Telecom Churn Prediction

Presented for Group

1) Saurabh B Ghavghave
2) Arpitha Balashankar
3) Akash Raaj Singh

Problem Statement

To reduce customer churn, telecom companies need to predict which customers are at high risk of churn. In this project, we will analyze customer-level data of a leading telecom build predictive models to identify customers at high risk of churn and identify the

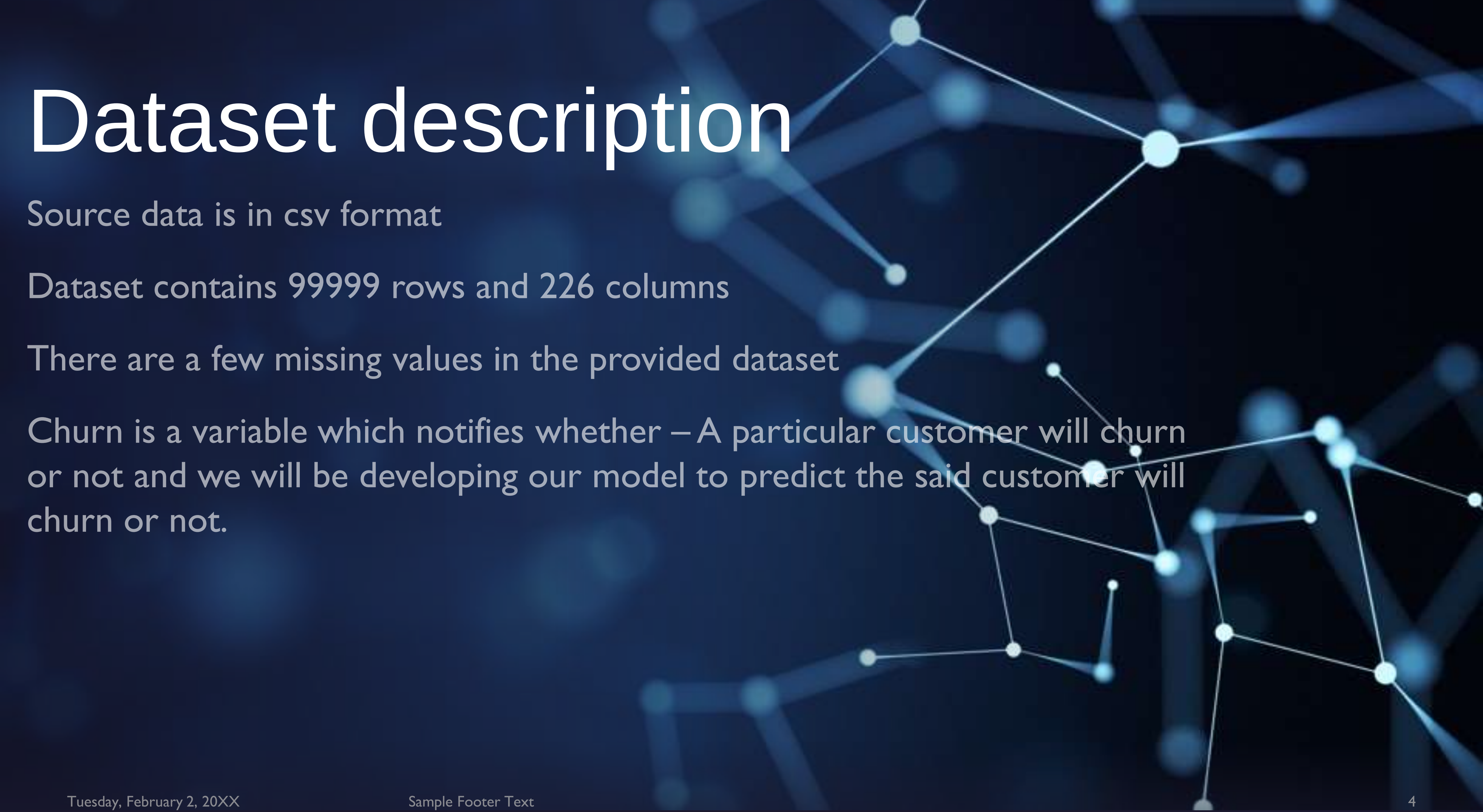
indicators of churn.

