Telecom Churn Case Study

(Domain Oriented Case Study)

Submitted by:

Pulaparthi Siddhartha Rachit

Problem Statement

- The telecom industry experiences approx 15-25% annual churn rate
- Since new customer acquisition costs 5-10 times than retaining an existing one, customer retention is very important
- Retaining highly profitable customers is even more important
- To reduce churn telecom companies need to predict which all customers are likely to churn in future.
- In this case study our objective is to build predictive models to identify customers at high risk of churn and also to identify main indicators of churn.

Overall Approach

- Usage based definition will be used to define churn for this study

- The dataset contains data for month 6,7,8 &9. First two represent ‘good phase’ wherein customers are happy. Third month is the ‘action’ phase wherein customer experience starts to sore and the fourth month is the ‘churn’ phase.

- Considering the above, import relevant python libraries and data files.

- Filter high value customers by defining high value customers as those who have recharged with an amount more or equal to the amount which is 70th percentile of average recharge amount in the first two months.

- Tag all churners and remove attributes related to churn phase.

Data understanding & preparation

- Missing values treatment: 74% values for recharge related data are missing in some variables. We can impute by 0, considering no recharges done by customers.

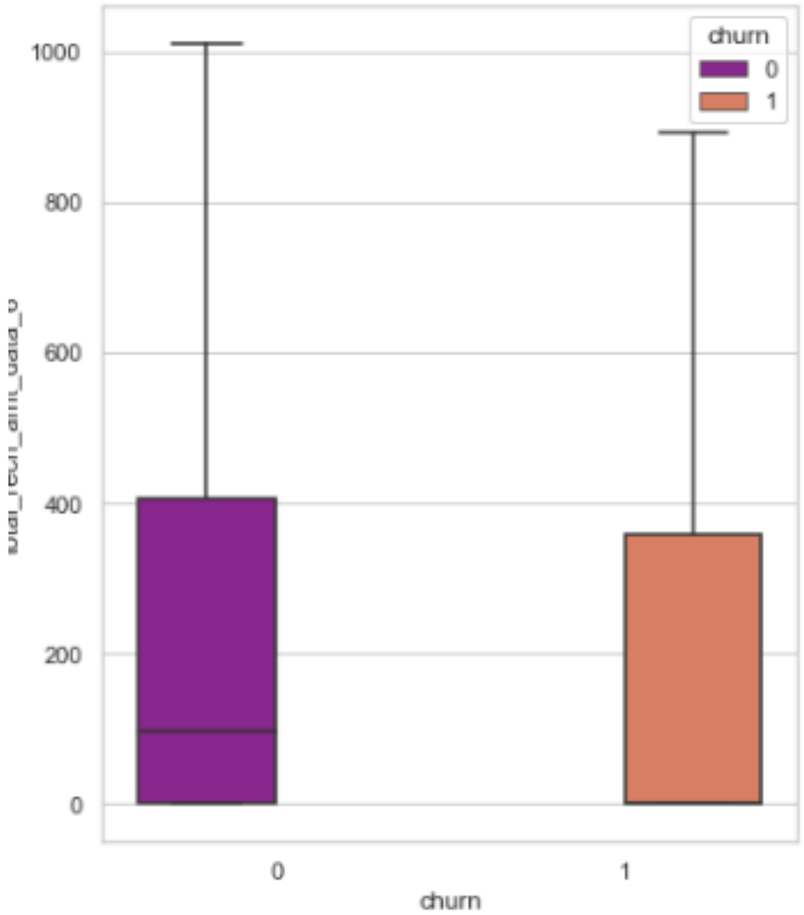
- Filtering high value customers: 70th percentile of 6th and 7th month average amount is 478. Filtering leads to 29953 rows.

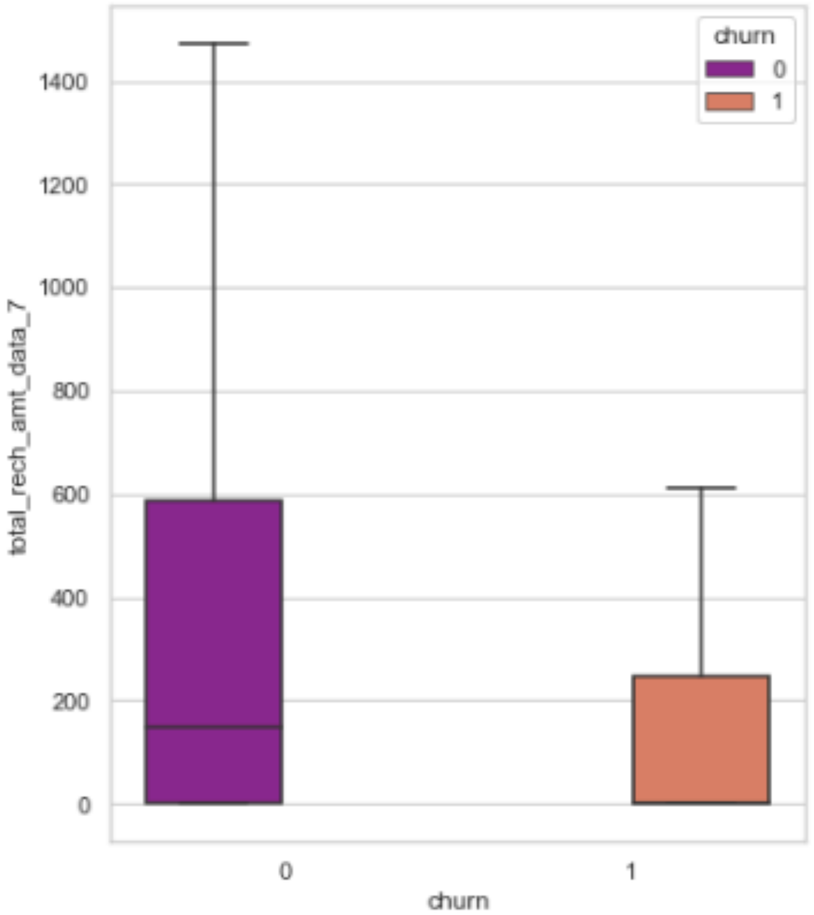
- Tag churn customers: Using 4th month data identify those customers who have not made any calls and not used internet even once. After tagging remove all attributes corresponding to churn phase.

