Telecom Churn Data Case Study

Problem 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.

Business Objective

- To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.
- In this project, you will analyze customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn and identify the main indicators of churn.

Understanding and Defining Churn

- There are two main models of payment in the telecom industry - postpaid (customers pay a monthly/annual bill after using the services) and prepaid (customers pay/recharge with a certain amount in advance and then use the services).
- In the postpaid model, when customers want to switch to another operator, they usually inform the existing operator to terminate the services, and you directly know that this is an instance of churn.
- However, in the prepaid model, customers who want to switch to another network can simply stop using the services without any notice, and it is hard to know whether someone has actually churned or is simply not using the services temporarily (e.g. someone may be on a trip abroad for a month or two and then intend to resume using the services again).
- Thus, churn prediction is usually more critical (and non-trivial) for prepaid customers, and the term ‘churn’ should be defined carefully.
- Also, prepaid is the most common model in India and southeast Asia, while postpaid is more common in Europe in North America. This project is based on the Indian and Southeast Asian market.

Details of files given

- telecom_Churn_data.ipynb : The python file showing coding and data analysis
- telecom_Churn_data.xlsx : Data Dictionary
- README.md : what Project does
- Telecom_churn_data_PP : PDF file that contains a presentation to present an analysis with both technical and business aspects.

Solution Methodology

1. Reading Data
2. Cleaning Data
3. EDA
4. PCA
5. Logistic Regression
6. Finding the important coefficients

Data cleaning and data manipulation

1. Check and handle duplicate data.
2. Check and handle NA values and missing values.
3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
4. Imputation of the values, if necessary.
5. Check and handle outliers in data

PCA

- Principal Component Analysis :

  - Is a Unsupervised learning algorithm technique used to examine the interrelations among a set of variables
  - It is a technique to draw strong patterns from the given dataset by reducing the variances.
  - PCA works by considering the variance of each attribute because the high attribute shows the good split between the classes, and hence it reduces the dimensionality.
  - The PCA algorithm is based on some mathematical concepts such as:
    - Variance
    - Covariance

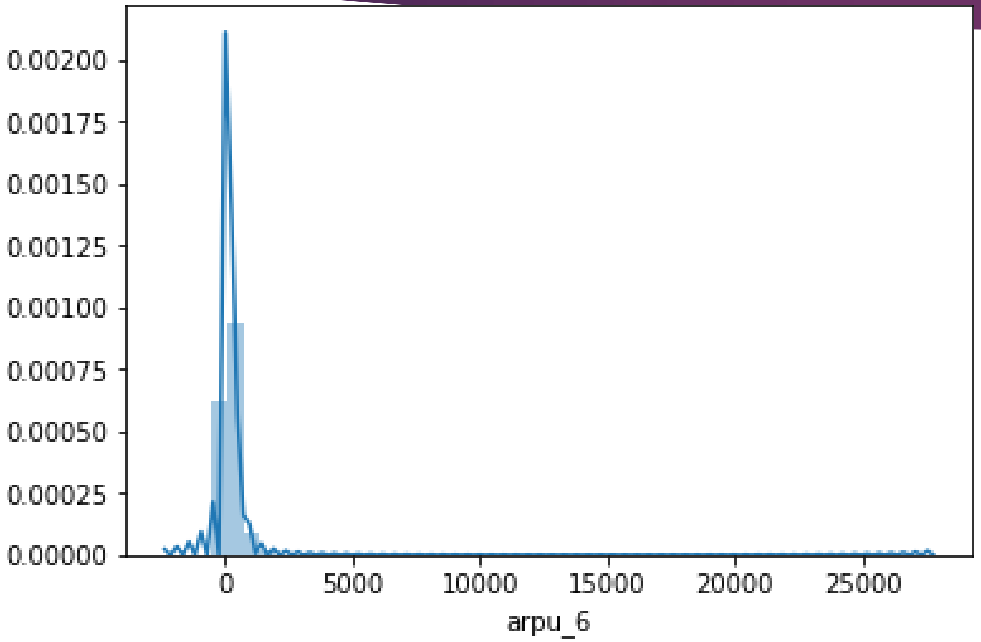
Logistic Regression

- Logistic regression is an example of supervised learning.
- It is used to calculate or predict the probability of a binary (yes/no) event occurring.
- It is often used for classification and predictive analytics.
- Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables.

