Feedback Based Telecom Churn Prediction using Machine Learning

LY Major Project Report

Submitted in partial fulfilment of the requirements of the Degree of Bachelor of Technology in Computer Engineering

by

Anirudh Bhattacharya

Saumya Shah

Mink Shethia

Supervisor

Dr. Madhura Phadke

Department of Computer Engineering

K. J. Somaiya Institute of Technology

An Autonomous Institute permanently affiliated to University of Mumbai

Ayurvihar, Sion, Mumbai -400022

2022-23

Feedback Based Telecom Churn Prediction using Machine Learning

LY Major Project Report

Submitted in partial fulfilment of the requirements of the Degree of Bachelor of Technology in Computer Engineering

by

Anirudh Bhattacharya (B-58)

Saumya Shah (B-69)

Mink Shethia (B-72)

Supervisor

Dr. Madhura Phadke

Department of Computer Engineering

K. J. Somaiya Institute of Technology

An Autonomous Institute permanently affiliated to University of Mumbai

Ayurvihar, Sion, Mumbai -400022

2022-23

CERTIFICATE

This is to certify that the project entitled "Feedback Based Telecom Churn Prediction using Machine Learning" is bonafide work of ***Anirudh Bhattacharya, Saumya Shah and Mink Shethia*** *submitted to the University of Mumbai in partial fulfilment of the requirement in Project, for the award of the degree of "Bachelors of Technology" in "Computer Engineering".*

Dr. Madhura Phadke
Project Guide
Assistant Professor
Department of Computer Engineering

Dr. Sarita Ambadekar
Head of Department
Dept. of Computer Engineering

Dr. Suresh K. Ukarande
Principal
KJSIT

Place: Sion, Mumbai-400022
Date:

PROJECT APPROVAL FOR LY

This project report entitled **Feedback Based Telecom Churn Prediction using Machine Learning** by

Anirudh Bhattacharya (B-58)

Saumya Shah (B-69)

Mink Shethia (B-72)

is an approved Last Year Project **in Computer Engineering**.

Examiners

1._____

2._____

Place: Sion, Mumbai-400022
Date:

DECLARATION

We declare that this written submission represents our ideas in our own words and where other's ideas or words have been included, we have adequately cited and referenced the sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Anirudh Bhattacharya _____

Saumya Shah _____

Mink Shethia _____

Date:

ACKNOWLEDGEMENT

Before presenting our LY project work entitled "**Feedback Based Telecom Churn Prediction using Machine Learning**", we would like to convey our sincere thanks to the people who guided us throughout the course for this project work.

First, we would like to express our sincere thanks to our beloved Principal **Dr. Suresh Ukarande** and Vice principal **Dr. Sunita Patil** for providing various facilities to carry out this report.

We would like to express our immense gratitude towards our Project Guide **Dr. Madhura Phadke** for the constant encouragement, support, guidance, and mentoring at the ongoing stages of the project and report.

We would like to express our sincere thanks to our **H.O.D. Dr. Sarita Ambadekar**, for the encouragement, co-operation, and suggestions progressing stages of the report.

Finally, we would like to thank all the teaching and non-teaching staff of the college, and our friends, for their moral support rendered during the course of the reported work, and for their direct and indirect involvement in the completion of our report work, which made our endeavour fruitful.

Anirudh Bhattacharya
Saumya Shah
Mink Shethia

Place: Sion, Mumbai-400022
Date:

ABSTRACT

The telecommunications industry is highly competitive, with numerous organisations vying for customers' attention and loyalty. One of the biggest challenges facing these companies is churn, which refers to the rate at which customers leave for various reasons, including dissatisfaction with service or pricing, better offers from competitors, or changes in their circumstances. To address this issue, companies must understand the underlying factors that contribute to churn and be able to predict when it is likely to occur. While past churn rates can provide some insight into this phenomenon, they are not always reliable predictors of future behaviour. Other factors, such as changes in the competitive landscape or shifts in customer preferences, can impact churn rates in unpredictable ways.

To overcome these limitations, we propose a method that combines past data predictions with customer feedback to generate a more accurate churn rate estimate for a particular period, such as a quarter or a year. This approach involves using machine learning algorithms to analyse past data and identify patterns and factors that contribute to churn. By analysing customer data, including demographics, location, job, and preferences, the algorithm can generate predictions about future churn rates. However, to further refine these predictions, the algorithm can also incorporate customer feedback, which can provide insights into the specific factors that are driving churn. Feedback can be collected through surveys, social media, or customer support interactions, and can provide valuable information about customers' experiences with the organisation. By combining past data predictions with customer feedback, organisations can gain a more accurate understanding of churn and take proactive steps to improve customer retention. This could involve changes to pricing or service offerings, improvements to customer support, or other strategies aimed at addressing the specific factors driving churn. By reducing churn, organisations can improve customer satisfaction and loyalty, ultimately benefiting both the company and the customer.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Sr. No.	Abbreviation	Description
1	SVM	Support Vector Machine
2	CNN	Convolutional Neural Networks
3	ANN	Artificial Neural Network
4	CPNN	Counter Propagation Neural Networks
5	CART	Classification and Regression Trees
6	MLP	Multi Layer Perceptron
7	NB	Naïve Bayes

CHAPTER 1
INTRODUCTION

The introduction chapter of this project report provides an overview of the project and its objectives. Specifically, we will outline the problem of telecom churn prediction and how our project aims to address this issue. We will also introduce the specific objectives that we have set out to achieve through our research and analysis. In addition to this, we will provide an overview of the structure of the report. This will include a brief summary of each chapter and section, as well as an explanation of how they all fit together to form a comprehensive understanding of the project. By the end of this chapter, readers will have a clear understanding of what the project is all about, what we aim to achieve, and how we plan to present our findings.

1.1 PROBLEM DEFINITION

Because 5G is gaining in popularity on a daily basis and the major telecommunications companies are all rushing to have their infrastructure modified so that it can support broad usage of 5G, these companies have a responsibility to ensure that their endeavours will continue to get sufficient funding both in the present and in the future. Seeing as how 5G is gaining in popularity on a daily basis and the major telecommunications companies are all rushing to have their infrastructure modified so that it can support broad usage of 5G. It is essential for these companies to guarantee that they will continue to get enough financing for their 5G-related initiatives since the 5G standard is becoming more and more popular on a daily basis. Because of this, the rate of churn, also known as the percentage of a company's current customers who decide to stop doing business with that company, is one of the most important aspects that these kinds of businesses pay special attention to. Customers are likely to be dissatisfied with the quality of service that they get from a firm, particularly if the personnel turnover rate at the company is quite low. Take for example Reliance Jio, which has quickly become the most significant player in the Indian telecommunications industry. They were able to reduce their customer turnover rate from 3.7% to 2% during the first three months of 2022, which led to a very lucrative period for them [18].

1.2 AIM AND OBJECTIVE

Therefore, in order to evaluate whether or not a firm will be successful, or even continue to exist, the organisation in question must make every effort to precisely estimate the turnover rate of a certain time period. Because of this, the company will be able to modify

its policies, improve the quality of service it provides, and eventually raise the total amount of money it generates. It is possible to get at these projections by taking into consideration the factors that were responsible for consumers departing the service during the time period in question and by conducting a study of the comments made by former customers. In doing so, it is possible to arrive at these projections.

It is an indication that a client may choose to terminate their membership to a service in the near future if they offer unfavourable feedback about a product or service that they have purchased or received. On review websites, it is said that consumers of Reliance Jio have written both a specified number of excellent ratings and a certain number of unfavourable evaluations about their experience with the firm and their interactions with the company. If a customer publishes a negative evaluation of the company between April and June of 2022, it is feasible to assume that the customer intends to discontinue utilising the services supplied by the company before the end of the current fiscal quarter. When attempting to perform a study of the turnover rates of a certain telecom company, one can arrive at the following conclusion: the feedback supplied by customers is a crucial component that has to be taken into consideration.

Consequently, in order to construct the system that will take into consideration prior variables that caused churn as well as input from customers, it is necessary to have an acceptable dataset. In this dataset, characteristics like Age, Gender, Package, and Used Services are recorded, along with a variety of additional information of a similar kind. On the basis of these variables alone, a machine learning model may be calibrated to provide a prediction about customer turnover for the upcoming time period. Now, if a web scraper is added to get reviews from retail sites and search engines, and a natural language processor is used to find the emotion behind the reviews, then the positivity and negativity of the reviews can be ascertained, and by the number of positive and negative reviews, the output from the machine learning model can be altered to give a more accurate prediction of the churn rate in the telecommunications industry. churn rate is the rate at which customers leave a service provider.

1.3 ORGANIZATION OF THE REPORT

In Chapter 2, the work done in Telecom Churn Prediction, Web Scraping and Natural Language processing was surveyed in order to build a more advanced system. This leads to the hardware and software specifications being analyzed in Chapter 3. Furthermore, the financial requirements are analyzed in the same chapter. Next, the design of the system is presented in Chapter 4. This proposed system's methodology is explained in Chapter 5 and in the next chapter, the implementation of this system is shown. Finally, in

Chapter 7, this system is compared to legacy systems and it’s understood how well this holds up against tried and true systems. This report concludes with Chapter 8 where the references and the accomplishments of this system are shown.

CHAPTER 2
REVIEW OF LITERATURE

The literature review chapter provides an overview of the existing research on telecom churn prediction. We conducted a survey of academic papers and websites to gather information on techniques and challenges in the field. Our findings are presented in an organized manner, highlighting key insights and conclusions. The chapter informs and contextualizes our own research in this project. By the end of the chapter, readers will have a solid understanding of the existing knowledge on the topic.

2.1 REVIEW OF PREVIOUS WORK IN TELECOM CHURN PREDICTION

In order to produce an estimate of the rate at which customers cancel their subscriptions, a wide range of separate algorithms have been put through their paces in terms of testing and evaluation to see how effectively they work on their own and how well they can predict the rate of client cancellations. This has been done in order to produce an estimate of the rate at which customers terminate their subscriptions, which is the primary purpose for why this has been done. The author employed the K-means clustering and support vector machine (SVM) techniques in order to produce a forecast about the total amount of money collected from customers. The particulars of 7043 different customers were included in the data that was collected from Kaggle, which was used as the source for the data that was gathered and acquired. Kaggle was the starting point for the collection and acquisition of the data that was done. After going through the process of converting qualitative data into dummy data, the investigation led to the discovery of a total of seventeen characteristics that distinguished one customer from the others. These qualities were discovered as a result of the inquiry. Cluster 1 had a total of 3955 records and a churn rate of 31.3, whereas cluster 2 had a total of 1671 records and a churn rate of 15.3. This was achieved by the process known as "dummy data conversion." The information was organised using a method known as k-means clustering, and the data were clustered using that method in order to make the structure of the data more transparent and simple to understand. The XGBoost model was shown to have the greatest performance all around, with an accuracy of 81%, a precision of 75%, and a recall of 90%, respectively [1].

The authors of the study compared the metrics of several different deep learning methods. In particular, they focused on the Convolutional Neural Network method, the hybrid probabilistic possibilistic fuzzy C-means clustering (PPFCM-ANN) method, the particle

classification optimization-based BP network for telecommunication customer churn prediction method (PBCCP), which was developed by Ruiyun Yu and Bo Jin, and the support vector machine in k-means clustering method. After taking into account each of these measurements, it was determined that the Precision metric was the one that contributed the most to CNN's overall performance. This was one of the reasons why this conclusion was reached. This was due to the fact that it provided the most exact measurements possible. The network was successful in achieving all of its objectives, as shown by the fact that it achieved an accuracy rate of 98.5%, a precision rate of 99%, an F-measure of 98%, and a recall rate of 99%, respectively. In line with those values, the PPFCM-ANN model was able to effectively accomplish the following results: a recall of 97%, precision of 98%, F-measure of 94%, and accuracy of 97%. The following degrees of success were attainable when the PBCPP was utilised: a recall rate of 98 percent, an accuracy rate of 98 percent, a precision rate of 99 percent, and an F-measure of 97 percent. As a point of contrast, the K-means+SVM had a precision of 95%, an F-measure of 89%, a recall of 95%, and an accuracy of 96% [2]

In 2013, for the very first time ever, a Bayesian Network was used in the process of constructing a model of customer turnover. This particular model was the very first one of its sort to ever be crafted into an actual model [3].

In this study, This particular layout was the very first of its kind to ever be manufactured. The authors of the study were able to construct a Bayesian network model with the aid of MATLAB and the BNT tools. The results of the study into the data that was employed by this model revealed that it was effective in forecasting the outcomes of events 69.8 percent of the time. This was proved by the findings of the investigation into the data that was utilised. This model takes into account a variety of different elements pertaining to the subscriber, such as the customer's address, the subscriber's complaint rate, the subscriber's utilisation of broadband service, and the visibility of the subscription fee. The information about the subscription fee is said to have a certain level of visibility depending on whether or not it was classified as information that is open to the public or information that is held in strict confidence. In addition to this, we took into account the subscriber's usage of the broadband service as well as the subscriber's corporate information as two additional factors in our research. On the basis of these criteria, the authors developed a categorization system for the many reasons why their customers ceased using the service. This enabled them to better arrange the information that they

gathered. They were able to arrange the information that they had obtained in a manner that was more efficient as a result of this.

In addition, the Random Forest method was used to solve the problem of calculating the amount of consumers who would continue to remain loyal. This was done in order to solve the task. This action was taken in the hope that it might lead to the discovery of a solution to the issue. In addition to Random Forest, the authors of this research made use of a wide number of additional algorithms during the course of their work. These algorithms ranged in complexity from simple to complex. Methods such as Random Tree, J48, Random Forest with a 10-fold cross validation, Decision Stump, Bagging + Random Tree, Naive Bayes, Multilayer Perceptron, Logistic Regression, IBK, and LWL are examples of those that fall under this area of statistical analysis techniques. In the field of machine learning, one of the most often used algorithms is called Random Tree. Within the parameters of this investigation, the strategy known as "Random Forest" emerged as the one that was capable of having the most amount of impact. With 88.63% of the total, the Random Forest technique had the highest percentage of occurrences that were successfully recognised on their own dataset. This was the most advanced method out of all the others. This strategy resulted in the most effective results when compared to the other possible courses of action that may have been taken. Both the J-48 technique and the AttributeSelectedClassifier finished in a tie for first place after being examined using the churn-bigml dataset [16]. Both of these methods were used to assess the data.

