Domain Specific Telecom Churn Case Study

Submitted by

BHAGYASHREE SHARMA

Business Statement

- In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate. Given the fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, **customer retention** has now become even more important than customer acquisition.
- For many incumbent operators, *retaining high profitable customers is the number one business goal*.
- To reduce customer churn, telecom companies need to **predict which customers are at high risk of churn.**

Business Understanding

- 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.

Business Objective

- 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 behavior during churn will be helpful.

Data Preparation

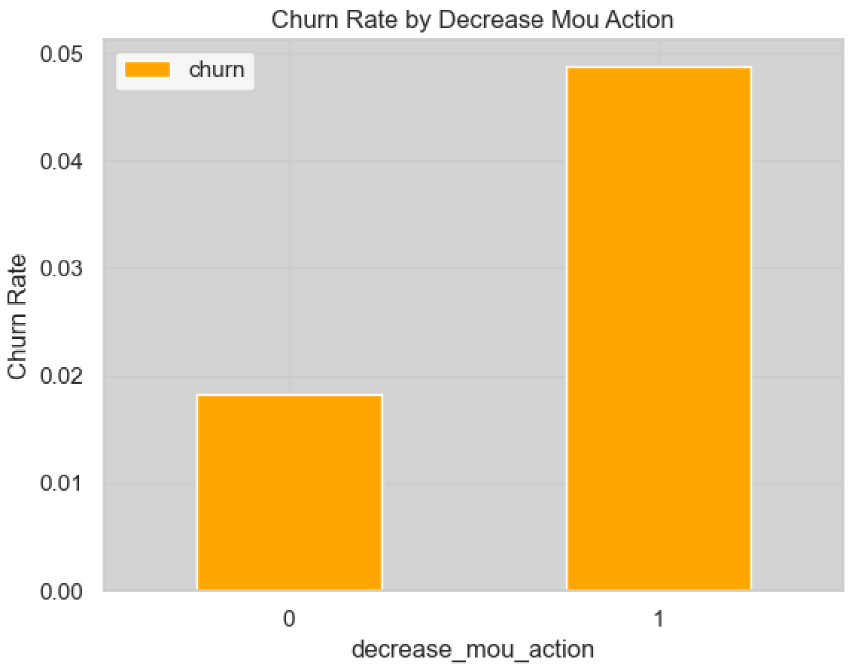
1. **Derive new features-** This is one of the most important parts of data preparation since good features are often the differentiators between good and bad models. Use your business understanding to derive features you think could be important indicators of churn.
2. **Filter high-value customers-** As mentioned above, you need to predict churn only for the high-value customers. 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).
3. **Tag churners and remove attributes of the churn phase-** Now tag 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 you need to use to tag churners are:
    1. total_ic_mou_9
    2. total_og_mou_9
    3. vol_2g_mb_9
    4. vol_3g_mb_9

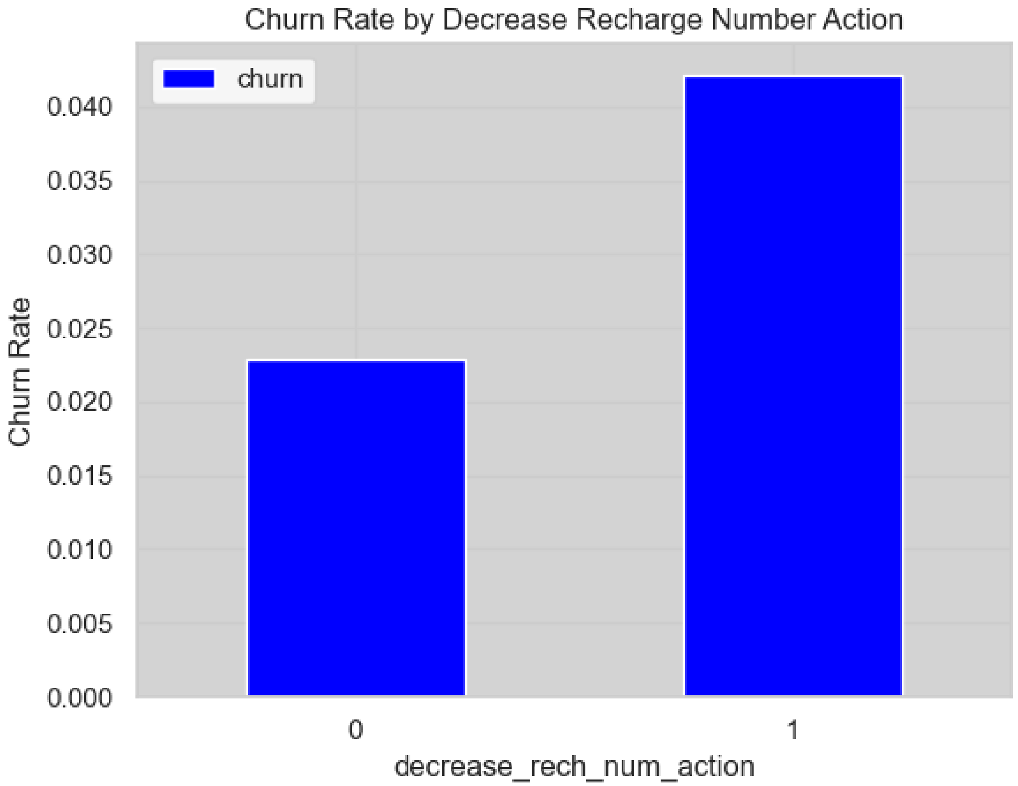
- After tagging churners, remove all the attributes corresponding to the churn phase (all attributes having ‘_9’, etc. in their names).

