Capstone Project – Telecom Churn

Using Logistic Regression

By A.R. Premkumar

BUSINESS OBJECTIVE

- In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate. Given the fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, **customer retention** has now become more important than customer acquisition.

- For many incumbent operators, *retaining highly profitable customers is the number one business goal*.

- To reduce customer churn, telecom companies need to **predict which highly profitable customers are at risk of churn.**

- The goal is to **develop a model** to predict customers who are likely to Churn

SOLUTION METHODOLOGY

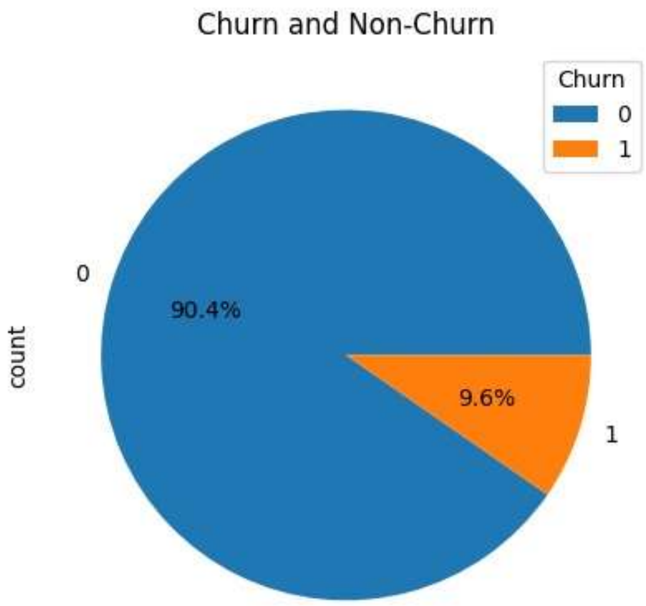
- Data Cleaning and manipulation
- Exploratory Data Analysis
- Model Building
- Model Evaluation
- Model Prediction on Testset
- Inferences
- Recommendation

DATA CLEANING

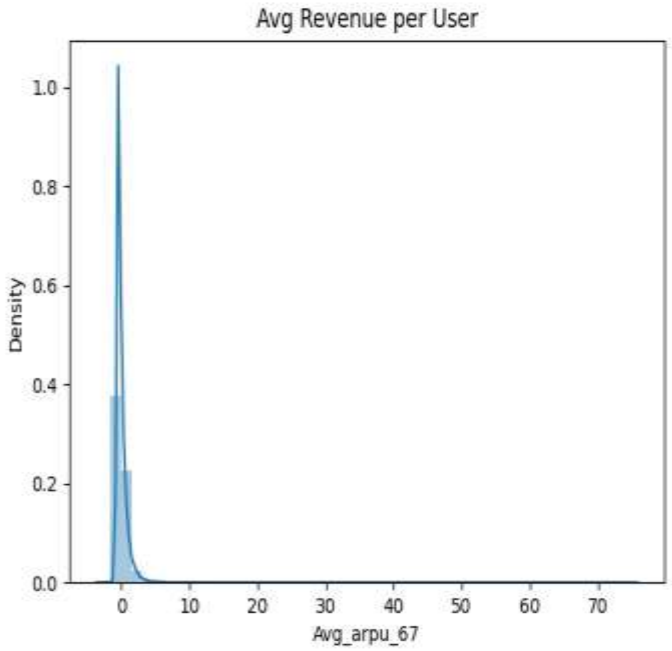
- There are 226 columns with high number of missing values and since we have around 99999 data points we can eliminate the columns that are less relevant to this project;
- We dropped mobile_number, circle_id, columns that end with “-9”, since they are of no use to us;
- Columns like 'loc_og_t2o_mou', 'std_og_t2o_mou', 'loc_ic_t2o_mou', 'last_date_of_month_6’, etc., which has unique values of 0s and 1s which are of no use;
- Filled Nan values with zeros on columns like date_of_last_rech_data_9, date_of_last_rech_data_6, date_of_last_rech_data_8, date_of_last_rech_data_7, av_rech_amt_data_6, fb_user_6, total_rech_data_6, et.
- Identified high collinearity between columns and deleted columns 'arpu_2g_6', 'arpu_2g_7', 'arpu_2g_8', 'arpu_2g_9', 'arpu_3g_6’, etc.
- Add column 'Avg_arpu_67’ by taking avg of columns 'arpu_6’ & 'arpu_7’, and deleted both columns 6 & 7.

