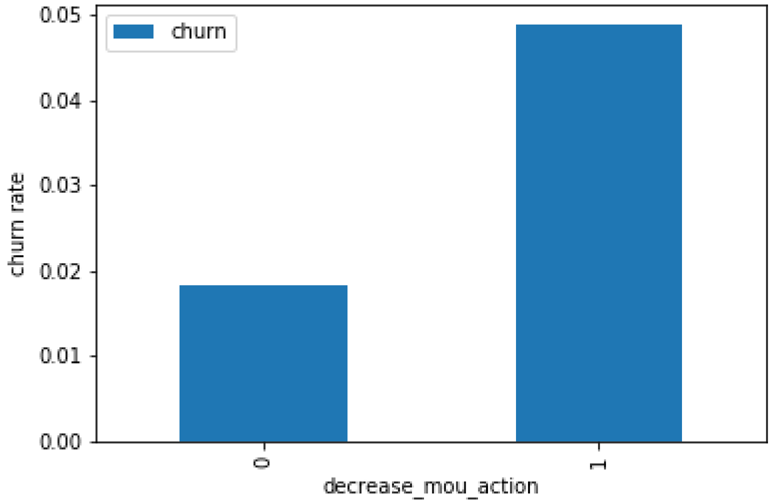
Telecom churn case study

Univariate analysis

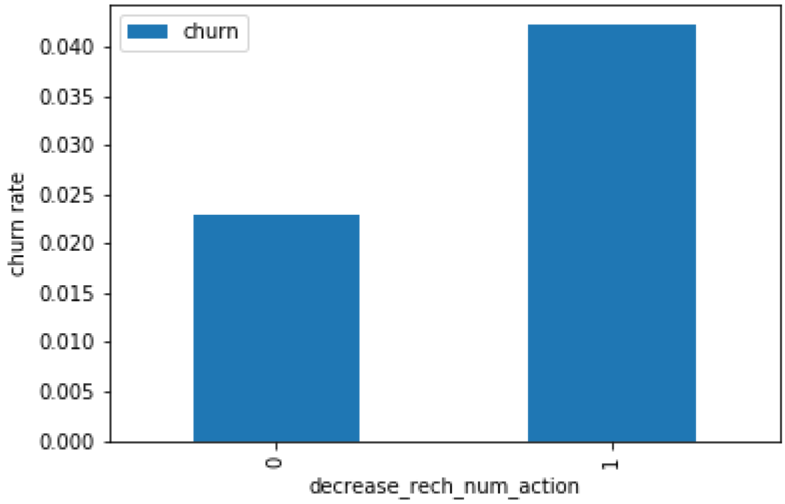
Churn rate on the basis whether the customer decreased her/his MOU in action month



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.

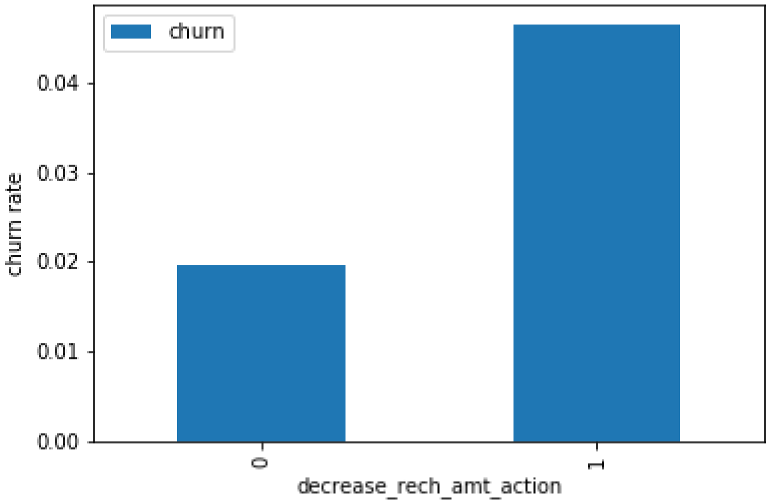
Churn rate on the basis whether the customer decreased her/his number of recharge in action month



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.