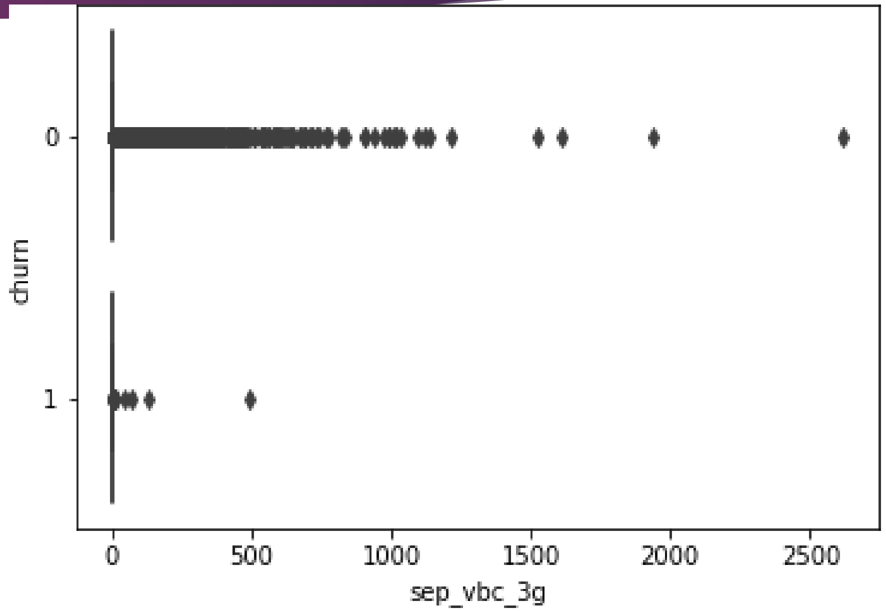
EDA

1. Univariate data analysis : value count, distribution of variable etc.
2. Bivariate data analysis : correlation coefficients and pattern between the variables etc

univariate(arpu_6)

