Telecommunication customer churn prediction using machine learning techniques

A PROJECT REPORT

Submitted by

DESIKA V - 312319205028
SURYAPRABHA S - 312319205171

of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

St. JOSEPH'S COLLEGE OF ENGINEERING

(An Autonomous Institution)

St. Joseph's Group of Institutions

Jeppiaar Educational Trust

OMR, Chennai 600 119

ANNA UNIVERSITY: CHENNAI

April-2023

ANNA UNIVERSITY: CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project report **“Telecommunication customer churn prediction using machine learning techniques**” is the bonafide work of **DESIKA V (312319205028)** and **SURYAPRABHA S (312319205171)** who carried out the project under my supervision.

SIGNATURE
Supervisor,
Ms. J. Divya,M.Tech.,(Ph.D),

Assistant Professor,
Department of IT,
St.Joseph’s College of Engineering,
OMR, Chennai- 600119.

SIGNATURE
Head of the department,
Dr. V. Muthulakshmi, M.E.,Ph.D,

Professor,
Department of IT,
St.Joseph’s College of Engineering,
OMR, Chennai- 600119.

CERTIFICATE OF EVALUATION

COLLEGE NAME **:** St. Joseph's College of Engineering, Chennai-600119.

BRANCH **: B.TECH., IT** (Information Technology)

SEMESTER VIII

SL. NO	**NAME OF THE STUDENT**	**TITLE OF THE PROJECT**	**NAME OF THE SUPERVISOR WITH DESIGNATION**
1 2	DESIKA V (312319205028) SURYAPRABHA S (312319205171)	Telecommunication customer churn prediction using machine learning techniques.	Ms. J. DIVYA, M.Tech.,(Ph.D)., ASSISTANT PROFESSOR

The report of the project work submitted by the above students in partial fulfillment for the award of Bachelor of Technology Degree in Information Technology of Anna University was confirmed to be report of the work done by the above students and then evaluated.

Submitted to Project and Viva Examination held on________________.

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

At the outset we would like to express our sincere gratitude to the beloved **Chairman, Dr.Babu Manoharan, M.A.,M.B.A.,Ph.D.,** for his constant guidance and support.

We would like to express our heartfelt thanks to our respected **Managing Director Mrs. S. Jessie Priya, M.Com.,** for her kind encouragement and blessings.

We wish to express our sincere thanks to our **Executive Director Mr. B. Shashi Sekar, M.Sc.,** for providing ample facilities in the institution.

We express our deepest gratitude and thanks to our beloved **Principal Dr.Vaddi Seshagiri Rao, M.E., M.B.A., Ph.D., F.I.E.,** for his inspirational ideas during the course of the project.

We wish to express our sincere thanks and gratitude to **Dr.V. Muthulakshmi, M.E.,Ph.D.,** Head of the Department, Department of Information Technology, St. Joseph's College of Engineering for her guidance and assistance in solving the various intricacies involved in the project.

It is with deep sense of gratitude that we acknowledge our indebt ness to our supervisor **Ms. J. Divya, M.Tech.,(Ph.D).,** for her expert guidance and connoisseur suggestion.

Finally we thank our department staff members who helped us in the successful completion of this project.

ABSTRACT

Customer churn is a significant issue and one of the top issues for big businesses. Companies are working to create methods to predict probable customer churn because it has a direct impact on their revenues, particularly in the telecom industry. The market for telecommunications services is becoming more and more competitive, and acquiring new consumers is significantly more expensive than keeping the ones you already have. In order to reduce customer churn, it is crucial to identify the variables that contribute to this churn. Researchers and analysts use Customer Relationship Management (CRM) data to identify customers likely to churn using machine learning models. The key contribution is the creation of a churn prediction model that helps telecom providers identify consumers who are most likely to experience churn. The cost sensitive learning algorithm handles the significant class imbalance that exists between the two classes. As a result, we investigated the machine learning model, Random Forest, and data transformation methods to improve the accuracy of customer churn prediction in the telecommunications industry. To optimize the prediction model, the best hyperparameters were compared and chosen. The Genetic Algorithm tuning in the Random Forest Classifier algorithm was used in this project to optimize the Machine Learning model, which aims to identify the best input values for the best possible output values or results. The random forest's efficiency has been increased from 73% to 80% accuracy after implementing genetic algorithm tuning with an AUC of 0.839.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBREVIATION	DEFINITION
CNN	Convolutional Neural Network
ANN	Artificial Neural Network
KNN	K Nearest neighbor
SVM	Support Vector Machine
RF	Random Forest
GBM	Gradient Boosting Machine
CRM	Customer Relationship Manage
ML	Machine Learning
CCP	Customer Churn Prediction
CLTV	Customer Lifetime Value
CCA	Cost of Client Acquisition
CP	Customer Prediction
CC	Churn Customer
DF	Decision Forest
LMT	Logistic Model Tree
FT	Functional Tree
KDD	Knowledge Discovery Database
RNN	Recurrent Neural Network
PCA	Principal Component Analysis

CHAPTER 1
INTRODUCTION

Churners are customers that discontinue using a product or service for a predetermined amount of time. The person who has chosen to no longer use a company's services is known as a churn in a telecommunications company. The churn model identified the person who is most likely going to leave the company soon. As a frequent use for data mining technique, many companies construct models akin to churns. Globally distributed mobile phone companies are almost ready to develop their own churn model. Furthermore, churn results can be effectively used for a variety of other objectives in addition to customer retention. The first step in creating a model is actually to use a churn management method. Instead of choosing a single approach that has the best lift, the project generally needs a churn model in the best way. So, as a long-term standard, we have created an automated application. The customer of one business could also be a customer ofone or more telecommunications companies in this digital age. Some of us might use various carriers depending on the distance, while others would do so dependingon the various plans each carrier offers. Customer experience often yields useful insights when machine learning is used for analysis. Some customers occasionally switch service providers. The calling rate may increase or decrease depending on the various job duties. Depending on the data's accessibility, numerous circumstances may be reflected.

1.1. SYSTEM OVERVIEW

The Telco customer churn data contains information about a fictional telco companythat provided home phone and Internet services to 7043 customers in California in Q3. It indicates which customers have left, stayed, or signed up for their service. Satisfaction Score, Churn Score, and Customer Lifetime Value (CLTV) index.

- To apply Exploratory Data analysis to infer patterns from the data collected.
- To predict the number of customers that are about to churn usingthe generated patterns.
- To create a novel model to implement and deploy on anapplication.

1.2. SCOPE OF THE PROJECT

The percentage of customers that ceased using a company's product or service duringa given time period is known as customer churn. Especially in businesses where the Cost of Client Acquisition (CAC) is high, such as technology, telecom, finance, and so on, predicting customer churn is a challenging but crucial business problem. For organizations, the capacity to foresee that a particular customer is very likely to leavewhile there is still time to take action is a large additional potential revenue source. The main objective of the customer churn predictive model is to actively engage high-risk churning customers in order to keep them as clients. A significant problem in the communications sector is customer attrition. A Research the University of Zambia found that telecom provider turnover rates can rise to 67% annually. One ofthe main contributing causes was dissatisfaction with various services. By classifying your consumers according to a classification rule and specific criteria, binary classification can assist with client segmentation while working with massive telecoms datasets. Using machine learning strategies and a telecommunications dataset, the number of users who will most likely choose not to use a service can be forecasted. With the help of multiple client touchpoints, most notably a CRM system, the model enables to identify at-risk customers.

CHAPTER 2
LITERATURE SURVEY

[1] Shabankareh, M.J., Shabankareh, M.A., Nazarian, A., Ranjbaran, A. and Seyyedamiri, N., 2022. "A Stacking-Based Data Mining Solution to Customer Churn Prediction". Journal of Relationship Marketing, 21(2), pp.124-147.

In today's competitive world, organizations are in a constant struggle to retain their current customers while attracting new customers through various methods. Customer churn is a major challenge in different industries and companies. Despite their initial successful attempts at attracting customers, organizations soon face the fact that their current customers may turn away from their rivals. By identifying churn candidates, organizations will be able to guarantee their future success by revising their customer relationship management policy. Analyzing the data of the telecommunications industries, this study provided an effective early-churn-detection solution using modern techniques by stacking data mining algorithms. Research findings indicate that integrating support vector machines (SVMs) with the chi- square automatic interaction detection (CHAID) decision tree can yield the best outcome. The results show the proper accuracy of the proposed churn prediction solution. In addition, stacking contributed to improved customer churn detection results.

[2] Sudharsan, R. and Ganesh, E.N., 2022. "A Swish RNN based customer churnprediction for the telecom industry with a novel feature selection strategy", Connection Science, 34(1), pp.1855-1876.

Owing to saturated markets, fierce competition, dynamic criteria, along with the introduction of new attractive offers, the considerable issue of customer churn was faced by the telecommunication industry. Thus, an efficient Churn Prediction (CP) model is required for monitoring customer churn. Therefore, a novel framework to predict customer churn has been proposed through a deep learning model namely Swish Recurrent Neural Network (S-RNN). Finally, SRNN is adapted to classify theChurn Customer (CC) and a normal customer. If the result is a churned customer, thenew utilization history is analyzed for the process. Whereas, the number of churn customers based on the area network usage has not recognized this framework Owing to saturated markets, fierce competition, dynamic criteria, along with the introduction of new attractive offers, the considerable issue of customer churn was faced by the telecommunication industry. Thus, an efficient Churn Prediction (CP) model is required for monitoring customer churn. Therefore, a novel framework to predict customer churn through a deep learning model has been proposed, namely Swish Recurrent Neural Network (S-RNN). Finally, S-RNN is adapted to classify the ChurnCustomer (CC) and a normal customer. If the result is a churn customer, network utilization history is analyzed for retention process.