UNIVARIATE ANALYSIS

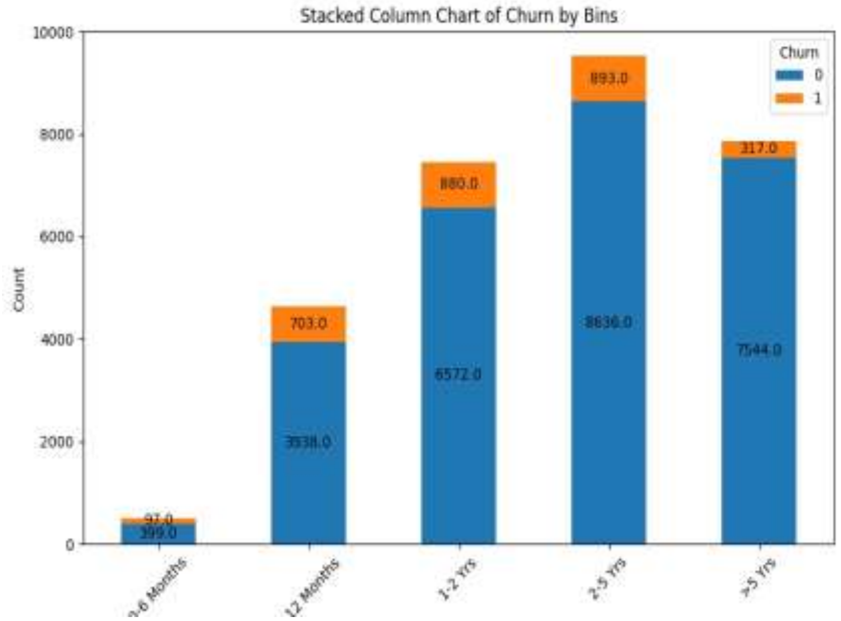
Churn Count



Average Revenue per user Months 6 & 7



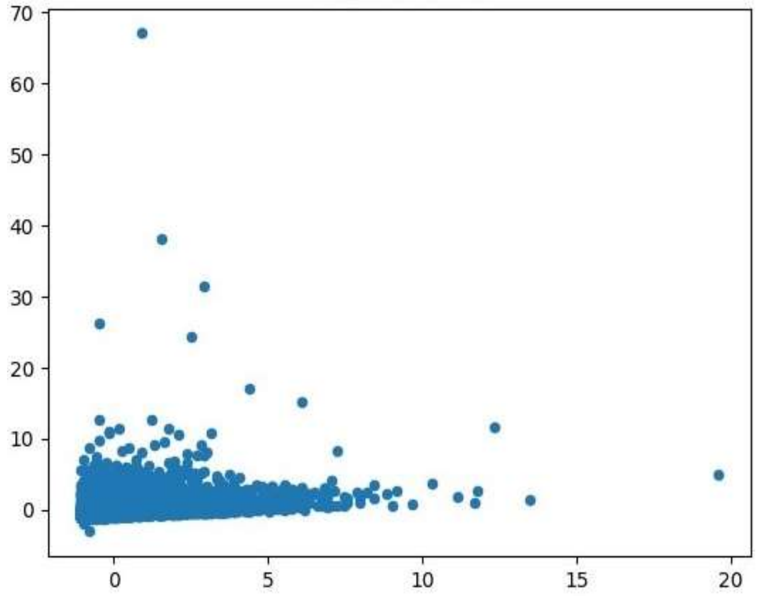
Churn vs Non-Churn



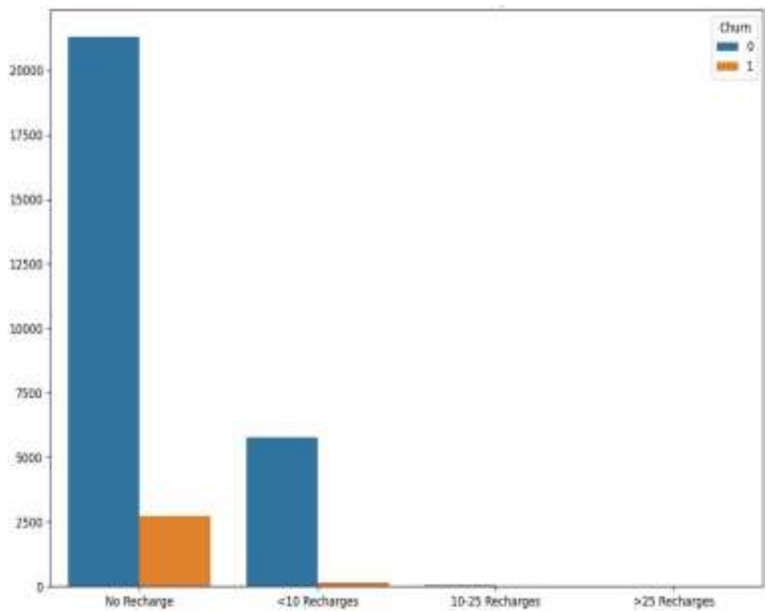
- Definitely there's class imbalance in Churn and Non-Churn counts. This is tackled using SMOTE before logistics regression.
- Avg Revenue per user for the good phase of months 6 & 7 has been arrived to identify the probability density function.
