IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, June 2024

Customer Churn Prediction using Machine Learning

Sakshat Salekar and Riddhesh Awade
Student, Department MCA
Late Bhausaheb Hiray S S Trust's Hiray Institute of Computer Application, Mumbai, India

Abstract: *Companies have to fight hard to lure in new customers from their suppliers. Client retention is a trendy issue for investigation since it directly impacts a business's revenue; early discovery of client churn allows organizations to take proactive steps to retain consumers. Thus, through customer retention programs, all businesses could employ a range of strategies to recognize their clientele early on. Consequently, this study tries to advise on the ideal machine- learning technique for early client churn prediction. All customer information dating back around nine months prior to the churn is included in the data used in this research. Anticipating the reactions of current clients is the aim in order to retain them. Several algorithms, including k- nearest neighbors, random forest, logistics regression etc have been tested in this work. As theThe aforementioned algorithms had accuracy rates of 78.1%, 82.6%, 83.9%, and 82.9%, respectively. By analyzing these algorithms and debating the best of the four from various angles, we have obtained the most efficient outcomes.*

Keywords: Customer Churn

I. INTRODUCTION

Churning, in marketing terms, refers to the number of customers who stopped using a particular product. Always the churn rate must be low. Customer churning is common with anyproduct when there are multiple options for a single problem. Usually, customers will churn when they face any difficulties or disappointments in the services rendered by the product. The churn rate is usually measured for a specific time. Any organization's primary motive should be satisfying customers and retaining existing customers. Retaining existing customers is equally important as gathering new customers. Customer churn prediction is the most important issue in adopting an industry's product. One of the biggest problems businesses have is managing client turnover, particularly for those who provide subscription-based services. Losing clients due to shifting preferences, improper customer relationship management, moving, and other factors is known as customer churn, also known as customer attrition. Businesses that are able to accurately forecast customer attrition can identify and target customers who are most likely to leave, giving them superior services. Therefore, in today's digital economy, a churn prediction model is a must. It is possible for a business to increase income and maintain a highclient retention rate.One of the biggest problemsbusinesses have is managing client turnover, particularly for those who provide subscription- based services. Customer loss, often known as customer attrition or customer chur. One of the biggest problems businesses have is managing client turnover, particularly for those who provide subscription-based services. Losing clients due to shifting preferences, improper customer relationship management, moving, and other factors is known as customer churn, also known as customer attrition. Businesses that are able to accurately forecast customer attrition can identify and target customers who are most likely to leave, giving them superior services. Therefore, in today's digital economy, a churn prediction model is a must. It is possible for a business to increase income and maintain a high client retention rate.One of the biggest problems businesses have is managing client turnover, particularly for those who provide subscription-based services. Customer loss,often known as customer attrition or customer churn, is brought on by

II. METHODOLOGY USE

The system involved in the analysis ofcustomer churning uses four differentalgorithms mentioned below.

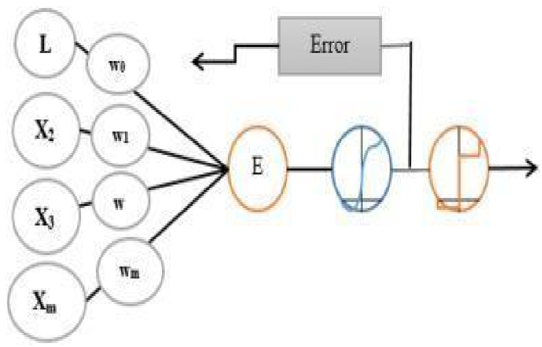
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, June 2024

- Logistic Regression
- Decision Tree
- Random Forest Classifier
- Support Vector Classifier

1 . LOGISTIC REGRESSION:

