Telecom Churn Case Study

Using Logistic Regression

Contents

Data cleaning

- Created new column for total data recharge as only average values were available
- Tagging churn
- Handling null values
  - Removed columns with missing values > 40%
  - Removed all the rows from the columns with missing values which were less than 5%
- Removing columns that do not play a role in analysis
  - Columns containing information about recharge
- Columns with only one value
  - These columns also don’t add any value to the analysis

- Deriving new features
  - Adding up ‘onnet_mou’ and ‘offnet_mou’ and creating ‘total_mou’ for each month
  - Adding up 6th month and 7th month data to good phase and treating the 8th month as action phase

Almost all columns had some outliers, while most of them were because they were 0, as the service war not used

These outliers were capped to proper upper limits values as we do not have high level of expertise in the subject matter

EDA Observations

- Many customer with more mou for std outgoing in good phase were churners
- Mou dropped significantly for churners in the action phase
- These users had other services than mou that were boositing revenue
- Users with less vbc were also generating high revenues
- Users with high local usage had less max recharge amount
- Users having max recharge amount < 200 churned more
- High max recharge and low incoming mou churned more

Data Preparation

- Standardization
  - StandardScaler was used from sklearn
- Handling class imbalance
  - SMOTE was used from imblearn
- Pca
  - Principal component analysis was done for 25 components

Model Building

1. Logistic Regression
    a. Unaltered X and y were used so that RFE could be used instead of PCA for comparison
    b. RFE was performed for 25 features
    c. Accuracy = 79.65%
    d. After removing two variable the accuracy dropped to 78.25%
    e. ROC curve - auc = 0.86
    f. Optimal cutoff point = 0.5
    g. Most of the critical features are from action phase

2. Decision Tree

a. Split performed as 0.7, 0.3
b. Initial accuracy = 87.98%
c. Recall = 89%
d. Auc = 93%

3. Decision Tree with Hyperparameter tuning

e. Accuracy = 90%
f. Recall = 93%
g. Auc = 100%

4. Random Forest

a. Accuracy = 87.98%
b. Recall = 93%
c. Auc = 96%

5. Random Forest with Hyperparameter tuning

d. Accuracy = 95%
e. Recall = 97%
f. Auc = 100%

6. Adaboost

a. Accuracy = 95%
b. Recall = 97%

Observations and Recommendation

We can see most of the top predictors are from the action phase, as the drop in engagement is prominent in that phase

Some of the factors we noticed while performing EDa which can be clubed with these inshigts are:

1. Users whose maximum recharge amount is less than 200 even in the good pahse, should have a tag and re-evaluated time to time as they are more likely to churn

2. Users that have been with the network less than 4 years, should be monitored time to time, as from data we can see that users who have been associated with the network for less than 4 years tend to churn more

3. MOU is one of the major factors, but data especially VBC if the user is not using a data pack if another factor to look out