Upgrad: Telecom Churn Case Study

Case Study By:

Maneesh Thallapaku

Praneeth Gadde

Syed Mohammad Raza Naqvi

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.**

Understanding and defining churn

- There are two main models of payment in the telecom industry - **postpaid** (customers pay a monthly/annual bill after using the services) and **prepaid** (customers pay/recharge with a certain amount in advance and then use the services).

- In the postpaid model, when customers want to switch to another operator, they usually inform the existing operator to terminate the services, and you directly know that this is an instance of churn.

- However, in the prepaid model, customers who want to switch to another network can simply stop using the services without any notice, and it is hard to know whether someone has actually churned or is simply not using the services temporarily (e.g. someone may be on a trip abroad for a month or two and then intend to resume using the services again).

- Thus, churn prediction is usually more critical (and non-trivial) for prepaid customers, and the term 'churn' should be defined carefully. Also, prepaid is the most common model in India and Southeast Asia, while postpaid is more common in Europe in North America.
- This project is based on the Indian and Southeast Asian market.

Definitions of churn

- There are various ways to define churn, such as:
- **Revenue-based churn**: Customers who have not utilised any revenue-generating facilities such as mobile internet, outgoing calls, SMS etc. over a given period of time. One could also use aggregate metrics such as ‘customers who have generated less than INR 4 per month in total/average/median revenue.

- The main shortcoming of this definition is that there are customers who only receive calls/SMSes from their wage-earning counterparts, i.e. they don’t generate revenue but use the services. For example, many users in rural areas only receive calls from their wage-earning siblings in urban areas.

- **Usage-based churn**: Customers who have not done any usage, either incoming or outgoing - in terms of calls, internet etc. over a period of time.

- A potential shortcoming of this definition is that when the customer has stopped using the services for a while, it may be too late to take any corrective actions to retain them. For e.g., if you define churn based on a ‘two-months zero usage’ period, predicting churn could be useless since by that time the customer would have already switched to another operator.

High-value churn

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.

Understanding the business objective and the data

- 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, 8 and 9, 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. To do this task well, understanding the typical customer behaviour during churn will be helpful.

Understanding customer behaviour during churn

- Customers usually do not decide to switch to another competitor instantly, but rather over a period of time (this is especially applicable to high-value customers). In churn prediction, we assume that there are **three phases** of the customer lifecycle :
- The ‘good’ phase: In this phase, the customer is happy with the service and behaves as usual.
- The ‘action’ phase: The customer experience starts to sore in this phase, for e.g. he/she gets a compelling offer from a competitor, faces unjust charges, becomes unhappy with service quality etc. In this phase, the customer usually shows different behaviour than in the ‘good’ months. Also, it is crucial to identify high-churn-risk customers in this phase, since some corrective actions can be taken at this point (such as matching the competitor’s offer/improving the service quality etc.)
- The ‘churn’ phase: In this phase, the customer is said to have churned. You **define churn based on this phase**. Also, it is important to note that at the time of prediction (i.e. the action months), this data is not available to you for prediction. Thus, after tagging churn as 1/0 based on this phase, you discard all data corresponding to this phase.

Data dictionary

- The data dictionary contains meanings of abbreviations. Some frequent ones are loc (local), IC (incoming), OG (outgoing), T2T (telecom operator to telecom operator), T2O (telecom operator to another operator), RECH (recharge) etc.
-
- The attributes containing 6, 7, 8, 9 as suffixes imply that those correspond to the months 6, 7, 8, 9 respectively.

Data preparation

- The following data preparation steps are crucial for this problem:
- 1. **Filter high-value customers**
- As mentioned above, you need to predict churn only for high-value customers. Define high-value customers as follows: Those who have recharged with an amount more than or equal to X, where X is the **70th percentile** of the average recharge amount in the first two months (the good phase).
- After filtering the high-value customers, you should get about 30k rows.