- Its evident that customer retention grows stronger over long duration. In other words, the Churn rate is high in first 6 months.

BI-VARIATE ANALYSIS

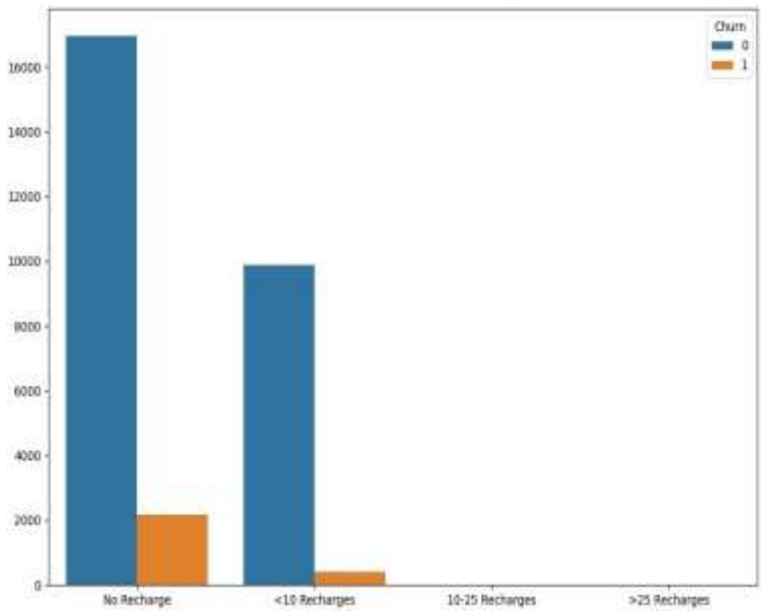
Avg Revenue vs Total Reach



Total Reach Data vs Data Group



Total Reach Number vs Number Group



- Avg Revenue in the action phase of 8th month is positively correlated with Total Reach
- Total recharge in the action phase of 8^{th} month has high correlation with Churn Rate
- Total Reach Number in the action phase of 8^{th} month has high correlation with Churn Rate

