TELECOM CHURN PREDICTION

BUSINESS PROBLEM OVERVIEW

- 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 even more important than customer acquisition.
- For many incumbent operators, retaining high profitable customers is the number one business goal.
- To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.
- In this project, we 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.

PROJECT OBJECTIVE

- Build models to predict churn. The predictive model that we are going to build will serve two purposes:

1. It will be used to predict whether a high-value customer will churn or not, in near future (i.e. churn phase). By knowing this, the company can take action steps such as providing special plans, discounts on recharge etc.
2. It will be used to identify important variables that are strong predictors of churn. These variables may also indicate why customers choose to switch to other networks.

- Now , after identifying important predictors, display them visually - we can use plots, summary tables etc. - whatever we think best conveys the importance of features.
- Finally, recommend strategies to manage customer churn based on our observations.

DATA SET

- Source dataset is in csv format.
- Dataset contains 99999 rows and 226 columns.
- Churn is the variable which notifies whether a particular customer is churned or not.
- Here , we will be developing our models to predict churn.

STEPS

Steps:-

1. Reading, understanding and visualizing the data.
2. Preparing the data for modelling.
3. Building the model.
4. Evaluate the model.

Reading, understanding and visualizing the data.

- Handling missing values in columns.
- **Filtering high-value customers:-**Creating column avg_rech_amt_6_7 by summing up total recharge amount of month 6 and 7. Then taking the average of the sum.
- Finding the 70th percentile of the avg_rech_amt_6_7.
- Filtering the customers, who have recharged more than or equal to X.
- Here , **after filtering** the high-value customers , we have around **30K rows.**
- **Tag churners**:-Now tag the churned customers (churn=1, else 0) based on the fourth month as follows:

Those who have not made any calls (either incoming or outgoing) AND have not used mobile internet even once in the churn phase.

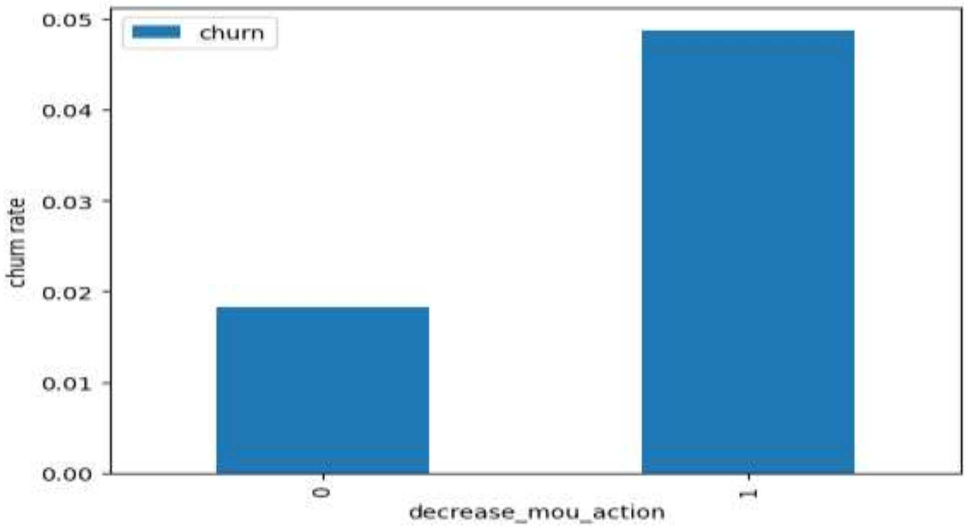
- Outliers treatment
- **Derive new features:-**

- **Deriving new column decrease_mou_action-** indicates whether the minutes of usage of the customer has decreased in the action phase than the good phase.
- **Deriving new column decrease_rech_num_action-** indicates whether the number of recharge of the customer has decreased in the action phase than the good phase.
- **Deriving new column decrease_rech_amt_action-** indicates whether the amount of recharge of the customer has decreased in the action phase than the good phase.
- **Deriving new column decrease_arpu_action-** indicates whether the average revenue per customer has decreased in the action phase than the good phase.
- **Deriving new column decrease_vbc_action-** indicates whether the volume based cost of the customer has decreased in the action phase than the good phase.

