Telecom Churn Case Study

- Kirti Rani

Following are the Top 5 Rows of telecom Churn data set. Showing Mobile number, Circle id….

```
# Reading the dataset
df = pd.read_csv('D:/Upgrad Assignment/Telecom Churn Case Study_Kirti Rani/telecom_churn_data.csv')
df.head()
```

	mobile_number	circle_id	loc_og_t2o_mou	std_og_t2o_mou	loc_ic_t2o_mou	last_date_of_month_6	last_date_of_month_7	last_date_of_month_8	last_date_of
0	7000842753	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	
1	7001865778	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	
2	7001625959	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	
3	7001204172	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	
4	7000142493	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	

Data set has 99999 Rows and 226 Column.

```
df.shape
```

```
(99999, 226)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99999 entries, 0 to 99998
Columns: 226 entries, mobile_number to sep_vbc_3g
dtypes: float64(179), int64(35), object(12)
memory usage: 172.4+ MB
```

```
df.describe()
```

	mobile_number	circle_id	loc_og_t2o_mou	std_og_t2o_mou	loc_ic_t2o_mou	arpu_6	arpu_7	arpu_8	arpu_9	onnet_mou_6	on
count	9.999900e+04	99999.0	98981.0	98981.0	98981.0	99999.000000	99999.000000	99999.000000	99999.000000	96062.000000	96
mean	7.001207e+09	109.0	0.0	0.0	0.0	282.987358	278.536648	279.154731	261.645069	132.395875	
std	6.956694e+05	0.0	0.0	0.0	0.0	328.439770	338.156291	344.474791	341.998630	297.207406	
min	7.000000e+09	109.0	0.0	0.0	0.0	-2258.709000	-2014.045000	-945.808000	-1899.505000	0.000000	
25%	7.000606e+09	109.0	0.0	0.0	0.0	93.411500	86.980500	84.126000	62.685000	7.380000	
50%	7.001205e+09	109.0	0.0	0.0	0.0	197.704000	191.640000	192.080000	176.849000	34.310000	
75%	7.001812e+09	109.0	0.0	0.0	0.0	371.060000	365.344500	369.370500	353.466500	118.740000	
max	7.002411e+09	109.0	0.0	0.0	0.0	27731.088000	35145.834000	33543.624000	38805.617000	7376.710000	8

Below showing missing value of column’s attributes in %

```
# Cheking percent of missing values in columns
df_missing_columns = (round(((df.isnull().sum()/len(df.index))*100),2).to_frame('null')).sort_values('null', ascending=False)
df_missing_columns
```

	null
arpu_3g_6	74.85
night_pck_user_6	74.85
total_rech_data_6	74.85
arpu_2g_6	74.85
max_rech_data_6	74.85
...	...
max_rech_amt_7	0.00
max_rech_amt_6	0.00
total_rech_amt_9	0.00
total_rech_amt_8	0.00
sep_vbc_3g	0.00

Now we have 99999 Rows and 226 Column after deleting column's which has more than 30% missing values

```
# List the columns having more than 30% missing values
col_list_missing_30 = list(df_missing_columns.index[df_missing_columns['null'] > 30])
```

```
# Delete the columns having more than 30% missing values
df = df.drop(col_list_missing_30, axis=1)
```