Retaining high profitable customers is the main business goal here.

Project Objective

- To predict customer churn
- Highlighting the main variables/factors influencing customer churn
- Use ML algorithm to build predictive model, evaluate the accuracy and performance of the built model based on various metrics.
- Providing executive Summary

Dataset description



Source data is in csv format

Dataset contains 99999 rows and 226 columns

There are a few missing values in the provided dataset

Churn is a variable which notifies whether – A particular customer will churn or not and we will be developing our model to predict the said customer will churn or not.

Steps / Methodologies

1. **Reading, understanding and visualizing the data**: Reading the dataset “telecom_churn.csv” and Understanding it . Performing basic data checks like shape, info, describe, converting data types wherever required, etc. Importing visualization libraries to visualize dependent and independent variables via Univariate and Bi-variate analysis.

2. Preparing the data for modelling: In case studies with machine learning models, Data cleaning plays an important role. The quality and efficiency of the model depends on the data cleaning. Hence it must be done precisely. a. Calculation of missing values for each column and dropping columns named date_cols (contains dates of all four months not required for our modelling)

b. Dropping the columns with high percentage (30%) of missing values.

c). **Filter high-value customers**: Since, we need to predict churn only for high-value customers, we define high-value customers as follows: Those 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). After filtering the high-value customers, you should get about 30k rows.

d). Removing rows having more than 50% missing values.

e). Dropping again all those columns of 6, 7, 8 & 9 MOU months having least percentages of missing

f) Tag churners and remove attributes of the churn phase: We then tagged the churned customers (churn=1, else 0) based on the fourth month as follows: Those who have not made any calls (either incoming or outgoing) AND have not used mobile internet even once in the churn phase. The attributes we used to tag churners are: total_ic_mou_9, total_og_mou_9, vol_2g_mb_9, vol_3g_mb_9. After tagging churners, we removed all the attributes corresponding to the churn phase (all attributes having ‘ _9’, etc. in their names). And hence we get about 96.6% of non-churn rate implying high class imbalance as opposed to only 3.4% of churn rate. g). **Data Cleaning**: We handled the outliers by capping the data above 90th percentile and below 10th percentile. h). **Deriving new features**: We derived some new features like decrease_mou_action, decrease_rech_num_ action, decrease_rech_amt_action, decrease_arpu_action, & decrease_vbc_action to analyse whether behaviour declined from good phase to action phase and performed visualization on the same and pointed out few observations.

3. Building the model: Before diving deep into model building, we did **train-test split**(80:20), dealt with **imbalanced data** using SMOTE methodology and done **feature scaling** namely standard scaling. Of all the models out there, we use logistic regression to solve this problem (to handle categorical target variable). Before model building, we had about 120 columns and we created generalized linear model taking into account all variables. Since there were lots of variables to deal with, we moved to automatic feature selection. We used both RFE and manual elimination methods to get the final list of columns. During which the most insignificant, highly correlated, based on p-value and VIF variables were dropped.

4). Evaluating the model: - . We know that the relationship between ln(odds) of ‘y’ and feature variable “X” is much more intuitive and easier to understand. The equation is: ---