❑EDA

▪Univariate analysis

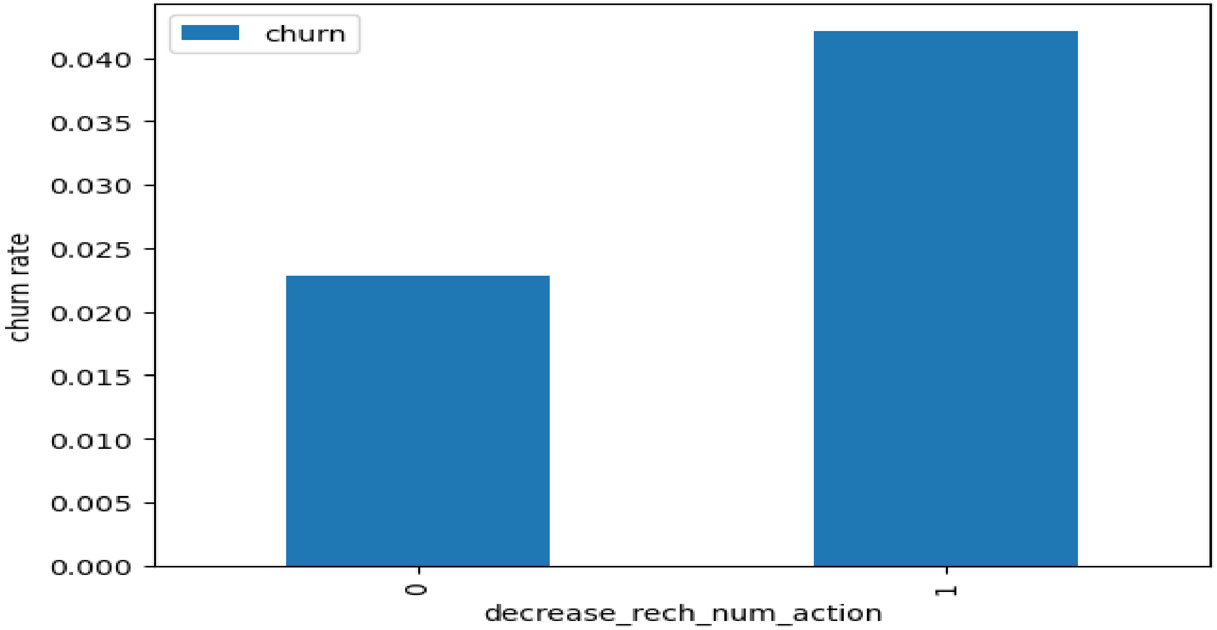
Churn rate on the basis whether the customer decreased her/his MOU in action month



Analysis:-

We can see that the churn rate is more for the customers, whose minutes of usage(mou) decreased in the action phase than the good phase.

Churn rate on the basis whether the customer decreased her/his number of recharge in action month



