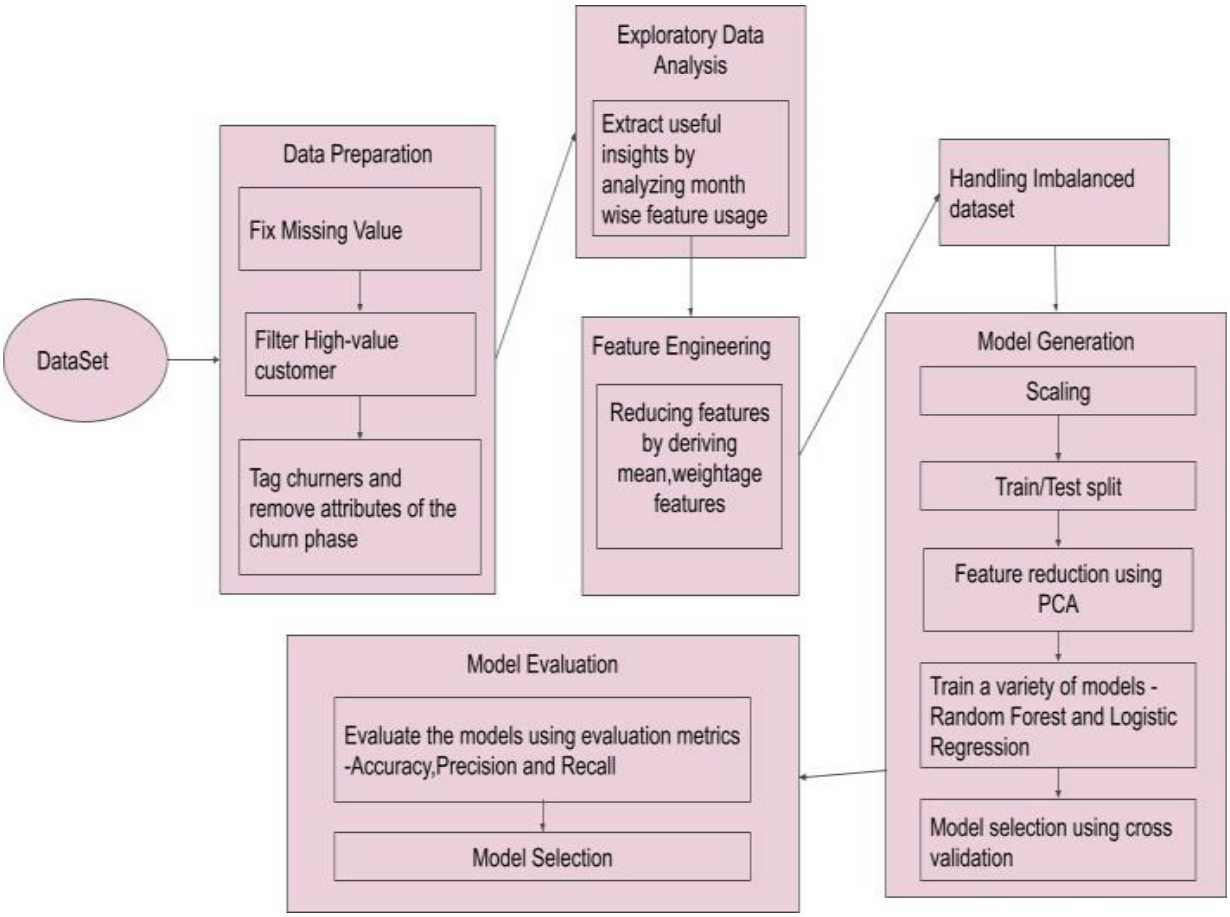
CUSTOMER CHURN PREDICTION

Phase2: Innovation

SYSTEM DESIGN

It is very crucial to make the data useful because unwanted or null values can cause unsatisfactory results or may lead toproducing less accurate results. In the data set, there are a lot of incorrect values and missing values. We analyzed thewhole dataset and listed out only the useful features. The listing of features can result in better accuracy and contains onlyvaluable features as to come up with the specific information like the owner, place of registration, addresseful features.



Architectural Design for Customer Churn Prediction Figure

Feature selection is a crucial step for selecting the required elements from the data set based on the knowledge. The datasetused here consists of many features out of which we chose the needed features, which enable us to

improve performance measurement and are useful for decision-making purposes while remaining will have less importance. The performance of classification increases if the dataset is having only valuable variables and which are highly predictable. Thus having only significant features and reducing the number of irrelevant attributes increases the performance of classification. Many techniques have been proposed for customer churn prediction in the telecommunication industry. Here by using logistic regression, Random Forest and KNN we can predict the probability of a churn i.e., the likelihood of a customer to cancel the subscription and we can evaluate the models using performance metrics like accuracy , precision and recall score.

IMPLEMENTATION

Load the dataset and print the first 5 records of the dataframe to check the loaded dataset. Here mobile number is the unique id column for each customer. It has about a lack of customer records and 226 columns. In order to filter the high value customer records, derived the column of average recharge amount of June and July month(the good phase), take only the records that is more than the 70th percentile of the average recharge amount .Drop the remaining records which is not required and print the count of rows and columns of newly filtered data.

HANDLING MISSING VALUE

In order to fix the missing value in dataset check for the count of missing values in the dataset and list the columns with the missing values. Then pass the dataframe to get_cols_split helper function and get the column categories and pass the month's column list to get_cols_sub_split helper function and get the columns sub-categories.here fb_user and night_pack_user columns are of nominal type 0 and 1. Since missing values could be of another type, imputing them as 2. Missing values for some set of columns seem to be as data not available. So imputing them with 0.Few date columns have

EXPLORATORY DATA ANALYSIS

Due to data imbalance churn rate is low in the overall dataset. In order to fix

it analysis is performed on certain important features column like age on network(AON), incoming calls usage, outgoing calls usage, operator wise calls usage, recharge amount, recharge count, average revenue per user and 2G and 3G. These columns seem to have outliers at the top percentile which is treated using outliers treatment. The outlier treatment is to cap the outliers at the 99th percentile for the above mentioned features column which derives some mandatory features.