TELECOM CHURN CASE STUDY

DS C50 – KUSH, SAHANA AND SANJEEV

INDEX

INTRODUCTION TO PROBLEM STATEMENT

- Customers have the option to choose from multiple telephone service providers. Being a highly competitive industry, the telecommunications industry has an average of 15-25% annual churn rate.
- The fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, makes customer retention very important when compared to customer acquisition.
- The way to solve this is to focus on customer experience of high net-worth individuals and identify the attributes contributing to the churn of such customers.
- Objective of the current study was to predict churn and identify the key drivers of churn in each business division using simulated customer data sets.

GOALS OF THE CASE STUDY

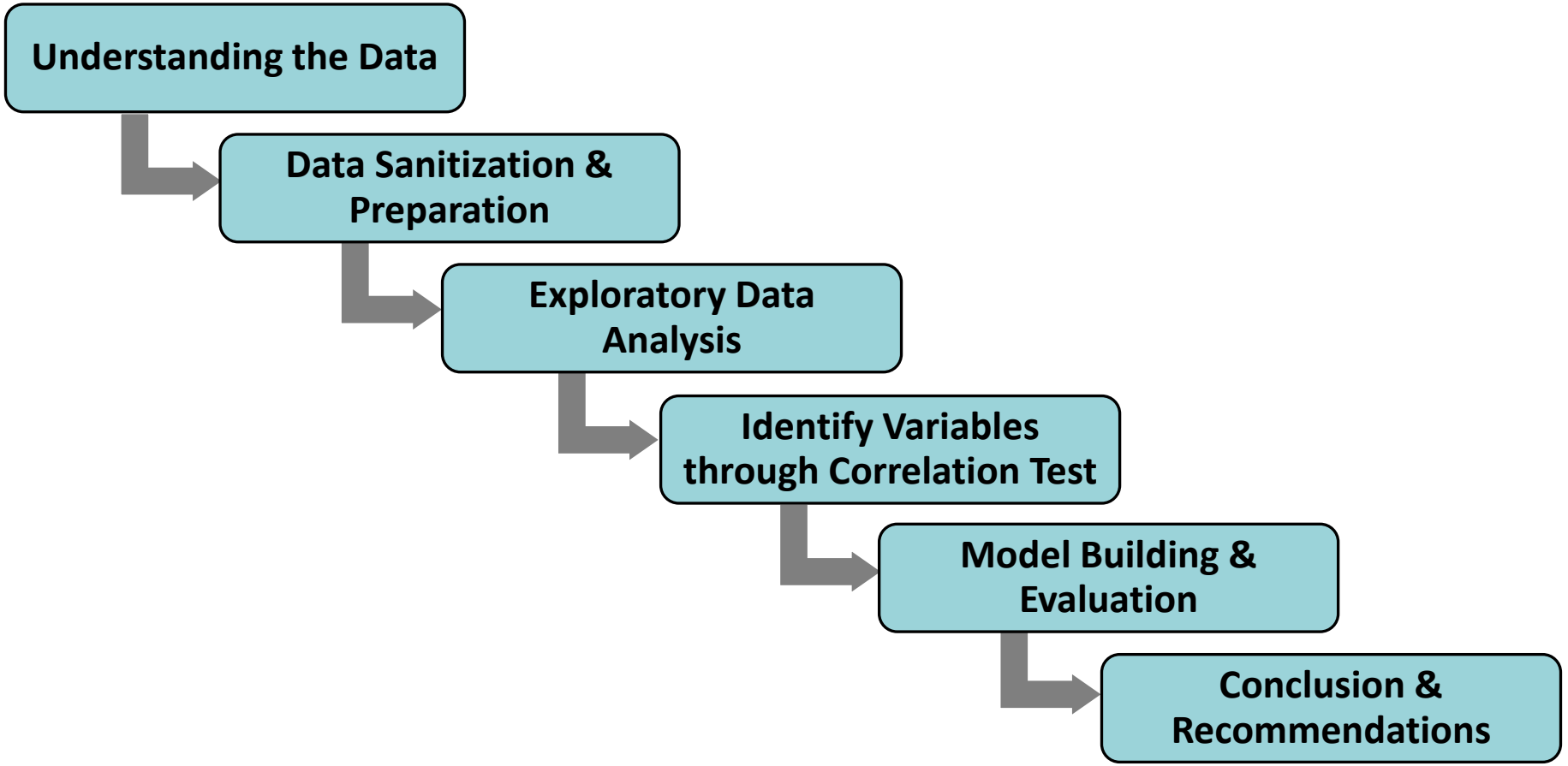
Logistic Regression Modelling:

- To build a logistic regression model to identify attributes that lead to churn of customer which can later be used to improve customer experience and retention.

Goals of the Project:

- To equip use-based churn as a strategy to identify the churn rate
- To identify the variables that are contributing the most to teh churn through exploratory data analysis
- Building a model to define high-value customers based on a certain metric (mentioned later below) and predict churn only on high-value customers

APPROACH TAKEN



DATA PREPARATION & SANITIZATION

- Null Value elimination
  - Is Empty() fn present in matlab was used to identify null values and row associated with it was removed

- Missing Value Elimination
  - Is NaN() fn present in matlab was used to identify missing values and row associated with it was removed

- Negative Value Elimination
  - Negative values cannot be part of the analysis being performed hence these values were identified and eliminated

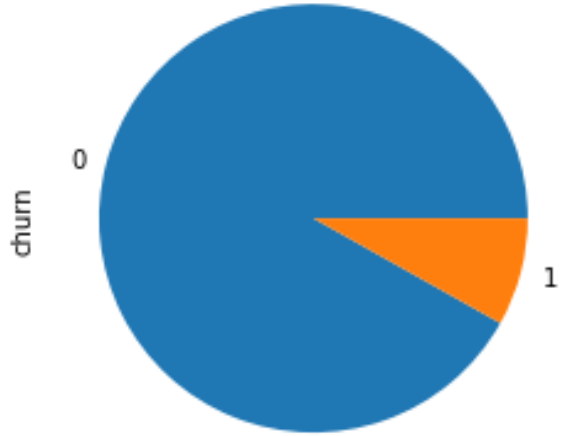
EDA – UNIVARIATE ANALYSIS

Insights From Univariate Analysis:

As we can see that 91% of the customers do not churn, there is a possibility of class imbalance.

Class imbalance has been handled later in the case study using the SMOTE method, post which the dataset can be used for further processing.

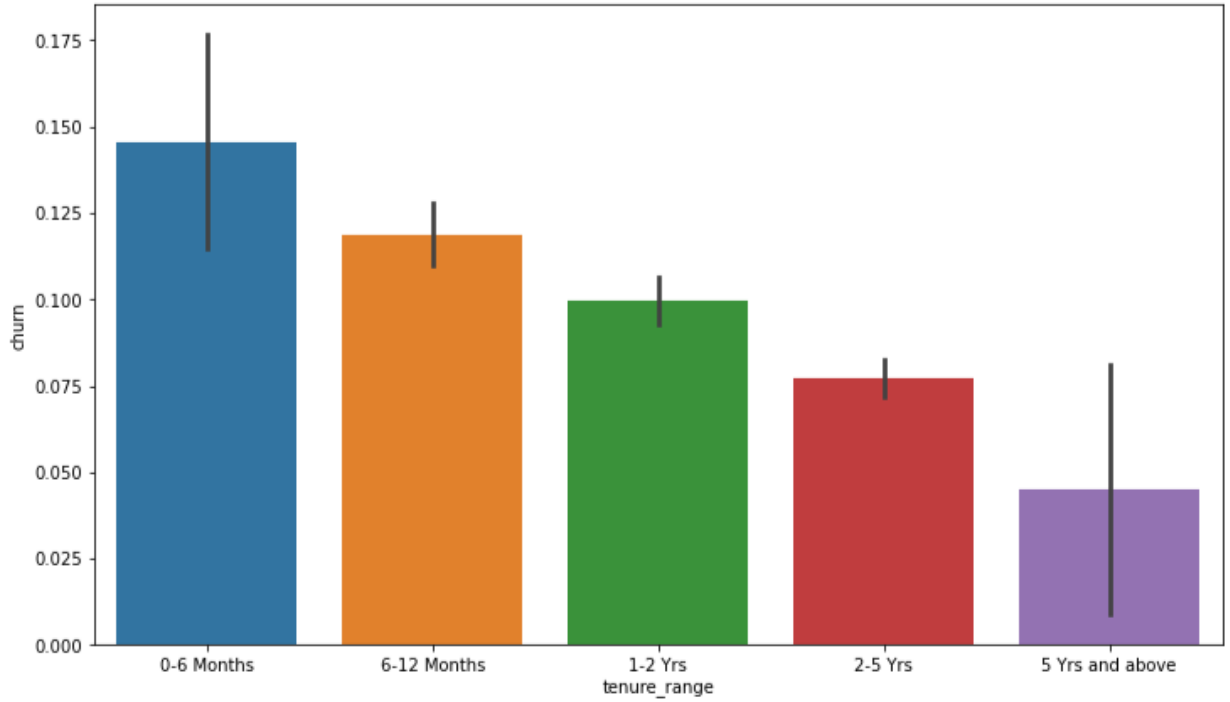
```
0    91.863605
1    8.136395
Name: churn, dtype: float64
```