[3] Samah Wael Fujo, Suresh Subramanian and Moaiad Ahmad Khder "Customer Churn Prediction in Telecommunication Industry Using Deep Learning ",Information Sciences Letters- An International Journal,2022

Without proper analysis and forecasting, industries will find themselves repeatedly churning customers, which the telecom industry in particular cannot afford. A predictable model for customers will allow companies to retain current customers and to obtain new ones. Deep-BP-ANN implemented in this study using two feature selection methods, Variance Thresholding and Lasso Regression, in addition, our model strengthened by early stopping technique to stop training at right time and prevent overfitting. We compared the efficiency of minimizing overfitting between dropout and activity regularization strategies for two real datasets: IBM Telco and Cell2cell. Different evaluation approaches used: Holdout, and 10-fold cross- validation to evaluate the model's efficiency. To solve unbalanced issue, the Random Oversampling technique was used to balance both datasets.The model implemented performs well with lasso regression for feature selection, early stopping technique to pick the epochs, and large numbers of neurons (250) into the input and hidden layers, and activity regularization to minimize overfitting for both datasets. In predicting customer churn, our findings outperform ML techniques: XG_Boost, Logistic_Regression, Naïve_Bayes, and KNN. Moreover, our Deep-BP- ANN model's accuracy outperforms the existing deep learning techniques that use holdout or 10- fold CV for the same datasets.

[4] Tianyuan Zhang ,Sérgio Moro and Ricardo F. Ramos, Big Data Analytics, "A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation",Privacy and Visualization,2022

Numerous valuable clients can be lost to competitors in the telecommunication industry, leading to profit loss. Thus, understanding the reasons for client churn is vital for telecommunication companies. The aim was to develop a churn prediction model to predict telecom client churn through customer segmentation. Data were collected from three major Chinese telecom companies, and Fisher discriminant equations and logistic regression analysis were used to build a telecom customer churn prediction model. According to the results, it can be concluded that the telecom customer churn model constructed by regression analysis had higher prediction accuracy (93.94%) and better results. The model will help telecom companies efficiently predict the possibility of and take targeted measures to avoid customer churn, thereby increasing their profits.

[5] Yajun Liu, Jingjing Fan, Jianfang Zhang, Xinxin Yin & Zehua Song , "Research on telecom customer churn prediction based on ensemble learning",journal of Intelligent Information Systems ,2022

As the market in the telecom industry becomes saturated and competition between telecom operators heats up, preventing customer churn has become a company's top concern. It is, therefore, crucial to identify customers who are likely to churn and the reasons, as it directly impacts the company's revenue. The main contribution lies in the multidimensional data preprocessing, feature extraction and processing of the dataset provided by the telecom operator. Then, the k-means algorithm is used to cluster different consumer groups, which in turn analyses the factors of concern to different consumer groups and makes targeted suggestions. Finally, to improve the effectiveness and robustness of the model, ensemble learning is introduced into the telecom customer churn field. The experimental results show that the extracted features and the experimental results are satisfactory. Ensemble learning was also applied to the dataset provided by S. Khotijah and it was found that the churn prediction accuracy rate improved regardless of whether the dataset was balanced,especially in the unbalanced dataset.

[6] FE Usman-Hamza, AO Balogun, LF Capretz,"Intelligent Decision Forest Models for Customer Churn Prediction" - Applied Sciences, 2022

Customer churn is a critical issue impacting enterprises and organizations, particularly in the emerging and highly competitive telecommunications industry. It is important to researchers and industry analysts interested in projecting customer behavior to separate churn from non-churn consumers. The fundamental incentive is a firm's intent desire to keep current consumers, along with the exorbitant expense of gaining new ones. Many solutions have been developed to address customer churn prediction (CCP), such as rule-based and machine learning (ML) solutions. However, the issue of scalability and robustness of rule-based customer churn solutions is a critical drawback, while the imbalanced nature of churn datasets has a detrimental impact on the prediction efficacy of conventional ML techniques in CCP. As a result, intelligent decision forest (DF) model has been developed for CCP in telecommunication. Specifically, the prediction performances of the logistic model tree (LMT), random forest (RF), and Functional Trees (FT) as DF models and enhanced DF (LMT, RF, and FT) models based on weighted soft voting and weighted stacking methods have been measured. Extensive experimentation was performed to ascertain the efficacy of the suggested DF models utilizing publicly accessible benchmark telecom CCP datasets. The suggested DF models efficiently distinguish churn from non-churn consumers in the presence of the class imbalance problem. In addition, when compared to baseline and existing ML-based CCP methods, comparative findings showed that the proposed DF models provided superior prediction performances and optimal solutions for CCP in the telecom industry. Hence, the development and deployment of DF-based models for CCP and applicable ML tasks are recommended.

[7] Syed Fakhar Bilal, Abdulwahab Ali Almazroi, Saba Bashir, Farhan Hassan Khan, Abdulaleem Ali Almazroi, PeerJ," An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry", Computer Science, 2022,

Mobile communication has become a dominant medium of communication over the past two decades. New technologies and competitors are emerging rapidly and churn prediction has become a great concern for telecom companies. A customer churn prediction model can provide the accurate identification of potential churners so that a retention solution may be provided to them. The proposed churn prediction model is a hybrid model that is based on a combination of clustering and classification algorithms using an ensemble. First, different clustering algorithms (i.e. K-means, K- medoids, X-means and random clustering) were evaluated individually on two churn prediction datasets. Then hybrid models were introduced by combining the clusters with seven different classification algorithms individually and then evaluations were performed using ensembles. The proposed research was evaluated on two different benchmark telecom data sets obtained from GitHub and Bigml platforms. The analysis of results indicated that the proposed model attained the highest prediction accuracy of 94.7% on the GitHub dataset and 92.43% on the Bigml dataset. State of the art comparison was also performed using the proposed model. The proposed model performed significantly better than state of the art churn prediction models.

[8] Shivani Vaidya and Rajesh kumar Nigam, ”An Analysis of Customer Churn Predictions in the Telecommunications Sector International”, Journal of Electronics Communication and Computer Engineering , Volume 13, 2022

The Telecommunications (telecom) Industry is saturated and marketing strategies are focusing on customer retention and churn prevention. Churning is when a customer stops using a company’s service thereby opting for the next available service provider. This churn threat has led to various Customer Churn Prediction (CCP) studies and many models have been developed to predict possible churn cases for timely retention strategies. So customer churn is an important area of concern. It aims at carrying out a literature review for the past decade reviewing around 50 research papers in the area of telecom churn with two perspectives: mining technique applied and publication year. The review looks at the existing models in the literature, using 30 selected CCP studies particularly from 2014 to 2020. Data preparation methods and churn prediction challenges have also been explored. It has been revealed that Support Vector Machines, Naïve Bayes, Decision Trees and Neural Networks are the mostly used CCP techniques. Feature selection is the most used data preparation method followed by Normalization and Noise removal. Under-sampling is the mostly preferred technique for handling telecom data class imbalances. Imbalanced, large and high dimensional datasets are the key challenges in telecom churn prediction.

[9] Mohammad TabrezQuasim, Adel Sulaiman, Asadullah Shaikh and Mohammed Younus,”Blockchain in churn prediction based telecommunication system on climatic weather application”,Sustainable Computing: Informatics and Systems Volume 35, 2022

For better customer service and customer retention, businesses take proactive measures, including troubleshooting and solving potential challenges promptly. Blockchain technology integrates different recognition techniques of distributed pattern for monitoring database of a dedicated network and has proven a very promising technology. An automated pattern recognition decentralizes and distributes customized specific data. Machine learning, originated from artificial intelligence is primarily related to recognition of behavior patterns. Methods like knowledge discovery in database (KDD) and data mining focus on unsupervised approaches and are widely used in business and climatic weather applications. Blockchain addresses data-security concerns and builds trust by creating distributed ledger. Theft, willful fraud, software and hardware are considered by blockchain in data protection. Blockchain has greater significant feature as it makes the secure data access without enabling the central management entity. Two design features of blockchain technology help in this task. Recommended customer data pattern recognition technique using blockchain may eliminate all these problems. Two kinds of cryptographic algorithms employed in blockchain are asymmetric-key algorithms and hash functions. The asymmetric cryptography approach has been analyzed along with key pair which supports in system security. Recurrent neural network (RNN) and support vector machine (SVM) classifier techniques consider both old customers and new customers as stable. The predictive model aids in identifying customers at churn risk in the telecommunication system. Existing proactive methods are unable to explain difficulties in customer interaction understanding and meeting their genuine needs. The proposed model organizes the customer situation and designs a customer proactive re-engagement over mobile-based telecommunication systems. Performance measures like churn prediction, classifier, confusion matrix, machine learning in the telecommunication system are used to evaluate and validate the results.

[10] P Jeyaprakaash , Sashirekha K, ”Accuracy Measure of Customer Churn Prediction in Telecom Industry using Adaboost over Decision Tree Algorithm”,JOURNAL OF PHARMACEUTICAL NEGATIVE RESULTS, Volume 13,2022

To enhance and predict the accuracy rate of customer churning in the telecommunication industry using Adaboost over Decision Tree algorithm.Materials and methods: Adaboost algorithm and Decision Tree algorithm with sample size (N=10) is executed with multiple training and testing splits for predicting the accuracy for customer churn prediction with 75% as g power value and threshold value as 0.000 and 95% as confidence interval . The performance of these algorithms are calculated based on the rate of accuracy using customer churn dataset.Results and Discussion: The accuracy of predicting customer churn using Adaboost algorithm(90%) and Decision Tree algorithm (73%) is obtained. There was a analytical difference between the Novel Adaboost and Decision Tree algorithm is (p=0.000).

[11] Sulaiman Olaniyi Abdulsalam, Micheal Olaolu Arowolo, Yakub Kayode Saheed, Jesutofunmi Onaope Afolayan, ”Customer Churn Prediction in Telecommunication Industry Using Classification and Regression Trees and Artificial Neural Network Algorithms”,Indonesian Journal of Electrical Engineering and infomatics, 2022

Customer churn is a serious problem, which is a critical issue encountered by large businesses and organizations. Due to the direct impact on the company's revenues, particularly in sectors such as the telecommunications as well as the banking, companies are working to promote ways to identify the churn of prospective consumers. Hence it is vital to investigate issues that influence customer churn to yield appropriate measures to diminish churn. The major objectiv is to advance a model of churn prediction that helps telecom operatives to envisage clients that are most probable to be subjected to churn. The experimental approach uses the machine learning procedures on the telecom churn dataset, using an improved Relief-F feature selection algorithm to pick related features from the huge dataset. To quantify the model's performance, the result of classification uses CART and ANN, the accuracy shows that ANN has a high predictive capacity of 93.88% compared to the 91.60% CART classifier.

