TELECOM CHURN

Assignment Submission

Student Name: Tarriq Ferrose Khan
Batch : DSC 68

upGrad

upGrad & IIITB | Data Science Program - May 2024

Business Problem (Overview)

Telecom industry experiences an average of 15-25% annual churn rate., *retaining high profitable customers is the number one business goal. To* reduce customer churn, this project, will analyze customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn and identify the main indicators of churn. There are various ways to define churn, such as: **Revenue-based churn, Usage-based churn and in** this project, **usage-based definition** churn is used for Analysis.

High-value churn are identified by a specified business logic through which significant revenue leakage will be reduced.

Customer behavior during churn:

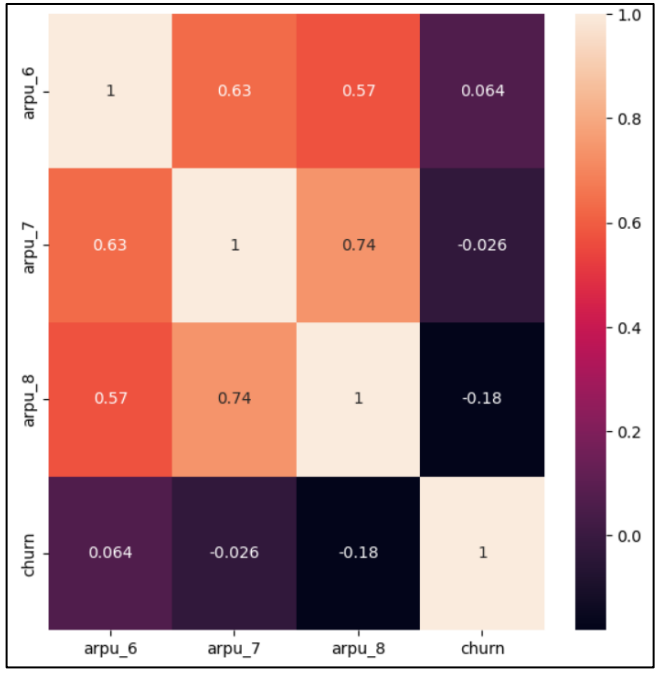
'Good' phase: Customer is happy with the service and behaves as usual.'
'Action' phase: The customer experience starts to sore in this phase, usually shows different behavior than in the 'good' months, it is crucial to identify high-churn-risk customers in this phase.
'**churn**' phase: The customer is said to have churned.

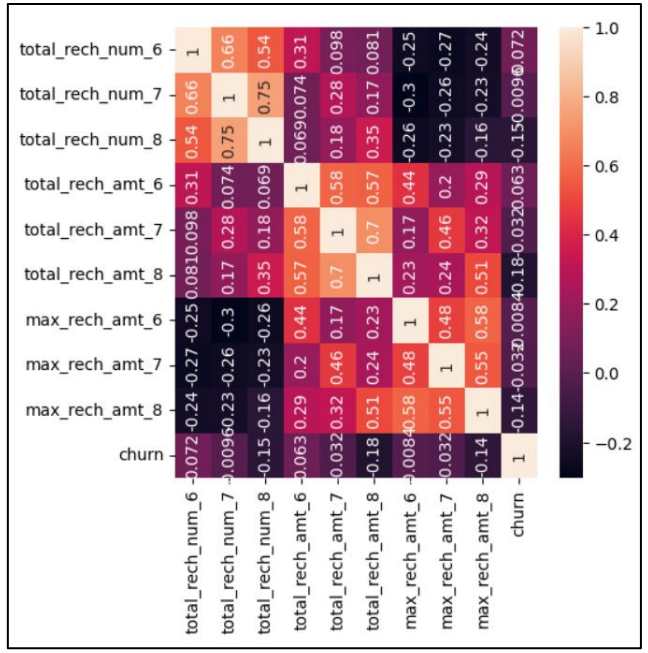
In this case, a four-month window, the first two months are the 'good' phase, the third month is the 'action' phase, and the fourth month is the 'churn' phase, considered.