We were able to get the largest possible percentage of occurrences that were correctly detected by combining these two different methods, which gave us a total of 91.91% of the potential total. After doing more study and reviewing the data, the authors came to their conclusion based on empirical evidence, which led them to the realisation that the Random Forest and J48 algorithms performed much better than the other algorithms. This realisation led them to the conclusion that the Random Forest and J48 algorithms performed significantly better than the other algorithms. This conclusion was arrived at by the authors after the completion of further research as well as the analysis of the data. The authors were able to arrive at this realization after determining that the Random Forest and J48 algorithms were superior to those of the other algorithms. This led them to the conclusion that this realisation was possible. As a direct consequence of possessing this insight, they were able to come to this comprehension. The writers then continue by providing commentary on the several ways in which churning consumers might be

recognised, profiled, and therefore kept on as customers of the firms that are being presented in this debate [4].

An additional method that has been tested extensively in a number of settings and shown to be effective in all of them. The objective of ensemble learning is to increase overall performance by merging numerous trained models into a single system with the aim of outperforming the performance of each individual component model. This is accomplished via the use of a technique known as "ensemble learning." If the errors do not correlate with one another, then an ensemble approach that takes the average of N distinct models has the ability to lower the error by an ideal factor of N. This is achieved via the process known as "ensemble learning." Having said that, the presumption that the errors do not have any type of link with one another is necessary for this to be a right conclusion. This is accomplished by a method that is referred to as "ensemble learning." The predicted improvement is far lower than what was at first assumed to be the case due to the fact that, in real practise, the variations in imprecision across the models are tightly related to one another. The estimated improvement is much lower than what was first anticipated to be the case as a consequence of this outcome.

The performance of the ensemble model, which is produced by averaging the outcomes of each of the component models, continues to be superior to the performance of the individual models. This is because the ensemble model is made by adding up all of the findings from the component models. Despite the fact that the outcomes of each of the component models are averaged before being included into the ensemble model, this is still the case. When creating the ensemble model, the results obtained from each of the different models are integrated into a single set. When applied to a collection of machine learning models that have been trained, SEM will only choose those models from the collection that are likely to deliver correct results. This is because SEM is a probabilistic algorithm. It will remove from the collection any models that have a high probability of producing inaccurate results and filter out the rest. When it is used on a group of machine learning models that have been instructed, it is called application. This transpires as a consequence of applying it to a collection of trained machine learning models, which causes the aforementioned conclusion.

As soon as this stage is finished, it will restrict the collection down to just those models that are likely to give the outcomes that are required, and it will choose them from among

those remaining models. The component that is referred to as the "selector" and is responsible for dynamically determining which models should be included and which models should be deleted is termed the "selector," and the word "selector" refers to that component. The selection of which models should be included and which models should be removed is done by the "selector." In addition, the "component" itself is referred to as the "selector" in certain contexts. The "selector," which is a model that is taught to learn the proper actions to take for each model by using machine learning, makes the dynamic choice on which models should be included and which models should be removed. The "selector" is a model. The "selector" is the component that is responsible for making the dynamic decision on which models should be included and which models should be omitted, and it is the responsibility of this model to determine what those suitable actions are going to be. This training is done on the models themselves in their actual role, and they are the ones who carry it out. When it comes time to make predictions about how things will turn out in the end, the selector will only take into account the models that have been picked for inclusion, and those models will be the only ones that will be taken into consideration by the selector.

The authors used a dataset that included information on 3,333 unique customers, and we divided the time we had available between training and testing in the proportion of 70 to 30, respectively. The models were created by the use of a number of different methods, namely CNN, ANN, SVM, and RF, in that particular sequence. The following is a breakdown of the results of the accuracy check: 87.4% for RF, 85.1% for ANN, 84.9% for SVM, and 89.9% for CNN appropriately. The Averaging Ensemble Model was effective in reaching not just an accuracy of 90.9% but also a Precision of 79.5% and a Recall of 81.3% as well. All three of these metrics were above the average. Each and every one of these metrics is a measure of either accuracy or recall. These three indicators all had values that were much higher than average within their respective categories. Both of these metrics, which are used in the process of evaluating the model, centre on the precision of the model's forecasts as its primary emphasis. The Selective Ensemble Model was able to achieve a larger degree of accuracy while also having a higher level of performance overall when compared to the other component models. This was made possible by its ability to outperform the other component models. Because of the way the model works with ensembles, this was made feasible. Because it was able to perform better than they could, this was the only reason why it was even somewhat plausible. Because the model is capable of accomplishing a greater degree of performance, which

made it feasible for this to become a reality, this was made achievable. It had an accuracy rate of 93.9 percent, a recall rate of 83.8 percent, and a precision rate of 83.8 percent [5].

In this study, a comparison and contrast is made between the benefits and drawbacks of the logistic regression methodology and the efficacy and precision of the Hoeffding algorithm. The Hoeffding algorithm comes out on top in both categories. It has been shown that the Hoeffding algorithm is better in both of these areas. Not only does this study take into account how effectively each option works, but it also takes into account how correctly it fulfils its function. Because one of the objectives in developing this approach was to be able to predict churn, one of the goals needed to be a logistic technique that was constructed with the idea of putting the data into the category of unsupervised machine learning. The incorporation of the data into the category served as the inspiration for the construction of this method. This was done in order to simplify things and provide more precise forecasts about the number of customers that would be coming in. It is feasible to make comparisons between the logic that is responsible for the operation of an algorithm using a Hoeffding tree and the logic that is responsible for the operation of a decision tree.

The Hoeffding tree is a decision-tree learning strategy that may be used in the process of model learning. This approach removes the need for the learner to make a difficult trade-off choice. It is feasible to achieve this goal because of the fact that it does not fluctuate, which enables it to do the same amount of work in the same amount of time for each sample. It is possible to arrive at this result because of the Hoeffding tree, which does the same amount of work in exactly the same amount of time for each and every sample. The impact of a probabilistic method in the Hoeffding algorithm, which includes adopting alternative tests at each given node, has an effect that exponentially decreases as the number of testers rises. This is because the probabilistic approach involves adopting alternative tests at each given node. This is due to the fact that the likelihood of a particular occurrence is influenced by the outcomes that have gone before it. This approach is referred to as "testing alternatives," which is also its name. In comparison, the precision of the Logistic algorithm is 0.865 and the recall is 0.98, while the accuracy of the Hoeffding technique is 0.855 and the recall is 1. Both of these aspects demonstrate that the Hoeffding technique is superior to the Logistic algorithm, which is the algorithm that is often used when attempting to find solutions to issues such as this one. To put it another way, the results that are generated by the Hoeffding technique are more accurate than

those generated by the Logistic approach. It was shown that the logistic algorithm performed a great deal better than the Hoeffding Tree Algorithm did when it came to forecasting the rate at which customers would stop employing a service. This is one of the areas in which the Hoeffding Tree Algorithm underperformed. This was the predicament that we found ourselves in while attempting to predict how quickly customers would stop using a service, and we had no idea how fast it would happen [6].

When a consumer has enrolled at a company, their choice regarding whether or not to continue getting treatment at the facility is impacted by a broad variety of various circumstances. Each of these considerations may or may not persuade the consumer to continue receiving treatment at the institution. In order to carry out an examination of these components, a study was carried out, and the findings of that research are reported. This summarises the findings obtained after doing the investigation. The idea behind beginning the development of a social graph was to make the process of determining whether or not various persons are connected to one another in any way simpler and easier to comprehend. Specifically, the goal was to streamline the procedure of determining whether or not different people know one another. It was generally agreed upon that this change had occurred, and it was not difficult to see the effect that it had had on the ties that people had with one another as a direct consequence of this transition. It was observed that whenever a person churned, there was a change in the graph, and this variation took place in some fashion each and every time it took place. This change took place in some manner each and every time it took place. It is hypothesised that the amount of churn impact that an influencer has on their social network would decrease proportionally if the influencer had access to a broader range of possibilities for pleasure, and this would lead to a reduction in the amount of churn effect that the influencer had. This is done in order to take advantage of the enormous advantages that come with maintaining a large number of subscribers while solely focussing on satisfying the expectations of a single subscriber (the influencer). The goal is to maximise the number of subscribers maintained while minimising the amount of attention paid to a single subscriber. These advantages are made possible by the fact that maintaining a large number of subscribers needs a significant amount of additional effort. Only data sets that consisted of 'calls' that were conducted via wifi were assessed by the researchers as possible building blocks for a social network and sources from which to draw predictive features. This was done so that the results might be used to make predictions. This was done in order to maintain the highest level of objectivity throughout the study. To put it another way, the researchers

were only interested in the data sets that included "calls" that were conducted over wifi. Those were the only ones that met their criteria.

Their research discovered a flaw in the methods that are currently being used for the forecast of customer churn, proposed a solution that was an improvement by taking into consideration and identifying the churn influencers, suggested a framework for the prediction of such influencers, and verified the framework using data from the real world. In addition, the methodologies that are now being applied for the forecasting of churn were found by their study to include an error. In spite of the fact that the aforementioned ideas have the potential to be studied as prospective alternatives in more research, the investigation that we carried out revealed a flaw in the approaches that are presently being used to predict churn. These technologies may be helpful for wireless carriers because they may aid wireless carriers in improving their methods for churn prediction and identifying potential customers who may be served as part of the companies' retention plan. In addition, these technologies may assist wireless carriers in discovering prospective consumers who may be served as part of the businesses' retention strategy. In addition, the carriers may get assistance from these technologies in the process of finding prospective consumers who may be served as part of the retention strategy for the firms. To put it another way, these technologies could be able to help cellular carriers keep more of their current customers if they are implemented. [7]

It is feasible to produce an accurate estimate about the amount of consumers who will cease supporting a certain firm by using a variety of techniques, such as decision trees and neural networks, amongst other available options. According to the findings, a combination model may not only have a greater capability to interpret data like a decision tree model, but it also may have a higher prediction accuracy rate like a neural network model. This is based on the fact that the combination model combines both a decision tree model and a neural network model. This is because a combination model is made up of two different models: one that is a decision tree model, and another that is a neural network model. This is as a consequence of the fact that a combined model is better equipped to compensate for the weaknesses of a single prediction model and get more trustworthy and accurate prediction results than a single prediction model employed on its own would get. The summary is as follows: The accuracy of the decision tree was found to be 93.47 percent, the accuracy of the neural network was found to be 96.42%, and the correctness of the whole process was found to be 98.87 percent. [8]

The authors of this article make an effort to provide a forecast regarding the number of customers who stop purchasing products or services from a particular business by utilising the following algorithms: Counter Propagation Neural Networks (CPNN), Classification and Regression Trees (CART), J48, and fuzzyARTMAP. This is done in an effort to provide a forecast regarding the number of customers who stop purchasing products or services from this particular business. Classification and Regression Trees are denoted by the abbreviation CART. In order for them to reach their objective in the context of the telecommunications firm in India, they need to make a distinction between consumers who would churn and customers who would not churn. Because of this, they would be able to do what they set out to do and attain their goals. The Indian Telecommunication Service Industry kindly provided the dataset that was utilised for the purpose of this investigation.

The writers of this piece conducted research on the topic of telecommunications in India so that they could get the information that was essential for the successful completion of the project. This album contains 125 fully unique samples, each of which may be downloaded in one of five different quality levels depending on your preference. A partitioning approach was used to the whole of the dataset, and it resulted in the dataset being split up into the aforementioned three distinct regions. Throughout the course of the experiment, a total of three rounds of the cross-validation method in its many guises were carried out. This allowed for the collection of data from a wide variety of sources. This was done so that we could capture the maximum amount of data feasible. The record contains information on the client's degree of dissatisfaction, the costs involved with switching services, the extent to which the services were used, the customer's status, and any further pertinent data that may be applicable. These are only a few of the many distinct aspects that are associated with the client. In addition to that, this package also contains other information that is relevant to the end user. The authors used the computer language C to develop the models, which was necessary in order to do so. It is possible to summarise the accuracy of the models as follows on an average basis: CPNN: 89.83%, fuzzyARTMAP: 91.67%, and J48: 82.07% [9]

The authors of the study offered a BaYcP system as a possible solution to the problem that they identified after doing research on the most current suggestions and usage of churn prediction in the telecoms industry. This research was done in order to find a

solution to the problem that they encountered. This investigation was carried out with the purpose of finding a remedy for the issue that they found. They developed a classification strategy that they call the naïve Bayes method and implemented it inside the framework of their model. The name of this strategy comes from the person who was responsible for coming up with it, and hence it is frequently referred to as the "naive Bayes method." They made use of a dataset that was accessible via the machine learning repository that is held at the University of California, Irvine. This repository was located in Irvine. There were 3,333 occurrences in this particular dataset, and the training set and the testing set were each divided 70/30 in line with the distribution of the occurrences. It was found that the model had a specificity of 98.34%, an accuracy overall of 97.89%, and that the total accuracy of the model was 97.89% [12].