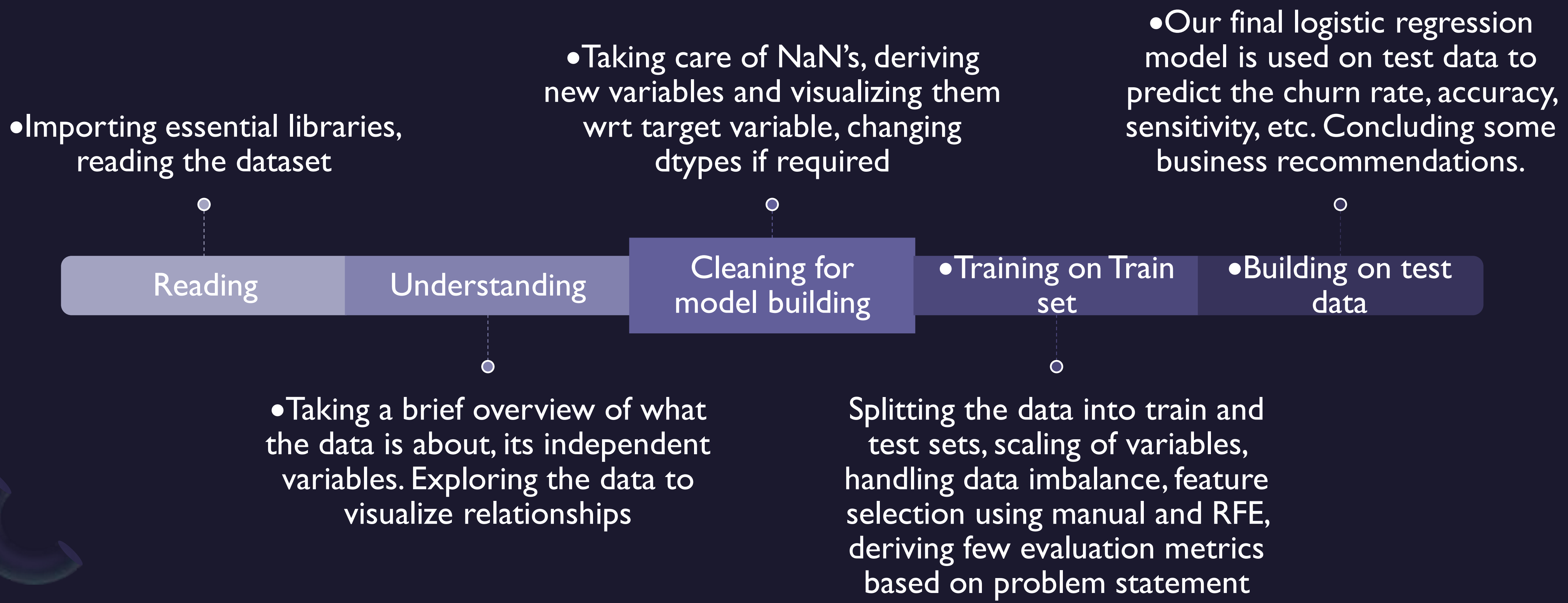
a) -1.3285 - 0.0963 * onnet_mou_8 + 0.6539 * offnet_mou_7 - 0.6897 * offnet_mou_8 - 1.2748 * isd_og_mou_8 - 1.8546 * og_others_7 - 0.5630 * loc_ic_t2f_mou_8 - 3.5787 * loc_ic_mou_8 - 0.7228 * std_ic_t2f_mou_8 - 1.6468 * ic_others_8 - 0.8806 * monthly_2g_8 - 0.9765 * monthly_3g_8

b) We then created the confusion matrix and evaluated accuracy, sensitivity and specificity scores for training data. We then went on to create Roc curve and in order to find optimal cut off point, plotted accuracy sensitivity and specificity for various probabilities.

c). We chose the cutoff probability as 0.50 from Accuracy, Sensitivity, Specificity curve and calculated y_train_pred_final based on this cut-off . The sensitivity of model (on train data) was around 90% giving ROC area of 0.91

d). Predictions were made on the test set using the same cut-off probability and calculated metrices such as Accuracy, Sensitivity and specificity. The sensitivity of our logistic regression model came out to be 83%

Timeline



Reading

- Importing essential libraries, reading the dataset

Understanding

- Taking a brief overview of what the data is about, its independent variables. Exploring the data to visualize relationships

Cleaning for model building

- Taking care of NaN’s, deriving new variables and visualizing them wrt target variable, changing dtypes if required

- Training on Train set

Splitting the data into train and test sets, scaling of variables, handling data imbalance, feature selection using manual and RFE, deriving few evaluation metrics based on problem statement

