Telecom Churn Case Study

-PRATEEK BAJORIA, RICHA JHUNJHUNWALA & RAVITEJA THRIPURARI

Introduction

As is widely known, in the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another.

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.

Retaining highly profitable customers is the main business goal here.

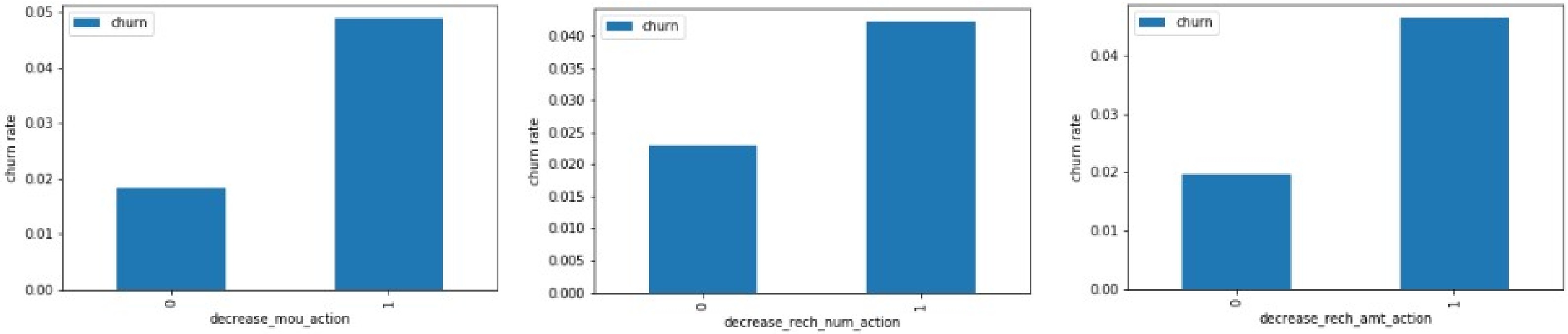
Data Understanding & Preparation

Understanding: The dataset contains customer-level information for a span of four consecutive months - June, July, August and September. The months are encoded as 6, 7,(good phase) 8(action phase) and 9(churn phase), respectively. The **business objective** is to predict the churn in the last (i.e. the ninth) month using the data (features) from the first three months.

Preparation: We filtered high value customers that is those having recharge amount >=70th percentile of the average recharge amount in the first two months (the good phase) and handled missing values. Next, we tagged the churned customers (churn=1, else 0) based on the fourth month and removed attributes of the churn phase. We performed outliers treatment and also derived various new variables for the purposes of our analysis. Please consider the Jupyter file attached for further details.

EDA-Univariate Analysis

Churn rate on the basis of minutes of usage, no. of recharge and amount of recharge in the action month respectively:



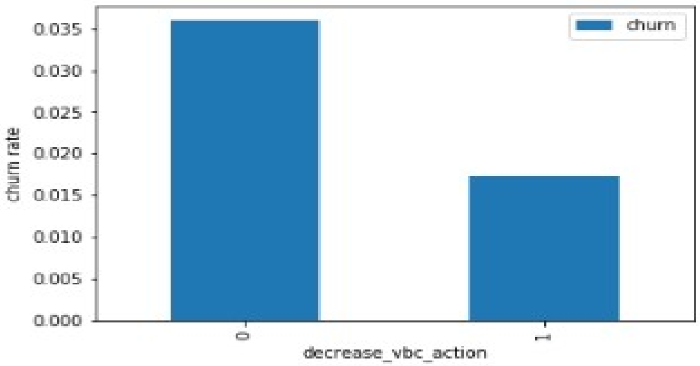
We can see that the churn rate is more for the customers, whose minutes of usage(MOU) decreased in the action phase than in the good phase.

Also, the churn rate is more for the customers, whose number of recharge in the action phase is lesser compared to the number in good phase.

Further**,** the churn rate is more for the customers, whose amount of recharge in the action phase is lesser than the amount in good phase.