[12] Maryam Sadeghi, Mohammad Naderi Dehkordi, Behrang Barekatain & Naser Khani , ”Improve customer churn prediction through the proposed PCA- PSO-K means algorithm in the communication industry”, The Journal of Supercomputing ,2022

Customer churn prediction is one of the areas in Customer Relationship Management that differentiates loyal customers from factors that have a negative impact on business growth. Hence, various machine learning-based methods have been developed by researchers to accurately predict customer churn. However, high dimensionality and low prediction accuracy are problems in identifying averse customers. It presents a new system called PCA-PSO-K Means algorithm, which combines three algorithms: principal component analysis (PCA) for data set feature reduction, K Means algorithm for classification, and particle swarm optimization (PSO) algorithm to optimize K Means in providing initial centroids. The experimental results in the data set of one of the fixed internet providers in Isfahan Province show the improvement of the accuracy of customer churn prediction. The proposed system has an accuracy of 99.77%, a sensitivity of 75%, a specificity of 99.81% and a correlation coefficient of 0.443 ± 0.271. Found.

[13] Teoh Jay Shen, Abdul Samad Bin Shibghatullah, ”Customer Churn Prediction Model for Telecommunication Industry”, Journal of Advances in Artificial Life Robotics Vol. 3(2), 2022

Customer churn is a constant issue that poses a serious threat and is one of the top issues for telecom businesses. The businesses are working to develop and build a strategy to anticipate customer attrition. This is why it is important to identify the sources of client churn. Churn prediction is the process of identifying which consumers are most likely to stop using a service or to cancel their subscription. Because getting new customers frequently costs more than keeping existing ones, it is an important prediction for many firms. The suggested models built use both deep learning and machine learning algorithms. These models were developed and tested using the Python environment and a publicly available dataset from www.kaggle.com. This dataset, which was used in the construction of the models' training and testing phases, includes 7043 rows of customer data with 21 features. Four different machine learning and deep learning algorithms were utilized by these models, including the Artificial Neural Network, Self-Organizing Map, Decision Tree and a hybrid model with the combination of the Self-Organizing Map and Artificial Neural network algorithms.

[14] Zeynep Hilal KİLİMCİ, ”The Effectiveness of Homogeneous Classifier Ensembles on Customer Churn Prediction in Banking, Insurance, and Telecommunication Sectors”,International Journal of Computational and Experimental Science and Engineering, Volume 8, 2022

The prediction of customer churn is a big challenging problem for companies in different sectors such as banking, telecommunication, and insurance. For this reason, analysts and researchers are focus on to investigate reasons behind of customer churn analyzing behaviors of them. An ensemble-based framework is proposed to predict the customer churn in various sectors, namely banking, insurance, and telecommunication. To demonstrate the effectiveness of proposed ensemble framework, k-NN, logistic regression, naïve Bayes, support vector machine, decision tree, random forest, multilayer perceptron algorithms are employed. Moreover, the effects of the inclusion of feature extraction process are investigated. Experiment results indicate that that random forest algorithm is capable to predict churn customers with 89.93% of accuracy in banking, 95.90% of accuracy in telecommunication, and 77.53% of accuracy in insurance sectors when feature extraction procedure is carried out.

[15] P. Ramesh, J. Jeba Emilyn & V. Vijayakumar , ”Hybrid Artificial Neural Networks Using Customer Churn Prediction”,Wireless Personal Communications volume 124, pages1695–1709, 2022

The current wave of technologies with increased awareness among customers and retaining customers has a vital role in the growth of the company. A good indicator of service satisfaction of customers and service quality is customer churn. In order to enable the organizations to understand customers for churning, intelligible and accurate models are needed. There have been several techniques of data mining that were applied for the prediction of churn. The extensive research in Artificial Intelligence has made it feasible to study and learn the aspects accounting for such customer churn. The work presents effective solutions to all these challenging problems in Customer Churn Prediction (CCP). The model uses datasets in the telecommunication industry, the Artificial Neural Networks (ANN), and the Random Forests (RF) to determine the factors that influence consumer churn. A hybrid ANN- based work is proposed for predicting CCP. The results of the experiment proved that the proposed method achieves better levels of performance. The classification accuracy of ANN-4 hidden layers improves its result compared to RF and ANN-2 hidden layers. The maximum accuracy attained by ANN-2 hidden layers is 88.14% and by ANN-4 hidden layers is 90.34%.

CHAPTER 3
SYSTEM ANALYSIS

3.1. EXISTING SYSTEM

In order to estimate customer turnover on an individual basis, telecommunication services use machine learning algorithms. To entice customers to stay, they may provide discounts, exclusive deals, or other benefits. A common categorization issue in the field of supervised learning is a customer churn analysis. Several methods, including data mining, machine learning, and hybrid technologies, have been used to anticipate customer attrition. These methods make it possible for businesses to recognise, foresee, and keep churning customers. Additionally, they support CRM and decision-making in industries. Most of them shared the usage of decision trees, which are a well-known technique for determining customer turnover but are inappropriate for complicated issues. However, prior research indicates that decreasing the data enhances the decision tree's accuracy. Data mining methods are sometimes used for historical analysis and consumer prediction.

3.1.1. DISADVANTAGES OF THE EXISTING SYSTEM

- One of the frequent issues with current customer churn prediction models is the imbalance dataset, and present models are less effective at handling the imbalance dataset since there is not enough data available for the training process.

- The overfitting issue that presents additional data instances is a drawback of existing models like Random Forest and Support Vector Machine. In the training phase, deep learning models like LSTM and CNN are readily over-fitted.

- The current customer churns prediction approach uses Support Vector Machine, but SVM is unable to manage the big dataset. In data relation analysis, Naive Bayes is limited to processing features that are not dependent on one another.

- Due to the lack of sufficient data instances to train the classifier, the majority of the existing models perform less effectively on the imbalance dataset.

3.2. PROPOSED SYSTEM

To predict whether or not a customer would depart, various Machine Learning (ML) techniques and algorithms have been researched. Due to this, the primary models won't use the data, but it will still be analyzed as part of the exploratory data analysis (EDA). Some performance metrics that will be used to evaluate the eventual performance of the best models are accuracy, recall, and precision. An ensemble methodology has been evaluated to identify whether a customer has left the business in the last month. By integrating the predictions of several base estimators built using a particular learning process, ensemble methods aim to boost robustness. The objectives of the proposed work are formulated below: Since the dataset is imbalanced the class-weighted/ cost-sensitive learning has been utilised. In cost- sensitive learning, a weighted cost function is used.

3.2.1. ADVANTAGES OF THE PROPOSED SYSTEM

- The methodology presents a wide range of possibilities for any organisation that wishes to gain from customer attrition prediction. The machine learning model has the ability to assess client behaviour and calculate the likelihood that they would leave.

- Case-sensitivity is inherently faster to parse (albeit only slightly) since it can compare character sequences directly without having to figure out which characters are equivalent to each other.

3.3. REQUIREMENTS SPECIFICATION

3.3.1.HARDWARE REQUIREMENTS

- Intel or ARM-based processor, minimum Pentium or equivalent
- Minimum 512MB RAM
- Minimum 4GB disk space

3.3.2. SOFTWARE REQUIREMENTS

Jupyter notebook:

The JupyterNotebook App is a server-client application that allows editing and running notebook documents via a web browser. The Jupyter Notebook App can be executed on a local desktop requiring no internet access (as described in this document) or can be installed on a remote server and accessed through the internet.

Pycharm or Visual Studio Code:

PyCharm is an integrated development environment (IDE) used for programming in Python. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems, and supports web development with Django. PyCharm is developed by the Czech company JetBrains.It is cross-platform, working on Microsoft Windows, macOS and Linux. PyCharm has a Professional Edition, released under a proprietary license and a Community Edition released under the Apache License. PyCharm Community Edition is less extensive than the Professional Edition.

Flask:

Flask is a web framework, it’s a Python module that lets you develop web applications easily. It’s has a small and easy-to-extend core: it’s a microframework that doesn’t include an ORM (Object Relational Manager) or such features. It does have many cool features like URL routing, template engine. It is a WSGI web app framework.

Pandas:

Pandas is a Python library used for working with data sets. It has functions for analysing , cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Anaconda:

Anaconda software helps you create an environment for many different versions of Python and package versions. Anaconda is also used to install, remove, and upgrade packages in your project environments. Furthermore, you may use Anaconda to deploy any required project with a few mouse clicks.

Tensorflow/Keras:

TensorFlow is an open source library created for Python by the Google Brain team. TensorFlow compiles many different algorithms and models together, enabling the user to implement deep neural networks for use in tasks like image recognition/classification and natural language processing. TensorFlow is a powerful framework that functions by implementing a series of processing nodes, each node representing a mathematical operation, with the entire series of nodes being called a "graph".

Numpy:

NumPy is a Python library used for working with arrays. It also has functions for working in the domain of linear algebra, Fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open-source project and you can use it freely.

Matplotlib:

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is easy to use Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

3.1. LANGUAGE SPECIFICATION

The python programming language is chosen for this project because it offers concise and readable code. Python is an interpreted high-level programming language, it offers multiple options for developing GUI (Graphical User Interface). Out of all the GUI methods, tkinter is most commonly used method. It is a standard Python interface to the Tk GUI toolkit shipped with Python. Python with tkinter outputs the fastest and easiest way to create the GUI applications. Creating a GUI using tkinter is an easy task. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented , imperative, functional and procedural, and has a large and comprehensive library. Python is a multi-paradigm programming language. Python's large standard library, commonly cited as one of its greatest strengths, provides tools suited too many tasks.Python is a programming language that sets itself apart from others by offering the adaptability, clarity, and dependable tools necessary to develop contemporary software. Python is best suited for machine learning since it is reliable and based on simplicity. Due to its independent platform and widespread use in the programming community, the Python programming languageis the most suitable for machine learning techniques.