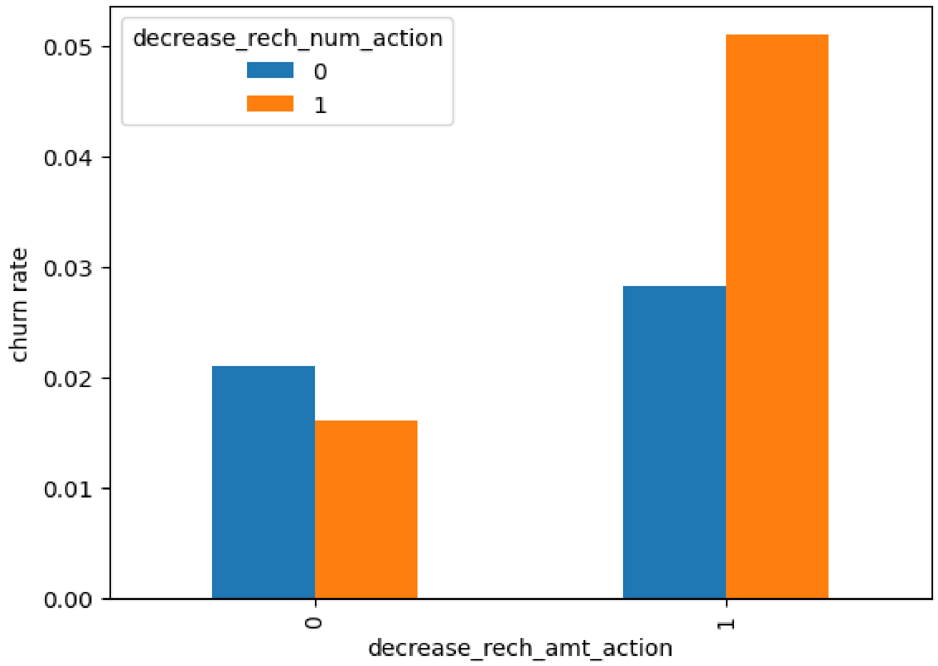
Analysis

As expected, the churn rate is more for the customers, whose number of recharge in the action phase is lesser than the number in good phase.

Note please: The remaining analyses is given in detail in the main Python file.

Bivariate analysis

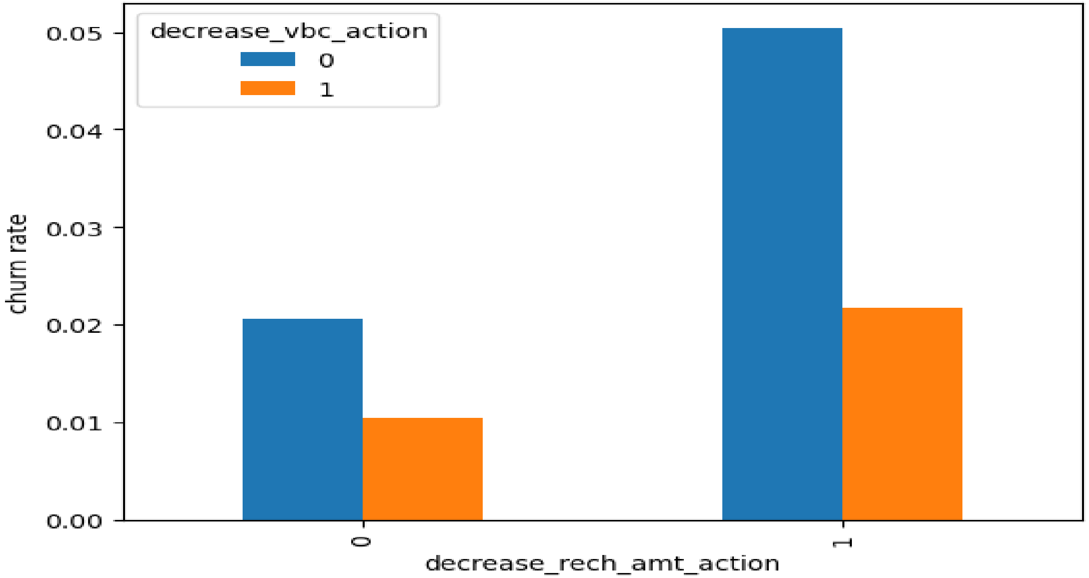
Analysis of churn rate by the decreasing recharge amount and number of recharge in the action phase