EDA – BIVARIATE ANALYSIS

The figure on the right shows the churn rate based on the tenure of the customer.
It can be noticed that the maximum churn rate happens within 0-6 month, but it gradually decreases with an increase in customer retention in the network.

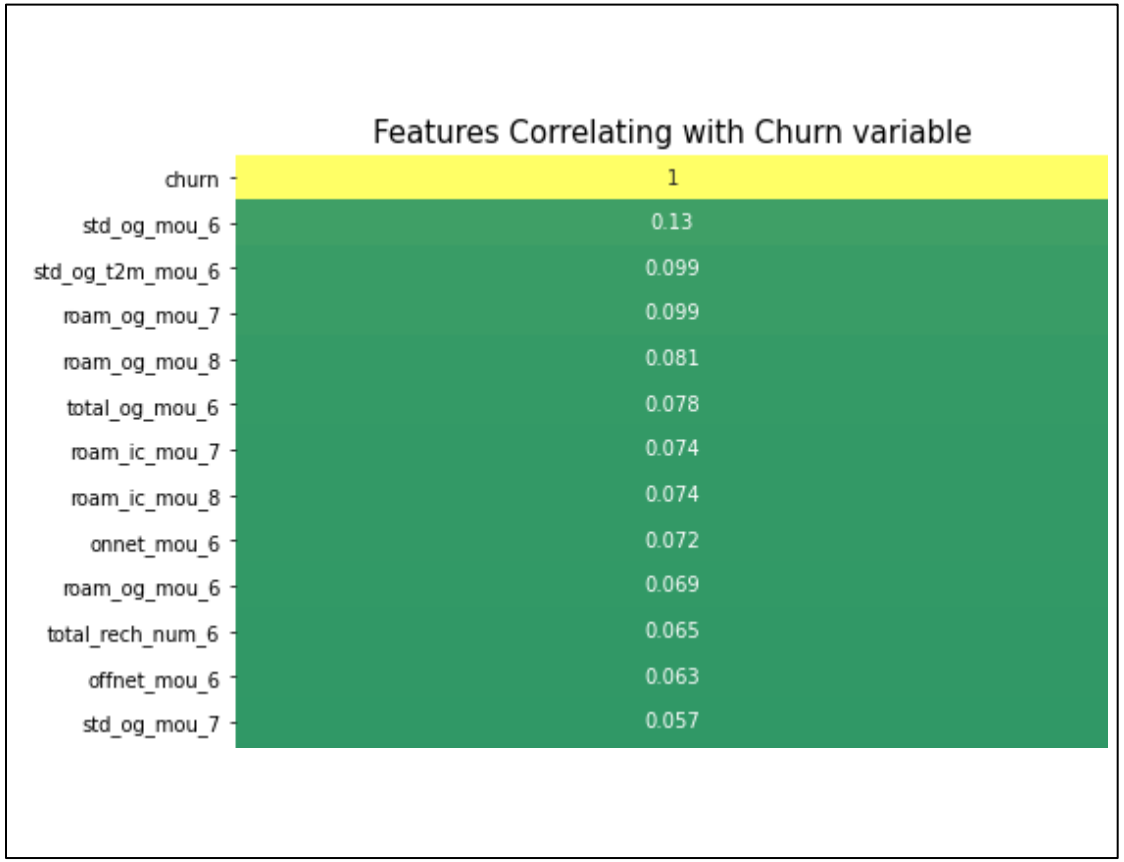
This shows that customer retention is an important factor in reducing churn.



