paid	4	4731.99042	0.0425
payment_method_id	26	3851.37325	0.0346
is_auto_renew	5	2204.14853	0.0198
Duration of Engagement	24	1973.54379	0.0177
plan_list_price	16	1249.58049	0.0112
Age_Udated	21	439.903415	0.0039
registered_via	7	71.4272675	0.0006



Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.5235	0.5274	0.5257	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6144	0.6185	0.6166	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1767	0.1759	0.1759	∑ -Log(ρ[j])/n
RMSE	0.2176	0.2171	0.2169	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.0947	0.0943	0.0943	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0597	0.0594	0.0594	∑ (ρ[j]≠ρMax)/n
N	286868	71717	119528	n

Confusion Matrix

Training

Actual	Predicted Count	
is_churn	0	1
0	249896	1980
1	15148	19844

Validation

Actual	Predicted Count	
is_churn	0	1
0	62443	473
1	3785	5016

Test

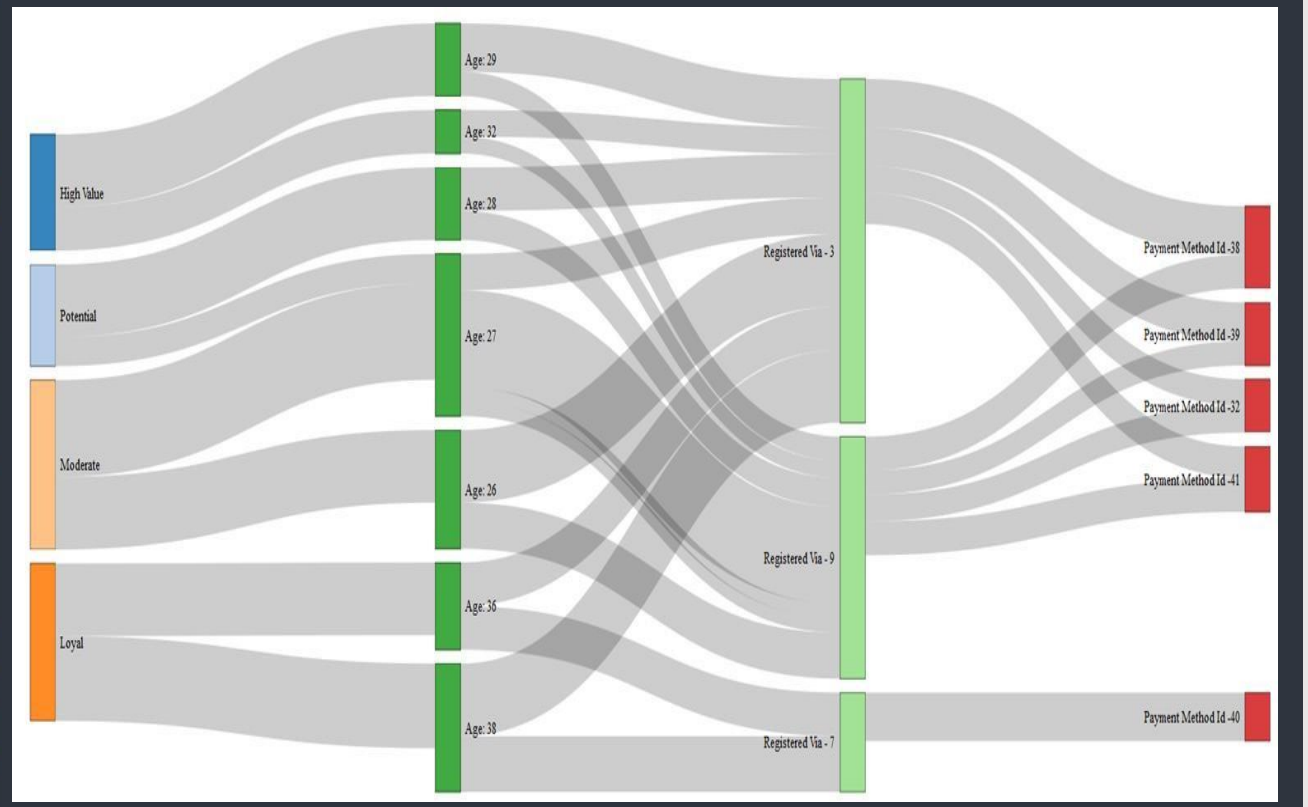
Actual	Predicted Count	
is_churn	0	1
0	104122	819
1	6285	8302