Analysis

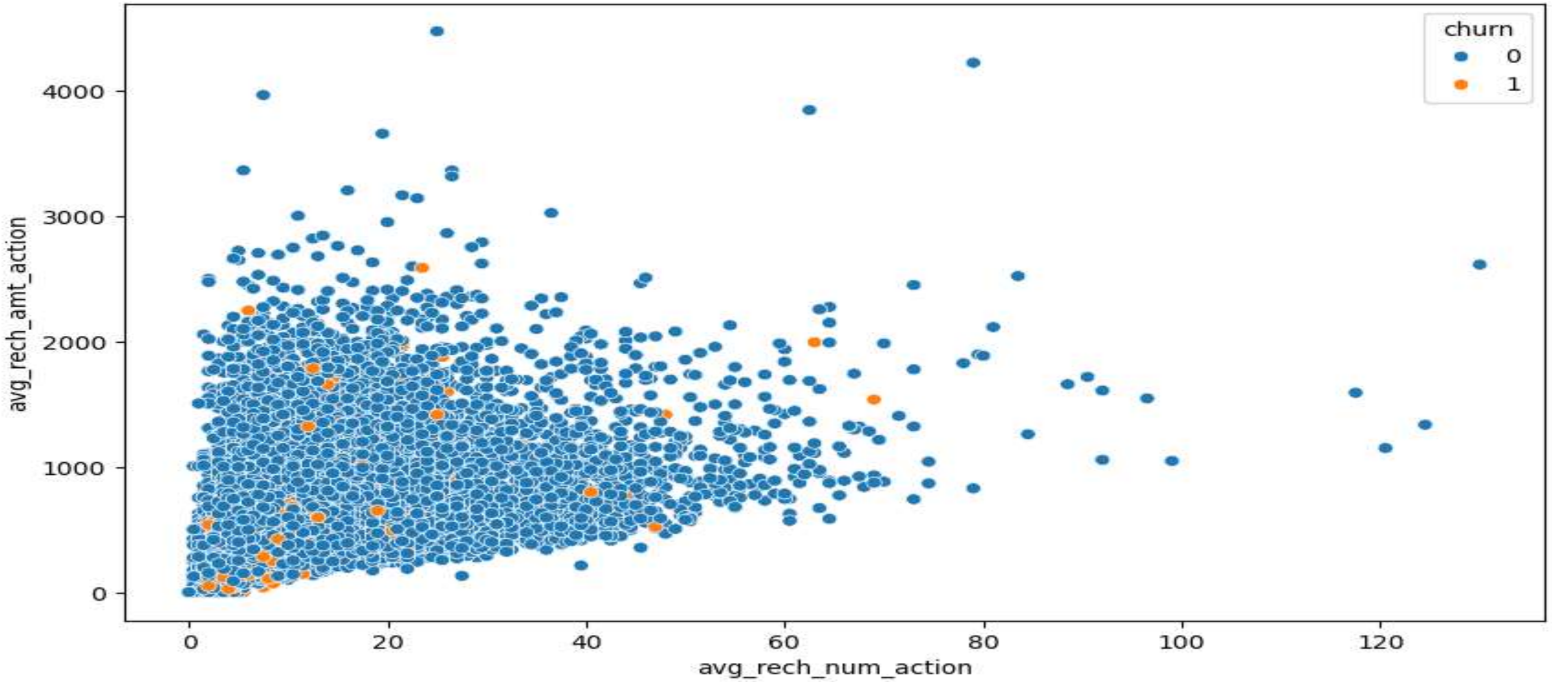
We can see from the above plot, that the churn rate is more for the customers, whose recharge amount as well as number of recharge have decreased in the action phase than the good phase.

Analysis of churn rate by the decreasing recharge amount and volume based cost in the action phase



Analysis
Here, also we can see that the churn rate is more for the customers, whose recharge amount is decreased along with the volume based cost is increased in the action month.

Analysis of recharge amount and number of recharge in action month



Analysis

We can see from the above pattern that the recharge number and the recharge amount are mostly proportional. More the number of recharge, more the amount of the recharge.