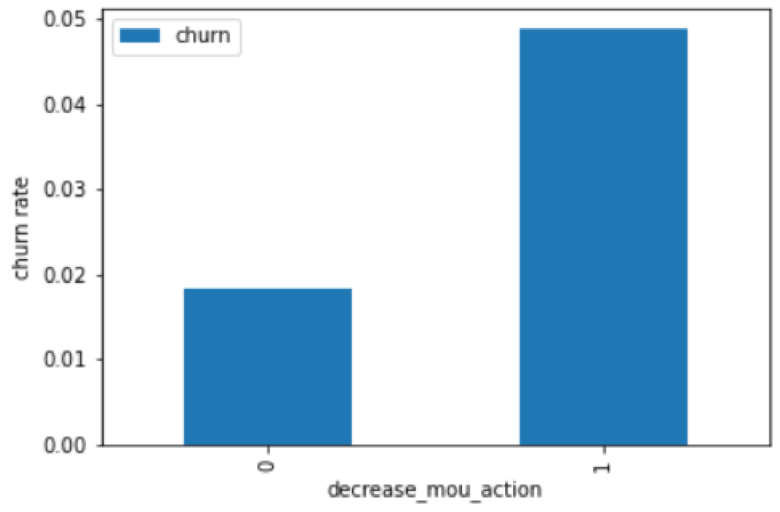
```
df.shape
```

```
(99999, 226)
```

Churn rate on the basis whether the customer decreased her/his MOU in action month

```
# Converting churn column to int in order to do aggfunc in the pivot table
data['churn'] = data['churn'].astype('int64')
```

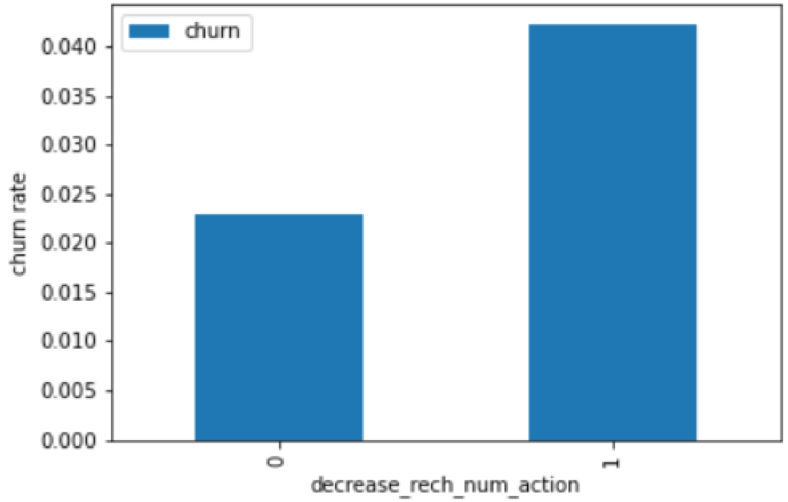
```
data.pivot_table(values='churn', index='decrease_mou_action', aggfunc='mean').plot.bar()
plt.ylabel('churn rate')
plt.show()
```



Analysis

We can see that the churn rate is more for the customers, whose minutes of usage(mou) decreased in the action phase than the good phase.

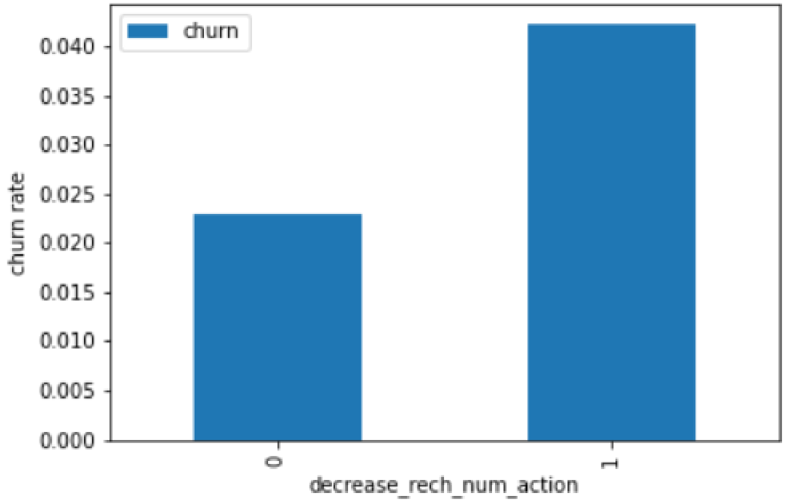
```
data.pivot_table(values='churn', index='decrease_rech_num_action', aggfunc='mean').plot.bar()
plt.ylabel('churn rate')
plt.show()
```



