

DTH CUSTOMER CHURN PREDICTION

FINAL SUBMISSION



January 14, 2022

sAI kOMARAVOLU

1. [**Introduction**](#Introduction)
2. [**Exploratory Data Analysis**](#EDA)
3. [**Data Cleaning & Pre Processing**](#Encoding)
4. [**Model Building**](#Modelbuilding)
5. [**Model Validation**](#ModelValidation)
6. [**Interpretation**](#Interpretation)
7. [**Recommendation**](#Recommendation)

index

1. **Need of the study**
   1. 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. Acquiring a new customer can cost 5x time more than retaining an existing customer.
   2. Increasing customer retention by 5% can increase profits from 25-95%.
   3. The success rate of selling a customer you already have is 60-70%, while the success rate of selling a new customer is only 5-20%.
   4. Loyal customers are 5x as likely to repurchase, 5x are likely to forgive, 4x are likely to refer, and 7x as likely to try a new offering from DTH company.
   5. Examples
      1. US companies lose $136.8 billion per year due to avoidable customer switching.
      2. American Express found 33% of customers will consider switch after 1 instance of poor customer service.
   6. Losing a account for DTH is losing more than one customer, that will compound to the loss of losing and gaining a customer.
   7. Above are the facts derived from [internet](https://www.outboundengine.com/blog/customer-retention-marketing-vs-customer-acquisition-marketing/).

introduction

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 to proceed with Linear Algorithms.

Exploratory data analysis - Preface

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 for ACCOUNtid and city tier

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 care more than 120 times.

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

eDA – Univariate analysis for customer service

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 longtime customers are present as outliers.

eDA – Univariate analysis for tenure

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 for account and revenue

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 for coupons and service connect

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 for cashback and service score

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 has many customers next to regular\_plus account\_segment.
4. Married customers are higher.

Chart, bar chart

Description automatically generated

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

eDA – Univariate analysis for GENDER and accout segment

A picture containing histogram

Description automatically generatedFeatures are scattered and do not show strong correlations with each other.

eDA – BIVARIATE analysis

**Churn:**

It is a Binary Classifier and it has states of 0 and 1. Where in, 0 is non-churning account and 1 is churning account. Data set has 83% non churning customers and 17% churning customers. This sets the minority class context of churning customers.

**City Tier:**

Chart, bar chart

Description automatically generated

Tier 1 cities have churn rate of 10% and 2% for Tier 3 customer accounts. Customers catering to Tier1 have direct affect on the DTH company.

eDA – FEature analysis

**Payment:**

Chart, bar chart

Description automatically generated

Customers who use Debit Card as a default transaction mode have higher churn and next in the line of higher churn is users with Credit Card. May be customers would have had significant issues in transacting with cards, probably failed transactions with money being deducted from customer’s account.

**Gender:**

**Chart, bar chart

Description automatically generated**

Male customers churn more than the female customers.

eDA – FEature analysis

**Service Score:**

Chart, bar chart

Description automatically generated

Customers who provided score of 2, 3 and 4 are the high churners for the DTH company.

**Users per account:**

**Chart, bar chart

Description automatically generated**

Accounts with 4 customers have churned higher along with accounts with 3 and 5 customers. This shall impact customer as they lose more with 1 account.

eDA – FEature analysis

**Marital Status:**

Chart, bar chart

Description automatically generated

Single customers have churned more than others, however each segments has it’s share of churn. This could not be a significant factor for churn.

**Yearly complaints:**

**Chart, bar chart

Description automatically generated**

Complained customer form non complained, there’s no significance for Churn, either of them are prone to disconnect the DTH service. I see no significance for this feature on churn prediction.

eDA – FEature analysis

**Login Device:**

Chart, bar chart

Description automatically generated

Customers transacting from Computer and Mobile churned, I see no significance of this feature on churn prediction as both of the features impact the churn.

eDA – FEature analysis

**Account Segment:**

Chart, bar chart

Description automatically generated

Customers opted for Regular Plus segment churned more, could be, DTH company channel package did not suit the audience.

