Case Study On Telecom Churn Prediction

Vandit Sardana | Damanpreet Baweja | Bhavya Sinha

Telecom Churn Problem Statement

- We have received customer-level information for a span of four consecutive months - June, July, August and September. For understanding customer churn pattern we have segregated the months as

1) Good Phase
2) Action phase
3) Churn phase

- **Business Objective** : To predict the churn in the last month (September) using the data from the first three months (June, July & August). To do this task well, understanding the typical customer behavior during churn will be helpful.

Problem Solving Strategy

Following are 5 Strategies -

- Data importing
- Data Inspection
- Data Cleaning
- Missing Value
- Treatment Outlier
- Treatment

- EDA
- Understanding the features and relationships among them
- Target feature

- Scaling Features
- Data Preparation
- Building different
- ML models to predict churn
- Identify Important features

- Test The model on
- Train Set Evaluate
- Model using different metrics
- Test The model on Test Set Evaluate
- Model using different metrics

- Suggestions and Recommendations

Analyzing the problem statement and approach towards solution

- The dataset contains customer level information for June , July , august and September . as the problem is concerning identifying potential churned customers and important fac tors influencing their behaviors, we need to build a suitable classification model machine learning techniques.
- As the dataset does not contain the churn or non-churn labels, we have to derive a feature (label) to train the model which can subsequently identify potential churned customers.
- We will also need to identify important fac tors influencing churn so one of the requirements is to build a simple and interpretable model through which important fac tors can be correlated with final business objective

High Value Customers

- In the Indian and South-Asian markets, approximately 80% of revenue comss from the top 20% of customers (called high value customers) . Focusing on them, we can reduce significant revenue leakage. Thus we have selected high value customers (in terms of recharge amount in June and July) and focused our analysis on them.
- Also as the dataset does not contain the churn or non-churn labels, we have tagged the high value customers as churn or non-churn based on following four distinct features.

1. Total minutes of usage of incoming calls for month of September.
2. Total minutes of usage of outgoing calls for month of September.
3. Total volume of usage of 2g internet for month of September.
4. Total volume of usage of 3g internet for month of September.

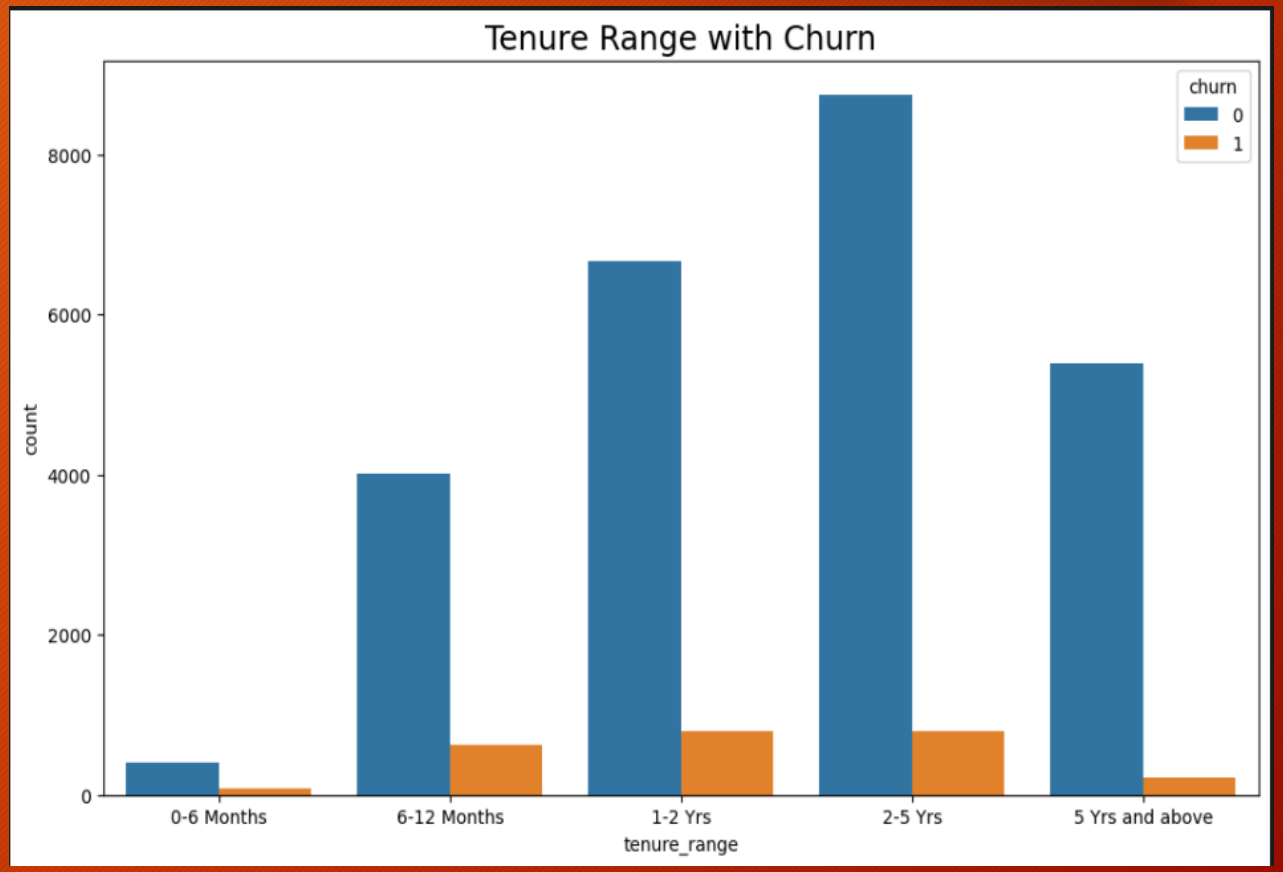
- We have subsequently removed all the records for month of September from our analysis.