Univariate Analysis Continued

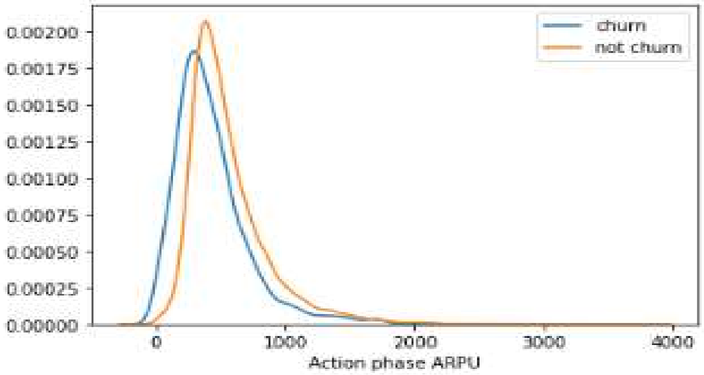
Churn rate on the basis whether the customer decreased her/his volume based cost in action month-

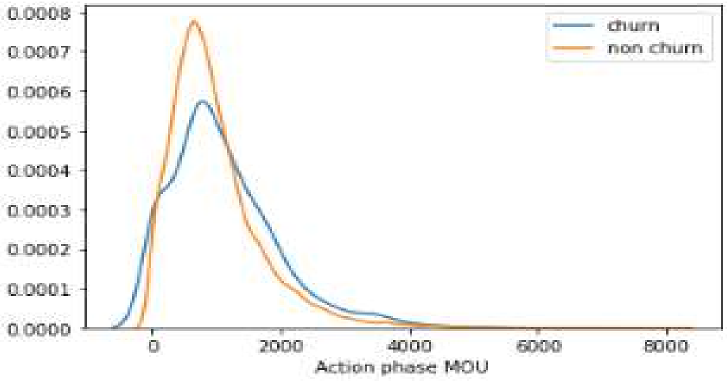


As expected, the churn rate is more for the customers, whose volume based cost in the action month is increased. This means that the customers do not do more monthly recharge when they are in the action phase.

Univariate Analysis Continued

Analysis of the average revenue per customer (churn and not churn) and minutes of usage in the action phase-





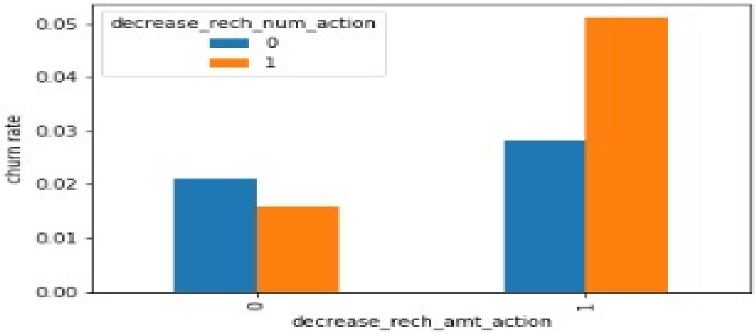
Average revenue per user (ARPU) for the churned customers is mostly dense between 0 to 900. The higher ARPU customers are less likely to be churned. ARPU for the not churned customers is mostly dense between 0 to 1000.

Minutes of usage(MOU) of the churn customers is mostly populated on the 0 to 2500 range. Higher the MOU, lesser the churn probability.

Bivariate Analysis

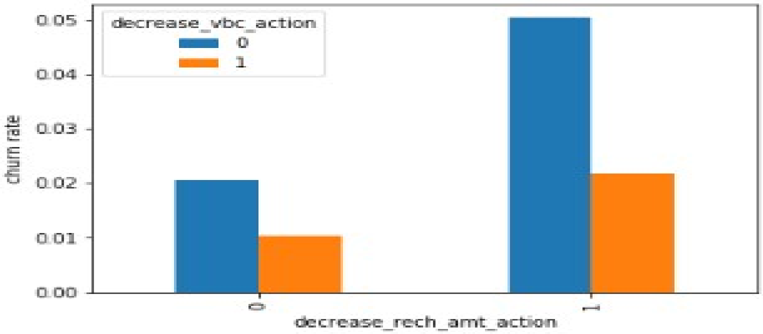
Analysis of churn rate by the decreasing recharge amount and number of recharge in the action phase-

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.