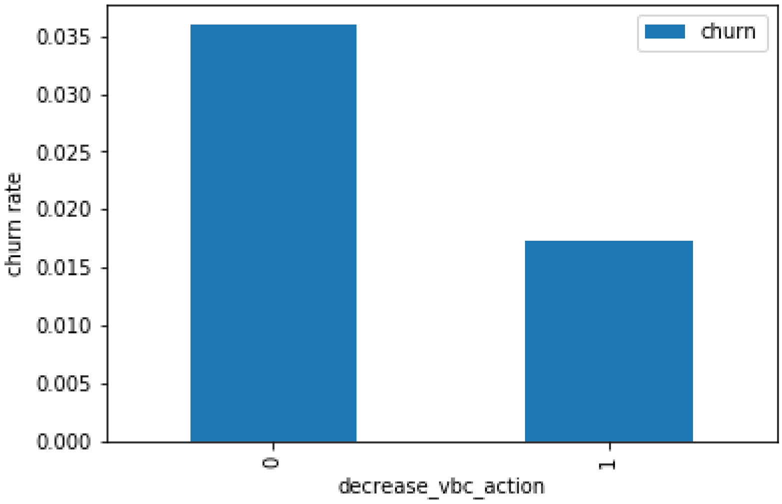
Churn rate on the basis whether the customer decreased her/his amount of recharge in action month



Analysis

Here also we see the same behaviour. The churn rate is more for the customers, whose amount of recharge in the action phase is lesser than the amount in good phase.

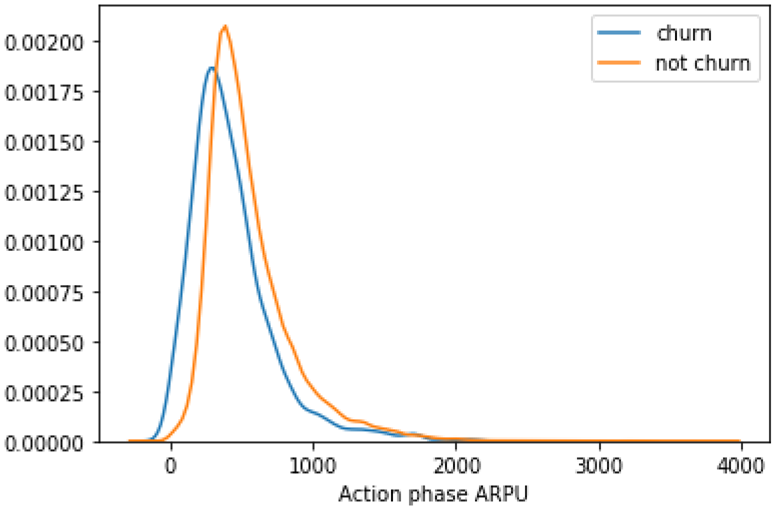
Churn rate on the basis whether the customer decreased her/his volume based cost in action month



Analysis

Here we see the expected result. The churn rate is more for the customers, whose volume based cost in action month is increased. That means the customers do not do the monthly recharge more when they are in the action phase.

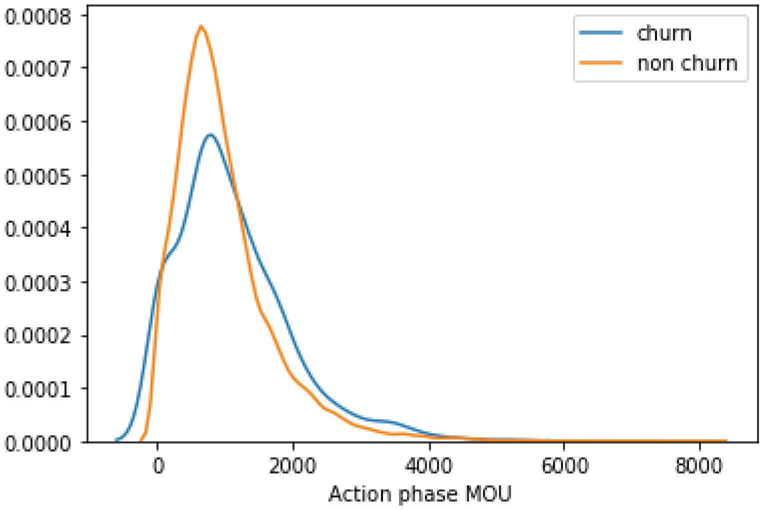
Analysis of the average revenue per customer (churn and not churn) in the action phase



Average revenue per user (ARPU) for the churned customers is mostly densed on the 0 to 900. The higher ARPU customers are less likely to be churned.