Analyzing Churn with Tenure

- We have analyzed the churn pattern with derived variable ‘ tenure’ by bucketing customers in the following bins:-

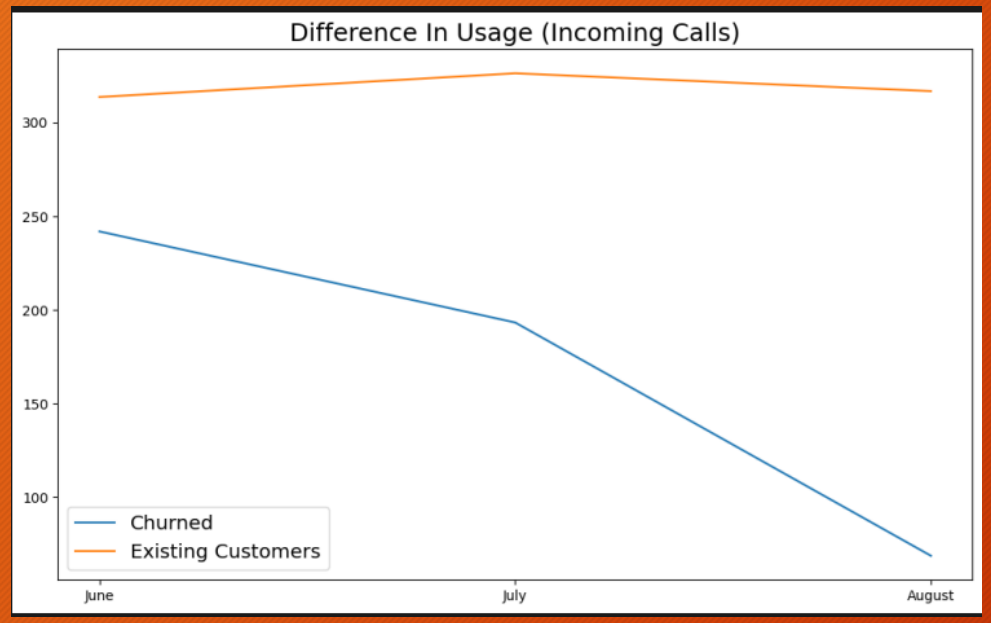
1. 0-6 months
2. 6-12 months
3. 1-2 years
4. 2-5 years
5. 5 years and above