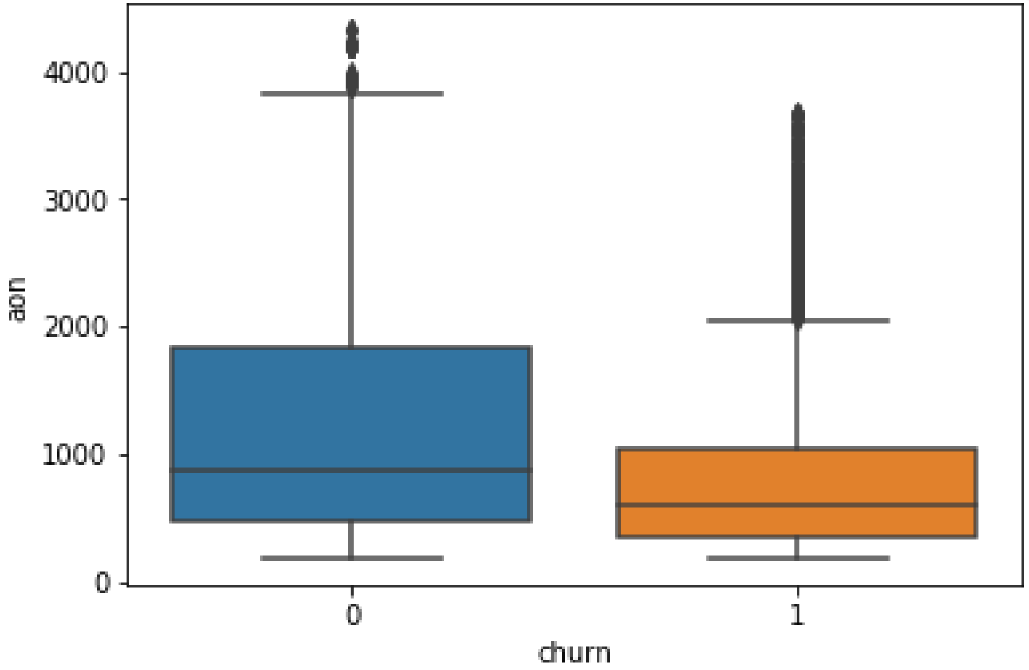


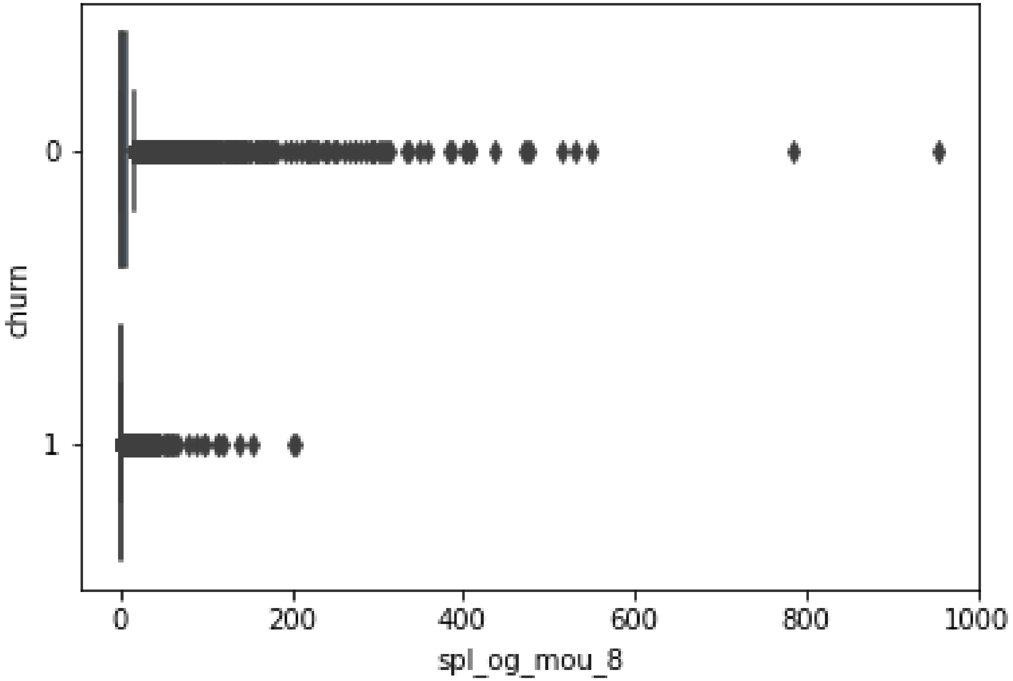
bivariate(sep_vbc _3g)



bivariate(hurn)

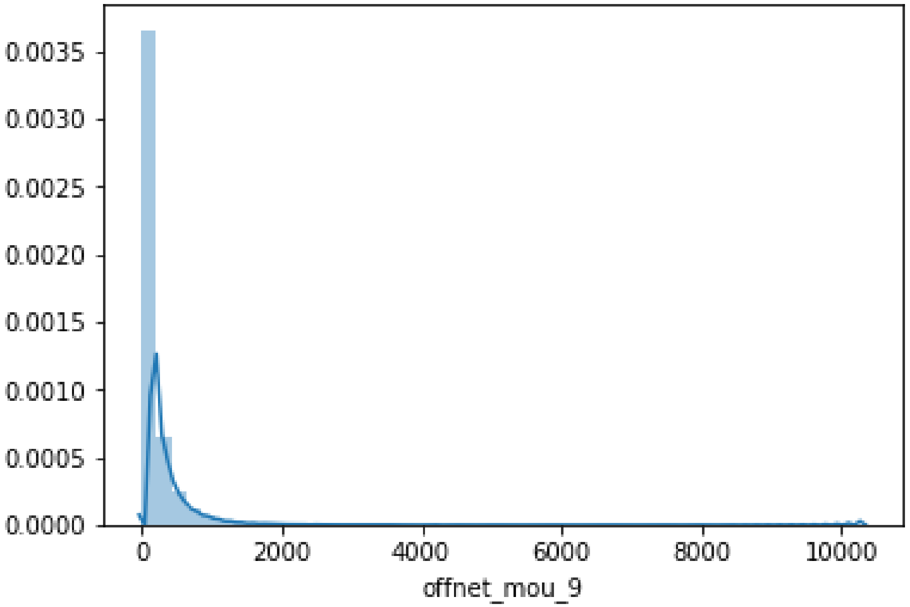
bivariate(spl_og_mou_8)

