Telecom Churn Case Study

06-Dec-2022

Problem Statement

Business Case:

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.

Problem Statement:

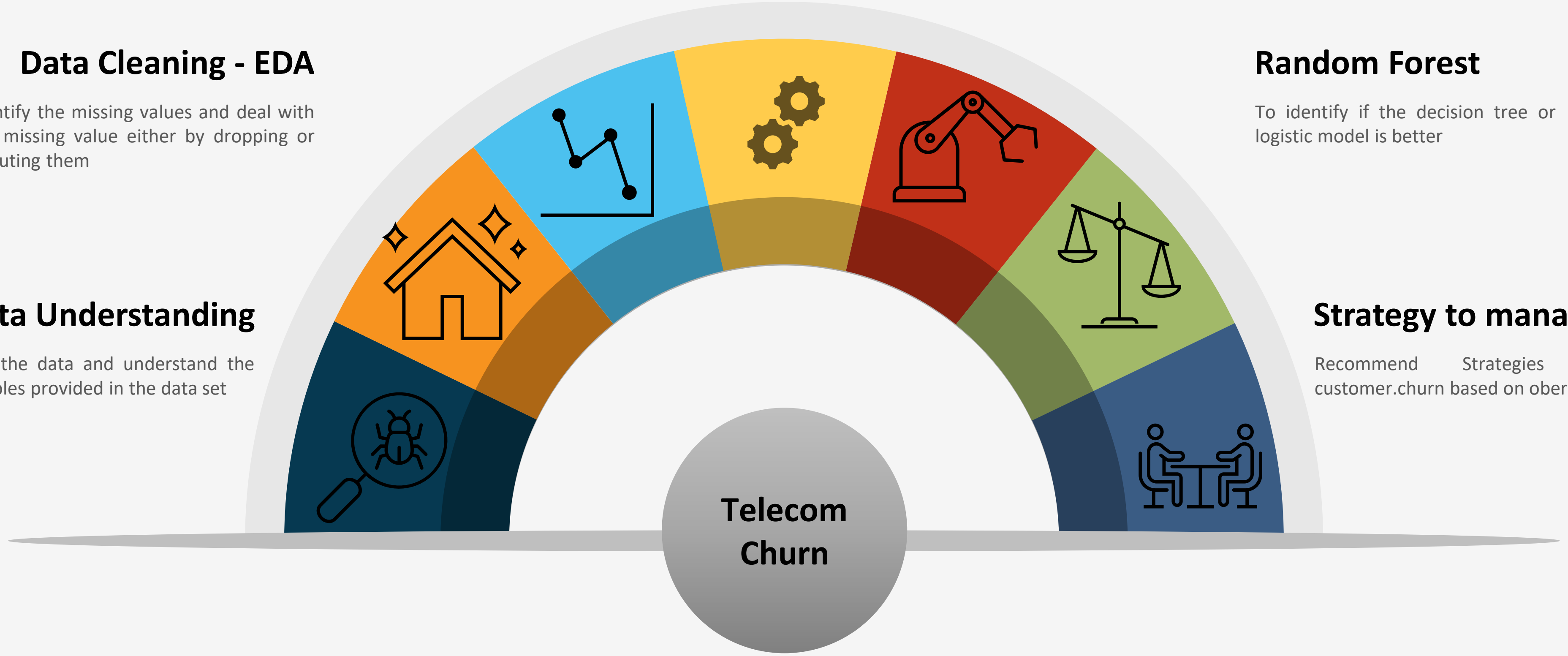
To reduce customer churn by identifying and predicting customers who are at high risk of churn

Objective:

There are two objective for

- ❑ Predict whether a high-value customer will churn or not, in near future by building up predictive model
- ❑ identify important variables that are strong predictors of churn and recommend strategies to manage customer churn. These variables may also indicate why customers choose to switch to other networks.

Case Study -Approach



Data Understanding

Load the data and understand the variables provided in the data set

Data Cleaning - EDA

Identify the missing values and deal with the missing value either by dropping or imputing them

Data Preparation & EDA

Define the high value customer > 70th percentile. Tag the churn and deal with the outliers.