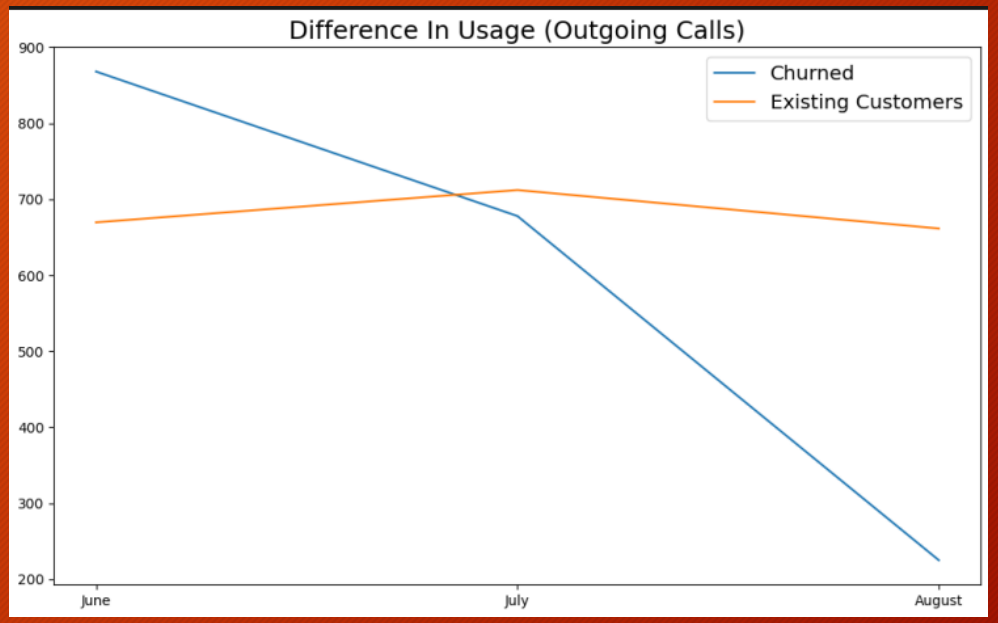
- Its evident that churn percentage with respect to non churn in a particular period is comparatively higher at the beginning i.e 0 - 6 months window.



Understanding Churn With Avg Minutes Of Usage And Avg Volume Of Internet Usage

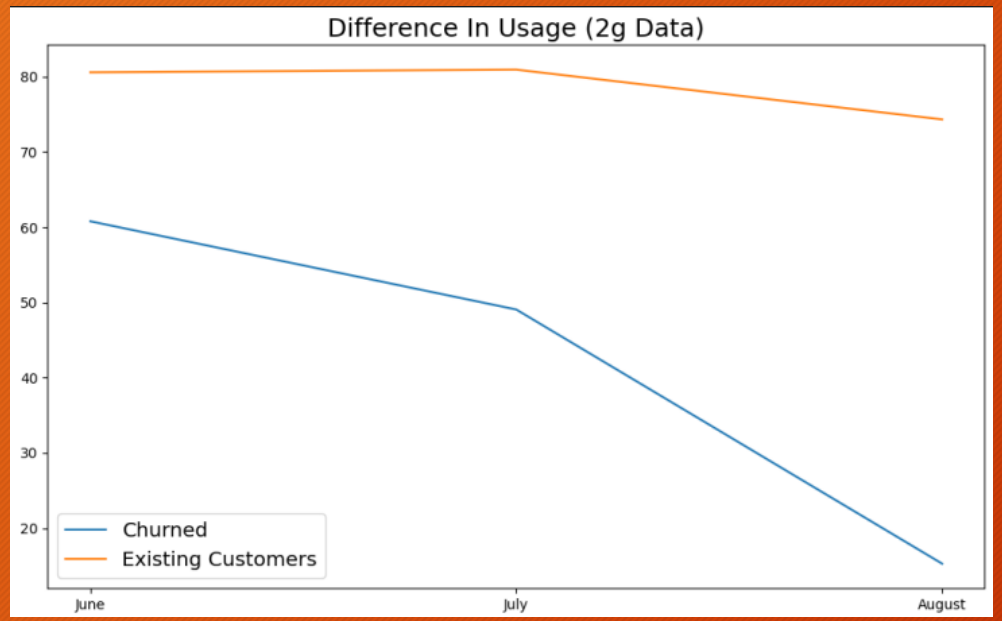
We have defined few derived variables to better understand the churn predictors. We have used metrics like average minutes of usage of incoming and outgoing calls and volumes of usage of 2g and 3g internet to understand churn and non churn pat tern