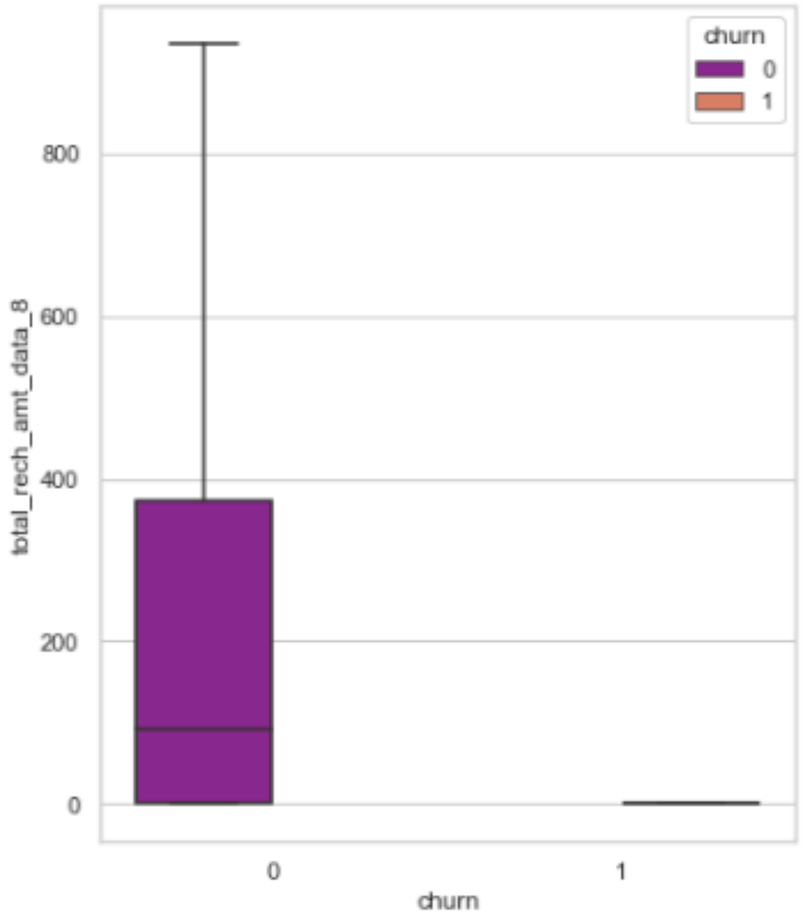
- Drop categorical columns having only one unique value as they will not add any value to model building and analysis

