Executive Summary (287 words)

Customer Churn indicates the process of predicting customers who will suspend a company service they receive, especially in the telecom sector, and taking up the necessary actions to prevent this cancellation. In this research, a small version of the original cell2cell customer churn dataset is used. The dataset contains all customer information and is used to train, test, and evaluate the system. We have applied different machine learning and deep learning techniques to model the customer data and developed a prediction model which helps the telecom operators to predict customers who are likely to be subject to churn. The model's performance is evaluated using the accuracy score, precision, recall, f1-score. Four machine learning and deep learning algorithms, i.e, Artificial Neural Networks (ANN), Decision Trees, Naïve Bayes, and Support Vector Machines (SVM), were applied to develop a customer churn prediction model. The machine learning and deep learning algorithms are evaluated while implementing both scenarios with and without feature selection. The efficiency of these models can be further improved by implementing the hyperparameter tuning i.e., Grid Search CV and Randomized Search CV. The result was then analyzed and compared to find an appropriate model in terms of accuracy score, precision, recall, f1-score. From the results, it can be inferred that the Support Vector Machine model performed better specifically in testing data before the grid search. The information gathered after working on this project will help to further increase the industry's awareness about the new trends in the field of machine learning that can be used in churn prediction. In the future, we can also create a web or mobile application to predict the churn rate and implement state-of-the-art machine learning models to improve the previous predictions further.

Task 1:

Introduction (557 words)

In the past few years, the telecom industry has become one of the world's fastest-growing industries, impacting around 90% of the world population and giving significant importance to the customer [1]. Due to the rapid increase in this industry, nowadays customers have multiple options available, from which they can choose to receive the services. The customers' choice to select a service provider depends upon various factors such as cost and customer service to name a few [2]. To meet the customers' needs, telecom companies are constantly working on developing policies and improving their services to retain and attract new customers to dominate the growing market. However, there exists the need to predict the possible churners and the reasons behind the decision for the service-providing telecom companies to take the

necessary measures and develop effective retention strategies. Predicting customer churn in the telecom industry is an area of interest for many researchers. Extensive researched have been carried out on various techniques to predict telecom customer churn [3]. Since customer churn directly affects the revenues and the growth of the companies, especially in the telecom sector, tremendous focus has been placed on researches in predictive factors for increased churning and coping mechanisms for companies facing this challenge [4].

The highly competitive nature of telecom companies encourages customers to switch between the service providers; moreover, poorly managed customer services, convoluted emails, and high plan prices of the telecom companies add to the frustration, hence the need for efficient customer churn prediction systems [4]. In the past few years, the need to automate identifying the customer churn has increased, as the telecom industry battles with the annual loss of approximately 25% of its customers resulting in a great revenue loss. It is important to note that the cost of adding a new customer to the system is between 5 to 10 times higher than that of retaining the old customers; thus it is crucial to maintain the existing customers and prevent customer churn [5].

AI has successfully tackled the customer churn problem by analyzing and modeling the large amounts of customers' data to find the hidden insights and the contributing features, thus predicting/forecasting the customer churn [6]. AI can analyze a wide variety of data, including the new data sources, along with relatively complex individual history to determine the risk involved. AI can also be used for recommendation purposes to indicate the best solution that will help retain the customer and moreover, to identify the possible reasons why a customer leaves the system [6].

Multiple types of research have shown that machine learning and deep learning algorithms have proven to be efficient in predicting customer churn [6]. In this research, four different machine learning and deep learning algorithms i.e Artificial Neural Networks, Decision Tree, Naïve Bayes, and Support Vector Machines are used to develop a customer churn prediction model while implementing both learning models with and without feature selection. The efficiency of these models was further explored by passing them through Grid Search CV and Randomized Search CV, and a comparative analysis was performed between the above-mentioned four machine learning and deep learning algorithms. The results have shown the approach's effectiveness and a satisfactory level of performance. However, we identified that the Support Vector Machine model performed better in the particular case of testing data before the grid search.

Task 2:

Literature Review (853 words)

Customer churn is a primary concern for large companies, as it affects their revenues, especially in the telecom sector. Hence, it is necessary to find the main reasons that contribute to customer churn to take the required actions to overcome the problem. A great deal of research has been done to address the customer churn prediction problem, with a significant number of publications produced over the past few years.

A neural network algorithm-based model was proposed, which aims to solve the customer churn problem in a large telecom company that has about six million customers [7]. The accuracy obtained on the test dataset was about 80%.

To solve the customer churn problem a novel approach was proposed which uses genetic programming and AdaBoost classifier to select an appropriate number of features. The efficiency of the proposed method was evaluated using the Orange Telecom dataset [8]. An accuracy of 63% was obtained using the proposed approach.

