Telecom Churn – Case Study

Our Team

Sriranjani S

Manager – Citi

Somake Mehrotra

Manager

Joshly Johnson

Designer

Problem Statement

To reduce customer churn, telecom companies need to predict which customers are at high risk of churn. In this project, we will analyse 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 high profitable customers is the main business goal here.
In the Indian and Southeast Asian markets, approximately 80% of revenue comes from the top 20% of customers (called high-value customers). Thus, if we can reduce the churn of high-value customers, we will be able to reduce significant revenue leakage.

Objective

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.

Retaining high profitable customers is the main business goal here.

Approach

1. Reading and understanding the data
2. Data Preparation for Modeling
3. Building the model
4. Evaluate the model

Step 1. Reading and understanding the data

- The Telecom data dump provided was about a lakh entries. The Data dictionary can be used to understand the column representation

Step 2. Data Preparation for Modeling

Step 2.1. Handling missing values in columns

- We will drop the variables with more than 40% of missing value.

Step 2.2. Filter 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.
- Filter the customers, who have recharged more than or equal to X.

Step 2.3. Handling missing values in rows

- Check the MOU values for all months and eliminate the records. This results in almost 7% loss in data however the remaining number of records are enough to proceed with the analysis.

Step 2.2. 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.

Step 2.3. Deleting all the attributes corresponding to the churn phase

Step 2.4. Checking churn percentage

Step 2.5. Outliers treatment

- In the filtered dataset except mobile_number and churn columns all the columns are numeric types. Hence, converting mobile_number and churn datatype to object.

Step 2.6. Derive new features

- Deriving new column decrease_mou_action: This column 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: This column 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: This column 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: This column indicates whether the average revenue per customer has decreased in the action phase than the good phase.

- Deriving new column decrease_vbc_action: This column indicates whether the volume based cost of the customer has decreased in the action phase than the good phase.

Step 3. Building & Evaluation of the Model – EDA

3.1. Univariate analysis

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 behavior. 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

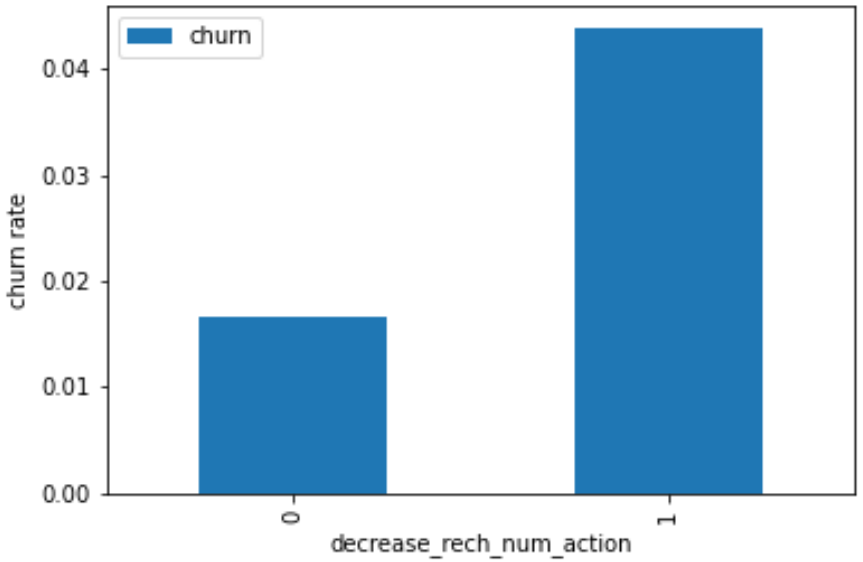
Analysis

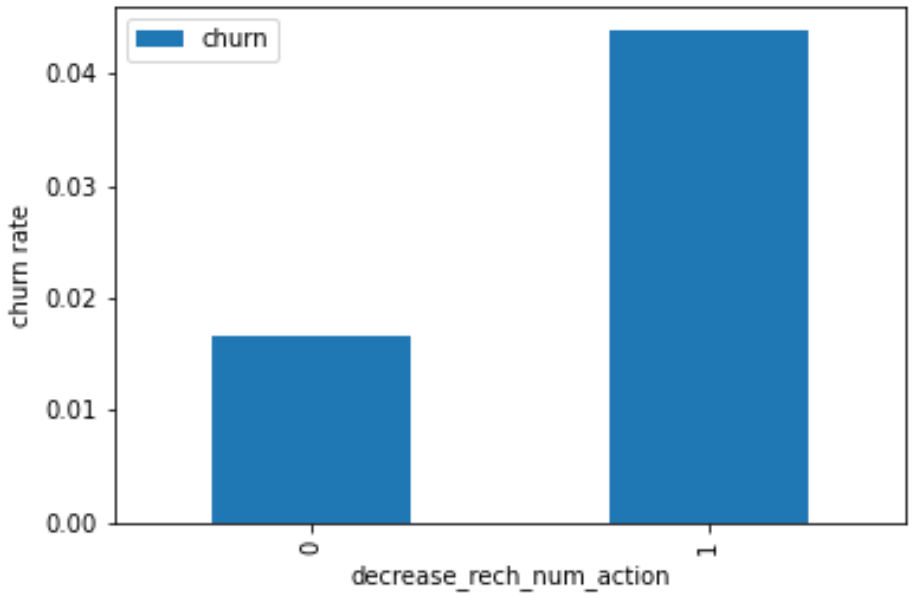
- Here we see the expected result. The churn rate is more for the customers, whose volume based cost in action month is increased. That means the customers do not do the monthly recharge more when they are in the action phase.

