**Mini Project Report on**



**CUSTOMER CHURN PREDICTION**



**Submitted in partial fulfilment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Student Name: DISHA JAIN** **University Roll No.:2021878**

***Under the Mentorship of***

**--**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**January-2024**



**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“Customer Churn Prediction”** in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **\_\_,** Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name: Disha Jain University Roll no: 2021878

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | 4-5 |
| Chapter 2 | Literature Survey | 6-7 |
| Chapter 3 | Methodology | 8-10 |
| Chapter 4 | Result and Discussion | 11-12 |
| Chapter 5 | Conclusion and Future Work | 13 |
|  | References | 14 |

**Chapter 1**

**Introduction**

In the following sections, a brief introduction and the problem statement for the work has been included.

* 1. **Introduction**

Customers are the foremost requirement or the demand for a successful business. Entrepreneurship is the outstanding domain which is currently outgrowing in this market. Building new business or outstanding with the existing ones demands some rules that need to be seriously follow to introduce a brand in a market. Everything is made for customers; every effort is done to attract them. Different ways of attracting new ideal potential customers to avail different services and products or to retaining the existing ones and gaining their loyalty, so that they won’t churn from the services and the products they are using should be one of the major concerns of a business or a startup. Businesses put most of the efforts to gain more and more customers and also make the customers not to churn. Now, what is churning, why it is one of the important terms in the world of businesses. Churning is the process of switching from one service to another. So, customers switching from one service that they currently using or is indulging in to another service or product due to their not satisfying services, weak customer retention, or maybe they are more interested in the offers or the benefits which other service is providing. So, their may be any number of reasons because of which customer wishes to churn from one service to another. Churn is a key driver for EBITDA (Earnings Before Interest, Taxes, Depreciation and Amortization. A churned customer gives less revenue or no revenue and increases market competition. Churn prediction means detecting which customers are likely to leave a service or to cancel a subscription to a service. It is a critical prediction for many businesses because acquiring new clients often costs more than retaining existing ones. One’s one is able to predict that customer is on the verge to churn, then they can show exact market strategy, use different ideas, market different ideas, offers different schemes to convince customer not to churn from the service he is using. This can only be possible if one is able to predict this beforehand from the behaviour of customer whether he will churn or not. There are different customers with different behaviour and have different reasons to churn for the current using service. Churning in marketing term , refers to when a customer stopped using a product. So, a churn rate of a business should be low enough as to avoid low EBITDA value. The more competition there is, the higher rates companies can charge. Firms in low-competition markets may charge lower rates than highly competitive markets. The number of services associations can provide also affects the fees charged to the associations. The more services the company provides, the higher the company can charge. If is one of the challenging tasks for the business to attract ideal potential customers to use their services and retain the loyalty of the existing customer to stop them from churning. So, machine learning once again came as a helping hand to build different model which on various insights, various criteria can predict that whether the respective customer will churn or not. This model can be proved to be of great heap if used precisely, helping businesses to retain their customers.

* 1. **Problem Statement**

We have tried making a model for helping business in their customer retention, i.e., customer churn prediction model. We have been given with a dataset in which their different criteria like senior citizen, gender, partnership, internet service, online service, online backup, and many more of a telecom industry. We will first be taking the insights from the dataset, cleaning it, and then training it on the criteria’s due to which the prediction will be affected the most. We will be following the complete model making procedure from data cleaning to taking insights to training the data to testing. We have to make the nearest precise churn prediction model which can result in prediction of churn rate as well and will be helpful in the real world.

**Chapter 2**

**Literature Survey**

In this chapter some of the major existing work in these areas has been reviewed.

Customer churning is not a completely unknown field. This is the most concerning field which already has the hands-on by some researchers. But since it is not that easy to make a complete precise model to predict whether the customer will churn from the availing services, i.e., whether customer will switch from one service to another or not. So, the research is ongoing till now. With the constant repetition of artificial intelligence, new ideas in business world, new ongoing trend coming with entrepreneurship, the researchers are continuously developing the ideas to solve business problems or to help business using these machine learning tools. The most important thing that paper should include is precision, proper theoretical knowledge and proper outcome that can prove to helpful in real world. This paper includes information about the dataset of customer churning based different aspects like senior citizen, gender, partnership, internet service, online backup, service prices and many more. In recent years, different work has been recorded in this area of Customer churn analysis. Different methods have been recorded by the researchers. Some of them are:

1- Omar Adwan et al. [1], new commercial competitors offer multi-layer perceptron neural networks (MLPNN): tor modelling and analysis, resulting in serious loss of profit. They have also used the data from the telecommunication industry but they use MLPNN technology to detect its precision.