A theoretical approach, i.e., rough set theory, uses a classification model to predict customer churn. After performing the comparative analysis of the theoretical approach with the traditional machine learning algorithms, it can be inferred that the rough set theory-based classification model performs better than the traditional machine learning algorithms i.e., Logistic Regression, Decision Tree, and KNN Classifier [9].

The issue of unbalanced datasets was addressed, and performance analysis was applied using random sampling, under-sampling, gradient boosting model, weighted random forest [10]. The accuracy score and lift metrics are used to evaluate the model's performance. From the results, it can be demonstrated that the under-sampling approach gives better results than other approaches.

The data mining-based approach relies upon call details of the customer's dataset, consisting of twenty-one different independent features. The independent features possess the information regarding the incoming and outgoing messages and the voicemail details of each customer [11]. As the dataset contains twenty-one dimensions, the principal component analysis is used to reduce the dimensions. Multiple machine learning algorithms were implemented i.e., Naïve Bayes classifier, Artificial Neural Network, and Support Vector Machine. The accuracy score is used as an evaluation metric to check the efficiency of the proposed approach and the performances of the machine learning algorithms. A comparative analysis was done while implementing the

state-of-the-art data mining techniques that are used for the data preparation and data analysis, at the same time some interesting observations were made with regards to the performance impact depending upon the choice of the algorithms [12].

A general review of the various data mining techniques for the customer churn prediction was carried out, taking into account the performance analysis of each approach on the available customer churn dataset [13]. The effectiveness of the data mining algorithms was analyzed through Logistic Regression, Decision Trees, and other algorithms. The performance metric for the evaluation was the accuracy score; it was concluded that the model performance is heavily dependent upon the choice of the data mining algorithm chosen for a particular problem; however, the future research direction was not provided in this research [14].

A big data platform was also used to address the customer churn prediction problem [15]. The main goal of the research was to show that the big data platform improves the overall process of predicting the customer churn using the volume along with the velocity of the data. A big data platform needed at a global telecom company to deal with operation and business support departments [15]. The selected machine learning model was Random Forest, and the performance of the model was analyzed while in the function of accuracy score as the evaluation metrics.

A detailed review of the state-of-the-art machine learning models, available datasets, prediction methods, and performance metrics was done [16, 17]. The main focus of this research was to list down the existing machine learning models, datasets, and evaluation metrics. From the results, it was inferred that the boosted versions of the classifiers perform better when compared to the traditional machine learningclassifiers. An accuracy of 85% was achieved when using SVM-Poly with the AdaBoost classifier [16, 17]. Moreover, a review on the social network classifiers using the call records dataset was done, and concluded that the network-only link-based classifier outperforms the other classifiers [18]. To create an efficient prediction model, traditional machine learning algorithms i.e., Decision Tree, and an Artificial Neural Network were also used. The effectiveness of both algorithms was evaluated with regards to the accuracy score [19]. A novel clustering algorithm was proposed, i.e., SDS clustering method for predicting customer churn [20]. In this clustering-based approach, different clusters were formed from the dataset, which was further used for the efficient prediction of churn.

From the literature review, it can be analyzed that multiple machine learning algorithms for churn prediction in the telecom sector have been considered, i.e., SVM, Artificial Neural Networks, Decision Tree, Naïve Bayes, Logistic Regression, Random Forest. A

comparative and performance analysis of these approaches have proven that SVM outperformed the other traditional machine learning algorithms, as SVM possesses the ability to deal with non-linearities. Moreover, the artificial neural network performed better than other conventional machine learning algorithms.

Task 3:

Research Design (514 words)

This research used a smaller version of the original cell2cell customer churn dataset to train and evaluate the machine learning and deep learning models. The dataset consists of 57 features containing information about the monthly revenue, monthly minutes, roaming calls, dropped calls, blocked calls, etc.

The entire research can be divided into four main parts. In the first part, data preprocessing and data wrangling were implemented. In the second part, data normalization was done and exploratory data analysis was performed. Multiple machine learning and deep learning models were designed and built on the full data. In the third part, a Genetic Algorithm is applied for feature selection. The machine learning and deep learning models are evaluated while undertaking both scenarios with and without feature selection in the fourth part. The efficiency of these models are further improved by applying Grid Search CV and Randomized Search CV. The proposed churn prediction framework is illustrated in Figure 1.

The specific process is as follows:

- In the data preprocessing stage, one-hot encoding and data labeling was performed; along with this, data types of the numerical columns were changed from object to float, and the dataset was cleaned by removing the missing values.
- In the second stage, data normalization was conducted. The exploratory data analysis was performed by importing Seaborn and Matplotlib libraries and creating bar plots, histograms, violin plots, and density plots. In the next step, train and test split applied using the full data, machine learning algorithms and deep learning algorithms i.e., Decision Tree Classifier, Random Forest Classifier, Naïve Bayes Classifier, Support Vector Machine Classifier, and Artificial Neural Network were designed and built, along with this hyperparameter tuning of each model was implemented using Grid Search CV and Randomized Search CV to

find the best parameters and to increase the accuracy score and the efficiency of the model.