**Service agent score**

**Chart, bar chart

Description automatically generated**

All agent scores have their share of Churn, without significance to a single agent score, this looks like non impacting feature for prediction.

eDA – FEature analysis

A picture containing chart

Description automatically generatedFeatures do not show strong correlations between each other, we can assume to use Naïve Bayes algorithm as there are no correlations.

eDA – COVARIATE analysis

**Dataset Encoding/Outlier Handling/Missing Value analysis**

Chart, bar chart

Description automatically generated

Summary:

1. COMPLAIN\_LY feature has highest invalid records and next to it is the DAY\_SINCE\_CC\_CONNECT. Reason might be the customer never contacted customer care on a yearly basis.
2. Mode is used to impute the values for COMPLAN\_LY.
3. Missing values have been imputed using Mean, Mode and Median appropriately.
4. Outliers have been corrected using IQR method.
5. Categorical Features were Dummy Encoded.
6. No variables were added or removed to understand feature explain ability while model building.

**Outliers:**

Chart

Description automatically generated

Summary:

1. Barring ACCOUNT\_ID and CHURN, some features have outliers.
2. All outliers have been treated using IQR logic.
3. Categorical columns were encoded using One Hot Encoding to proceed with Machine Learning.

**Model Building**

**(Target Variable Imbalance 85% and 15%)**

1. This is a Supervised Learning & Binary Classification problem, hence Logistic Regression, Tree Models will be attempted to produce optimal results.
2. Data imbalance exists with the feature of interest as “Churn-1” which is also a Minority, hence, Recall score will be the parameter to determine best model.
3. SMOTE also will be attempted to visualize the model performance with balanced Target variable.

**Naïve Bayes with Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Naïve Bayes - With Outliers** | | | | | | | | |
| **Accuracy Train: 0.79** | | | | **Accuracy Test: 0.79** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 0.92 | 0.91 | 0.82 | 0.84 | 0.87 | 0.87 | 7023 | 2341 |
| 1 | 0.42 | 0.42 | 0.63 | 0.59 | 0.5 | 0.5 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
|  |  |

Summary:

1. With Outliers, Naïve Bayes has same accuracy score of 79% for both train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 63% and Test data is 59%.
3. On train data 63 of 100 times algorithm can predict customer churned, on test data only 59 times of 100 algorithm is able to predict.
4. This score is not the best one to perform prediction, this algorithm will not be considered.

**Naïve Bayes without Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Naïve Bayes - Without Outliers** | | | | | | | | |
| **Accuracy Train: 0.77** | | | | **Accuracy Test: 0.78** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 0.93 | 0.93 | 0.79 | 0.80 | 0.86 | 0.86 | 7023 | 2341 |
| 1 | 0.41 | 0.41 | 0.70 | 0.70 | 0.51 | 0.52 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
|  |  |

Summary:

1. Without Outliers, Naïve Bayes has same accuracy score of 77% and 78% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 70% and Test data is 70%.
3. On train data 70 of 100 times algorithm can predict customer churned, on test data only 70 times of 100 algorithm is able to predict.
4. Train and Test sensitivity is good having same score for both, will be considered as a best model, and further evaluated.

**Linear Discriminant Analysis with Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Linear Discriminant Analysis - With Outliers** | | | | | | | | |
| **Accuracy Train: 0.87** | | | | **Accuracy Test: 0.86** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 0.98 | 0.88 | 0.97 | 0.97 | 0.93 | 0.92 | 7023 | 2341 |
| 1 | 0.71 | 0.69 | 0.40 | 0.35 | 0.51 | 0.46 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
|  |  |

Chart, bar chart

Description automatically generated

Summary:

1. With Outliers, Linear Discriminant Analysis has accuracy score of 87% and 86% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 40% and Test data is 35%.
3. On train data 40 of 100 times algorithm can predict customer churned, on test data only 35 times of 100 algorithm is able to predict.
4. Train and Test sensitivity is poor, and this model will not be considered further.
5. Feature Importance:
   1. As per LDA Account\_User\_Count and Complaints per year are the best features that impact the model.
   2. On the contrary, Performance metrics are not impressive for LDA and will not describe the feature importance further.

