# *Model Selection Case Study*
# *Telecom Churn*

Shilpa Kulkarni
Shibsankar Saha
Shaik sohaib

# Problem Statement Objectives

- Analyse customer-level data of a leading telecom firm
- Build predictive models to identify customers at high risk of churn
- Identify the main indicators of churn
- To use usage-based definition to define churn

# Summary

- The dataset contained 99,999 rows and 226 columns
- The dataset was inspected for data types and Null values
- Few columns had very high % of Null values. Those columns were deleted since they will not help in the prediction
- For the remaining columns which had less than 10% of Null values , appropriate Null value imputation was done
- The high value customers were filtered (> 70 percentile of the average recharge over the first 2 months)
- The above resulted in ~30,000 rows were were used to build the prediction models
- The Churn customers were identified basis those who have not made any calls and have not used mobile internet even once in the churn phase i.e. month 9
- Around 91% of the customers turned out as Not Churned whereas only 9% resulted in Churned. This implies that this is an unbalanced data and would need special handling before we can begin with model building

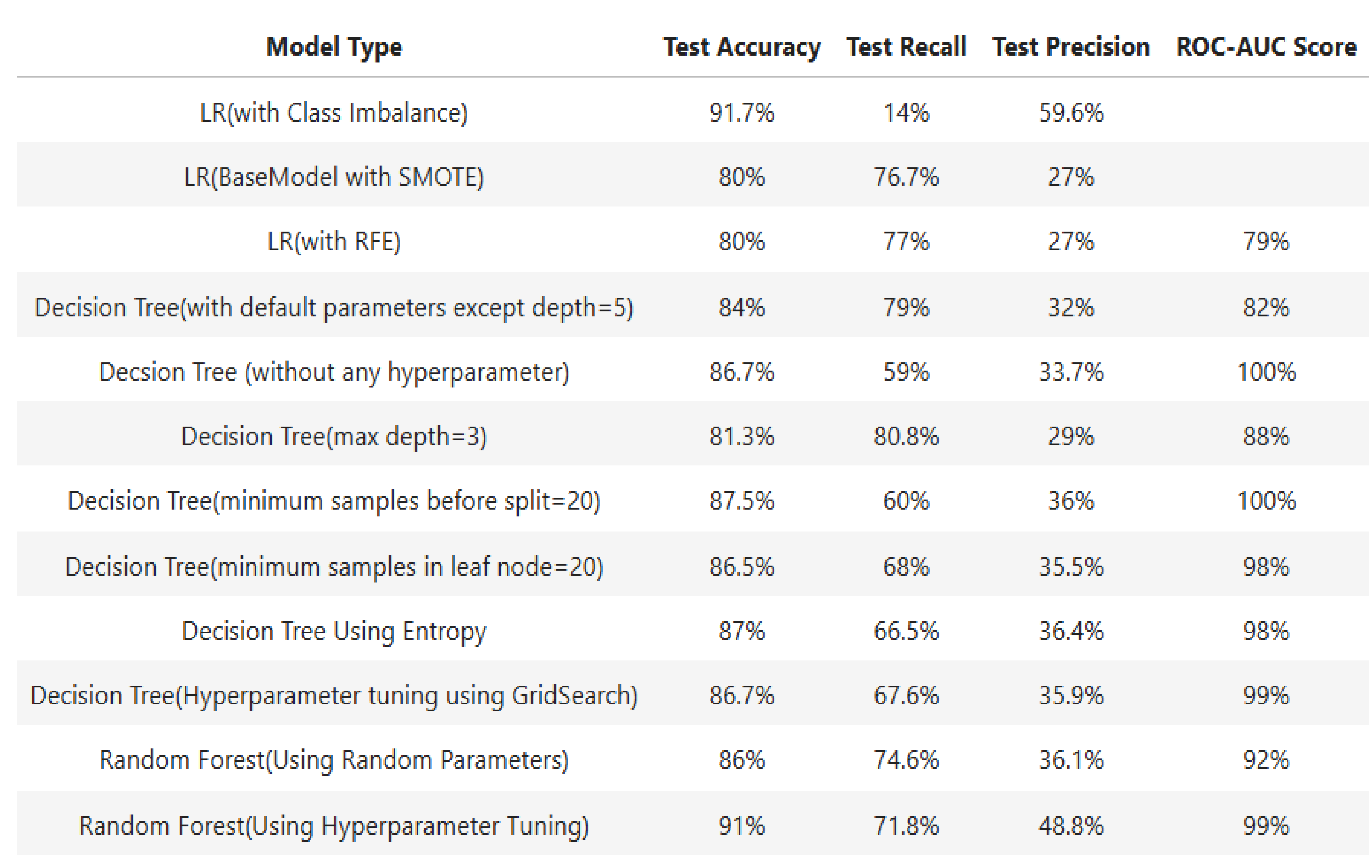
# Summary

- Since we want to make predictions in the Churn Phase i.e. month 9 , we deleted all the attributes corresponding to this month
- As part of feature engineering new columns were derived for the total of business specific columns namely – og , ic , recharge no , recharge amount to evaluate if these values have increased or decreased in the action phase (month 8) in comparison to the good phase (months 6 and 7)
- In the EDA, Univariate and Bivariate analysis was done to understand the distribution of each attribute. This led to the following observation
  - For the churn customers there is a significant drop in the below in the action phase just before they leave the operator
    - Recharge Numbers
    - Minutes of usage(MOU)
    - Average Revenue per user(ARPU)
  - ▪ There is a slight drop in the Volume Based Cost(VBC) in the action phase
  - ▪ The churned customers had less number of association year with the operator. So newer customers are churning more
- The heat map showed that correlated variables are present in the dataset

# Summary

- Before beginning with the actual model building
  - the data was scaled
  - The simple model built using Logistic Regression gave a Recall value of only 14%
  - SMOTE technique was applied to handle the Class Imbalance which increased the Recall value to 76%
- In the actual model building , various models were built using the below techniques to further increase the Recall value
  - Logistic Regression with RFE
  - Multiple models using Decision Tree
  - Multiple models using Random Forest

# Model Comparison



| Model Type | Test Accuracy | Test Recall | Test Precision | ROC-AUC Score |
|---|---|---|---|---|
| LR(with Class Imbalance) | 91.7% | 14% | 59.6% | |
| LR(BaseModel with SMOTE) | 80% | 76.7% | 27% | |
| LR(with RFE) | 80% | 77% | 27% | 79% |
| Decision Tree(with default parameters except depth=5) | 84% | 79% | 32% | 82% |
| Decsion Tree (without any hyperparameter) | 86.7% | 59% | 33.7% | 100% |