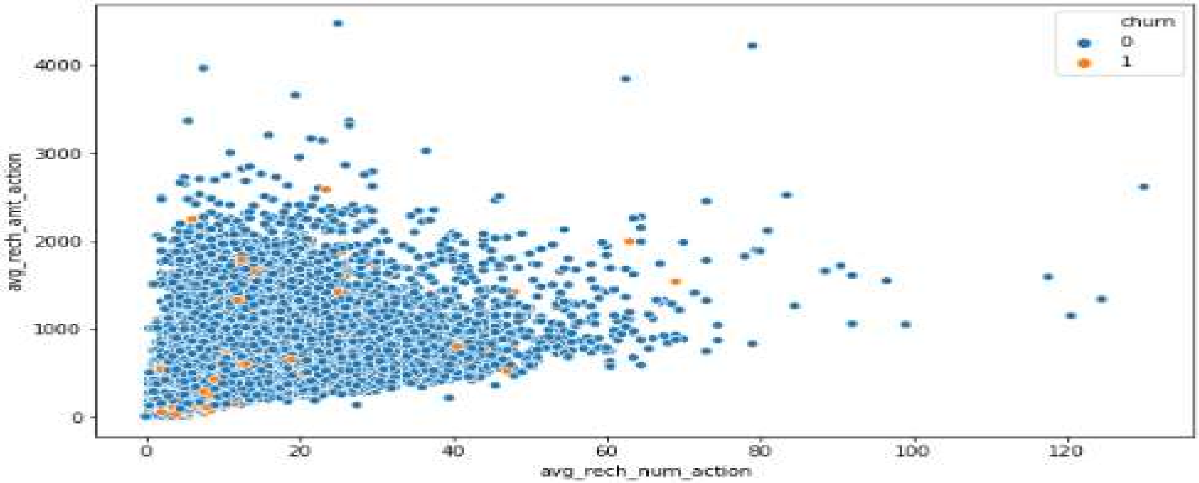


Analysis of churn rate by the decreasing recharge amount and volume based cost in the action phase-

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



Analysis of recharge amount and number of recharge in action month



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

Train-Test Split, Dealing with Data Imbalance & Feature Scaling

Post EDA and dropping certain columns which were no longer required for further analysis, we set out to split the data into train and test data.

Also, dealt with data imbalance using Synthetic Minority Oversampling Technique as we noticed there was a very little percentage of churn rate on data inspection.

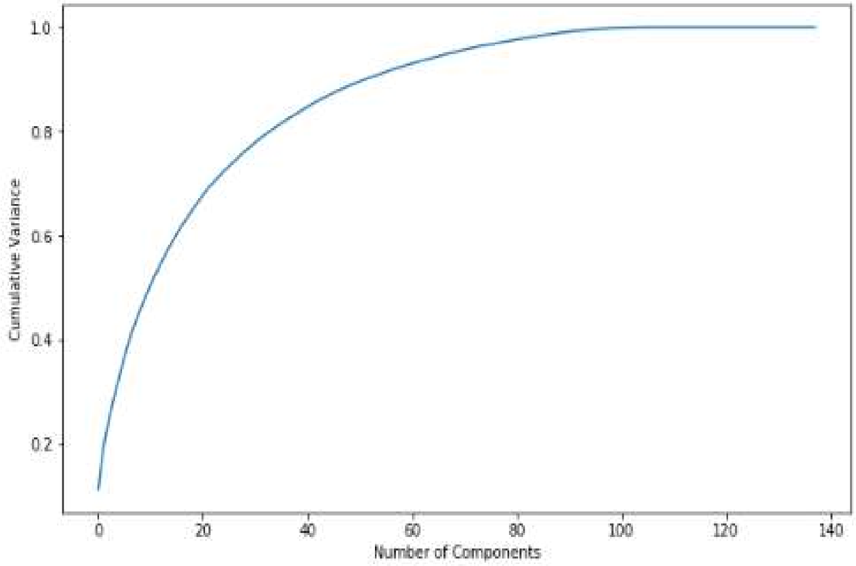
We had fit the train data into scaler and transformed it. As for the test data, we don't fit scaler on the test set. We only transform the test set. Thus, we progressed accordingly.

Model with PCA

We can see that 60 components explain 90% of Data variance.

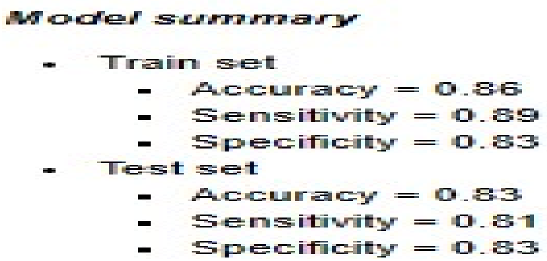
So, we performed PCA with 60 components.

We were more focused on higher Sensitivity/Recall score than the accuracy as we need to care more about churn cases than the non churn cases. The main goal is to retain the customers, who have the possibility to churn. There should not be a problem, if we consider few non churn customers as churn customers and provide them some incentives for retaining them. Hence, the sensitivity score is more important here.