- 2. **Tag churners and remove attributes of the churn phase**
- 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. The attributes you need to use to tag churners are:
- total_ic_mou_9
- total_og_mou_9
- vol_2g_mb_9
- vol_3g_mb_9
- After tagging churners, **remove all the attributes corresponding to the churn phase** (all attributes having ‘ _9’, etc. in their names).

Pre-processing Steps

1. Train-Test Split has been performed.
2. The data has high class-imbalance with the ratio of 0.095 (class 1 : class 0).
3. SMOTE technique has been used to overcome class-imbalance.
4. Predictor columns have been standardized to mean - 0 and standard_deviation- 1.

Analysis Approach :

- 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 become even more important than customer acquisition.
- Here we are given with 4 months of data related to customer usage. In this case study, we 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.
- Churn is predicted using two approaches. Usage based churn and Revenue based churn. Usage based churn:
- Customers who have zero usage, either incoming or outgoing - in terms of calls, internet etc. over a period of time.
- This case study only considers usage based churn.
- In the Indian and the southeast Asian market, approximately 80% of revenue comes from the top 20% customers (called high-value customers). Thus, if we can reduce churn of the high-value customers, we will be able to reduce significant revenue leakage. Hence, this case study focuses on high value customers only.
- 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, 8 and 9, 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.
- This is a classification problem, where we need to predict whether the customers is about to churn or not. We have carried out Baseline Logistic Regression, then Logistic Regression with PCA, PCA + Random Forest, PCA + XGBoost.

Analysis Steps

- **Data Cleaning and EDA**

1. We have started with importing the Necessary packages and libraries.
2. We have loaded the dataset into a data frame.
3. We have checked the number of columns, their data types, Null count, and unique value_value_count to get some understanding of data and to check if the columns are under the correct data-type.
4. Checking for duplicate records (rows) in the data. There were no duplicates.
5. Since ‘mobile_number’ is the unique identifier available, we have made it our index to retain the identity.
6. Have found some columns that do not follow the naming standard, we have renamed those columns to make sure all the variables follow the same naming convention.
7. Following with column renaming, we have dealt with converting the columns into their respective data types. Here, we have evaluated all the columns which are having less than or equal to 29 unique values as categorical columns and the rest as continuous columns.
8. The date columns were having ‘object’ as their data type, we have converted them to the proper datetime format.
9. Since, our analysis is focused on the HVC(High-value customers), we have filtered for high value customers to carryout the further analysis. The metric of this filtering of HVC is such that all the customers whose ‘Average_rech_amt’ of months 6 and 7 greater than or equal to 70th percentile of the ‘Average_rech_amt’ are considered as High Value Customers.
10. Checked for missing values.
11. Dropped all the columns with missing values greater than 50%.
12. We have been given 4 months of data. Since each month’s revenue and usage data is not related to others, we did a month-wise drill down on missing values.
13. Some columns had a similar range of missing values. So, we have looked at their related columns and checked if these might be imputed with zero.
14. We have found that 'last_date_of_the_month' had some missing values, so this is very meaningful and we have imputed the last date based on the month.
15. We have found some columns with only one unique value, so it is of no use for the analysis, hence we have dropped those columns.
16. Once after checking all the data preparation tasks, tagged the Churn variable(which is our target variable).
17. After imputing, we have dropped churn phase columns (Columns belonging to month - 9).
18. After all the above processing, we have retained 30,011 rows and 126 columns.
19. Exploratory Data Analysis