- Doing Train-Test split
- **Dealing with data imbalance:-**

➢ We are creating synthetic samples by doing upsampling using SMOTE(Synthetic Minority Oversampling Technique).

- Feature Scaling

MODELS

Final conclusion for Models with PCA

- ➢After trying several models we can see that for acheiving the best sensitivity, which was our ultimate goal, the classic Logistic regression or the SVM models preforms well.
- ➢ For both the models the sensitivity was approx 81%. Also we have good accuracy of apporx 85%.
- ❑ ***Model-3 log_no_pca_3 will be the final model***.(explained in detail in the python file)

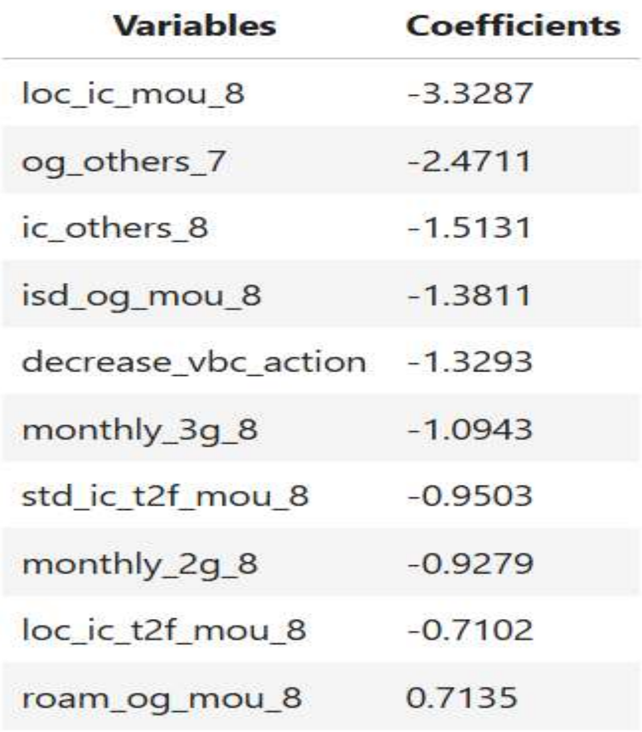
Final conclusion for Models with no PCA

- ➢We can see that the logistic model with no PCA has good sensitivity and accuracy, which are comparable to the models with PCA.
- ➢ So, we can go for the more simplistic model such as **logistic regression with PCA** as it explains the important predictor variables as well as the significance of each variable.
- ➢The model also helps us to identify the variables which should be acted upon for making the decision upon the 'to be churned customers'.
- ➢ Hence, the model is more relevant in terms of explaining to the business.

BUSINESS RECOMMENDATIONS

Top predictors

Below are few top variables selected in the logistic regression model.



Variables	Coefficients
loc_ic_mou_8	-3.3287
og_others_7	-2.4711
ic_others_8	-1.5131
isd_og_mou_8	-1.3811
decrease_vbc_action	-1.3293
monthly_3g_8	-1.0943
std_ic_t2f_mou_8	-0.9503
monthly_2g_8	-0.9279
loc_ic_t2f_mou_8	-0.7102
roam_og_mou_8	0.7135

- We can see most of the top variables have negative coefficients. That means, the variables are inversely correlated with the churn probability.