Analysis

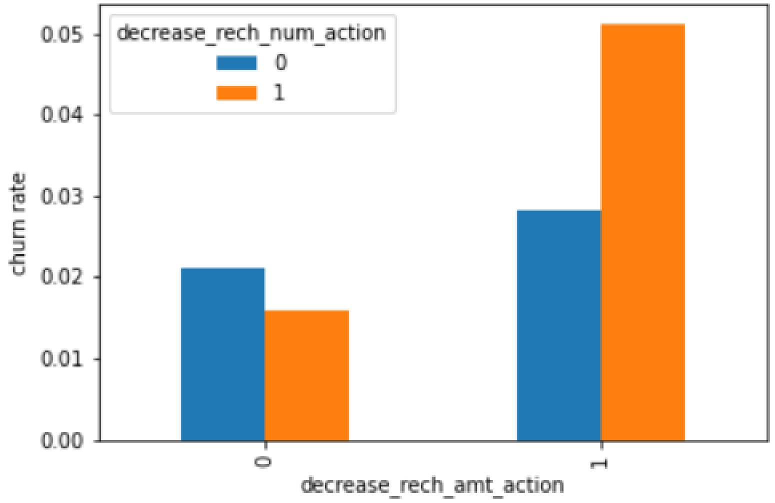
As expected, the churn rate is more for the customers, whose number of recharge in the action phase is lesser than the number in good phase.

```
data.pivot_table(values='churn', index='decrease_rech_num_action', aggfunc='mean').plot.bar()
plt.ylabel('churn rate')
plt.show()
```



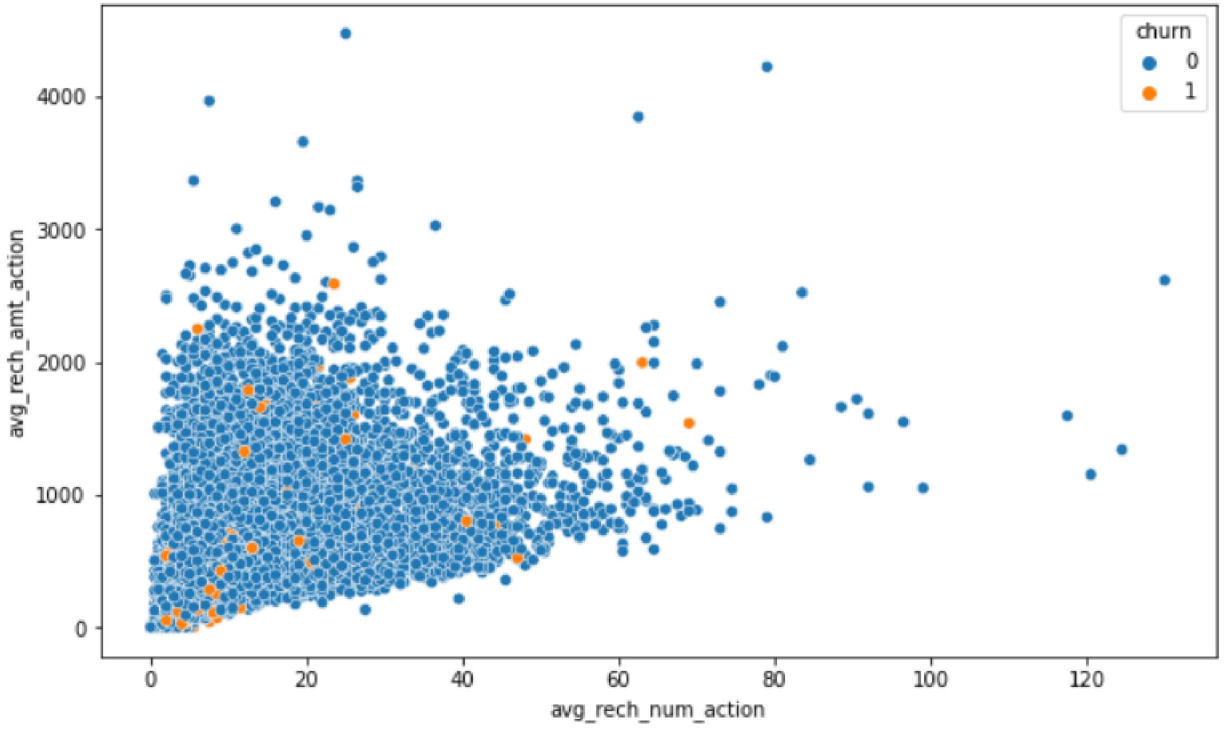
Analysis

As expected, the churn rate is more for the customers, whose number of recharge in the action phase is lesser than the number in good phase.