Using feature selection is still another workable alternative that may be used while carrying out churn. This is a possibility that may be considered. This is only one of many potential approaches that might be used in the situation. This particular strategy may be considered and implemented at any time. An Information Gain computation was done on each of the attributes, and an attribute evaluator was called for their feedback in order to arrive at the conclusion that would ultimately be chosen. This was done in order to facilitate the selection of the features that would, in the end, be used. After being graded, the characteristics were compared to one another, and those that had an overall rating that was lower than the others were disqualified from further consideration. The Support Vector Machine (SVM), Multi Layer Perceptron (MLP), Random Forest (RF), and Naive Bayes (NB) are the names of the four techniques that were used for the purpose of fulfilling the aim of this particular piece of research. Each of these methods was given its own acronym to indicate its full name. When there was no feature selection carried out, the accuracy of the algorithms was as follows: 88.24% for NB, 92.92% for RF, 92.10% for SVM, and 85.86% for MLP. Following are some percentages that illustrate how carefully the qualities were selected: 88.46% for NB, 95.02% for RF, 92.74% for SVM, and 86.86% for MLP. The accuracy of the SVM algorithm when feature selection was not used was 73.70 percent, the accuracy of the MLP technique when feature selection was not used was 91.60 percent, the accuracy of the RF algorithm when feature selection was not used was 92.50 percent, and the accuracy of the NB methodology when feature selection was not used was 86.80 percent. Every one of these numbers was arrived at without making use of any kind of feature selection. The following is a breakdown of the accuracy of the feature selection in respect to SVM, MLP, RF, and NB as follows: 85.60%

for SVM, 92.40% for MLP, and 87.10% for NB. Following is a breakdown of the F-measure for the algorithms that did not pick the features: 87.0% for NB, 92.50% for RF, and 91.60% for SVM. Following the selection of the features, the F-measure was computed, and the subsequent findings were as follows: 87.30 percent for NB, 82.30 percent for SVM, 92.40 percent for MLP, and 94.70 percent for RF. The standard methods that have been used in the past to study the same material provide findings that are not quite as good as the technique that is being recommended, which produces results that are far better. This may be seen by examining its level of precision as well as the performance it has in relation to the F-measure. [13]

2.2 REVIEW OF WEB APPS WITH MACHINE LEARNING MODULES

In article [10], the authors create a web application for a college community that is responsive and allows college students, teachers, and alumni to connect on a single platform. Their goal is to develop the functioning capability of the web site by utilising Django for the backend, ReactJS for the frontend, and a database. The classification model was built using SVM, and it has an accuracy of 94% when predicting whether or not a given text included profanity. Because of its text search functionality, PostgreSQL was chosen as the database to utilize. After that, the authors uploaded their application to the Amazon Web Services platform.

2.3 REVIEW OF WEB SCRAPING TECHNIQUES

In order to collect data from a variety of websites and save it in a csv file for later use, a piece of software known as a web scraper is used in the process. Using Python, the author produced a project [11] to illustrate how data from a social networking site, in this case Reddit, is crawled and saved in the database of the goods by applying the method of XPath to identify the particulars of each component of the Frequent Searches. In this particular example, the data was crawled and placed in the database of the goods. Because Reddit is one of the most prominent social networking sites on the internet, the author decided to use it to illustrate this idea rather than any other platform. The findings of the research were presented in a way that was based on percentages, and they placed a focus on the sections of the website that served as the basis for the evaluation and had the highest volume of visitors. Their thesis makes a recommendation for a prototype system that is entirely autonomous and reliant on the domain, and that is able to uncover and integrate the data that is hidden behind the search form. This idea is presented in the form

of a system. Even though its original goal was to gather online data from its source, the Scrapy project may also be used to remove the data, despite the fact that it was designed with the concept of online data scraping in mind when it was first built. One of the most important advantages of using Scrapy is the way in which requests are organised and carried out in a sequential fashion. Because of this, it is able to send another request or carry out other activities in the meantime, which are both very useful things to be able to do. The most important outcomes of their research were user-friendly search interfaces, indexing, query processing, effective data extraction techniques based on site structure, form submission analysis, and unique submission processes. These outcomes were highlighted in the previous sentence. These are the tasks that were finished off successfully.

2.4 REVIEW OF WORK DONE IN NATURAL LANGUAGE PROCESSING

Using a collection of transformer-based language models, the authors of [14] develop a model in order to comprehend the range of feelings that may be extracted from text. This paradigm is used in order to identify feelings. This picture illustrates the model. At the conclusion of the day, they decided that the BERT, RoBERTa, and XLNet models would be the best options for them to use. They were able to do this by generating a stacked ensemble of the models outlined earlier and integrating the results of each model using a process that averages the findings. This allowed them to arrive at the correct conclusion. Before the model was subjected to cross validation and fine-tuning, the texts that were a part of the dataset, which had a total of 7666 items, were preprocessed. There were a total of 7666 items included in the dataset. The models were, at long last, put through what felt like a lifetime's worth of paces after what seemed like an eternity had passed. The Ensemble Model has the highest metrics for evaluating grief, rage, and guilt when compared to other algorithms such as SVM, Naive Bayes, and Random Forest. These algorithms measure emotion using probabilities. When it came to distinguishing disgust, the BERT model and the FastTextmb approach had the highest metrics, however the BERT algorithm was the most accurate when it came to recognising shame and fear.

An Emotion EMbedding model was used, as was the case in [15], in order to carry out the process of emotion identification from text. This is one of the options available to you. Tweets containing expressions of emotion, in addition to hashtags relating to those emotions, are employed in the building of an emotion embedding model. Words were extracted with the use of the NLTK VADER sentiment analyzer, and those words were

then utilised in the formation of this model. A database was created using the information that was gleaned from 144,701 tweets. Plutchik identifies the following as the eight fundamental forms of emotions: anger, anticipation, disgust, fear, joy, trust, and sadness, as well as surprise. Because we need to vectorize the text, we have no choice but to use the CNN learning technique, and one of the steps that must be taken is the installation of an embedding layer. In this specific case, the GloVe embedding model is what's being used to embed the content. It was shown that feelings of wrath (14.4%), anticipation (9.5%), disgust (2.37%), fear (16.4%), joy (22.89%), trust (4.94%), grief (19.88%), and surprise (9.63%) were prevalent across all 137,052 narrative texts that were analysed using this methodology.

CHAPTER 3
REQUIREMENT SPECIFICATION

The Requirements Specification chapter outlines the specific hardware, software, and financial resources required for the successful completion of the project. It lists the necessary equipment and infrastructure, software applications, and tools, as well as the financial budget needed. The chapter serves as a critical guide for all stakeholders to ensure cost-effective and efficient execution. By detailing these requirements, all parties involved have a clear understanding of the resources needed. The chapter provides direction for project managers, developers, and reviewers.

3.1 INTRODUCTION:

To be used efficiently, all computer software needs certain hardware components or the other software resources to be present on a computer. These prerequisites are known as(computer) system requirements and are often used as a guideline as opposed to an absolute rule. Most software defines two sets of system requirements: minimum and recommended. With increasing demand for higher processing power and resources in newer versions of software, system requirements tend to increase over time. Industry analysts suggest that this trend plays a bigger part in driving upgrades to existing computer systems than technological advancements.

3.2 HARDWARE REQUIREMENTS:

The most common set of requirements defined by any operating system or software application is the physical computer resources, also known as hardware. A hardware requirements list is often accompanied by a hardware compatibility list (HCL), especially in case of operating systems. An HCL lists tested, compatibility and sometimes incompatible hardware devices for a particular operating system or application. The following subsections discuss the various aspects of hardware requirements.

HARDWARE REQUIREMENTS FOR PRESENT PROJECT:

PROCESSOR : Intel Pentium dual core or above.

RAM : 2 GB

HARD DISK : 160 GB

3.3 SOFTWARE REQUIREMENTS:

Software Requirements deal with defining software resource requirements and prerequisites that need to be installed on a computer to provide optimal functioning of an

application. These requirements or pre-requisites are generally not included in the software installation package and need to be installed separately before the software is installed.

SOFTWARE REQUIREMENTS FOR OUR PROJECT:
OPERATING SYSTEM : Windows XP and above, Ubuntu v12.04 and above.
SOFTWARE INSTALLATIONS NEEDED : Python, React

3.4 COST ESTIMATION:

We have analyzed the estimated costs of developing, testing and deploying our project. The costs were taken from popular testing and deployment services and have been shown in Table 3.1

Table 3.1: Estimated Cost of the Project

Server Software	Price per month	Quantity	Total (yearly)	Supporting information
Training, Testing	7000	2	14000	On Plutora
Production	317	12	3806	For an Amazon EC2 instance
Total	-	-	17806	

As visible in Table 3.1, We have taken only training and testing costs into consideration as there is no cost of developing the project, as all of the software used in this project is open-source. It specifies the budget required per month of training, testing and production of the project. The testing would be carried out twice in a year. The platforms on which the hosting would be carried out are also specified.

CHAPTER 4
PROJECT ANALYSIS AND DESIGN

The system's requirements, specifications, and design are thoroughly discussed in this chapter, providing insight into the reasoning behind the system's architecture. Additionally, the chapter highlights the potential benefits and drawbacks of the proposed system, as well as any challenges that may arise during implementation. Through this analysis a clear understanding of the system's objectives and how it aims to meet them can be found. This chapter sets the foundation for the rest of the report, providing context for the subsequent chapters that delve deeper into the development and implementation of the system.

4.1 INTRODUCTION:

The system which has been proposed and analyzed is built with reference to the literature survey in Chapter 3 and the observed defects have been removed by adding additional modules.

4.2 PROPOSED SYSTEM:

The method that has been suggested for usage in order to fit a model upon IBM's Telco Customer Churn (WA_Fn-UseC_-Telco-Customer-Churn) dataset is called the Adaptive Boosting Classifier. This approach employs a decision tree classifier as the weak learner, which is the method that has the best performance on this dataset. The goal of this endeavour is to ensure that the model accurately represents the data.

This dataset was chosen in large part because of the high degree of user adoption it has had; it has been referenced in more than 900 notebooks that have been made public on Kaggle, and it was also used in [1]. In addition to this, when compared to the other supervised learning models that were fitted on this dataset, the Adaptive Boosting Classifier performed the best with an accuracy of over 81%; hence, it was selected as the candidate for the machine learning model since it was the top performer overall.

However, the fitting procedure for the model only takes use of 19 of those characteristics, despite the fact that the dataset has 21 features. Due to the fact that the features are required for properly fitting the model, the dataset does not include any information that is repeated. As a direct result of this fact, dimensionality reduction is not required since all features are equally important.

The technique shown in Figure 4.1 may be fraught with trouble due to the fact that [16] illustrates that datasets may be afflicted by a variety of flaws. The dataset taken into consideration could also have these same problems.. Because of this, while examining the forecast of customer churn, additional factors need to be taken into consideration so that the accuracy of the prediction may be increased [7]. One possibility for the implementation of this feature is derived from the feedback and recommendations offered by consumers about a certain supplier of telecommunications services.

As can be seen in [17], the ratings and comments that have been provided by consumers are highly significant to the prognosis. Therefore, taking into consideration the fact that churning clients cancel their contracts with a certain telecom operator at some point throughout the course of a quarter, it is probable that the consumer would submit a review on a website such as Google or Yelp. A poor review will serve as evidence that there are issues to be addressed at the organisation, and it will also contribute to an increase in the customer turnover rate by a certain percentage. If, on the other hand, the review is positive, this indicates that the consumers are content with the product or service during the time period in question, and there will not be any rise in the churn rate as a direct result of this factor's presence in the equation.

The reviews that are considered will have an effect on the overall Churn rate.. In order to arrive at an accurate estimate of the rate at which customers leave a business, it will be required to make use of both the output of the machine learning model and customer reviews that have been made available on the internet. The requirements are: a machine learning model, a web scraper, a natural language processor, and a database in order to put this system into action. Additionally, a database is required.

Web scraping is currently a practicable activity because of the accessibility of many technological tools, which made it possible for the technique to become widespread. The strategy that is referred to as XPath was implemented in the Scrapy project, which was created by David Mathew Thomas and the other individuals that worked with him [11]. In addition to this, the Beautiful Soup library has the capability of extracting data from HTML files and saving it to a database.

Because of this, the web scraper is in a position to get customer evaluations from a website, and the natural language processor is in a position to make use of these reviews

in order to do more research. In order to determine the mood that is conveyed by a piece of literature, the writers of [14] used the BERT, RoBERTa, and XLnet models. The natural language processing (NLP) approach may also be used to determine the emotion of a review, which, depending on the surrounding circumstances, can be either happy or sad.

The solution that has been proposed uses the Telco Customer Churn dataset that is provided by IBM in order to aid the fitting of a model with the help of the Adaptive Boosting approach. This was done in order to make the process easier. Table 4.1 provides insight into the possible outcomes that may be extrapolated from the model after the incorporation of additional independent variables. After that, the web scraper will gather customer feedback from unbiased third-party websites like Google and Yelp, and then save it in Table 4.2.. It is possible to save the URL, the review, and the star rating (if there is one) in the columns that are most relevant to each piece of information that you have entered. The Natural Language Processor may be used to do an analysis on the Review column, and the findings of that analysis will be stored in the Alt column when they have been generated.

The amount by which it is predicted that the churn rate will either rise or decrease will serve as the basis for determining the value that will be allotted to Alt in the future. It is possible to map the column known as Alt to Table I of the database, in this report, Table 4.1, where it will have an effect on the value known as MP. This is something that can be done. In the event that this mapping is successful, the database will operate in the manner that was anticipated.

Table 4.1: Table 1 of Database

	Database Columns		
No.	**MP**	**Alt**	**FP**
1	2.2	0.9	3.1
2	2.1	0.5	2.7
…			
n	2.5	0.0	2.5

MP: The prediction made by the machine learning model.

Alt: The percentage by which the prediction by the machine learning model should be altered.

FP: The final prediction, inclusive of the prediction made by the machine learning model and the customer's feedback.

Note: Values in the table are only for reference.

Table 4.2: Table 2 of Database

	Database Columns			
No.	**Site**	**StarValue**	**Review**	**Alt**
1	TrustPilot	1	Bad	0.9
2	G2	3	Fair	0.5
…				
n	MouthShut	5	Good	0.0

Site: The site from which the reviews are being scraped.

StarValue: If a star value exists it is populated in this column.

Review: The classification of the review in terms of good, bad and fair.

Alt: The percentage by which the prediction by the machine learning model should be altered.

For example: If the predicted rate of churn in Table 4.1 (MP) is 2.2% and Alt is 0.6%, the final prediction (FP) will be MP + Alt = FP = 2.8%. This will give the final prediction of customer churn for the telecommunication company in question. In Tables 4.1 and 4.2, the proposed schema for the database is given.