ARPU for the not churned customers is mostly densed on the 0 to 1000.

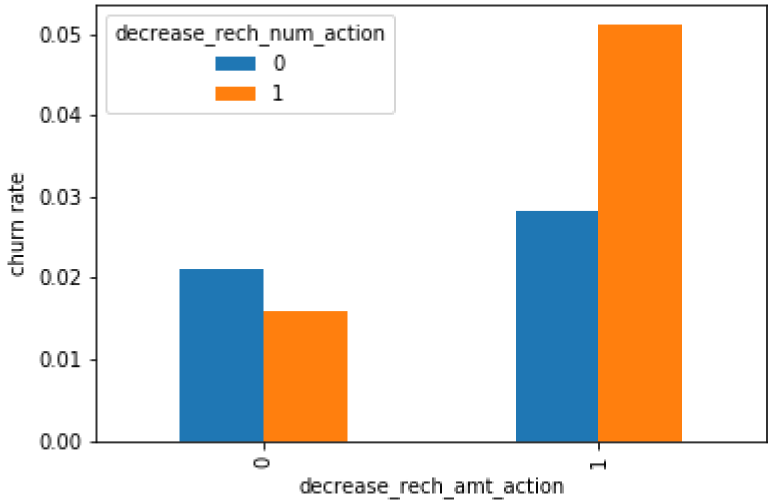
Analysis of the minutes of usage MOU (churn and not churn) in the action phase



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

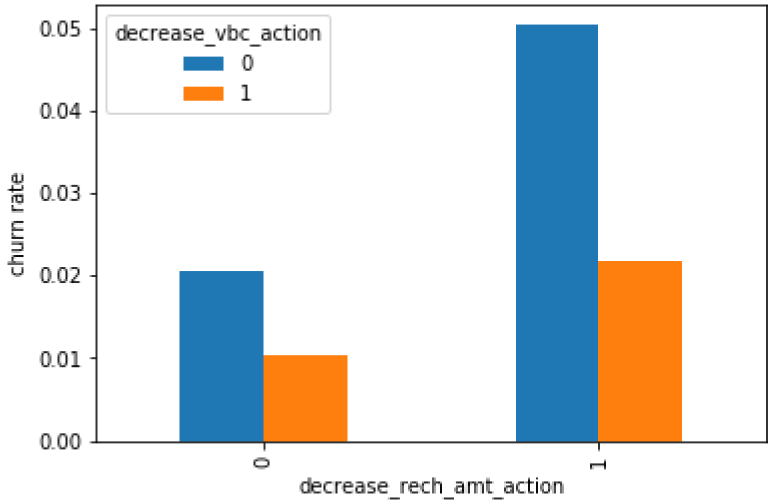
Analysis of churn rate by the decreasing recharge amount and number of recharge in the action phase



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.

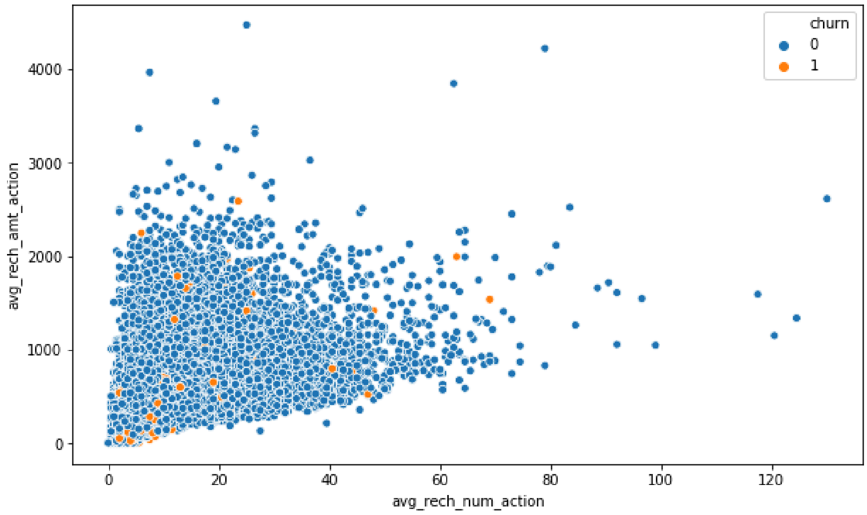
Analysis of churn rate by the decreasing recharge amount and volume based cost in the action phase



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.