Analysis

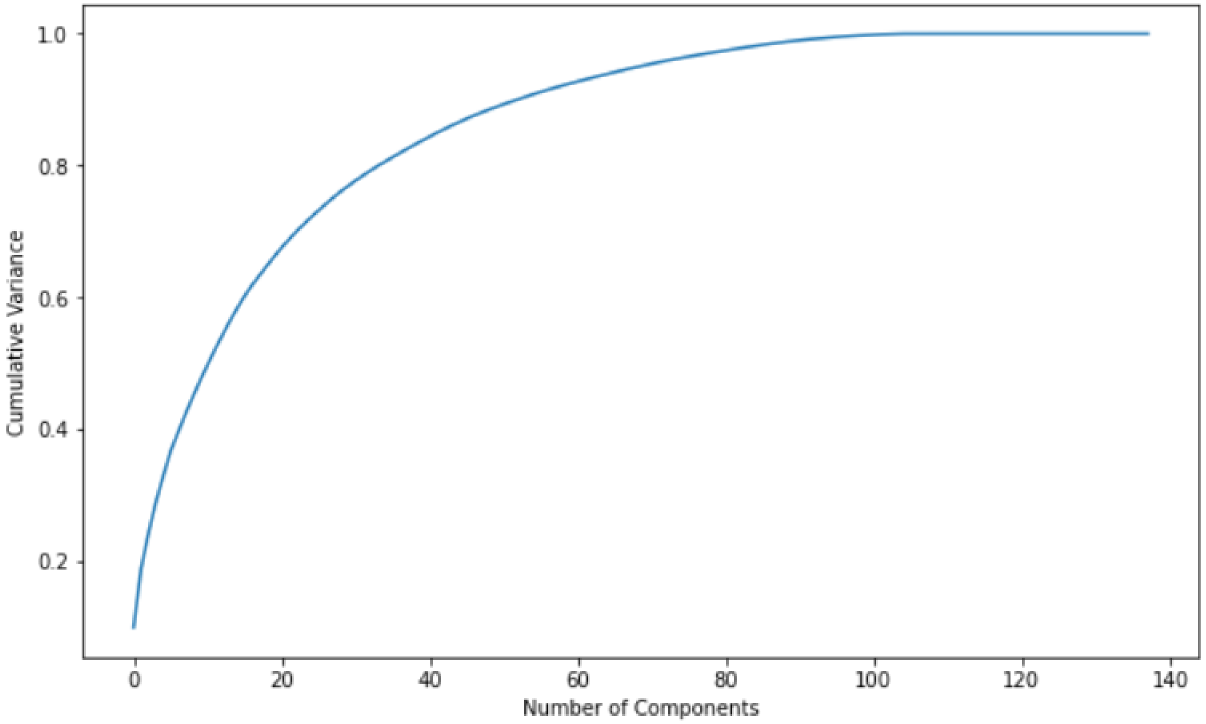
We can see from the above plot, that the churn rate is more for the customers, whose recharge amount as well as number of recharge have decreased in the action phase than the good phase.