In Table 4.2, we see the main analysis of the reviews being conducted. The data is taken form the Site, the Star Values (if existing) are populated in the second column and the quality of the Review is shown in the third column. Finally, Alt, is the percentage by which the prediction by the machine learning model should be altered and this is mapped to the Alt column in Table 1, shown in Table 4.1. Again, the values in this table are given only for reference and are not an indication of the data in the database.

The proposed predictor system will work as such:

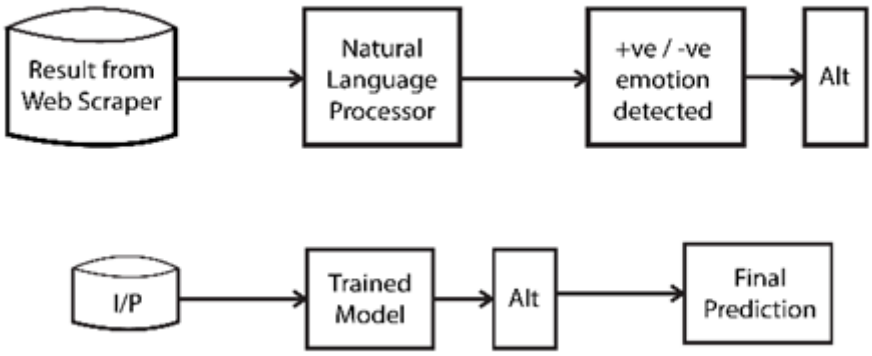


Figure 4.1: Proposed Predictor

In Figure 4.1, we see the workflow of the prediction system. As seen in the first line of the flowchart, we get the data gathered from the web scraping application, which is then segregated into positive and negative emotion values. This segregation is done by the Natural Language Processing Module, using the TextBlob library. This gives us the value, Alt, by which the prediction of the model-based on historical data-is affected and we mathematically add it to the prediction given by the model. The result of this addition will give us the Final Prediction, a combination of the result of historical data and feedback. Thus, with this prediction system in place, a more accurate prediction will be generated which will help better predict the churn rate of a telecom company.

4.3 FEASIBILITY STUDY

The feasibility of the project has been analyzed as the project implementation is done. Three key considerations that will consider for the feasibility analysis are:

4.3.1 Economic Feasibility

This project will be hosted on the internet for the users to view. The only cost involved for the user is that of setting up a computer and internet connection.

To deal with requirements to handle completion of the project we are having a strong resource of knowledge over the required technologies among our group members. Furthermore, these technologies are being taught in depth in WT tutorials to overcome any of the difficulties. Also the technologies required are economically and legally feasible for implementation purposes.

4.3.2 Operational Feasibility

This application gives a more accurate prediction of the rate of customer attrition

4.4 SOFTWARE SPECIFICATION

For our proposal to be implemented, three modules need to be used: a machine learning model, a web scraping tool, a parsing tool and a natural language processor.

4.4.1 Boosting in Machine Learning:

The purpose of the ensemble modelling method known as "boosting" is to build a reliable classifier by merging the findings of a large number of other classifiers whose accuracy is lower than the others. It is feasible to achieve this goal by using a series of less robust models one after the other in order to build a model. To begin, a model is built by using the data that will later be used for training purposes. After then, the second model has been built in an effort to improve upon the deficiencies of the first model that was previously developed. In this method, models are added until either the whole of the training data set can be successfully predicted or the maximum number of models has been added. If the training data set can be accurately predicted, then the process ends. ADABoost was the name given to the very first boosting algorithm that was created for binary classification, and it was a big success. The technique was designed to enhance the accuracy of binary classification.

Extreme Gradient Boosting, often known as XGBoost, is a library for distributed gradient boosting that was developed to be exceptionally effective, adaptable, and portable. The Gradient Boosting framework is used in order to facilitate the process of constructing machine learning algorithms. This method makes use of parallel tree boosting. This method may be utilized to solve a variety of issues that arise in the field of data science in a timely and precise way.

4.4.2 Web Scraping:

We used a web scraper and a parser to get data from the internet. Web scraping is a method that includes the use of computers to collect large amounts of information from a variety of websites. This information may be used for a variety of purposes. The vast bulk of this material is stored in HTML format, despite the fact that it is in an unstructured manner. However, in order for this information to be exploited in a variety of applications, it will first need to be transformed into structured data and stored in a spreadsheet or database. Only then will it be usable. Web scraping requires the use of not one but two unique components in order for it to be successful. These components are referred to as a scraper and a crawler. The term "crawler" refers to a kind of software that uses artificial intelligence to search the internet by following connections in order to locate specified

material. Crawlers are also referred to as spiders. On the other hand, the scraper is a one-of-a-kind piece of software that was developed especially for the purpose of gathering information from the website. This is the only reason it was ever created. To guarantee that the data is extracted in the manner that is both accurate and efficient to the greatest extent feasible, it is quite probable that the architecture of the data scraper will need to go through major revisions. The scale and scope of the project will determine the answer to this question. Web scrapers may gather all of the information that is shown on a certain website, or they can collect just the information that a user specifically asks from that website. Either way, they are able to acquire all of the information that is displayed on the website. This will help ensure that the web scraper meets the requirements.

In order for a web scraper to successfully gather data from a website, the web scraper must first be provided with the URLs of the website in question. After that, the HTML code that is unique to each of the websites is loaded into each of the browser tabs. It is feasible that a scraper with a higher level of knowledge might also extract all of the CSS and Javascript components. The information that the data scraper needs to work is then extracted from the HTML code, and it is produced in the format that the user has chosen for it to be generated in. It is common practice to save the information in the form of an Excel spreadsheet or a CSV file; however, it is not difficult to store the data in other formats, such as a JSON file.

In this system, the scraped reviews are directly stored to the database using the Django ORM.

4.4.3 Natural Language Processing:

By using technology that analyses natural language, it is feasible to ascertain the sentiments that were sent by the customer. The objective of the branch of "artificial intelligence" (AI) within the discipline of computer science that is known as "natural language processing" (NLP) is to equip computers with the ability to grasp spoken and written words in a way that is equivalent to how humans do so. Computational linguistics, also known as the rule-based modelling of human language, is combined with statistical, machine learning, and deep learning models within the domain of natural language processing (NLP). This may be thought of as the rule-based modelling of human language.

The NLP library used in this project is TextBlob and it analyses the reviews and gives them a binary value, either positive or negative. These reviews are stored in the database in the “reviews” table. They are POSTed to the front end using a RESTful API.

CHAPTER 5

METHODOLOGY

In this chapter, we see the methodology of this system and we understand the working of the system developed in this project. The various components that are used to make are described in this chapter. Additionally, the chapter provides a detailed explanation of how these components work together to achieve the desired outcome. An introduction to the technologies and programming languages used in the development of the system is defined. Overall, this chapter provides a comprehensive overview of the system and its inner workings, which is crucial for understanding the subsequent chapters in the project.

5.1 INTRODUCTION:

The defects observed in the literature survey have been removed in the previous chapter and now, we understand the working of the entire system, how the requirements impact the project and the working and how the modules connect to each other.

5.2 METHODOLOGY OF THE SYSTEM:

This system is made up of a web-based application that has been constructed, as we learned how to include a machine learning model into a web-based application in [10], with the prediction being done on the server side, thanks to the machine learning model. In addition to this, the programme has a client-side application, and it is this application's responsibility to provide the prediction to the user. As we have seen, there are a number of factors [7] that it is possible for us to not be able to take into account when trying to predict the actual churn rate that will occur during a certain time period. This is because a number of factors cannot be taken into consideration by a machine learning model. As a consequence of this, the turnover rate has a strong propensity to be very unreliable and unpredictable.

By making use of the feedback that customers provided at the right time, we arrive at an estimate of the rate of customer turnover that a certain telecommunications business experienced during the relevant time period in question. This would require that we utilise the input that customers provided at the proper time. Consumers throughout the relevant time period in question provided their input for the purposes of analysis and it’s feasible that these dissatisfied customers have left the company.

The dataset has important information such as subscriber information, which includes the location of the subscriber, the existence of partners and dependents, information on the

subscriber's contract, which may include the subscriber's tenure, charges, international plan, and voicemail use, as well as other information of a similar nature. This is standard data which is required for the prediction of churn in the telecommunication industry. There is a potential that in the not-too-distant future, further material that is relevant will be made available. In order to guarantee that the model shows these one-of-a-kind characteristics in the most realistic way possible, it may be altered in response to the data that is supplied in order to make adjustments for any variations that may occur in the future. If another dataset is going to be used to train the model, then not only does the new dataset need to be cleaned, but it also has to display advantageous qualities that are equal to those that were found in the first dataset.

After the model has been integrated into an application, the results that it creates are recorded and stored in a database so that they may be referred to at a later time. This allows the results to be accessed in a more convenient manner. Because of this, the findings may be incorporated into several additional applications that make use of the model.

The results of the model are recorded into the database so that they may be easily retrieved at a later time if that becomes necessary. This is done so that the results will be readily available. We are able to harvest data that is relevant to our project and include it into our application by using a web scraper such as Scrapy. The use of web scrapers makes this accomplishment conceivable. In order for this outcome to even have a sliver of a chance of occurring, it was necessary to make use of a web scraper. This aim could not have been accomplished without the assistance of a web scraper, which is why it was necessary to get one in the first place.

Data may be "scraped" from websites in order to get the information that is relevant. There is a possibility that each and every one of these testimonies will be entered into the database on an individual basis, with each one being given its very own column in the area of the table that has been established specifically for the purpose of achieving that activity.

The natural language processor will carry out an analysis of these testimonies and make conclusions whenever that particular function is called upon. Emotions are only one of the many potential study areas that may be examined via research, but they are also one

of the many possible research subjects. Because we are able to apply this information to the processor that is being used in this online application, we are able to forecast the emotional intensity of the testimonials that are being provided to us as a direct result of this. Specifically, we are able to predict the degree to which the individuals who are providing us with these testimonials are pleased with the companies' services. If the reviews also include star ratings, then those ratings will surely be recorded in a separate column of the database.

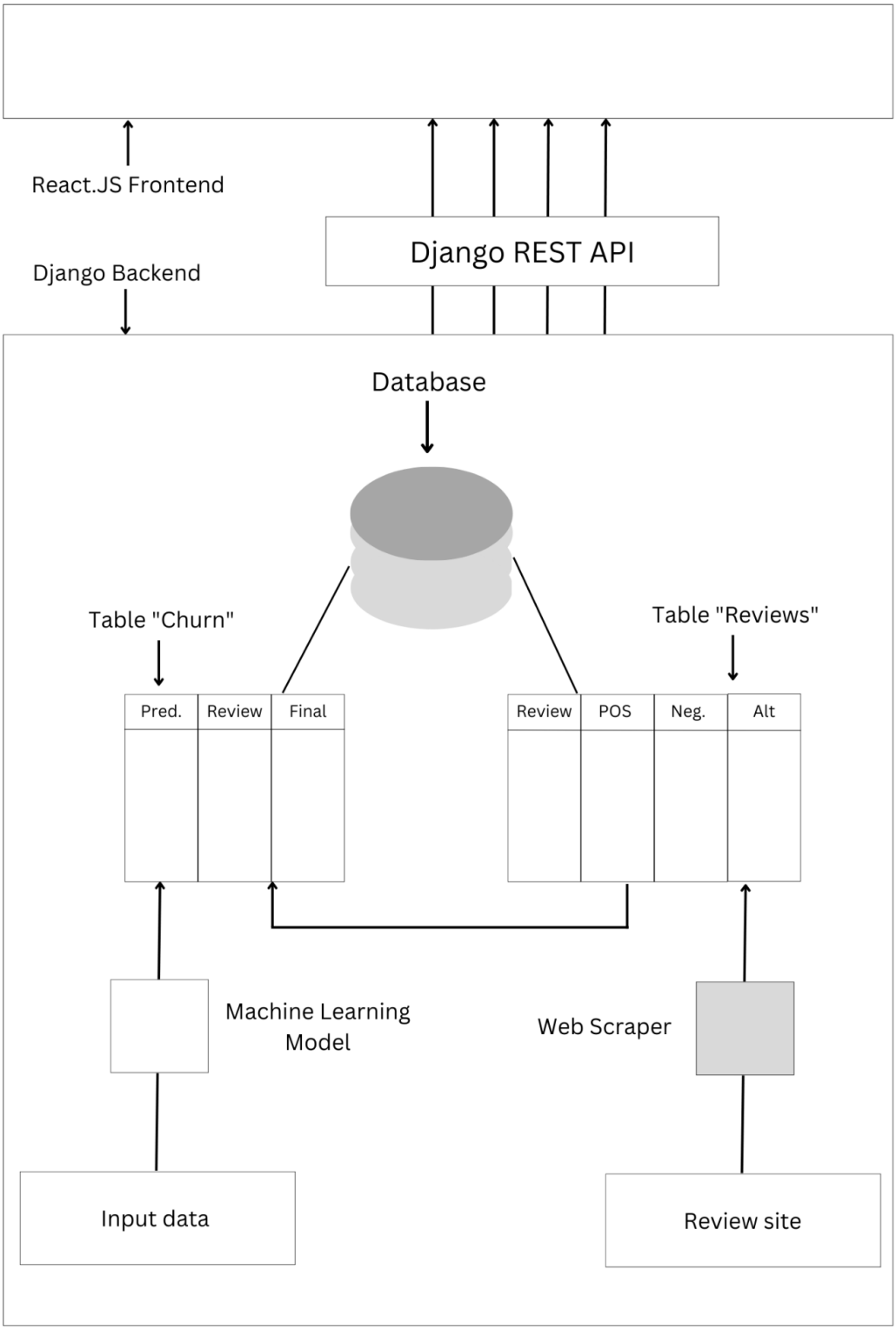


Fig 5.1: The architecture of this system.

Once stored, the NLP module will analyse the texts and save it in a different column of the same table of the database. This analysis will affect the final value. The entire architecture of the system is shown in Fig. 5.1, with the tables of the database, the static files, the web scraper and the input data. The input data and the review site give us data

which can be used to generate the prediction. The input data is put through the Machine Learning Model. The prediction generated by the Machine Learning Model's predictions are saved to the Pred table. The data from the review site is scraped using Scrapy and is saved to the review column in the database. The natural language component, TextBlob segregates the reviews as positive and negative and this generates Alt. This generates the final prediction. Django REST Framework creates APIs from the database and the data is POSTed to the frontend.