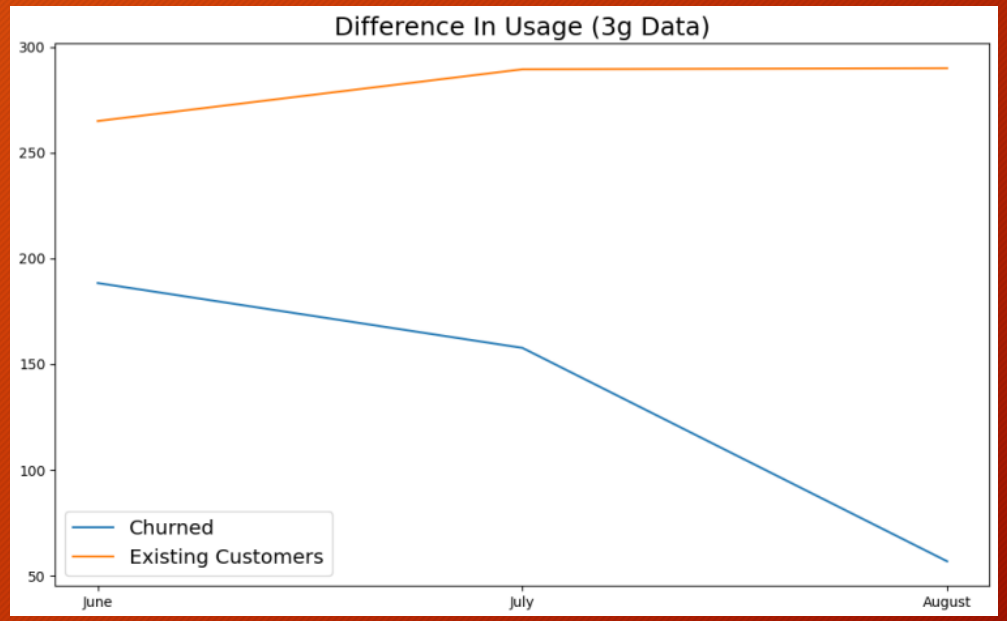




Continue

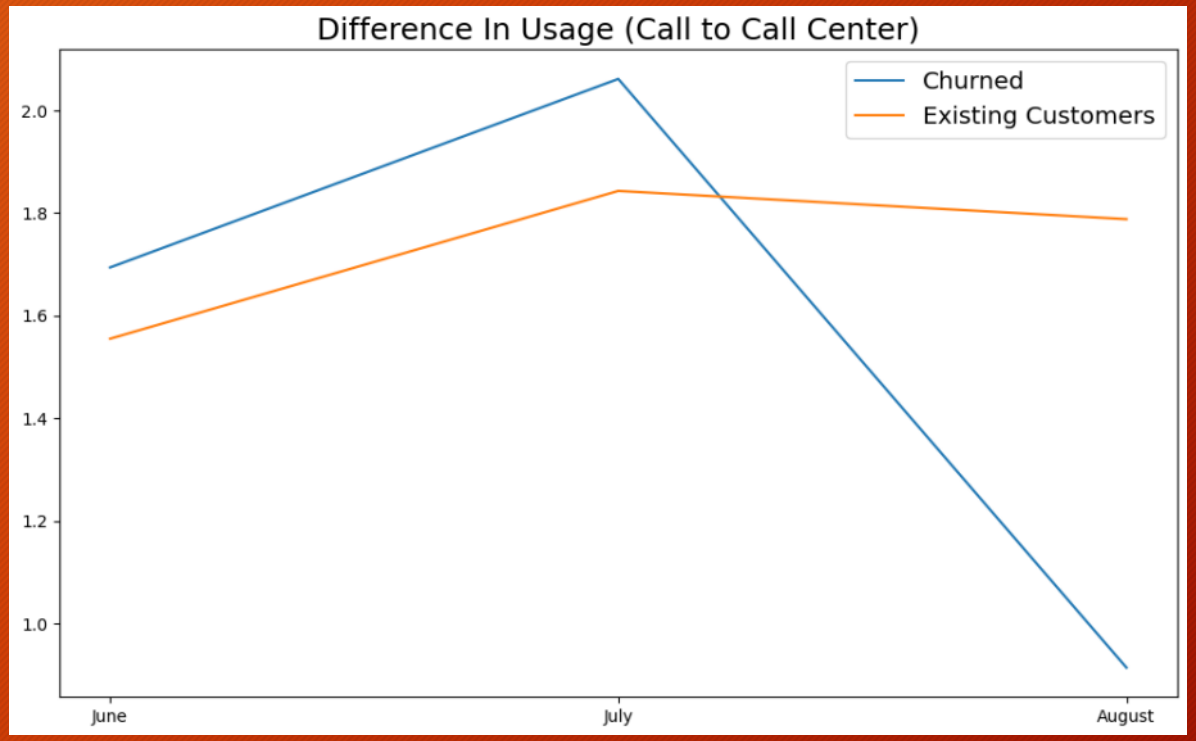
For all the metrics its pretty evident churned customers tend to have a decreasing pat tern in terms of usage of telecom services. The drop is significant specially in the month of august .