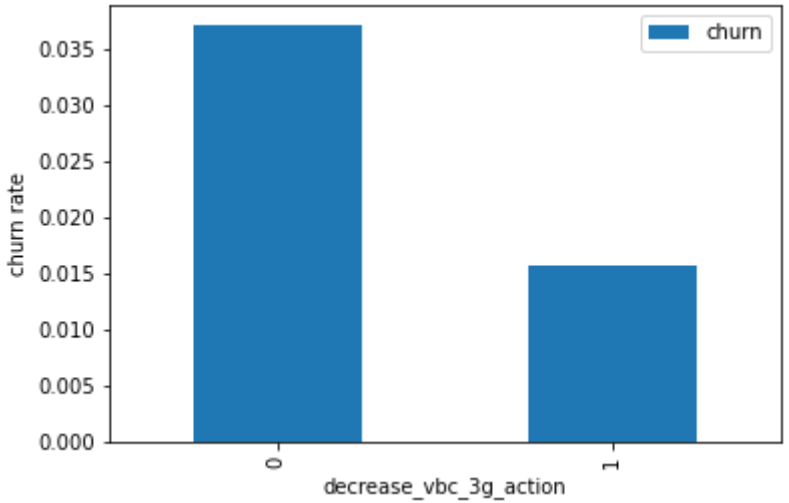
Analysis of the average revenue per customer (churn and not churn) in the action phase

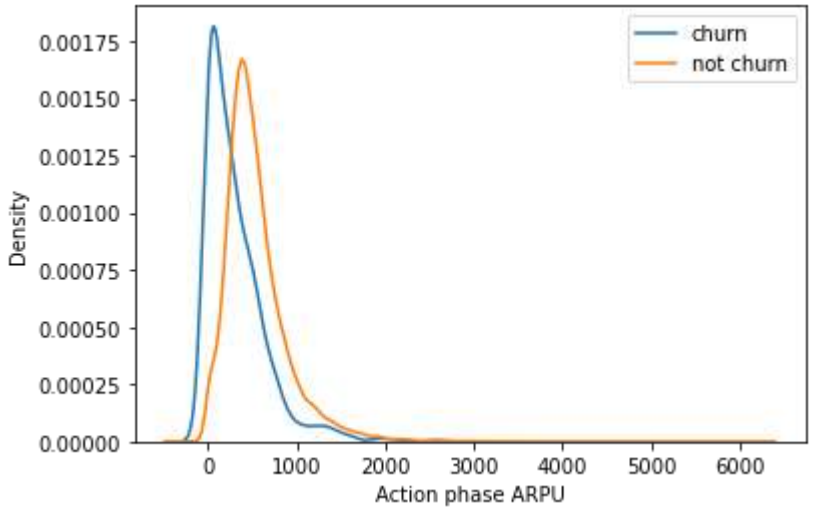
- Average revenue per user (ARPU) for the churned customers is mostly densed on the 0 to 900. The higher ARPU customers are less likely to be churned.
- ARPU for the not churned customers is mostly densed on the 0 to 1000.
- Analysis of the minutes of usage MOU (churn and not churn) in the action phase
- Minutes of usage(MOU) of the churn customers is mostly populated on the 0 to 2500 range. Higher the MOU, lesser the churn probability.

Graphical Representation of Univariate Analysis









3.2. Bivariate analysis

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.

