DOMAIN ORIENTED TELECOM CHURN CASE STUDY

SUBMITTED BY- RAGUL

PRIYANKA SANGA

NEELANJAN ROY

CONTENTS

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.

BUSINESS OBJECTIVE

- To build a predictive model that will be used to predict whether a high-value customer will churn or not, in near future (i.e. churn phase). By knowing this, the company can take action steps such as providing special plans, discounts on recharge etc.
- It will be used to identify important variables that are strong predictors of churn. These variables may also indicate why customers choose to switch to other networks.

MODEL BUILDING APPROACH

- Importing dataset
- Data cleaning and Data preparation
- Filtering high value customers
- Handling Missing values and Outliers
- Exploratory Data Analysis(EDA)
- Train-Test dataset split
- Handling Class Imbalance
- Feature Scaling
- Model building with top 15 features selected through RFE
- Model fine tuning
- Model evaluation
- Model prediction on Test dataset
- Defining Model Summary and providing insightful recommendations

DATA PREPARATION

- Firstly we filtered out the Customers who have recharged with an amount more than or equal to X, where X is the 70th percentile of the average recharge amount in the first two months (the good phase). We call this the dataset of the high value customers
- This returns 30011 records of high value customers out of 99999 total rows
- Then we identified the Churned/Not Churned customers from the new dataset
- Removed all the attributes corresponding to the churn phase (all attributes having ‘ _9’, etc. in their names)

HANDLING MISSING VALUES

- We dropped the columns which had more than 30% of the data missing i.e.Null. Also we dropped the columns which had only unique value in each and every record
- For the features having less than 30% Null records, we replaced the null values with the Median of that particular feature using ***SimpleImputer*** package from *sklearn*

EXPLORATORY DATA ANALYSIS

- None of the variables are in normal distribution
- Average revenue per user for all 3 months has more values between 0 to 3000.
- Further, distribution of the three variables for three months looks almost similar which proves that customers follows same pattern of recharges and services utilization which generates more or less same amount of revenue.
- Through correlation matrix, we could see high correlation between arpu(Average revenue per user) and total_rech_amt

HANDLING CLASS IMBALANCE AND RESAMPLING

We have utilized the following resampling techniques:

- Random Under-Sampling
- Random Over-Sampling
- SMOTE - Synthetic Minority Oversampling Technique
- ADASYN - Adaptive Synthetic Sampling Method
- SMOTETomek - Over-sampling followed by under-sampling

All the above resampling techniques have been run on the logistic regression & random forest model and performance of all resampling techniques have been compared.

From the observation, we could see ADASYN with logistic regression model has attained recall value of 82.5%

HANDLING CLASS IMBALANCE AND RESAMPLING

Performance of ADASYN with logistic regression model:
Accuracy: 82.11
Recall: 82.5

As the business objective is to find the potential customers who could churn out, it is important to focus on recall value as the inclusion error could be tolerated.

Distribution of churn classes on the resampled training set with ADASYN:
[(0, 19192)
(1, 18669)]

```
Accuracy:  0.8211905819635718
F1 score:  0.4436765722183829
Recall:  0.8251928020565553
Precision:  0.30340264650283555

 clasification report:
              precision    recall  f1-score   support

           0       0.98      0.82      0.89      8226
           1       0.30      0.83      0.44       778

    accuracy                           0.82      9004
   macro avg       0.64      0.82      0.67      9004
weighted avg       0.92      0.82      0.85      9004
```

MODEL BUILDING AND EVALUATION

- Feature selection done through RFE Feature selection from which 5 features were deleted by looking at the p-value and VIF score and re-fit the model after deletion of each feature
- Model evaluation on the Train dataset-

