

# Telecom Churn- Domain Oriented Case Study

Group Members- Anjali Mishra

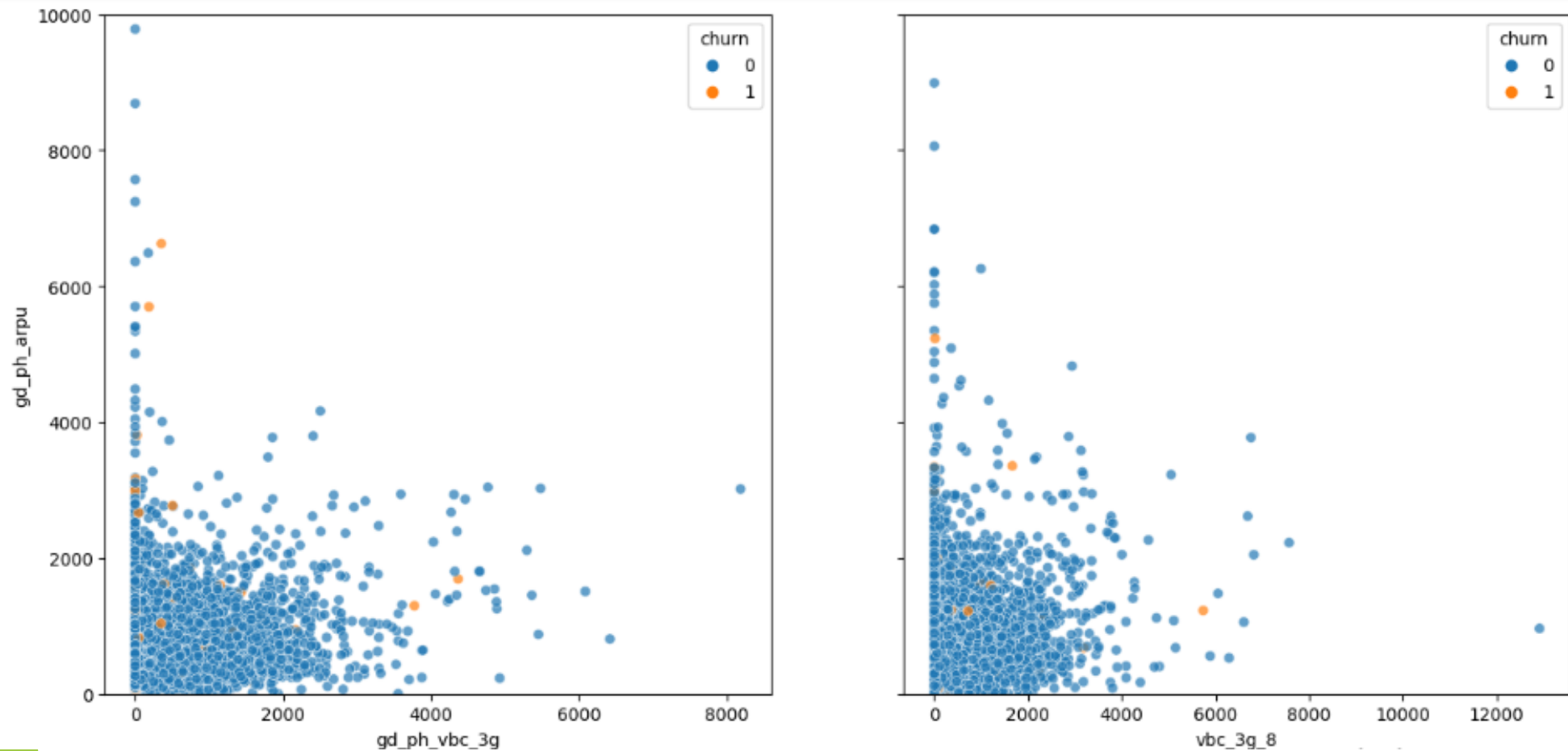
Dhruva E Seshasayee

Kyatham Nikhil

# Problem Statement

- ▶ To analyze customer level data of a leading telecom firm and to build predictive models to identify customers at a high risk of churn
- ▶ To identify the main indicators of churn.

