Telecom Churn Case Study

- Introduction

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.

Objective

The objective of the case study is to predict churn in telecom customers.
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.

In this project, we analyzed 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.

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 customer lifecycle :

1. **The 'good' phase:** In this phase, the customer is happy with the service and behaves as usual.
2. **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 the 'good' months.
3. **The 'churn' phase:** In this phase, the customer is said to have churned. We 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 us for prediction

Data Preprocessing

- Cleaning and preparing the data for analysis, handling missing values, and encoding categorical variables.

Data Cleaning and EDA

1. We have started with importing Necessary packages and libraries.
2. We have loaded the dataset into a dataframe.
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. We have been given 4 months data. Since each months revenue and usage data is not related to other, we did month-wise drill down on missing values.
6. We have found that 'last_date_of_the_month' had some misisng values, so this is very meaningful and we have imputed the last date based on the month.
7. 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.
8. Once after checking all the data preparation tasks, tagged the Churn variable(which is our target variable).
9. After all the above processing, we have retained 30,011 rows and 126 columns.
10. After imputing, we have dropped churn phase columns (Columns belonging to month - 9)

Feature Engineering

- **Creating new features from the existing ones, scaling numerical features, and feature selection.**

1. Correlation analysis has been performed.
2. Outlier treatment has been performed. We have looked at the quantiles to understand the spread of Data.
3. We have checked categorical variables and contribution of classes in those variables. The classes with less ccontribution are grouped into 'Others'.
4. Dummy Variables were created.

Pre-processing Steps

a) Train-Test Split has been performed.

b) The data has high class-imbalance with the ratio of 0.095 (class 1 : class 0).

Model Training(With PCA)

Model with PCA

60 components explain amost more than 90% variance of the data. So, we performed PCA with 60 components.

Logistic regression with PCA:

Higher values of hyperparameter C corresponded to less regularization.

Model summary:

Train set: *Accuracy* = 0.87 *Sensitivity* = 0.90 *Specificity* = 0.83

Test set: *Accuracy* = 0.83 *Sensitivity* = 0.80 *Specificity* = 0.83

Decision tree with PCA

Five fold model was used along with hyperparameter tuning.

Model summary:

Train set: *Accuracy* = 0.89 *Sensitivity* = 0.92 *Specificity* = 0.86

Test set: *Accuracy* = 0.84 *Sensitivity* = 0.64 *Specificity* = 0.85

Model Training(Without PCA)

A. There were few features have positive coefficients and few had negative.
B. Many features had higher p-values and hence became insignificant in the model.
C. Feature selection was done and model were built so that all the variables are significant and there is no multicollinearity among the variables.

Accuracy stood at 0.6

Sensitivity was decreasing with the increased probablity.

Specificity was increasing with the increasing probablity.

Model summary:

Train set: *Accuracy* = 0.88 *Sensitivity* = 0.66 *Specificity* = 0.85

Test set: *Accuracy* = 0.84 *Sensitivity* = 0.76 *Specificity* = 0.84

Conclusion

- Summarizing the findings and discussing potential business implications.

Conclusion with PCA

- After trying several models it is shown that for acheiving the best sensitivity, 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 approx 85%.

Conclusion with no PCA

- The logistic model with no PCA has good sensitivity and accuracy, as compared to the models with PCA. The model also hels us to identify the variables which should be act upon for making the decision of the to be churned customers.

Recommendations

A. Target the customers, whose outgoing others charge in July and incoming others on August are less.
B. Customers, whose monthly 3G recharge in August is more, are likely to be churned.
C. Customers decreasing monthly 2g usage for August are most probable to churn.
D. 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.
E. 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.