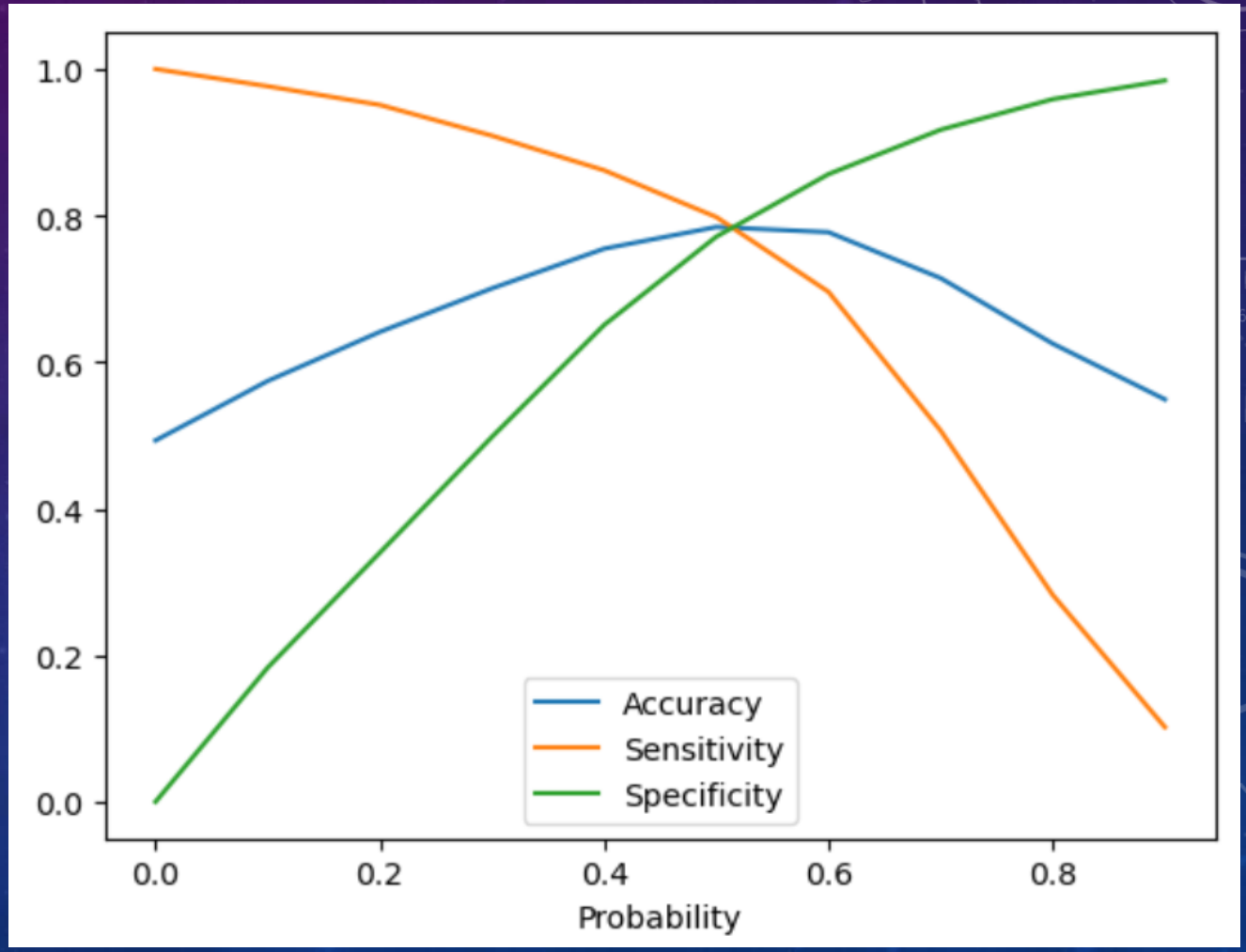
Accuracy	78.43%
Sensitivity	79.79%
Specificity	77.11%

MODEL SELECTION

- Based on the logistic regression models built, the model logm7 has been chosen as the P-value of the coefficients of the variables are 0.00 which is less than the threshold limit of 5%. Hence, we could infer that coefficients are statistically significant.
- Further, VIF values of the features are within the permissible limit of 5% which took away the problem of multicollinearity.
- logm7 has been built with 10 variables which makes the model less complex with high interpretability.
- Performance of model on the training set and test dataset is almost similar which states that model is consistent.

- Plotting the Accuracy, Sensitivity and Specificity we could figure out that the Probability cut-off for getting the optimal result is 0.5
- Results after using cut-off 0.5 increased sensitivity

Accuracy	78.43%
Sensitivity	79.79%
Specificity	77.11%
Recall	79.79%



PREDICTION ON TEST DATASET

Accuracy	77.32%
Sensitivity	83.16%
Specificity	76.76%
Recall	83.16%

- Top features which are responsible for the churn or not churn decision of the customers are:

➢ total_og_mou_8
➢ total_og_mou_7
➢ loc_og_t2m_mou_8
➢ total_rech_amt_6
➢ total_rech_num_8
➢ loc_og_mou_6
➢ loc_ic_t2m_mou_8
➢ roam_og_mou_8
➢ roam_ic_mou_8
➢ last_day_rch_amt_8

INFERENCES

- As per the final logistic model, we could see that features total_og_mou_8, last_day_rch_amt_8 & loc_ic_t2m_mou_8 plays crucial role in deciding the churn of the customers. Hence, it is recommended to reach out to the customers who haven't made any outgoing calls, minimal recharge amount and customers who have dropped minutes of incoming calls during the action phase.
- Coefficients of the above variables strongly indicates that these features has significant effect on churn of customers. Hence, drop in the either of the above variable's numbers should be given an importance.
- total_rech_num_8 and loc_og_mou_6 are other crucial indicators of the churn.

BUSINESS RECOMMENDATIONS

- Offering competitive pricing and plans
- Personalizing the customer experience
- Improving network quality and reliability
- Implementing loyalty programs to reward long-term customers with special incentives, discounts, or exclusive services.
- Investing in self-service options
- Offering long-term contracts with discounts and benefits to encourage commitment from customers.