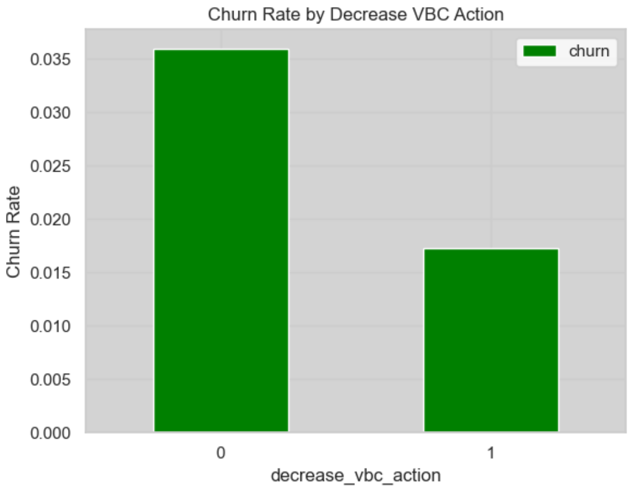
Steps Followed:

1. Reading & Understanding Data
2. Data Cleaning
3. EDA & Visualization
4. Train & Test Split
5. Model With PCA
6. Data Preparation
7. Conclusion

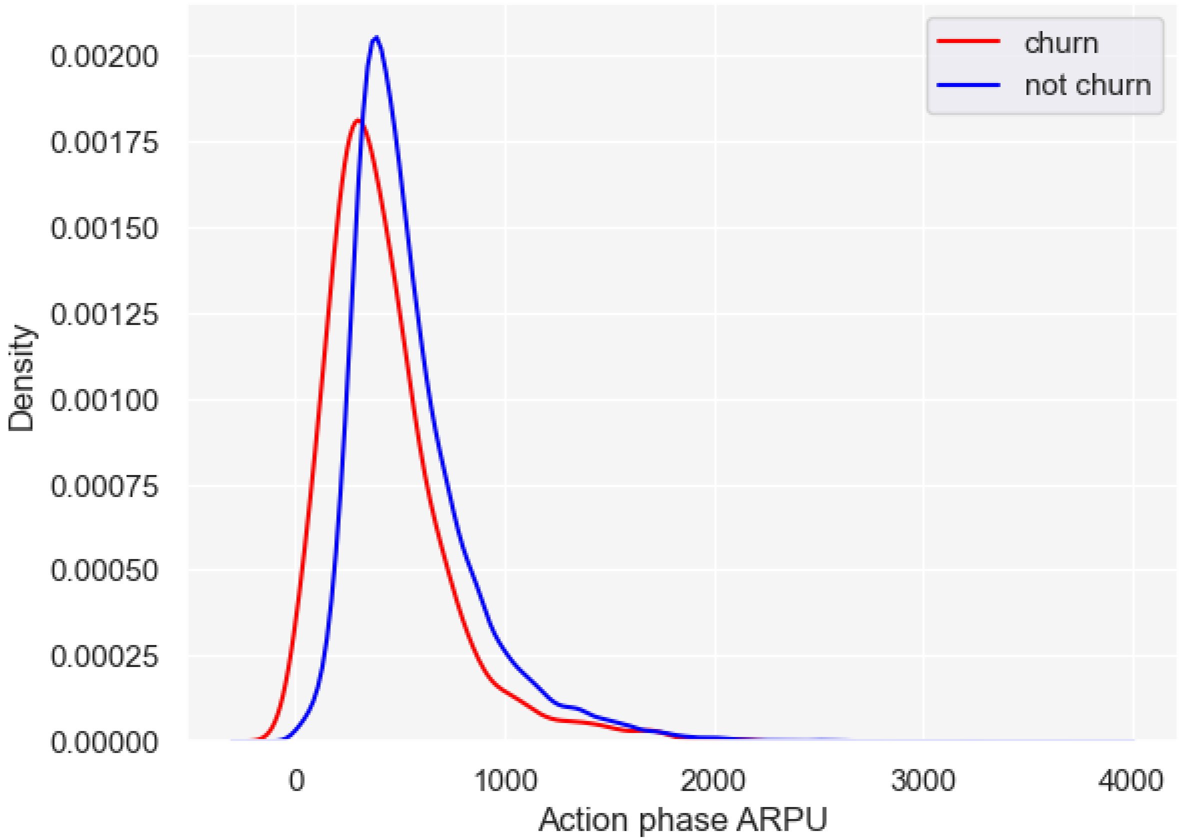
EDA- Univariate Analysis

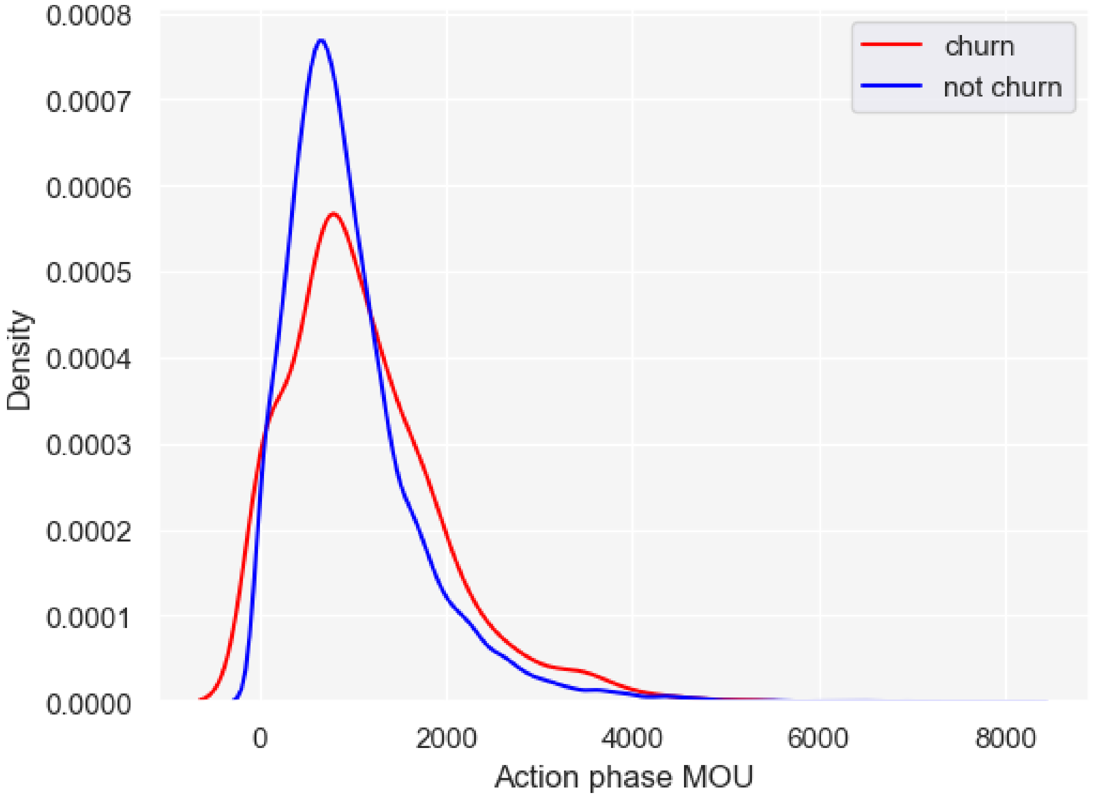