Approach Note

Data Cleaning and Preparation:

- **High Valued Customers**: Source Data (Telecom_churn.csv) contained 100,000 records and 226 columns, filtering the High Valued Customers returned 30K records, out of which Churn is tagged basis the business logic provided and this data is used for further analysis. g
- **Data Cleaning: Null Columns with null% > 60 are removed and columns that will not add value like Circle_id and date based columns are removed.**
- **Outlier Handling:** As there are huge number of columns, a user defined function is written to identify and eliminate the Outlier values.
- **Correlation :** Key group of attributes were correlated and removed columns where all values where same like 0.
- Total Columns were created for each group of columns like incoming, outgoing to understand the total movement of data.





Churn Rate and Handling Imbalance

- Churn Rate was ~8% and hence class imbalance technique were used to create a balanced Sample.
- The results of the metrics were stored in a data frame.
- Checked for the highest recall score
- Logistic Regression with ADASYN turned out to be the best sample and hence used for resampling
- After resampling , the churn rate turned out to be ~49%.

	Group	Method	Perc_churn	Recall	Rank
0	LR	ADASYN	23.516759	0.826754	0
1	LR	Under Sampling	23.964868	0.817982	1
2	RF	Under Sampling	17.207385	0.811404	2
3	LR	SMOTE+TOMEK	21.688475	0.811404	3
4	LR	Over Sampling	22.262054	0.811404	4
5	LR	SMOTE	21.598853	0.807018	5
6	DT	Under Sampling	26.707295	0.767544	6
7	RF	SMOTE	9.715003	0.671053	7
8	RF	SMOTE+TOMEK	9.930095	0.668860	8
9	RF	ADASYN	9.840473	0.660088	9
10	DT	ADASYN	13.335723	0.576754	10
11	DT	SMOTE+TOMEK	12.385732	0.561404	11
12	DT	SMOTE	12.367808	0.524123	12
13	RF	Over Sampling	5.807492	0.504386	13
14	DT	Base	9.535759	0.489035	14
15	DT	Tomek Links	9.195196	0.471491	15
16	RF	Tomek Links	5.431081	0.451754	16
17	LR	Tomek Links	5.807492	0.416667	17
18	DT	Over Sampling	7.653701	0.414474	18
19	RF	Base	4.749955	0.410088	19
20	LR	Base	4.265997	0.324561	20

Feature Scaling

- Most of variables are of Similar scale.
- `StandardScaler` was used to fit and Scale the values.

	arpu_6	arpu_7	arpu_8	onnet_mou_6	onnet_mou_7	onnet_mou_8	offnet_mou_6	offnet_mou_7	offnet_mou_8	roam_ic_mou_6	...	sep_vbc_3g	recharge_first_2_months	high_valued_cust
0	0.193591	0.603352	1.410365	1.004738	0.354050	1.123915	-0.049903	1.041829	1.661525	-0.234376	...	-0.089694	0.780368	
1	-0.048069	-0.012171	1.049418	0.914238	1.722133	2.493447	-0.683018	-0.732165	-0.541454	-0.234376	...	-0.089694	0.039260	
2	-0.523957	-0.646955	0.490368	-0.531563	-0.570986	-0.345337	-0.523694	-0.532783	-0.233264	-0.234376	...	-0.089694	-1.099517	
3	-0.462010	-0.235626	-0.317984	-0.513780	-0.544658	-0.511032	0.979239	1.734622	0.463736	-0.234376	...	-0.089694	-0.376484	
4	-0.475845	-0.787024	-0.296214	-0.788651	-0.769303	-0.584629	-0.221689	0.156693	0.663432	0.603367	...	-0.089694	-1.058201	