PCA

To handle multicollinearity PCA for feature selection

Identify Model

Based on churn variables to be predicted through logistic regression or family of tree model

Random Forest

To identify if the decision tree or logistic model is better

Strategy to manage churn

Recommend Strategies to manage customer.churn based on obersation

Data Understanding, Preparation & EDA

Data Understanding:

Total no of Unique customer: 99999
No of variables: 226
Object: 12
Numeric: 214

EDA:

Missing value

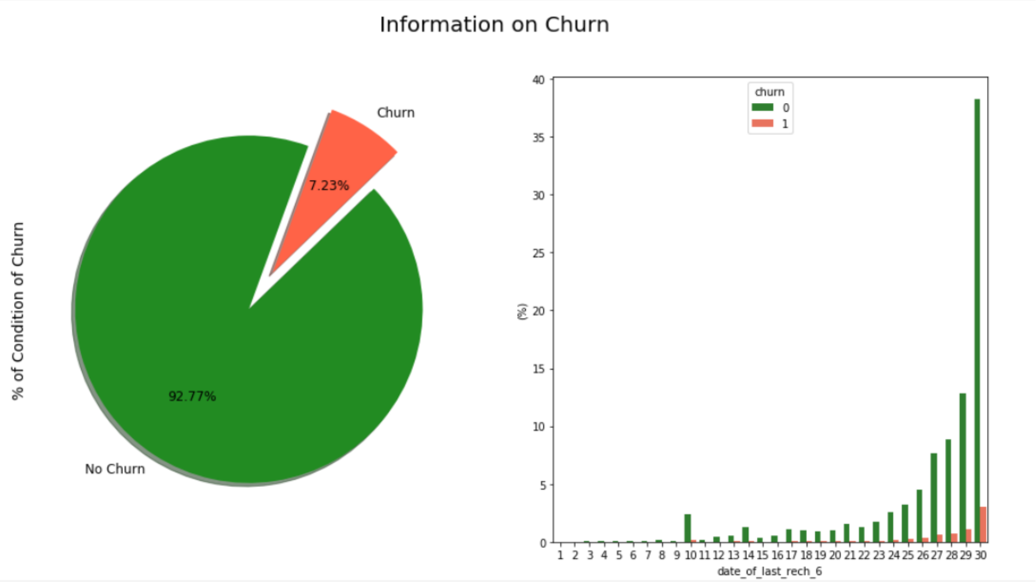
- 40 variable have more than 50% missing value
- The missing values have been imputed with 0 as the variable are significant for analysis
- The missing values in date column been dropped

Filter High Value customers:

No of High value customers:29906
% of High value customers: 30%

Churn:

- 2418 are tagged as churns out of 29906 high value customers
- From the above values we can see that there are just **8.09% churn** cases.
- Dataset is **highly imbalanced**, with the non-churners constituting the majority (91.91%) and the churn instances being the minority (8.09%).
- Post outlier and missing value treatment non churns constitute 92.77% and churn 7.23%

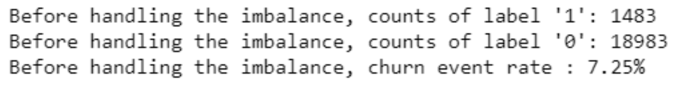


Information on Churn

Model Building

Handling Class imbalance:

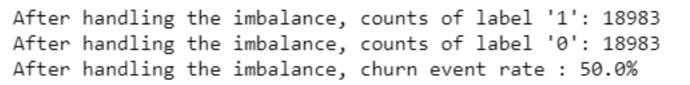
Before handling imbalance -



```
Before handling the imbalance, counts of label '1': 1483
Before handling the imbalance, counts of label '0': 18983
Before handling the imbalance, churn event rate : 7.25%
```

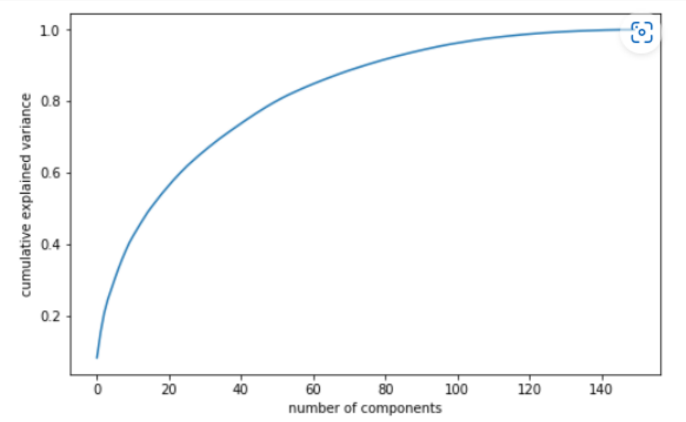
Considering there is significant amount of imbalance SMOTE technique is used to reduce class imbalance

After handling the imbalance –

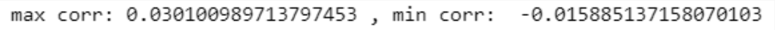


```
After handling the imbalance, counts of label '1': 18983
After handling the imbalance, counts of label '0': 18983
After handling the imbalance, churn event rate : 50.0%
```