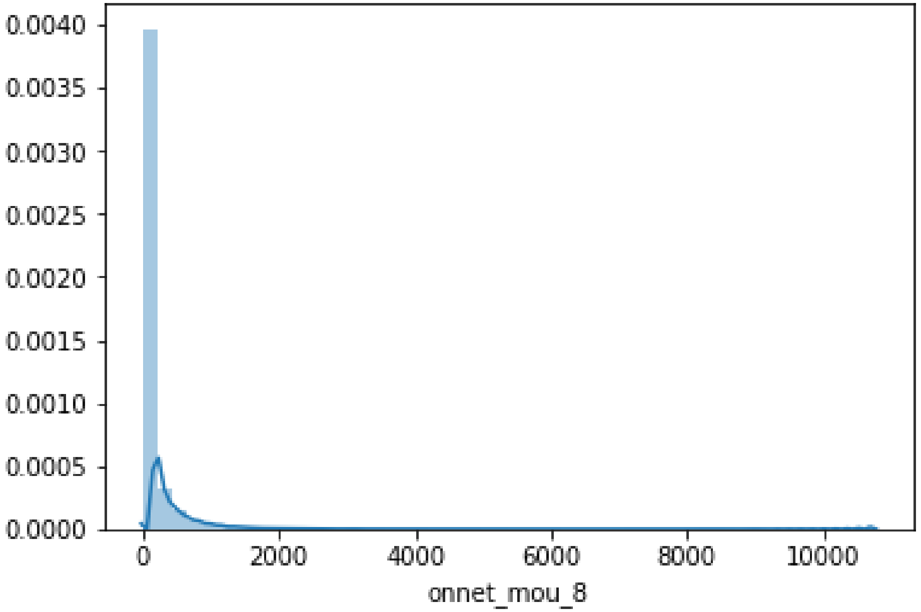




univariate(offnet_mou_9)

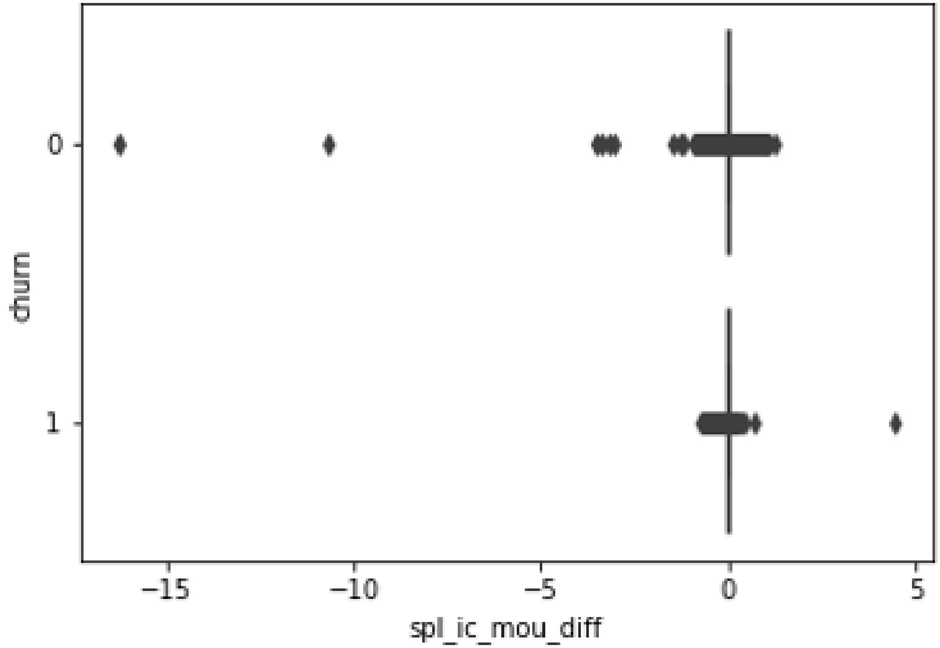
univariate(onnet_mou_8)