**Linear Discriminant Analysis without Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Linear Discriminant Analysis - Without Outliers** | | | | | | | | |
| **Accuracy Train: 0.79** | | | | **Accuracy Test: 0.79** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 0.90 | 0.90 | 0.97 | 0.97 | 0.93 | 0.93 | 7023 | 2341 |
| 1 | 0.76 | 0.75 | 0.49 | 0.46 | 0.59 | 0.57 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
|  |  |

Summary:

1. Without Outliers, Linear Discriminant Analysis has accuracy score of 79% and 79% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 49% and Test data is 46%.
3. On train data 49 of 100 times algorithm can predict customer churned, on test data only 46 times of 100 algorithm is able to predict.
4. Train and Test sensitivity is poor and this model will not be considered further.

**KNN With Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: KNN - With Outliers** | | | | | | | | |
| **Accuracy Train: 0.84** | | | | **Accuracy Test: 0.81** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 0.86 | 0.84 | 0.98 | 0.97 | 0.92 | 0.90 | 7023 | 2341 |
| 1 | 0.69 | 0.32 | 0.18 | 0.07 | 0.29 | 0.11 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
|  |  |

Summary:

1. With Outliers, KNN algorithm has accuracy score of 84% and 81% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 18% and Test data is 7%.
3. Train and Test sensitivity is poor and this model will not be considered further.

**KNN Without Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: KNN - Without Outliers** | | | | | | | | |
| **Accuracy Train: 0.85** | | | | **Accuracy Test: 0.81** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 0.86 | 0.84 | 0.98 | 0.97 | 0.92 | 0.90 | 7023 | 2341 |
| 1 | 0.70 | 0.32 | 0.19 | 0.07 | 0.30 | 0.11 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
|  |  |

Summary:

1. Without Outliers, KNN algorithm has accuracy score of 85% and 81% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 19% and Test data is 7%.
3. Train and Test sensitivity is poor and this model will not be considered further.

**Decision Tree with Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Decision Tree - With Outliers** | | | | | | | | |
| **Accuracy Train: 1** | | | | **Accuracy Test: 0.92** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 1 | 0.96 | 1 | 0.95 | 1 | 0.96 | 7023 | 2341 |
| 1 | 1 | 0.77 | 1 | 0.79 | 1 | 0.78 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
|  |  |

Summary:

1. With Outliers, Decision Tree has accuracy score of 100% and 92% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 100% and Test data is 79%.
3. On train data 100 of 100 times algorithm can predict customer churned, on test data only 79 times of 100 algorithm is able to predict.
4. Deviation of 20% for recall score between is train and test is not a good parameter for a prediction model.
5. However, Train and Test sensitivity is good having same score for both, will be considered as a best model, and further evaluated.

**Decision Tree without Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Decision Tree - Without Outliers** | | | | | | | | |
| **Accuracy Train: 1** | | | | **Accuracy Test: 0.94** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 1 | 0.96 | 1 | 0.95 | 1 | 0.96 | 7023 | 2341 |
| 1 | 1 | 0.82 | 1 | 0.80 | 1 | 0.81 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
|  |  |

Summary:

1. Without Outliers, Decision Tree has accuracy score of 100% and 94% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 100% and Test data is 80%.
3. On train data 100 of 100 times algorithm can predict customer churned, on test data only 80 times of 100 algorithm is able to predict.
4. Marginally improved by 1% after addressing outliers. Deviation of 20% for recall score between is train and test is not a good parameter for a prediction model.
5. However, Train and Test sensitivity is good having same score for both, will be considered as a best model, and further evaluated.

**Random Forest with Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Random Forest - With Outliers** | | | | | | | | |
| **Accuracy Train: 1** | | | | **Accuracy Test: 0.96** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 1 | 0.96 | 1 | 1 | 1 | 0.98 | 7023 | 2341 |
| 1 | 1 | 0.99 | 1 | 0.80 | 1 | 0.89 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
|  |  |

**AUC/ROC:**

A picture containing graphical user interface

Description automatically generated

This model has 99% AUC, it explains 99 times this model will be effective to predict a Churning customer.