- The optimization technique implemented in this research is based on the Genetic Algorithm, which finds a subset of the relevant features, as the implemented optimization technique performs much better than the available feature selection techniques; as it can handle the datasets with a large number of features as our dataset consists of 57 different features. The genetic algorithm is implemented using Genetic Selection CV. In the Genetic Algorithm, a population is being generated while using subsets of the possible features and using this population. In the next step, each of the subsets is being evaluated using different predictive models, the algorithm runs for a certain number of iterations, after this, the optimal member of the population is thus being selected as the features.

- To select the subset of relevant features, in the next step, the performance analysis of each of the machine learning and deep learning algorithms is done while implementing both the scenarios with and without feature selection. Through hyperparameter tuning, the best parameters for each of the models were selected. The evaluation metrics used for the evaluation of the accuracy score, precision, recall, f1-score. We have implemented four different machine learning and deep learning algorithms, and detailed result analysis is provided in the next section.

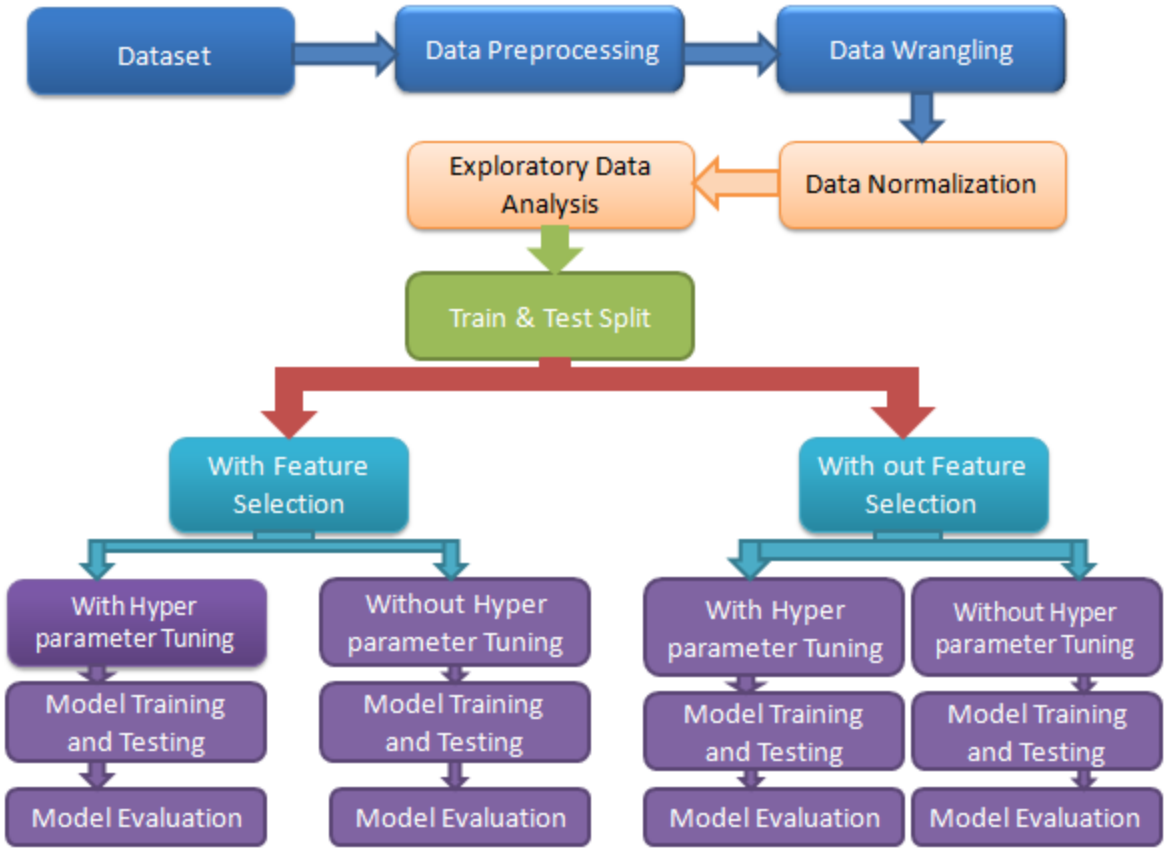

Figure 1. Proposed Churn Prediction Framework

Task 4:

Experimental Results and Analysis (1064 words)

In this research, machine learning and deep learning techniques were applied to explore and model the customer data, which can help to reduce the churn rate; it can also help to forecast customer behavior and identify trends in customer churn in the telecom companies. The customer churn dataset for this research consists of 6380 attributes and 57 different features. Four machine learning and deep learning algorithms were used to classify the patterns; each of the implemented algorithms was evaluated using the performance accuracy, f1-score, and recall score.