2- Ning Lu [2] suggests the use of boosting algorithms Improve customer churn prediction models a customer are grouped based on boosting. The weight of the algorithm. It was a high-risk squad saw.

These reviews lack some respect to the theoretical knowledge, new introduced methods and techniques which can be used in this field. Since they have some or more deficiencies in their prediction model on fake news detection. So, the search for the precise one is still being continued. Based on the above analysis, in order to take the responsibility to solve these deficiencies by making our model more prone to be believed by people out there. There is a compressed overview of the research situation in this field of churn analysis. The model which is being represented by this review paper will definitely be able to make people more prone to be more helpful to secure their customers from churning and may be helpful to the business out there. Some specific contributions are as follows:

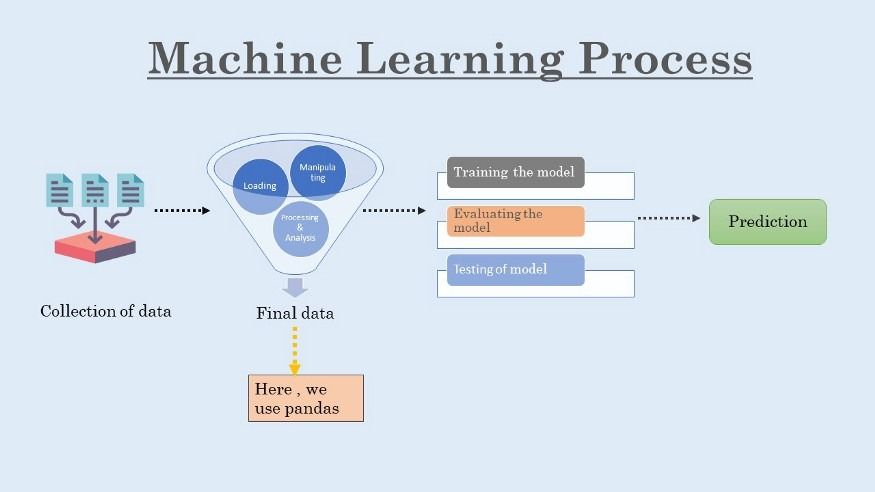
* We have divided the model into two parts, one which takes the insight from the dataset and other one trains the model on the basis of the best insights to get the model which can predict whether the customer will churn or not with the most accurate result.
* We have made use of machine learning algorithms. The extra point is that we have added multiple classification algorithms of machine learning and on the bases of their accuracy we have trained our model.
* We have also used the confusion matrix while predicting the accuracy of each algorithm which makes it more understandable to another person about the model. Confusion matrices make our model more precise, readable, and understandable.

This paper focus on the techniques, different ways to find insights from the dataset, to choose the most affected areas, and then to train the model on the basis of those categories, as training initial models on the basis of many categories may be difficult to predict and make the model precise, but filtering them with the most affected one differentiate this paper from others. The model is made on the dataset of a telecommunication industry, to help those business to secure their customer and get the best for them and helping out various businesses in the real world.

**Chapter 3**

**Methodology**

This section presents the methodology which is used in this model to build it from the very start to the end where it can be declared as a model to detect customer churn. It contains all the algorithms, methods, figures used to make this model. It contains detailed information about each step from data collecting to data testing.

A systematic process is being followed to build a model which we will be discussing in this section. The first step in this is the dataset collection phase, followed by preprocessing, then implementing selection features and then training and testing the dataset and then finally running the model to predict whether the customer will churn or not from the availing services. 