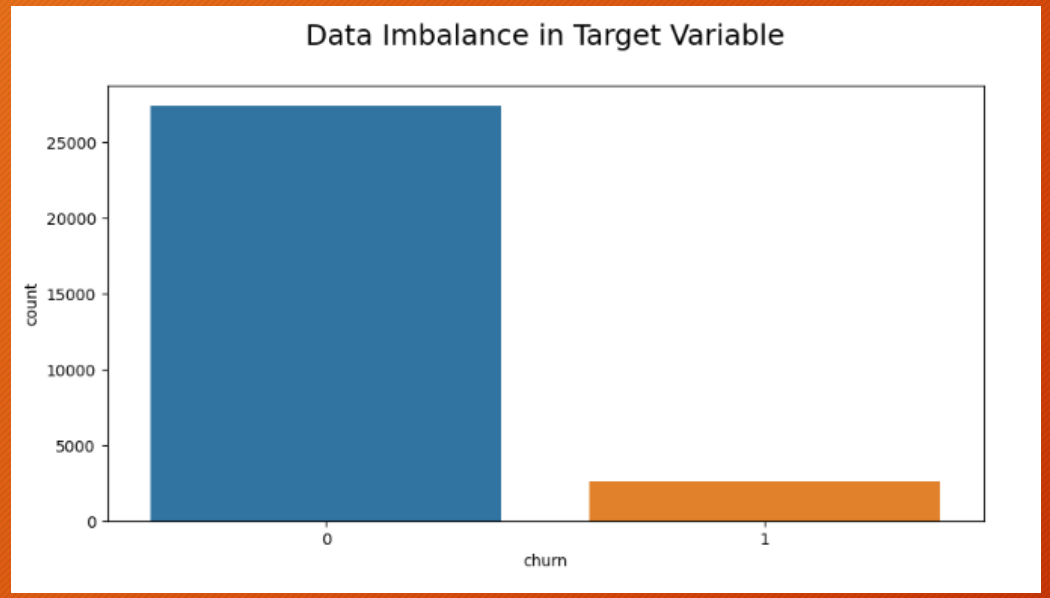
Understanding Customer Behavior With Avg Call To Call Centre Metric

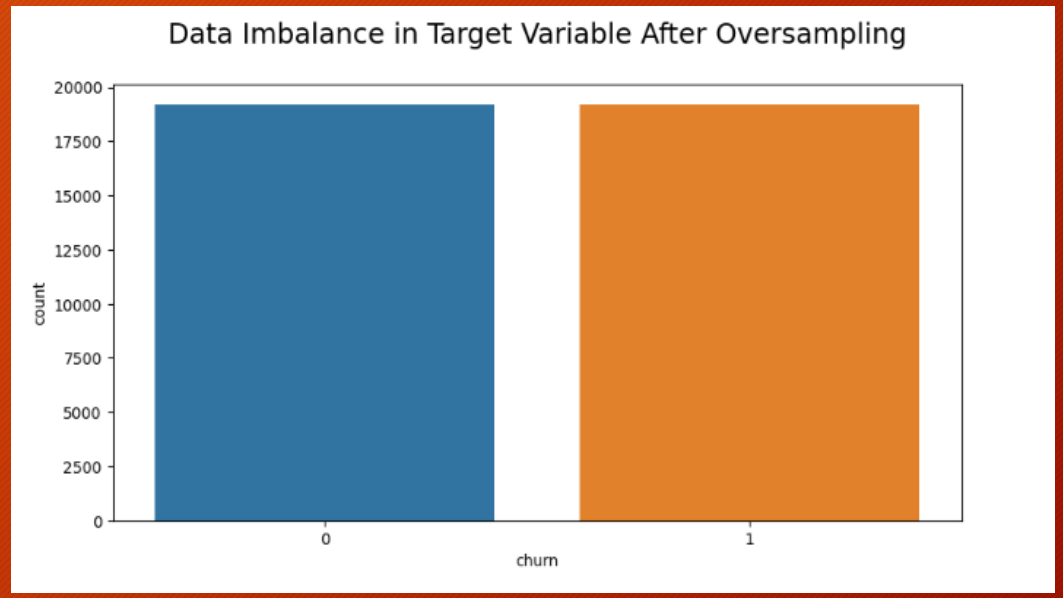
As can be seen, for churned customers call to call centers peaked in July and dropped abruptly in august probably indicating their dissatisfaction with the service from operator