CORRELATION TEST

- The heatmap as shown in the slide is gives an overview of the correlation of each attribute with the churn variable.
- It is important to note that the figure shown herein is only an extract of the actual figure as per the python notebook and hence does not give the complete picture.
- On thorough analysis of the entire heatmap as plotted, we have identified that:

  1. Avg Outgoing Calls & calls on roaming for 6 & 7th months are positively correlated with churn.
  2. Avg Revenue, No. Of Recharge for 8th month has negative correlation with churn.



MODEL BUILDING

- A logistic Regression Model has been used for predicting the categorical variable which is ‘churn’ in the case study.

- The process involved two stages for selecting appropriate features: RFE (Recursive Feature Elimination) with coarse tuning and manual fine-tuning using p-values and VIFs (Variance Inflation Factors).

- The steps undertaken in building the model are as follows:
  - Creation of Target Variable
  - Splitting the Dataset into train and test set
  - Scaling of Features
  - Correlation Checking
  - Feature Elimination based on Correlation Checking
  - Feature Selection Using RFE (Recursive Feature Elimination)

- Using statsmodel, a detailed model is built. The process is repeated 4 times and columns with high p-value above the accepted threshold of 0.05 p-value are dropped.

- Model 4 is ultimately the final model.

- Model 4 is stable and has significant p-values within the threshold (p-values less than 0.05) and thus can be used for further analysis.

FINAL MODEL

Generalized Linear Model Regression Results

Dep. Variable:	churn	**No. Observations:**	38576
Model:	GLM	**Df Residuals:**	38557
Model Family:	Binomial	**Df Model:**	18
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-15852.
Date:	Mon, 01 Mar 2021	**Deviance:**	31704.
Time:	15:41:29	**Pearson chi2:**	8.51e+10
No. Iterations:	7		
Covariance Type:	nonrobust		

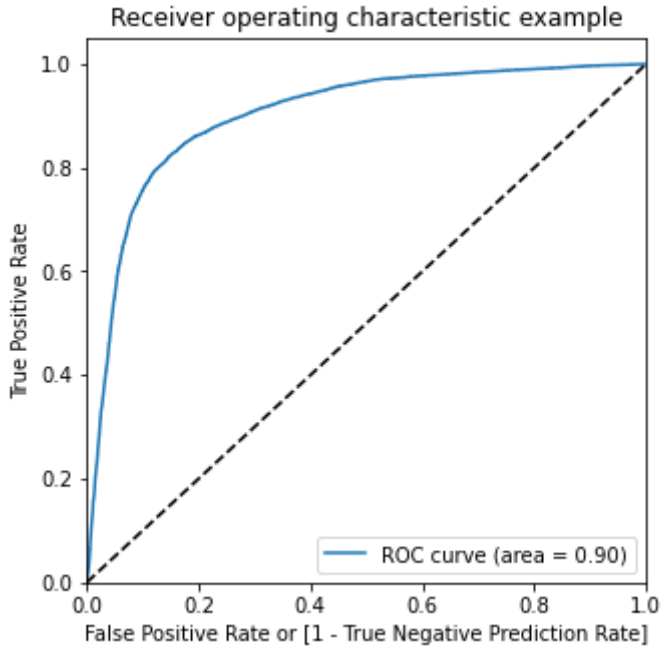
	Features	VIF
9	spl_ic_mou_8	83.90
7	loc_ic_mou_8	42.86
0	arpu_8	18.96
6	loc_ic_mou_6	18.68
5	total_og_mou_8	5.46
12	total_rech_data_8	3.58
4	std_og_mou_7	3.27
8	std_ic_mou_8	2.88
15	monthly_2g_8	2.76
3	loc_og_t2m_mou_8	2.54
14	vol_2g_mb_8	2.06
13	av_rech_amt_data_8	1.76
2	roam_og_mou_8	1.56
16	aug_vbc_3g	1.35
17	avg_arpu_6_7	1.33
1	roam_ic_mou_7	1.30
10	total_rech_num_8	1.15
11	last_day_rch_amt_8	1.05