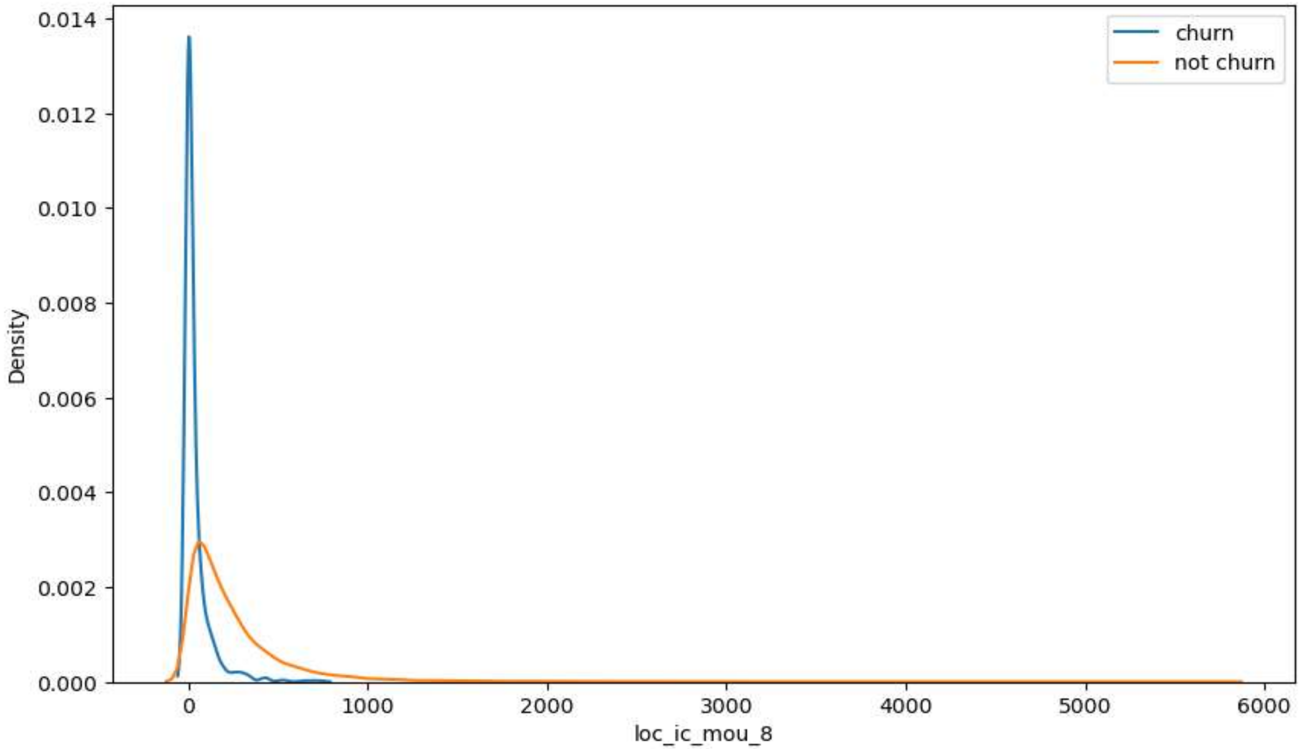
- E.g.:-If the local incoming minutes of usage (loc_ic_mou_8) is lesser in the month of August than any other month, then there is a higher chance that the customer is likely to churn.

Business Recommendations

1. Target the customers, whose minutes of usage of the incoming local calls and outgoing ISD calls are less in the action phase (mostly in the month of August).
2. Target the customers, whose outgoing others charge in July and incoming others on August are less.
3. Also, the customers having value based cost in the action phase increased are more likely to churn than the other customers. Hence, these customers may be a good target to provide offer.
4. Customers, whose monthly 3G recharge in August is more, are likely to be churned.
5. Customers having decreasing STD incoming minutes of usage for operators T to fixed lines of T for the month of August are more likely to churn.
6. Customers decreasing monthly 2g usage for August are most probable to churn.
7. Customers having decreasing incoming minutes of usage for operators T to fixed lines of T for August are more likely to churn.
8. roam_og_mou_8 variables have positive coefficients (0.7135). That means for the customers, whose roaming outgoing minutes of usage is increasing are more likely to churn.