➢The next step was logistic regression with PCA i.e. tuning hyper parameter C which is the inverse of regularization strength in Logistic Regression. Higher values of C correspond to less regularization.

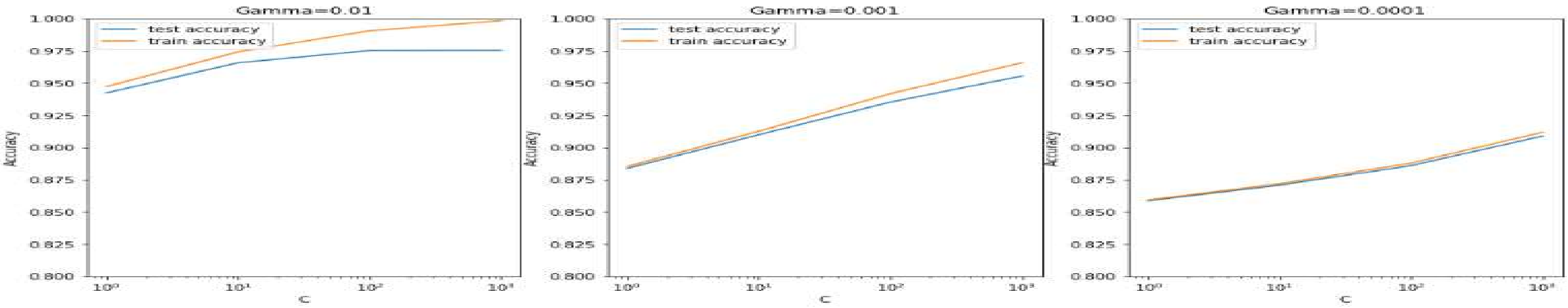
➢Finally, we found that the model was performing well in the test set, based on what it had learnt from the train set.



Model summary

- Train set
  - Accuracy = 0.86
  - Sensitivity = 0.89
  - Specificity = 0.83
- Test set
  - Accuracy = 0.83
  - Sensitivity = 0.81
  - Specificity = 0.83

Support Vector Machine(SVM) with PCA



From the above plot, we can see that higher value of gamma leads to overfitting the model. With the lowest value of gamma (0.0001) we have train and test accuracy almost same.

Also, at C=100 we have a good accuracy and the train and test scores are comparable.

Though sklearn suggests the optimal scores mentioned above (gamma=0.01, C=1000), one could argue that it is better to choose a simpler, more non-linear model with gamma=0.0001. This is because the optimal values mentioned here are calculated based on the average test accuracy (but not considering subjective parameters such as model complexity).

We can achieve comparable average test accuracy (~90%) with gamma=0.0001 as well, though we'll have to increase the cost C for that. So to achieve high accuracy, there's a tradeoff between:

High gamma (i.e. high non-linearity) and average value of C

Low gamma (i.e. less non-linearity) and high value of C

We argue that the model will be simpler if it has as less non-linearity as possible, so we choose gamma=0.0001 and a high C=100.

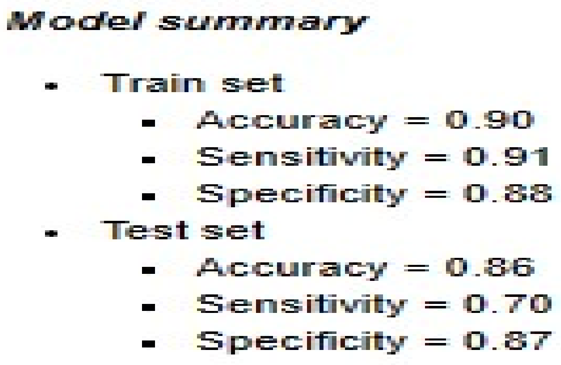
Based on this we build a model on optimal hyper parameters and arrived at the model summary presented to the right.

Model summary

- Train set
  - Accuracy = 0.89
  - Sensitivity = 0.92
  - Specificity = 0.85
- Test set
  - Accuracy = 0.85
  - Sensitivity = 0.81
  - Specificity = 0.85

Decision Tree with PCA

The end result of the captioned approach is presented below. We can see from the model performance that the Sensitivity has decreased while evaluating the model on the test set. However, the accuracy and specificity is quite good in the test set.