MODEL BUILDING

- Slitting the data into train and test split with 70:30 ratio
- Scale numerical feature using MinMax scaler
- Use Recursive feature Elimination (RFE) to identify 20 most important feature
- Use p-value and Variance inflation factor to eliminate statistically insignificant features
- Finally, we ended up with 19 features for the model

MODEL EVALUATION

[94]:

Generalized Linear Model Regression Results

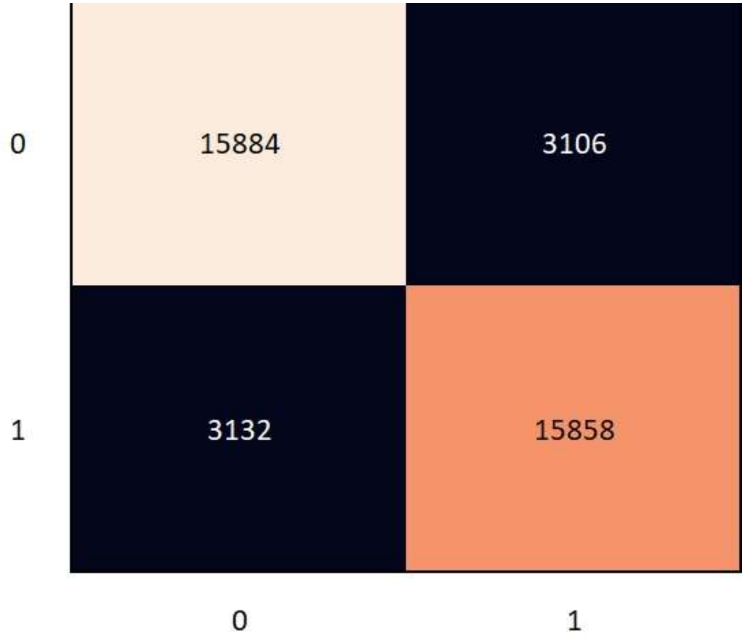
Dep. Variable:	Churn	**No. Observations:**	37980
Model:	GLM	**Df Residuals:**	37960
Model Family:	Binomial	**Df Model:**	19
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-14904.
Date:	Tue, 02 Apr 2024	**Deviance:**	29808.
Time:	13:07:09	**Pearson chi2:**	6.40e+05
No. Iterations:	9	**Pseudo R-squ. (CS):**	0.4520
Covariance Type:	nonrobust		

[92]:

	Features	VIF
12	total_rech_amt_8	16.84
0	arpu_8	16.31
1	offnet_mou_8	8.63
8	loc_ic_mou_8	8.25
9	total_ic_mou_8	6.08
3	std_og_t2m_mou_8	5.96
5	total_og_mou_8	3.59
7	loc_ic_t2t_mou_8	3.47
14	max_rech_data_8	2.55
6	loc_ic_t2t_mou_7	2.17
16	monthly_3g_8	2.17
11	total_rech_num_8	1.78
18	Avg_arpu_67	1.77