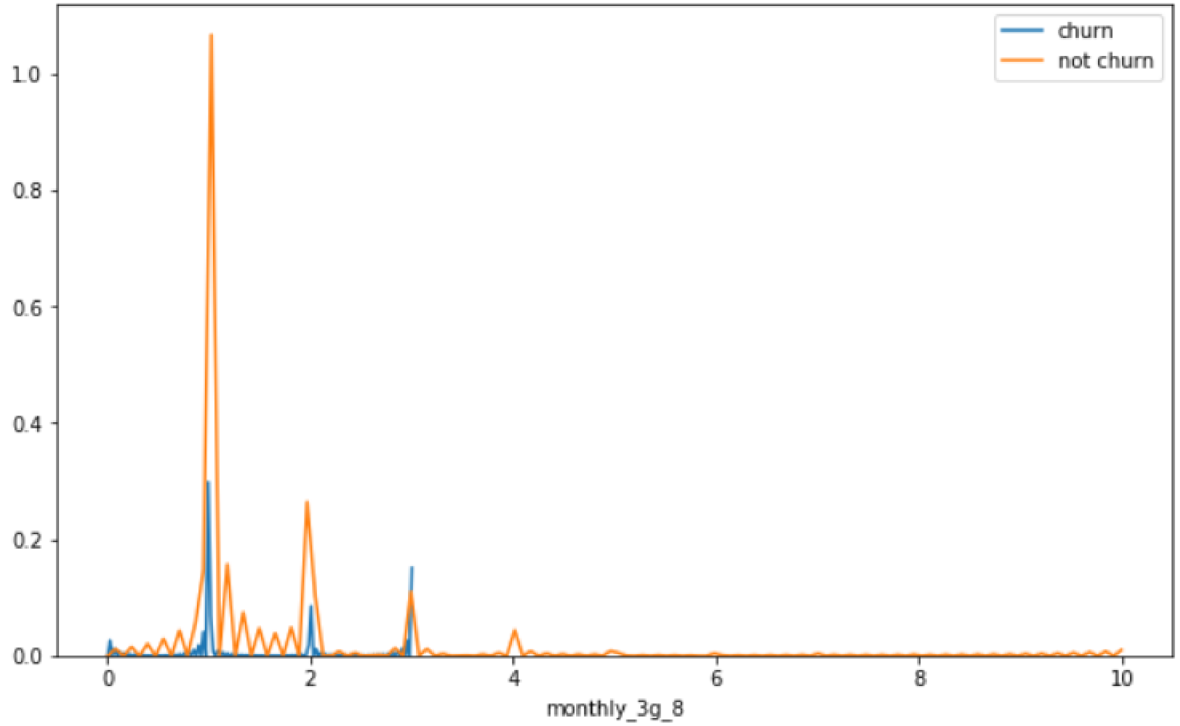
Analysis

We can see from the above pattern that the recharge number and the recharge amount are mostly propotional. More the number of recharge, more the amount of the recharge.

```
Text(0, 0.5, 'Cumulative Variance')
```



We can see that `60 components` explain amost more than 90% variance of the data. So, we will perform PCA with 60 components.



The number of mothly 3g data for August for the churn customers are very much populated aroud 1, whereas of non churn customers it spreaded accross various numbers.

Similarly we can plot each variables, which have higher coefficients, churn distribution.

Model analysis

1.We can see that there are few features have positive coefficients and few have negative.
2.Many features have higher p-values and hence became insignificant in the model.

Coarse tuning (Auto+Manual)

We'll first eliminate a few features using Recursive Feature Elimination (RFE), and once we have reached a small set of variables to work with, we can then use manual feature elimination (i.e. manually eliminating features based on observing the p-values and VIFs).

Recommendations-

1.Target the customers, whose minutes of usage of the incoming local calls and outgoing ISD calls are less in the action phase (mostly in the month of August).

2.Target the customers, whose outgoing others charge in July and incoming others on August are less.

3.Also, the customers having value based cost in the action phase increased are more likely to churn than the other customers. Hence, these customers may be a good target to provide offer.

4.Cutomers, whose monthly 3G recharge in August is more, are likely to be churned.

1.Recommendations-

5. Customers having decreasing STD incoming minutes of usage for operators T to fixed lines of T for the month of August are more likely to churn.

6. Cutomers decreasing monthly 2g usage for August are most probable to churn.

7. Customers having decreasing incoming minutes of usage for operators T to fixed lines of T for August are more likely to churn.

8. roam_og_mou_8 variables have positive coefficients (0.7135). That means for the customers, whose roaming outgoing minutes of usage is increasing are more likely to churn.

Thank You !!