Churn customer prediction problem

Retail industries want to grow their customer’s base and profit. In genereal, a larger segment of market obtains a big share of customers who have similar features in age, hobby. Within segment, there may consists of customers in varied financial abilities and habbit which may change over time. Adding new customers is costlier than retaining customer. In addition, understanding reason of churning can help company to adapt the strategy in new environment. For example, knowing why customer leaves can help to retain customers. Predicting the probability of churn scores for paricular customers can help to detect customer who needs more promoting campaign with discount (email bu nott spam).Churn rate can be computed by dividing the number of customers cease/ cancel their activities within a time period by the number of active customers at the start of that period.

Among four main types of churn (<https://www.datascience.com/blog/what-is-a-churn-analysis-and-why-is-it-valuable-for-business>), the most difiicult is voluntary (as well as non-contractual) while others churn may be easier observed (contractual/ autopay and involuntary as credit card expired) . When selected, a churn threshold will be used to detect churn event for a period of inactivity. It is practical to focus on one particular type of churn in company’s strategy. One solution is to use a period of twice longer than the churn period. Given a churn period of 30 days, we need to data for a period of 60 days.

Prediction the probability of churn customer is a main task in many companies’strategies to success. The output of such predictive model indicates the future risk of customer cancellation.

Suggestion:

Step 1: Fill missing data (e.g: mean), remove duplicate. Some information may be ignored but useful such as gender. Other information, especially category may be fill with unknown value.

Note 1: two information to be considered: churn period and churn threshold

Note 2:

|  |  |
| --- | --- |
| Date Diff | Difference between two dates |
| Distinct\_X | Number of distinct entries per user for different category |
| Sum\_X | Aggregate numeric value for each user |
| StDev\_X | StandardDeviation of a specific numeric feature for each user |
| TimeAvg | Average time between transactions for each users |
| Recency | Difference between the last transaction and the last day of a time period each user |
| Advance num | New numeric features |
| Advance cat | New categorical features |

Some numeric data we may be interested to sum/ aggregate as well as standard deviation are Quantity, Value while categorical data can be used to compute number of unique entries (location, address, product category) for each user.

**Timestamp** may be a factor such as high activities (active transaction) during specific holiday season but inactive after that.