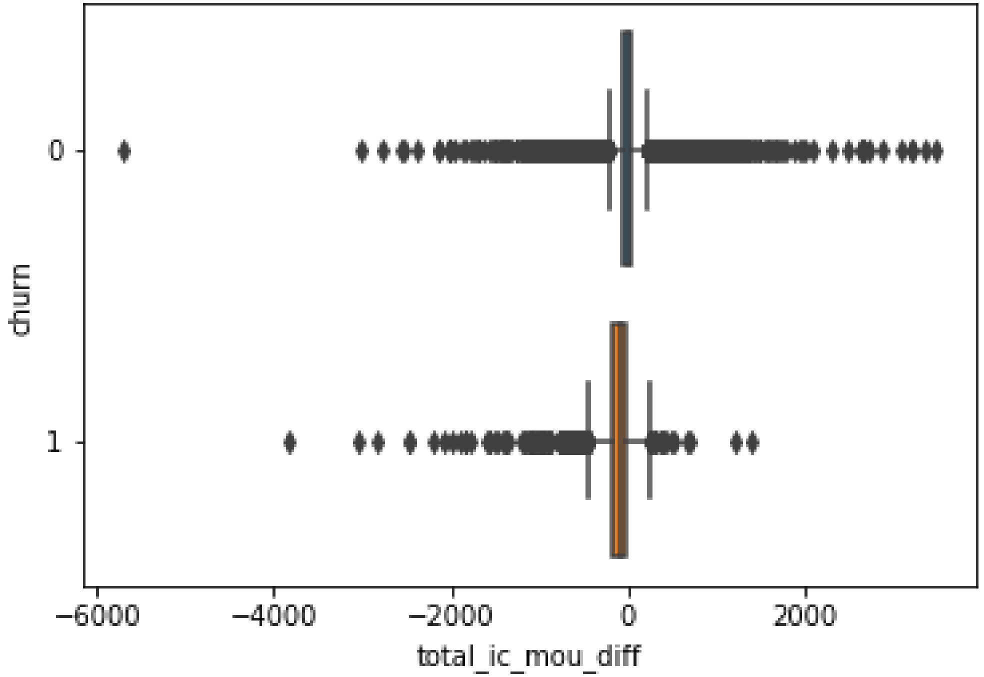




bivariate(spl_ic_mou_diff)

bivariate(total_ic_mou_diff)





EDA - Summary

- Calls Revenue(3 columns):

  - Invalid Values : Having minimum values as negatives, indicating some customers are making loss to the company. These columns are either invalid or not adding value to our prediction, can be dropped from the dataset.
  - Standardize: Revenue columns can be rounded to 2 decimal places.
  - Minutes of usage(60+ columns):
    - Usage minutes is generally 0 except for few outliers, for below variables:
    - Roaming Incoming ISD Incoming Special Incoming Others STD incoming T2F STD outgoing T2F Outgoing Others ISD Outgoing Local Outgoing T2C (Customer care calls)

- Most of the columns have outliers.

- Aggregating Columns based on Incoming and Outgoing, or Aggregating based on Each Type of Incoming Calls and Outgoing Calls and looking at the metrics will give a better understanding of the data.

- Recharge (12 Numeric + 3 Date columns)

- Data Type Conversion:
  - Data in numeric columns are integers, so can be converted to int type.
  - Date columns need to be converted to date type
  - Data 2G And 3G(22 Columns)

- Most of the columns have median as O and have outliers

- vbc_3g columns need column renaming as it needs month to be encoded to its number.
- Standardize: Columns can be rounded off to 2 decimal places.
- Age on Network (1 Column)
- Feature can be derived from AON column.
- Churn (Dependent Variable)
- There exists a Class Imbalance in the dataset, where actual churn customers are only 6% of the dataset.

- Reviewing the Dropped Columns:

- More columns will be lost because of dropping missing value columns, while it can be handled to be imputed by considered 0 as missing values follow a pattern where Calls only users have blanks for Data related columns and the vice versa.