The dependent features representing the potential churner are binary, where 1 represents the “YES” and 0 represents the “NO”. The 56 other independent variables

contain the object, float, and integer data types. The object data types are converted into either integer or float after doing one-hot encoding or converting the object data type column to float or integer.

After loading the dataset and removing the missing values, performing the exploratory data analysis, and implementing the data normalization, the purpose of the data normalization is to change the values of the numerical features to a common scale between 0 to 1 without causing any distortion in the range of the values. Each machine learning algorithm is being evaluated in both scenarios,, with and without feature selection. We have used 75% of the data for the training purpose and 25% of the data to evaluate the model. To further increase the model's efficiency, hyperparameter tuning was implemented using the Randomized Search CV and Grid Search CV to find the best parameters for each models used for the evaluation. The optimization technique used in this research is the Genetic Algorithm, as it finds a subset of the relevant features; in addition, it performs much better than the available feature selection techniques. Our main aim was to evaluate the existing machine learning and deep learning models for churn prediction. The Support Vector Machine model performs better in this particular case than the other models on the testing data before the Grid Search. A detailed analysis of each machine learning and deep learning models used in this research is presented in the next section. After performing the hyperparameter tuning of the following parameters i.e., max depth of the trees, the minimum sample size of the leaf, and the criterion in the decision tree classifier using the Grid Search CV and choosing fourfold cross-validation, the accuracy score obtained was 71.61%, which was previously 61.4% before the hyperparameter tuning, so hyperparameter tuning helps us to find those optimal parameters which are giving us the best accuracy. The confusion matrix and ROC curve obtained after hyperparameter tuning are shown in Figures 2 and 3. The confusion matrix compares the actual target values with those predicted by the machine learning model. After applying feature selection to find the relevant features using genetic algorithm as an optimization technique, the accuracy score obtained was 71.55% which means that selecting the relevant features using genetic algorithm does not have any impact on the accuracy; however, hyperparameters had a very significant impact over the accuracy of the decision tree classifier model as it can be seen from the results that the accuracy has increased from 61.4% to 71.5%.