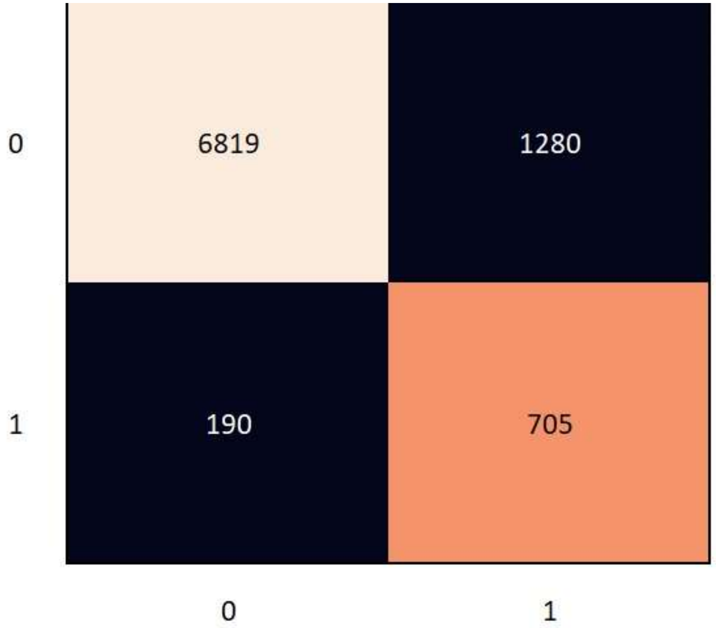
MODEL EVALUATION

TRAINING SET



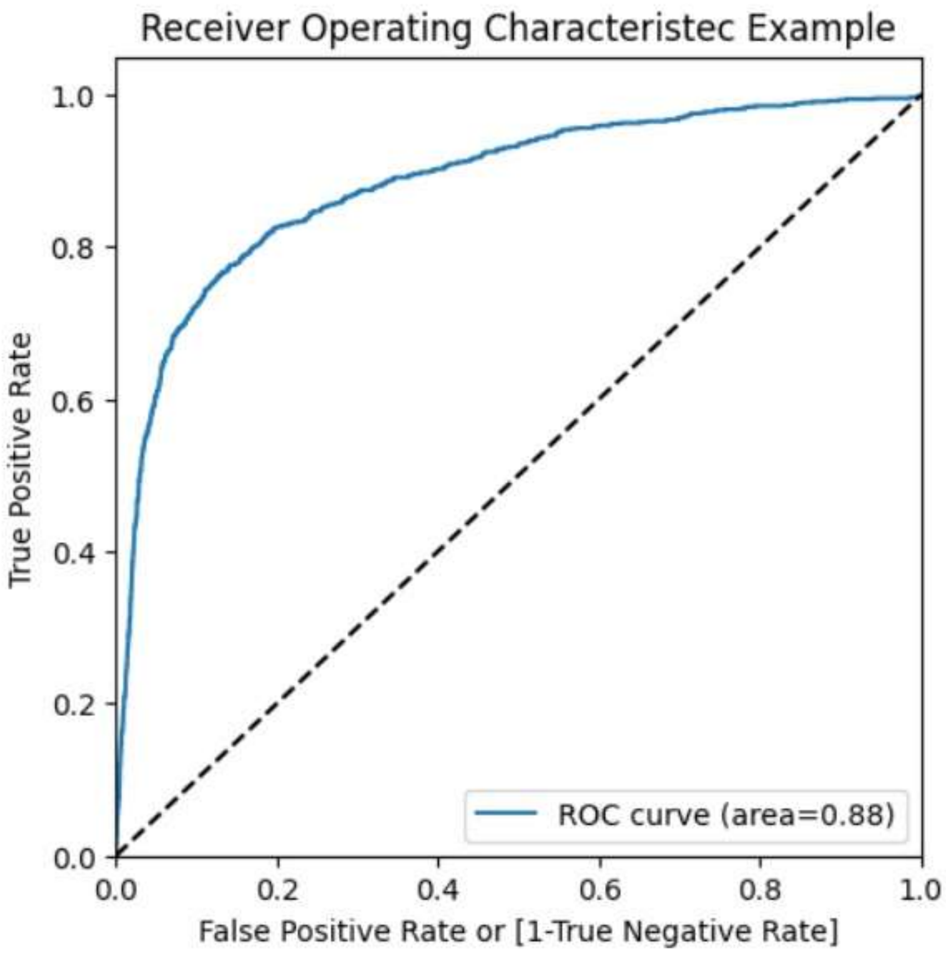
Accuracy	83.58%
Sensitivity	83.51%
Specificity	83.64%
Precision	35.52%
Recall	83.62%