- Feature Engineering - Thoughts

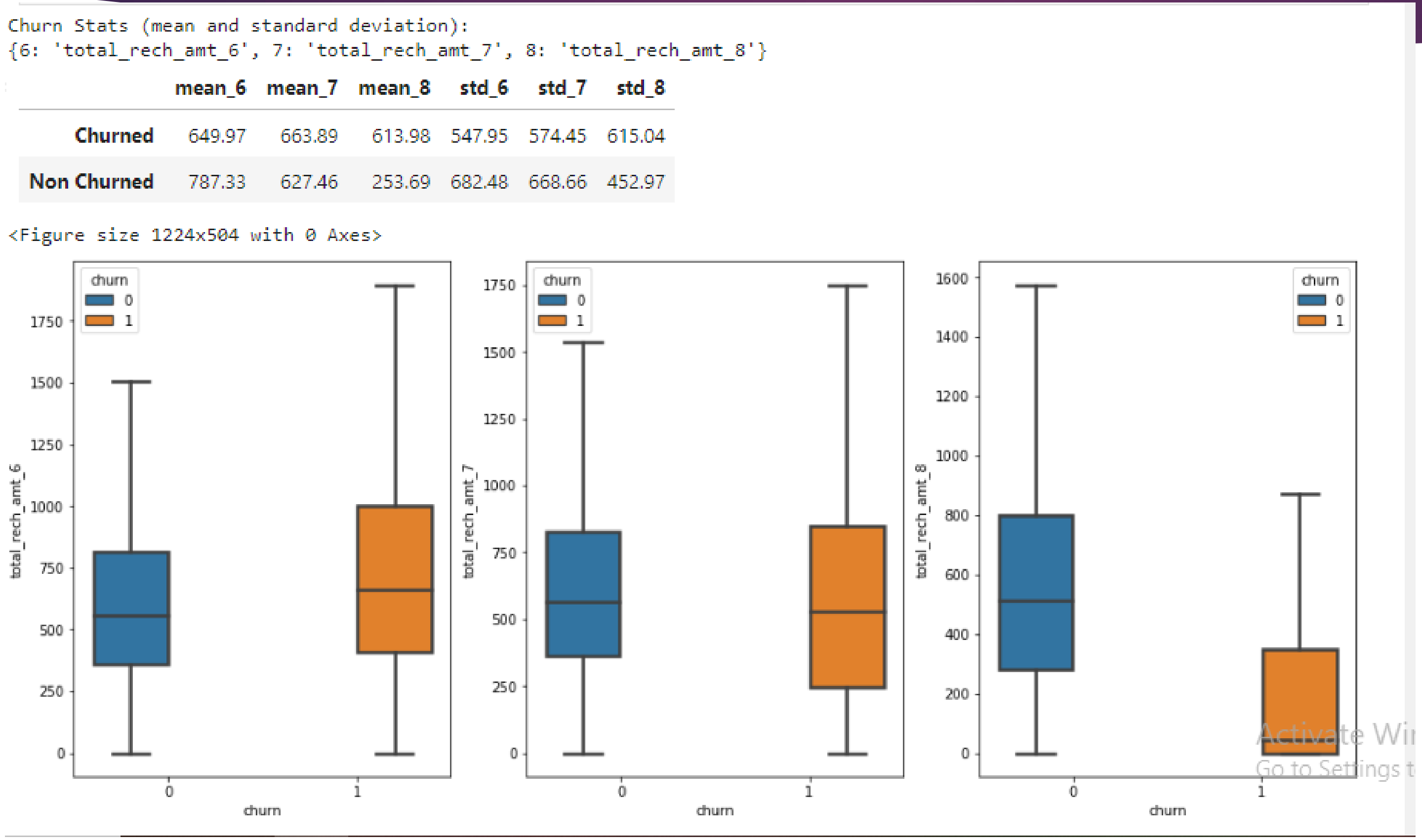
- Derive no. of years the customer is using network from AON

- Derive fields to indicate the type of user the customer is: Uses Both Calls and Data, Only Calls, Only Data, Only Incoming calls, Only Outgoing calls, etc.

- Bin the customers into different segments based on Service usage, Recharge amount, Usage/Recharge pattern.