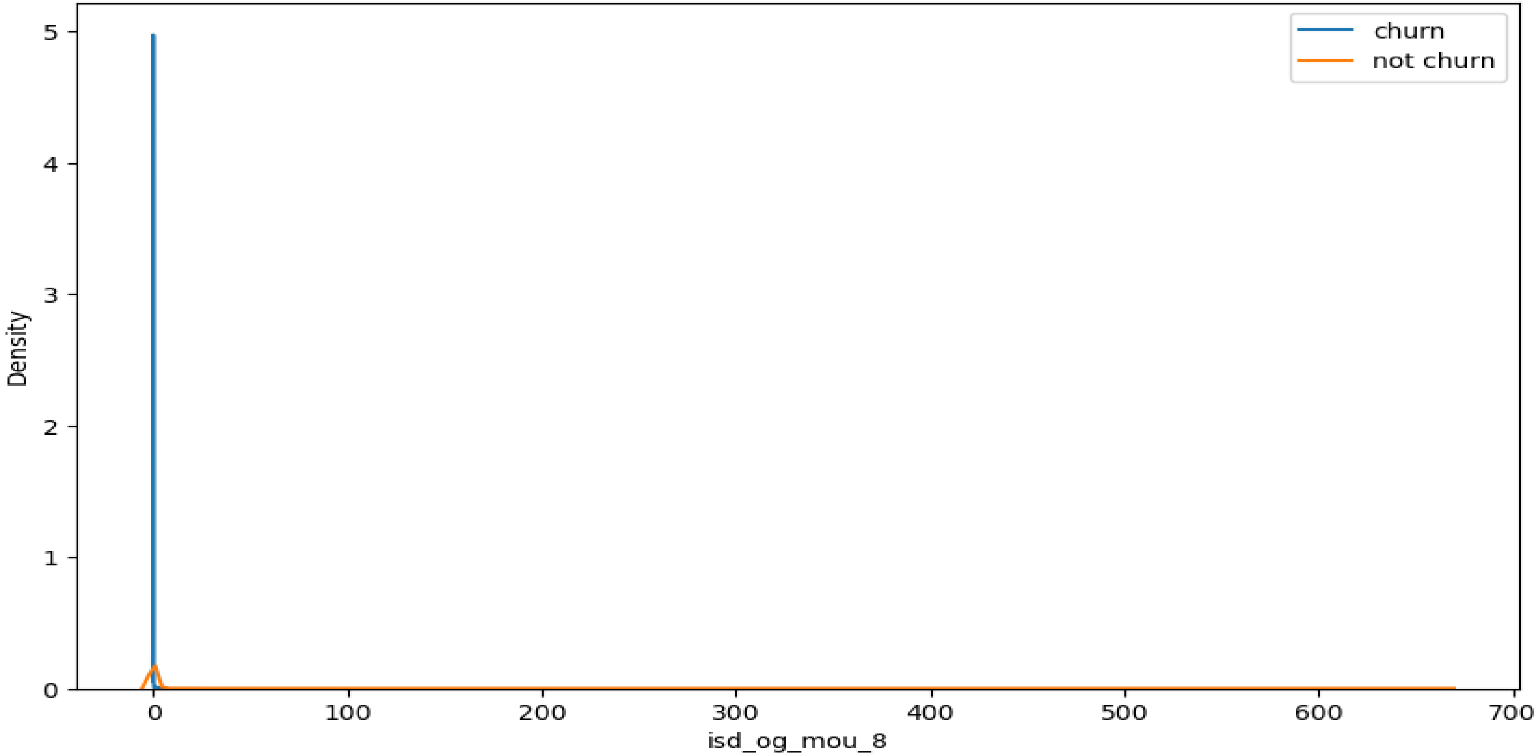
Plots of important predictors for churn and non churn customers

