Telecom Churn Case Study

Devishree

Faizal

Reshma

Problem Statement

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.

In this project, you 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.

Data Cleaning and Preparation

1. High value customers were filtered and tagged for churn or not churn. New features we derived since the data is too granular like total recharge amount for data, average recharge amounts for month 6 & 7. etc,.
2. Churn or not churn was identified and tagged basis on the incoming/outgoing calls information and mobile data usage.
3. Attributes with more than 40% of values missing was identified and dropped and those records with less than 5% missing data is dropped.
4. Columns with only one unique values is identified and dropped.
5. Correlation between features is calculated and highly correlated ones are removed.
6. New features:
Total minutes of usage (mou) is derived by adding onnet and offnet values.
Good phase features are derived by averaging out all values for month 6 and 7.
vbc columns have also been averaged out.
7. Outliers are handled by capping the upper limits.

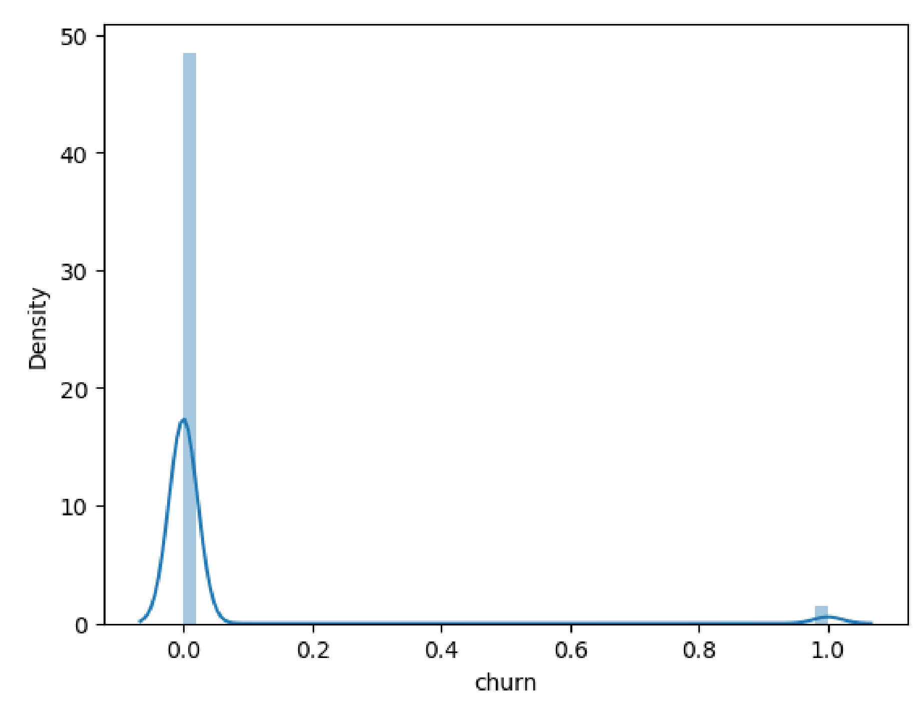
EDA

1. Majority of the churners have less than 4 years of tenure.
2. MOU have dropped significantly for the churners in the action phase i.e 8th month, thus hitting the revenue generated from them.
3. Users who were recharging with high amounts were using the service for local uses less as compared to user who did lesser amounts of recharge.
4. Users who had the max recharge amount less than 200 churned more.

Data Preparation for Modelling

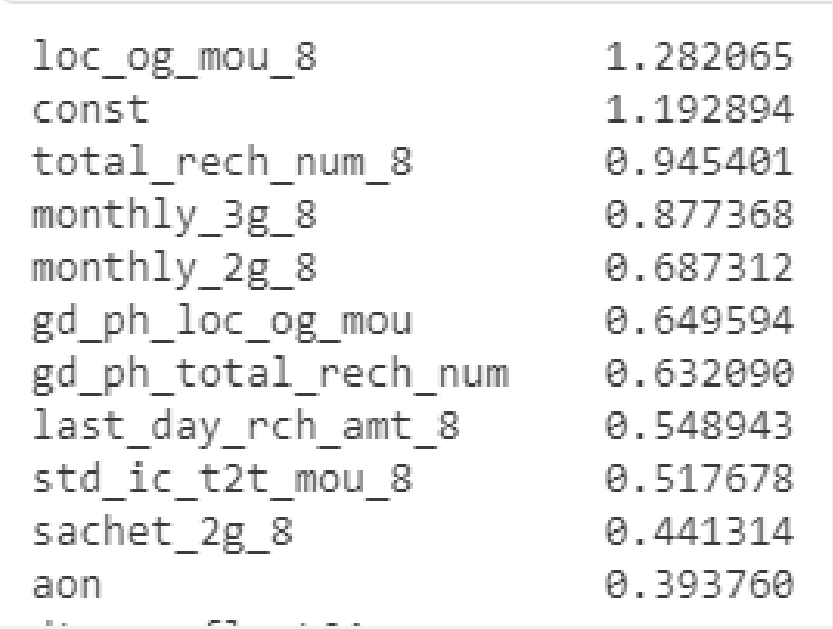
Standardization: StandardScaler
Class Imbalance: SMOTE
PCA used for feature selection keeping to 25 features.