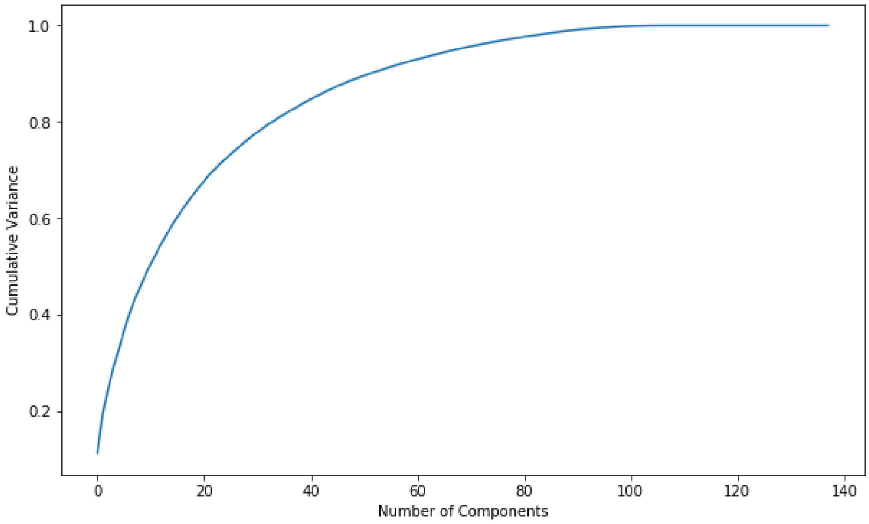
Analysis of recharge amount and number of recharge in action month



Analysis

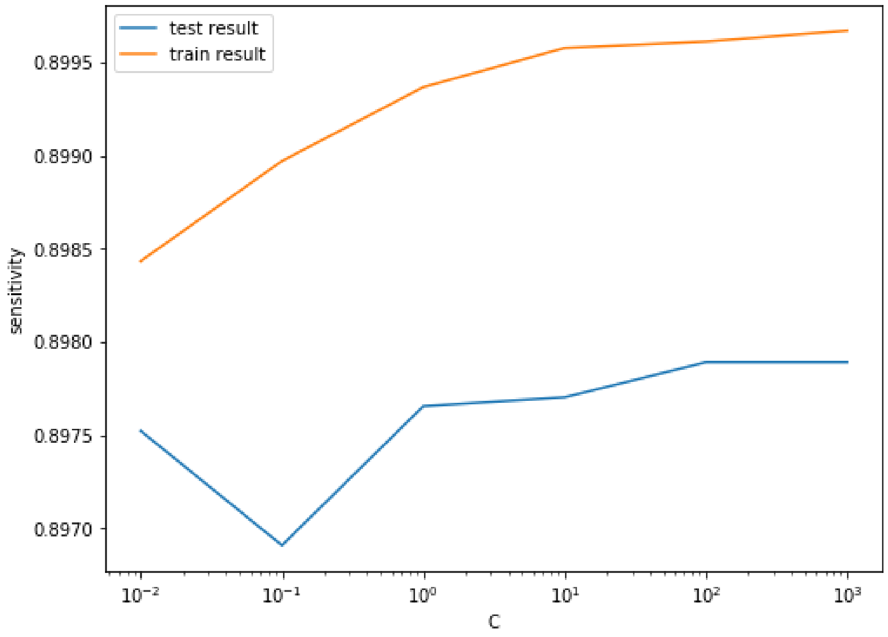
We can see from the above pattern that the recharge number and the recharge amount are mostly proportional. More the number of recharge, more the amount of the recharge.