CHAPTER 4

SYSTEM IMPLEMENTATION

4.1. SYSTEM ARCHITECTURE

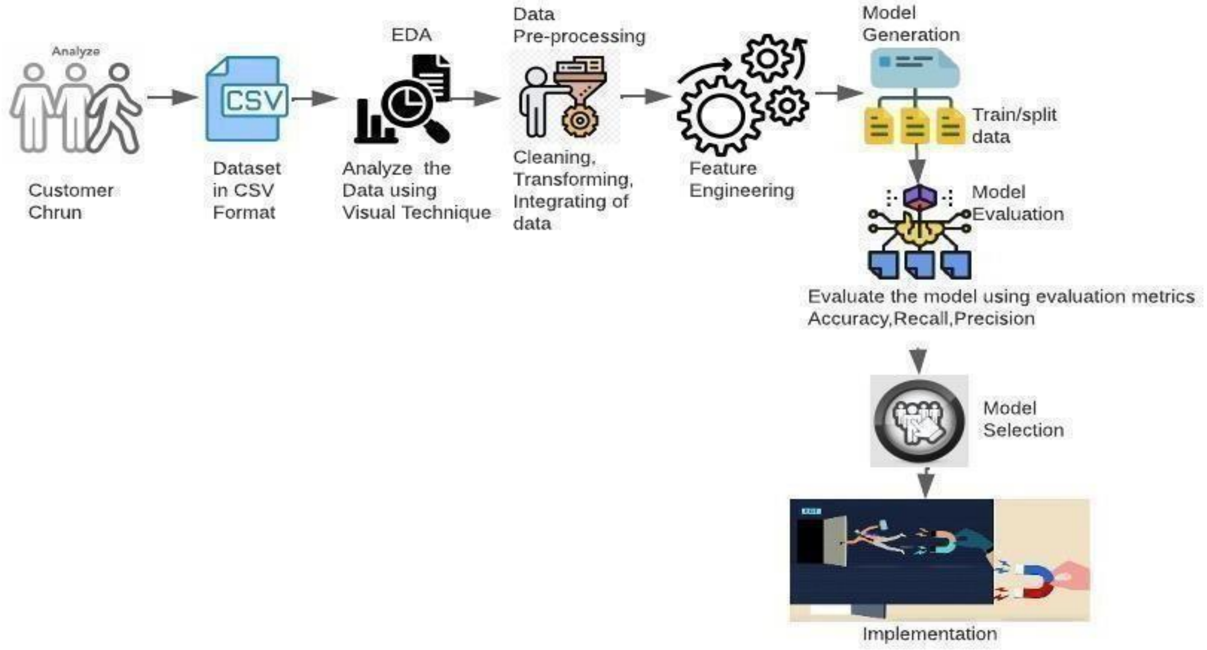


Fig: 4.1.1.a) System Architecture

a) Dataset Collection:

The process of gathering data depends on the type of project, if the project uses real-time data, then a system that uses different sensor data can be built. The data set can be collected from various sources such as a file, database, sensor, and many other such sources but the collected data cannot be used directly for performing the analysis process as there might be a lot of missing data, extremely large values, unorganized text data or noisy data. Therefore, to solve this problem Data Preparation is done.

b) Data pre-processing:

Data pre-processing is one of the most important steps in machine learning. It is the most important step that helps in building machine learning models more accurately. In machine learning, there is an 80/20 rule. Every data scientist should spend 80% time on data pre-processing and 20% time actually performing the analysis. Data pre-processing is a process of cleaning the raw data i.e. the data is collected in the real world and is converted to a clean data set. In other words, whenever the data is gathered from different sources it is collected in a raw format and this data isn't feasible for analysis. Therefore, certain steps are executed to convert the data into a small clean data set, this part of the process is called as data pre-processing.

c) Data Cleaning:

Data pre-processing is a process of cleaning the raw data into clean data, so that can be used to train the model. So, data pre-processing is needed to achieve good results from the applied model in machine learning and deep learning projects.

Most of the real-world data is messy, some of these types of data are:

1. Missing data: Missing data can be found when it is not continuously created or due to technical issues in the application (IOT system).

2. Noisy data: This type of data is also called outliners, this can occur due to human errors (humans manually gathering the data) or some technical problem of the device at the time of collection of data.

3. Inconsistent data: This type of data might be collected due to human errors (mistakes with the name or values) or duplication of data.

Data transformation is the process of converting, cleansing, and structuring data into a usable format that can be analyzed to support decision making processes, and to propel the growth of an organization. Data transformation is used when data needs to be converted to match that of the destination system

d) Exploratory Data Analysis:

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions. EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. EDA techniques continue to be a widely used method in the data discovery process today.

e) Feature Extraction:

Feature extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. It yields better results than applying machine learning directly to the raw data. Feature extraction can be accomplished manually or automatically.

f) Building and Training the model:

Building an ML Model requires splitting of data into two sets, such as a ‘training set’ and ‘testing set’ in the ratio of 80:20 or 70:30; A set of supervised (for labeled data) and unsupervised (for unlabeled data) algorithms are available to choose from depending on the nature of input data and business outcome to predict. You train the classifier using ‘training data set’, tune the parameters using ‘validation set’ and then test the performance of the classifier on an unseen ‘test data set’. An important point to note is that during training the classifier only the training and/or validation set is available. The test data set must not be used during the training of the classifier. The test set will only be available during testing of the classifier.

Training set: The training set is the material through which the computer learns how to process information. Machine learning uses algorithms to perform the training part. A set of data used for learning, that is to fit the parameters of the classifier.

Validation set: Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. A set of unseen data is used from the training data to tune the parameters of a classifier.

Test set: A set of unseen data used only to assess the performance of a fully-specified classifier.

g) Model Deployment:

The process of deploying a model is thought to be difficult for data scientists. This is because it is frequently not regarded as their primary duty and because model creation, training, and the organizational tech stack, including versioning, testing, and scaling, make deployment challenging, differ technologically and psychologically from one other. With the appropriate model deployment frameworks, tools, and procedures, these technological and organizational silos can be broken down.The models can be deployed in production environments to perform prediction either in batch inference mode or on-line inference mode. The batch inference can be achieved by scheduling as a job to run at a time interval and send the results via email to the intended users. The on-line inference can be achieved by exposing the model as a web service using frameworks such as python flask library or streamlit library to develop interactive web applications and invoke the model using its HTTP endpoint.

CHAPTER 5

5. MODULE DESCRIPTION

5.1. MODULES

- Dataset Preparation
- Exploratory Data Analysis
- Data Pre-processing/Cleaning
- Model Building
- Comparison of models
- Deploying the model

5.1.1. Dataset Preparation

The Dataset was collected from IBM sample datasets, The Tele communication Customer Churn dataset comprises details about a fictitious telco business that served 7043 clients in California in Q3 with home phone and Internet services. It shows which clients have cancelled, stayed, or joined their service. Each customer also has a Customer Lifetime Value (CLTV) index, a Satisfaction Score, a Churn Score, and several significant demographics.

The data set includes information about:

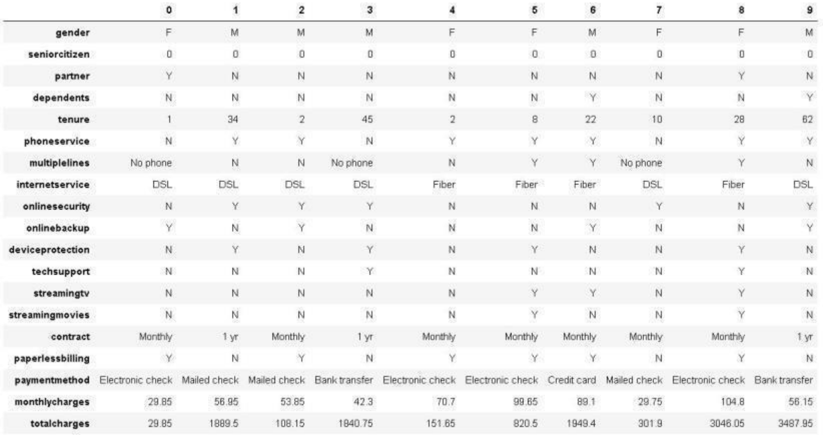
- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they’ve been a customer,

contract, payment method, paperless billing, monthly charges, and total charges

- Demographic info about customers – gender, age range, and if they have partners and dependents

5.1.2. Exploratory Data Analysis

Several actions are taken to explore the data after data gathering. Understanding the data structure, performing basic preprocessing, cleaning the data, identifying patterns and discrepancies in the data (such as skewness, outliers, and missing values), and developing and validating hypotheses are the objectives of this module. EDA's initial section evaluates the data frame's structure, columns it contains, and data kinds. The objectives of this step are to gain a general grasp of the data set, assess domain knowledge, and gather preliminary suggestions for research topics.



	0	1	2	3	4	5	6	7	8	9
gender	F	M	M	M	F	F	M	F	F	M
seniorcitizen	0	0	0	0	0	0	0	0	0	0
partner	Y	N	N	N	N	N	N	N	Y	N
dependents	N	N	N	N	N	N	Y	N	N	Y
tenure	1	34	2	45	2	8	22	10	28	62
phoneservice	N	Y	Y	N	Y	Y	Y	N	Y	Y
multiplelines	No phone	N	N	No phone	N	Y	Y	No phone	Y	N
internetservice	DSL	DSL	DSL	DSL	Fiber	Fiber	Fiber	DSL	Fiber	DSL
onlinesecurity	N	Y	Y	Y	N	N	N	Y	N	Y
onlinebackup	Y	N	Y	N	N	N	Y	N	N	Y
deviceprotection	N	Y	N	Y	N	Y	N	N	Y	N
techsupport	N	N	N	Y	N	N	N	N	Y	N
streamingtv	N	N	N	N	N	Y	Y	N	Y	N
streamingmovies	N	N	N	N	N	Y	N	N	Y	N
contract	Monthly	1 yr	Monthly	1 yr	Monthly	Monthly	Monthly	Monthly	Monthly	1 yr
paperlessbilling	Y	N	Y	N	Y	Y	Y	N	Y	N
paymentmethod	Electronic check	Mailed check	Mailed check	Bank transfer	Electronic check	Electronic check	Credit card	Mailed check	Electronic check	Bank transfer
monthlycharges	29.85	56.95	53.85	42.3	70.7	99.65	89.1	29.75	104.8	56.15
totalcharges	29.85	1889.5	108.15	1840.75	151.65	820.5	1949.4	301.9	3046.05	3487.95

Fig: 5.1.2.a) Features in the data set

Correlation, r, measures the linear association between two quantitative variables. Correlation measures the strength of a linear relationship only. (See the following Scatterplot for display where the correlation is 0 but the two variables are obviously related.)