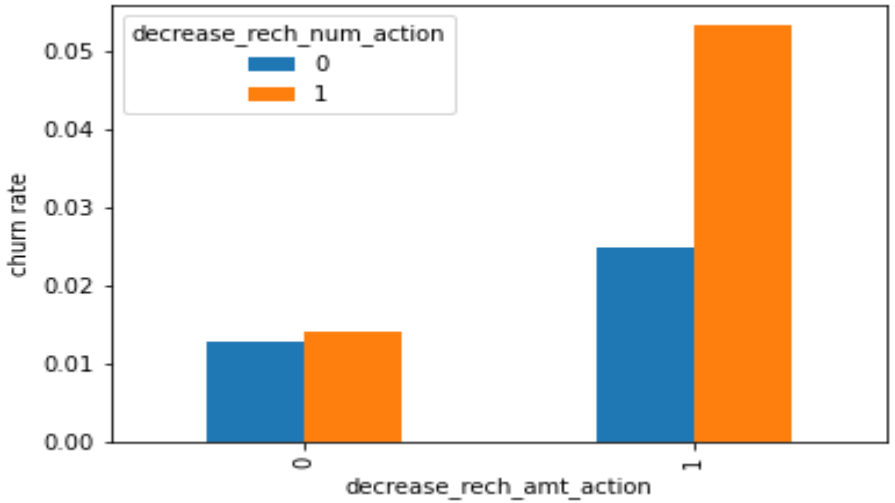
Analysis of recharge amount and number of recharge in action month
__Analysis__
We can see from the above pattern that the recharge number and the recharge amount are mostly proportional. More the number of recharge, more the amount of the recharge.

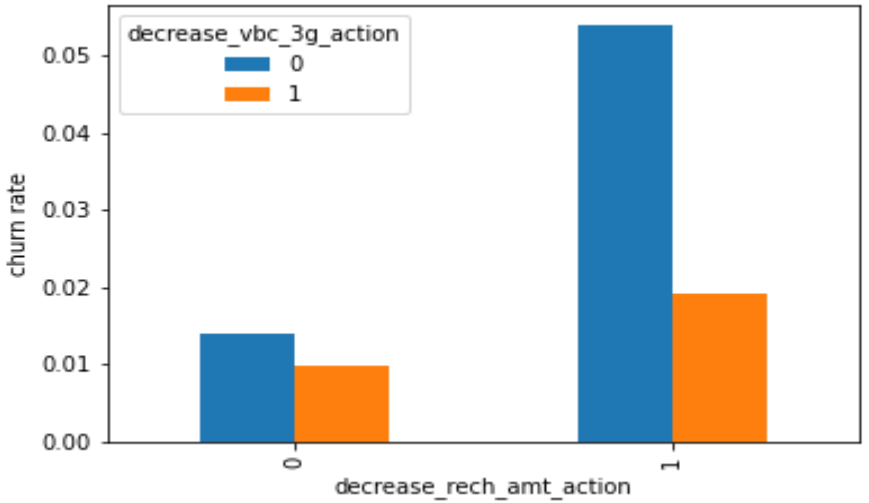
Dropping few derived columns, which are not required in further analysis

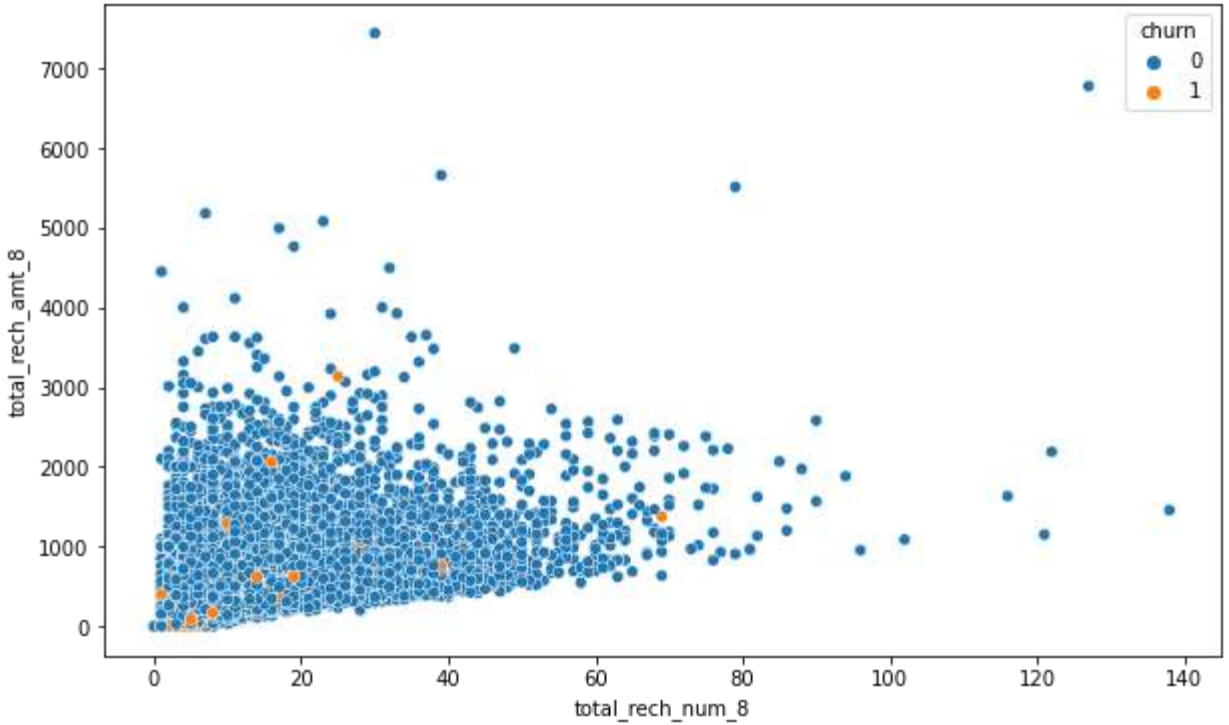
Train-Test Split:
Dealing with data imbalance
We are creating synthetic samples by doing up-sampling using SMOTE(Synthetic Minority Oversampling Technique).

