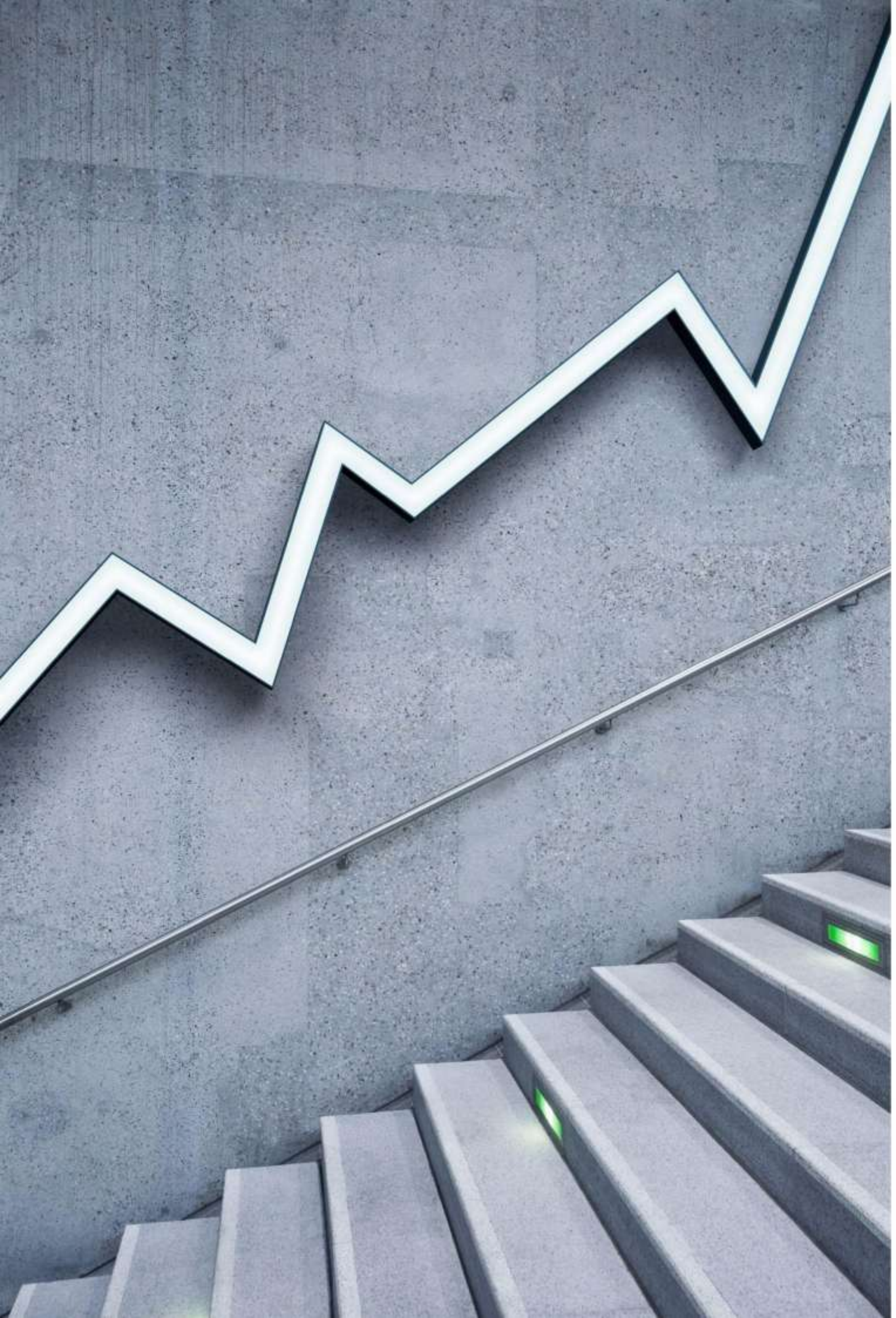


Telecom Churn Analysis

TEJESH N

04/02/2023

Background

- 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, our objective is to **predict which customers are at high risk of churn.**

Data Understanding (1)

- A total of 99999 samples were given in input file with 226 features. Most of the features (214/226) are of numerical types.

```
In [4]: # Checking the dimensions of the dataset
        telecom_data.shape

Out[4]: (99999, 226)

In [5]: telecom_data["mobile_number"].nunique()

Out[5]: 99999

In [8]: cat_cols = telecom_data.select_dtypes("object")
        print("categorical columns: {}".format(len(cat_cols.columns)))
        print("numerical columns: {}".format(len(telecom_data.columns) - len(cat_cols.columns)))

        categorical columns: 12
        numerical columns: 214
```

Data Understanding (2)

- There are 16 columns containing unique values for each row. Hence they can be dropped being included in training data
- There are 40 columns with more than 70% missing values.

