Telecom churn- Case Study

Submitted by- Shubhangi Khare & Shubham Goel

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 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.
- 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 period. 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 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 about data and to check if the columns are under 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 don't 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 rest as continuous columns.
8. After all the above processing, we have retained 30,011 rows and 126 columns.

Analysis Steps:

- **EDA**

- 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.
- 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.
- 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 total_ic_mou_8 > 2000 are very less likely to churn.

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.

Modelling

MODEL 1 SUMMARY

- ***Most important predictors of Churn , in order of importance and their coefficients are as follows:***

monthly_2g_8_1	-4.4332
monthly_3g_8_1	-3.8982
const	1.4828
loc_og_t2f_mou_6	-0.5478
monthly_3g_6_0	-0.4510
std_og_t2f_mou_8	-0.4457
sachet_2g_7_0	-0.2535
total_rech_num_8	-0.1954
sachet_2g_8_0	-0.1740
monthly_2g_6_0	-0.1303
total_rech_num_6	0.0758

- ***With the performance being:***

Train Performance:

Accuracy : 0.763
Sensitivity: 0.763
Specificity: 0.763
Precision: 0.763
F1-score: 0.763

Test Performance :

Accuracy : 0.762
Sensitivity: 0.694
Specificity: 0.768
Precision: 0.222
F1-score: 0.336

Modelling

MODEL 2 SUMMARY- LOGISTIC REGRESSION MODEL WITH PCA

- It is clear that 95% of variance in the train set can be explained by first 18 principal components and 100% of variance is explained by the first 45 principal components.

Modelling

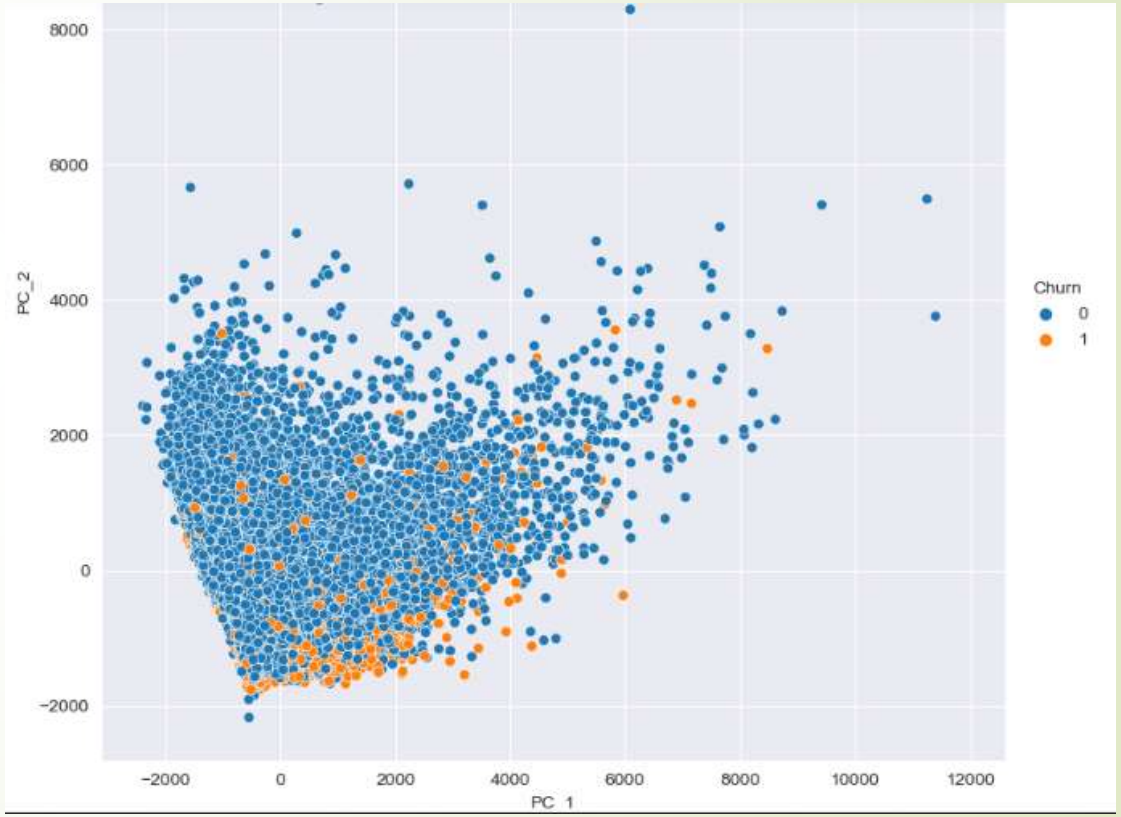
- ***With the performance being:***

```
Train Performance :

Accuracy : 0.578
Sensitivity: 0.933
Specificity:  0.545
Precision: 0.161
F1-score: 0.275

Test Performance :

Accuracy : 0.147
Sensitivity: 0.985
Specificity:  0.067
Precision: 0.092
F1-score: 0.168
```



Modelling

MODEL 3 SUMMARY- RANDOM FOREST WITH PCA

- ***With the performance being:***

Train Performance :

```
Accuracy : 0.885
Sensitivity: 0.822
Specificity:  0.891
Precision: 0.413
F1-score: 0.413
```

Test Performance :

```
Accuracy : 0.892
Sensitivity: 0.055
Specificity:  0.971
Precision: 0.156
F1-score: 0.081
```

Recommendations

- **Most important predictors of Churn, in order of importance and their coefficients are as follows:**

monthly_2g_8_1	-4.4332
monthly_3g_8_1	-3.8982
const	1.4828
loc_og_t2f_mou_6	-0.5478
monthly_3g_6_0	-0.4510
std_og_t2f_mou_8	-0.4457
sachet_2g_7_0	-0.2535
total_rech_num_8	-0.1954
sachet_2g_8_0	-0.1740
monthly_2g_6_0	-0.1303
total_rech_num_6	0.0758

- ***From the above the following can be inferred:***
  - *The best model for prediction if the churners was logistic regression with PCA.*
  - *Customers still using 2g services are very less likely to churn for telecom company.*
  - *Customers who churn show lower avg monthly local incoming calls from fixed line in the action period by 0.55 Standard deviation compared to users who don't churn.*
  - *Customers relying on sachet recharges are also less likely to churn.*

Recommendations

- **Recommendations for the company:**

- *Focus more on customers with there 0.55 standard deviations lower than avg incoming calls for fixed line.*
- *Concentrate on users who recharge a greater number of times in the 6th month. They are more likely to churn*

THANK YOU