PCA:



It can be observed that 80 components account for 90 % of the variance.
Considering large data incremental PCA is the model used



```
max corr: 0.030100989713797453 , min corr:  -0.015885137158070103
```

As can be seen, there is almost no link between the two components. Our data was successfully cleaned of multicollinearity, and as a result, our models will be significantly more reliable

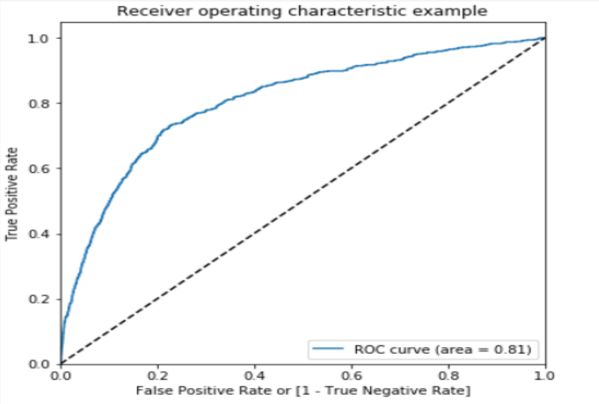
Model Building – Logistic Regression

Logistic Regression Model:

Logistic regression model has been chosen considering churn is a classification data

ROC curve:

Area under ROC curve is 0.81 which is closer to one which shows the model is better performing model when it comes to classification

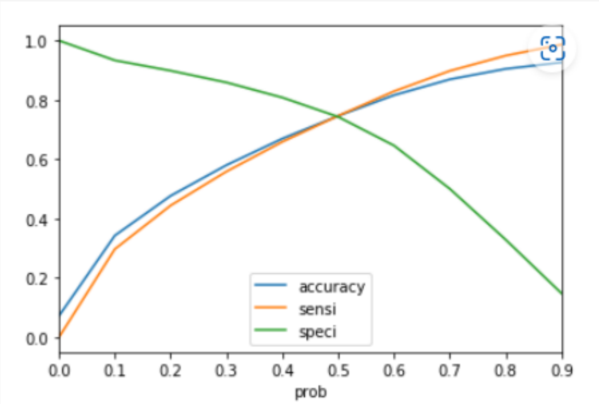


Churn with predicted y with x> 0.49:

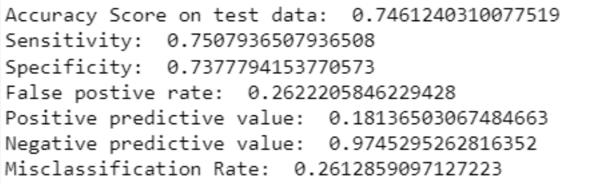
```
0    8142
1     630
```

Optimal probability threshold:

0.49 is the optimal threshold for logistic regression



Metrics of logistic regression model:



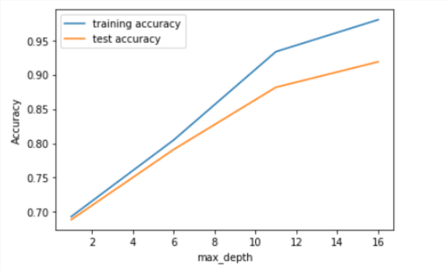
```
Accuracy Score on test data:  0.7461240310077519
Sensitivity:  0.7507936507936508
Specificity:  0.7377794153770573
False postive rate:  0.2622205846229428
Positive predictive value:  0.18136503067484663
Negative predictive value:  0.9745295262816352
Misclassification Rate:  0.2612859097127223
```

Model Building – Random Forest

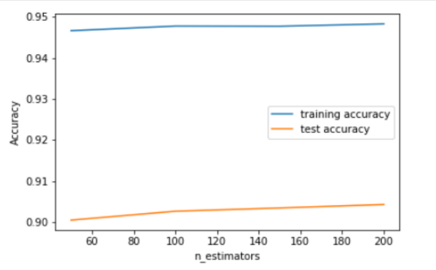
Random Forest:

Hyperparameter tuning

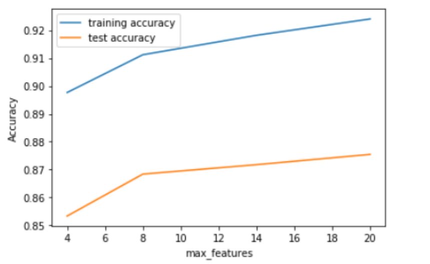
Max_depth:



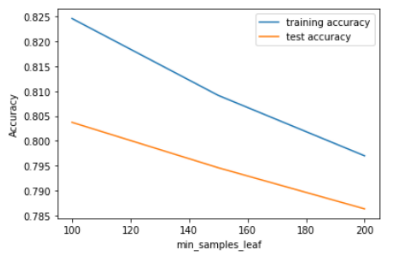
N-estimators:



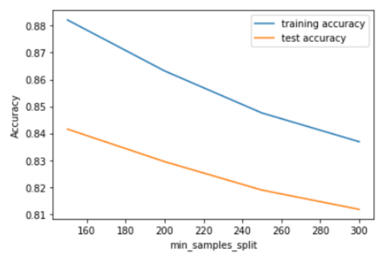
Max_features:



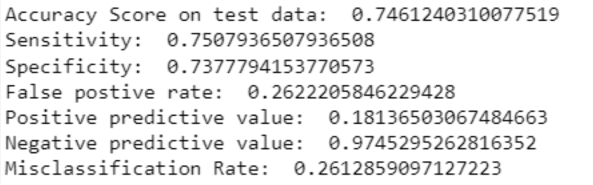
Min sample leaf:



Min sample split:

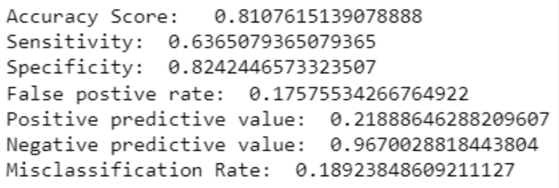


Metrics of logistic regression model:



```
Accuracy Score on test data:  0.7461240310077519
Sensitivity:  0.7507936507936508
Specificity:  0.7377794153770573
False postive rate:  0.2622205846229428
Positive predictive value:  0.18136503067484663
Negative predictive value:  0.9745295262816352
Misclassification Rate:  0.2612859097127223
```

Metrics of Random forest model:

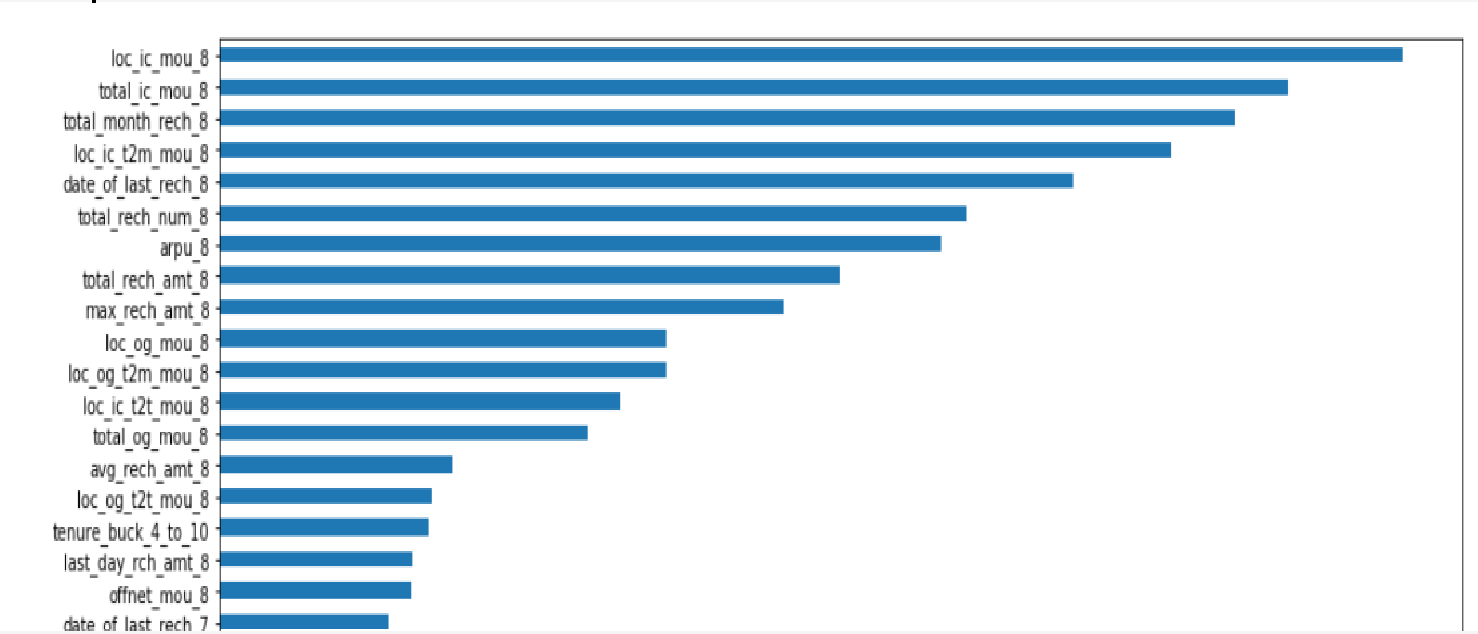


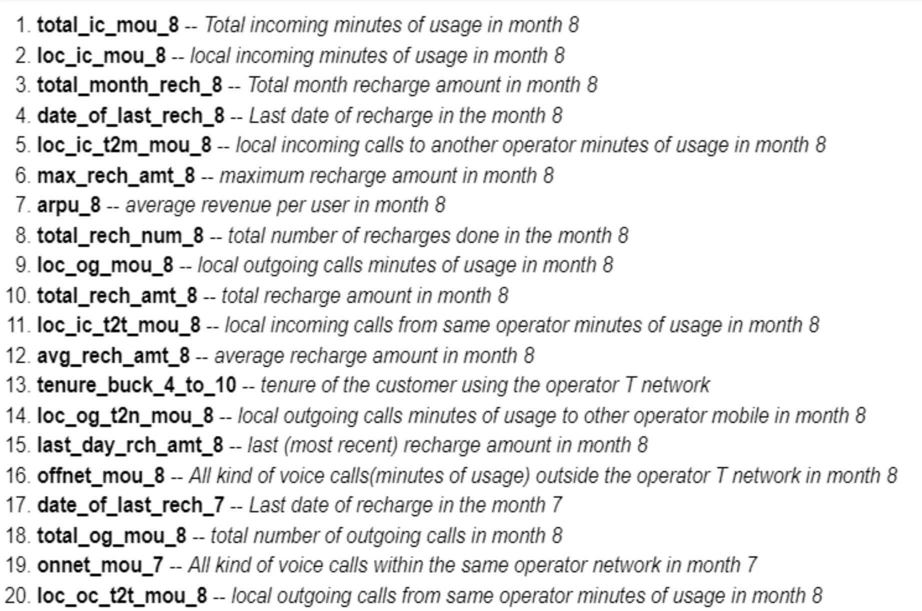
```
Accuracy Score:   0.8107615139078888
Sensitivity:  0.6365079365079365
Specificity:  0.8242446573323507
False postive rate:  0.17575534266764922
Positive predictive value:  0.21888646288209607
Negative predictive value:  0.9670028818443804
Misclassification Rate:  0.18923848609211127
```

Conclusion : The results of the customer churn analysis showed that Logistic Regression outperformed Decision Tree. As Logistic Regression has a higher sensitivity and better specificity score, we may be able to predict prospective client attrition.

Observations and Recommendations

Top Variables:





1. **total_ic_mou_8** -- *Total incoming minutes of usage in month 8*
2. **loc_ic_mou_8** -- *local incoming minutes of usage in month 8*
3. **total_month_rech_8** -- *Total month recharge amount in month 8*
4. **date_of_last_rech_8** -- *Last date of recharge in the month 8*
5. **loc_ic_t2m_mou_8** -- *local incoming calls to another operator minutes of usage in month 8*
6. **max_rech_amt_8** -- *maximum recharge amount in month 8*
7. **arpu_8** -- *average revenue per user in month 8*
8. **total_rech_num_8** -- *total number of recharges done in the month 8*
9. **loc_og_mou_8** -- *local outgoing calls minutes of usage in month 8*
10. **total_rech_amt_8** -- *total recharge amount in month 8*
11. **loc_ic_t2t_mou_8** -- *local incoming calls from same operator minutes of usage in month 8*
12. **avg_rech_amt_8** -- *average recharge amount in month 8*
13. **tenure_buck_4_to_10** -- *tenure of the customer using the operator T network*
14. **loc_og_t2n_mou_8** -- *local outgoing calls minutes of usage to other operator mobile in month 8*
15. **last_day_rch_amt_8** -- *last (most recent) recharge amount in month 8*
16. **offnet_mou_8** -- *All kind of voice calls(minutes of usage) outside the operator T network in month 8*
17. **date_of_last_rech_7** -- *Last date of recharge in the month 7*
18. **total_og_mou_8** -- *total number of outgoing calls in month 8*
19. **onnet_mou_7** -- *All kind of voice calls within the same operator network in month 7*
20. **loc_oc_t2t_mou_8** -- *local outgoing calls from same operator minutes of usage in month 8*

Recommendations:

Top characteristics are those that are predominantly associated with month 8, or the action phase, as determined by our Random Forest implementation.

Providing discounts and additional service packs during phase to customers would help in better churn

Thank You