Model with PCA

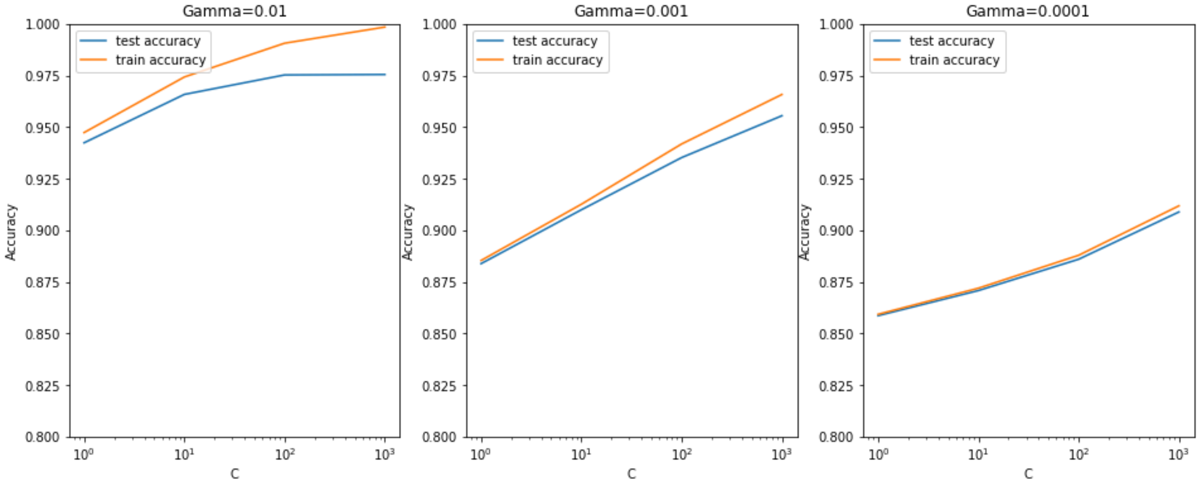


We can see that `60 components` explain almost more than 90% variance of the data. So, we will perform PCA with 60 components.

Logistic regression with PCA



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

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

Logistic regression with No PCA

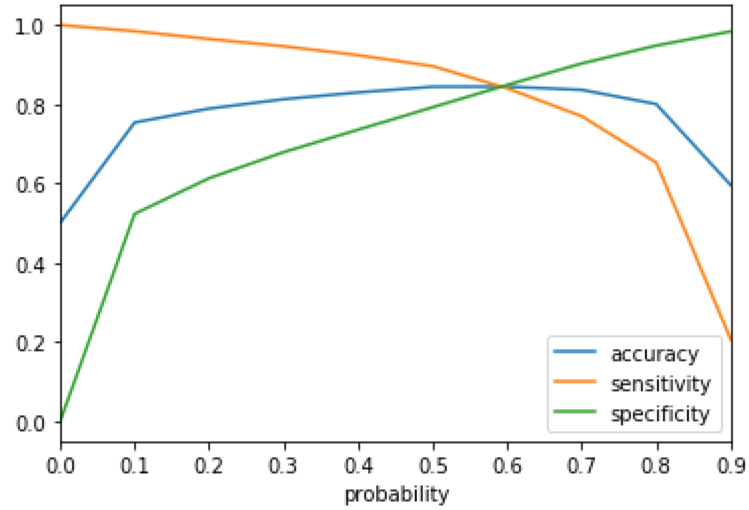
Model analysis

1. We can see that there are few features have positive coefficients and few have negative.
2. Many features have higher p-values and hence became insignificant in the model.

Coarse tuning (Auto+Manual)

We'll first eliminate a few features using Recursive Feature Elimination (RFE), and once we have reached a small set of variables to work with, we can then use manual feature elimination (i.e. manually eliminating features based on observing the p-values and VIFs).

calculate the accuracy sensitivity and specificity for various probability cutoffs.



Analysis of the above curve

Accuracy - Becomes stable around 0.6

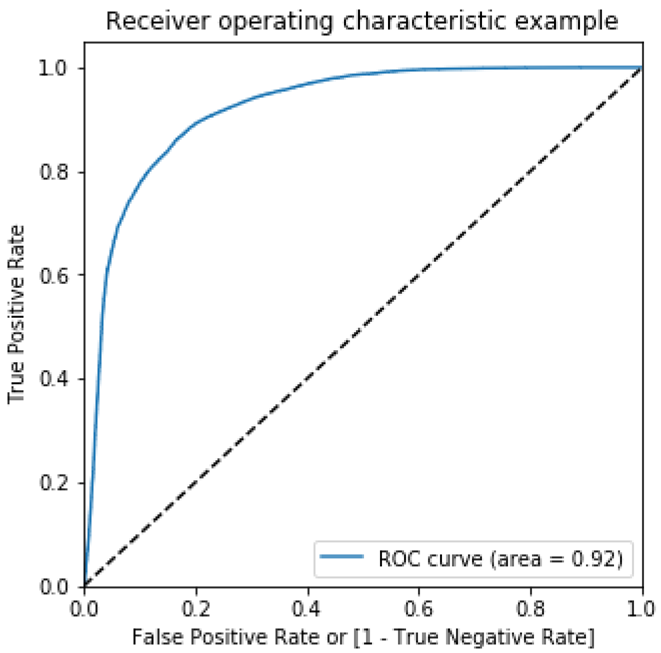
Sensitivity - Decreases with the increased probablity.