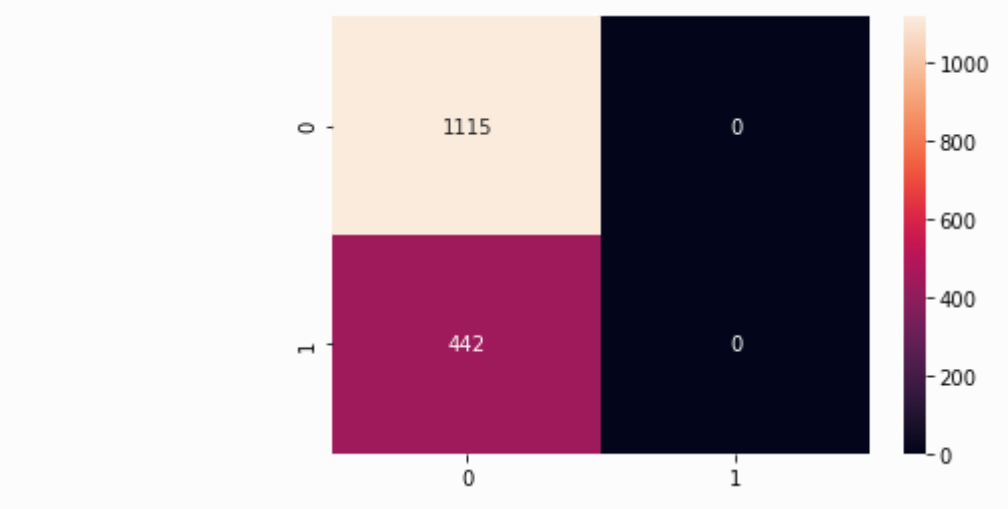

Figure 2. Confusion Matrix for Decision Tree Classifier

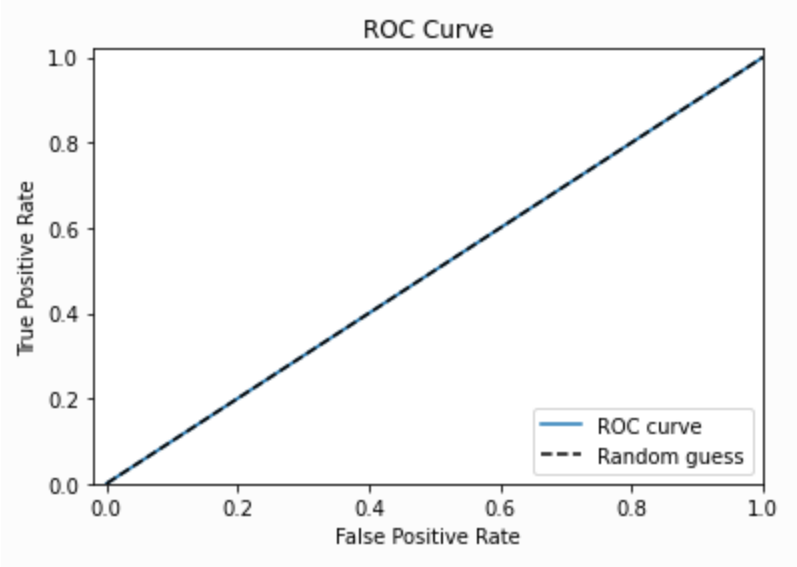

Figure 3. ROC Curve for Decision Tree Classifier

A Naïve Bayes classifier was applied to create a customer churn prediction model; in the first step, the Gaussian naïve Bayes model was trained on the training dataset; we have used 75% data for the training purpose and 25% data for the testing purpose. After training the model, we evaluate the effectiveness of our model using the test dataset. After evaluating the model on the test dataset, the accuracy obtained was 38.02%. In the next step, we perform the hyperparameter tuning using Randomized Search CV by passing a range of user-defined values to the distribution's variance which is referred to by the variable "var_smooting" and using 5 fold cross-validation and

20 number of iterations, after choosing the best parameters obtained after hyperparameter tuning and passing them to the machine learning model the accuracy score obtained was 71.61%. The confusion matrix and ROC curve obtained after hyperparameter tuning is shown in Figure 4 and 5, respectively. In the next step, after applying feature selection to select the relevant features and removing the irrelevant features, the accuracy score obtained was 71.5%.

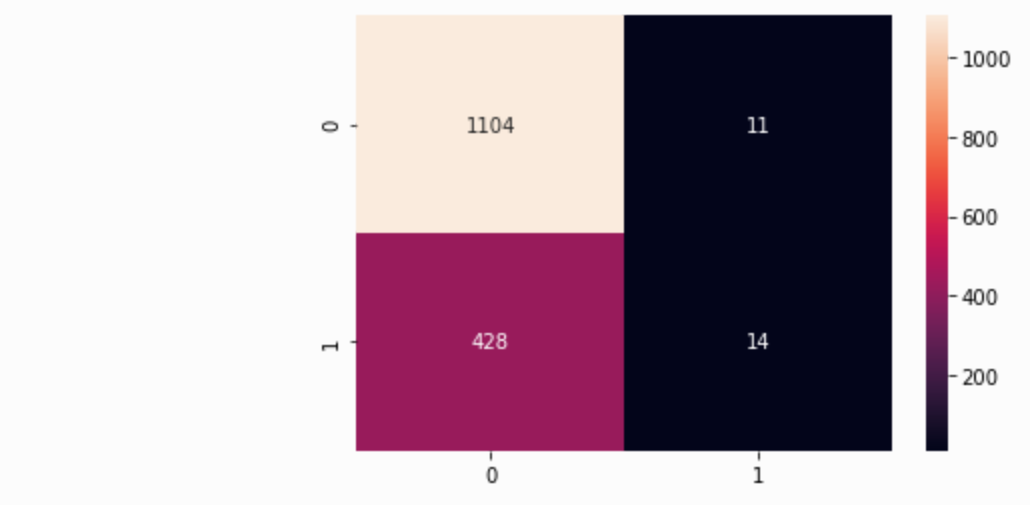

Figure 4. Confusion Matrix for Naive Bayes Classifier

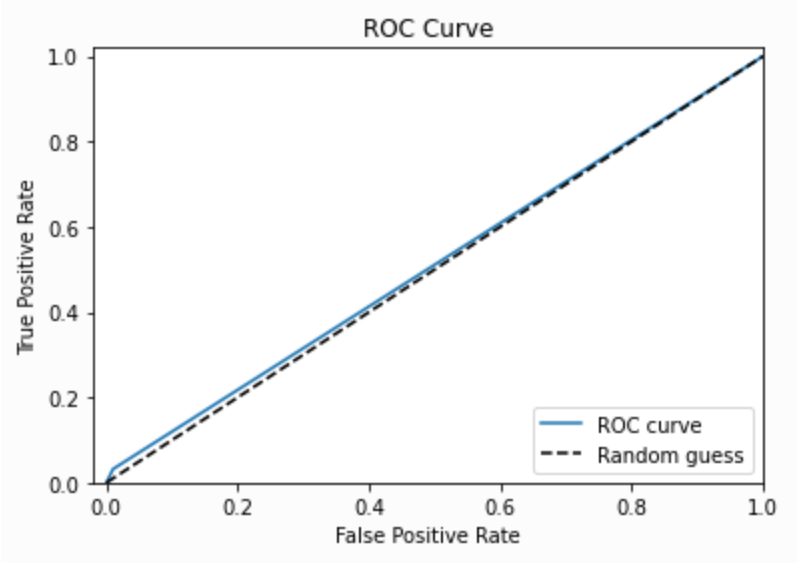

Figure 5. ROC Curve for the Naive Bayes Classifier

Furthermore, we have implemented artificial neural networks and support vector machine classifiers in this research to create a customer churn prediction model. After implementing the hyperparameter tuning (choosing the optimal parameters) and the feature selection, the accuracy score obtained in the case of the artificial neural network was 64.43%. In comparison, in the case of the support vector classifier, the accuracy score was 71.612%. The ROC curve and the confusion matrix for the artificial neural network and the support vector machine classifier are shown in the Figures below.