- Convert date columns to date format

Understand recharge amount and churn relation

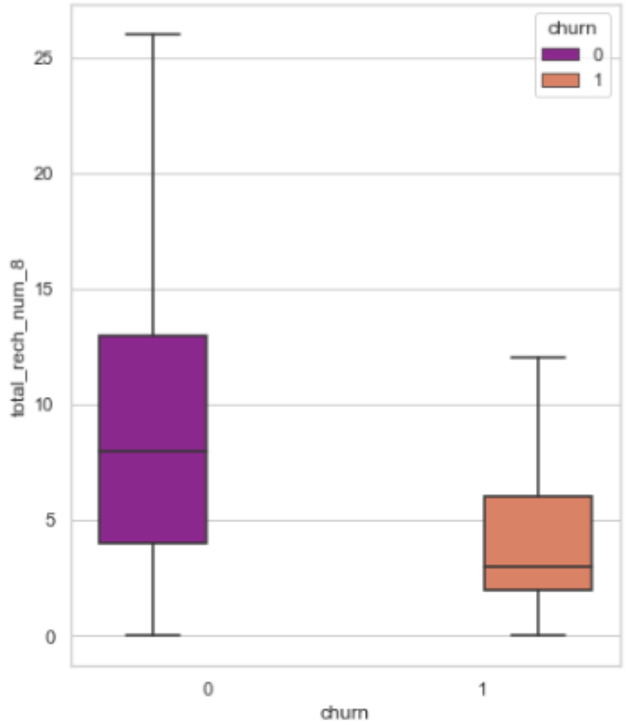


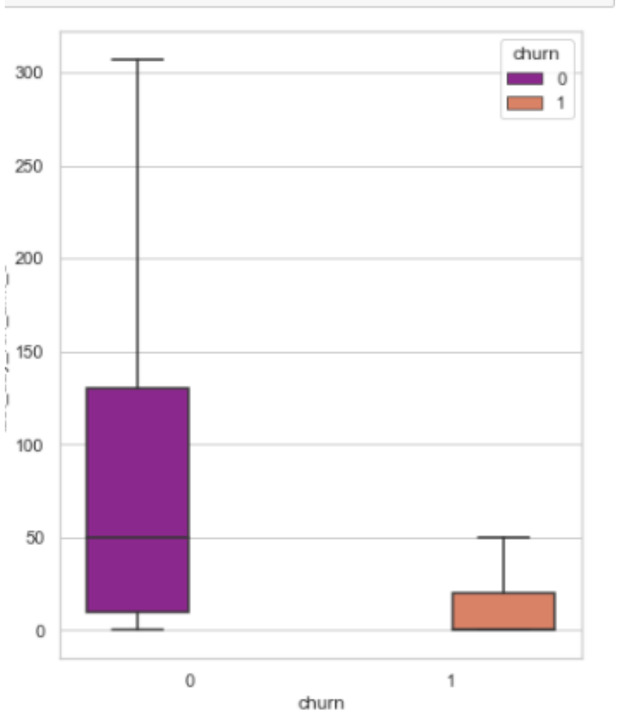


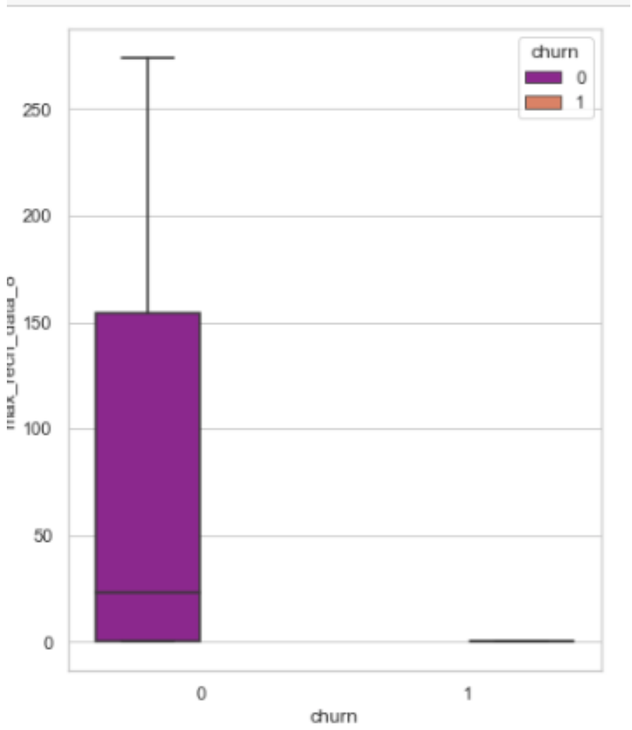


- We can see a drop in the total recharge amount for churned customers in the 8th Month (Action Phase).
- We can see that there is a huge drop in total recharge amount for data in the 8th month (action phase) for churned customers.
- We can see that there is a huge drop in maximum recharge amount for data in the 8th month (action phase) for churned customers.