**Feature Importance:**

Chart, bar chart

Description automatically generated

**Tenure is the feature with huge impact on the customer churn**, this explains that customers with shorter tenure impact DTH company as No.of customers join and leave in a short rate, causing more onboarding and retaining losses.

**Summary:**

1. With Outliers, Random Forest algorithm has accuracy score of 100% and 96% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 100% and Test data is 80%.
3. On train data 100 of 100 times algorithm can predict customer churned, on test data only 80 times of 100 algorithm is able to predict.
4. Deviation of 20% for recall score between is train and test are not a good parameter for a prediction model.
5. However, Train and Test sensitivity is good having same score for both, will be considered as a best model, and further evaluated.
6. **Tenure feature impacts the predictability of the DTH customer.**

**Random Forest without Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Random Forest - Without Outliers** | | | | | | | | |
| **Accuracy Train: 1** | | | | **Accuracy Test: 0.96** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 1 | 0.96 | 1 | 1 | 1 | 0.98 | 7023 | 2341 |
| 1 | 1 | 0.99 | 1 | 0.79 | 1 | 0.88 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
| Chart  Description automatically generated |  |

Summary:

1. Without Outliers, Random Forest algorithm has accuracy score of 100% and 96% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 100% and Test data is 79%.
3. On train data 100 of 100 times algorithm can predict customer churned, on test data only 79 times of 100 algorithm is able to predict.
4. No impact of Outlier treatment is observed. Deviation of 20% for recall score between is train and test are not a good parameter for a prediction model.
5. However, Train and Test sensitivity is good having same score for both, will be considered as a best model, and further evaluated.

**Logistic Regression with Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Logistic Regression - With Outliers** | | | | | | | | |
| **Accuracy Train: 0.88** | | | | **Accuracy Test: 0.88** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 0.90 | 0.89 | 0.97 | 0.97 | 0.93 | 0.93 | 7023 | 2341 |
| 1 | 0.78 | 0.74 | 0.44 | 0.40 | 0.56 | 0.52 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
|  |  |

Summary:

1. With Outliers, Logistic Regression algorithm has accuracy score of 88% and 88% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 44% and Test data is 40%.
3. On train data 44 of 100 times algorithm can predict customer churned, on test data only 40 times of 100 algorithm is able to predict.
4. Train and Test sensitivity is poor, and this model will not be considered further.

**Logistic Regression without Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Logistic Regression - Without Outliers** | | | | | | | | |
| **Accuracy Train: 0.89** | | | | **Accuracy Test: 0.89** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 0.91 | 0.90 | 0.97 | 0.97 | 0.94 | 0.94 | 7023 | 2341 |
| 1 | 0.78 | 0.78 | 0.51 | 0.49 | 0.62 | 0.60 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
|  |  |

Summary:

1. With Outliers, Logistic Regression algorithm has accuracy score of 89% and 89% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 51% and Test data is 49%.
3. On train data 51 of 100 times algorithm can predict customer churned, on test data only 49 times of 100 algorithm is able to predict.
4. Though recall improved after outlier treatment, Train and Test sensitivity is poor and this model will not be considered further.

**Model Tuning**

* I have used Ensemble Modelling to try other models for optimal performance, also GridSearchCv on Radom Forest was applied, and it did not result in significant upshift in Performance Measures.
* I have used SMOTE to increase the density of churning customer to check the model performance with a balanced data set. Random Forest has no effect on the SMOTE data set.