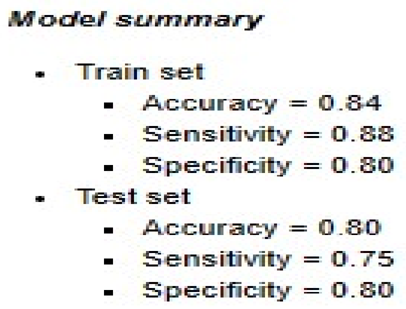


Model summary

- Train set
  - Accuracy = 0.90
  - Sensitivity = 0.91
  - Specificity = 0.88
- Test set
  - Accuracy = 0.86
  - Sensitivity = 0.70
  - Specificity = 0.87

Random forest & Final conclusion with PCA

We can see from the model performance that the Sensitivity has decreased while evaluating the model on the test set. However, the accuracy and specificity is quite good in the test set.



Model summary

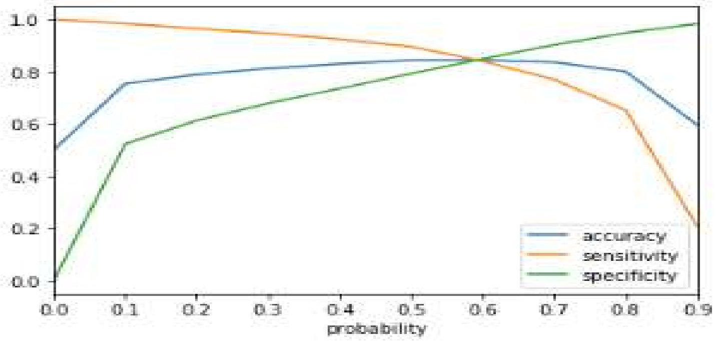
- Train set
  - Accuracy = 0.84
  - Sensitivity = 0.88
  - Specificity = 0.80
- Test set
  - Accuracy = 0.80
  - Sensitivity = 0.75
  - Specificity = 0.80

Final conclusion with PCA

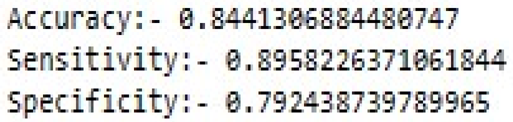
After trying several models we can see that for acheiving the best sensitivity, which was our ultimate goal, the classic Logistic regression or the SVM models preforms well. For both the models the sensitivity was approx 81%. Also we have good accuracy of apporx 85%.

Without PCA

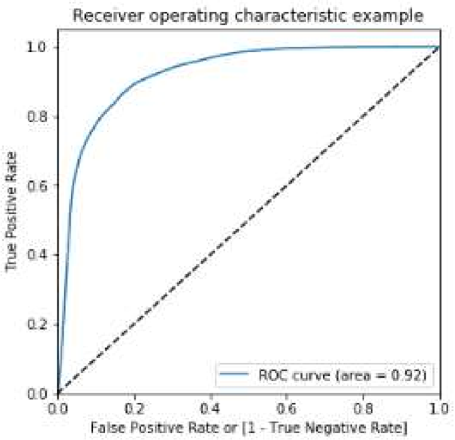
We also explored the model building without PCA and noted the following-



Accuracy becomes stable around 0.6. With increased probability sensitivity decreases while specificity increases. At 0.6, where all parameters intersect, we find the required balance among the 3 with good accuracy. But our main objective is better sensitivity. So, we took an optimal probability cut off point at 0.5. Results for train set are presented below along with ROC curve. We can see area of the ROC curve is closer to 1 which is the Gini of the model.

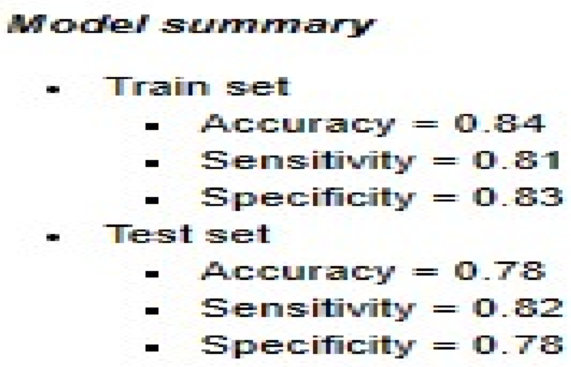


```
Accuracy:- 0.8441306884480747
Sensitivity:- 0.8958226371061844
Specificity:- 0.792438739789965
```



Conclusion 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 of the to be churned customers. Hence, the model is more relevant in terms of explaining to the business owners.