EDA- Univariate Analysis

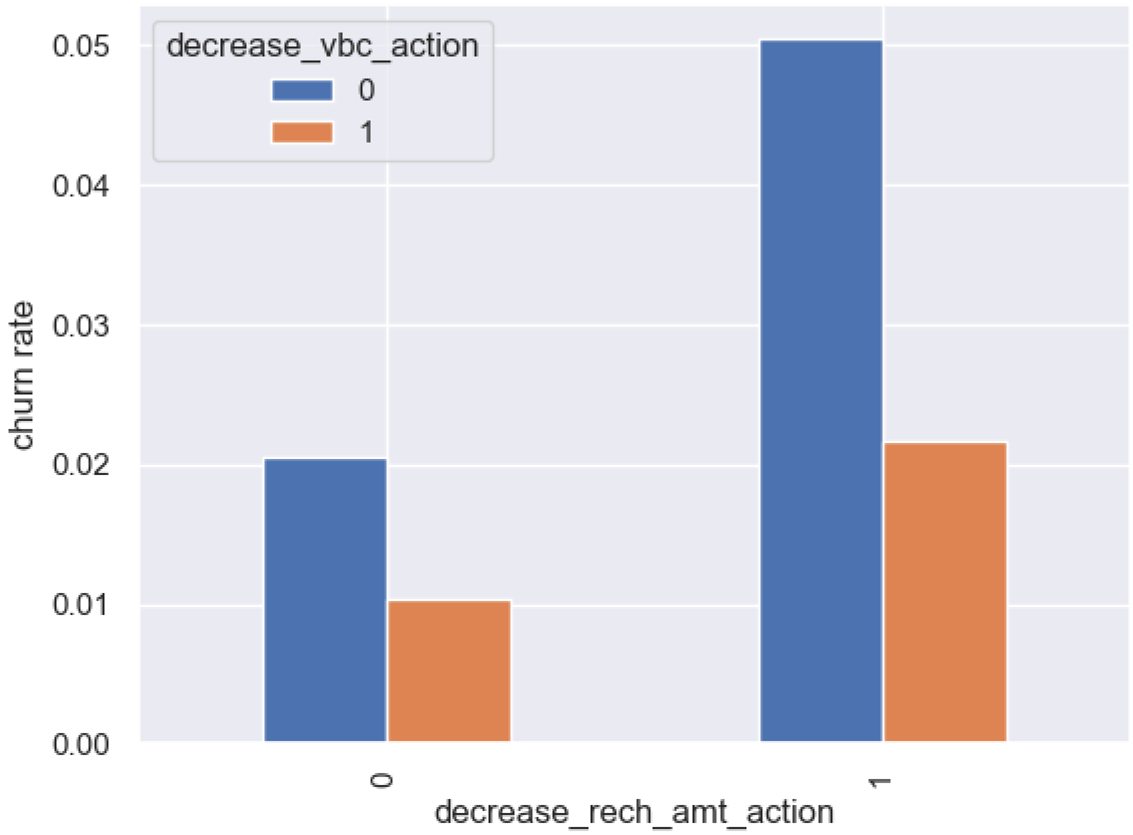


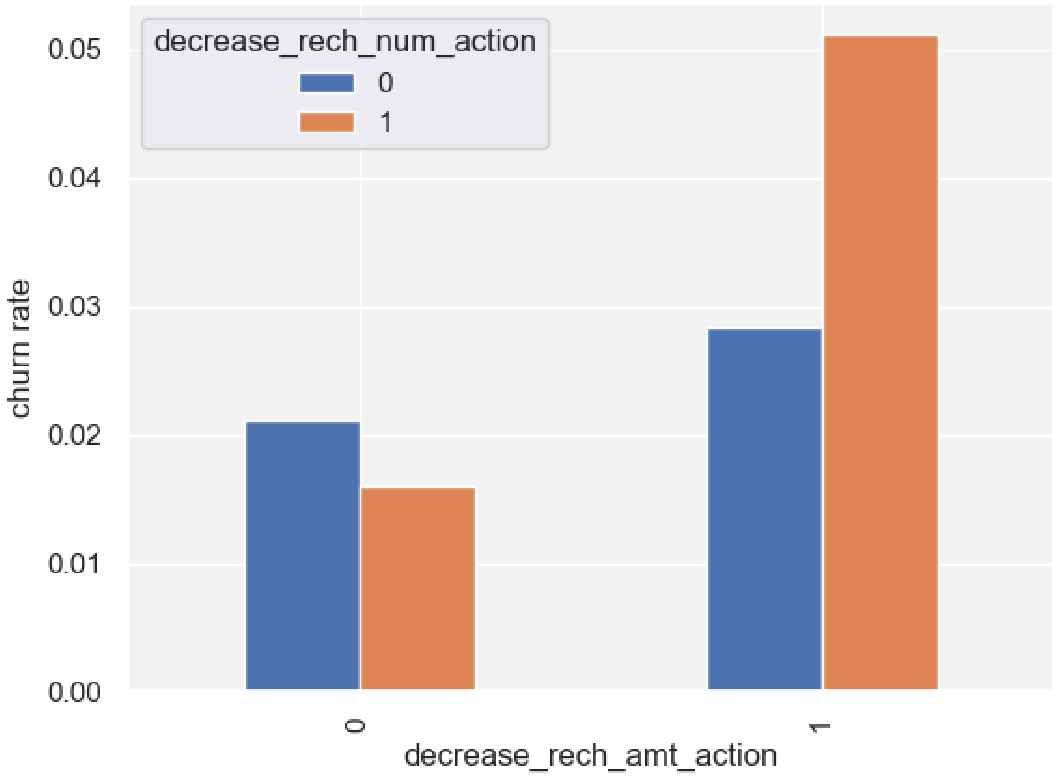


* The higher the ARPU(Avg Revenue Per User) the less likely to be churned.
* ARPU with less churned are in between 0 to 900.
* Maximum number of unchurned customers are when their ARPU is nearly 500.