Specificity - Increases with the increasing probablity.

`At point 0.6` where the three parameters cut each other, we can see that there is a balance bethween sensitivity and specificity with a good accuracy.

Here we are intended to acheive better sensitivity than accuracy and specificity. Though as per the above curve, we should take 0.6 as the optimum probability cutoff, we are taking ***0.5*** for acheiving higher sensitivity, which is our main goal.

Plotting the ROC Curve (Trade off between sensitivity & specificity)



We can see the area of the ROC curve is closer to 1, whic is the Gini of the model.

Testing the model on the test set

Plots of important predictors for churn and non churn customers

Top predictors

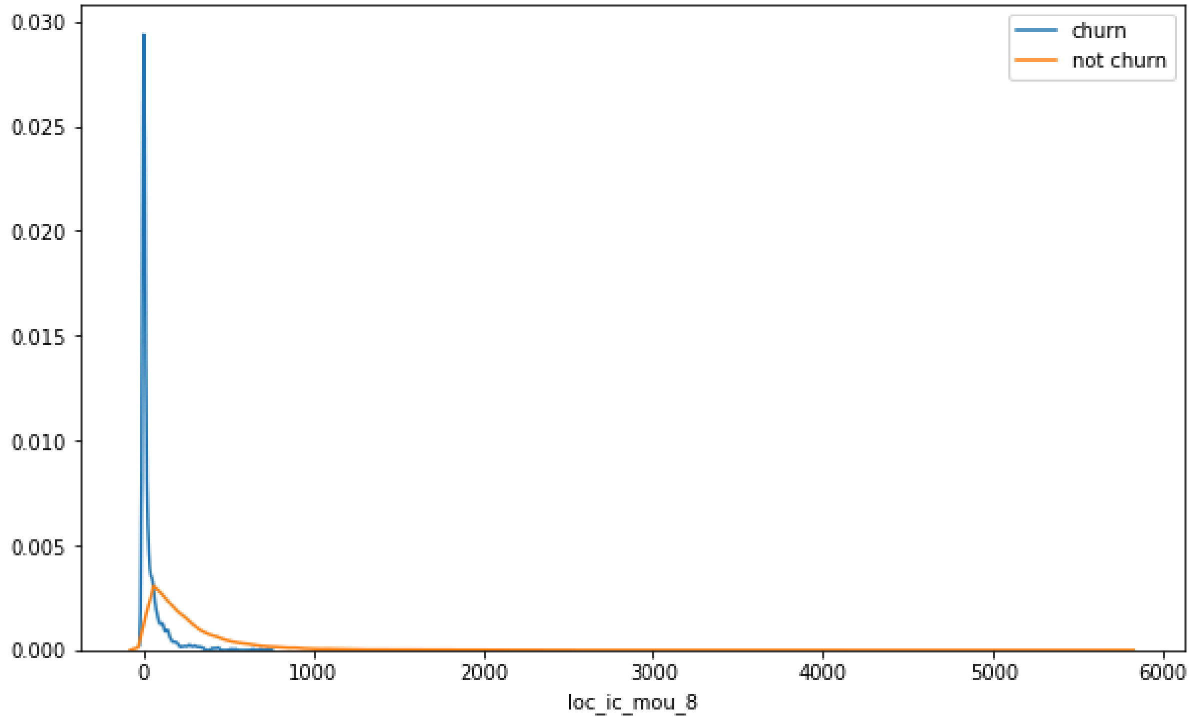
Below are few top variables selected in the logistic regression model.