Feature Scaling - Scaling the test set
We don't fit scaler on the test set. We only transform the test set.

Graphical Representation of Bivariate Analysis







3.3. Model with PCA

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.

Performing PCA with 60 components

Applying transformation on the test set: We are only doing Transform in the test set not the Fit-Transform. Because the Fitting is already done on the train set. So, we just have to do the transformation with the already fitted data on the train set.

Emphasize Sensitivity/Recall than Accuracy

- We are more focused on higher Sensitivity/Recall score than the accuracy.
- Because we need to care more about churn cases than the not 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 not churn customers as churn customers and provide them some incentives for retaining them. Hence, the sensitivity score is more important here.

3.3. Model with PCA (Contd..)

Logistic regression with PCA

Tuning hyperparameter C: C is the inverse of regularization strength in Logistic Regression. Higher values of C correspond to less regularization.

The highest test sensitivity is 0.8978916608693863 at C = 100

Logistic regression with optimal C:

__Model summary__

•**Train set**

- Accuracy = 0.86
- Sensitivity = 0.89
- Specificity = 0.83

•**Test set**

- Accuracy = 0.83
- Sensitivity = 0.81
- Specificity = 0.83

Overall, the model is performing well in the test set, what it had learnt from the train set.

3.3. Model with PCA (Contd..)

Decision tree with PCA

Model summary

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

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

3.4. Model without PCA

Logistic regression with No PCA

1. ***Model analysis***
    - We can see that there are few features have positive coefficients and few have negative.
    - Many features have higher p-values and hence became insignificant in the model.
2. ***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).
3. Feature Selection Using RFE
4. RFE with 15 columns

Model Creation:

- We will build model using RFE elimination with some random 15 variables selected by RFE.

Checking VIFs and P- Value:

We will eliminate features with high p-value or VIF till we get the optimal model for our analysis.

Optimal Model – Model 4:

We will use model 4 to proceed with out analysis as in this model we don’t have any variable with high p-value of VIF.

3.4. Model without PCA (Contd..)

Model without PCA

Model performance on the train set:

Creating a data frame with the actual churn and the predicted probabilities

Finding Optimal Probability Cutoff Point:

- Now let's 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 probability.
- Specificity - Increases with the increasing probability.
- At point 0.6 where the three parameters cut each other, we can see that there is a balance between sensitivity and specificity with a good accuracy.
- Here we are intended to achieve 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 achieving higher sensitivity, which is our main goal.

3.4. Model without PCA (Contd..)

Metrics

- We have got good accuracy, sensitivity and specificity on the train set prediction.
- Plotting the ROC Curve (Trade off between sensitivity & specificity)
- Testing the model on the test set
- Predictions on the test set with final model

__Model summary__

•**Train set**

- Accuracy = 0.83
- Sensitivity = 0.88

•**Test set**

- Accuracy = 0.77
- Sensitivity = 0.80

Overall, the model is performing well in the test set, what it had learnt from the train set.

Final 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 act upon for making the decision of the to be churned customers. Hence, the model is more relevant in terms of explaining to the business.

Business Recommendations

i. 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).
ii. Target the customers, whose outgoing others charge in July and incoming others on August are less.
iii. 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.
iv. Customers, whose monthly 3G recharge in August is more, are likely to be churned.
v. 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.
vi. Customers decreasing monthly 2G usage for August are most probable to churn.
vii. Customers having decreasing incoming minutes of usage for operators T to fixed lines of T for August are more 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.
- The number of monthly 3g data for August for the churn customers are very much populated around 1, whereas of non churn customers it spreaded across various numbers.
- Similarly we can plot each variables, which have higher coefficients, churn distribution.

Thank You..!