Analyse behaviour of customer

		Category	High Value	Potential	Loyal	Moderate
Population Distribution	Is Churn Absolute	0	11	2	178,185	255,495
		1	5,160	7,030	11,289	20,941
		Total	5,171	7,032	189,474	276,436
	Is Churn%	0	0.213%	0.028%	94.042%	92.425%
		1	99.787%	99.972%	5.958%	7.575%
		Total %	1.082%	1.471%	39.630%	57.818%

Clustering Varibales	Duration of Engagement	2,160	1,622	3,055	1,187
	payment_plan_days	400	175	30	31
	plan_list_price	$ 1,658	$ 711	$ 149	$ 147
	actual_amount_paid	$ 1,658	$ 710	$ 148	$ 146
	Recency	275	268	271	298
	Age	29	27	36	26

Profiling Variables	city	13,5,4,15	13,5,4,15	13,5,4,15	13,5,4,15
	registered_via	9,3	9,3	9,7	9,3
	Duration of Engagement	2,160	1,622	3,055	1,187
	payment_method_id	32,38	32,38	39,40	39,41
	payment_plan_days	400	175	30	31
	plan_list_price	$1,658	$711	$149	$147
	actual_amount_paid	$1,658	$710	$148	$146
	is_auto_renew	-	-	95.60%	83.83%
	Recency	275	268	271	298
	is_cancel	-	-	2.35%	3.30%
	Age_Updated	29	27	36	26



Findings:

- From “Potential” customers –
  - ➢ 99.9% customers are likely to move out
  - ➢ 75% of the customers are from 27 – 30 years age group
- From “Loyal” customers –
  - ➢ 2.35% customers are likely to move out
  - ➢ ~75% of the customers are from 36 to 40 years age group

Insights:

- Potential customers are very similar to most engaged customers in terms of the age group, city, registration channel, and payment method
- But, they have low “Duration of Engagement” as compared to “Most Engaged”
- With migration of a ”Potential” customer to “Most Engaged”, company would make additional $943 per customer
- “High Value” and “Potential” customers didn’t opt for auto renewal and didn’t cancel subscription

Analyse behavior of customer

Iterative Clustering

Cluster Comparison

Method	NCluster	CCC	Best
K-Means Clustering	4	205.385	Optimal CCC

Columns Scaled Individually

Control Panel

K Means NCluster=4

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	5171	23	0
2	7032		
3	189474		
4	276436		