- Calls to Customer Care is a Key indicator that customer is not happy with the services, derive columns like time over call

Anlyze total_rech_amt across 6th, 7th and 8th month '

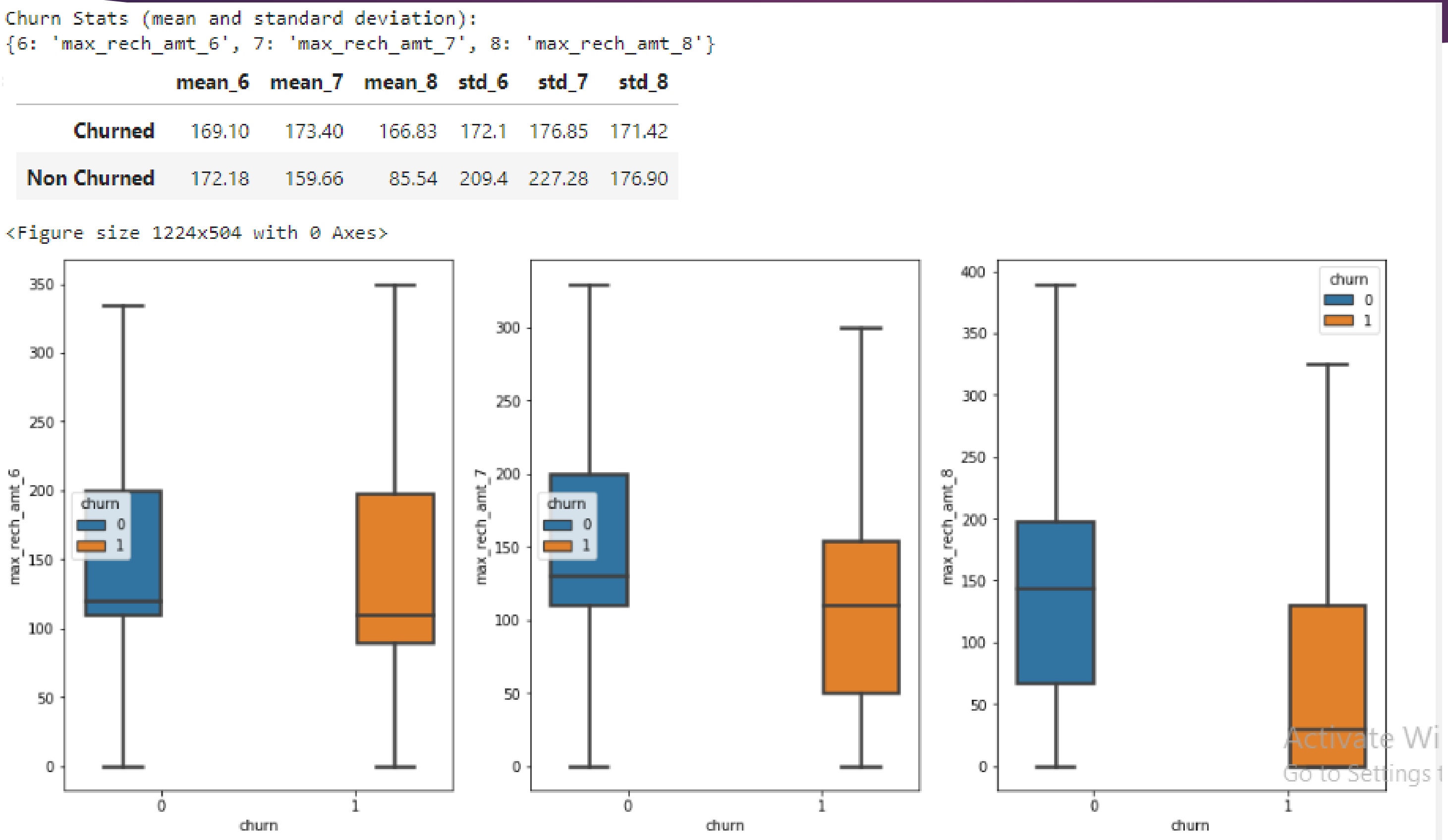


```
Churn Stats (mean and standard deviation):
{6: 'total_rech_amt_6', 7: 'total_rech_amt_7', 8: 'total_rech_amt_8'}
```

	mean_6	mean_7	mean_8	std_6	std_7	std_8
Churned	649.97	663.89	613.98	547.95	574.45	615.04
Non Churned	787.33	627.46	253.69	682.48	668.66	452.97

```
<Figure size 1224x504 with 0 Axes>
```