In [13]:
```
# Lets check the columns unique values and drop such columns with its value as 1
unique_col=[i for i in telecom_data.columns if telecom_data[i].nunique() == 1]
telecom_data.drop(unique_col, axis=1, inplace = True)
print("\n The following Columns are dropped from the dataset as their unique value is 1. (i.e.)It has no variance in the model\n'
      unique_col)
```

```
 The following Columns are dropped from the dataset as their unique value is 1. (i.e.)It has no variance in the model
 ['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_month_9', 'std_og_t2c_mou_6', 'std_og_t2c_mou_7', 'std_og_t2c_mou_8', 'std_og_t2c_mou_9', 'std_ic_t
2o_mou_6', 'std_ic_t2o_mou_7', 'std_ic_t2o_mou_8', 'std_ic_t2o_mou_9']
```

In [13]:
```
# Checkng the overall missing values in the dataset
null_df = ((telecom_data.isnull().sum()/telecom_data.shape[0])*100).round(2).sort_values(ascending=False)
```

In [14]:
```
print("There are {} columns with more than 70% missing values.".format(len(null_df[null_df > 70])))
```

```
There are 40 columns with more than 70% missing values.
```

Data Understanding (3)

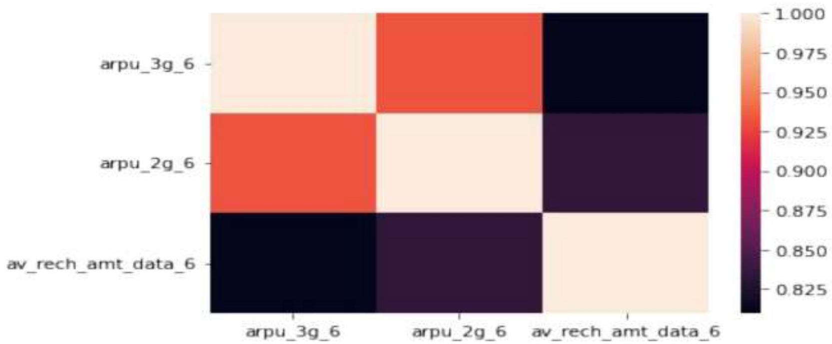
- The below columns have been dropped, since they can be explained from the 'total_rech_data' colum:
  - 'count_rech_2g_6'
  - 'count_rech_3g_6'
  - 'count_rech_2g_7'
  - 'count_rech_3g_7'
  - 'count_rech_2g_8'
  - 'count_rech_3g_8'
  - 'count_rech_2g_9'
  - 'count_rech_3g_9'
- The below columns have been dropped, since they are meaningless:
  - 'fb_user_6'
  - 'fb_user_7','
  - 'fb_user_8'
  - 'fb_user_9'
  - 'night_pck_user_6'
  - 'night_pck_user_7'
  - 'night_pck_user_8
  - 'night_pck_user_9'

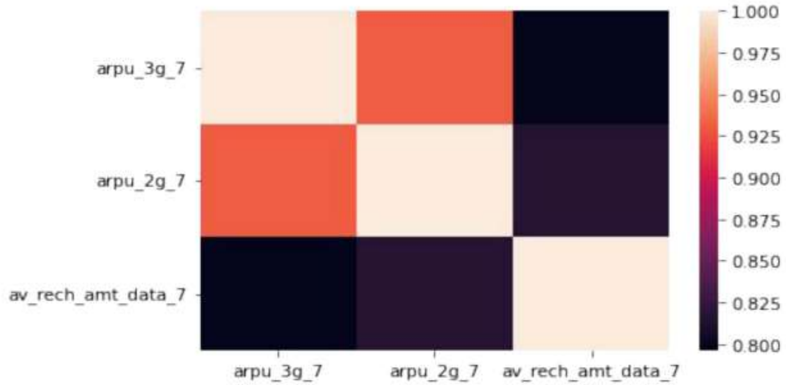
Data Understanding (4)

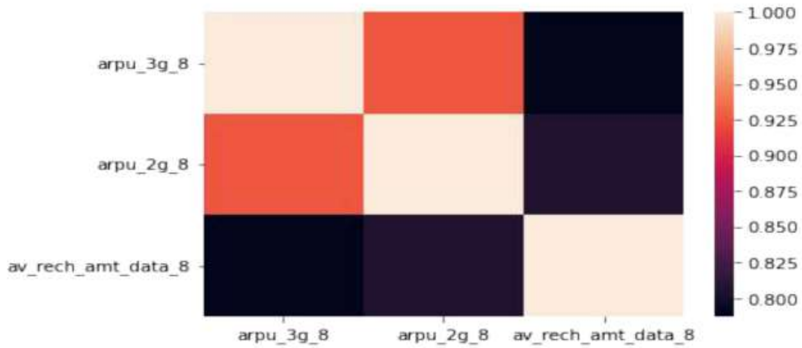
- The below columns are dropped as they have no significance to the data:
  - 'date_of_last_rech_6'
  - 'date_of_last_rech_7'
  - 'date_of_last_rech_8'
  - 'date_of_last_rech_9'