Distribution of the target variable



Modelling – Logistic Regression

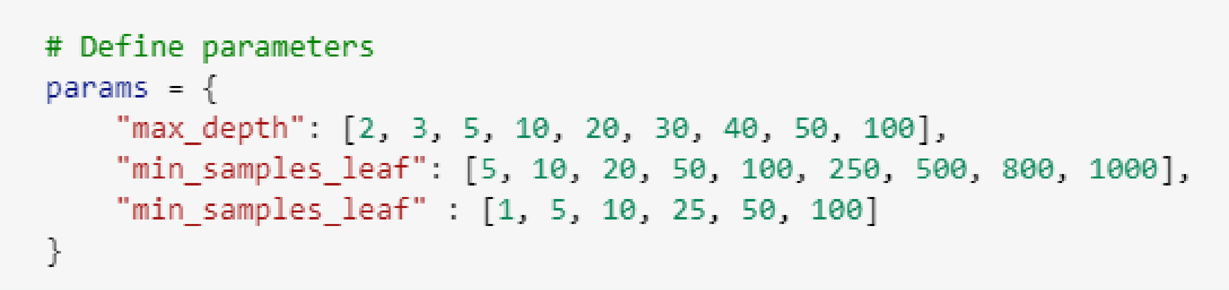
1. RFE used for feature selection
2. 0.5 as optimum threshold value
3. Accuracy of 78.5% on train data and 78.8% on test data
4. Most of the critical features are form the `action` phase, which is inline with the business understanding that `action` phase needs more attention.
Top 10 predictors:



```
loc_og_mou_8            1.282065
const                   1.192894
total_rech_num_8        0.945401
monthly_3g_8            0.877368
monthly_2g_8            0.687312
gd_ph_loc_og_mou        0.649594
gd_ph_total_rech_num    0.632090
last_day_rch_amt_8      0.548943
std_ic_t2t_mou_8        0.517678
sachet_2g_8             0.441314
aon                     0.393760
```

Modelling - Decision Tree

1.max_depth: 10.
2. Hyperparameter tuning done using 4 fold Gridsearch CV and accuracy is ~88%.
3. The parameters defined are as follows:

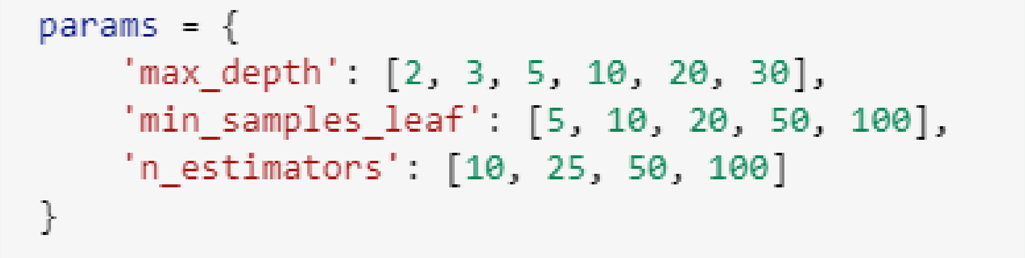


```
# Define parameters
params = {
    "max_depth": [2, 3, 5, 10, 20, 30, 40, 50, 100],
    "min_samples_leaf": [5, 10, 20, 50, 100, 250, 500, 800, 1000],
    "min_samples_leaf" : [1, 5, 10, 25, 50, 100]
}
```

4. The best estimator is the one with depth=50.

Modelling – Random Forest

1. number of estimators: 15, max_depth: 10, max_features=5.
2. Hyperparameter tuning done using 4 fold Gridsearch CV and accuracy is ~94%.
3. The parameters defined are as follows:

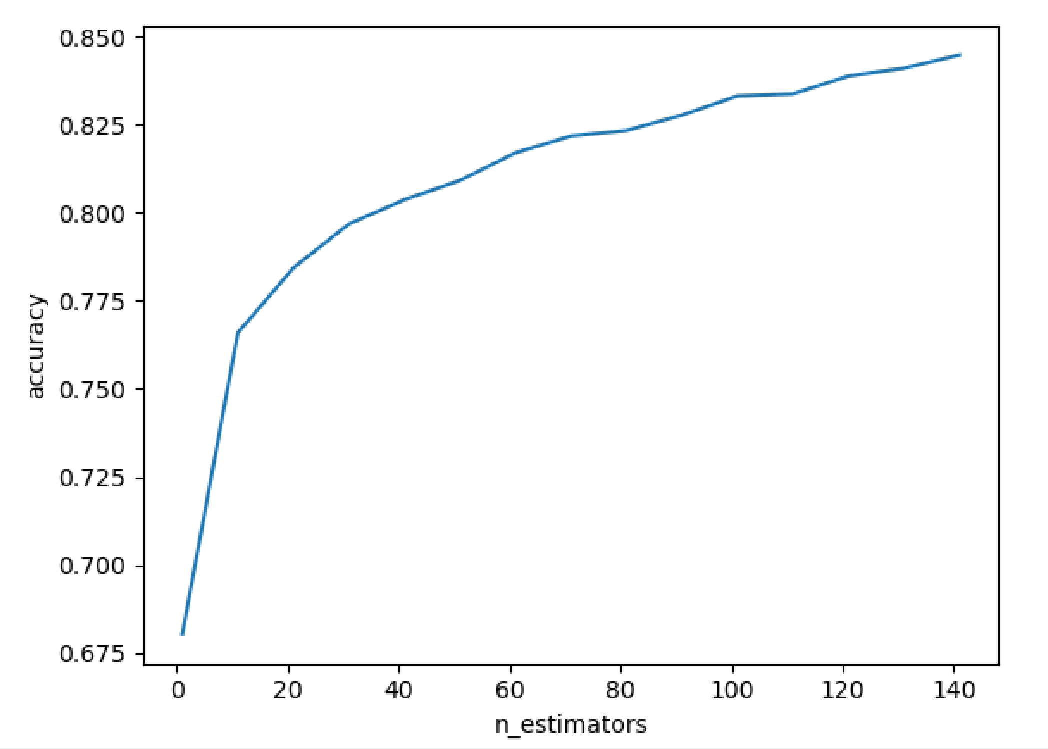


```
params = {
    'max_depth': [2, 3, 5, 10, 20, 30],
    'min_samples_leaf': [5, 10, 20, 50, 100],
    'n_estimators': [10, 25, 50, 100]
}
```

4. The best estimator is the one with depth=30.

Modelling - Adaboost

1. max_depth: 2
2. A shallow decision tree used as the base estimator.
3. Accuracy scores compared to the number of estimators :



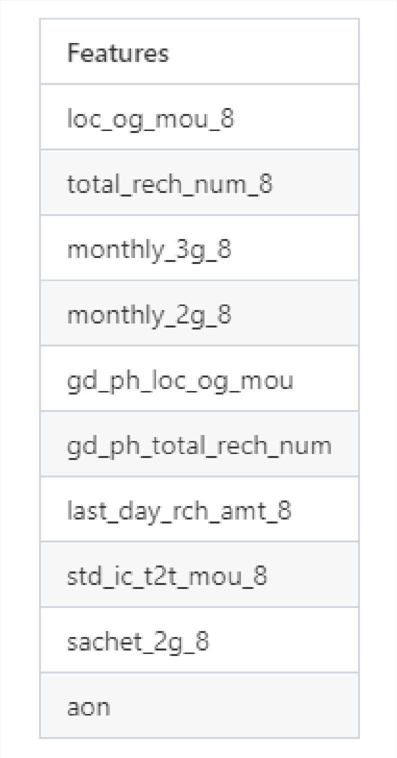
Conclusions

1. Given our business problem, to retain their customers, we need higher recall. As giving an offer to user not going to churn will cost less as compared to losing a customer and bring new customer, we need to have high rate of correctly identifying the true positives, hence recall.

2. When we compare the models trained we can see the tuned random forest and ada boost are performing the best, which is highest accuracy along with highest recall i.e. 95% and 97% respectively. So, we will go with random forest instead of adaboost as that is comparatively simpler model.

Strategies to Manage Customer Churn

The top 10 predictors



Features
loc_og_mou_8
total_rech_num_8
monthly_3g_8
monthly_2g_8
gd_ph_loc_og_mou
gd_ph_total_rech_num
last_day_rch_amt_8
std_ic_t2t_mou_8
sachet_2g_8
aon

Some of the factors we noticed while performing EDA which can be clubbed with these insights are:

1. Users whose maximum recharge amount is less than 200 even in the good phase, 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