•MOU(Minutes Of Usage) of the churn customer is in between 0 to 2000.
•Higher the MOU lesser the Churn.

EDA- Bivariate Analysis

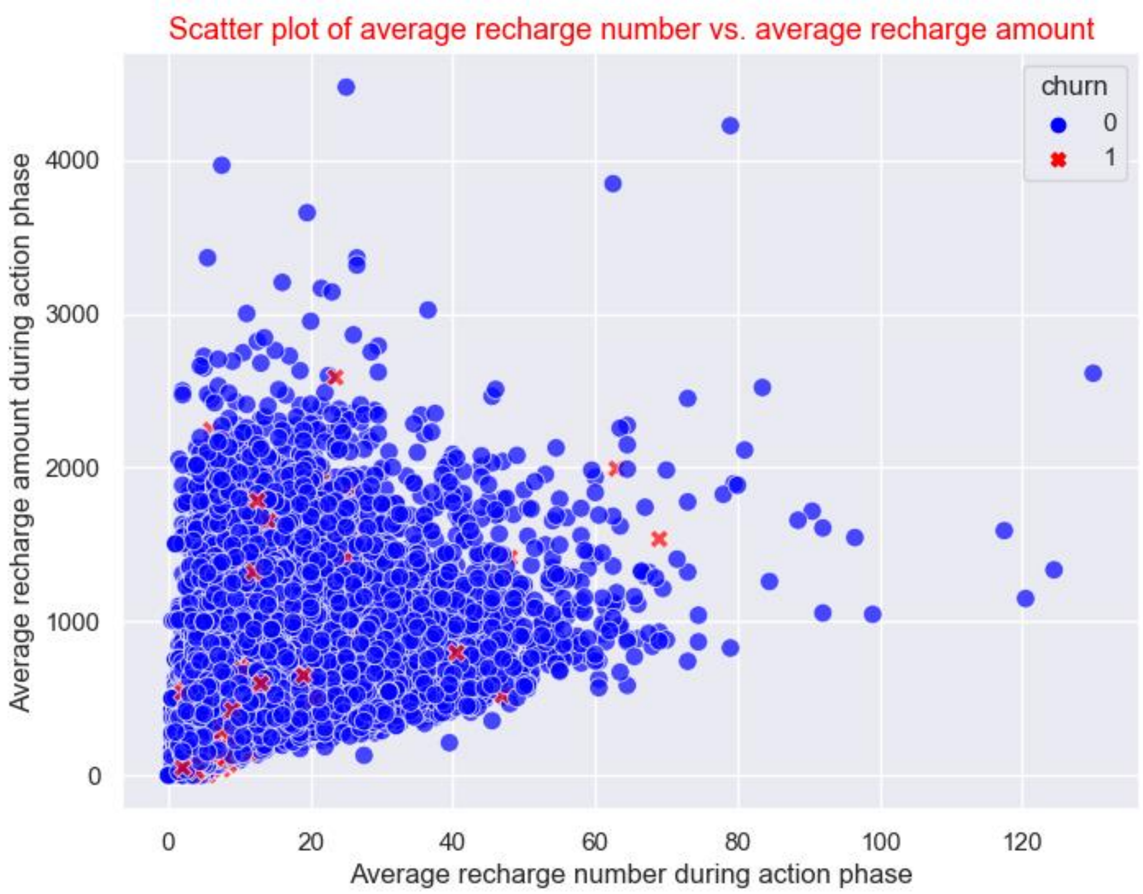




The analysis shows that customers who experienced a decrease in recharge amount along with an increase in volume-based cost during the action month have a higher churn rate.

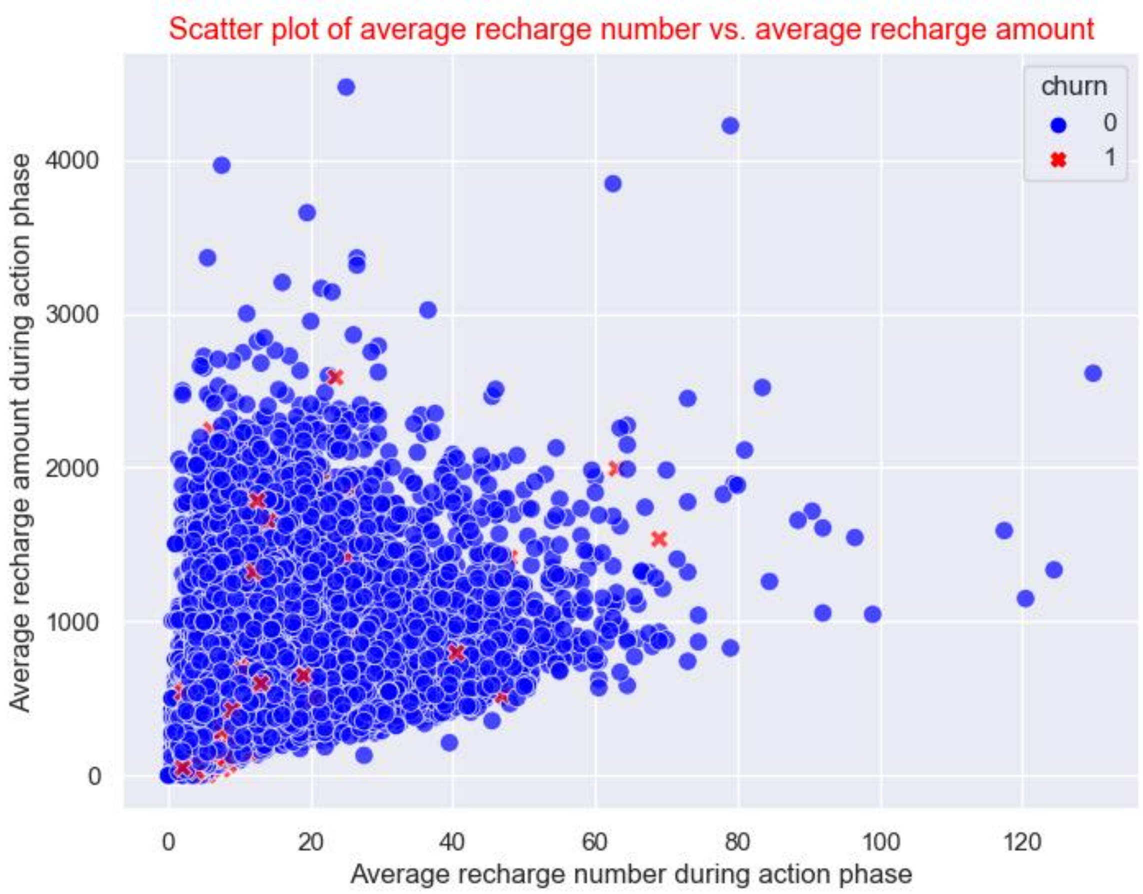
•The plot indicates that customers who have experienced a decrease in both the amount and number of recharge during the action phase have a higher churn rate compared to those who did not experience a decrease during this phase.