MODEL EVALUATION

- Tools used in evaluating the model:
  - Confusion Matrix
  - Accuracy
  - Sensitivity and Specificity
  - Threshold determination using ROC & Finding Optimal cutoff point
  - Precision and Recall

- **An ROC curve demonstrates:**
  1. any increase in sensitivity will be accompanied by a decrease in specificity. It shows the tradeoff between sensitivity and specificity .
  2. The closer the curve follows the top-left hand border and then the top border of the ROC space, the more accurate the test.
  3. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

  (Area under ROC curve is 0.9 out of 1 which indicates a good predictive model)

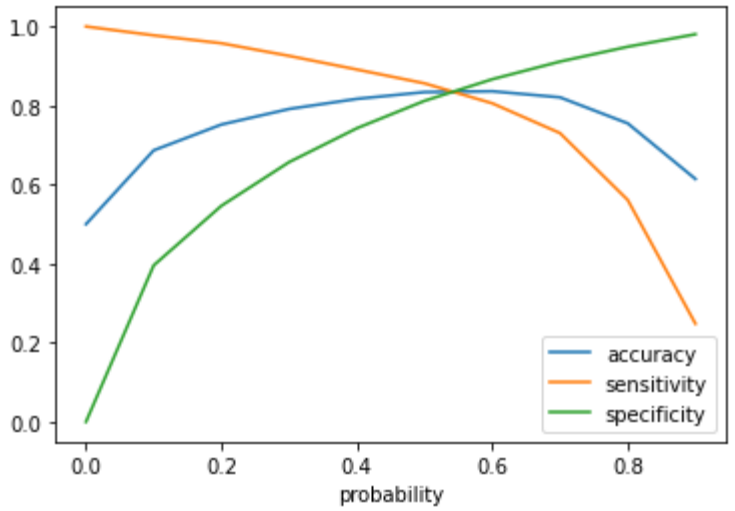


ROC Curve

MODEL EVALUATION

Optimal Cut-Off Point:

To determine the optimal cutoff point or probability, it is necessary to identify the threshold that achieves a balance between sensitivity and specificity.
From the graph shown it is understood that the point 0.5 (approx.) is the optimal cut-off value.



Optimal Cut-off Point

Given beneath the figure is the plotting of accuracy, sensitivity and specificity for various probabilities.

	probability	accuracy	sensitivity	specificity
0.50	0.50	0.834042	0.856128	0.811956
0.51	0.51	0.835001	0.851669	0.818333
0.52	0.52	0.835675	0.846796	0.824554
0.53	0.53	0.836038	0.841611	0.830465
0.54	0.54	0.836245	0.836375	0.836116
0.55	0.55	0.836064	0.830983	0.841145
0.56	0.56	0.837075	0.826991	0.847159
0.57	0.57	0.837179	0.821910	0.852447
0.58	0.58	0.836219	0.815896	0.856543
0.59	0.59	0.835831	0.810452	0.861209

CONCLUSIONS & RECOMMENDATIONS

Given below are a few recommendations which can be implemented in order to decrease the churn rate of the telecom service company:-

- Focus should be given to the features having positive as well as high coefficients to enhance target market strategies.
- Professionals can be engaged for customized services and engagement of customers to ensure high retention rates.
- Optimize communication channels to improve lead engagement impact.
- Commissions to sales force on sale of additional services to customers can improve efficiency significantly

THANK YOU