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

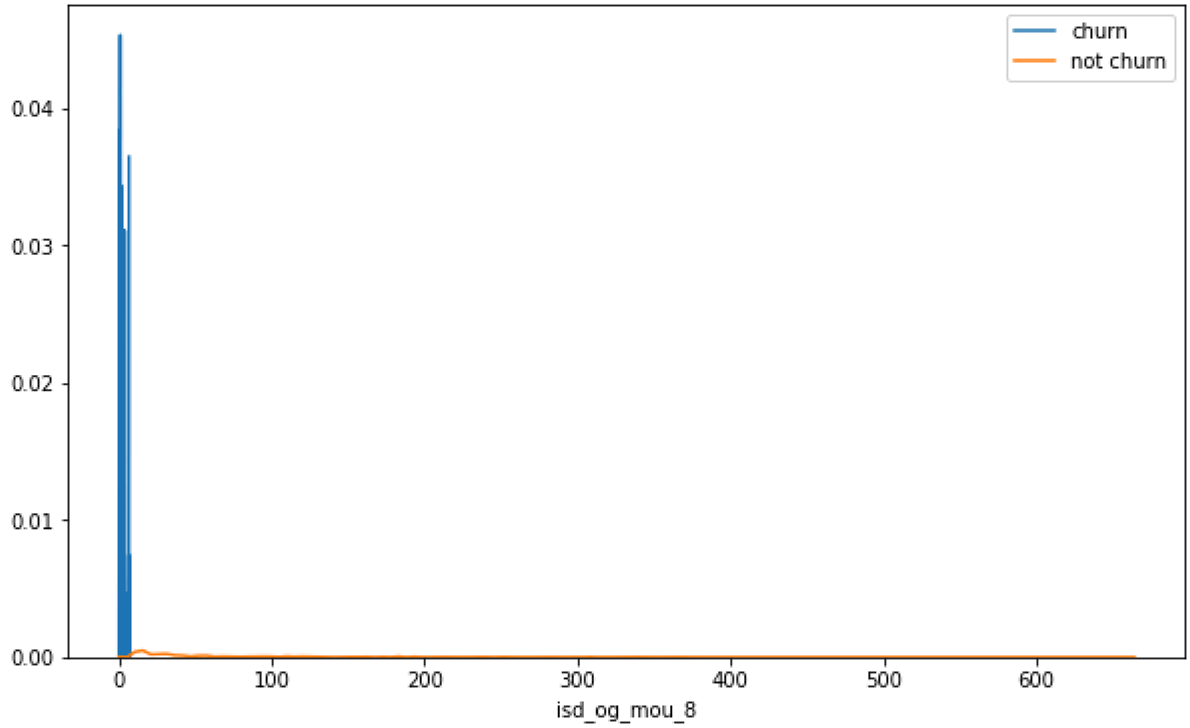
monthly_2g_8	-0.9279
loc_ic_t2f_mou_8	-0.7102
roam_og_mou_8	0.7135