| Decision Tree(max depth=3) | 81.3% | 80.8% | 29% | 88% |
| Decision Tree(minimum samples before split=20) | 87.5% | 60% | 36% | 100% |
| Decision Tree(minimum samples in leaf node=20) | 86.5% | 68% | 35.5% | 98% |
| Decision Tree Using Entropy | 87% | 66.5% | 36.4% | 98% |
| Decision Tree(Hyperparameter tuning using GridSearch) | 86.7% | 67.6% | 35.9% | 99% |
| Random Forest(Using Random Parameters) | 86% | 74.6% | 36.1% | 92% |
| Random Forest(Using Hyperparameter Tuning) | 91% | 71.8% | 48.8% | 99% |

# Conclusion

- Among all the models, Decision tree with max depth of 3 provides the best recall rates (80.8%). Since we are interested in increasing the True Positives and reducing the False Negatives, Recall is the metric we are focused on
- Random Forests score very high on the accuracy and ROC but not very good on the Recall metrics when compared to the above Decision Tree
- The Logistic Regression Model with RFE scored better than Random Forests on the Recall parameter
- Hence after Decision Tree the preferred model for this problem statement is Logistic Regression using RFE
- The ROC score for all the models have shown a pretty high value > 79%

# Business Recommendations

## Important predictors to identify customer Churn

We used the output of the Decision Tree with max Depth 3 as our final model to identify important predictors

1.A sharp decline (measured by usage in current month minus average usage in prior 2 months) in the total recharge amount of a customer (total_rech_amt) has a large contribution towards the odds of the customer churning next month.

2.A considerable decline (measured by current month outgoing call in roaming zone minus average of prior 2 months) in the outgoing calls in roaming zone (roam_og_mou) is the second biggest contributor to the odds of the customer churning.

3. A decline in the incoming calls (measured by current month incoming call minus average of prior 2 months) (loc_ic_t2m_mou,loc_ic_t2t_mou,roam_ic_mou) are the next contibutors to the odds of the customer churning.

4. Age on network (aon) is the next biggest contributer to the odds of the customer churning. This means that newer customers have a higher probability to churn than the old customers

# Business Recommendations

## Strategies to manage customer churn

**1.Target identified probable churn customers in the action month so that they don't churn**

a. Offer targeted incentives to these customers to continue with the operator. Since the decline in recharge is one of the indicators of churn, these customers could be offered recharge related incentives

b. Also they can be offered call related (incoming/outgoing) incentives since those are the next important predictor parameters

c. Since the newer customers have a higher probability to churn, target them with monthly recharges or call related deals so that they stick for a longer period

**2.Continuous monitoring and model updates**

a. Continuously monitor the outcome of the incentives on the churn behavior and input the same into the model. If needed re-tune the hyper parameters to make better predictions

b. In case of any significant event impacting churn, redevelop the model factoring the parameters describing the significant event.

# THANK YOU