**Fig. 3.1 Procedure to build a model**

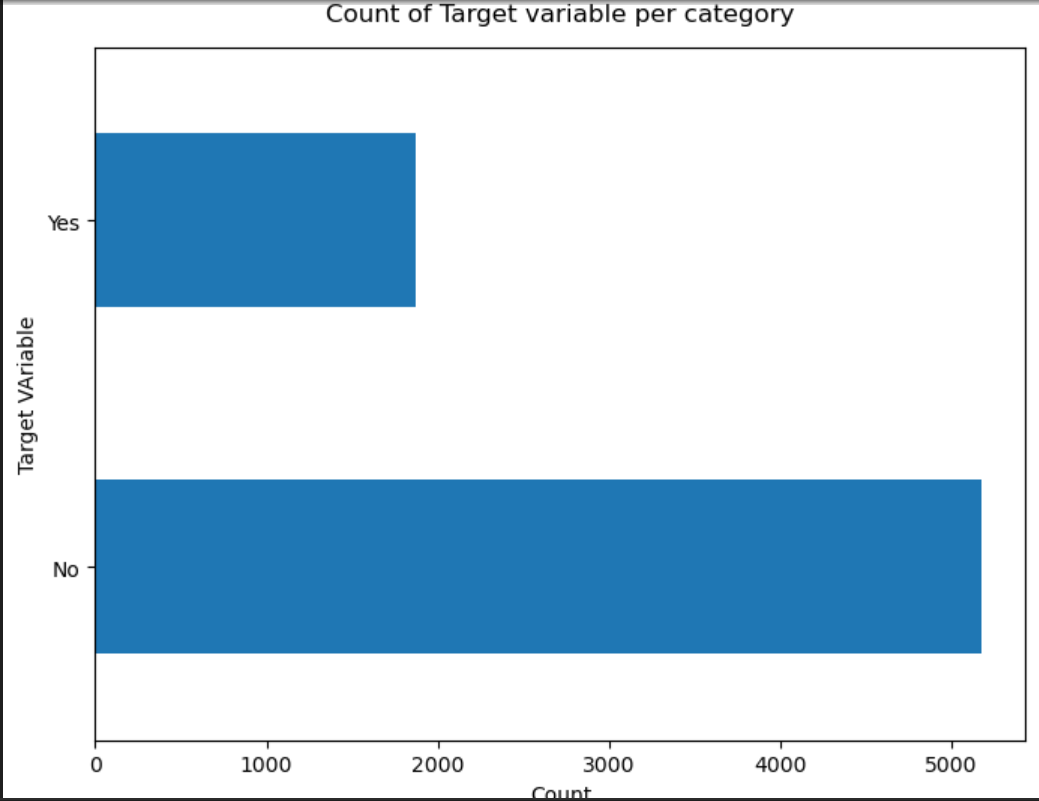
The very first step which we performed is the collection of data from the dataset of the telecom industry which provides a dataset having a collection of various categories such as gender, partnership, senior citizen, different services according to which it will provide that whether a customer will churn or not. In this dataset, \*Yes\* indicates that customer will churn basis on those criteria’s and \*No\* indicates that customer will won’t churn on the basis of those categories. After collection of data, we used pandas, a library of python to input the dataset. After that all the work of preprocessing, collecting and cleaning of data is done by the pandas library only of python.

Data preprocessing is one of the major steps in this which further include the steps to be carried out.