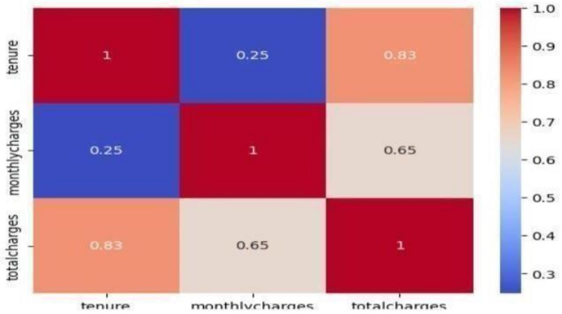


Fig:5.1.2.b) Correlation between numerical features

Since "totalcharges" is the sum of all monthly charges across a customer's tenure, as shown by the correlation matrix and regplots, "totalcharges" is strongly connected with "monthlycharges" and "tenure".

When looking for a correlation between two nominal variables, Cramer's V is a better choice than Pearson correlation. The Cramer's V measure is used in this situation.

The Churn plot reveals that the dataset is unbalanced (bottom right corner).

It is possible to display the frequency distribution of data in a table or graph. Frequency tables, histograms, and bar charts are a few popular ways to display frequency distributions. A frequency table is an easy way to show how frequently a certain value or feature occurs.

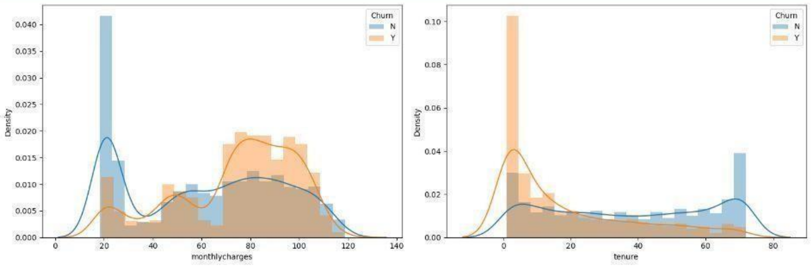


Fig:5.1.2.c) Frequency table

We can state that the majority of consumers who churn have monthly fees between

70 and 110 and tenures between 0 and 20 months.

Short duration clients make up 50% of the churned customers and 50% of the churned revenue. This also implies that clients with short duration pay higher monthly fees than those with long tenure. Therefore, a significant portion of the revenue can be kept if the ML models are successful in detecting the low tenure consumers.

We can infer that the majority of the customers who are churned do not have monthly contracts, are not senior citizens, do not have dependents, do not subscribe to online security services or technical support, do not have Fiber internet service, do not have paperless billing, and do not have phone service subscriptions.

Categorical Value

For Gender, we can observe that males are more susceptible to churning and not churning. The data count is similar.

For the classification of Senior Citizen, we can see that non-senior citizens are more susceptible to churning and not churning. However, the dataset contains more data on non-senior citizens than seniors.

For the Partner classification, we can see that those without a partner are more susceptible to churning, whilst those who have a partner are more susceptible to not churning.

For the Dependents classification, here we can observe that customers who are not financially dependent are more susceptible to churning and not churning.

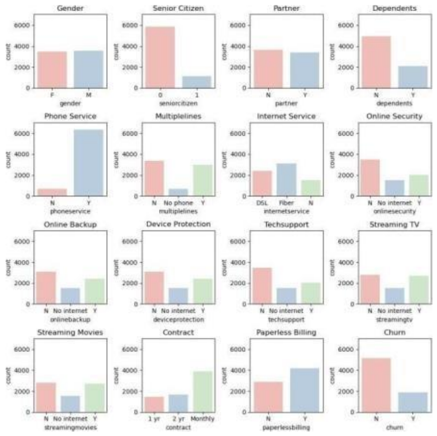


Fig:5.1.2.e) Churn among categorical values

For the Phone Service classification, we can ascertain that those with phone service are more susceptible to churning and not churning. For the Multiple Lines classification, we can see that those with no multiple lines are more susceptible to churning and not churning. For the Internet Service classification, we can understand that those with Fiber Optic service are more susceptible to churning, whilst those who use a DSL service are more susceptible to not churning. For the Online Security classification, we can observe that customers who have no online security are more susceptible to churning and not churning.

For the Online Backup classification, we can ascertain that those without Online Backup are more susceptible to churning, whilst those with it are not. For the Device Protection classification, we can see that those with no protection are more susceptible to churning and not churning. For the Tech Support classification, we can understand that those without such service are more susceptible to churning and not churning. For the Streaming TV classification, we can observe that customers who stream and do not stream tv are security are not susceptible to churning, with those who do not stream more susceptible to churning.

5.1.3. Data Pre-processing

Data pre-processing is one of the most important stages of information discovery activities. It necessitates a number of steps, including data transformation and reduction. When raw data is converted into low-quality data, the efficiency and accuracy of learning algorithms are compromised. Thus, by performing proper data preprocessing steps and selecting appropriate learning algorithms, the collected data can be correctly analysed. Missing values, non-numeric features, inconsistent feature scales, and other issues plague telecom datasets. As a result, it is critical to pre- process the data before implementing a learning model. Handling missing data and encoding categorical variables using label encoding are all part of the pre-processing phase. Furthermore, using the normalisation technique to scale the high variance values and removing the unnecessary features that do not affect the dependent variable. The EDA shows that both datasets are imbalanced, and we used the cost sensitive learning technique to address this issue.

Figure 3 describes that there is a Class Imbalance in the target of our dataset, where data entries for "No" is counted at 5174, whilst data entries for "Yes" is counted at only 1869.

Data Imbalance Interpretation:

Ratio of 20 - 40% = Mild Imbalance Ratio of 1 - 20% = Moderate Imbalance Ratio of Below 1% = Extreme Imbalance

In our case, the data entries for those with target outcome "Yes" (or Class 1) is only 26.53% of the dataset as a whole. Therefore, there is a mild imbalance in our data.

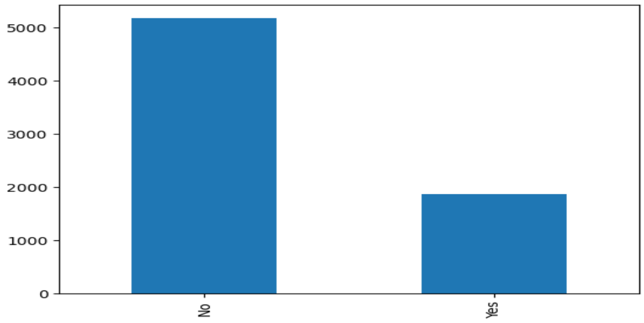


Fig:5.1.3.a) Class imbalance

1) Missing data:

Missing data has an impact on statistical analysis due to knowledge loss and irregularities in data patterns. Because the dataset contains missing values, it is critical to understand the percentage of missing values and their location. The total charges have 0.16% missing values, which is less than 30%. As a result, missing data in both data sets are handled using the simple imputer technique. Simple Imputer is a programme that is used to impute or replace missing continuous and categorical variables with various means, medians, most frequent values, and constant values. Because the missing values in the dataset are less than 30%, the best strategy for dealing with them is to fill the null with the mean of each specific feature.

2) Data Tranformation:

Data transformation techniques can significantly improve the overall performance of churn prediction. On the data set, three different data transformation methods were used: Normalization and label encoding.

Label Encoding (LE) is a fundamental technique for mapping categories as continuous integers into nominal attributes. The LE technique is more useful for categorical variables with only two distinct values, such as (yes or no) and (male or female). Gender, partner, dependents, phone service, paperless billing, and churn are six categorical variables with two unique values in the IBM Telco data set. These characteristics were encoded using the LE technique, with yes replaced by one and no replaced by zero, and male replaced by one and female replaced by zero.

Normalization: The dataset was normalised to improve the results of machine learning methods for customer churn prediction. MinMaxScaler is one of the most commonly used normalisation methods. By rescaling variables into the range (0,1), MinMaxScaler scales and translates each feature individually. When categorical features are already encoded, the MinMax Scaler is applied only to the high variance features, not the entire data set; otherwise, it causes over rescaling. There are many other techniques for rescaling values, such as standardisation and binarization, but after some testing, normalisation yields the best performance results.

3) Handling Imbalanced dataset:

One of the predictive modelling classification problems used in this study is imbalanced classification. An imbalance occurs when one or more classes have low proportions in the training results in comparison to the other classes. In our case, the churn class has much lower proportions than the non-churn class.

Customers' features, for example, have 80 instances of customers who have not churned and 20 instances of customers who have churned, and our training dataset only contains these instances. This is an example of an unbalanced classification problem. The CCP faces new challenges as a result of class disparities. As a result, the IBM Telco dataset was discovered to be unbalanced, as the percentage of the second (minority) class representing churn customers in IBM Telco' dataset is approximately 26.5%. To tackle this problem, cost sensitive learning technique was used. When training a machine learning model, cost-sensitive learning takes into account the costs of prediction errors (as well as potentially other costs). It is a branch of mathematics that is closely related to imbalanced learning and is concerned with classification on datasets with skewed class distributions. As a result, many of the concepts and techniques developed and applied for cost-sensitive learning can be adopted for imbalanced classification problems.