TEST SET

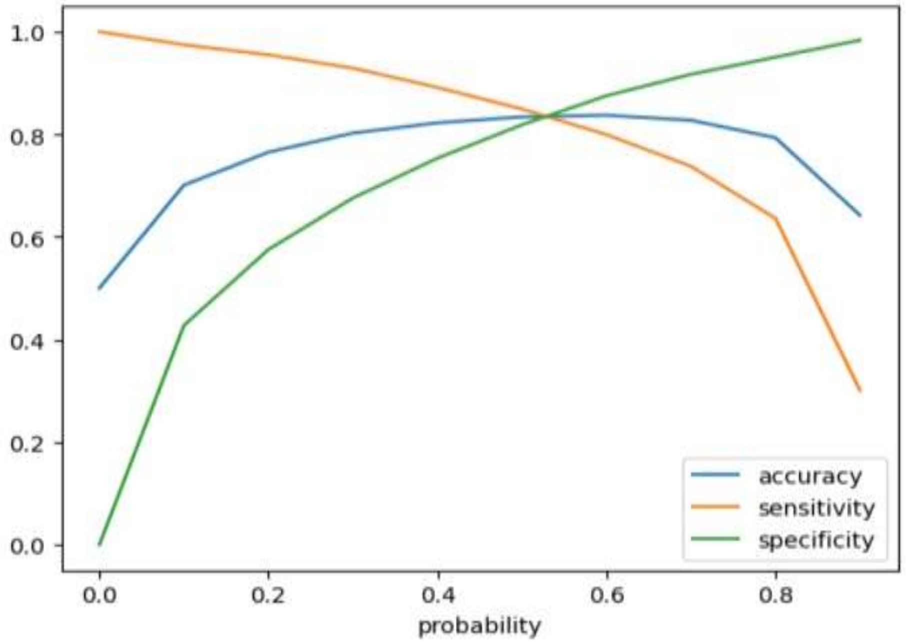


Accuracy	83.66%
Sensitivity	78.77%
Specificity	84.19
Precision	35.52%
Recall	78.77%

MODEL EVALUATION - ROC/CUTOFF



	probability	accuracy	sensitivity	specificity
0	0.0	50.00%	100.00%	0.00%
0.1	0.1	70.09%	97.45%	42.73%
0.2	0.2	76.51%	95.49%	57.53%
0.3	0.3	80.23%	92.89%	67.57%
0.4	0.4	82.22%	89.13%	75.32%
0.5	0.5	83.38%	84.92%	81.84%
0.6	0.6	83.72%	79.93%	87.50%
0.7	0.7	82.73%	73.77%	91.69%
0.8	0.8	79.32%	63.61%	95.03%
0.9	0.9	64.22%	30.09%	98.35%



INFERENCES

Top three variables in your model which contribute most towards the probability of a customer getting churned

a. ***Age on Network 'aon',***
b. ***Total Reach Data 'total_rech_data_9',***
c. ***Total Reach Num Group 'total_rech_num_9'***

Top categorical/dummy variables in the model which should be focused the most on in order find the maximum probability of Churn

a. **'total_rech_data_group_8'**
b. ***'total_rech_num_group_8'***

RECOMMENDATION

Depending on the requirements the model needs to be tweaked such that

- New clients are more likely to churn
- Clients with higher Monthly Charges are also more likely to churn
- Tenure and Monthly Charges are probably important features
- Customers with the first 4 additional services (Security, Backup, Protection, Tech support) are less likely to churn
- Streaming services are not likely to associate with churn
- Marketers should be careful with the tradeoff between precision and recall.
- We recommend future tuning out prediction model before we offer discount to retain customers.
- Limited Data (7,043 observations with 26 variables)
- Imbalanced Data (26.54% of churned customers)
- Bias: a point in time
- More features and more data to train model
- Not possible to retain high precision when aiming high recall

Thank You