EDA- Bivariate Analysis



•The pattern observed in the above analysis suggests that there is a positive correlation between recharge number and recharge amount. In other words, as the number of recharges increases, so does the total amount of recharge.

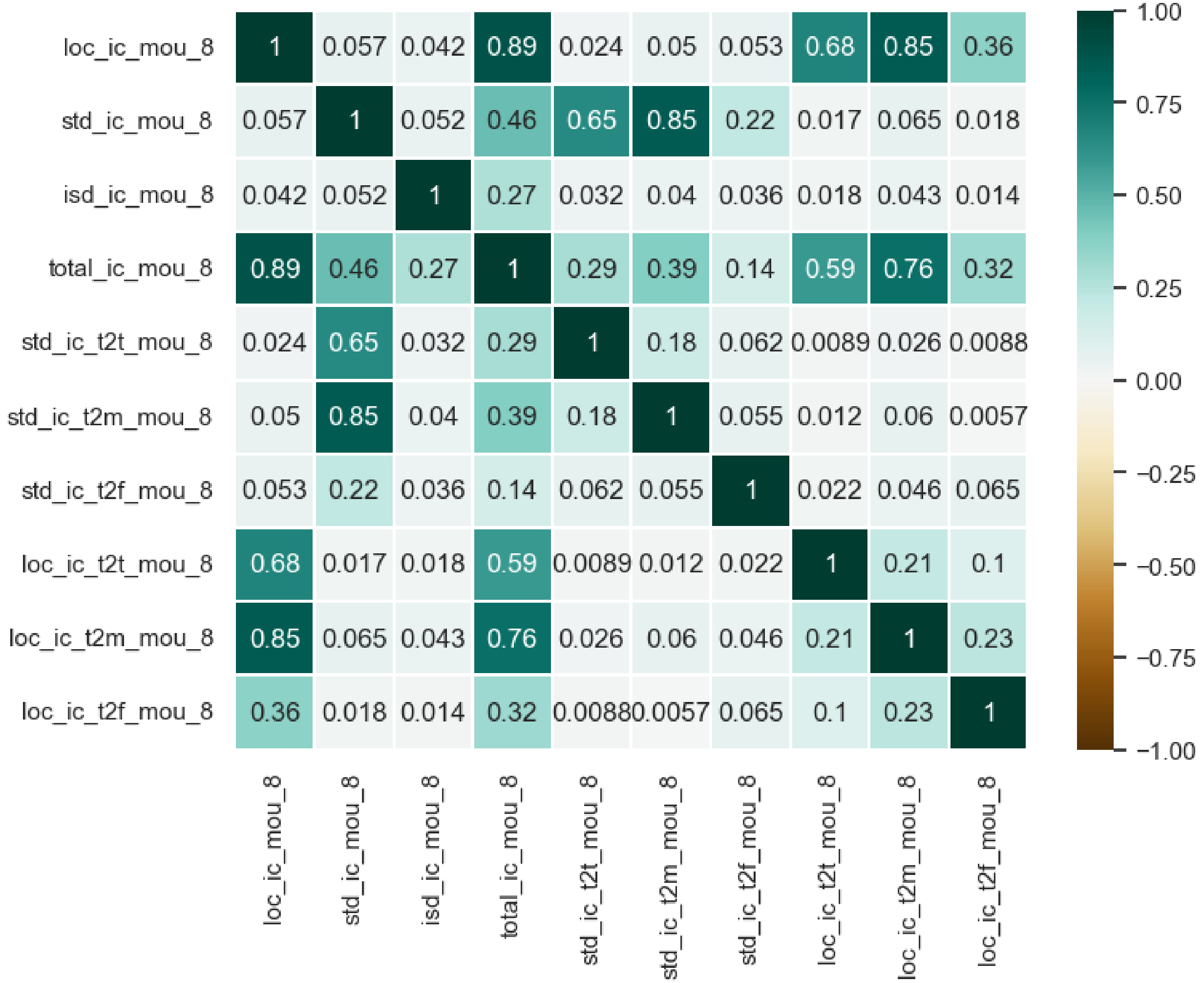
EDA- Bivariate Analysis



•The pattern observed in the above analysis suggests that there is a positive correlation between recharge number and recharge amount. In other words, as the number of recharges increases, so does the total amount of recharge.