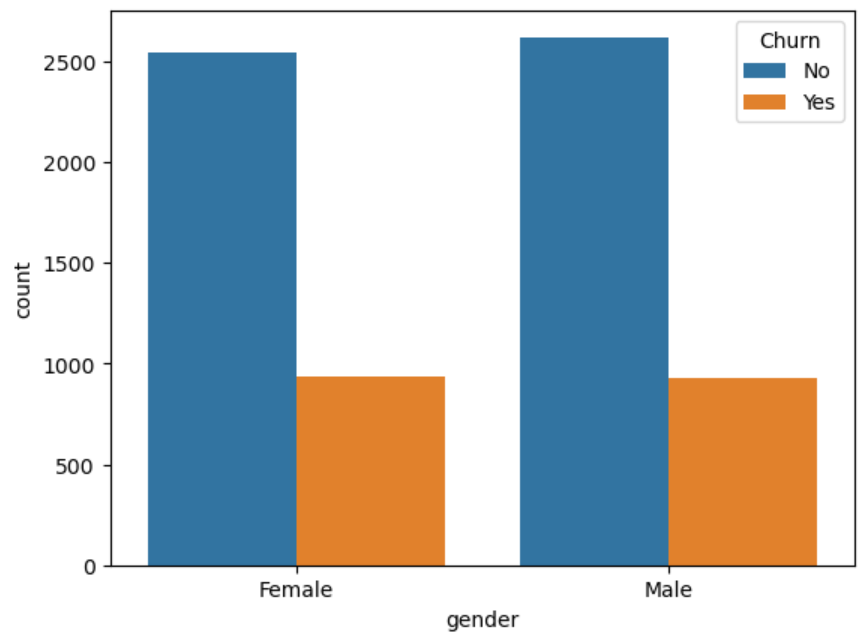
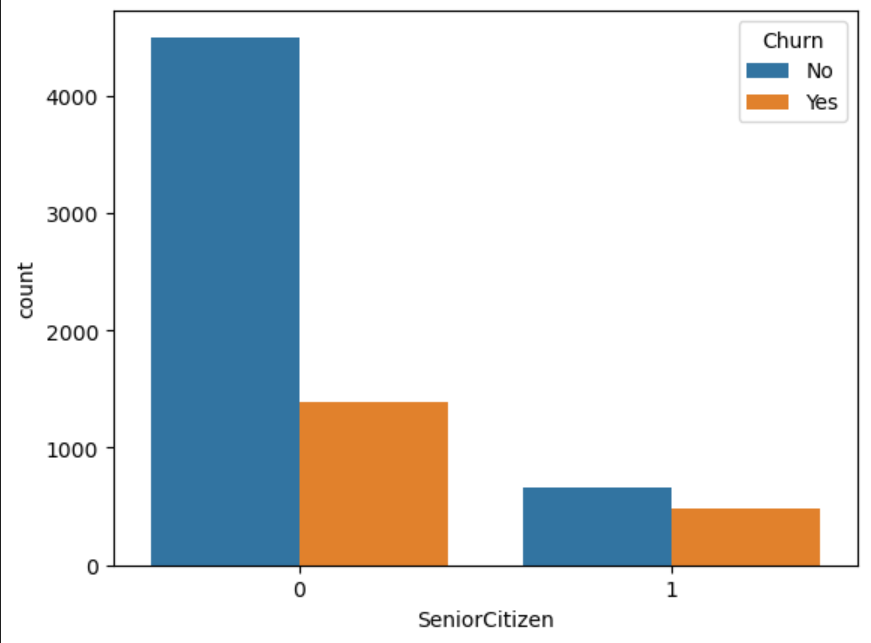
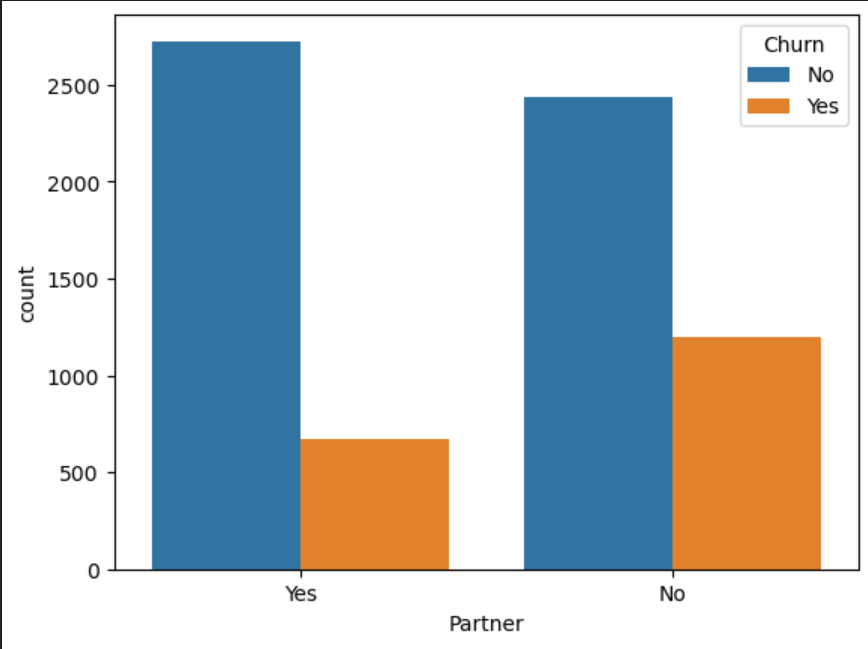
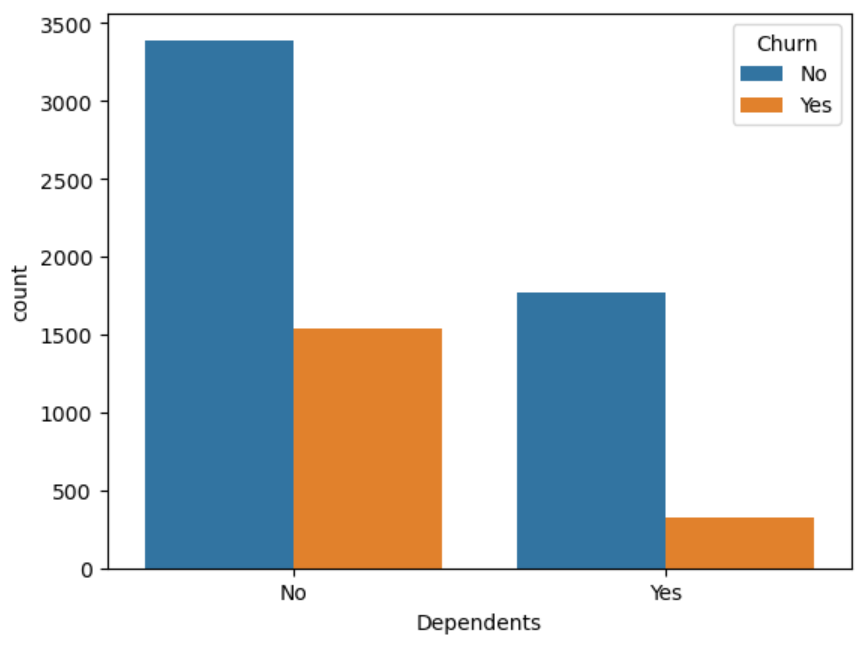
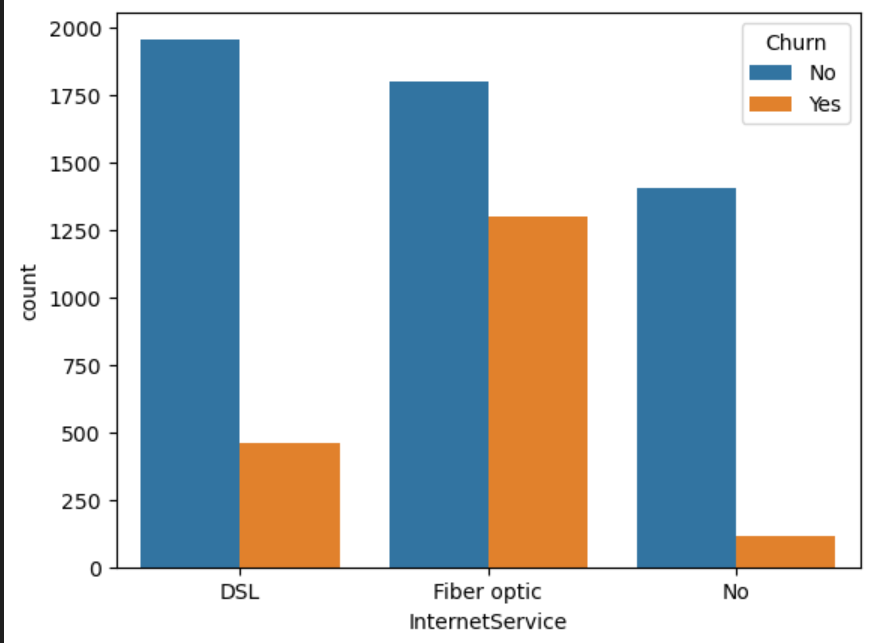
**Fig. 3.2 Data before pre-processing Fig. 3.3 Data after preprocessing**