CHAPTER 6

IMPLEMENTATION

In this chapter, we can explore the practical implementation of the project by utilizing a wide range of technologies, programming languages, and libraries. With the advancement of technology, the possibilities for creating innovative and efficient solutions have become endless. The execution of the project requires a deep understanding of the available technologies, programming languages, and libraries. By utilizing the right combination of these resources, we can develop an efficient and effective solution that meets the project objectives.

6.1 INTRODUCTION:

In this chapter, it is seen how we have built the coding aspect of the project, in a manner that would be visible to a user.

6.2 IMPLEMENTATION:

Python 3 was employed all throughout the process of developing the application's server side, whereas React.JS was utilised all throughout the process of developing the client side [10, 19], and [20] in that particular order. Throughout the whole of the process of constructing the program's server side, version 3 of Python was used. During the whole of the constructing process, these two languages were used in a number of different contexts, each of which has its own individual characteristics. The dataset was preprocessed with the assistance of scikit-learn, which is a machine learning toolkit that is open-source and free to use. The toolkit was used in the preprocessing of the dataset. During the process of preparing the dataset, the toolbox was put to good use. After all was said and done, the machine learning model was constructed according to the guidelines that were supplied by the scikit-learn programme. Following that, the model was built with the assistance of joblib by deriving the information that was included inside the dataset. This was done in order to bring the process to a successful conclusion. After being extracted, this model was then saved in the Django web application framework by making use of the framework's included static file storage system. After that, the model was stored in a file that was given the extension.sav when it was saved.

Scrapy, the web scraper, has been integrated into Django in order to provide users with access to the database, which is a SQLite database designed specifically for Django. The link between the React.JS frontend and the Django backend has been formed via the use of the Django REST framework, which was imported into a file that was put inside the

system. This allowed for the connection to be made. TextBlob was them implemented to find the sentiment of reviews. As a direct result of this, the algorithm is able to identify the importance of a statement.

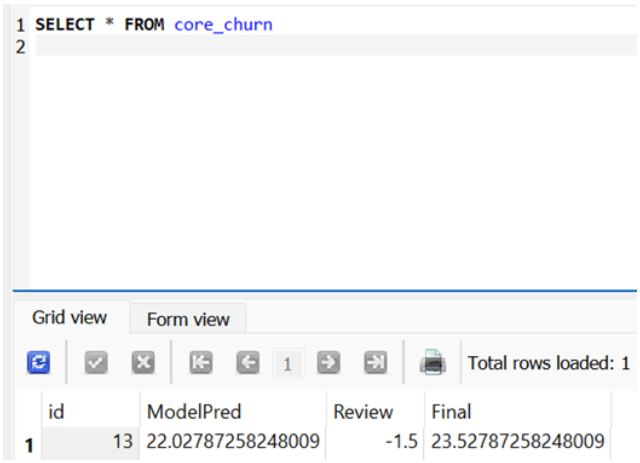


Fig 6.1: Table “churn” of the database

In figure 6.1, we observe the “churn” table of the database, which shows the ModelPred, Review and Final values, which have been calculated using the machine learning model and web scraper. These values are converted into APIs and are then POSTed to the front-end of the website.

The web crawler that was constructed for the second application with the aid of Scrapy [11] is a component of the second application, and the web crawler that is the subject of this discussion is a part of the second application. This web crawler connects to a trustworthy website, in this instance TrustPilot, in order to get data from that domain. After that, the data are put into the "reviews" table of the database, and after they've been stored there, the process is complete. This programme also incorporates TextBlob, a natural language processor, albeit the manner in which it does so varies [21]. It's possible that this will take the form of an import or an embed. In the paragraphs that follow, we are going to talk about one of the functions that it has, and then we will go on to the next topic. TextBlob will next evaluate the scraped reviews and, based on the substance of the reviews, will come to a conclusion as to whether or not they are favourable evaluations of the product or whether or not they are negative evaluations of the product. After the first stage of the procedure, which entails calculating the number of positive and negative comments, has been successfully completed, the difference between the two totals is then split into 10 equal halves in order to successfully bring the procedure to a successful finish. This value, which is stored in the "Alt" column of the "reviews" database, has been

mapped to the "Review" field that is included inside the "churn" table. Both of these tables are a part of the "reviews" database.

Fig 6.2: Table “reviews” of the database

In Fig. 6.2, we see the reviews table of the database where we have the columns: Plaintext reviews, PosFeedback, NegFeedback and Alt, which have the text of the reviews, the number of positive feedbacks, the number of negative feedbacks and the value by which the prediction of the model changes, respectively. These are also POSTed to the front screen through a RESTful API.

The third application is known as the Frontend, and each of its constituent components is known collectively as the Frontend. The term "Frontend" is also used to refer to the third application. The developers made use of React.JS in order to design the frontend of the application, and the Django REST Framework was used in order to make connections between the various components. This framework will be used to assist in the development of application programming interfaces (APIs), which stand for "application programming interfaces," for the content that will be given. The graphical user interface (GUI) of the application is connected to the database, so the values may be seen on the display at any moment.

A customised command known as "runmodel" has been developed in order to facilitate the speeding up of the processes of the object relationship model (ORM), the scraper, and the machine learning model. This was done in order to save time. This command was developed in order to make it easier to speed up the actions that were taking place. This command was developed in order to make the method less complicated to understand and more uncomplicated to put into practice. In addition, the mathematical calculation ((number of positive reviews - number of negative reviews) / 10) be carried out as a reaction to this instruction in order to provide an estimate for the overall churn. It is right that a lower overall number of positive evaluations from consumers will lead to a bigger

churn estimate; however, it is also valid that the opposite will occur in the event of a larger churn projection.

An increase in the total number of adverse customer reviews is likely to have the effect of lowering the expected rate of customer turnover, which is likely to happen as a result. We anticipate that our method will result in an accuracy that is between six and seven percent more accurate than any accuracy that has ever been recorded in the past. These precise details are drawn from research that was done in the past. In comparison to any prior accuracies that may have been acquired, this is a significant leap forward in terms of precision.

The home page of the site is shown in Fig. 6.3. This is the landing page and is visible when the server is started.

Fig: 6.3: Home page of the website

The landing page visible in Fig 6.3 has been built with React and it will have the features usually associated with home pages. It will have a hyperlink, linking it to the churn rate of telecommunication companies.

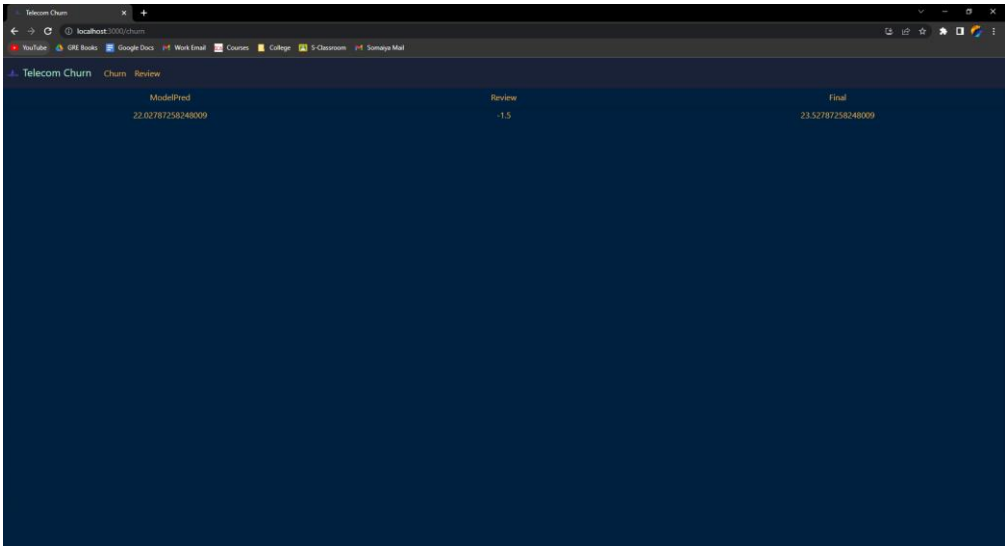


Fig 6.4: The “churn” table seen on the frontend

Fig. 6.4 shows the contents of the database which have been POSTed from the database, thus verifying the database connection. The ModelPred, Review and Final columns of the churn table in the database are called through the RESTful API and is displayed on the screen for the user to see.

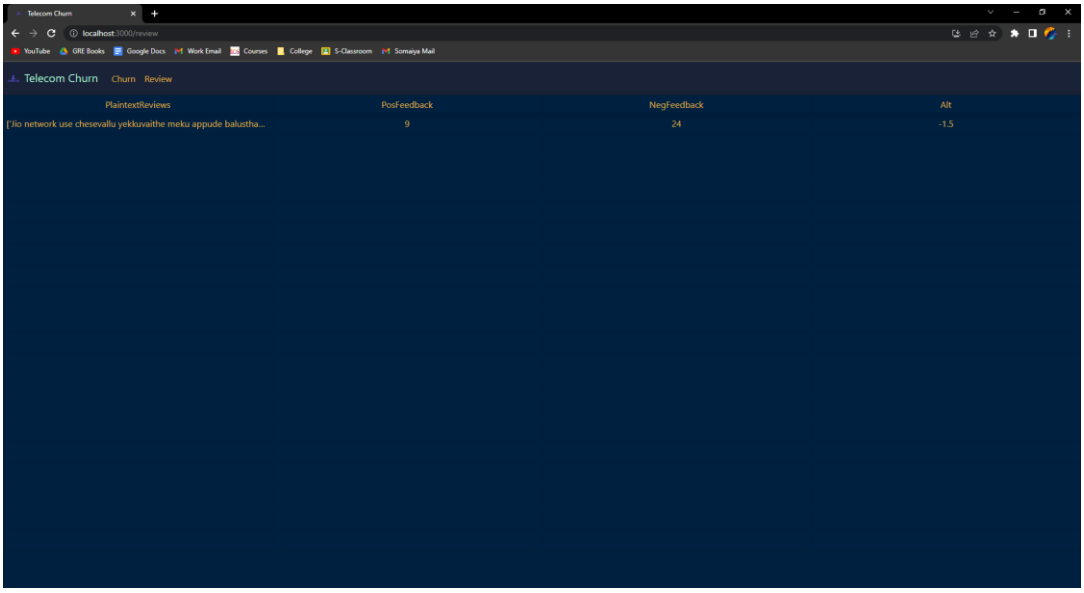


Fig 6.5: The “reviews” table seen on the frontend

In Fig 6.5, the “reviews” table is POSTed to the screen and is visible to the user. Using this we can deduce that the “reviews” table has been successfully connected to the React.JS frontend using the RESTful API. The PlaintextReviews, PosFeedback, NegFeedback and Alt columns of the review table in the database are called through the RESTful API and is displayed on the screen for the user to see.

CHAPTER 7

RESULT ANALYSIS

The completed project comprises three essential components, namely the machine learning model, the web scraper, and the natural language processing unit. Each of these components plays a crucial role in the overall functionality of the project. Therefore, analyzing their performance is necessary to ensure that the project delivers the desired outcomes. The performance of these three components is critical to the success of the project. By analyzing their performance, we can identify any issues or shortcomings that need to be addressed to improve the project's functionality and efficiency.

7.1 PERFORMANCE ANALYSIS:

In this chapter, we review the various performance metrics on all the different aspects of the project.

7.1.1 Machine learning model:

IBM decided to utilise the adaptive boosting classifier for its Telco Churn dataset since it had a better performance than other popular algorithms that are used to forecast customer churn on the same dataset. These other algorithms were used to make the prediction of customer churn. These strategies are also used to forecast the amount of customers who will leave an organisation. These extra methods may also be used to make predictions on the percentage of clients who cancel their orders. Predictions about the proportion of customers who end up cancelling their purchases may also be made with the use of these additional approaches. It turned out that the method had an accuracy of 81.7974%, a precision of 65.6160%, and a recall value of 53.3799% after being subjected to more study and having the necessary literature combed through. After the approach was investigated further and the relevant literature was reviewed, these numbers were calculated as a result. Throughout the whole of the development process for the project, these metrics were kept up to date, and the.sav file that was produced by utilising joblib exhibited the exact same metrics as well. It has been shown that this model is capable of performing to an acceptable level by meeting the rigorous requirements that were established for it prior to its release.

7.1.2 Web Scraper:

Because of the persistent efforts of the spider, Scrapy [11] was selected to fulfil the role of web scraper for this particular architectural design. Because it was specified in the robots.txt file of the website, TrustPilot was set apart from other big sites in a manner that

was recognisable because we were allowed permission to submit many requests to the website that performs reviews. This separated TrustPilot apart from other large sites in a way that was identifiable. Additionally, the permission for TrustPilot to communicate with us and provide us its responses was authorised. After the scraper has been turned on, it will instantly start collecting all of the testimonials, and after they have been reviewed, they will be stored in the database. This procedure will start as soon as the scraper is given permission to operate. These evaluations often include non-essential characters like apostrophes and spaces, and in order for them to be processed, it is necessary to eliminate both of these characters. In order to make use of them, however, these characters will first need to be erased. The use of a simple method is all that is required to accomplish the level of cleanliness that has been accomplished here. The data that was scraped may then be evaluated with the assistance of the NLP tool once it has been stored in the database that it was scraped from. After the data has been scraped, this step may finally be completed. This scraper provides an adequate level of performance, and if any problems should arise with it, it can either be updated or replaced for a scraper that provides quality that is of a higher grade. Regardless of which option is chosen, the scraper will continue to provide an adequate level of performance.