5 rows × 134 columns

Model Building, Evaluation , Feature Selection and Hyper Parameter Tuning

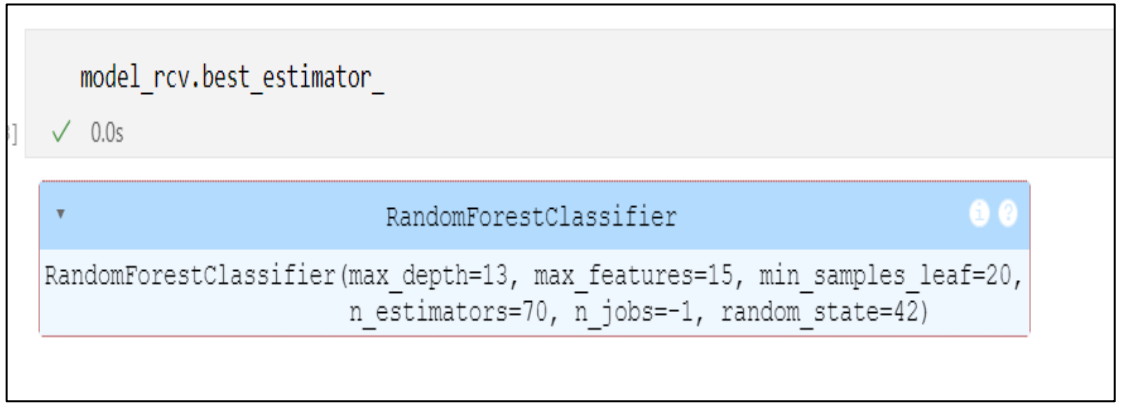
- Both Logistic Regression (LR) and RandomForestClassifier (RF) were Analysed for Accuracy , Recall and other metrics.
- Cross Validation Score for Accuracy and Recall were checked
- Overall Random Forest was performing well.
- Still RFECV was used for Feature Selection on LR and RF
- RFECV selected 26 attributes for LR and for RF it selected 133 attributes.
- `RandomizedSearchCV` used for fine tuning hyper parameters basis which the Final Model was generated
- Final Important Features are derived

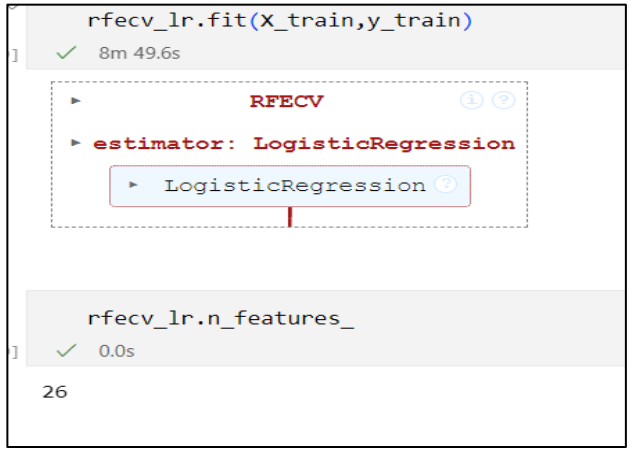
Cross_val_score:Accuracy

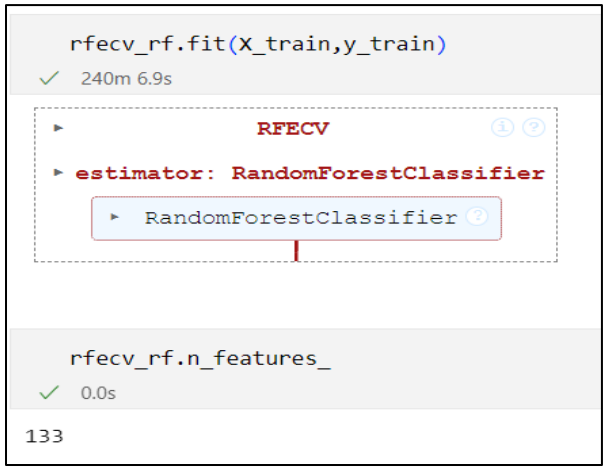
1. lr_cross_val_score Mean (Accuracy)=0.9178331678927952
2. rf_cross_val_score Mean (Accuracy)=0.9418908751742766
3. rf.oob_score_=0.9642692194650879