**Ensemble Modelling:**

**AdaBoost With Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Ada Boost - With Outliers** | | | | | | | | |
| **Accuracy Train: 0.89** | | | | **Accuracy Test: 0.90** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 0.92 | 0.92 | 0.96 | 0.96 | 0.94 | 0.94 | 7023 | 2341 |
| 1 | 0.76 | 0.77 | 0.60 | 0.59 | 0.67 | 0.67 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
|  |  |

Summary:

1. With Outliers, AdaBoost algorithm has accuracy score of 89% and 90% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 60% and Test data is 59%.
3. On train data 60 of 100 times algorithm can predict customer churned, on test data only 59 times of 100 algorithm is able to predict.
4. Train and Test score is average, and this model will be evaluated after removing outliers.

**AdaBoost Without Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Ada Boost - Without Outliers** | | | | | | | | |
| **Accuracy Train: 0.89** | | | | **Accuracy Test: 0.90** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 0.92 | 0.92 | 0.96 | 0.97 | 0.94 | 0.94 | 7023 | 2341 |
| 1 | 0.77 | 0.79 | 0.58 | 0.60 | 0.66 | 0.68 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
|  |  |

Summary:

1. Without Outliers, AdaBoost algorithm has accuracy score of 89% and 90% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 58% and Test data is 60%.
3. On train data 58 of 100 times algorithm can predict customer churned, on test data only 60 times of 100 algorithm is able to predict.
4. Train and Test score is average, and this model will not be considered for Prediction.

**Gradient Boost with Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Gradient Boost - With Outliers** | | | | | | | | |
| **Accuracy Train: 0.91** | | | | **Accuracy Test: 0.91** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 0.93 | 0.93 | 0.97 | 0.97 | 0.95 | 0.95 | 7023 | 2341 |
| 1 | 0.82 | 0.80 | 0.63 | 0.62 | 0.71 | 0.70 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
| Chart, treemap chart  Description automatically generated | Chart, treemap chart  Description automatically generated |

Summary:

1. With Outliers, Gradient Boost algorithm has accuracy score of 91% and 91% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 63% and Test data is 62%.
3. On train data 63 of 100 times algorithm can predict customer churned, on test data only 62 times of 100 algorithm is able to predict.
4. Train and Test score is average, and this model will be evaluated after removing outliers.

**Gradient Boost without Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Gradient Boost - Without Outliers** | | | | | | | | |
| **Accuracy Train: 0.92** | | | | **Accuracy Test: 0.91** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 0.93 | 0.93 | 0.97 | 0.97 | 0.95 | 0.95 | 7023 | 2341 |
| 1 | 0.83 | 0.81 | 0.62 | 0.62 | 0.71 | 0.70 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
|  |  |

Summary:

1. Without Outliers, AdaBoost algorithm has accuracy score of 92% and 91% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 62% and Test data is 62%.
3. On train data 62 of 100 times algorithm can predict customer churned, on test data only 62 times of 100 algorithm is able to predict.
4. Train and Test score is average, and they are same, further evaluation is required to consider this model for prediction.

**Bagging With Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Bagging - With Outliers** | | | | | | | | |
| **Accuracy Train: 1** | | | | **Accuracy Test: 0.96** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 1 | 0.97 | 1 | 0.99 | 1 | 0.98 | 7023 | 2341 |
| 1 | 1 | 0.95 | 1 | 0.83 | 1 | 0.89 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
|  |  |

Summary:

1. With Outliers, Bagging algorithm has accuracy score of 100% and 96% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 100% and Test data is 83%.
3. On train data 100 of 100 times algorithm can predict customer churned, on test data only 83 times of 100 algorithm is able to predict.
4. Deviation of 17% for recall score between is train and test are not a good parameter for a prediction model.
5. However, Train and Test sensitivity is good having same score for both, will be considered as a best model, and further evaluated.

**Bagging Without Outliers**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Bagging - Without Outliers** | | | | | | | | |
| **Accuracy Train: 1** | | | | **Accuracy Test: 0.96** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 1 | 0.97 | 1 | 0.99 | 1 | 0.98 | 7023 | 2341 |
| 1 | 1 | 0.95 | 1 | 0.83 | 1 | 0.89 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
| Chart  Description automatically generated | Chart, treemap chart  Description automatically generated |

Summary:

1. Without Outliers, Bagging algorithm has accuracy score of 100% and 96% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Outlier treatment has no effect.
3. Recall score for Train data is 100% and Test data is 83%.
4. On train data 100 of 100 times algorithm can predict customer churned, on test data only 83 times of 100 algorithm is able to predict.
5. Deviation of 17% for recall score between is train and test are not a good parameter for a prediction model.
6. However, Train and Test sensitivity is good having same score for both, will be considered as a best model, and further evaluated.