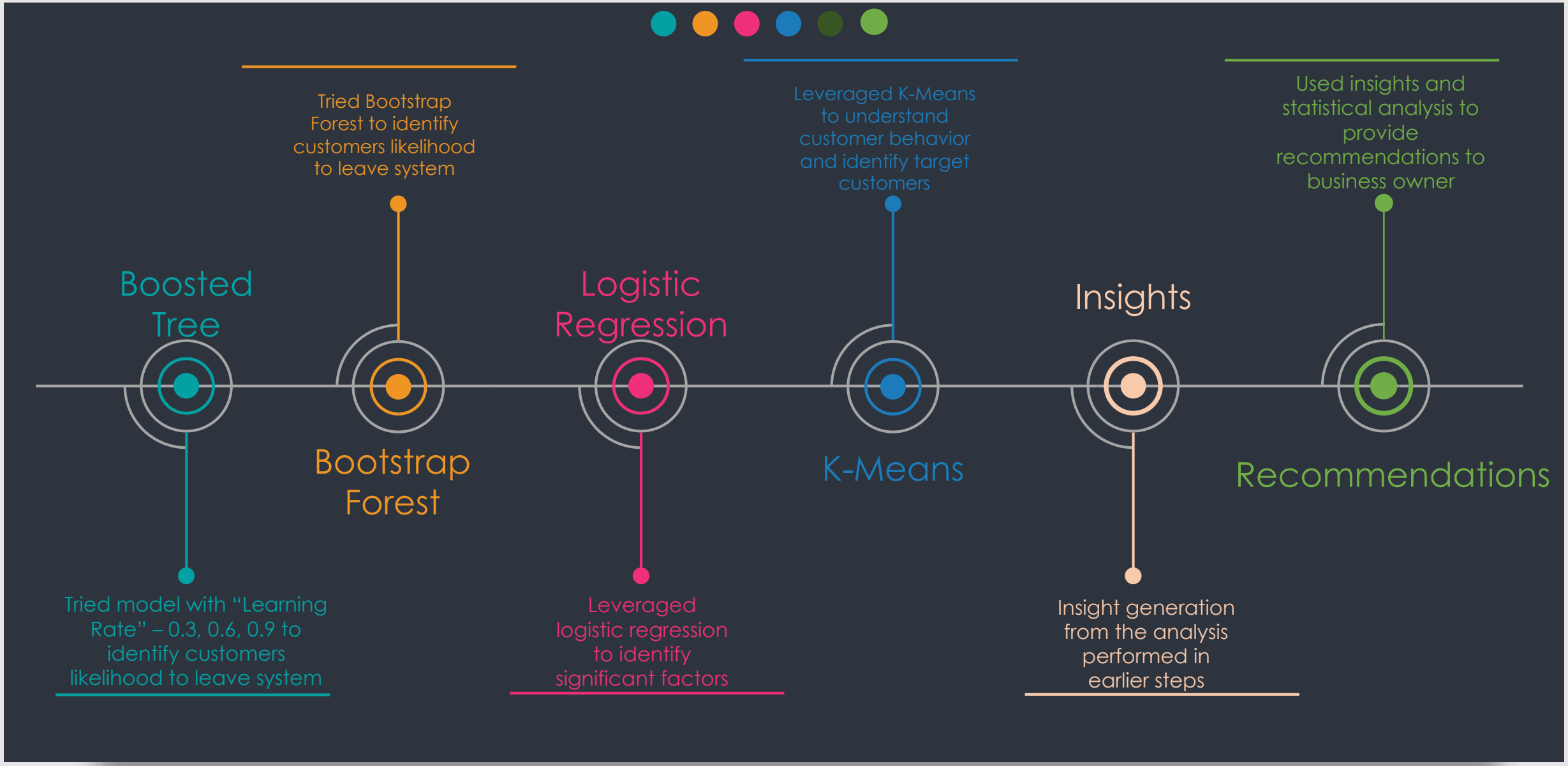
Cluster Means

Cluster	Duration of Engagement	payment_plan_days	plan_list_price	actual_amount_paid	Recency	Age_Udated
1	2159.80101	399.581319	1657.9855	1657.9855	275.276349	28.968217
2	1621.9963	175.123009	710.513936	710.286547	268.499716	27.0430745
3	3055.14862	30.0399685	148.819353	148.303852	271.263683	35.8418049
4	1186.71071	30.758606	147.213105	145.830916	297.793811	26.247933

Cluster Standard Deviations

Cluster	Duration of Engagement	payment_plan_days	plan_list_price	actual_amount_paid	Recency	Age_Udated
1	1189.78152	20.4480775	229.36904	229.36904	61.4214718	8.89267872
2	1131.13964	45.1609097	232.153629	232.066341	38.3145846	8.06394777
3	1004.79544	3.6305251	19.4779351	20.8524481	76.8675268	8.55817648
4	685.321028	10.733759	32.3248692	34.4615684	136.808018	5.42958174

Project Progress



Recommendations

Recommendations:

- We need to focus on top two categories of which more than 90% customers are like to churn. These are the categories which are generating a significant share of overall revenue
- Customers can be offered discounts and credits under a marketing campaign whenever their subscription is going to end
- Targeting of customers should be as per the level of engagement in case of limited budget
- Capture data at service level so that pre post analysis can be performed to see the impact of implemented marketing campaigns

Case 1:

Findings:

- "High Value" customers didn't opt for auto-renewal
- 99.7% of them are likely to churn out
- ~75% of the customers are from 27-30 years age group

Insights:

- These customers are premium customers who contributes most to the company revenue
- Every time their subscription ends they look for available options in market
- Age shows that these customers are highly active and updated
- Every new customer gets some initial discounts or credits in all the companies.

Recommendation:

Whenever subscriptions of "High Value" customer ends, marketing or sales team needs to follow up with them for new subscriptions and services. They must be offered discounts and extra credits. This would help to avoid migration of customers to other competitors.