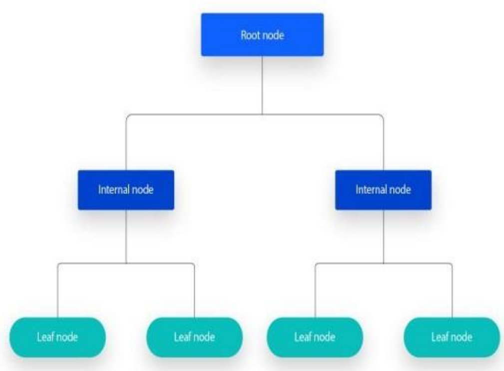
Logistic regression is a statistical method that isused for building machine learning models where the dependent variable is dichotomous:

i.e. binary. Logistic regression is used to describe data and the relationship between one dependent variable and one or more independentvariables.



2. DECISION TREE

A decision tree is a non-parametric supervisedlearning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of aroot node, branches, internal nodes and leaf nodes.

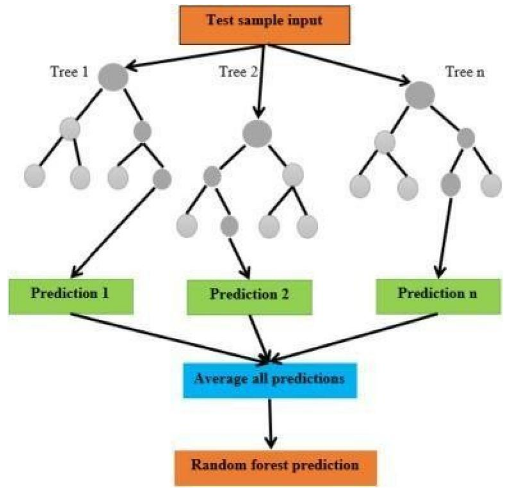


3. RANDOM FOREST CLASSIFIER

Random forest is a commonly-used machine learning algorithm trademarked by Leo Breimanand Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification regression problems.

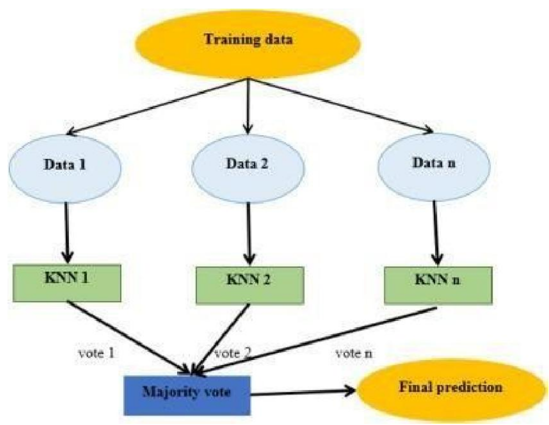
DOI: 10.48175/IJARSCT-18969

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, June 2024

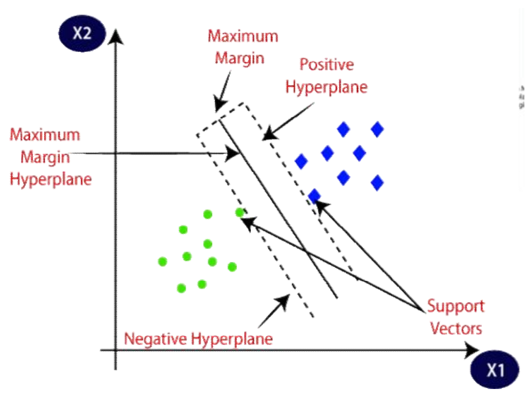


4. KNN CLASSIFIER

As we saw above, the KNN algorithm can be used for both classification and regression problems. The KNN algorithm uses 'feature similarity' to predict the values of any new data points. This means that the new point is assigneda value based on how closely it resembles the points in the training set.



5. SUPPORT VECTOR CLASSIFIER



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, June 2024

A support vector machine (SVM) is a type of supervised learning algorithm used in machine learning to solve classification and regression tasks; SVMs are particularly good at solving binary classification problems, which require classifying the elements of a data set intotwo groups.