Data Imbalance Present In Churn Rate

In the dataset, we see high imbalance in churn cases as expected. Churn cases represent only around 9% of the total customers. We have used oversampling technique (smote) to tackle this issue and remove bias from our analysis.





Building Machine Learning Model

- As we want to understand important factors influencing customer churn behavior, we built a classic logistic regression model which has high interpretability and thus provides useful insights into the solution.
- In general, in the logistic regression models there is trade off between interpretability and performance , so we have built multiple high performance models to increase classifying capability of our analysis.
- Firstly, we used principal component analysis techniques to reduce the number of features and built logistic regression , decision tree and random forest classifier models using reduced number of features and hyper tuning the parameters to compare performance of the models

Modal Evaluation

- There are multiple model evaluation metrics available such as accuracy , sensitivity , specificity , precision , f1 score etc. Which we have used to evaluate our model's overall performance.
- Given our objective is to identify potential churn customers and retain them, overall accuracy will not help us assess the performance well. Here , we are more interested to correctly identify churned customers than non-churn customers. As misclassifying non-churn customers as churn customers to a certain extent will result in focusing more on few non-churn customers, but misclassification of churn customers as non churn will directly result to loss of high value customers.
- Thus, proper metric to be used in the analysis is sensitivity or recall which represents correct classification of churn cases among all the actual churns.

Modal Evaluation - Metrics

- The evaluation metrics (accuracy and sensitivity) have been summarized below for train and test data (unseen data) with different ml models.

	TRAIN DATA			
	LOGREG WITHOUT PCA	LOGREG WITH PCA	DECISION TREE WITH PCA	RANDOM FOREST WITH PCA
ACCURACY	0.8	0.85	0.85	0.87
SENSITIVITY	0.86	0.92	0.87	0.86

	TEST DATA			
	LOGREG WITHOUT PCA	LOGREG WITH PCA	DECISION TREE WITH PCA	RANDOM FOREST WITH PCA
ACCURACY	0.55	0.8	0.8	0.87
SENSITIVITY	0.87	0.85	0.69	0.69

As can be seen, logistic regression (LOGREG) with PCA provides better accuracy and sensitivity on unseen data.
So we have performed final classification with this model

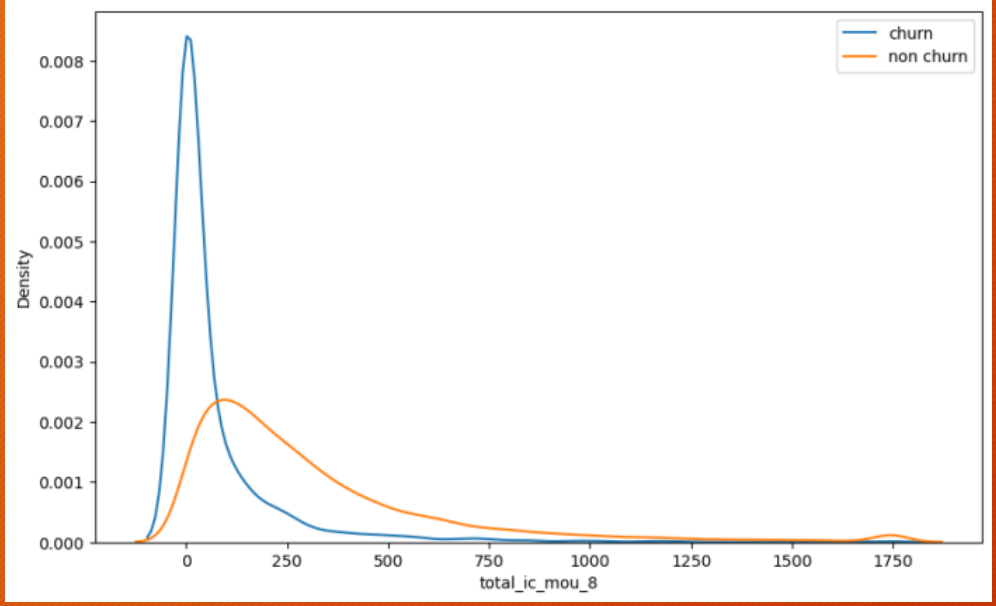
Important factors influencing churn behavior