Understand recharge number, data and churn relation

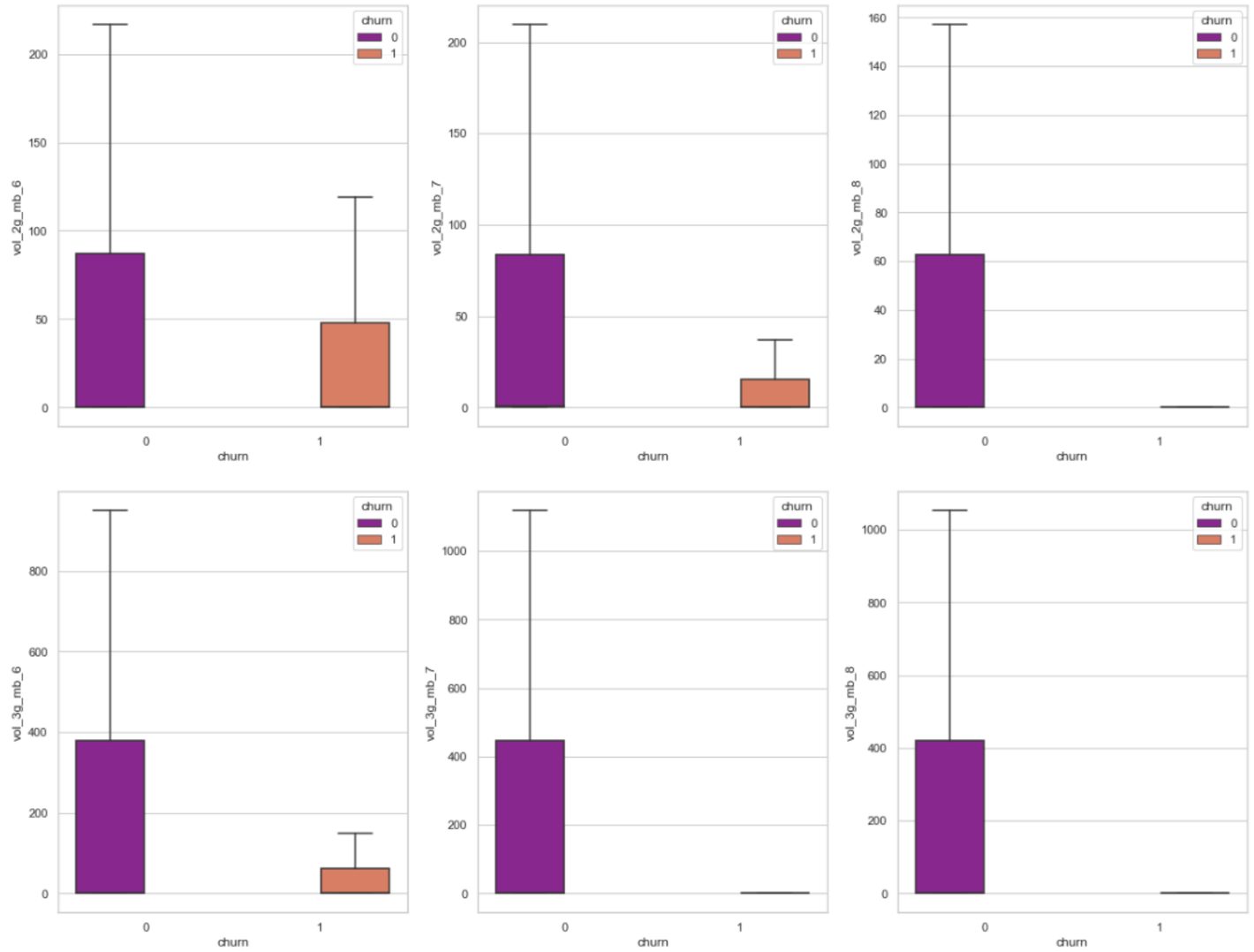






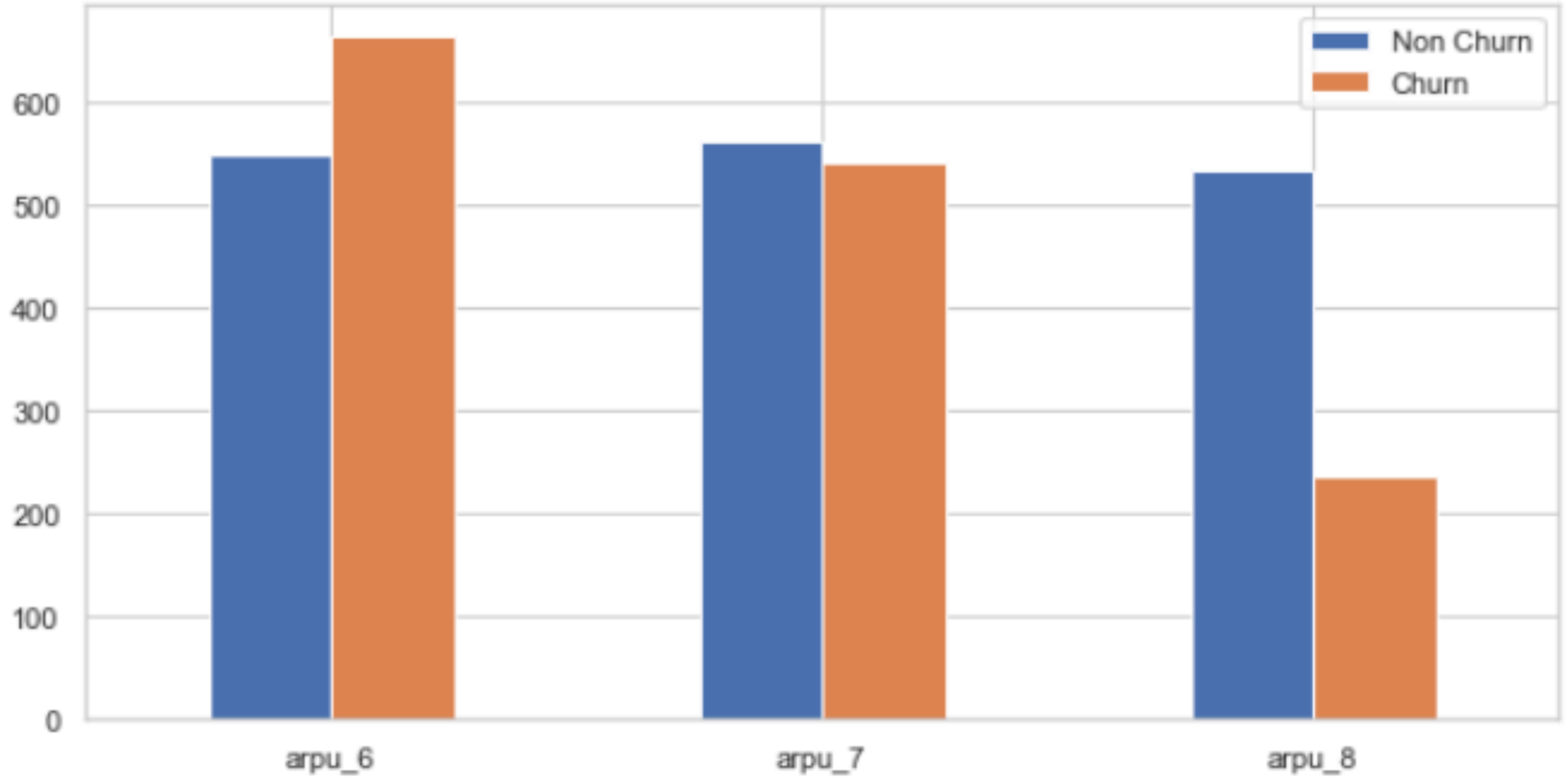
- there is a huge drop in total recharge number also in the 8th month (action phase) for churned customers.
- there is a huge drop in maximum recharge for data also in the 8th month (action phase) for churned customers.
- We are getting a huge drop in 8th month recharge amount for churned customers.