Appendix

Prediction of customer attrition

Specifications

Target Column:	is_churn	Number of training rows:	286868
Validation Column:	Validation	Number of validation rows:	71717
Number of Layers:	50	Number of test rows:	119528
Splits per Tree:	3		
Learning Rate:	0.3		
Overfit Penalty:	0.0001		

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.4931	0.4998	0.4991	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5849	0.5918	0.5909	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1880	0.1862	0.1858	∑ -Log(ρ[j])/n
RMSE	0.2231	0.2220	0.2216	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.1024	0.1016	0.1015	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0620	0.0615	0.0609	∑ (ρ[j]≠ρMax)/n
N	286868	71717	119528	n

Confusion Matrix

Training

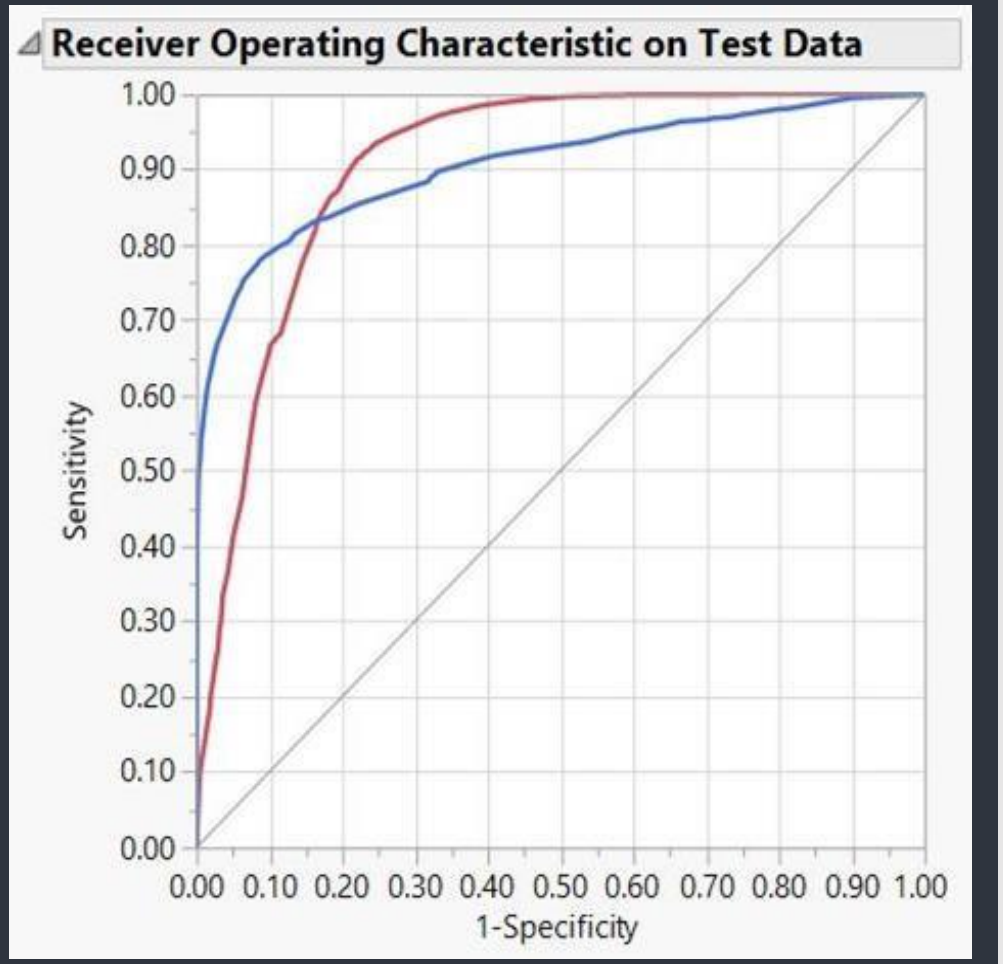
Actual	Predicted Count	
is_churn	0	1
0	249404	2472
1	15300	19692

Validation

Actual	Predicted Count	
is_churn	0	1
0	62302	614
1	3795	5006

Test

Actual	Predicted Count	
is_churn	0	1
0	103966	975
1	6308	8279



Details:

Boosted tree is used with a learning rate of 0.6 which led to a "Misclassification Rate" of 0.0601 on test data set

Misclassification Rate:

Training : 0.0609
Validation : 0.0607
Test : 0.0601

Prediction of customer attrition

Specifications

Target Column:	is_churn	Number of training rows:	286868
Validation Column:	Validation	Number of validation rows:	71717
Number of Layers:	50	Number of test rows:	119528
Splits per Tree:	3		
Learning Rate:	0.6		
Overfit Penalty:	0.0001		

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.5047	0.5099	0.5102	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5963	0.6017	0.6016	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1837	0.1825	0.1817	∑ -Log(ρ[j])/n
RMSE	0.2212	0.2204	0.2198	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.0988	0.0983	0.0981	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0609	0.0607	0.0601	∑ (ρ[j]≠ρMax)/n
N	286868	71717	119528	n