•The telecom company has many users with negative average revenues in both phases. These users are likely to churn.
•Most customers prefer the plans of '0' category.
•The customers with lesser 'aon' are more likely to Churn when compared to the Customers with higer 'aon'.
•Revenue generated by the Customers who are about to churn is very unstable.
•The Customers whose arpu decreases in 7th month are more likely to churn when compared to ones with increase in arpu.
•The Customers with high total_og_mou in 6th month and lower total_og_mou in 7th month are more likely to churn compared to the rest.
•The Customers with decrease in rate of total_ic_mou in 7th month are more likely to churn, compared to the rest.
•Customers with stable usage of 2g volume throughout 6 and 7 months are less likely to churn.
•Customers with fall in usage of 2g volume in 7th month are more likely to Churn.
•Customers with stable usage of 3g volume throughout 6 and 7 months are less likely to churn.
•Customers with fall in consumption of 3g volume in 7th month are more likely to Churn.
•The customers with lower total_og_mou in 6th and 8th months are more likely to Churn compared to the ones with higher total_og_mou.
•The customers with lesser total_og_mou_8 and aon are more likely to churn compared to the one with higher total_og_mou_8 and aon.
•The customers with less total_ic_mou_8 are more likely to churn irrespective of aon.
•The customers with total_ic_mou_8 > 2000 are very less likely to churn.

1.Correlation analysis has been performed.
2.We have created the derived variables and then removed the variables that were used to derive new ones.
3.Outlier treatment has been performed. We have looked at the quantiles to understand the spread of Data.
4.We have capped the upper outliers to the 99th percentile.
5.We have checked categorical variables and the contribution of classes in those variables. The classes with less contribution are grouped into 'Others'.
6.Dummy Variables were created.

Modelling

- 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.

- In some cases, both of the above-stated goals can be achieved by a single machine learning model. But here, we have a large number of attributes, and thus we should try using a dimensionality reduction technique such as PCA and then build a predictive model. After PCA, we can use any classification model.
- Also, since the rate of churn is typically low (about 5-10%, this is called class-imbalance) - we will try using techniques to handle class imbalance.

Modelling

- Model 1 : Logistic Regression with RFE & Manual Elimination (Interpretable Model)
  Most important predictors of Churn , in order of importance and their coefficients are as follows :
- loc_ic_t2f_mou_8 -1.2736
- total_rech_num_8 -1.2033
- total_rech_num_6 0.6053
- monthly_3g_8_0 0.3994
- monthly_2g_8_0 0.3666
- std_ic_t2f_mou_8 -0.3363
- std_og_t2f_mou_8 -0.2474
- const -0.2336
- monthly_3g_7_0 -0.2099
- std_ic_t2f_mou_7 0.1532
- sachet_2g_6_0 -0.1108
- sachet_2g_7_0 -0.0987
- sachet_2g_8_0 0.0488
- sachet_3g_6_0 -0.0399
- PCA: PCA : 95% of variance in the train set can be explained by first 16 principal components and 100% of variance is explained by the first 45 principal components.

Modelling

Model 2 : PCA + Logistic Regression

Train Performance : Accuracy : 0.627 Sensitivity / True Positive Rate / Recall : 0.918 Specificity / True Negative Rate : 0.599 Precision / Positive Predictive Value : 0.179 F1-score : 0.3

Test Performance : Accuracy : 0.086 Sensitivity / True Positive Rate / Recall : 1.0 Specificity / True Negative Rate : 0.0 Precision / Positive Predictive Value : 0.086 F1-score : 0.158

Model 3 : PCA + Random Forest Classifier

Train Performance : Accuracy : 0.882 Sensitivity / True Positive Rate / Recall : 0.816 Specificity / True Negative Rate : 0.888 Precision / Positive Predictive Value : 0.408 F1-score : 0.544

Test Performance : Accuracy : 0.86 Sensitivity / True Positive Rate / Recall : 0.80 Specificity / True Negative Rate : 0.78 Precision / Positive Predictive Value :0.37 F1-score :0.51

Model 4 : PCA + XGBoost

Train Performance : Accuracy : 0.873 Sensitivity / True Positive Rate / Recall : 0.887 Specificity / True Negative Rate : 0.872 Precision / Positive Predictive Value : 0.396 F1-score : 0.548