- Building on test data

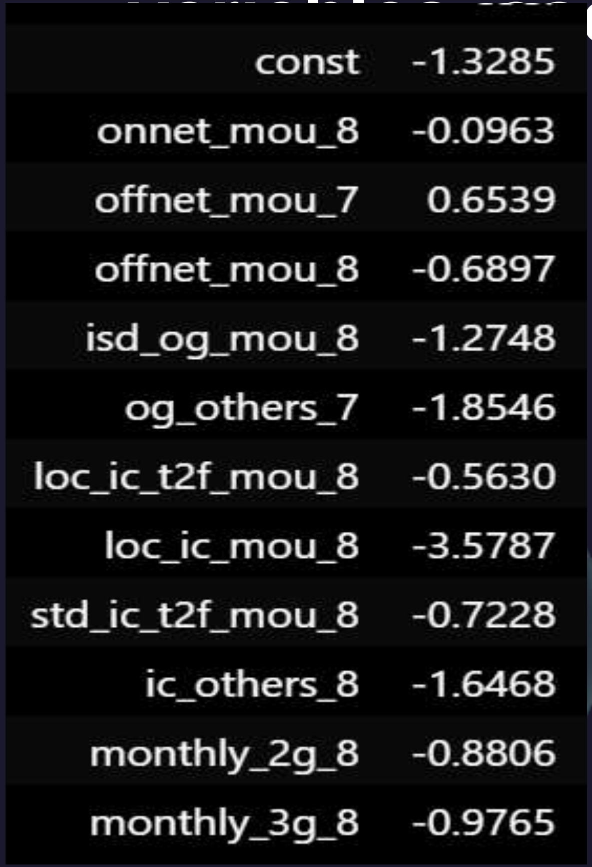
- Our final logistic regression model is used on test data to predict the churn rate, accuracy, sensitivity, etc. Concluding some business recommendations.

Business recomendation

Top predictors

Below are the few top variables selected in the logistic regression model

variables coefficients



const	-1.3285
onnet_mou_8	-0.0963
offnet_mou_7	0.6539
offnet_mou_8	-0.6897
isd_og_mou_8	-1.2748
og_others_7	-1.8546
loc_ic_t2f_mou_8	-0.5630
loc_ic_mou_8	-3.5787
std_ic_t2f_mou_8	-0.7228
ic_others_8	-1.6468
monthly_2g_8	-0.8806
monthly_3g_8	-0.9765

Title

Findings and suggestions

We can see most of the top variables have negative coefficients. That means, the variables are inversely correlated with the churn probablity.

E.g.:- If the local incoming minutes of usage (loc_ic_mou_8) is lesser in the month of August than any other month, then there is a higher chance that the customer is likely to churn.

Recomendations

1. Target the customers, whose minutes of usage of the incoming local calls and outgoing ISD calls are less in the action phase (mostly in the month of August).

2. Target the customers, whose outgoing others charge in July and incoming others on August are less.

3. Also, the 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.

4. Cutomers, whose monthly 3G recharge in August is more, are likely to be churned.

5. 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.

Findings and suggestions

6. Cutomers decreasing monthly 2g usage for August are most probable to churn.

7. Customers having decreasing incoming minutes of usage for operators T to fixed lines of T for August are more likely to churn.

8. offnet_mou_7 variables have positive coefficients (0.65). That means for the customers, whose all types of calls (incoming, outgoing, isd, etc) minutes of usage for July is increasing are more likely to churn.

Title

Final Predictions

Our dataset had: 0’s (non-churn): 96.6% 1’s (churn): 3.4%

After Splitting the dataset in 80:20

ON TRAIN SET

- Accuracy: 82%
- Sensitivity: 89%
- Specificity: 75%
- Confusion Matrix: Predicted

```
Actual              Non-Churn Churn
Non-churn        [[16171 5254]
Churn             [ 2304 19121]]
```

ON TEST SET

- Accuracy: 75%
- Sensitivity: 83.41%
- Specificity: 74.7%
- Confusion Matrix: Predicted

```
Actual              Non-Churn Churn
Non-churn        [[3995 1353]
Churn              [  32 161]]
```

Summary

Of all the evaluation metrices available viz Accuracy, Sensitivity, Specificity, Precision, Recall, etc., we provided higher weightage to Sensitivity because we were more interested in finding churn rate and extending promotions to those who are most likely to churn and we used logistic Regression model to do so as our target variable was Categorical

Thank You

Presented for group: Saurabh Ghavghave, Arpitha Balashankar & Akash Raaj Singh

By Saurabh Ghavghave