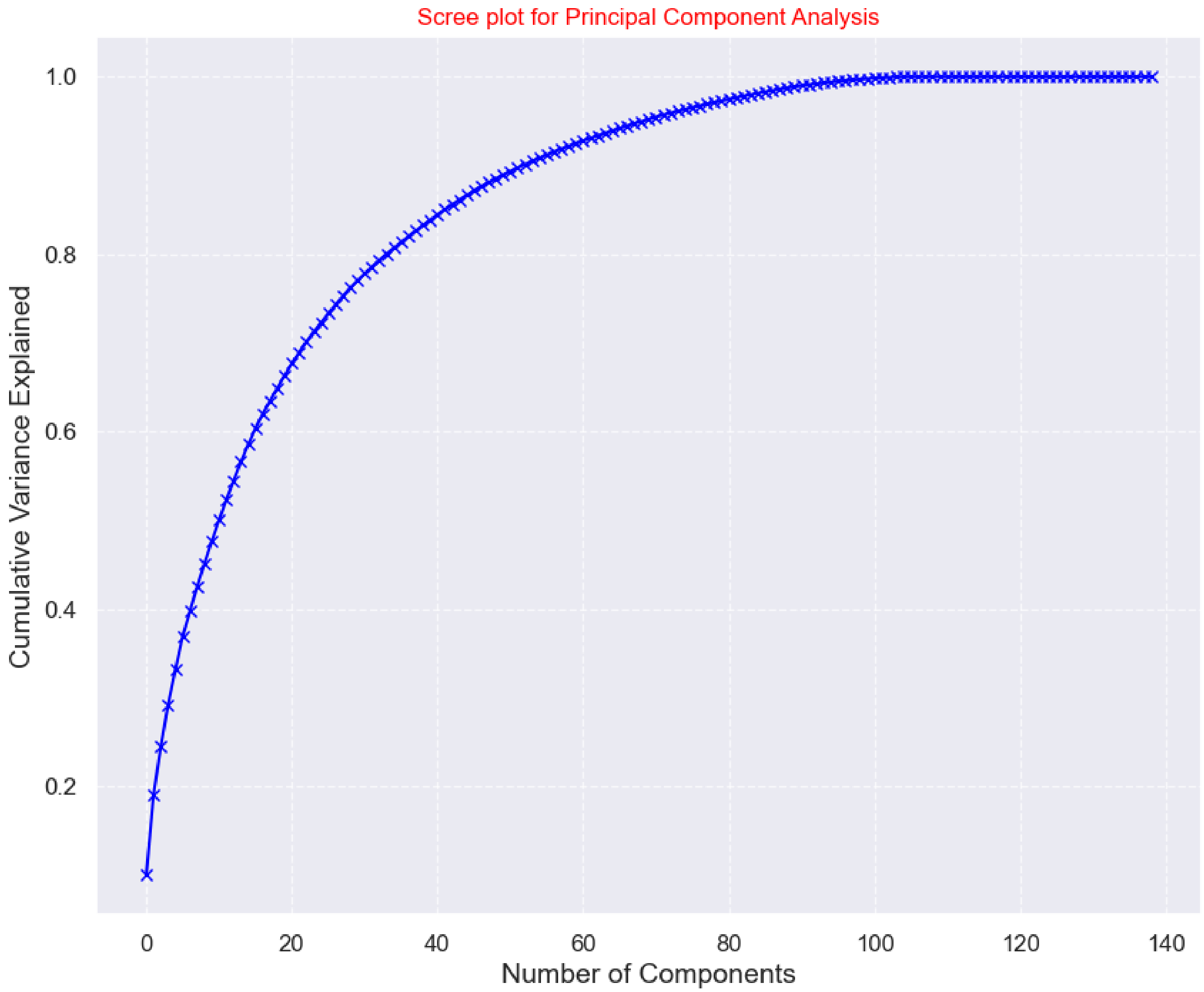
Categorical Variables

We observe that loc_ic_mou_8, std_ic_mou_8 and total_ic_mou_8 have strong correlation with other features. These features should be taken care of to handle multicollinearity.



Model Selection

- Logistic Regression
- SVM
- Random Forest
- Decision Tree



FEATURE ENGINEERING

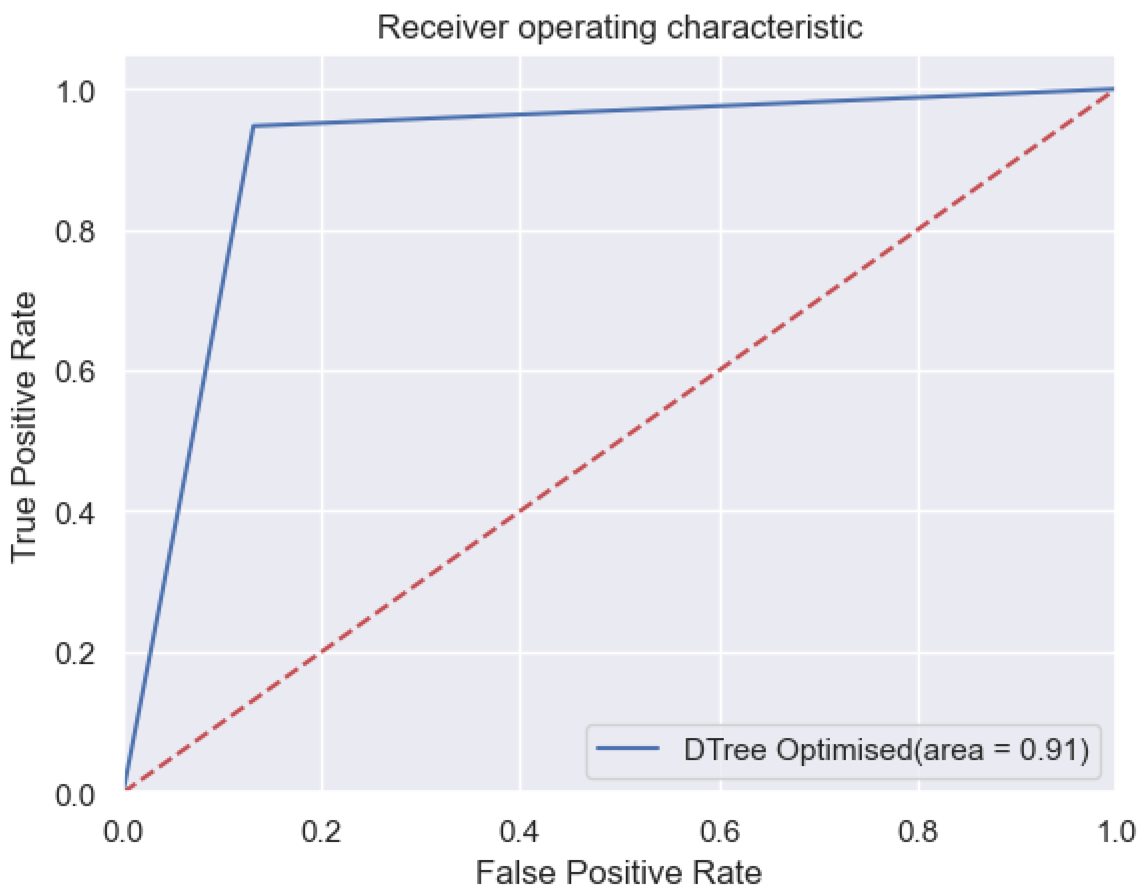
- **Logistic Regression:** - Sensitivity/Recall:0.89, - Specificity:0.83, - ROC AUC Score:0.86
- **SVM**: - Sensitivity/Recall:0.83, - Specificity:0.83, - ROC AUC Score:0.90
- **Random Forest**: - Sensitivity/Recall:0.65, - Specificity:0.88, - ROC AUC Score:0.89