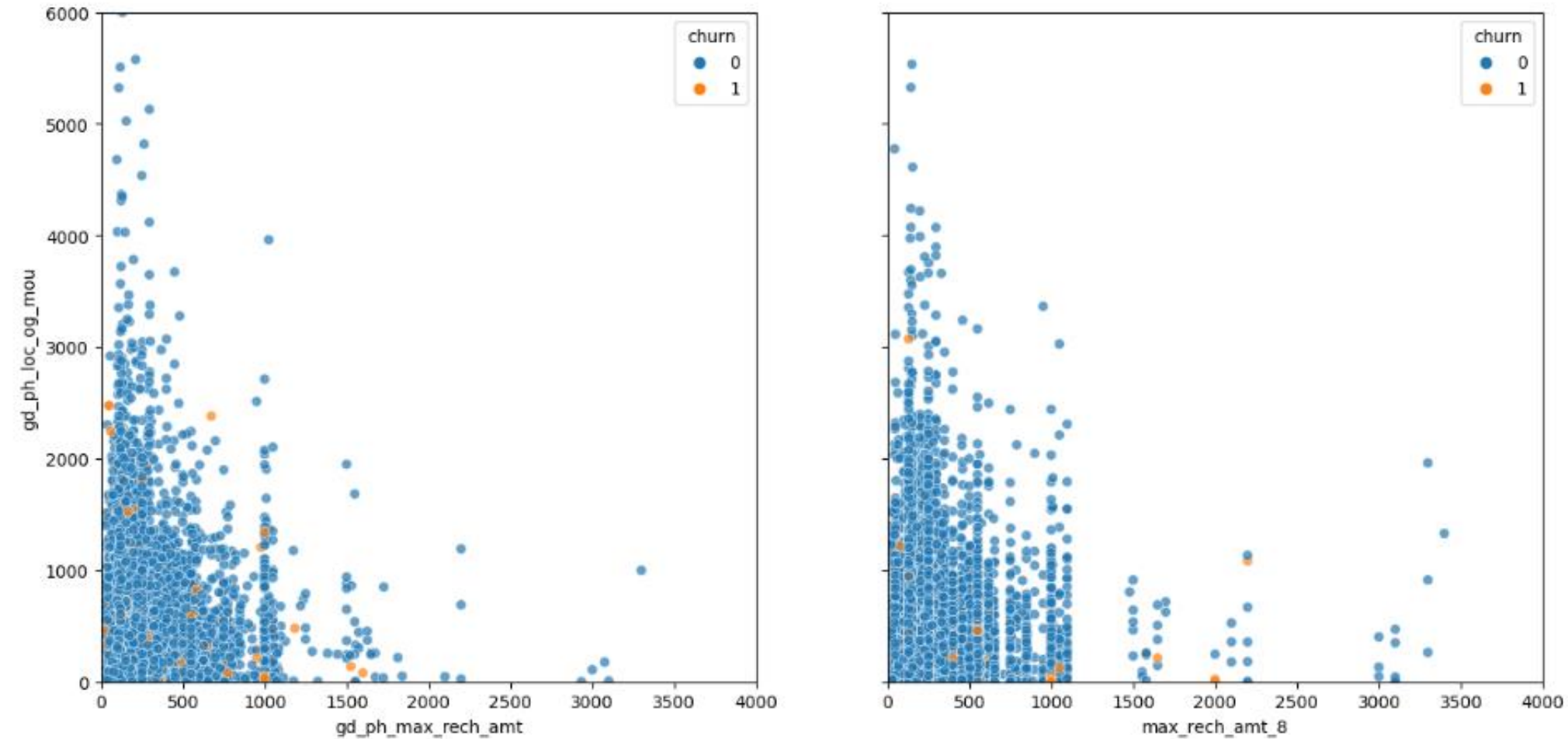
## Checking how total Minutes of Usage(MoU) affects revenue



### Observation

- We can see that the users who were using very less amount of VBC data and yet were generating high revenue churned
- Yet again we see that the revenue is higher towards the lesser consumption side