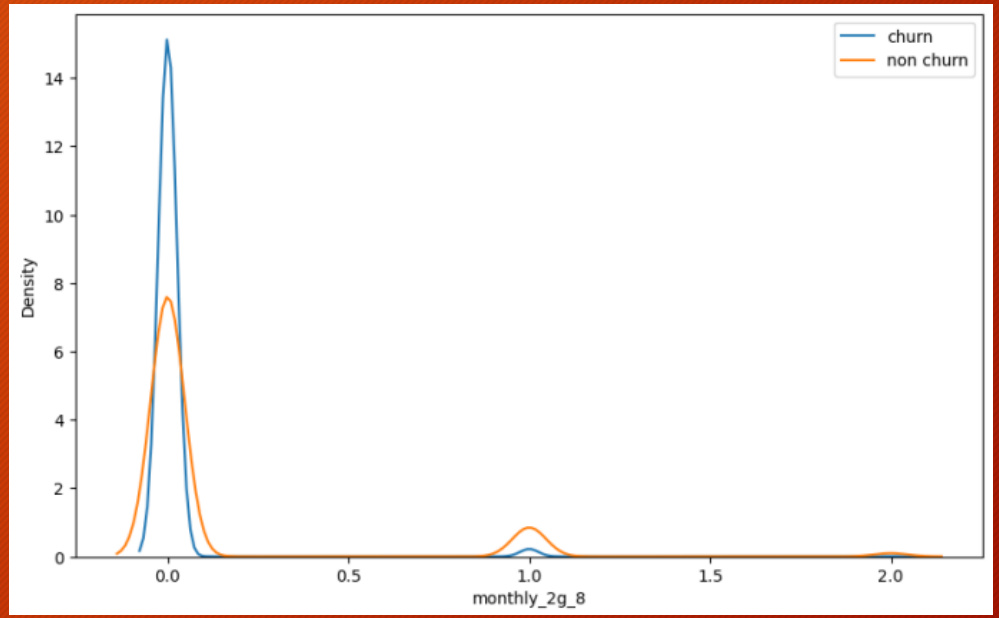
- Most important factors have been listed below. The negative (-ve) sign indicates inverse relation with churn behavior and vice versa.

IMPORATANT FACTORS	COEFFICIENTS FROM INTERPRETABLE MODEL
total_ic_mou_8	-2.76
monthly_2g_8	-1.01
total_ic_mou_7	0.92
monthly_3g_8	-0.62
sachet_2g_8	-0.5
std_og_t2t_mou_8	-0.5
spl_ic_mou_8	-0.47
std_og_t2m_mou_8	-0.42
std_og_t2m_mou_6	0.34
std_og_t2t_mou_7	0.22
std_og_t2t_mou_6	0.16
offnet_mou_7	0.07

Plots of important features for churn and non-churn customers

In left plot, we see, for churn customers the minutes of usage of incoming calls for the month of august is mostly populated on the lower side of non churn customers.





Suggestions and Recommendations

As seen earlier, most of the coefficients have are negative indicating inverse relation with churn behavior.

- The customers whose total minutes of usage for incoming calls is less in the month of august (8th month) have high chance of churning.
- Customers who have a diminishing monthly usage of 2g internet for august as compared to June and July, are likely to churn.
- There is significant decrease in monthly 3g internet usage in august for the customers with churn behavior.
- Customers having decreasing outgoing minutes of usage within same and other operator are likely to churn
- Customers having reduced special incoming calls are likely to churn as well.
- The company should also target the customers for which the total outgoing calls has reduced significantly in the month of August as found in exploratory data analysis

Thank You