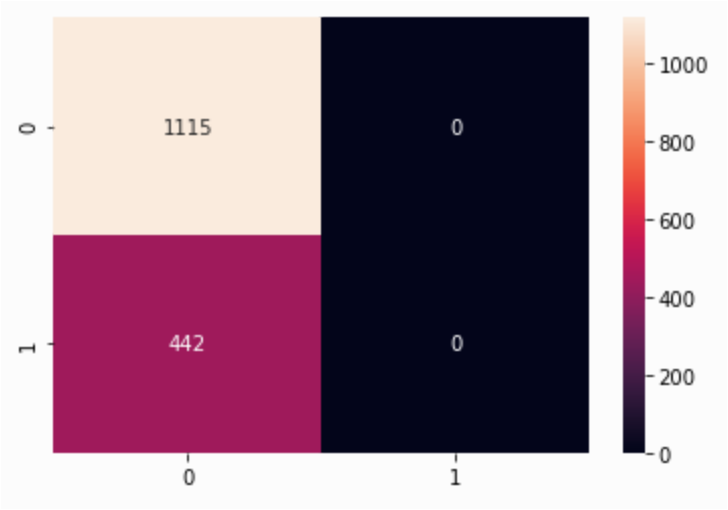

Figure 6. Confusion Matrix for Support Vector Machine Classifier

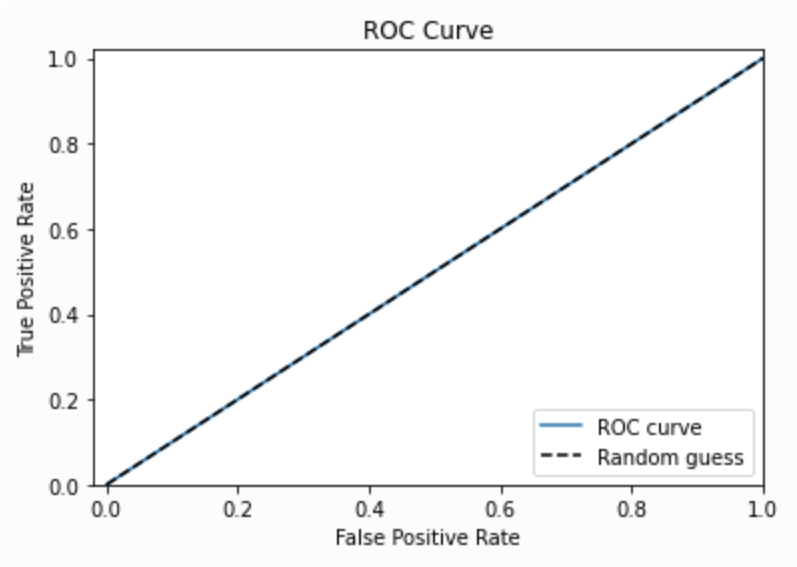

Figure 7. ROC Curve for the Support Vector Machine Classifier

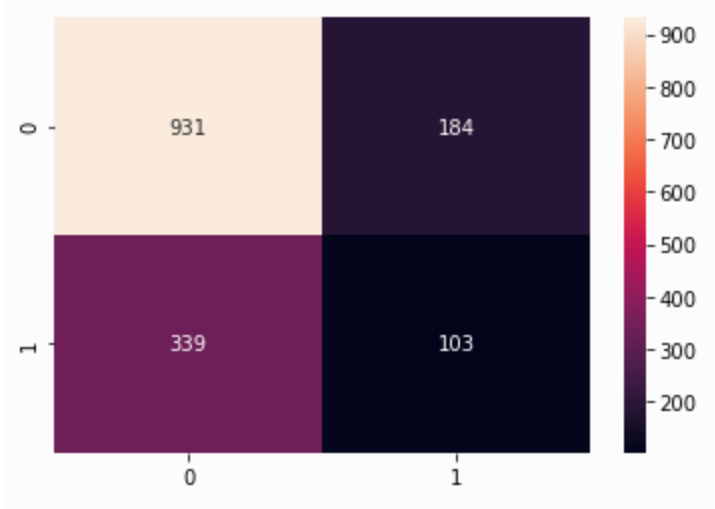

Figure 8. Confusion Matrix for the ANN

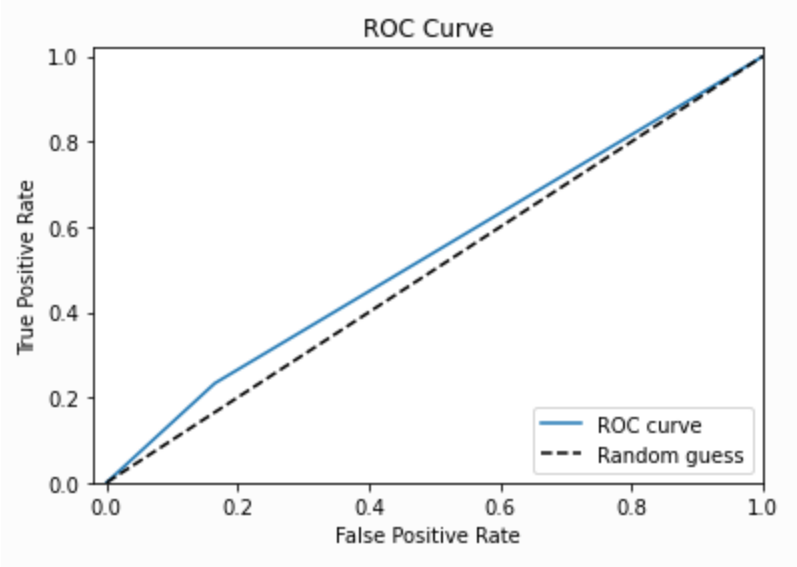

Figure 9. ROC Curve for the ANN

After evaluating each of the four models in terms of accuracy score, precision, recall, f1-score, the results show that the Support Vector Machine model performs better in this particular case on the testing data before the grid search.

In this research, an efficient customer churn prediction model was designed using both scenarios with and without feature selection. Exploratory data analysis was performed through creating bar charts, scatter plots, violin plots, and histograms and explored the relationship of the attributes with the target variables along with this correlation heat maps was also created; and comparative analysis of each of the model was performed. After analyzing each of the attributes, exploring the relationship among them and the information gathered in efforts to create an efficient customer churn prediction model will not only help to increase the realization in the industry and explore the new trends in the field of artificial intelligence and the machine learning that can be used to create a more effective churn prediction model to address the customer churn problem. The customer churn analysis will help us identify customers who will discontinue utilizing a service or the product [21]. The customer churn prediction model and analysis can be advantageous to create an efficient and robust strategy to retain the customers and reduce the churn as churn analysis helps in understanding the behaviors of the customers and predicting the customers that will discontinue a company service they receive, especially in the telecom sector, and taking up the necessary steps to prevent this cancellation [22].

Task 5:

Conclusion (447 words)

In the 21st century, the telecom sector has seen a radical increase in growth; it has emerged as one of the world's fastest-growing industries, impacting around 90% of the world population. With the increase in technology and the rapid growth and increase in services in the telecom sector, customers nowadays have multiple options. It has become difficult for telecom companies to