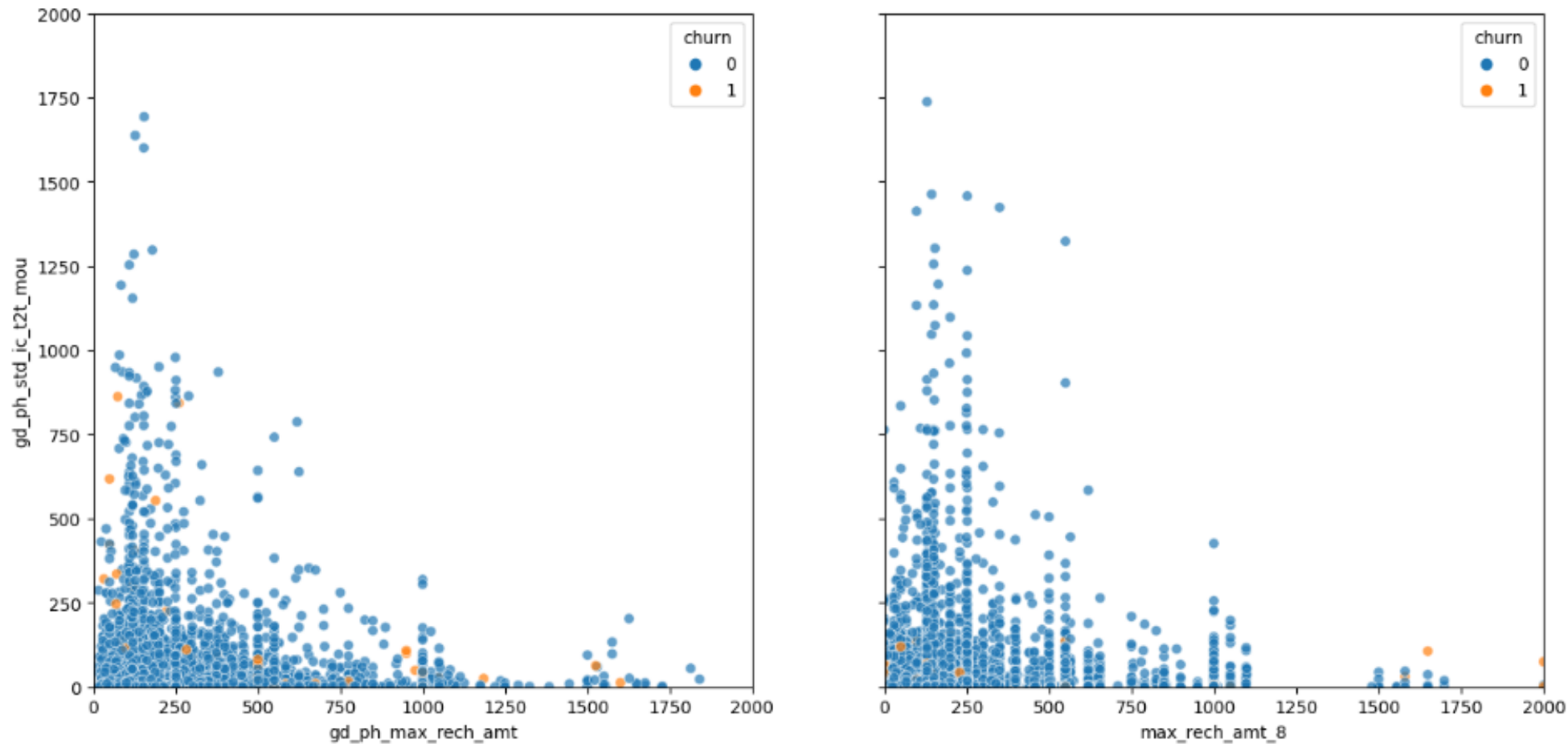
# Relation between recharge amount and local outgoing calls



## Observations

- Users who were recharging with high amounts were using the service for local uses less as compared to user who did lesser amounts of recharge
- People whose max recharge amount as well as local out going were very less even in the good phase churned more