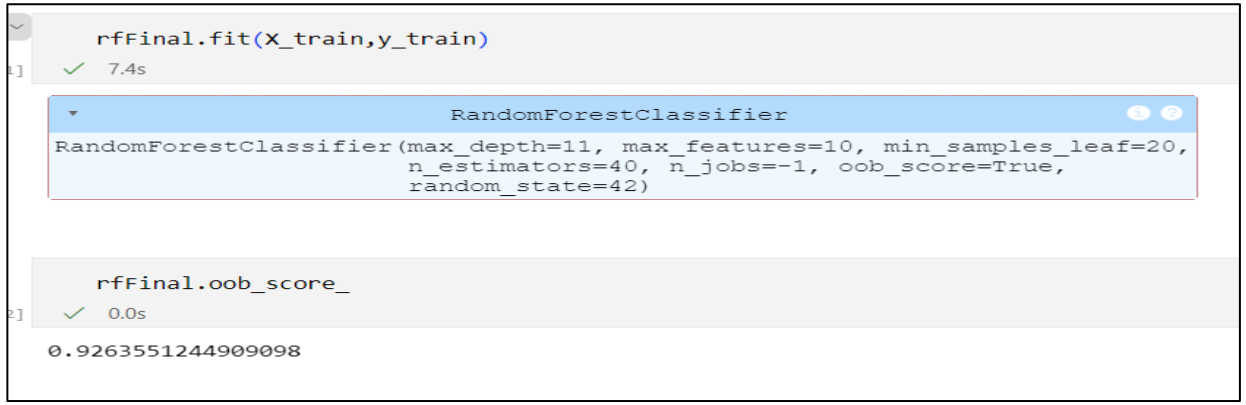
Cross_val_score:Recall

1. lr_cross_val_score_recall Mean (Recall)=0.9151498659270911
2. rf_cross_val_score_recall Mean (Recall)=0.9350331285939012

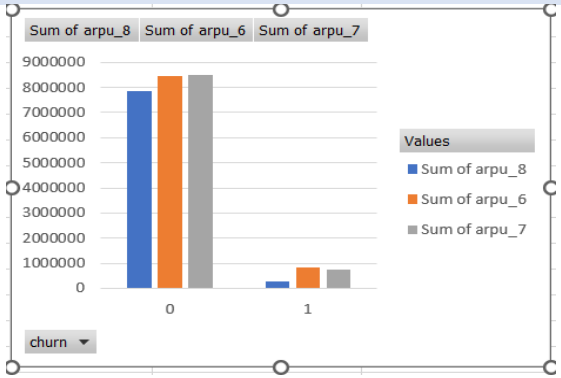


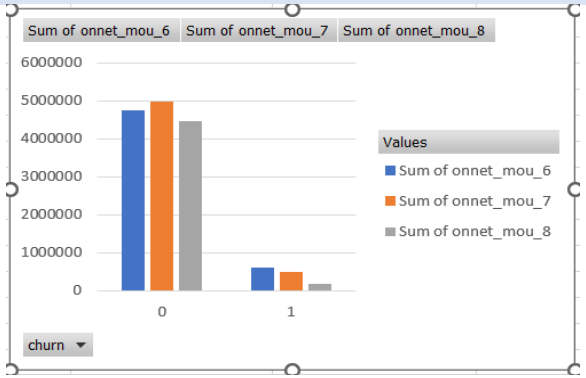


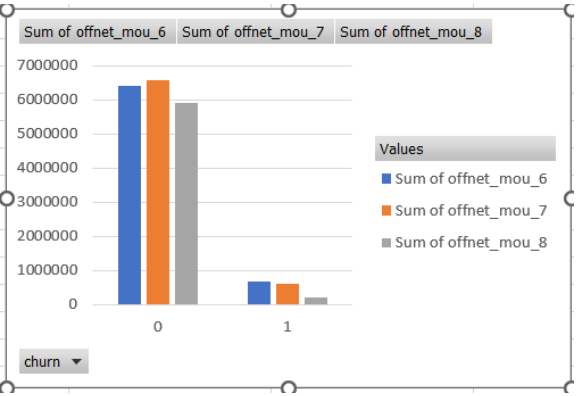


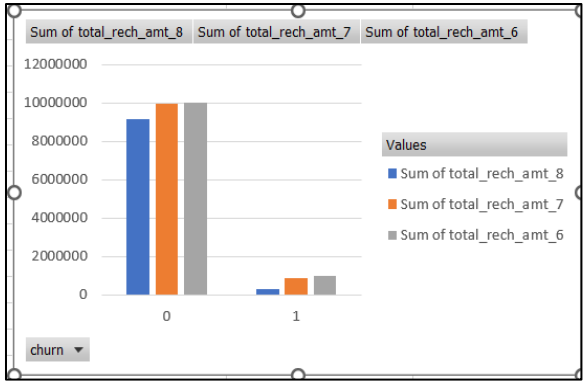


Analysis for RandomForest Selected Important Features: Incoming Vs Outgoing Vs Recharge