In data preprocessing, we use various techniques for data cleaning where we find the missing or null values, merge some columns and drop some to reduce the complexity in the dataset and make it more reliable and readable. These steps make data more precise and simpler for the machine to understand. The more the data is short and precise, the more it will be better trained by algorithms and then better be tested.



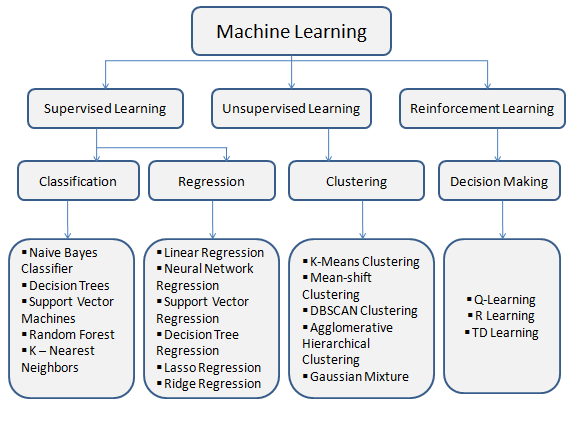
**Fig. 3.4 Count of no. of customers who will churn or not**

After doing pre-processing of data, treating missing values, cleaning the data, we have done data exploration for taking the data insights. So, to select the best categories and examine them among gender, senior citizen, partnership, internet service, online service, online backup and many more that among these criteria’s which should be used for training the model because o which the model will be most affected.