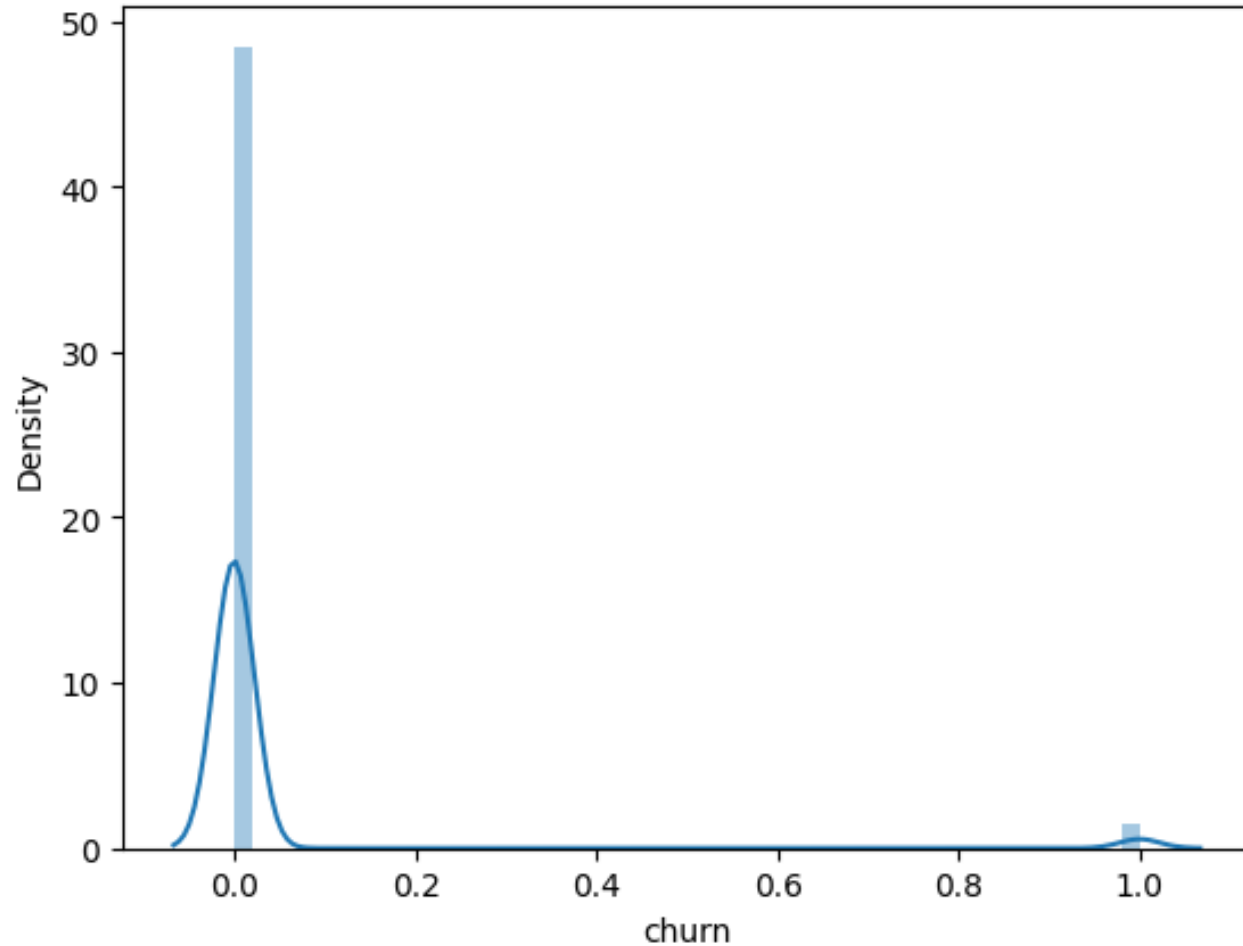
## Incoming from the same service provider vs the recharge amount



### Observation

- Users who have max recharge amount on the higher end and still have low incoming call MoU during the good phase, churned out more

# Distribution of target variable



## Observation

- Though the variable is not skewed it is highly imbalanced, the number of non-churners in the dataset is around 94%
- We will handle this imbalance using SMOTE algorithm

# Achieved parameters after analysis

- ▶ Using Logistic regression we are getting an accuracy of 78.5% on train data and 78.8% on test data
- ▶ We are getting an accuracy of 90% on test data, with decision tree
- ▶ We are getting an accuracy of 95% on test data, with Random forest

# Conclusion

- ▶ Given our business problem, to retain their customers, we need higher recall. As giving an offer to an user not going to churn will cost less as compared to loosing a customer and bring new customer, we need to have high rate of correctly identifying the true positives, hence recall.
- ▶ When we compare the models trained we can see the tuned random forest and Ada boost are performing the best, which is highest accuracy along with highest recall i.e. 95% and 97% respectively. So, we will go with random forest instead of Ada boost as that is a simpler model.