

DTH CUSTOMER CHURN

PROJECT NOTES 1



November 28, 2021

sAI kOMARAVOLU

1. Problem Understanding.
   1. Need of the study/project
   2. Understanding business/social opportunity
2. Data Report
   1. Data collection
   2. Visual inspection
   3. Data shape
   4. Data summary
   5. Data dictionary
   6. Data cleansing
3. Exploratory Data Analysis
   1. Univariate Analysis
   2. Bivariate Analysis
   3. Drop un-impacting variables
   4. Data balancing
   5. Data Correlation
   6. Missing value treatment
   7. Categorical variable profiling
   8. Outlier treatment
   9. Dummy variables
4. Business Insights

AGENDA

1. **Need of the study**
   1. A DTH company is facing a lot of competition in the current market, and it has become a challenge to retain existing customers.
   2. Company wants to develop a model through which they can do churn prediction of the accounts and provide segmented offers to the potential churners.
   3. In this company, account churn is a major thing because 1 account can have multiple customers. hence by losing one account the company might be losing more than one customer.
   4. I have been assigned to develop a churn prediction model for this company and provide business recommendations on the campaign.
   5. I need to suggest a very clear idea on the campaign offer, which should not negatively impact the revenue of the company.
2. **Business/Social Opportunity**
   1. Company’s motive is to decrease the Customer’s churn away rate by offering good deals and cashbacks.
   2. Essential features must be identified, that help company to achieve less churn.
   3. Identify similar cluster of customers who churn, this will help to curate offers and retain them.
   4. By studying Company’ data I can help to retain the right customers and accordingly provide lucrative promotional offers to make them permanent customers.

Problem understanding

1. Data collection
   1. Data was provided to me in a excel from customer.
   2. Dataset consists of all customer accounts, and this is not a timeseries dataset.
2. Visual inspection, Data shape & summary
   1. Dataset is presented with 11260 rows and 19 features.
   2. AccountID is the identifier that uniquely points to a customer.
   3. Churn feature is a categorical column, which with value 0 is not churned and 1 is churned.
   4. Dataset has 5 float, 2 int and 12 object typed features, however few object typed columns need to be converted to numeric.
3. Data cleansing
   1. Features have been converted to upper case for the ease of analysis.
   2. Features like Tenure, Account\_user\_count, Rev\_per\_month, rev\_growth\_yoy, coupon\_used\_for\_payment and day\_since\_cc\_connect are converted to numeric.

DATA REPORT

1. Missing Values
   1. Dataset has missing values with most of the features which will be imputed.
   2. Tenure has invalid category “#” which is imputed by max category “1”.
   3. Gender feature has F and M which were imputed to be “Female” and “Male”.
   4. Account\_User\_count has “@” which was imputed to category “4”.
   5. Coupon\_used\_for\_payment has “\*”, “#” and “$” which were imputed to category “1”.
   6. Login\_Device has “&&&&” which was imputed to be “Mobile”.
   7. Day\_Since\_cc\_connect has “$” which was imputed to be category “3”.
   8. Cashback column had many empty cells, which were imputed to value 0.
   9. Cashback column had “$” for 2 cells, which were imputed to value 0.
2. Categorical Features
   1. Payment feature has 5 categories; UPI, Cash on Delivery, E wallet, Credit Card and Debit Card. Maximum account holders have debit card and next maximum is credit card.
   2. Gender feature has 2 categories, Male and Female. Male category accounts for maximum account holders.
   3. Account\_Segment feature has 7 categories, Super category caters to maximum account holders.
   4. Marital\_Status features has 3 categories, Married category caters to maximum account holders.
   5. Login\_Device feature has 2 categories, where in Mobile category has maximum account holders.

exploratory data analysis

Chart, histogram

Description automatically generated

Account\_ID feature is normally distributed, box plot does not hold any importance as this feature is a primary key for all transactions.

Chart

Description automatically generatedCity\_tier feature has customers from City\_tier “1” and less customers from City\_tier “2”.

eDA – Univariate analysis

Graphical user interface

Description automatically generated with medium confidence1. CC\_Contacted\_12m feature is a right skewed, at peak, 20 times customer had contacted customer care. This feature also has outliers with some customers reaching customer cate in excess of 120 times.

2. SERVICE\_SCORE feature is categorical and many customers felt that service is average.

eDA – Univariate analysis

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

1. Service\_Score feature is categorical, and customer’s popular opinion is that the service is average.
2. Tenure feature is a right skewed feature. Most of the customers are new and long time customers are present as outliers.

eDA – Univariate analysis

Chart

Description automatically generated

1. Account\_user\_count feature is categorical and many customers have 4 co-customers using the DTH.
2. Rev\_per\_month feature is normally distributed with outliers, some customers seem to bring huge revenue in comparison with others.
3. Rev\_growth\_yoy is consistent and right skewed with declining growth in the last 12 months.

eDA – Univariate analysis

Chart, histogram

Description automatically generated

1. Coupon\_used\_for\_payment feature is a right skewed feature without consistent peaks. Many customers have used 1 or 2 coupons in the time with DTH. Some customers have heavily used coupons which can be seen as outliers.
2. Day\_since\_cc\_connect is a right skewed feature without consistent peaks and outliers. Some customers have even contacted customer care 40+ times, this could be the reason for churn.

eDA – Univariate analysis

Graphical user interface, square

Description automatically generated with medium confidence

1. Cc\_agent\_score have rated the service to be average.
2. Complain\_12m, early customers have raised many complaints.
3. Cashback is high for customers is normally distributed, some customers have received maximum cashback.

eDA – Univariate analysis

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

1. Customers who hold debit cards are higher than other payment modes.
2. Male customers are higher among customers.
3. Account\_segment super have many customers next to regular\_plys account\_segment.
4. Married customers are higher.

Chart, bar chart

Description automatically generated

1. Customers that are logged in to dth by mobile are higher.

eDA – Univariate analysis

A picture containing histogram

Description automatically generated1. All features are scattered and do not show any correlations with each other.

eDA – BIVARIATE analysis

A picture containing chart

Description automatically generated1. No features have strong correlations with each other.

2. Some features have very weak correlations like coupon\_used\_for\_payment.

eDA – BIVARIATE analysis

1. REV\_GROWTH\_YOY, COUPON\_USED\_FOR\_PAYMENT, CITY\_tIER, CC\_CONTACTED\_LY, SERVICE\_SCORE, REV\_PER\_MONTH and CASHBACK have near to 0 correlation with CHURN, need to investigate to curate more features from them or to drop them.
2. Assumptions:
   1. Outliers were obtained for Cashback, day\_since\_cc\_connect, coupon\_used\_for\_payment, rev\_per\_month, tenure, cc\_contacted\_ly.
   2. These outliers are acceptable and I believe they show the real situation of the customers hold by DTH and should not transform the outliers.
3. Data Balance:
   1. Churned customers are 16.83% and Not-Churned customers are 83.16%. This is highly imbalanced target class.
   2. Usage of SMOTE or K-Means to under or oversample needs to be identified.

eDA – Data IMPORTANCE

1. Yes the data is unbalanced, we need to get the balance for Y to 45% and N to 55% for optimal modelling and prediction.
2. We can use SMOTE to oversample Y or use K-Means to undersample N.
3. Need to perform feature engineering, apply PCA and see combined effect of features on the target class.
4. We can perform LDA to generate features by considering the target class.
5. From the analysis, customer service, features play an important role in churn, many customers have contacted customer care in excess of 120 times.

Business insights