Confusion Matrix

Training

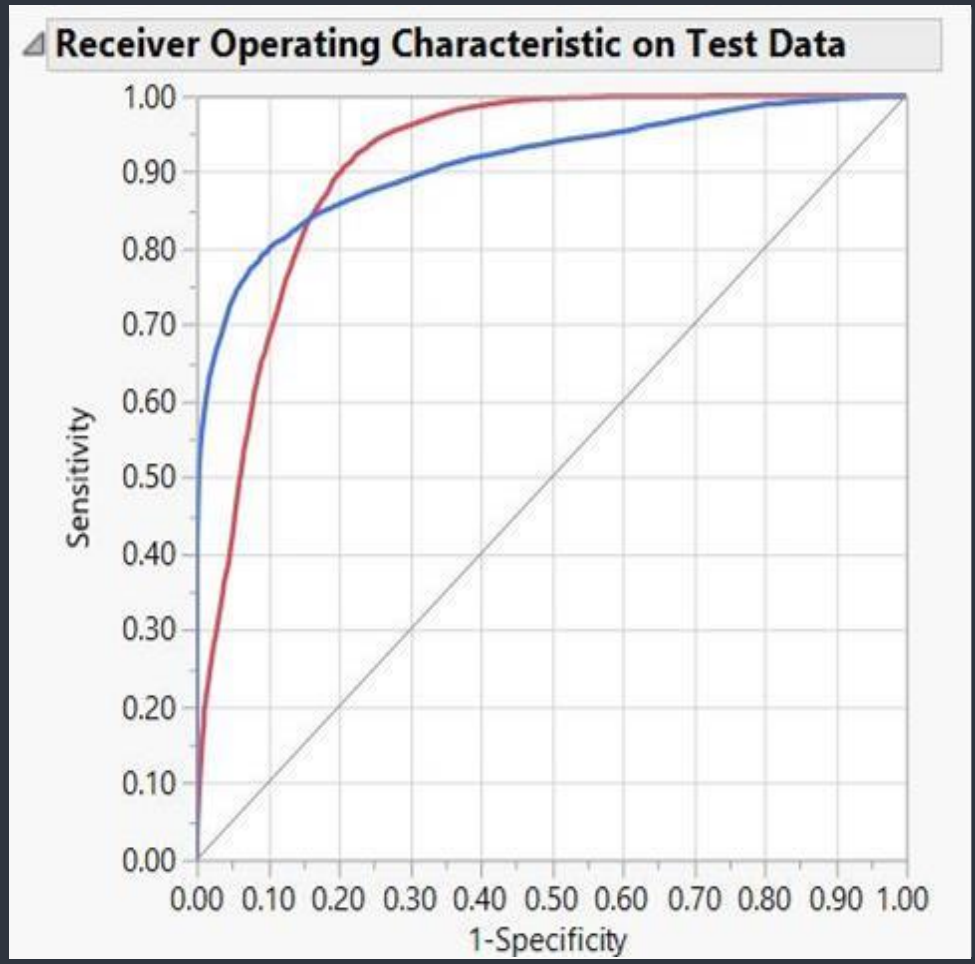
Actual	Predicted Count	
is_churn	0	1
0	249293	2583
1	14900	20092

Validation

Actual	Predicted Count	
is_churn	0	1
0	62277	639
1	3717	5084

Test

Actual	Predicted Count	
is_churn	0	1
0	103903	1038
1	6143	8444



Details:

Boosted tree is used with a learning rate of 0.9 which led to a "Misclassification Rate" of 0.0605 on test data set

Misclassification Rate:

Training : 0.0615
Validation : 0.0615
Test : 0.0605

Prediction of customer attrition

Specifications

Target Column:	is_churn	Number of training rows:	286868
Validation Column:	Validation	Number of validation rows:	71717
Number of Layers:	50	Number of test rows:	119528
Splits per Tree:	3		
Learning Rate:	0.9		
Overfit Penalty:	0.0001		

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.5083	0.5130	0.5143	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5998	0.6047	0.6056	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1823	0.1813	0.1802	∑ -Log(ρ[j])/n
RMSE	0.2208	0.2204	0.2193	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.0975	0.0971	0.0968	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0615	0.0615	0.0605	∑ (ρ[j]≠ρMax)/n
N	286868	71717	119528	n