predict those customers who are likely to churn [23]. Churn prediction is a challenging task; telecom companies need to develop such policies that they can meet the customer's needs and further improve their services to retain and attract new customers to influence the increasing market. In the past few years, a lot of research has been done by researchers in the customer churn prediction area. Almost every industry had to face the problem of customer churn, but in each sector, the reasons for the customer churn are very different. The customer churn in the telecom sector is unavoidable because of many reasons such as cost, customer service, monthly minutes, and service area to name a few.

This research has trained and evaluated four different machine learning and deep learning algorithms, i.e., Decision Tree Classifier, Naïve Bayes Classifier, Support Vector Machine Classifier, and Artificial Neural Network. The machine learning and

deep learning models were evaluated after implementing both scenarios with and without feature selection; the efficiency of these models is further improved by applying hyperparameter tuning using Grid Search CV and Randomized Search CV. The model's performance is evaluated through the accuracy score, precision, recall, f1-score. After evaluating each of the four models, the results have shown that the Support Vector Machine model performs better in this particular case on the testing data before the Grid Search. An efficient customer churn prediction model was designed for both scenarios, i.e., with and without feature selection, and exploratory data analysis was performed to explore the relationship among the attributes to create an efficient customer churn prediction model. This research work will not only increase the realization in the industry about how artificial intelligence and machine learning can be used to identify the factors that are causing customer churn and the necessary steps which can be taken to prevent the customer churn.

Along with this, it will also help to explore the new trends in the field of deep learning and machine learning so that a more effective churn prediction model can be created in the future. In the future, we can also create a web application or a mobile application to predict the churn rate. The user will upload the previous customer churn data, and using the web or mobile application; they can predict/ forecast new trends.