**Model Validation and Comparison**

**Evaluation Parameters:**

* This is a Binary Classification problem with Minority class prediction. **Precision/Recall will be used** to define best model, i.e., **Recall score shall explain the predictive capabilities of the model to identify minority class**, in this case minority class is a Churning customer with value 1.
* **Along with Recall, I will use Accuracy and AUC score**. AUC score defines how well model can identify a churning customer with value 1.
* **Random Forest Classifier** with 80% recall score and 99% AUC is the best model.



**Final Interpretation**

**Chart, bar chart

Description automatically generated**

* Random Forest Classifier is the best model to predict customer churn.
* **Random Forest Classifier has 80% churn predictability, and it is not affected by data imbalance/outliers.**
* **Random forest classifier determines Tenure to be significant feature** that defines customer churn.
* EDA shows high number of customers who have younger Tenure and very less older Tenure customers.
* Social impact of losing a young tenure customer puts 5x cost than retaining him, in DTH company’s case, young-tenure customers churning is found to be major problem of the revenue loss.

**Recommendation**

* **Tenure is the feature with huge impact on the customer churn**, this explains that customers with shorter tenure impact DTH company as No.of customers join and leave in a short rate, causing more onboarding and retaining losses.
* **DTH company must attract young tenure customers with a priority service and curated, custom DTH packages.**
* Once a customer expresses his concern to customer care, he/she must be pacified with younger Tenure customers treated well with extra channels of HD upgrade at nominal costs.
* If a customer decides to leave DTH, he/she must be provided with a selective and competitive offer than the competitor, there by reducing the 5x cost of obtaining another customer at a small cost.

**SMOTE to verify model performance on Balanced Data**

**SMOTE on Naïve Bayes**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Naïve Bayes with SMOTE** | | | | | | | | |
| **Accuracy Train: 80** | | | | **Accuracy Test: 0.68** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 0.86 | 0.88 | 0.71 | 0.71 | 0.78 | 0.79 | 7023 | 2341 |
| 1 | 0.75 | 0.27 | 0.88 | 0.53 | 0.81 | 0.36 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
|  |  |

Summary:

1. Naïve Bayes has same accuracy score of 80% and 68% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 88% and Test data is 53%.
3. On train data 88 of 100 times algorithm can predict customer churned, on test data only 53 times of 100 algorithm is able to predict.
4. Train and Test sensitivity has been degraded, SMOTE did not help improving models’ efficiency, hence it will not be a prediction model.

**SMOTE on Random Forest**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm: Random Forest with SMOTE** | | | | | | | | |
| **Accuracy Train: 1** | | | | **Accuracy Test: 0.95** | | | | |
| **Churn** | **Precision** | | **Sensitivity/Recall** | | **F1-Score** | | **Support** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 0 | 1 | 0.96 | 1 | 0.99 | 1 | 0.97 | 7023 | 2341 |
| 1 | 1 | 0.92 | 1 | 0.82 | 1 | 0.86 | 1422 | 474 |

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| **Train** | **Test** |
|  |  |

Summary:

1. Random Forest algorithm has accuracy score of 100% and 95% respectively for train and test data. As we have imbalanced data, we will not use this metric to determine if the model is a good model.
2. Recall score for Train data is 100% and Test data is 80%.
3. On train data 100 of 100 times algorithm can predict customer churned, on test data only 82 times of 100 algorithm is able to predict.
4. Deviation of 18% for recall score between is train and test are not a good parameter for a prediction model.
5. However, Train and Test sensitivity increased by 2% and it is good to see this score improving after SMOTE, will be considered as a best model for Prediction.