Confusion Matrix

Training

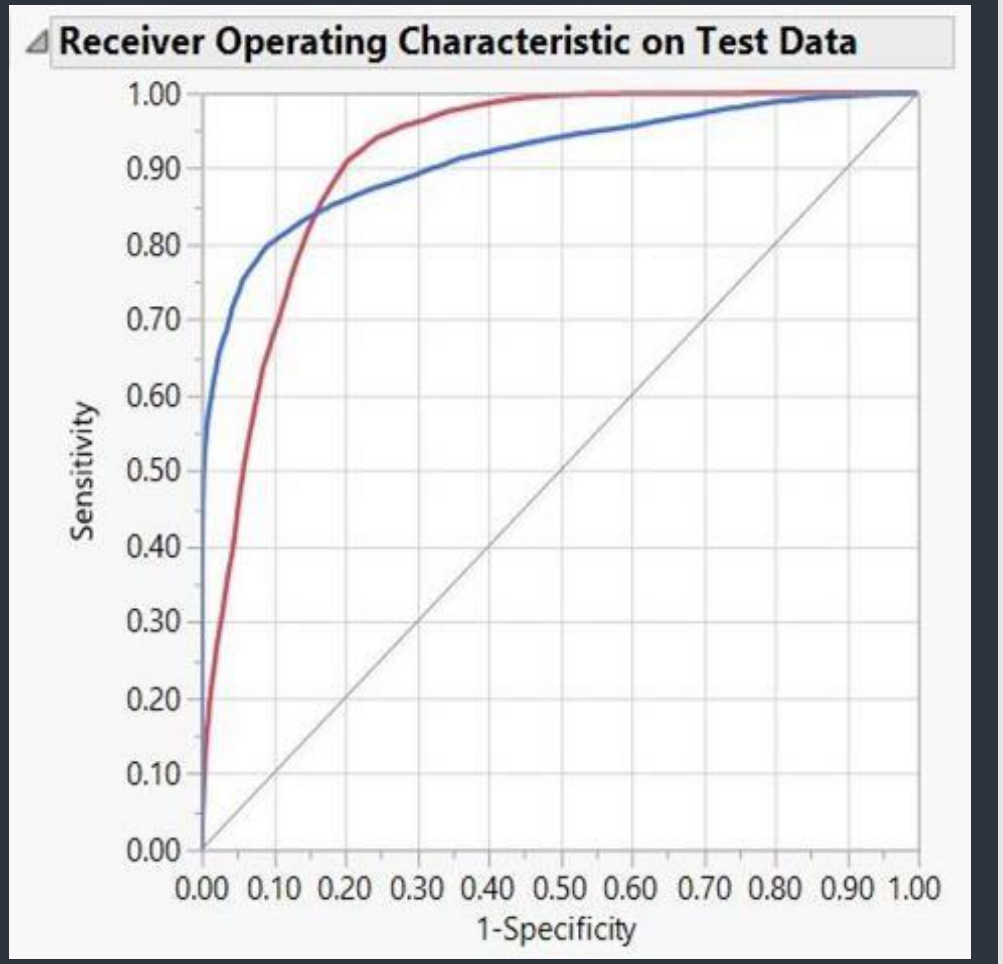
Actual	Predicted Count	
is_churn	0	1
0	248852	3024
1	14622	20370

Validation

Actual	Predicted Count	
is_churn	0	1
0	62163	753
1	3657	5144

Test

Actual	Predicted Count	
is_churn	0	1
0	103733	1208
1	6022	8565



Details:

Boosted tree is used with a learning rate of 0.9 which led to a “Misclassification Rate” of 0.0605 on test data set

Misclassification Rate:

Training : 0.0615
Validation : 0.0615
Test : 0.0605

Prediction of customer attrition

Bootstrap Forest for is_churn

Specifications

Number of Trees in the Forest:	16	Number of Terms:	11
Number of Terms Sampled per Split:	2	Bootstrap Samples:	286868
		Minimum Splits per Tree:	10
		Minimum Size Split:	478

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.4367	0.4409	0.4412	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5283	0.5330	0.5329	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2089	0.2081	0.2073	∑ -Log(ρ[j])/n
RMSE	0.2372	0.2367	0.2361	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.1269	0.1265	0.1263	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0736	0.0738	0.0733	∑ (ρ[j]≠ρMax)/n
N	286868	71717	119528	n