5.1.4. Model Building

Since this classification problem lacks linearity, the tress-based Ensemble technique has been applied for modelling. A class weightage of 1:3 has also been applied to address the 1:3 class imbalance, which means false negatives now cost three times as much as false positives. With the understanding that there has been no data leakage, the model was constructed using 80% of the data and verified using the remaining 20%. The genetic algorithm was used to tune the random forest model's many hyperparameters since it proved to be the most effective tuning technique while simultaneously preventing overfitting.

A Random Forest model is a special classifier composed of a collection of simple classifiers (Decision Trees), each of which votes for the most popular class in the input. Random Forest does not require the use of cross-validation techniques or verification on a separate set of variables in order to provide an impartial evaluation of the error, as this is inherent in the method. In fact, each tree is constructed using a different bootstrap sample of the original data, while the cases that were not chosen for tree construction are used to estimate the model's errors. This model is simple to use because it requires only two parameters to be entered (the number of variables in the subset of random variables used in each node and the number of trees in the forest) and is not overly sensitive to their values. The performance of a classification model is then assessed by creating a confusion matrix, which is a table. A calculation of accuracy was made using the confusion matrix.

5.1.5. Model Optimization

Hyperparameter tuning has been used to optimize the learning model. Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm

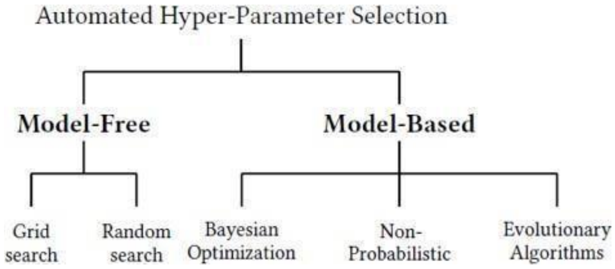


Fig:5.1.5.a) Hyper parameter tuning

The selection of machine learning hyperparameters can have a significant impact on an algorithm's performance, making it a time-consuming yet essential operation. Important machine learning pipeline phases like feature engineering and result interpretation are delayed by manual tuning. Although hands-off, grid and random searches take a long time to complete because they waste time assessing unproductive parts of the search field. The process of adjusting hyperparameters is increasingly done automatically, with no manual work required beyond the initial setup. The goal is to locate the best hyperparameters quickly using an educated search. While using Automatic Hyperparameter Tuning, the model hyperparameters to use are found using approaches such as:

Bayesian Optimization, Gradient Descent and Evolutionary Algorithms.

Random Search: A parameter space with a certain distribution can have a specific number of candidates sampled from it. Following a description of these tools, we go over best practises that apply to both strategies. The grid of hyperparameters used in Random Search was built, and we trained/tested our model using only a haphazard combination of these hyperparameters. The training set has undergone cross-validation. To ensure that our model is not overfitting our data while utilising Cross- Validation, the training set was split into N additional partitions. K-Fold Validation is one of the Cross-Validation techniques that is most frequently employed. In K- Fold, our training set has been divided into N partitions and then iteratively trained the model using N-1 partitions and tested it with the left-over partition (at each iteration we change the left-over partition). Once the model has been trained N times, we average the training outcomes from each iteration to get our overall training performance results.

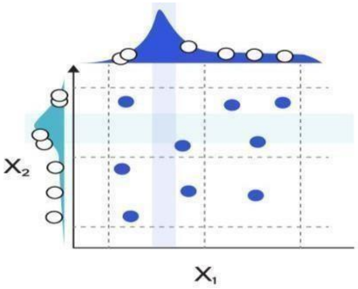


Fig:5.1.5.b) Random search

GridSearchCV: For the specified values, GridSearchCV carefully takes into account all possible parameter combinations. A grid of parameter values defined by the param grid parameter is generated exhaustively by the grid search offered by GridSearchCV. Grid Search trains and tests the model on all conceivable combinations using a grid of hyperparameters. We can now look at the parameters that worked best with Random Search to determine the ones to utilise in Grid Search and create a grid based on those parameters to see if a better combination can be found.

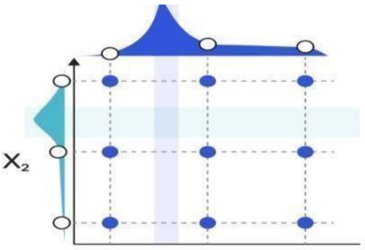


Fig:5.1.5.c) Grid search

Bayesian Optimization

While the ultimate goal of Bayesian optimization, a model-based method for finding the minimum of a function, is to identify the input value to a function that can yield the lowest feasible output value, it outperforms random search in terms of performance while requiring fewer iterations. Hence, Bayesian Optimization may result in faster optimization and improved testing performance. By selecting the input values while taking into account the results of previous iterations, Bayesian Optimization can lower the number of search iterations. By doing this, we may focus our search right away on values that are near to the results we want.

Genetic Algorithm

A genetic algorithm (GA) is a metaheuristic inspired by natural selection that belongs to the larger class of evolutionary algorithms in computer science and operations research (EA). Genetic algorithms, which rely on biologically inspired operators such as mutation, crossover, and selection, are commonly used to generate high-quality solutions to optimization and search problems. Natural selection mechanisms are attempted to be applied to Machine Learning contexts by Genetic Algorithms. Because they are inspired by the Darwinian process of Natural Selection, they are also known as Evolutionary Algorithms. Assume we build a population of N Machine Learning models with predefined Hyperparameters. The TPOT Auto Machine Learning library was used to implement Genetic Algorithms in Python. TPOT is based on the scikit-learn library and can perform both regression and classification tasks.

5.1.5. Evaluation of Model

The performance of models, which divide the original data into sections, is evaluated via cross-validation. The test set is used to assess the performance of models, whereas the training set is used to train the model. Since the out-of-sample data are also evaluated, cross-validation is helpful in producing a stable and reliable model. The average of all exams makes up the overall performance.

To rate the performance of models, accuracy, precision, recall, F1-score, and AUC are taken into consideration. The percentage of samples that the model properly predicts for all samples is known as accuracy. Precision is the percentage of samples that are actually positive compared to those that are expected to be positive.

Specificity measures how many negative examples in the real sample are accurately predicted, while recall measures how many positive examples in the actual sample are correctly predicted. The F1-score is produced by expanding on the basis of these four metrics. The outcomes of Precision and Recall are combined. The equations are written as, Where, TP (True Positive) signifies that the sample is both really and correctly expected to be positive.

False Negative (FN) signifies that the sample is actually positive even though it was expected to be negative.

FP (False Positive) denotes a sample that was expected to be positive but is really negative.

TN (True Negative) denotes that the sample is both genuinely and predictably negative.

Confusion Matrix was made to represent a tabular visualization of the model predictions versus the ground-truth labels.

5.1.7. Deployment of Model

Deploying a machine learning model simply refers to integrating the model into an already-existing production environment that can accept an input and produce a useful output for business decision-making. One component of the data is the model that is built/trained using different techniques on a large dataset. However, implementing machine learning in the actual world requires two steps, the first of which is employing these models in various applications.

The model needs to be deployed online so that users from the outside world can use it in order to forecast the new data. With the help of streamlit, a web application was built utilizing the machine learning model. We will build a simple HTML webpage to accept the measurements as input and classify the variety based on the classification model. Although there are competing frameworks on the market, such as FastAPI, streamlit is one among the most popular and well-respected framework among machine learning experts for deploying models.Using straightforward Python scripts, Streamlit enables you to develop apps for your machine learning project. Hot reloading is also supported, allowing your app to update in real time as you change and save your file. Using the Streamlit API, an app may be created in just a few lines of code (as we'll see below). Declaring a variable and adding a widget are the same thing. Writing a backend, specifying many routes, or managing HTTP requests are not necessary. It is simple to deploy and control.

CHAPTER 6

6.1. CONCLUSION

Customer retention is essential in a market that is competitive. This project used a genetic approach to implement hyperparameter adjustment in the Random Forest Classifier while taking the dataset into account. According to the telecom customer churn prediction model built using a random forest method with genetic algorithm tuning, churn can be anticipated when customers are dissatisfied with the service that is being delivered. The model achieved 80% accuracy. Also, as our findings indicate that customers with short contracts are more likely to churn, the telecom sector should make a compelling offer to entice customers to sign up for one or two years. Since our findings indicate that fibre optic service subscribers are more likely to churn than other consumers, the telecom sector ought to offer a good discount to these customers. The telecom sector should steer clients towards using mailed checks, bank transfers, or credit cards instead than electronic check methods for paying financial receivables. The findings demonstrate that customers who use electronic checks as their primary form of payment are more likely to leave. Since customers who are not provided with technical help are more likely to leave a telecom company, the industry should pay greater attention to this area.

6.2. FUTURE ENHANCEMENT

Using hybrid classification algorithms, future research in this area will concentrate on the connections between client lifetime value and churnprediction that already exist. By selecting the appropriate datasetvariables, it is important to consider the retention policies. Due to the passive and dynamic nature of the sector, data mining will play a bigger role in the future of the telecommunications business.