These are the bar graphs produced on exploring data on various categories. They show how much customers will churn or not on the basis of particular criteria.

After getting insights from this information, we used machine learning to train the model. They are categorised into two, that is, supervised and unsupervised. We have used supervised machine learning in classification of the dataset. We have used supervised machine learning because we want a precise end result that is whether the customer will churn or not. So, supervised machine learning contains those algorithms which make sure that the model will be trained to that extent till we get the precise result. Further in supervised learning we have used the classification algorithms since our model is about either yes or no. So, for this purpose we used the various classification algorithms to predict its accuracy. Thus, the classification algorithm proved the one to be used for training the model.



**Fig. 3.5 Machine Learning Algorithms classification**

We used various supervised algorithms to train our model like decision tree, random forest, decision tree classifier and random forest classifier.

Majority of them are classification algorithms but two of them are regression algorithms. But all of them are supervised machine learning algorithms. After training the dataset with these algorithms and then predicting the accuracy based on them, we will decide which algo will be used for final training of the model. After this process of training, finally the testing stage comes where it is being tested customer will churn or not. If it gives the correct result, that means we have trained the dataset properly. In this way, using a dataset we made a model to predict the churning basis of a customer to help businesses in the market.

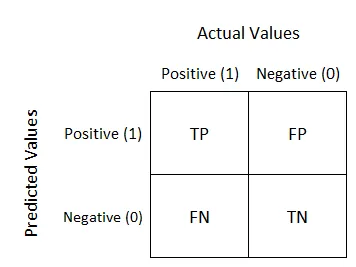
**Chapter 4**

**Result and Discussion**

This section contains the results that are produced after preprocessing and then training of the data. We have used various supervised machine learning algorithms to train our model. We have two algorithms to predict the accuracy and to train our model on the basis of it. These algorithms are:

* Decision tree classifier
* Random forest classifier

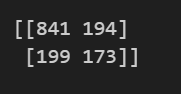
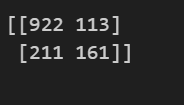
We have used mainly classification and also regression algorithms to prove classification algorithms to be more precise in this case. We have used the accuracy rate and also confusion matrix and classification reports to select the best algorithm to be used in training.



**Fig. 4.1 Confusion matrix**

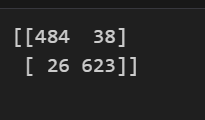
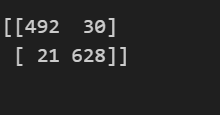
Confusion matrix is ​​a performance analysis tool in machine learning, representing the accuracy of classification models. It displays the number of true positives, true negatives, false positives, and false negatives.

So, we have displayed the confusion of these four algorithms. Since linear regression is used for algebraic data, it does not create a confusion matrix. Confusion matrices of rest four algorithms as shown as follows:

**Fig. 4.2 Decision tree classifier Fig. 4.3 Random Forest classifier**

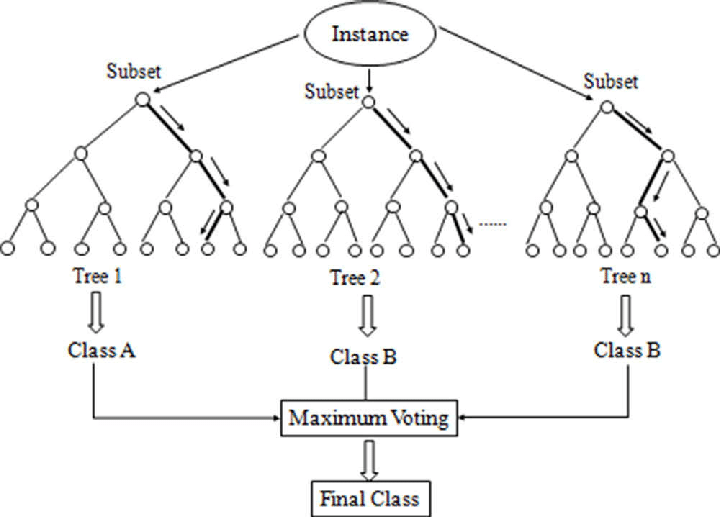
We have used techniques name SMOTE-ENN on decision tree classifier and random forest classifier to avoid overlapping in result. That is this technique adds the observation to minority classes to balance the class frequencies and thus avoid overlapping when there is n number of categories. After applying this technique, confusion matrix are as follows:

** **

**Fig. 4.2 Decision tree classifier Fig. 4.3 Random Forest classifier**

As it can clearly be depicted by the confusion matrix that on training the model with random forest supervised machine learning algorithm, it gives 21 false negative results as compared to 26 false negative given by decision classifier which can be concluded to be the most precise of them all.