2G/3G usage and churn relation



- 2G and 3G usage for churned customers drops in 8th month
- We also see that 2G/3G usage is higher for non-churned customers indicating that churned customers might be from areas where 2G/3G service is not properly available.

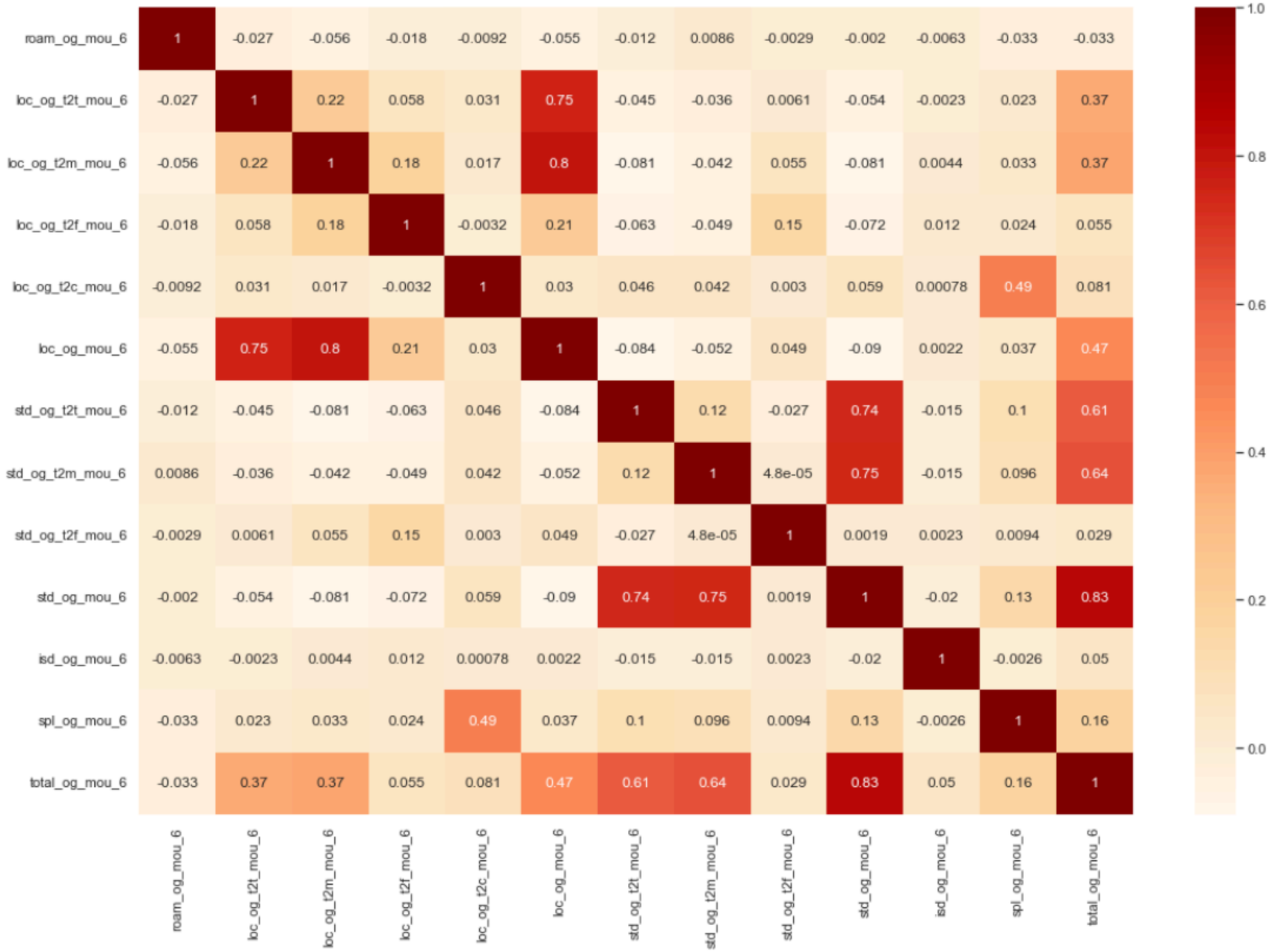
Average revenue per user and churn relation