APPENDIX 1

Depolyment

App.py

```
#Import libraries
import streamlit as st
import pandas as pd
import numpy as np
from PIL import Image
#load the model from disk
import joblib
model = joblib.load("modeldechm.sav")
#Import python scripts
from preprocessing import preprocess
def main():
#Setting Application title
st.title('Telco Customer Churn Prediction App')
#Setting Application description st.markdown(""":dart: This Streamlit app is made to
predict customer churn in a ficitional telecommunication use case.The application is
functional for both online prediction and batch data prediction.\n""")
st.markdown("<h3></h3>", unsafe_allow_html=True)
#Setting Application sidebar default image = Image.open('App.jpg') add_selectbox =
st.sidebar.selectbox("How would you like to predict?", ("Online", "Batch"))
st.sidebar.info('This app is created to predict Customer Churn')
st.sidebar.image(image)
if add_selectbox == "Online": st.info("Input data below")
#Based on our optimal features selection st.subheader("Demographic data")
seniorcitizen = st.selectbox('Senior Citizen:', ('Yes', 'No')) dependents =
st.selectbox('Dependent:', ('Yes', 'No'))
st.subheader("Payment data")
tenure = st.slider('Number of months the customer has stayed with the company',
min_value=0, max_value=72, value=0)
contract = st.selectbox('Contract', ('Month-to-month', 'One year', 'Two year'))
paperlessbilling = st.selectbox('Paperless Billing', ('Yes', 'No'))
PaymentMethod = st.selectbox('PaymentMethod',('Electronic check', 'Mailed check',
'Bank transfer (automatic)','Credit card (automatic)'))
monthlycharges = st.number_input('The amount charged to the customer monthly',
min_value=0, max_value=150, value=0)
totalcharges = st.number_input('The total amount charged to the
customer',min_value=0, max_value=10000, value=0)
st.subheader("Services signed up for")
mutliplelines = st.selectbox("Does the customer have multiple lines",('Yes','No','No
phone service'))
phoneservice = st.selectbox('Phone Service:', ('Yes', 'No')) internetservice =
st.selectbox("Does the customer have internet service",
('DSL', 'Fiber optic', 'No'))
onlinesecurity = st.selectbox("Does the customer have online security",('Yes','No','No
internet service'))
onlinebackup = st.selectbox("Does the customer have online backup",('Yes','No','No
internet service'))
techsupport = st.selectbox("Does the customer have technology support",
('Yes','No','No internet service'))
streamingtv = st.selectbox("Does the customer stream TV", ('Yes','No','No internet
service'))
streamingmovies = st.selectbox("Does the customer stream movies", ('Yes','No','No
internet service'))
data = {
'SeniorCitizen': seniorcitizen,
'Dependents': dependents,
'tenure':tenure,
'PhoneService': phoneservice,
'MultipleLines': mutliplelines,
'InternetService': internetservice,
'OnlineSecurity': onlinesecurity,
'OnlineBackup': onlinebackup,
'TechSupport': techsupport,
'StreamingTV': streamingtv,
'StreamingMovies': streamingmovies,
'Contract': contract,
'PaperlessBilling': paperlessbilling,
'PaymentMethod':PaymentMethod,
'MonthlyCharges': monthlycharges,
'TotalCharges': totalcharges
}
features_df = pd.DataFrame.from_dict([data])
st.markdown("<h3></h3>", unsafe_allow_html=True)
st.write('Overview of input is shown below')
st.markdown("<h3></h3>", unsafe_allow_html=True)
st.dataframe(features_df)
#Preprocess inputs
preprocess_df = preprocess(features_df, 'Online')
prediction = model.predict(preprocess_df)
```

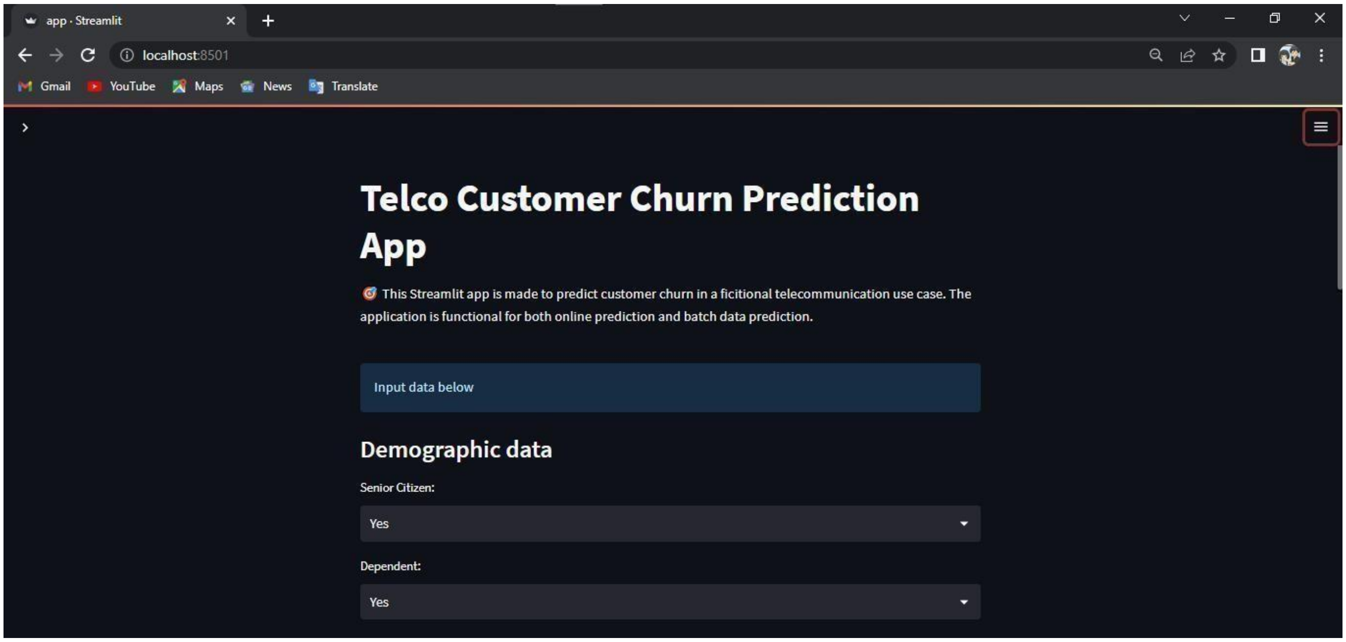
```
if st.button('Predict'):
 if prediction == 1:
st.warning('Yes, the customer will terminate the service.') else:
st.success('No, the customer is happy with Telco Services.')
else:
st.subheader("Dataset upload")
uploaded_file = st.file_uploader("Choose a file")
if uploaded_file is not None:
data = pd.read_csv(uploaded_file)
#Get overview of data
st.write(data.head())
st.markdown("<h3></h3>", unsafe_allow_html=True)
#Preprocess inputs
preprocess_df = preprocess(data, "Batch")
if st.button('Predict'):
#Get batch prediction
prediction = model.predict(preprocess_df)
prediction_df = pd.DataFrame(prediction, columns=["Predictions"]) prediction_df =
prediction_df.replace({1:'Yes, the customer will terminate the service.',
0:'No, the customer is happy with Telco Services.'})
st.markdown("<h3></h3>", unsafe_allow_html=True)
st.subheader('Prediction')
st.write(prediction_df)
```

Pre-processing

```
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
```

```
def preprocess(df, option): """
This function is to cover all the preprocessing steps on the churn dataframe. It
involves selecting important features, encoding categorical data, handling missing
values,feature scaling and splitting the data
"""
#Defining the map function
def binary_map(feature):
return feature.map({'Yes':1, 'No':0})
# Encode binary categorical features
binary_list = ['SeniorCitizen','Dependents', 'PhoneService', 'PaperlessBilling']
df[binary_list] = df[binary_list].apply(binary_map)
#Drop values based on operational options
if (option == "Online"):
columns = ['SeniorCitizen', 'Dependents', 'tenure', 'PhoneService','PaperlessBilling',
'MonthlyCharges', 'TotalCharges', 'MultipleLines_No_phone_service',
'MultipleLines_Yes', 'InternetService_Fiber_optic', 'InternetService_No',
'OnlineSecurity_No_internet_service', 'OnlineSecurity_Yes',
'OnlineBackup_No_internet_service', 'TechSupport_No_internet_service',
'TechSupport_Yes', 'StreamingTV_No_internet_service', 'StreamingTV_Yes',
'StreamingMovies_No_internet_service', 'StreamingMovies_Yes',
'Contract_One_year', 'Contract_Two_year', 'PaymentMethod_Electronic_check']
#Encoding the other categorical categoric features with more than two categories
df = pd.get_dummies(df).reindex(columns=columns, fill_value=0) elif (option ==
"Batch"):
pass
df =
df[['SeniorCitizen','Dependents','tenure','PhoneService','MultipleLines','InternetServic
e','OnlineSecurity',
'OnlineBackup','TechSupport','StreamingTV','StreamingMovies','Contract'
,'PaperlessBilling','PaymentMethod','MonthlyCharges','TotalCharges']]
columns = ['SeniorCitizen', 'Dependents', 'tenure', 'PhoneService', 'PaperlessBilling',
'MonthlyCharges', 'TotalCharges','MultipleLines_No_phone_service',
'MultipleLines_Yes', 'InternetService_Fiber_optic', 'InternetService_No',
'OnlineSecurity_No_internet_service', 'OnlineSecurity_Yes',
'OnlineBackup_No_internet_service', 'TechSupport_No_internet_service',
'TechSupport_Yes', 'StreamingTV_No_internet_service', 'StreamingTV_Yes',
'StreamingMovies_No_internet_service', 'StreamingMovies_Yes',
'Contract_One_year', 'Contract_Two_year', 'PaymentMethod_Electronic_check']
#Encoding the other categorical categoric features with more than two categories
df = pd.get_dummies(df).reindex(columns=columns, fill_value=0) else:
print("Incorrect operational options")
#feature scaling sc = MinMaxScaler()
df['tenure'] = sc.fit_transform(df[['tenure']]) df['MonthlyCharges'] =
sc.fit_transform(df[['MonthlyCharges']]) df['TotalCharges'] =
sc.fit_transform(df[['TotalCharges']]) return df
```

APPENDIX 2



Telco Customer Churn Prediction App

🎯 This Streamlit app is made to predict customer churn in a ficitional telecommunication use case. The application is functional for both online prediction and batch data prediction.