7.1.3 Natural language processor:

TextBlob is going to be the natural language processor that is used for this project. This decision was made due to the fact that TextBlob is widely employed in emotional analysis. The principal job that comes within its purview within the context of this system is the counting and analysis of the total number of positive and negative assessments. Within the context of this specific ecosystem, this is the most basic role that it plays. It was decided to employ TextBlob, a translation service, since it provides translations both into and out of Hindi. Because some of the evaluations on the website in question were written in Hindi, this was a need for the site. TextBlob satisfies this criterion by providing translation services in the other direction, from Hindi into Hindi. It now supports a range of languages, such as English, Hindi, Bengali, and Gujrati; but, at this moment, it does not support any of the other significant Indian languages. Because of this, the performance of this library is adequate; yet, it is still capable of being replaced in the event that a processor that is more dependable becomes available in the future.

7.2 REQUIRED COMMANDS:

`python manage.py migrate:`

This updates any changes made to the database using the ORM

```
python manage.py runserver, npm start
```

This starts the server, including the React.JS frontend

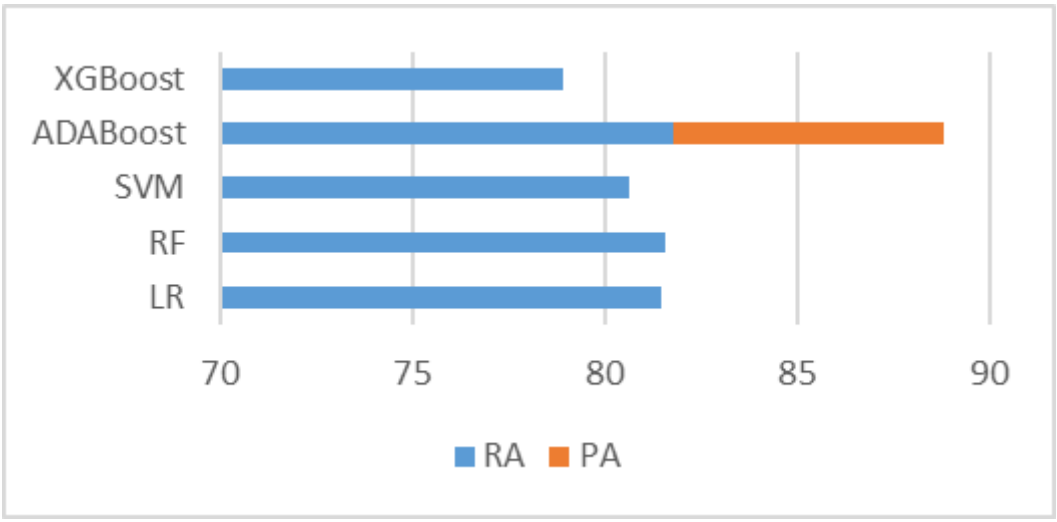
```
python manage.py runmodel
```

This runs the machine learning model, the web scraper and the natural language processor all at once, in order to save time.

7.3 COMPARATIVE ANALYSIS:

According to the findings of the research that was carried out as well as the findings of the analysis of the pertinent literature that was carried out, the degree of accuracy of previously established practises for churn prediction in the telecommunications sector ranges anywhere from sixty percent to ninety percent depending on the specific situation that is being addressed. The findings of the study laid the groundwork for the acquisition of this information. Due to the fact that the major source of information that is used in these forecasts is data gathered over the course of the years that have come before, these projections are likewise rather static. They are unable to take into account alterations that have taken place as a result of real-time advancements in the mentalities of consumers or in the procedures that organisations apply. Because it takes into consideration information from both the recent past and the current present in order to offer a forecast that is far more accurate, this method is a significant advancement over the one that was being used before it. According to the results of this machine learning model, the final estimate will probably have a degree of accuracy that is 6-7% closer to the mark than what was first supposed to be the case. This discovery contradicts what was previously believed to be the case.

The graphs that are shown below provide a contrast between the accuracy of the forecasts generated by this system and the accuracy of the predictions generated by other systems.



LR: Logistic Regression
RF: Random Forest
SVM: Support Vector Machine
RA: Recorded Accuracy
PA: Predicted Accuracy

Fig: 7.1: Comparison with other algorithms on Telco Churn

In Fig. 7.1, this system has been compared to other algorithms that have been applied to IBM's Telco Churn dataset. The purpose of the comparison is likely to evaluate the performance of the system relative to other existing approaches in predicting customer churn in the telecommunications industry. The graph suggest that the study compares the performance of different algorithms in predicting customer churn in the telecommunications industry, and explores the potential impact of incorporating reviews or feedback data on algorithm accuracy.

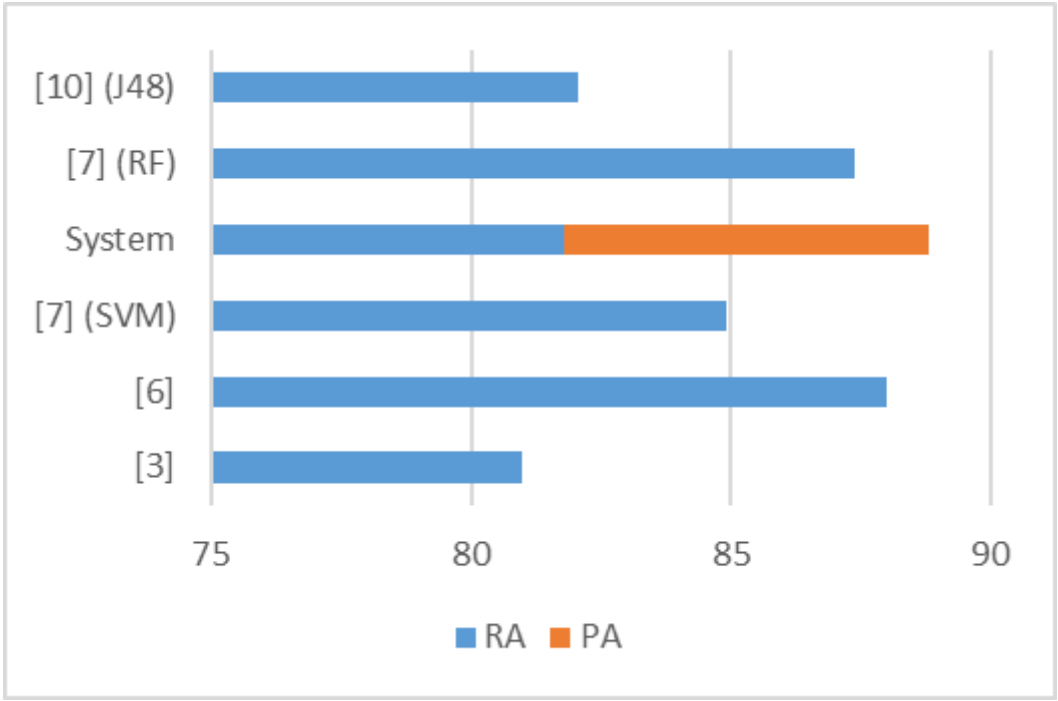


Fig: 7.2: Comparison with other papers

In Fig. 7.2, it is compared to other systems using machine learning techniques, which was observed in our literature survey. Comparison to other systems that use machine learning techniques, based on a literature survey conducted by the authors. The purpose of the comparison is likely to evaluate the performance of the system relative to other existing approaches in predicting some outcome variable or solving a particular problem. The study aims to compare the performance of the system with other existing systems that use machine learning techniques, based on a literature survey. Additionally, the impact of reviews on algorithm accuracy is being explored, which could potentially be used to improve the performance of the system.

CHAPTER 8
CONCLUSION AND FUTURE SCOPE

The churn rate of a telecommunication company is the rate at which existing customers leave the service, in a particular time period. The churn rate for a telecom company is important as a company needs to consider how they will cope with the loss of revenue and how much will their revenue be impacted by such leaving customers.

Analysis of customer churn using various machine learning algorithms has been done in the past but for the most accurate prediction, customer feedback needs to be taken into account as well, because when customers are dissatisfied with a service, they tend to review the companies as such on popular sites. Thus, with a system consisting of a machine learning model, a web scraper, and a natural language processor to analyze the emotion in a review, the churn rate of a telecom company can be accurately predicted. With the best algorithm in use and the most reliable, unbiased sites being used for reviews, the accuracy of the prediction will increase.

A working model of the proposed system has been developed, with the technologies of machine learning, web development, web scraping and natural language processing having been implemented. This implementation has been made with reference to several previously made systems and is more robust due to its use of components and APIs. This system can yield a better prediction than previously researched methods and can be a significant asset to telecommunication companies and their customers.

In the future, this system can be generalized, with reviews from various sites affecting the predictions given by machine learning models. For example, if a company is selling a product, they may take historical data to fit a machine learning model to predict sales. They can also use our concept and let reviews influence their predictions, such that they can get a more accurate prediction. Thus, whenever data and feedback is available, such a system may be used.

REFERENCES

1. Pan Tang (2020), “Telecom Customer Churn Prediction Model Combining K-means and XGBoost Algorithm”, International Conference on Mechanical, Control and Computer Engineering (ICMCCE)

2. Priya Gopal, Dr. Nazri Bin MohdNawi (2021), “A Survey on Customer Churn Prediction using Machine Learning and data mining Techniques in E-commerce”, IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)

3. Peng Sun, Xin Guo, Yunpeng Zhang, Ziyan Wu (2013), “Analytical Model of Customer Churn Based On Bayesian Network”, 2013 Ninth International Conference on Computational Intelligence and Security

4. Irfan Ullah , Basit Raza1, Ahmad Kamran Malik, Muhammad Imran, Saif ul Islam, Sung Won Kim (2019), “A Churn Prediction Model using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector” IEEEAccess

5. Ahmad Hammoudeh, Malak Fraihat, Mahmoud Almomani (2019), “Selective Ensemble Model for Telecom Churn Prediction”, IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)

6. Mr. Anurag Bhatnagar, Dr. Sumit Srivastava (2020), “Performance Analysis of Hoeffding and Logisitic Algorithm for Churn Prediction in Telecom Sector”, International Conference on Computation, Automation and Knowledge Management (ICCAKM) Amity University

7. Sara Motahari, Taeho Jung, Hui Zang, Krishna Janakiraman, Xiang-Yang Li, Kevin Soo Hoo (2014), “Predicting the Influencers on Wireless Subscriber Churn”, IEEE WCNC'14 Track 4 (Services, Applications, and Business)

8. Xin Hu, Yanfei Yang, Lanhua Chen, Siru Zhu (2020), “Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network”, IEEE 5th International Conference on Cloud Computing and Big Data Analytics

9. Ramakanta Mohanty, Jhansi Rani K (2015), “Application of Computational Intelligence to predict churn and non-churn of customers in Indian Telecommunication”, International Conference on Computational Intelligence and Communication Networks

10. Ankit Verma, Chavi Kapoor, Abhishek Sharma, Biswajit Mishra (2021), “Web Application Implementation with Machine Learning”, 2nd International Conference on Intelligent Engineering and Management (ICIEM)

11. David Mathew Thomas, Sandeep Mathur (2019), “Data Analysis by Web Scraping using Python”, Third International Conference on Electronics Communication and Aerospace Technology [ICECA 2019]
12. Pronaya Bhattacharya, Akhilesh Ladha, Ashwani Kumar, Ashwin Verma, Umesh Bodkhe, (2020) “BaYcP: A novel Bayesian customer Churn prediction scheme for Telecom sector”, 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)
13. Yakub K. Saheed, Moshood A. Hambali (2021), “Customer Churn Prediction in Telecom Sector with Machine Learning and Information Gain Filter Feature Selection Algorithms”, 2021 International Conference on Data Analytics for Business and Industry (ICDABI)
14. F. A. Acheampong; H. Nunoo-Mensah; Wenyu Chen (2021), “Recognizing Emotions from Texts Using an Ensemble of Transformer-Based Language Models”, International Conference on Wavelet Active Media Technology and Information Processing (ICWAMTIP)
15. Seo-Hui Park, Byung-Chull Bae, Yun-Gyung Cheong (2020), “Emotion Recognition from Text Stories Using an Emotion Embedding Model”, International Conference on Big Data and Smart Computing (BIGCOMP)
16. Philip Hart; Lijun He, Tianyi Wang, Vijay S. Kumar, Kareem Aggour, Arun Subramanian, Weizhong Yan (2022), “Application of Big Data Analytics and Machine Learning to Large-Scale Synchrophasor Datasets: Evaluation of Dataset ‘Machine Learning-Readiness’”, IEEE Open Access Journal of Power and Energy
17. Jo Mackiewicz, Dave Yeats, Thomas Thornton (2016),” The Impact of Review Environment on Review Credibility”, IEEE Transactions on Professional Communication
18. Tanay Singh Thakur (2022), “Reliance Jio Subscriber Churn Dipping, ARPU Rising a Good Thing”, TelecomTalk
19. Joel Vainikka (2018), Full-stack web development using Django REST framework and React (Thesis), Metropolia University of Applied Sciences
20. Shen Zhao (2022), Design and Implementation of Big Data Crawling and Visualization System Based on COVID-19 Data, 2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)
21. Orissa Octaria, Danny Manongga, Ade Iriani, Hindriyanto Dwi Purnomo, Iwan Setyawan (2022), Mining Opinion Based on Tweets about Student Exchange with

Tweepy and TextBlob, International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)

PUBLISHED PAPERS

Sr. No.	Paper title/ poster title	Technical Paper/ Poster	Venue	Date	Achievement (winner/runner up/ participation certificate)
1	Feedback Based Telecom Churn Prediction using Machine Learning	Paper	KJSIT	02/12/2022	Certificate of Acceptance, Publication
2	Telecom Churn Prediction: A Reviews and Machine Learning Based Approach	Paper	CONIT (The International Conference for Intelligent Technologies), Hubballi, Karnataka , India.	03/06/2023	Paper under review

2022 5th International Conference on Advances in Science and Technology (ICAST)

Feedback Based Telecom Churn Prediction Using Machine Learning

Dr. Madhura Phadke
Department of Computer Engineering
K J Somaiya Institute of Engineering and Information Technology
Mumbai, India
madhura.phadke@somaiya.edu

Anirudh Bhattacharya
Department of Computer Engineering
K J Somaiya Institute of Engineering and Information Technology
Mumbai, India
anirudh.b@somaiya.edu

Mink Shethia
Department of Computer Engineering
K J Somaiya Institute of Engineering and Information Technology
Mumbai, India
mink.shethia@somaiya.edu

Saumya Shah
Department of Computer Engineering
K J Somaiya Institute of Engineering and Information Technology
Mumbai, India
saumya.ss@somaiya.edu

2022 5th International Conference on Advances in Science and Technology (ICAST) | 978-1-6654-9263-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAST55766.2022.10039530

***Abstract*—In an industry with stiff competition among individual organizations, Churn is an important factor to be considered by the company itself as well as by prospective customers. In the telecommunication industry, churn can be affected by multiple factors: a customer's preference, location, job, and so on. Thus, customer churn in the telecom industry is a widely studied subject. However, churn based on previous rates alone is not enough to predict future churn and there must be additional factors considered. We propose a method to overcome this inadequacy. By using the predictions generated by past data, a rough estimate of the churn rate for a certain time period can be generated, such as a quarter or a year. A Machine Learning algorithm can be used for the same to get the value of the prediction. This value can be further tweaked by incorporating customer feedback which can affect the churn rate. Thus, the value generated by the predictions and the feedback will be more accurate.**

Keywords—Customer churn, Machine Learning, Natural Language Processing, Web Scraping

I. Introduction

With the popularity of 5G increasing day by day and all of the major telecommunication companies scrambling to get their infrastructure altered for the widespread use of 5G, these companies must make sure that their ventures continue to be well funded for now as well as in the future. Thus, a major factor that these companies take into consideration is their rate of churn, which is the rate at which existing customers leave a company. If a company has a low churn rate, it could mean that its customers are unsatisfied with it's service. Consider the Indian telecommunication sector's giant: Reliance Jio. The first quarter of 2022 was particularly fruitful for them as their churn rate fell to 2% from 3.7% [18].