III. RESULT AND DISCUSSION

The results were obtained using Python byutilizing the Jupyter Libraries from Anaconda. The various libraries used include numpy, pandas, matplotlib and seaborn.The results obtained in comparing the performance of the various algorithmsare narrated step by step.

TEST AND TRAIN DATASET SPLIT:

The customer churn dataset is split intotraining and testing data

:th	number_vmail_messages	total_day_minutes	total_day_calls	total_day_charge	total_eve_minutes	total_eve_calls	total_e
00	4250.000000	4250.000000	4250.000000	4250.000000	4250.000000	4250.000000	4:
:35	7.631765	180.259600	99.907294	30.644682	200.173906	100.176471	
01	13.439882	54.012373	19.850817	9.182096	50.249518	19.908591	
00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
00	0.000000	143.325000	87.000000	24.365000	165.925000	87.000000	
00	0.000000	180.450000	100.000000	30.680000	200.700000	100.000000	
00	16.000000	216.200000	113.000000	36.750000	233.775000	114.000000	
00	52.000000	351.500000	165.000000	59.760000	359.300000	170.000000	

TRAIN INFORMATION:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4250 entries, 0 to 4249
Data columns (total 20 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   state                          4250 non-null   object
 1   account_length                 4250 non-null   int64
 2   area_code                      4250 non-null   object
 3   international_plan             4250 non-null   object
 4   voice_mail_plan                4250 non-null   object
 5   number_vmail_messages          4250 non-null   int64
 6   total_day_minutes              4250 non-null   float64
 7   total_day_calls                4250 non-null   int64
 8   total_day_charge               4250 non-null   float64
 9   total_eve_minutes              4250 non-null   float64
 10  total_eve_calls                4250 non-null   int64
 11  total_eve_charge               4250 non-null   float64
 12  total_night_minutes            4250 non-null   float64
 13  total_night_calls              4250 non-null   int64
 14  total_night_charge             4250 non-null   float64
 15  total_intl_minutes             4250 non-null   float64
 16  total_intl_calls               4250 non-null   int64
 17  total_intl_charge              4250 non-null   float64
 18  number_customer_service_calls  4250 non-null   int64
 19  churn                          4250 non-null   object
dtypes: float64(8), int64(7), object(5)
memory usage: 664.2+ KB
```


DOI: 10.48175/IJARSCT-18969
ISSN 2581-9429 IJARSCT

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, June 2024

TEST INFORMATION:

```
test.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 750 entries, 0 to 749
Data columns (total 20 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   id                            750 non-null    int64
 1   state                         750 non-null    object
 2   account_length                750 non-null    int64
 3   area_code                     750 non-null    object
 4   international_plan            750 non-null    object
 5   voice_mail_plan               750 non-null    object
 6   number_vmail_messages         750 non-null    int64
 7   total_day_minutes             750 non-null    float64
 8   total_day_calls               750 non-null    int64
 9   total_day_charge              750 non-null    float64
 10  total_eve_minutes             750 non-null    float64
 11  total_eve_calls               750 non-null    int64
 12  total_eve_charge              750 non-null    float64
 13  total_night_minutes           750 non-null    float64
 14  total_night_calls             750 non-null    int64
 15  total_night_charge            750 non-null    float64
 16  total_intl_minutes            750 non-null    float64
 17  total_intl_calls              750 non-null    int64
 18  total_intl_charge             750 non-null    float64
 19  number_customer_service_calls 750 non-null    int64
dtypes: float64(8), int64(8), object(4)
```

PREDICTION OF LOGISTIC REGRESSION
Accuracy : 0.8541176470588235
Overall_Error_Rate :0.14588235294117646
Precision : 0.2222222222222222
Sensitivity Recall :0.05504587155963303
Specificity : 0.97165991902834F1 Score : 0.08823529411764706

PREDICTION OF DECISION TREE:
Accuracy : 0.9176470588235294
Overall_Error_Rate :0.08235294117647063
Precision : 0.6666666666666666
Sensitivity Recall :0.7155963302752294
Specificity : 0.9473684210526315F1 Score : 0.6902654867256638

PREDICTION OF RANDOM FORESTCLASSIFIER:
Accuracy : 0.9552941176470588
Overall_Error_Rate :0.04470588235294115
Precision : 0.9382716049382716
Sensitivity Recall :0.6972477064220184
Specificity : 0.9932523616734144F1 Score : 0.8

PREDICTION OF KNN CLASSIFIER:
Accuracy : 0.8929411764705882
Overall_Error_Rate :0.10705882352941176
Precision : 0.7045454545454546
Sensitivity Recall :0.28440366972477066
Specificity : 0.9824561403508771F1 Score : 0.40522875816993464

IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 3, June 2024

PREDICTION OF SUPPORT VECTORCLASSIFIER:

Accuracy : 0.8729411764705882
Overall_Error_Rate :0.12705882352941178
Precision : 1.0 Sensitivity Recall :0.009174311926605505
Specificity : 1.0
F1 Score : 0.018181818181818184

IV. CONCLUSION

To determine which Random Forest Classifier model is better, we need to consider the specificcontext and requirements of our problem because the choice of the "best" model can depend on various factors. Here are some key points to consider:

Accuracy: The Random Forest Classifier using the 'corr' features has a higher accuracy (0.955) compared to the one using mutual information (0.918). Higher accuracy generally indicates better overall performance, but it might not be the sole criterion for selecting the best model.

Precision: The 'corr' features model has a higher precision (0.938) compared to the mutual information model (0.783). Precision is crucial if minimizing false positives is a top priority. In some applications, like medical diagnoses, precision is of utmost importance.

Sensitivity (Recall): The 'corr' features model has a higher sensitivity (0.697) compared to the mutual information model (0.495). Sensitivity is essential when correctly identifying positive cases (e.g., detecting diseases) is critical. A higher sensitivity means fewer false negatives.

Specificity: The 'corr' features model has a higher specificity (0.993) compared to the mutual information model (0.980). Specificity is essential when minimizing false positives is a priority, especially in applications where the cost of false positives is high.

F1 Score: The 'corr' features model has a higher F1 score (0.800) compared to the mutual information model (0.607). The F1 score is the harmonic mean of precision and recall and provides a balanced measure of a model's performance.

ROC Area: The 'corr' features model has a higher ROC Area (0.85) compared to the mutual information model (0.74). A higher ROC Area indicates a better ability to distinguish between positive and negative cases.

In summary, both models have their strengths and weaknesses:

If overall accuracy is the primary concern and false positives and false negatives are of roughly equal concern, the 'corr' features model might be preferred.

If minimizing false positives is more critical, the 'corr' features model with higher precision and specificity should be considered.

If correctly identifying positive cases (high sensitivity) is of utmost importance, and we can tolerate some false positives, the 'corr' features model might still be the better choice.

REFFERENCES

[1]. https://neptune.ai/blog/how-to- implement-customer-churn- prediction
[2]. https://www.projectpro.io/artic le/churn-models/709
[3]. h ttps://userpilot.com/blog/chur n-prediction/
[4]. https://towardsdatascience.com/churn-prediction-with-machine-learning-ca955d52bd8c
[5]. https://github.com/topics/custo mer-churn-prediction
[6]. https://github.com/otavio-s-s/data_science/blob/master/mac hine%20learning%20for%20churn%20prediction.ipynb
[7]. https://www.leewayhertz.com/ai- and-ml-in-customer-churn- prediction/
[8]. https://www.diva- portal.org/smash/get/diva2:1574 424/fulltext01.pdf
[9]. https://addepto.com/blog/machin e-learning-predict-reduce- customer-churn/