	arpu_6	arpu_7	arpu_8
Non Churn	549.546959	562.929990	532.869746
Churn	663.709368	541.146131	237.655478

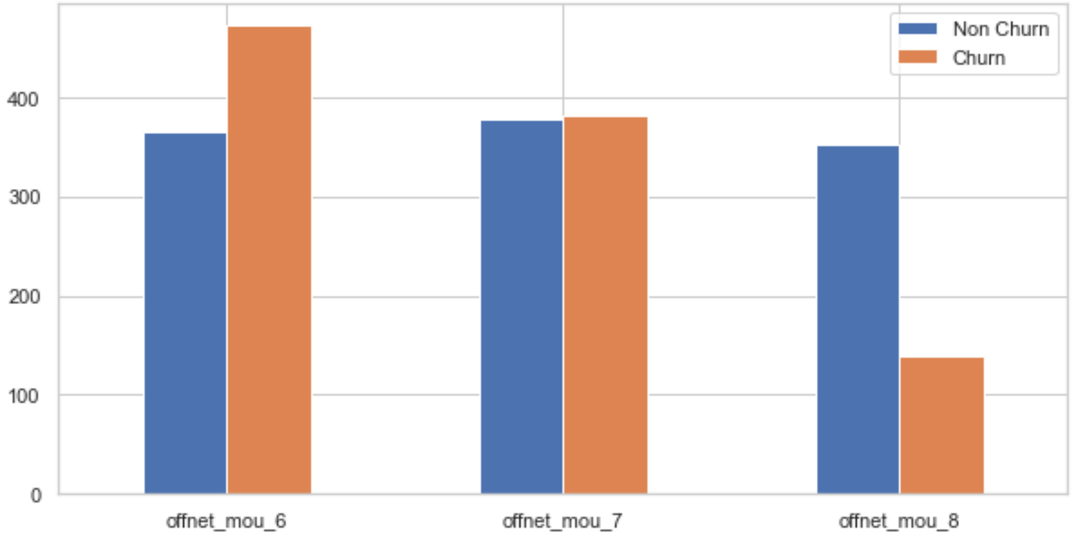
- There is a huge drop in ARPU in the 8th month for churned customers

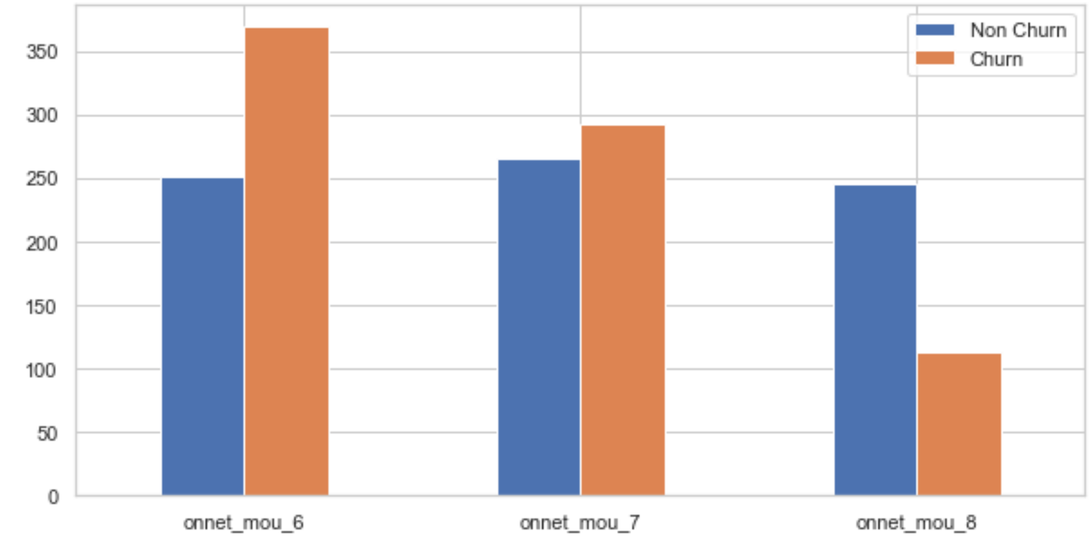
Check correlation and drop columns accordingly



- Here total_og_mou_6, std_og_mou_6 and loc_og_mou_6 seems to have strong correlation with other fileds and they needs to be inspected to avoid any multicolinearity issues.
- Here total_og_mou_6, std_og_mou_6 and loc_og_mou_6 is a combination of other variables present in dataset.So we can remove these columns for all months from the data set