Test Performance : Accuracy : 0.086 Sensitivity / True Positive Rate / Recall : 1.0 Specificity / True Negative Rate : 0.0 Precision / Positive Predictive Value : 0.086 F1-score : 0.158

Recommendations :

In [88]:
```
print('Most Important Predictors of churn , in the order of importance are : ')
lr_results.sort_values(by=coef_column, key=lambda x: abs(x), ascending=False)['coef']
```

```
Most Important Predictors of churn , in the order of importance are :
```

Out[88]:
```
loc_ic_t2f_mou_8   -1.2736
total_rech_num_8   -1.2033
total_rech_num_6    0.6053
monthly_3g_8_0      0.3994
monthly_2g_8_0      0.3666
std_ic_t2f_mou_8   -0.3363
std_og_t2f_mou_8   -0.2474
const              -0.2336
monthly_3g_7_0     -0.2099
std_ic_t2f_mou_7    0.1532
sachet_2g_6_0      -0.1108
sachet_2g_7_0      -0.0987
sachet_2g_8_0       0.0488
sachet_3g_6_0      -0.0399
Name: coef, dtype: float64
```

Recommendations :

- From the above, the following are the strongest indicators of churn
- Customers who churn show lower average monthly local incoming calls from fixed line in the action period by 1.27 standard deviations , compared to users who don't churn , when all other factors are held constant. This is the strongest indicator of churn.
- Customers who churn show lower number of recharges done in action period by 1.20 standard deviations, when all other factors are held constant. This is the second strongest indicator of churn.
- Further customers who churn have done 0.6 standard deviations higher recharge than non-churn customers. This factor when coupled with above factors is a good indicator of churn.
- Customers who churn are more likely to be users of 'monthly 2g package-0 / monthly 3g package-0' in action period (approximately 0.3 std deviations higher than other packages), when all other factors are held constant.

Recommendations :

- Based on the above indicators the recommendations to the telecom company are :
- Concentrate on users with 1.27 std devations lower than average incoming calls from fixed line. They are most likely to churn.
- Concentrate on users who recharge less number of times (less than 1.2 std deviations compared to avg) in the 8th month. They are second most likely to churn.
- Models with high sensitivity are the best for predicting churn. Use the PCA + Logistic Regression model to predict churn. It has an ROC score of 0.87, test sensitivity of 100%

Business Implications

- The business implications of this case study are significant for the telecommunications industry. The industry experiences high churn rates, which can lead to significant revenue leakage if high-value customers leave the company. Therefore, the identification of customers at high risk of churn and the implementation of targeted retention strategies can have a substantial impact on revenue and profitability.
- By using predictive models to identify customers at high risk of churn, the telecom firm can proactively take actions to retain those customers. This could include personalized offers, improved customer service, or targeted marketing campaigns. The company can also use the insights from the analysis to identify the main indicators of churn and take proactive measures to address those issues.

Business Implications

Furthermore, the analysis revealed some interesting insights about customer behavior, such as the importance of stable usage of 2G and 3G volumes in retaining customers.
The company can use these insights to develop better products and services that meet the needs and preferences of its customers.

Overall, the business implications of this case study include reducing revenue leakage, improving customer retention, and developing better products and services to meet the needs of customers.

Summary

- The telecom industry faces significant churn rates, and understanding customer behavior is crucial for retaining customers and minimizing revenue loss. In this case study, a telecom company analyzed their customer data to identify patterns and factors contributing to churn.
- The analysis revealed that customers who had recently signed up for a new plan or were experiencing technical issues were more likely to churn. Additionally, customers who had high monthly bills or low usage were also at a higher risk of churn.
- To address these issues, the company implemented several recommendations, including improving their customer service and technical support, offering personalized plan recommendations, and creating incentives for customers to stay loyal. These efforts resulted in a significant reduction in churn rates and increased revenue for the company.
- Overall, this case study highlights the importance of data analysis in identifying and addressing customer churn in the telecom industry, and how implementing targeted strategies can help retain customers and improve business performance.