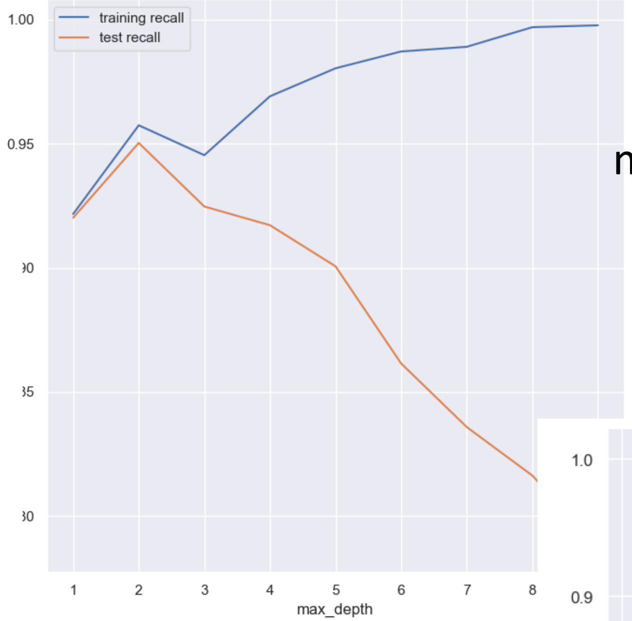
- **Observation**

- We choose logistic regression from all the above models as it has less time complexity and take less memory compared to all above models. Moreover, its sensitivity is very good which is our prime requirement in this case study.