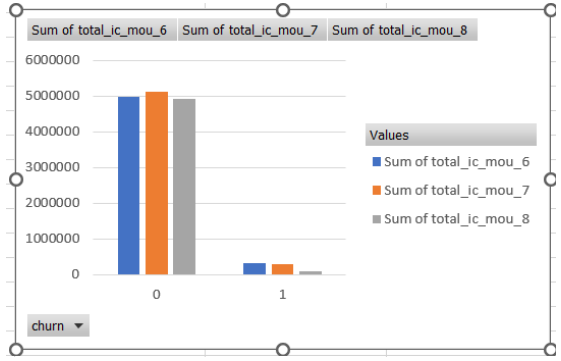


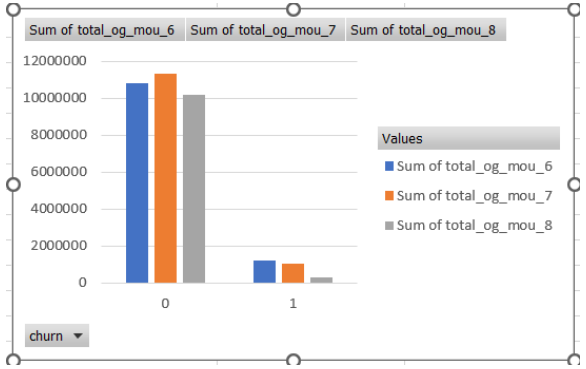


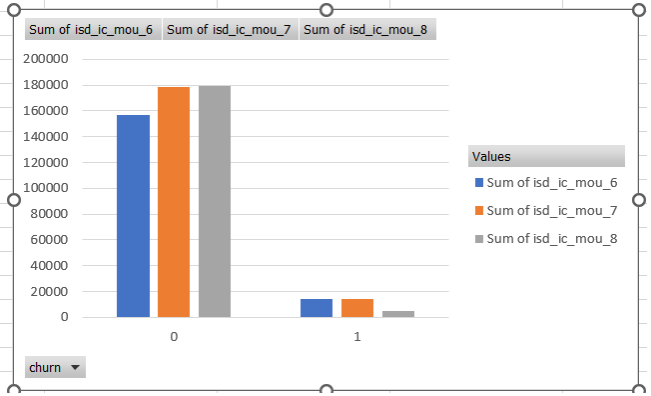


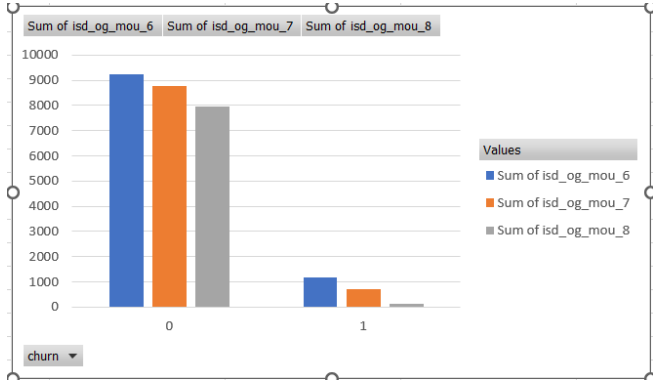


- Though Recharge is increased in Action phase there is a drop in incoming and outgoing mou (Minutes of Usage)
- Which indicates the current Validity of the talk time given for the customers may be one of the key attributes for the churn as there are **unused voice call talk time** in both incoming and outgoing calls , and hence the drop in trend.
- This will lead to customer frustration and tend to switch to competitors.
- **ISD incoming shows high trend vs ISD Outgoing, which indicates an offer on Night Talk time , would retain customers**