Confusion Matrix

Training

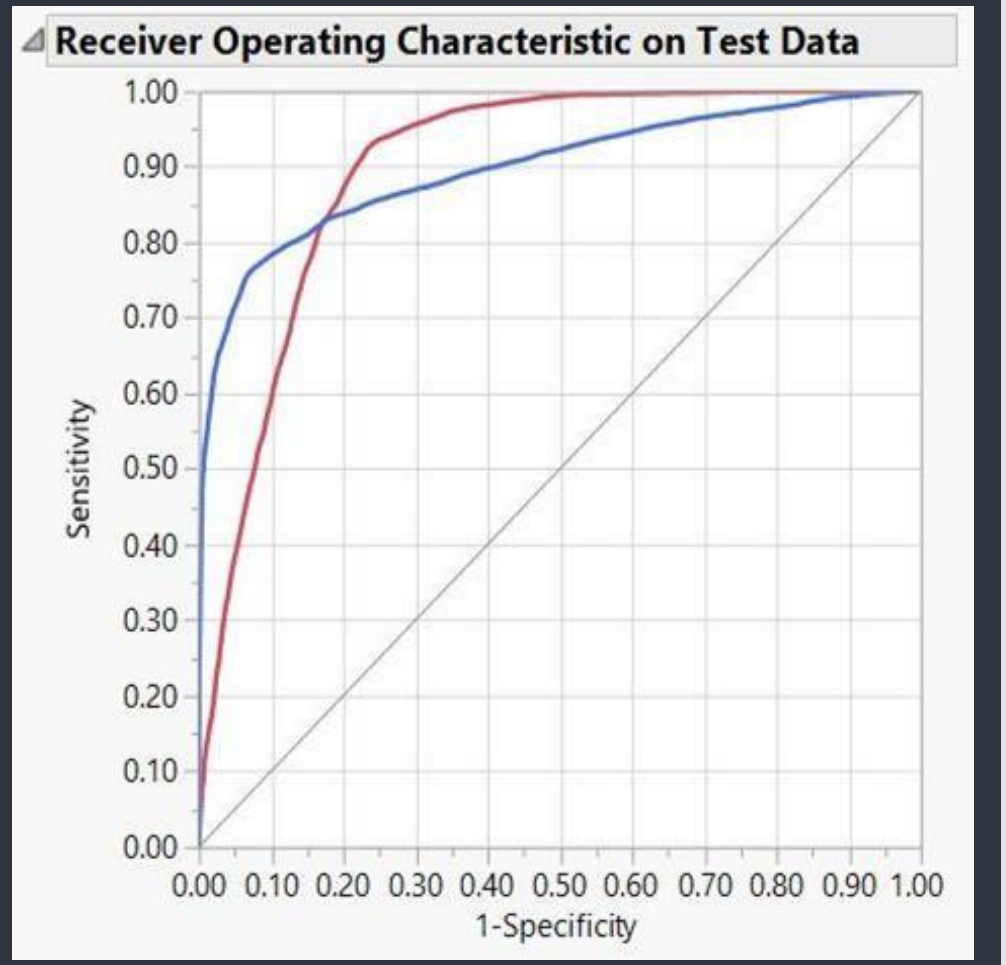
Actual	Predicted Count	
is_churn	0	1
0	250970	906
1	20209	14783

Validation

Actual	Predicted Count	
is_churn	0	1
0	62688	228
1	5064	3737

Test

Actual	Predicted Count	
is_churn	0	1
0	104553	388
1	8368	6219



Details:

Bootstrap forest is used which led to a “Misclassification Rate” of 0.0733 on test data set

Misclassification Rate:

Training : 0.0736
Validation : 0.0738
Test : 0.0733

Prediction of customer attrition

	Most Likely is_churn					
	0			1		
	Validation			Validation		
is_churn	Training	Validation	Test	Training	Validation	Test
0	248852	62163	103733	3024	753	1208
1	14622	3657	6022	20370	5144	8565

	Most Likely is_churn 2					
	0			1		
	Validation			Validation		
is_churn	Training	Validation	Test	Training	Validation	Test
0	250698	62633	104453	1178	283	488
1	18161	4508	7540	16831	4293	7047

	Cutoff_.35					
	0			1		
	Validation			Validation		
is_churn	Training	Validation	Test	Training	Validation	Test
0	245756	61376	102469	6120	1540	2472
1	12288	3076	5070	22704	5725	9517

	Boot_cut_off_.4					
	0			1		
	Validation			Validation		
is_churn	Training	Validation	Test	Training	Validation	Test
0	249734	62378	104070	2142	538	871
1	16164	4015	6677	18828	4786	7910