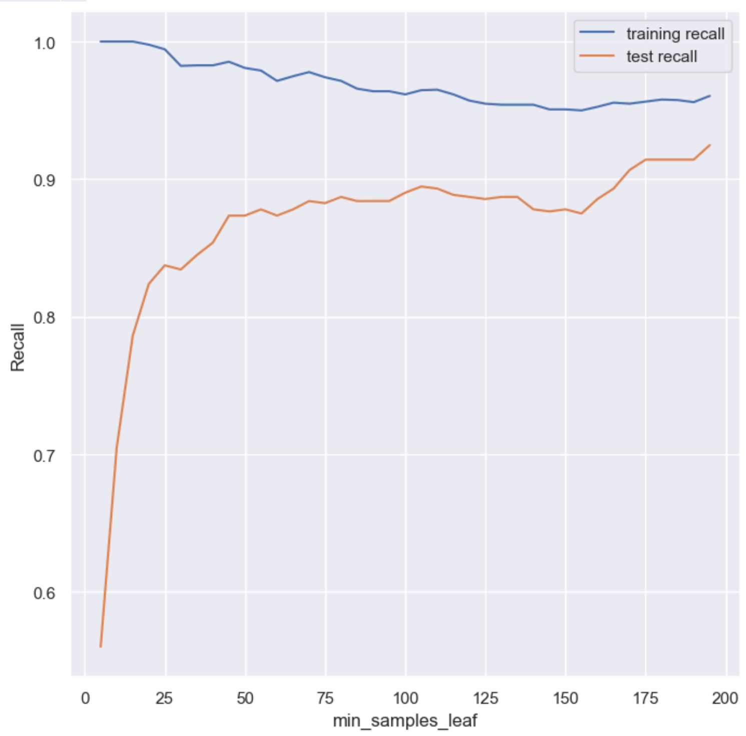
Model Building

ROC CURVE



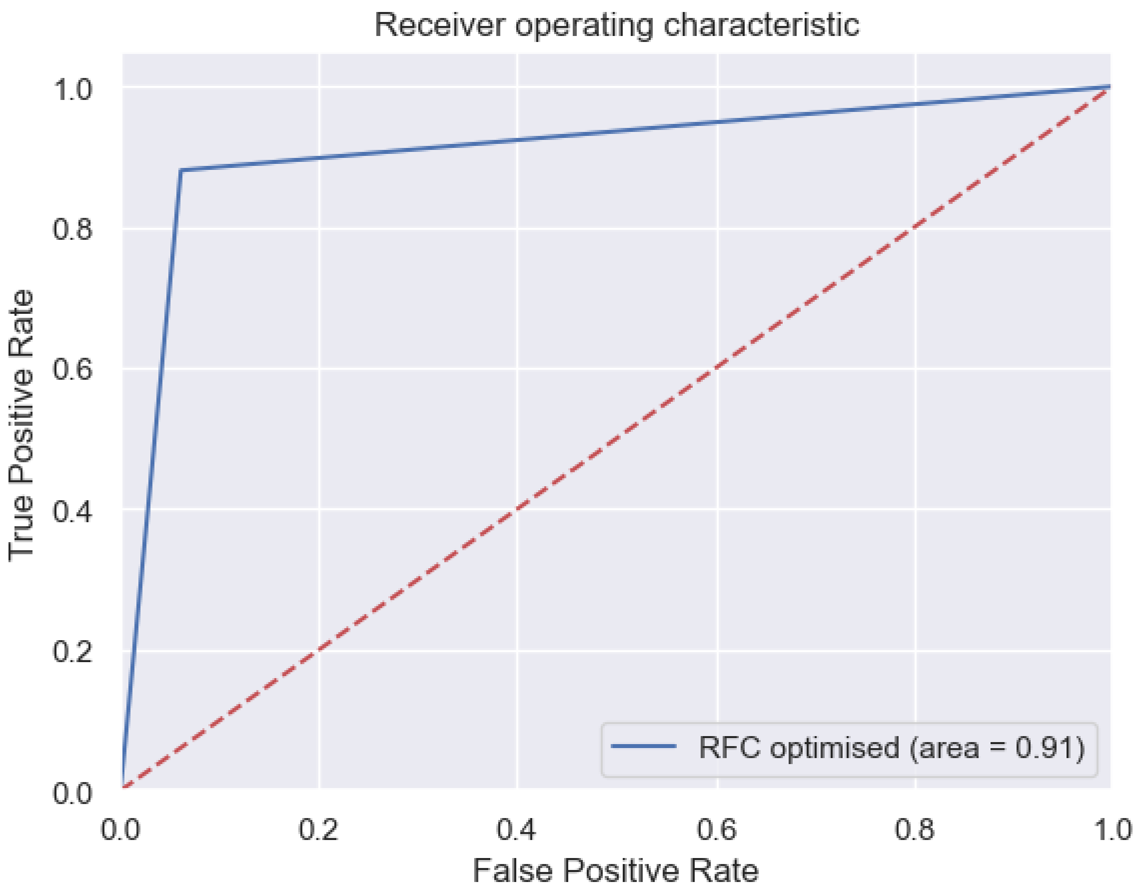


min_samples_leaf

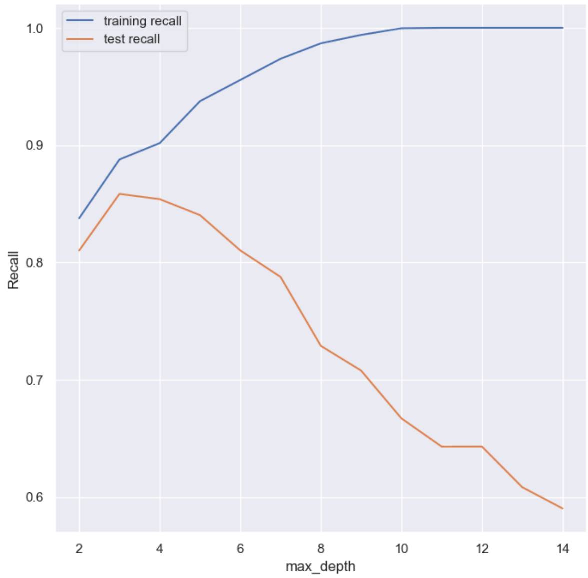


Model 2 Random Forest

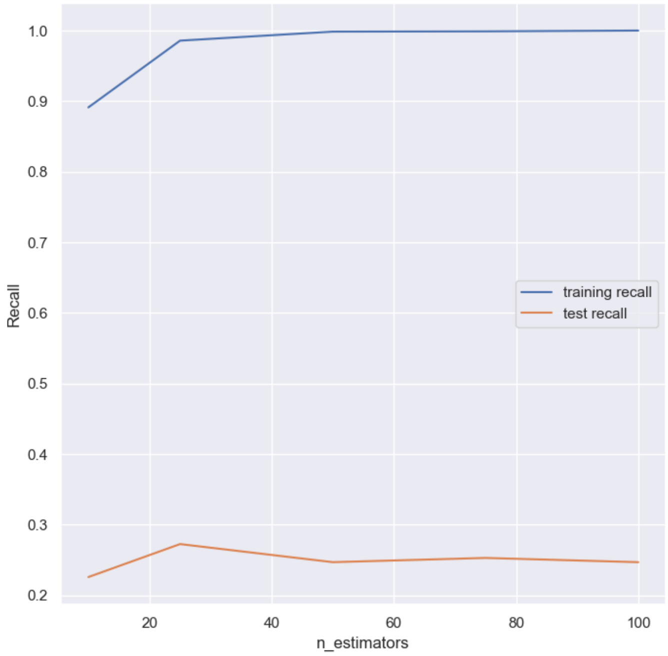
RFC



Max depth



N estimators



Model Summary

- Decision tree with tuned hyperparameters outperforms all the other models in terms of recall and has a pretty decent accuracy and AUC score. Let us choose this model to find out the most important features affecting churn.

	model	Recall	Test accuracy	Roc_auc_score
1	DTree - Tuned Param	0.84	0.82	0.83
3	RFC - Tuned Param	0.78	0.85	0.85
0	DTree - Default Param	0.48	0.92	0.71
2	RFC - Default Param	0.43	0.94	0.71

Conclusion

- The telecom industry experiences an annual churn rate of 15-25%, making customer retention more important than customer acquisition due to the high cost of acquiring new customers. To manage High Value Customer Churn, we predicted customers likely to churn and identified factors influencing high churn.
- A considerable drop in recharge, call usage, and data usage in the 8th month (Action Phase) was observed during exploratory analysis.
- Important predictors affecting churn include 'arpu_7', 'max_rech_amt_6', 'std_og_t2m_mou_8', 'loc_og_t2m_mou_8', 'max_rech_data_8', 'last_day_rch_amt_8', 'total_data_rech_8', 'total_amt_8', 'roam_og_mou_8', 'loc_ic_t2m_mou_8'.
- The average revenue per user in the 7th month plays a vital role in predicting churn.
- Local and STD minutes of usage (incoming and outgoing) are the most influential features on customer churn.
- The last day of recharge amount in the action phase and the maximum recharge for calling data in the 6th and 8th months should be focused on to prevent churn.
- The last day of recharge, total recharge for data done, and the total amount spent on calls and data in the 8th month also play a crucial role in indicating churn.
- Outgoing roaming calls made by clients in the 8th month also play a key role in predicting churn.
- Strategies to prevent churn include improving network and customer satisfaction, providing customized plans, routine feedback calls, introducing attractive offers, and promotional offers.