- Plotting loc_ic_mou_8 predictor for churn and not churn customers



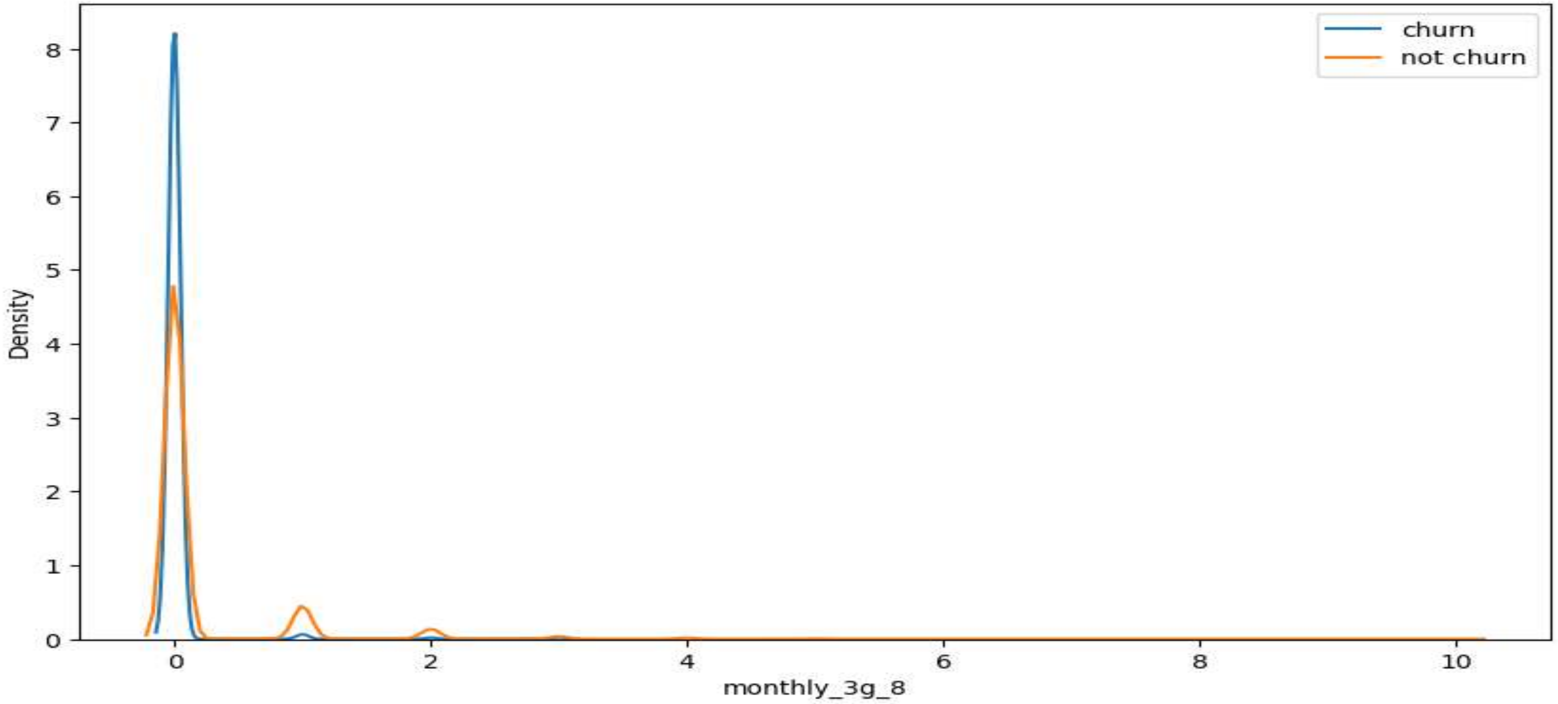
➢ We can see that for the churn customers the minutes of usage for the month of August is mostly populated on the lower side than the non churn customers.

Plotting isd_og_mou_8 predictor for churn and not churn customers



- We can see that the ISD outgoing minutes of usage for the month of August for churn customers is dense at ‘approximately to zero’.
- On the other hand, for the non churn customers, it is little more than the churn customers.

Plotting monthly_3g_8 predictor for churn and not churn customers



- The number of monthly 3g data for August for the churn customers are very much populated around 1, whereas for non churn customers it is spread across various numbers.
- Similarly ,we can plot each variables, which have higher coefficients, churn distribution.

THANK YOU