Thus, to ascertain the success or even the survival of a company, companies must aim to predict the churn rate of a certain time period so that they can accordingly alter policies and improve their service and increase their revenue. These predictions can be made by considering the factors which led to customers leaving the service in the previous time period and by analyzing the feedback given by previous customers.

When a customer gives negative feedback on a product or service, it is an indication that they may choose to leave said service in the near future. On review sites, it's noted that for Reliance Jio, there are a certain number of positive reviews and a certain number of negative reviews from customers of Reliance Jio. If a customer has posted a negative review between April and June 2022, it can be summarized that said customer intends to leave the company's services before the quarter ends. Thus, it can be concluded that the feedback given by customers is an important consideration when one tries to analyze the churn rates of any telecom company.

Thus, to build the system which will take into account previous factors affecting churn as well as customer feedback, an adequate dataset is needed, where attributes such as Age, Gender, Package, Services Used and other such details are listed, a machine learning model can be fit to predict the churn for the next time period solely based on these attributes. Now, if a web scraper is added to get reviews from retail sites and search engines, and a natural language processor is used to find the emotion behind the reviews, the positivity and negativity of the reviews can be ascertained and by the number of positive and negative reviews, the output from the machine learning model can be altered to give a more accurate prediction of the churn rate in the telecommunication industry.

II. Related Work

Methods such as Machine Learning algorithms, Deep learning algorithms, Probabilistic Graphical Models, and Ensemble models were used to predict customer churn in the Telecommunication industry.

The machine learning techniques which were used by other authors in [4], [5], [6], [13] were Random Forest (RF), Logistic Regression (LR), Hoeffding algorithm and Support Vector Machine (SVM) among others. These algorithms performed relatively well with the accuracy of the models ranging from 84 to 95. In [4], the authors concluded that the Random Forest and J48 algorithms performed the best when compared to the other algorithms when tested on the churn-bigml dataset. In [5], SVM and RF were used to fit a model on a dataset consisting of 3333 users, in a 70-30 training and

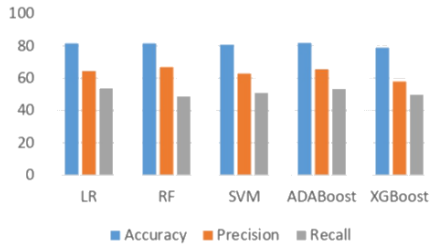
testing split respectively. Among SVM and RF individually, the RF algorithm performed better. The Hoeffding and Logistic Regression algorithms were compared in [6]. The LR algorithm was found to be 1.1% more precise than Hoeffding algorithm and the Recall of LR was greater by 2.04%. When the CART algorithm was fitted on a dataset given by Indian Telecommunication Service Providers, the model was 94.3% accurate. Additionally, when J48 was used, the model was 82.07% accurate. In [13], the Feature Selection method outperformed all of the other machine learning algorithms used.

The Deep Learning techniques which were used by other authors proved to be superior to the Machine Learning method by 5.77%. The Deep Learning algorithms also produced much higher metrics, as observed in papers [2], [8], [9], and [13]. The PPFCM-ANN, PBCPP, and CNN methods were used in [2] with the accuracy of the algorithms lying between 96 and 99 which are much higher than the values observed when the datasets were fitted with machine learning algorithms. When CPNN and fuzzyARTMAP were fit on a dataset in [9], the accuracies were 89.89% and 91.67% respectively.

Probabilistic Graphical Models were used to fit models on datasets in [3], [12], and [13], which yielded very good accuracies. In [3], the authors created a Bayesian Network which was found to be 69.8% accurate. According to certain factors, the authors categorized the motivation for customers to leave the service. Pronaya Bhattacharya *et al* used Naive-Bayes classification in their model [12], based on a dataset that had 3333 instances. The model was 97.89% accurate. Naive Bayes was also analyzed in [13] and it yielded an accuracy of 88.46% with feature selection and 86.80% without.

The averaging ensemble model, which was a combination of a few different algorithms, developed in [5] outperformed all of the individual models by at least 1% accuracy and the selective ensemble model achieved an accuracy of 93.9%, the highest observed in the study.

In order to find the metrics of more algorithms, IBM's WA_Fn-UseC_-Telco-Customer-Churn, as used in [1], and the churn-bigml-80 dataset were used, as in [4]. This analysis helped in knowing the effectiveness of different algorithms on the datasets used in previous papers. The first dataset contained information such as customer age, gender, the existence of partners and dependents, and finally their usage of the services provided by the telecommunication company. The dataset had 7043 rows and 21 columns. The data was then split into 75% data available for training and 25% for testing, after which the models were created. The Accuracy, Precision, and Recall metrics were used. The results were as such: The Logistic Regression algorithm had an accuracy of 81.4562%, a precision of 64.4257%, and a recall of 53.6130%. The Random Forest algorithm had an accuracy of 81.5699%, a precision of 66.7731%, and a recall of 48.7179%. The Support Vector Machine algorithm had an accuracy of 80.6598%, a precision of 62.8242%, and a recall value of 50.8158%. The ADABoost algorithm had an accuracy of 81.7974%, a precision of 65.6160%, and a recall value of 53.3799%. The XGBoost algorithm had an accuracy of 78.8964%, a precision of 57.8804%, and a recall value of 49.6503%. The results are shown in graphical form in Fig. 1:

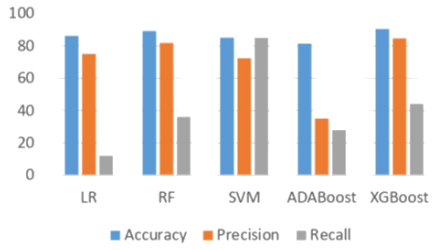


LR: Logistic Regression
RF: Random Forest
SVM: Support Vector Machine

Fig. 1: Analysis of algorithms on IBM's dataset.

On observation of Fig 1, the conclusion is that for this dataset the Adaptive Boosting classifier seems to be the best algorithm for this dataset.

The churn-bigml-80 dataset was then taken into consideration which contained information such as the customer's location, their contract details with the company, and also their activity. The data had 667 rows and 20 columns. The data was pre-processed and models were fit. In this case, the Logistic Regression algorithm was 86.2275% accurate, 75% precise, and had a recall value of 12%. The Random Forest algorithm was 89.2215% accurate, 81.8181% precise, and had a recall value of 36. The Support Vector Machine algorithm had an accuracy of 85.0299%, it was 72.3009% precise, and had a recall value of 85.0299. The ADABoost classifier was 81.4371% accurate, 35% precise, and had a recall value of 28% and the XGBoost algorithm was 90.4191% accurate, 84.6153% precise, and had a recall value of 28%. The results are shown in graphical form in Fig. 2.



LR: Logistic Regression
RF: Random Forest
SVM: Support Vector Machine

Fig. 2: Analysis of algorithms on churn-bigml

From Fig. 2, it was observed that the Support Vector Machine classifier seems to be the best algorithm for this dataset.

The workflow for the prediction is currently as such:

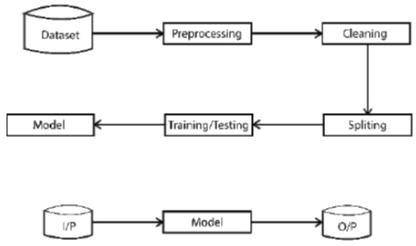


Fig. 3: Original workflow of the prediction

From the entirety of the research taken into consideration, it could be concluded that. that Probabilistic Graphical Models and Deep Learning Models are the best algorithms to predict the rate of customer churn, it's observed that when an algorithm like CNN is used on different datasets, there may be varying accuracy. Thus, if the dataset has problems, that too needs to be taken into account.

In order to find out the problems in datasets, in [16], the authors attempted to find the readiness of a dataset for machine learning. When the data was divided into "interconnects", they found that the first interconnect itself had 71.2653% bad data. Upon further analysis, there were more quality issues detected which would be harder to find for a non-specialist. Hence, while data may be pre-processed and cleaned, the dataset may still have underlying issues. Thus, with this in mind, the quality of the dataset needs to be considered before considering that an algorithm is the best for it.

An attempt to find the influences on customer churn was made in [7]. A social graph was built to discover a person's relationships which would alter in some way when a person churned. The authors assumed that restoring an influencer's pleasure will remove their churn influence on the social network. The study highlighted that there is a weakness in the current customer churn methods.

Many factors could affect the rate of customer churn in a certain city. If a company ceases its operations in a certain city, the churn rate for that quarter or the time period in consideration will be exceptionally high. There could also be a stoppage of certain services provided by the company which could displease some customers and would also result in churn. Additionally, a representative of the company could be involved in a controversial scenario which could also result in churn. A method to take into consideration must be developed in order to get the best prediction possible.

Thus, it's been determined that in order to get the best prediction possible, an additional factor needs to be introduced, which will help improve the accuracy of the prediction and yield better metrics overall.

III. Proposal

The proposed system is: the Adaptive Boosting Classifier using a decision tree classifier as the weak learner can be used to fit a model on IBM's Telco Customer Churn (WA_Fn-UseC_-Telco-Customer-Churn) dataset.

This dataset was chosen as it is quite popular with users; it has been used in over 900 notebooks that have been published on Kaggle, and it was also used in [1]. Additionally, when compared to the other supervised learning models fit on this dataset, the Adaptive Boosting Classifier performed the best with an accuracy of over 81% and that's the reason it was chosen for the machine learning model.

The dataset contains 21 features, out of which 19 are used for fitting the model. The features are crucial for fitting the model, and thus, there is no redundancy in the dataset, eliminating the need for dimensionality reduction.

As seen from [16], the dataset itself may have some issues, making the workflow shown in Fig. 3 problematic. Thus, to improve the prediction, another factor needs to be considered while considering the prediction of customer churn [7]. This factor could be the reviews left by customers about a particular telecommunication service provider.

The reviews left by customers are important for the prediction, as seen in [17] Thus if it's considered that during a quarter, a customer cancels their contract with a particular telecom company, they are likely to leave a review on sites such as Google and Yelp. A negative review will be an indicator of problems at the company, and would increase the churn rate by a certain percentage. On the other hand, if the review is positive, it shows that for the time period in question, the customers are satisfied and there will not be any increase in churn rate.

The total rate of Churn will be affected by the reviews taken into consideration. The calculation of customer churn will require the result from the machine learning model as well as customers' reviews left on the internet, which will not be considered while fitting the model as the customers' reviews are not mapped to the comments they leave on the internet. For the implementation of this system, a machine learning model, a web scraper, a natural language processor, and a database will be required. There are technologies available that can enable one to do web scraping. David Mathew Thomas *et al* used the XPath technique to develop a project called Scrapy [11]. The Beautiful Soup library can also get data from HTML files and save it to a database. Thus, the web scraper can get customer reviews from a website and they can be used for further analysis, by the natural language processor. In [14], the authors used BERT, RoBERTa, and XLnet models in order to find the emotion of a sentence. NLP technology can also be applied to find the emotion of a review, be it happiness or sadness.

The methodology of the proposed system is explained below:

The proposed system will use IBM's Telco Customer Churn dataset to fit a model using the Adaptive Boosting algorithm. The prediction from the model using new independent variables can be stored in MP in Table I. Next, the web scraper will take reviews from sites such as Google and Yelp and save them in Table II. The site, review, and star-value (if exist) can be saved in their respective columns. The Review column can be analyzed using the Natural Language Processor, the output of which will be saved in Alt. Alt will be the value by which, the predicted churn rate will increase or decrease. This column, Alt, can be mapped to Table I of the database, where it will affect the value of MP.

For example: If the predicted rate of churn in Table I (MP) is 2.2% and Alt is 0.6%, the final prediction (FP) will be MP + Alt = FP = 2.8%. This will give the final prediction of customer churn for the telecommunication company in question. In Tables, I and II, the proposed schema for the database is given.

TABLE I. TABLE 1 OF DATABASE

	Database Columns		
No.	**MP**	**Alt**	**FP**
1	2.2	0.9	3.1
2	2.1	0.5	2.7
…			
n	2.5	0.0	2.5

MP: The prediction made by the machine learning model.
Alt: The percentage by which the prediction by the machine learning model should be altered.
FP: The final prediction, inclusive of the prediction made by the machine learning model and the customer's feedback.
Note: Values in the table are only for reference.