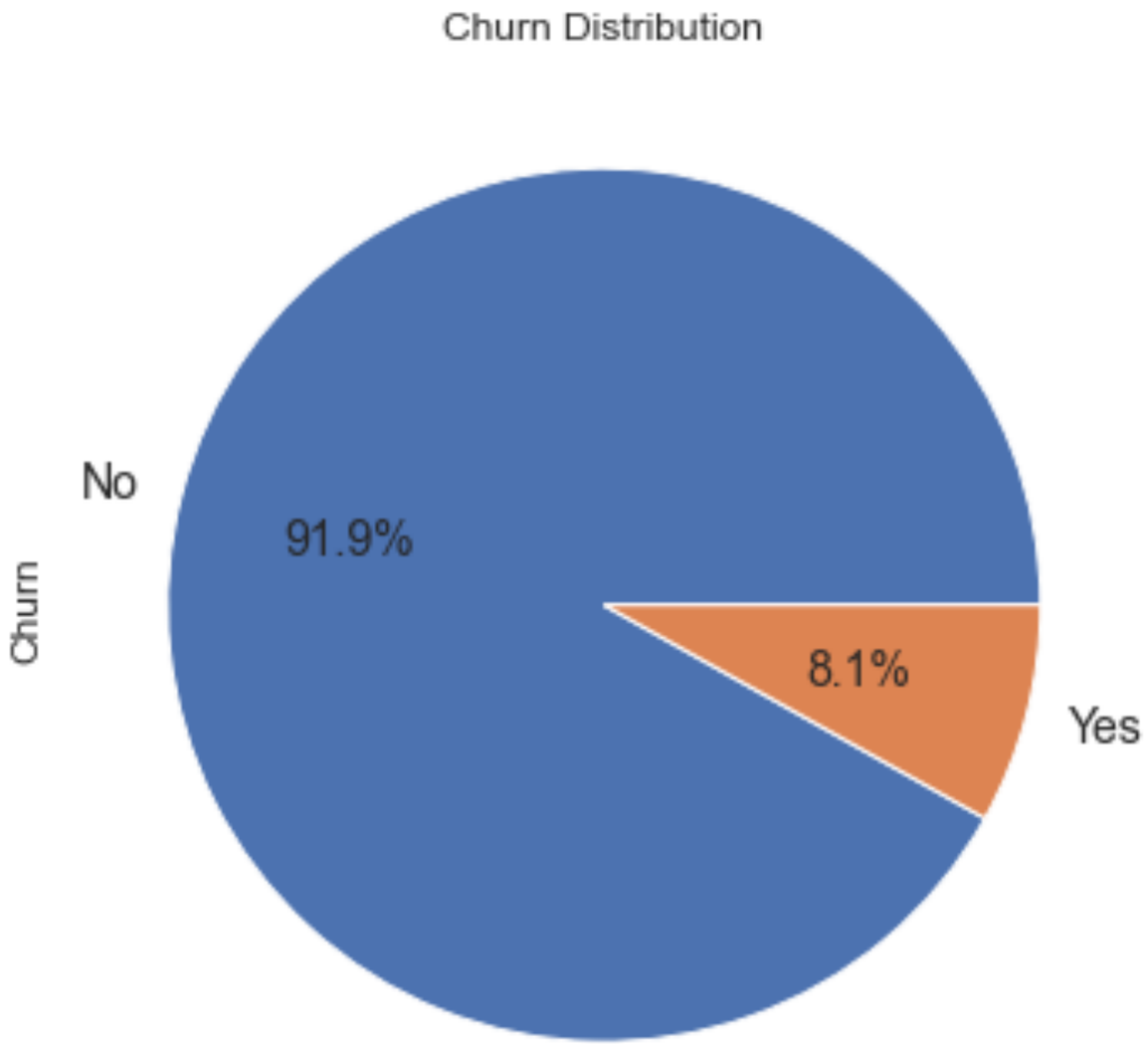
Offnet and Onnet data check





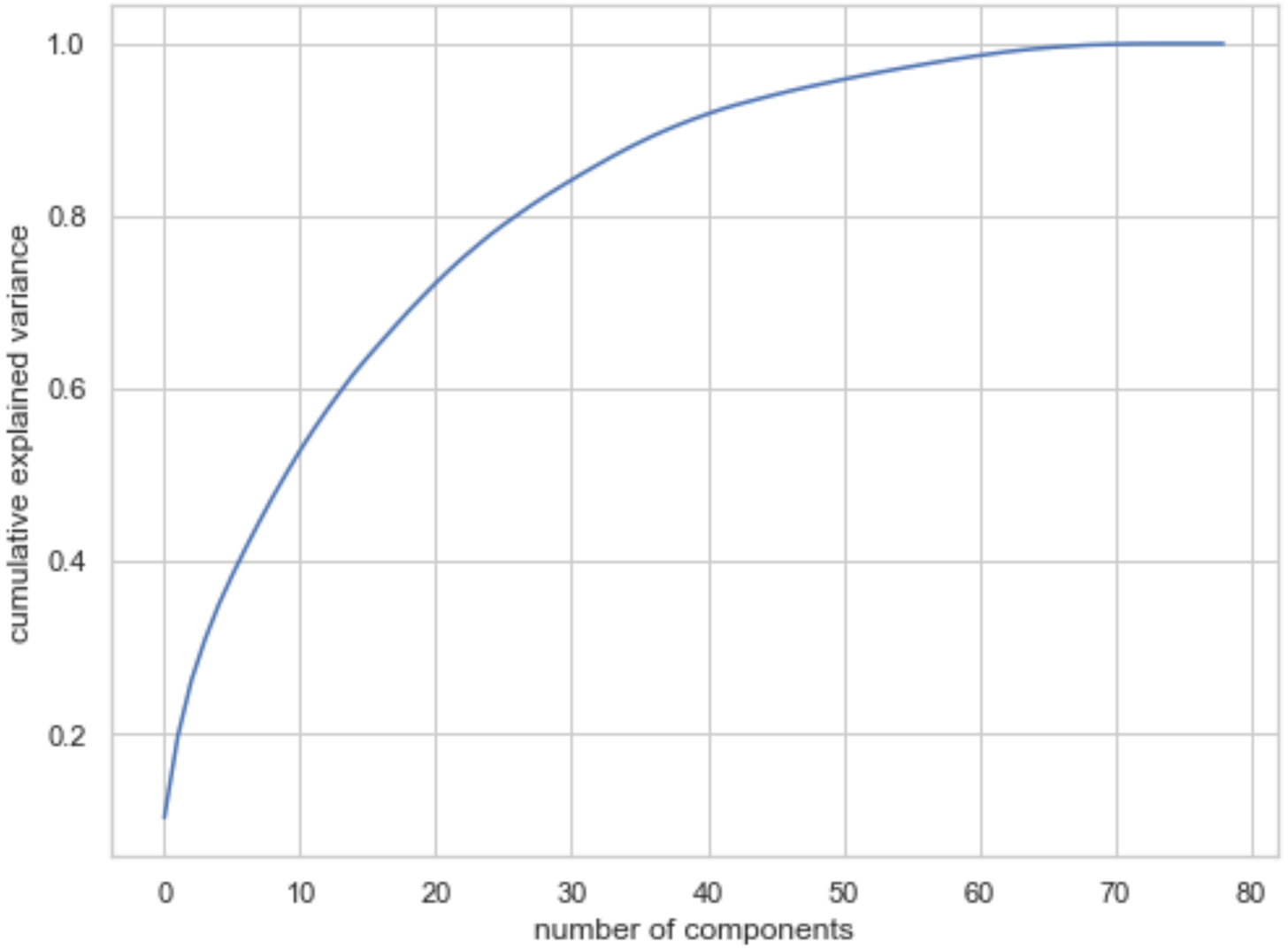
- There is a drop in both Offnet and Onnet usage in the 8th month for churned customers

Check ratio of non-churn and churn customers



- 8.1% of customers are tagged as Churn

PCA: Principal Component Analysis



- From the chart it seems that 60 components are enough to describe 95% of the variance in the dataset.We'll choose 60 components for our modeling

Further model building

- SVM Regression modeling: Make two basic models- Liner and Non-linear with default hyper parameters and compare accuracies
- Linear model accuracy score is 83.0 while non linear accuracy score is 87%. So we choose hyper parameters corresponding to non-linear.
- Analysis shows that:
- Non-linear models (high gamma) perform *much better* than the linear ones
- At any value of gamma, a high value of C leads to better performance
- Model with gamma = 0.1 tends to overfit and rest of the values seems to be good.
- This suggests that the problem and the data is **inherently non-linear** in nature, and a complex model will outperform simple, linear models in this case

- Basis the above , carry out regression using random forest and do hyper parameter tuning.

As we increase the value of max_depth, both train and test scores increase till a point, but after that test score becomme stagnant. The ensemble tries to overfit as we increase the max_depth. Thus, controlling the depth of the constituent trees will help reduce overfitting in the forest. 12 and 18 value have peek convergens and can be used for grid view search.

Recommendations

Business Insights

- Less number of high value customer are churing but for last 6 month no new high valued cusotmer is onboarded which is concerning and company should concentrate on that aspect.
- Customers with less than 4 years of tenure are more likely to churn and company should concentrate more on that segment by rolling out new schems to that group.
- Average revenue per user seems to be most important feature in determining churn prediction.
- Incoming and Outgoing Calls on romaing for 8th month are strong indicators of churn behaviour
- Local Outgoing calls made to landline , fixedline , mobile and call center provides a strong indicator of churn behaviour.
- Better 2G/3G area coverage where 2G/3G services are not good, it's strong indicator of churn business

Model Insights

- SVM with tuned hyperparameters produce best result on this dataset with 0.92 accuracy.
- Random forest also produce good accuracy with 0.91 (default overfit model) and 0.90 with tuned hyperparameters.
- As per our analysis SVM and Random forest produce best accuracy and models can be selected to predict churn data for future dataset or production.