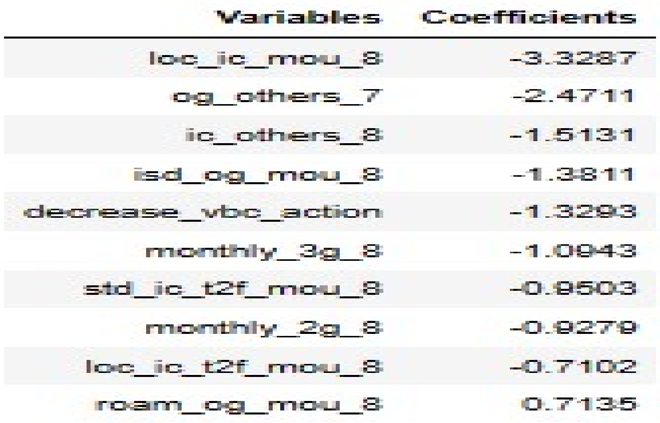


Model summary

- Train set
  - Accuracy = 0.84
  - Sensitivity = 0.81
  - Specificity = 0.83
- Test set
  - Accuracy = 0.78
  - Sensitivity = 0.82
  - Specificity = 0.78

Business recommendations-Top predictors

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

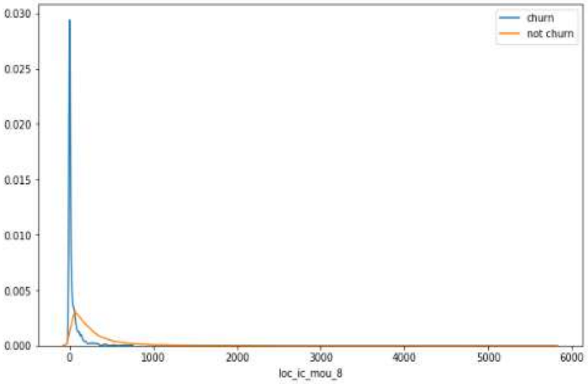


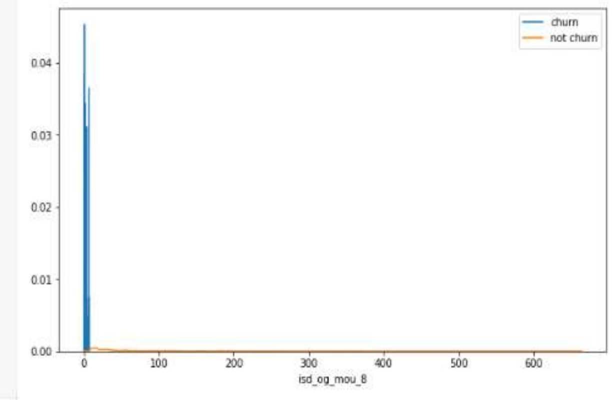
Variables	Coefficients
loc_ic_mou_8	-3.3287
og_others_7	-2.4711
ic_others_8	-1.5131
isd_og_mou_8	-1.3811
decrease_vbc_action	-1.3293
monthly_3g_8	-1.0943
std_ic_t2f_mou_8	-0.9503
monthly_2g_8	-0.9279
loc_ic_t2f_mou_8	-0.7102
roam_og_mou_8	0.7135

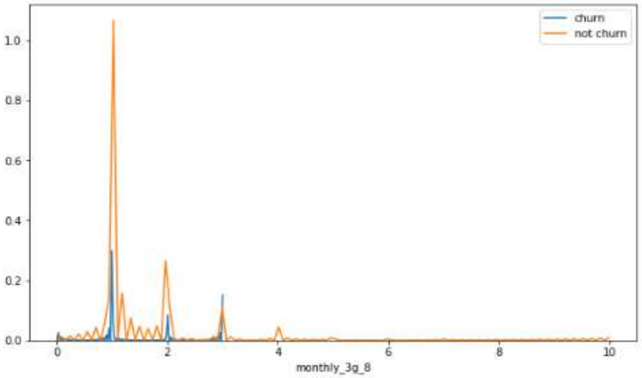
We can see most of the top variables have negative coefficients. That means, the variables are inversely correlated with the churn probability.

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.

Plots of important predictors for churn and non 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.

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

The number of monthly 3g data for August for the churn customers are very much populated around 1, whereas of non churn customers it spread across various numbers.

Similarly we can plot each variables, which have higher coefficients, churn distribution.

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.

THANK YOU