References:

[1] Pushkar Bhuse, Aayushi Gandhi, Parth Meswani, Riya Muni, Neha Katre. "Machine Learning Based Telecom-Customer Churn Prediction", 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020

[2] Chih Ping Wei, I-Tang Chiu, "Expert Systems with Applications", Volume 23, Issue 2, August 2002, Pages 103–112.

[3] Pradeep B, Sushmitha Vishwanath Rao, Swati M Puranik, Akshay Hegde, —Analysis of Customer Churn prediction in Logistic Industry using Machine Learning‖, International Journal of Scientific and Research Publications, Volume 7, Issue 11, November 2017 ISSN 2250- 3153.

[4] Iqbal Hanif, —Implementing Extreme Gradient Boosting (XGBoost) Classifier to Improve Customer Churn Prediction‖, Proceedings of the 1st International Conference on Statistics and Analytics, ICSA 2019, 2-3 August 2019, Bogor, Indonesia.

[5] 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 Conference on Cloud Computing and Big Data Analytics.

[6] Abinash Mishra, U. Srinivasulu Reddy, —A Comparative Study of Customer Churn Prediction in Telecom Industry Using Ensemble Based Classifiers ‖ , Proceedings of the International Conference on

Inventive Computing and Informatics (ICICI 2017) IEEE Xplore Compliant - Part Number: CFP17L34-ART, ISBN: 978-1-5386-4031-9.

[7] He Y, He Z, Zhang D,— A study on prediction of customer churn in fixed communication network based on data mining. In: Sixth international conference on fuzzy systems and knowledge discovery, vol. 1. 2009. p. 92

[8] Idris A, Khan A, Lee YS,— Genetic programming and adaboosting based churn prediction for telecom. In: IEEE international conference on systems, man, and cybernetics (SMC). 2012. p. 1328–32.

[9] Makhtar M, Nafis S, Mohamed M, Awang M, Rahman M, Deris M. —Churn classification model for local telecommunication company based on rough set theory. J Fundam Appl Sci. 2017;9(6):854–68.

[10] Burez D, den Poel V. —Handling class imbalance in customer churn prediction. Expert Syst Appl. 2009;36(3):4626–36.

[11] Brandusoiu I, Toderean G, Ha B. —Methods for churn prediction in the prepaid mobile telecommunications industry. In: International conference on communications. 2016. p. 97–100.

[12] Huang F, Zhu M, Yuan K, Deng EO. —Telco churn prediction with big data. In: ACM SIGMOD international conference on management of data. 2015. p .607–18.

[13] P. Kisioglu and Y. I. Topcu, —“Applying Bayesian belief network approach to customer churn analysis: a case study on the telecom industry of Turkey,” Expert Systems with Applications, vol. 38, no. 6, pp. 7151–7157

[14] M. H. U. Rehman, V. Chang, A. Batool, and T. Y. Wah, —“Big data reduction framework for value creation in sustainable enterprises,” International Journal of Information Management, vol. 36, no. 6, pp. 917–928, 2016.

[15] J. Z. Feng and S. Q. Cai,— “Research on the model and example of engaged customer identification in virtual brand community,” Chinese Journal of Management, vol. 17, no. 9, p. 1364, 2020.

[16] K. J. Trainor, J. Andzulis, A. Rapp, and R. Agnihotri, —“Social media technology usage and customer relationship performance: a capabilities-based examination of social CRM,” Journal of Business Research, vol. 67, no. 6, pp. 1201–1208, 2014.

[17] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, —“A comparison of machine learning techniques for customer churn prediction,” Simulation Modelling Practice and Theory, vol. 55, pp. 1–9, 2015.

[18] S. A. Neslin, S. Gupta, W. Kamakura, J. Lu, and C. H. Mason, —“Defection detection: measuring and understanding the predictive accuracy of customer churn models,” Journal of Marketing Research, vol. 43, no. 2, pp. 204–211, 2006.

[19] W. Bi, M. Cai, M. Liu, and G. Li,— “A big data clustering algorithm for mitigating the risk of customer churn,” IEEE Transactions on Industrial Informatics, vol. 12, no. 3, pp. 1270–1281, 2016.

[20] J. Z. Feng and S. Q. Cai, —“Research on the model and example of engaged customer identification in the virtual brand community,” Chinese Journal of Management, vol. 17, no. 9, p. 1364, 2020.

[21] Idris A, Khan A, Lee Y S.—“Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost based ensemble classification”. Journal of Applied Intelligence, 2013. 39(3):659-672

[22] Yu W, Jutla DN, Sivakumar SC. — “A churn-strategy alignment model for managers in mobile telecom”. In: Communication networks and services research conference, vol. 3. 2005. p. 48–53.

[23] Umayaparvathi V, Iyakutti K. — “A survey on customer churn prediction in the telecom industry: datasets, methods and metrics”. Int Res J Eng Technol. 2016;3(4):1065–70.