Random forest is a known supervised machine learning algorithm that is used to solve the cases related to regression and classification as well. In this case it is used as a classification algorithm in which it has to differentiate whether the news is true or false.



**Fig. 4.6 Random Forest working process**

Random forest thus proves to be a better algorithm since it makes it easy to visualise the data and is a further classified version of decision tree. So, it is a more improved one. It works according to the voting procedure; this works for the majority. Thus, with an accuracy score of 95.64% random forest algorithm proved to be the best to train our model.

After training of model using these algorithms and techniques, we have deployed the model using streamlet and made a model which on inputting entries such as gender, partner, senior citizen, internet service, online backup, online service, etc can predict that whether the customer will churn or not from the service he is currently enrolled in and thus can be helpful in managing customer rate.

**Chapter 5**

**Conclusion and Future Work**

Business is not just a field; this world of start-ups demands more and more emerging ideas and new enlightened technologies. They ask for every way out so, that they make their successful and stable business. Retaining customers from churning to another service can be proved to be very helpful for them. These emerging technologies is a big push to the business world as well. Customer churn prediction model has proved to be a great use for new startups or the existing business. We have used different machine learning techniques and SMOTE-ENN technique to build the model. SMOTE-ENN is a technique to avoid overlapping content. This technique helps in precision of the result. But, making these machine-based model to produce exact precise result to be implemented and get the correct output is not easy. Since, these types of models are trained based on large amount of data having different categories and criteria. Predicting human behaviour through machine is a less precise thing. So, saying that every model containing human behaviour is accurate is incorrect to say. The only thing we fight for the most accurate result, the close one can get to the most optimal result. And that’s why, we this tech field try to work on closest correct prediction in every way possible. Also, our model, even after using different machine leaning algorithms and then training it more on the based of SMOTE-ENN technique, we get maximum of 95.64% precision only, which is definitely not a bad number but also not a good one to be that precise to be implemented. The only reason behind this is customer unpredictive behaviour which have no limits and criteria.

Future work can include, models being made by working on more re-fined dataset pre-processed, cleaned and having more decisive results. Applying more optimized techniques to increase the precision of the output is yet need to be examined. As this business is not an area of experiments, because they are the most affected fields, they work on accuracy only.

In this field, a lot more is left to be analysed and be looked into. Regularly emerging technologies and methods can be searched for. Searching into unsupervised machine learning algorithms and considering them can be a point that can be taken in the future. Using those methods which will help in the successful development of the start-ups and businesses. Other python libraries can also be used for graphics, statistics to make the model more readable. So, this topic does not end here, it has a wide scope to be followed in the coming future.

**References**

[1] Omar Adwan, Hossam Faris, Khalid Zaradat, Osama Harfoushi, Najih Ghatasheh and others

Predicting subscriber churn in the telecom industry using multilayer preceptron neural networks: observations and analysis

[2] Ning Lu, Hua Lin, Ji Lu, Guangquan Zhang “A consumer Churn Prediction Model in Telecommunication Industry Using Boosting”, IEEE Transactions on Industrial Informatics, vol. 10, no. 2, May 2014 in the.

[3] Ammara Ahmed, D.Sc. Maheshwari Textiles Research and analysis of churn forecasting techniques for customer retention in telecommunication industries 2017 Fourth International Conference on Advanced Computing and Communications Systems (ICACCS), IEEE (2017)

[4]L. Ning, L. Hua, L. Ji, Z. Guangquan Predictive Modeling in Telecom Industry Using Boosting”, IEEE Definition. Ind. Please give suggestions. 10 (2014) 1659–1665.

[5] Coen W. de Bock, Dirk van den Poel, “Findings Analysis of variable-based group classification for the patient churn prediction”, Expert Systems Management 38 (2011); 12293-1 is available