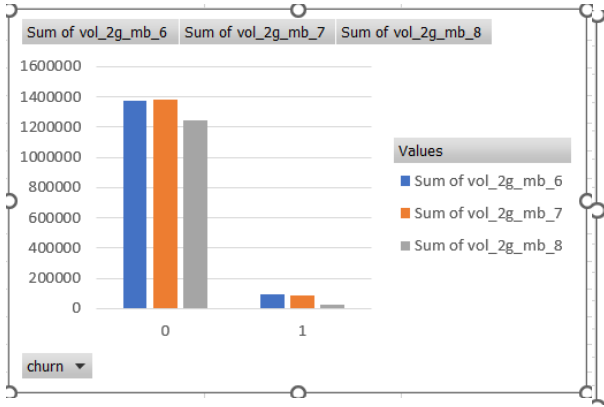


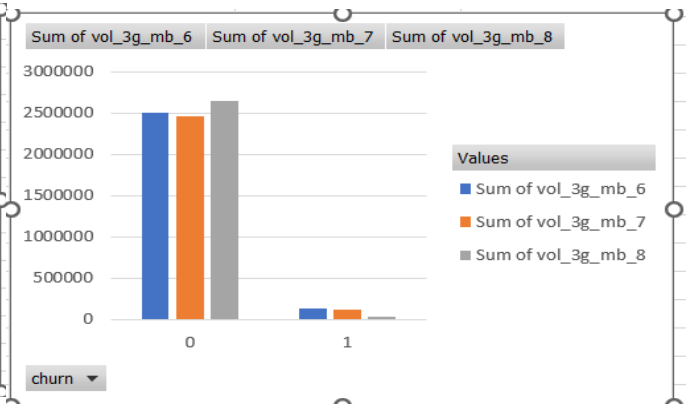


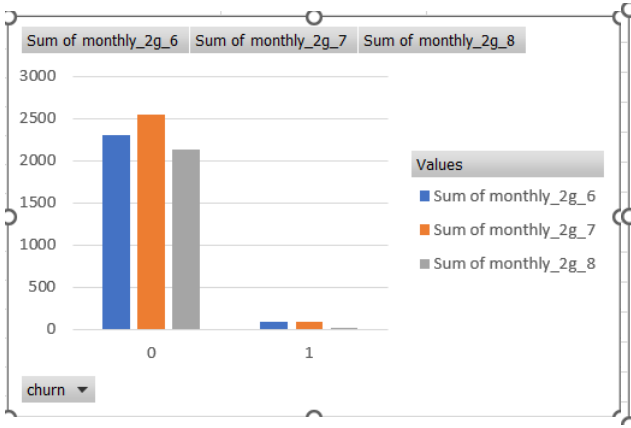


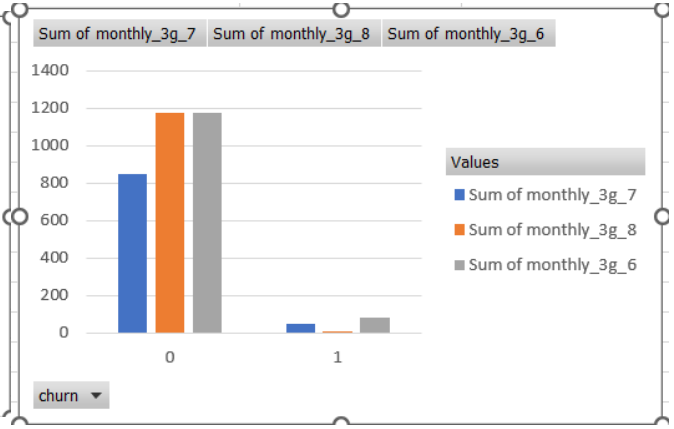


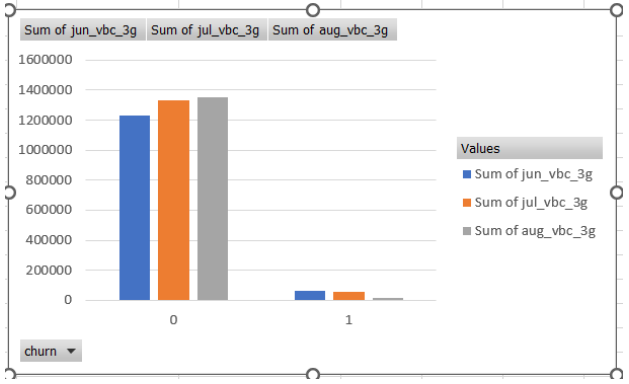
Charts for RandomForest Selected Important Features: Incoming Vs Outgoing Vs Recharge











- Volume Based Cost and 3G shows increased Trend in Action phase.
- There is a clear drop in 2g , which may be due to customers adopting to latest mobile devices and prefer 3G services.
- Targeting offer for 3G services would pay off in terms of customer retention.
- Volume Based Cost make them satisfied as customers don’t have to worry about the unused data.
- For Service provider also the burden of carrying the unused data forward wont’ be there.

Analysis Summary

- From the current model selected and analysing the data its clear that customers are actually active on the network
- They in-fact pay more but use less may due to the poor service plan options provided to the customers.
- They are of more ISD call receivers and may be binge watchers, showing the respective trends.
- Revising the plans to target these customers would RETAIN them and continue with the current Service Providing Company.

THANK YOU