Telecom Churn Case Study

Group Members:

Amit Pal Singh

Varun Eknath

Business Problem:

To reduce customer churn, telecom companies need to **predict which customers are at high risk of churn.**

Data Characteristics:

The dataset contains customer-level information for a span of four consecutive months - June, July, August and September. The months are encoded as 6, 7, 8 and 9, respectively.

The **business objective** is to predict the churn in the last (i.e. the ninth) month using the data (features) from the first three months. To do this task well, understanding the typical customer behaviour during churn will be helpful.

Resolution to be achieved:

We need to filter out the High value customers (top 70 percentile) and recognise a pattern that takes into consideration their spending/business behaviour with the Telecom company.

From this filtered dataset, we need to find customers who are about to churn, i.e. in the Action phase and what are the values or characteristics of there customers, so that an effective predictive model can be built to predict future high-value customer churn which has high precision, specificity & optimal sensitivity

Predictive model used:

Logistic Regression

Note: Trying Classification led to failure due to implementation error

Understanding the Data:

The dataset contains the charges record incurred by the customers for various types of services like Standard, ISD, Roaming, Special, incoming & outgoing call records & internet data recharges for 2G & 3G. This is recorded over a span of 4 months i.e. 6^{th}, 7^{th}, 8^{th} & 9^{th} month.

Data Preparation:

1. All necessary libraries are imported, especially metrics & sklearn
2. A new column for the end dates of the months 6,7,8,9 are created which is in date format
3. Columns total_og_mou & total_ic_mou null value entries are imputed with 0, meaning this is for the Churned customers and can be removed later on.
4. New columns are created for the total data recharge amount for the 6th, 7th, 8th, 9th month by multiplying total recharge data & average recharge amount for those respective months

Data Preparation:

5. To check the high paying customers we add the values from the columns total_data_rech_amt & total_rech_amt for the 6,7,8,9 months to get the amount that the Telecom company receives in full by the customers.
6. Now we add up the total amount received by the company for the 6^{th}, 7^{th} & 8^{th} month, excluding the 9^{th} month to get filter out the churned customers at the end of month 9.
7. After applying the 70 percentile filter, we calculate the churned high paying customers by assigning a Churn column with value 0 for every value that is zero in vol_3g_mb & vol_2g_mb at the end of 9^{th} month
8. From this dataset we only consider 'mobile_number', 'total_ic_mou_9', 'total_og_mou_9', 'vol_2g_mb_9','vol_3g_mb_9','Churn' columns as they are the determining factors for data behaviour and drop any persisting values with Churn = 0.
9. This is assigned to a dataset prime1 which will be the basis of our Data modelling and Prediction using Logistic Regression

Data Modelling:

1. The dataset prime1 is divided into prime1_class1 which has all the data in prime1 except the Churn column & prime1_class2 which only has the Churn column.
2. For prime1_class1, we focus on the type of services and decide to assign the categorical classification for the months 6,7 & 8.
3. The classifications will be for
    1. Local, Roaming
    2. Offnet, On Net
    3. Standard-ISD, Special

4. The values of 1 & 0 will be applied for the above columns to create the required categorical classification for the 6^{th}, 7th , 8^{th} month. After this we remove the obsolete columns.

5. For numerical data like 3G/2G data used. We created a separate column for the three months and put values that is the sum up of all data used values

Regression using RFE:

1. We first split the prime1_class1 & prime1_class2 into X & y respectively. Then we create X_train, y_train & X_test, y_test by dividing them into a 70-30 split

2. We call upon logistic regression & RFE from sklearn package. Using RFE to filter out columns with high importance into Regression, we fit the model for X_train & y_train. We then obtain the statistical results for Train set