Anlyze total recharge amount for data

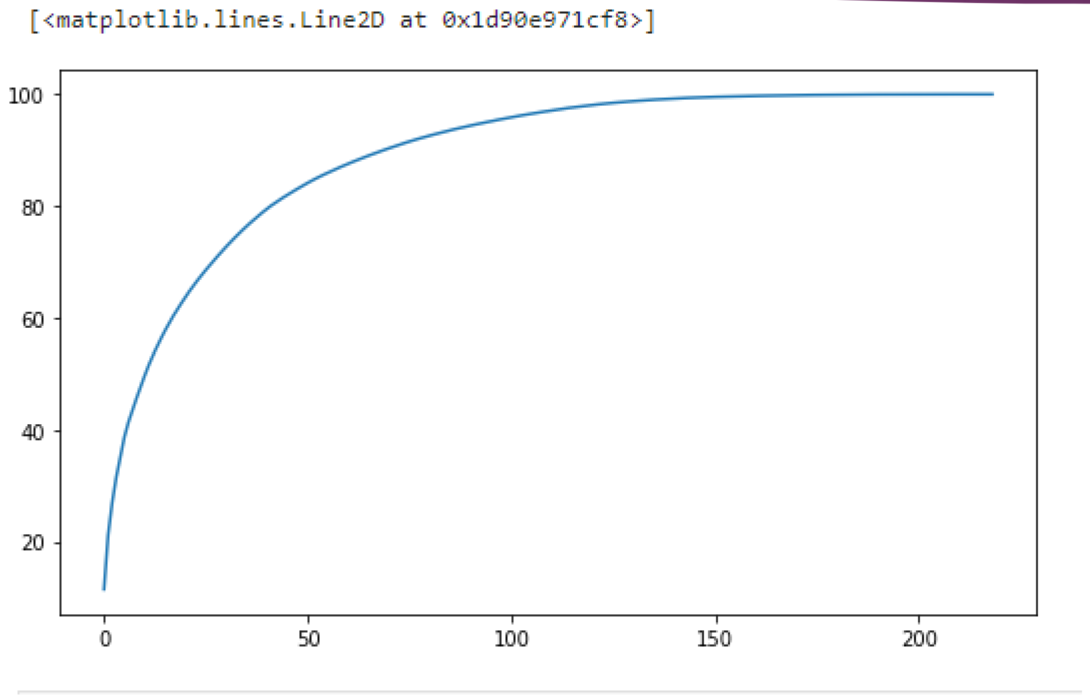


```
Churn Stats (mean and standard deviation):
{6: 'max_rech_amt_6', 7: 'max_rech_amt_7', 8: 'max_rech_amt_8'}
```

	mean_6	mean_7	mean_8	std_6	std_7	std_8
Churned	169.10	173.40	166.83	172.1	176.85	171.42
Non Churned	172.18	159.66	85.54	209.4	227.28	176.90

```
<Figure size 1224x504 with 0 Axes>
```

Plot feature variance



```
[<matplotlib.lines.Line2D at 0x1d90e971cf8>]
```

Main indicator of churn

- From analysis it is clear that the factors affecting the churn are:

  - total_ic_mou_8 (Total incoming call: Minutes of usage in the action phase)
  - total_rech_amt_diff (Total recharge amount difference)
  - total_og_mou_8 (Total outgoing call: Minutes of usage in the action phase)
  - arpu (Average revenue per user)
  - roam_ic_mou_8 (Roaming incoming call: Minutes of usage in the action phase)
  - roam_og_mou_8 (Roaming outgoing call: Minutes of usage in the action phase)
  - std_ic_mou_8(STD incoming call: Minutes of usage in the action phase)
  - std_og_mou_8 (STD outgoing call: Minutes of usage in the action phase)
  - av_rech_amt_data_8(average recharge amount in the action phase).

Conclusion :

Steps to help reduce churn

- Give special; discounts to customers according to their usage
- Provide additional internet services on recharge.
- Speak to customers to fulfil their desires.
- Lower tariffs on data usage, a better 2G area coverage where 3G is not available.
- Expansion of 3G network where 3G is currently not available.