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

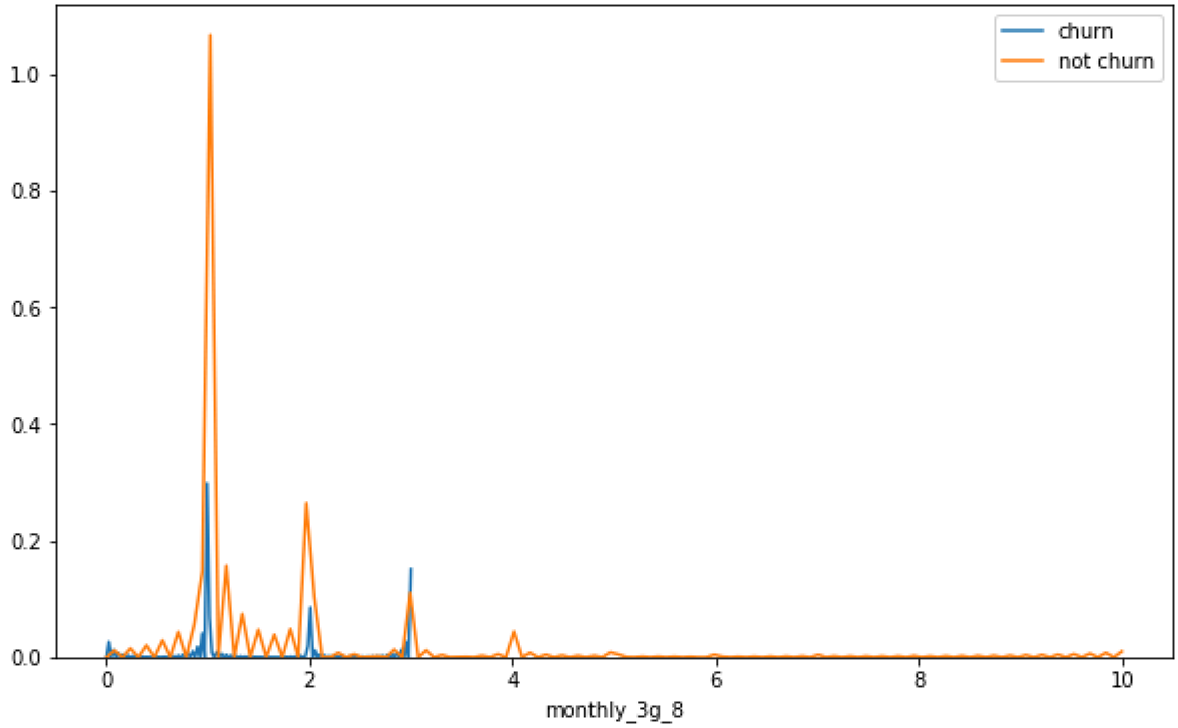
Plots of important predictors for churn and non churn customers



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 densed approximately to zero. On the onther hand for the non churn customers it is little more than the churn customers.



The number of mothly 3g data for August for the churn customers are very much populated aroud 1, whereas of non churn customers it spreaded accross various numbers.

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

Overview from a Business Perspective

Univariate Analysis Insights:

- MOU (Minutes of Usage) Decrease: Customers who experience a decrease in MOU during the action phase are more likely to churn. This suggests dissatisfaction or reduced engagement with services.
- Recharge Amount and Number Decrease: Similarly, a decrease in recharge amount and number during the action phase correlates with higher churn rates, indicating a potential loss of interest or satisfaction.
- Volume-Based Cost Increase: Customers whose volume-based costs increase in the action phase are more prone to churn. This could signal dissatisfaction with pricing or perceived value.
- ARPU (Average Revenue per User) Analysis: Churned customers tend to have lower ARPU, particularly concentrated below 900, indicating that higher-paying customers are less likely to churn.
- MOU Analysis: Churned customers predominantly exhibit lower MOU, suggesting that higher usage correlates with lower churn probability.

Bivariate Analysis Insights:

- Recharge Amount and Number: Customers with decreased recharge amounts and numbers in the action phase are more likely to churn, indicating a clear pattern of disengagement.
- Recharge Amount and Volume-Based Cost: A decrease in recharge amount alongside an increase in volume-based cost leads to higher churn rates, indicating dissatisfaction with pricing or perceived value.

Modeling Insights:

- Logistic Regression and SVM Models: These models, particularly with PCA, provide good sensitivity and accuracy for predicting churn. Features like incoming/outgoing minutes, ISD usage, and data usage appear to be strong predictors.
- Probability Cutoff Analysis: Choosing a probability cutoff of 0.5 prioritizes sensitivity, aiming to capture churners effectively while maintaining a reasonable level of accuracy.
- Important Predictors: Variables such as incoming/outgoing minutes, roaming usage, and data usage are identified as significant predictors of churn, providing actionable insights for targeted retention strategies.

Business Implications:

- Retention Strategies: The analysis highlights key indicators of churn, allowing telecom companies to develop targeted retention strategies. These may include personalized offers, improved service quality, or proactive customer support.
- Segmentation: Understanding customer segments based on usage patterns and revenue levels can help tailor retention efforts effectively. High-value customers may require different strategies compared to lower-value segments.
- Service Optimization: Insights into usage patterns and satisfaction levels can inform service optimization efforts, ensuring that offerings align with customer needs and preferences.
- Competitive Positioning: Monitoring churn indicators can provide valuable insights into market competitiveness and customer perception. Addressing pain points and differentiating services can help maintain a competitive edge in the market.

Overall, leveraging data-driven insights to understand churn behavior and implement targeted retention strategies is crucial for maximizing customer lifetime value and sustaining business growth in the telecom industry.