Dep. Variable:	Churn	**No. Observations:**	21000
Model:	GLM	**Df Residuals:**	20973
Model Family:	Binomial	**Df Model:**	26
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3648.1
Date:	Sun, 03 Mar 2024	**Deviance:**	7296.2
Time:	20:35:02	**Pearson chi2:**	3.73e+05
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.2006
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.4755	0.134	3.555	0.000	0.213	0.738
arpu_6	0.0004	7.07e-05	5.717	0.000	0.000	0.001
total_og_mou_8	-0.0003	7.27e-05	-4.303	0.000	-0.000	-0.000
total_ic_mou_7	0.0017	0.000	10.708	0.000	0.001	0.002
total_ic_mou_8	-0.0046	0.000	-13.921	0.000	-0.005	-0.004
last_day_rch_amt_6	0.0006	0.000	2.715	0.007	0.000	0.001
last_day_rch_amt_8	-0.0027	0.000	-6.663	0.000	-0.003	-0.002
roam_any_8	0.9599	0.070	13.697	0.000	0.823	1.097
local_any_6	-0.6454	0.169	-3.810	0.000	-0.977	-0.313
local_any_7	-0.3148	0.161	-1.952	0.051	-0.631	0.001
local_any_8	-1.5729	0.121	-12.988	0.000	-1.810	-1.336
std_any_6	0.0867	0.124	0.700	0.484	-0.156	0.330
std_any_7	0.2799	0.125	2.245	0.025	0.036	0.524
std_any_8	-0.6120	0.112	-5.455	0.000	-0.832	-0.392
spl_any_6	0.2730	0.090	3.031	0.002	0.096	0.449
spl_any_7	0.0971	0.083	1.168	0.243	-0.066	0.260
spl_any_8	-0.2273	0.085	-2.669	0.008	-0.394	-0.060
data_used_8	-0.0005	8.31e-05	-6.080	0.000	-0.001	-0.000

Taking VIF values & Confusion Matrix:

To Calculate VIF:

1. We call upon the variance_inflation_factor from statsmodel.
2. Make a empty dataframe vif, in which we run the variance_inflation_factor upon the Train data set, after reshaping the vif dataframe so that it fits.
3. After displaying vif data, we find high VIF values for local-any_7,6,8 & std_any_7,6,8. Instead of removing or reducing this value we try to find a cut-off for confusion matrix.
4. Assuming values for cut-off from 0-1, i.e. 0.0, 0.1, 0.2…1.0, we select 0.1 as the cut-off for our first trial and get below confusion matrix.

	Features	VIF
17	local_any_7	44.71
16	local_any_6	39.69
18	local_any_8	29.67
20	std_any_7	12.89
19	std_any_6	12.43
21	std_any_8	11.63
22	spl_any_6	5.44
3	total_ic_mou_8	5.32
2	total_ic_mou_7	5.20
24	spl_any_8	5.11
23	spl_any_7	5.05
25	data_used_8	3.53

```
[[18906   354]
 [  917   823]]
```

Confusion Matrix without cut-off

```
array([[17220,  2040],
       [  359,  1381]], dtype=int64)
```

Confusion Matrix with cut-off 0.1

Converting Probability values and finding Precision, Specificity & Sensitivity

1. We convert the probability values in the y_train and build a prediction model from it called y_train_pred.

2. From the confusion matrix obtained after applying cut-off, we use y_train_pred to calculate Precision, Specificity & Sensitivity to get their respective values.

   Precision = ~79%
   Specificity = ~89%
   Sensitivity = ~40%

	Converted	Converted_prob	predicted	final_predicted
0	1	0.598133	1	1
1	0	0.148086	0	1
2	0	0.008357	0	0
3	0	0.044076	0	0
4	0	0.008998	0	0

```
# Sensitivity = ~40%
# Specificity = ~89%
# Precision = ~ 79%
```

Conclusion:

A predictive model, to predict the churn of High Value customers in the "Action Phase" was created using previous data & Logistic regression with optimal values for Precision, Specificity & Sensitivity.

We also find while referring the statistical Results for the regression model, the **Standard data charges** for **Roaming** & **Local** service types, have a strong correlation with Churn data. Hence its advisable to manipulate the prices to retain high value customers from churning.