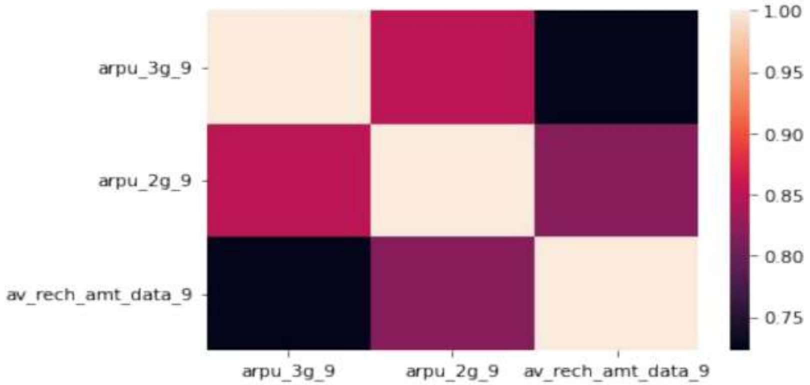
Data Understanding (4)

- The columns 'arpu_3g_6', 'arpu_2g_6', 'arpu_3g_7', 'arpu_2g_7', 'arpu_3g_8', 'arpu_2g_8', 'arpu_3g_9', 'arpu_2g_9' are dropped from the dataset due to high corelation between their respective arpu_* variables in the dataset.



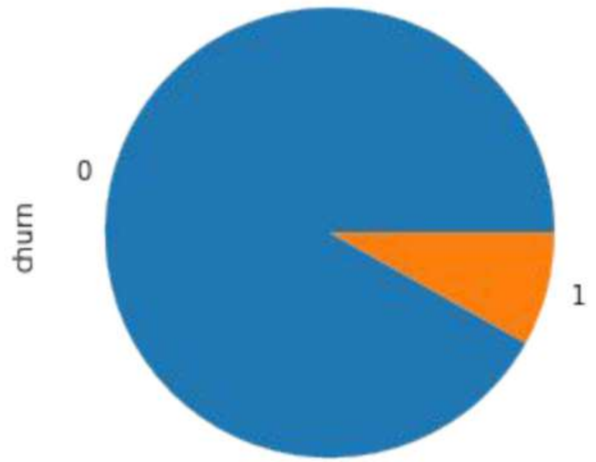






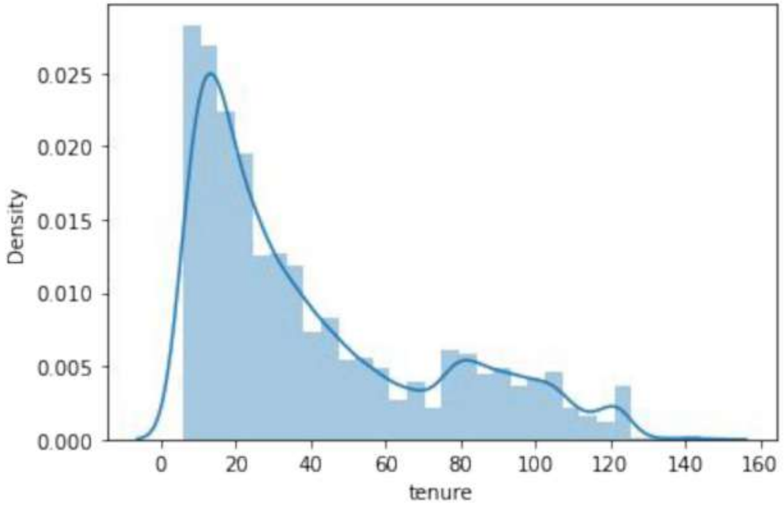
Data Understanding (5)

- The maximum recharge takes place in the month of June and July.
- The 70th quantile value to determine the High Value Customer is: 478.0.
- The data is highly imbalanced. 91% data indicates non-churn customer:



Data Understanding (6)

- Most of the entries are for a tenure of 0-40 days.



Model Finding

The predictive model that we are going to build will serve two purposes:

1. It will be used to predict whether a high-value customer will churn or not, in near future (i.e. churn phase). By knowing this, the company can take action steps such as providing special plans, discounts on recharge etc.
2. It will be used to identify important variables that are strong predictors of churn. These variables may also indicate why customers choose to switch to other networks

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.
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.