TABLE II. TABLE 2 OF DATABASE

	Database Columns			
No.	**Site**	**StarValue**	**Review**	**Alt**
1	TrustPilot	1	Bad	0.9
2	G2	3	Fair	0.5
…				
n	MouthShut	5	Good	0.0

Site: The site from which the review has been scraped
StarValue: The rating given by a customer (out of 5)
Review: Plaintext of the review given by the customer.
Alt: The percentage by which the prediction by the machine learning model should be altered.

Note: Values in the table are only for reference.

The proposed predictor system will work as such:

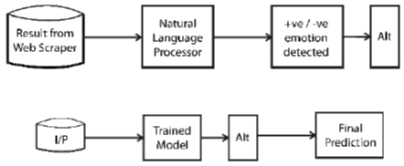


Fig. 4: Workflow of proposed system

Fig 4 shows the workflow of the prediction system. It shows that the data gathered from the web scraping application will be segregated into positive or negative emotion values based on the Natural Language Processor. This gives the value (Alt) by which the model prediction is affected. This, in turn, gives the Final Prediction.

Thus, with this prediction system in place, a more accurate prediction will be generated which will help better predict the churn rate of a telecom company.

IV. CONCLUSION AND FUTURE WORK

The churn rate of a telecommunication company is the rate at which existing customers leave the service, in a particular time period. The churn rate for a telecom company is important as a company needs to consider how they will cope with the loss of revenue and how much will their revenue be impacted by such leaving customers. Analysis of customer churn using various machine learning algorithms has been done in the past but for the most accurate prediction, customer feedback needs to be taken into account as well, because when customers are dissatisfied with a service, they tend to review the companies as such on popular sites. Thus, with a system consisting of a machine learning model, a web scraper, and a natural language processor to analyze the emotion in a review, the churn rate of a telecom company can be accurately predicted. With the best algorithm in use and the most reliable, unbiased sites being used for reviews, the accuracy of the prediction will increase.

In the future, we will be able to build a fully functional web-based application that can accurately predict the rate of customer churn for a given telecommunication company.

REFERENCES

1. Pan Tang (2020), Telecom Customer Churn Prediction Model Combining K-means and XGBoost Algorithm, *International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*
2. Priya Gopal, Dr. Nazri Bin MohdNawi (2021), A Survey on Customer Churn Prediction using Machine Learning and data mining Techniques in E-commerce, *IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*
3. Peng Sun, Xin Guo, Yunpeng Zhang, Ziyan Wu (2013), Analytical Model of Customer Churn Based On Bayesian Network, *2013 Ninth International Conference on Computational Intelligence and Security*
4. Irfan Ullah, Basit Raza1, Ahmad Kamran Malik, Muhammad Imran, Saif ul Islam, Sung Won Kim (2019), A Churn Prediction Model using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector *IEEE Access*
5. Ahmad Hammoudeh, Malak Fraihat, Mahmoud Almomani (2019), Selective Ensemble Model for Telecom Churn Prediction, *IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*
6. Mr. Anurag Bhatnagar, Dr. Sumit Srivastava (2020), Performance Analysis of Hoeffding and Logisitic Algorithm for Churn Prediction in Telecom Sector, *International Conference on Computation, Automation and Knowledge Management (ICCAKM) Amity University*
7. Sara Motahari, Taeho Jung, Hui Zang, Krishna Janakiraman, Xiang-Yang Li, Kevin Soo Hoo (2014), Predicting the Influencers on Wireless Subscriber Churn, *IEEE WCNC'14 Track 4 (Services, Applications, and Business)*
8. Xin Hu, Yanfei Yang, Lanhua Chen, Siru Zhu (2020), Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network, *IEEE 5th International Conference on Cloud Computing and Big Data Analytics*
9. Ramakanta Mohanty, Jhansi Rani K (2015), Application of Computational Intelligence to predict churn and non-churn of customers in Indian Telecommunication, *International Conference on Computational Intelligence and Communication Networks*
10. Ankit Verma, Chavi Kapoor, Abhishek Sharma, Biswajit Mishra (2021), Web Application Implementation with Machine Learning, *2nd International Conference on Intelligent Engineering and Management (ICIEM)*
11. David Mathew Thomas, Sandeep Mathur (2019), Data Analysis by Web Scraping using Python, *Third International Conference on Electronics Communication and Aerospace Technology [ICECA 2019]*
12. Pronaya Bhattacharya, Akhilesh Ladha, Ashwani Kumar, Ashwin Verma, Umesh Bodkhe, (2020) BaYcP: A novel Bayesian customer Churn prediction scheme for Telecom sector, *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*
13. Yakub K. Saheed, Moshood A. Hambali (2021), Customer Churn Prediction in Telecom Sector with Machine Learning and Information Gain Filter Feature Selection Algorithms, *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*
14. F. A. Acheampong; H. Nunoo-Mensah; Wenyu Chen (2021), Recognizing Emotions from Texts Using an Ensemble of Transformer-Based Language Models, *International Conference on Wavelet Active Media Technology and Information Processing (ICWAMTIP)*
15. Seo-Hui Park, Byung-Chull Bae, Yun-Gyung Cheong (2020), Emotion Recognition from Text Stories Using an Emotion Embedding Model,

International Conference on Big Data and Smart Computing (BIGCOMP)

16. Philip Hart; Lijun He, Tianyi Wang, Vijay S. Kumar, Kareem Aggour, Arun Subramanian, Weizhong Yan (2022), Application of Big Data Analytics and Machine Learning to Large-Scale Synchrophasor Datasets: Evaluation of Dataset 'Machine Learning-Readiness', *IEEE Open Access Journal of Power and Energy*
17. Jo Mackiewicz, Dave Yeats, Thomas Thornton (2016), The Impact of Review Environment on Review Credibility, *IEEE Transactions on Professional Communication*
18. Tanay Singh Thakur (2022), Reliance Jio Subscriber Churn Dipping, ARPU Rising a Good Thing, *TelecomTalk*

COMPETITIONS

Sr. No.	Project Competition	Venue	Date	Achievement (winner/runner up/participation certificate)
1	IEEE ICAST	KJSIT	02/12/2022	Participation Certificate
2	INTECH-2022	KJSIT	25/03/2023	Participation Certificate
3	XZIBIT	KCCEMSR	8/4/2023	Participation Certificate

CERTIFICATES

SOMAIYA VIDYAVIHAR
K J Somaiya Institute of Engineering & Information Technology
An Autonomous Institute permanently affiliated to University of Mumbai.

University of Mumbai

Organizes

5th IEEE-International Conference on Advances in Science and Technology (ICAST)-2022

ICAST KJSIEIT

CERTIFICATE

IEEE BOMBAY SECTION

This is to certify that

Anirudh Bhattacharya

has participated/presented a paper titled

Feedback Based Telecom Churn Prediction Using Machine Learning

Published at **IEEE Xplore** on **"5th IEEE International Conference on Advances in Science and Technology (ICAST)-2022 "** organized by **K J Somaiya Institute of Engineering and Information Technology (KJSIEIT) , Sion, Mumbai-22** in association with **University of Mumbai** and technically co-sponsored by **IEEE Bombay Section** on 2nd and 3rd of December, 2022.

Dr. Sunita R. Patil
Convenor-ICAST 2022
Vice Principal, KJSIEIT

Dr. Suresh K. Ukarande
Chairperson-ICAST 2022
Principal, KJSIEIT

SOMAIYA VIDYAVIHAR
K J Somaiya Institute of Engineering & Information Technology
An Autonomous Institute permanently affiliated to University of Mumbai.

University of Mumbai

Organizes

5th IEEE-International Conference on Advances in Science and Technology (ICAST)-2022

ICAST KJSIEIT

CERTIFICATE

IEEE BOMBAY SECTION

This is to certify that

Saumya Shah

has participated/presented a paper titled

Feedback Based Telecom Churn Prediction Using Machine Learning

Published at **IEEE Xplore** on **"5th IEEE International Conference on Advances in Science and Technology (ICAST)-2022 "** organized by **K J Somaiya Institute of Engineering and Information Technology (KJSIEIT) , Sion, Mumbai-22** in association with **University of Mumbai** and technically co-sponsored by **IEEE Bombay Section** on 2nd and 3rd of December, 2022.

Dr. Sunita R. Patil
Convenor-ICAST 2022
Vice Principal, KJSIEIT

Dr. Suresh K. Ukarande
Chairperson-ICAST 2022
Principal, KJSIEIT

SOMAIYA VIDYAVIHAR
K J Somaiya Institute of Engineering & Information Technology
An Autonomous Institute permanently affiliated to University of Mumbai.

University of Mumbai

Organizes

5th IEEE-International Conference on Advances in Science and Technology (ICAST)-2022

CERTIFICATE

IEEE BOMBAY SECTION

This is to certify that

Mink Shethia

has participated/presented a paper titled

Feedback Based Telecom Churn Prediction Using Machine Learning

Published at **IEEE Xplore** on "**5th IEEE International Conference on Advances in Science and Technology (ICAST)-2022** " organized by **K J Somaiya Institute of Engineering and Information Technology (KJSIEIT) , Sion, Mumbai-22** in association with **University of Mumbai** and technically co-sponsored by **IEEE Bombay Section** on 2nd and 3rd of December, 2022.

Dr. Sunita R. Patil
Convenor-ICAST 2022
Vice Principal, KJSIEIT

Dr. Suresh K. Ukarande
Chairperson-ICAST 2022
Principal, KJSIEIT

SOMAIYA VIDYAVIHAR
K J Somaiya Institute of Technology

IET The Institution of Engineering and Technology

Participation Certificate

Anirudh Bhattacharya

has participated in

National Level Poster cum Project Competition "**KJSIT-IET-INTECH 2K23**"

organized by

K J Somaiya Institute of Technology, Sion, Mumbai

in association with

The Institution of Engineering and Technology, Mumbai Local Network

on **March 25, 2023**

Vikrant Sankhe

Mr. Vikrant Sankhe
Chairman, IET Mumbai LN

Dr. Sunita Patil
Vice Principal, KJSIEIT

Dr. Suresh Ukarande
Principal, KJSIEIT

SOMAIYA VIDYAVIHAR
K J Somaiya Institute of Technology

The Institution of Engineering and Technology

Participation Certificate

Saumya Shah

has participated in

National Level Poster cum Project Competition "**KJSIT-IET-INTECH 2K23**"

organized by

K J Somaiya Institute of Technology, Sion, Mumbai

in association with

The Institution of Engineering and Technology, Mumbai Local Network

on **March 25, 2023**

Mr. Vikrant Sankhe	Dr. Sunita Patil	Dr. Suresh Ukarande
Chairman, IET Mumbai LN	Vice Principal, KJSIEIT	Principal, KJSIEIT

SOMAIYA VIDYAVIHAR
K J Somaiya Institute of Technology

Participation Certificate

Mink Shethia

has participated in

National Level Poster cum Project Competition "**KJSIT-IET-INTECH 2K23**"

organized by

K J Somaiya Institute of Technology, Sion, Mumbai

in association with

The Institution of Engineering and Technology, Mumbai Local Network

on **March 25, 2023**

Mr. Vikrant Sankhe	Dr. Sunita Patil	Dr. Suresh Ukarande
Chairman, IET Mumbai LN	Vice Principal, KJSIEIT	Principal, KJSIEIT


EXCELSSIOR EDUCATION SOCIETY'S
K.C. COLLEGE OF ENGINEERING &
MANAGEMENT STUDIES & RESEARCH,
THANE(E)
CERTIFICATE OF PARTICIPATION
THIS CERTIFICATE IS PRESENTED TO
Anirudh Bhattacharya
For participating in
XZIBIT-2023 NATIONAL LEVEL PROJECT COMPETITION
organized by the Department of Electronics & Telecommunication,
Information Technology & Computer Engineering of
K.C. College of Engineering Management Studies & Research, Thane (E)
on 8th & 10th April 2023
Mrs. Amarja Adgaonkar
(Convenor)
Dr. Vilas Nitnaware
(Principal)


EXCELSSIOR EDUCATION SOCIETY'S
K.C. COLLEGE OF ENGINEERING &
MANAGEMENT STUDIES & RESEARCH,
THANE(E)
CERTIFICATE OF PARTICIPATION
THIS CERTIFICATE IS PRESENTED TO
Saumya Shah
For participating in
XZIBIT-2023 NATIONAL LEVEL PROJECT COMPETITION
organized by the Department of Electronics & Telecommunication,
Information Technology & Computer Engineering of
K.C. College of Engineering Management Studies & Research, Thane (E)
on 8th & 10th April 2023
Mrs. Amarja Adgaonkar
(Convenor)
Dr. Vilas Nitnaware
(Principal)


EXCELSSIOR EDUCATION SOCIETY'S
K.C. COLLEGE OF ENGINEERING &
MANAGEMENT STUDIES & RESEARCH,
THANE(E)
CERTIFICATE OF PARTICIPATION
THIS CERTIFICATE IS PRESENTED TO
Mink Shethia
For participating in
XZIBIT-2023 NATIONAL LEVEL PROJECT COMPETITION
organized by the Department of Electronics & Telecommunication,
Information Technology & Computer Engineering of
K.C. College of Engineering Management Studies & Research, Thane (E)
on 8th & 10th April 2023
Mrs. Amarja Adgaonkar
(Convenor)
Dr. Vilas Nitnaware
Made for free with Certify'em

PLAGIARISM REPORT

churn report

ORIGINALITY REPORT

15%	7%	11%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Madhura Phadke, Anirudh Bhattacharya, Mink Shethia, Saumya Shah. "Feedback Based Telecom Churn Prediction Using Machine Learning", 2022 5th International Conference on Advances in Science and Technology (ICAST), 2022 Publication	9%
2	Submitted to Korea National University of Transportation Student Paper	2%
3	Submitted to Somaiya Vidyavihar Student Paper	2%
4	www.researchgate.net Internet Source	<1%
5	www.science.gov Internet Source	<1%
6	Xin Hu, Yanfei Yang, Lanhua Chen, Siru Zhu. "Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network", 2020 IEEE 5th International	<1%