	Cutoff_.4					
	0			1		
	Validation			Validation		
is_churn	Training	Validation	Test	Training	Validation	Test
0	247099	61710	102987	4777	1206	1954
1	13155	3307	5435	21837	5494	9152

	Boot_cut_off_.35					
	0			1		
	Validation			Validation		
is_churn	Training	Validation	Test	Training	Validation	Test
0	246411	61535	102790	5465	1381	2151
1	13380	3279	5529	21612	5522	9058

Neural

Model Launch

Model NTanH(3)

Training

is_churn

Measures	Value
Generalized RSquare	0.5735299
Entropy RSquare	0.4815734
RMSE	0.226493
Mean Abs Dev	0.1026658
Misclassification Rate	0.0652844
-LogLikelihood	55152.875
Sum Freq	286868

Confusion Matrix

Actual	Predicted Count	
is_churn	0	1
0	247810	4066
1	14662	20330

Confusion Rates

Actual	Predicted Rate	
is_churn	0	1
0	0.984	0.016
1	0.419	0.581

Validation

is_churn

Measures	V
Generalized RSquare	0.579
Entropy RSquare	0.48
RMSE	0.225
Mean Abs Dev	0.10
Misclassification Rate	0.064
-LogLikelihood	1368
Sum Freq	7

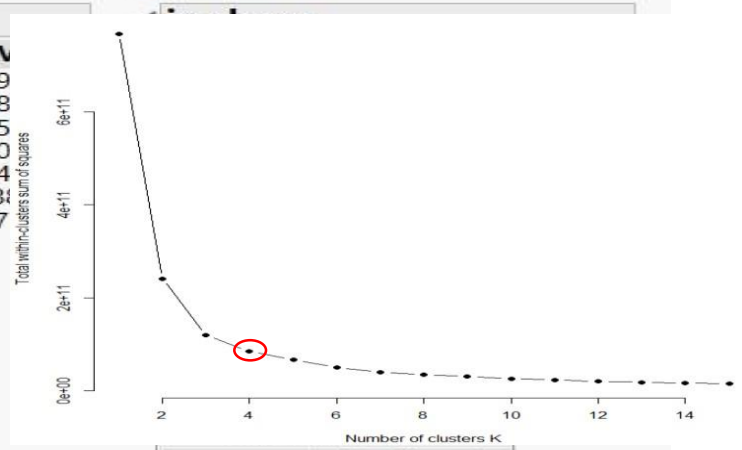
Confusion Matrix

Actual	Predicted Count	
is_churn	0	1
0	61919	997
1	3650	5151

Confusion Rates

Actual	Predicted Rate	
is_churn	0	1
0	0.984	0.016
1	0.415	0.585

Test



Actual	Predicted Rate	
is_churn	0	1
0	0.985	0.015
1	0.417	0.583

Model NTanH(3)NLinear(1)NGaussian(1)NTanH2(3)NLinear2(1)NGaussian2(1)

Training

is_churn

Measures	Value
Generalized RSquare	0.5910705
Entropy RSquare	0.4993924
RMSE	0.2228062
Mean Abs Dev	0.099435
Misclassification Rate	0.0618647
-LogLikelihood	53257.198
Sum Freq	286868

Confusion Matrix

Actual	Predicted Count	
is_churn	0	1
0	249005	2871
1	14876	20116

Confusion Rates

Actual	Predicted Rate	
is_churn	0	1
0	0.989	0.011
1	0.425	0.575

Validation

is_churn

Measures	Value
Generalized RSquare	0.598496
Entropy RSquare	0.5066469
RMSE	0.2216983
Mean Abs Dev	0.0988603
Misclassification Rate	0.0617009
-LogLikelihood	13172.878
Sum Freq	71717

Confusion Matrix

Actual	Predicted Count	
is_churn	0	1
0	62202	714
1	3711	5090

Confusion Rates

Actual	Predicted Rate	
is_churn	0	1
0	0.989	0.011
1	0.422	0.578

Test

is_churn

Measures	Value
Generalized RSquare	0.5962575
Entropy RSquare	0.5046785
RMSE	0.221358
Mean Abs Dev	0.0986739
Misclassification Rate	0.0608811
-LogLikelihood	21963.006
Sum Freq	119528

Confusion Matrix

Actual	Predicted Count	
is_churn	0	1
0	103800	1141
1	6136	8451

Confusion Rates

Actual	Predicted Rate	
is_churn	0	1
0	0.989	0.011
1	0.421	0.579