Input data below

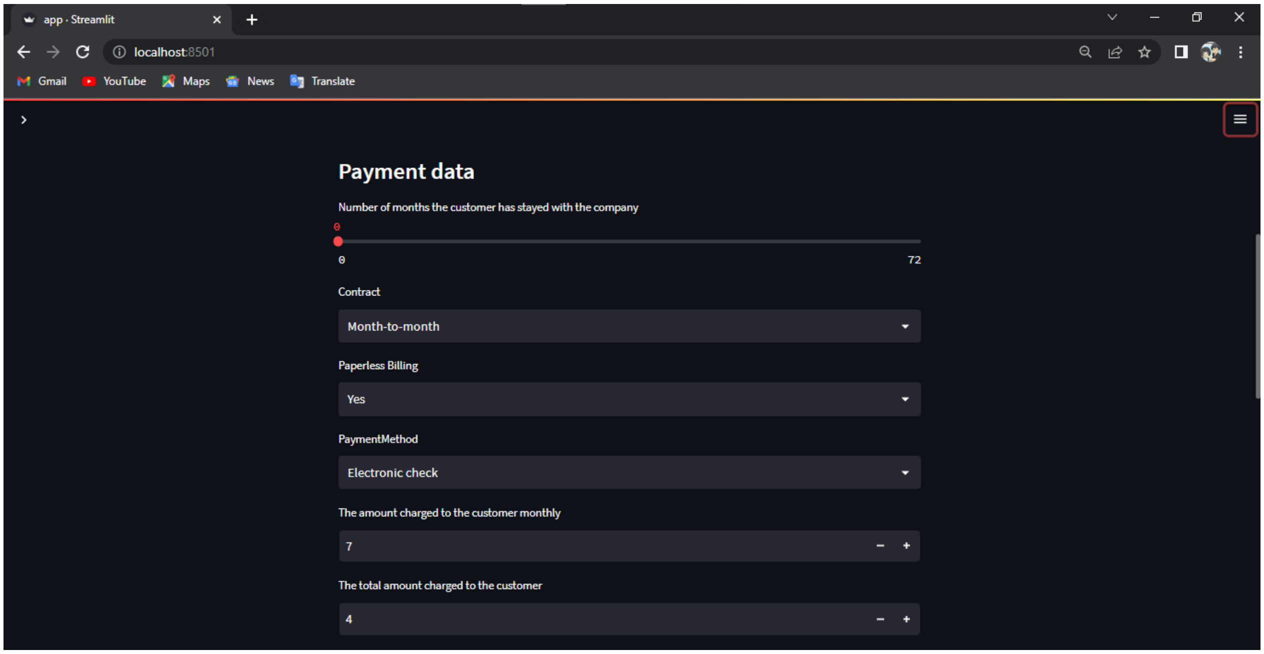
Demographic data

Senior Citizen:

Yes

Dependent:

Yes



Payment data

Number of months the customer has stayed with the company

0

0 72

Contract

Month-to-month

Paperless Billing

Yes

PaymentMethod

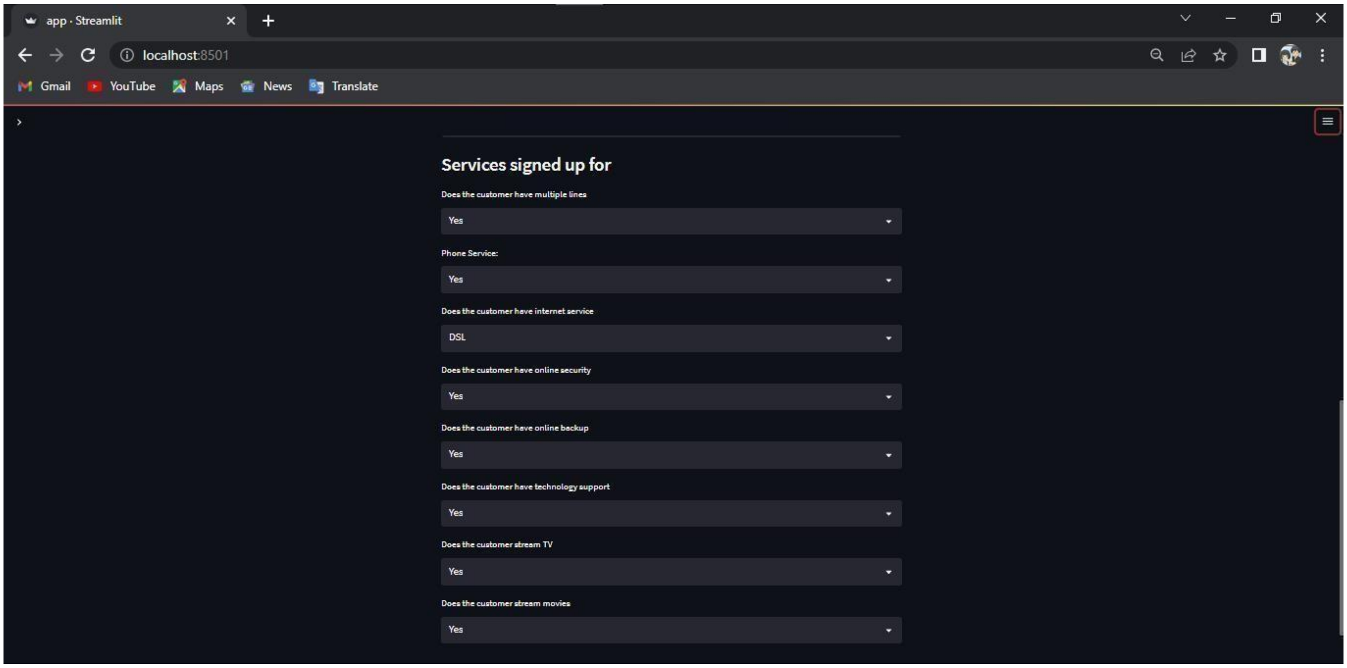
Electronic check

The amount charged to the customer monthly

7 − +

The total amount charged to the customer

4 − +



Services signed up for

Does the customer have multiple lines

Yes

Phone Service:

Yes

Does the customer have internet service

DSL

Does the customer have online security

Yes

Does the customer have online backup

Yes

Does the customer have technology support

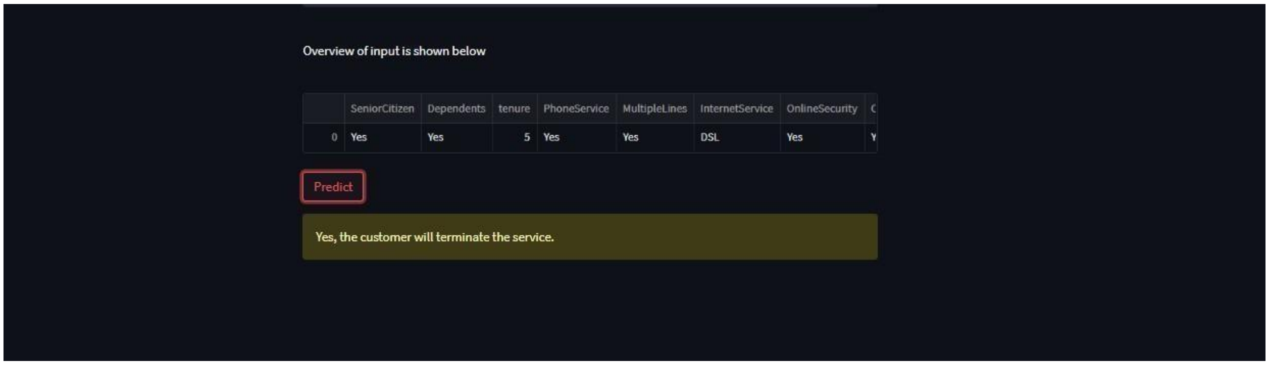
Yes

Does the customer stream TV

Yes

Does the customer stream movies

Yes



Overview of input is shown below

	SeniorCitizen	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	C
0	Yes	Yes	5	Yes	Yes	DSL	Yes	Y

Yes, the customer will terminate the service.

Reference

[1] Shabankareh, M.J., Shabankareh, M.A., Nazarian, A., Ranjbaran, A. and Seyyedamiri, N., 2022. A Stacking-Based Data Mining Solution to Customer Churn Prediction. Journal of Relationship Marketing, 21(2), pp.124-147.

[2] Sudharsan, R. and Ganesh, E.N., 2022. A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy. Connection Science, 34(1), pp.1855-1876.

[3] Samah Wael Fujo, Suresh Subramanian and Moaiad Ahmad Khder , Information Sciences Letters- An International Journal,2022

[4] Tianyuan Zhang ,Sérgio Moro and Ricardo F. Ramos, Big Data Analytics, Privacy and Visualization,2022

[5] Yajun Liu, Jingjing Fan, Jianfang Zhang, Xinxin Yin & Zehua Song , Journal of Intelligent Information Systems ,2022

[6] FE Usman-Hamza, AO Balogun, LF Capretz - Applied Sciences, 2022

[7] Syed Fakhar Bilal, Abdulwahab Ali Almazroi, Saba Bashir, Farhan Hassan Khan, Abdulaleem Ali Almazroi, PeerJ Computer Science, 2022,

[8] Shivani Vaidya and Rajesh kumar Nigam, International Journal of Electronics Communication and Computer Engineering , Volume 13, 2022

[9] Mohammad TabrezQuasim, Adel Sulaiman, Asadullah Shaikh and Mohammed Younus, Sustainable Computing: Informatics and Systems Volume 35, 2022

[10] P Jeyaprakaash , Sashirekha K,JOURNAL OF PHARMACEUTICAL NEGATIVE RESULTS, Volume 13,2022

[11] Sulaiman Olaniyi Abdulsalam, Micheal Olaolu Arowolo, Yakub Kayode Saheed, Jesutofunmi Onaope Afolayan, Indonesian Journal of Electrical Engineering and infomatics, 2022

[12] Maryam Sadeghi, Mohammad Naderi Dehkordi, Behrang Barekatain & Naser Khani , The Journal of Supercomputing ,2022

[13] Teoh Jay Shen, Abdul Samad Bin Shibghatullah, Journal of Advances in Artificial Life Robotics Vol. 3(2), 2022

[14] Zeynep Hilal KİLİMCİ, International Journal of Computational and Experimental Science and Engineering, Volume 8, 2022

[15] P. Ramesh, J. Jeba Emilyn & V. Vijayakumar , Wireless Personal Communications volume 124, pages1695–1709, 2022