International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 04 | Apr 2023 www.irjet.net p-ISSN: 2395-0072

Machine Learning Approaches to Predict Customer Churn in Telecommunications Industry

B Vamsi Krishna[1], K V Bramha Reddy[2], U Sai Praneeth[3]

[1] *Electronics and Communication, Vellore Institute of Technology*
[2] *Electronics and Communication, Vellore Institute of Technology*
[3] *Electronics and Communication, Vellore Institute of Technology*

Abstract - *This project aims to develop machine learning models for predicting customer churn in the telecommunications industry. The project will analyze various customer behavior and demographic data, such as tenure, payment method,monthly charges, total charges, etc to identify patterns and build predictive models. The project will use advanced techniques, such as logistic regression, decision trees, support vector machine and random forests, to predict customer churn accurately. The study will help the telecommunications industry to understand the reasons behind customer churn and implement effective strategies to reduce customer churn rates. The results of this project can be useful for improving customer retention and enhancing the overall customer experience in the industry.*

Key Words: **Machine Learning, Logistic Regression, Random Forest, Decision Tree, Support Vector Machine**

1. INTRDUCTION

The telecommunications industry is highly competitive, with companies vying to provide the best services to attract and retain customers. One of the most significant challenges faced by this industry is customer churn-the rate at which customers switch to another service provider. Customer churn can be a costly issue for telecom companies, as it can lead to lost revenue, decreased profitability, and damage to their reputation.To address this challenge, companies are increasingly turning to machine learning approaches to predict customer churn and take appropriate measures to retain customers. This project aims to develop machine learning models using decision tree, random forest, logistic regression, and SVM algorithms to predict customer churn in the telecommunications industry.The project will use a comprehensive data set containing customer attributes such as gender, senior citizen status, tenure, phone service, multiple lines, internet service, and various other service usage patterns. By analyzing this data using machine learning models, we aim to identify the most critical factors that contribute to customer churn and develop predictive models to reduce churn rates and improve customer retention. The project's findings could be beneficial to the telecommunications industry, enabling them to take a more proactive approach towards customer churn. By anticipating customer churn, companies can implement preventive measures, such as offering incentives, personalized services, or discounts, to retain customers. The results of this project can also help to enhance the overall customer experience by addressing the issues that lead to churn.

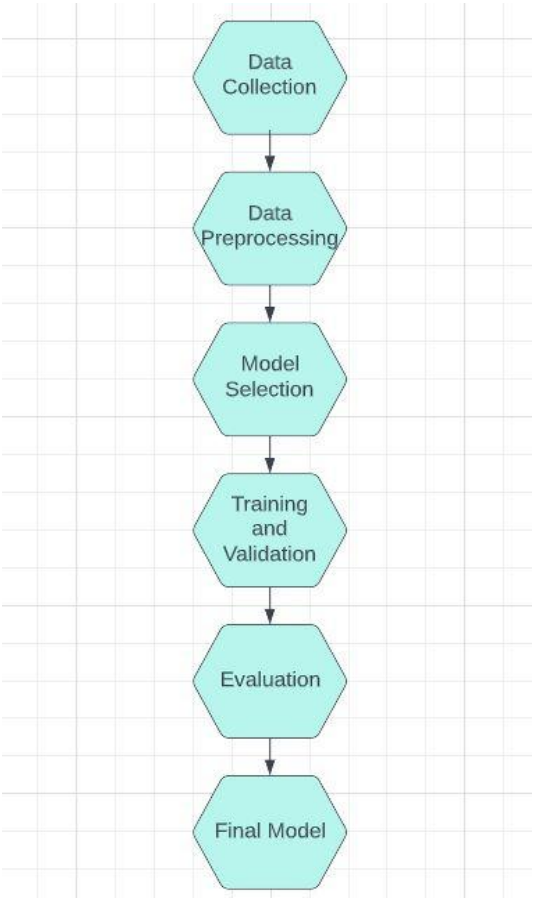
2. Tools

A. SOFTWARE

- JUPYTER NOTEBOOK
- PYTHON
- LIBRARIES-NUMPY,PANDAS,MATPLOLIB

3. METHODOLOGY

Designing of system is the process in which it is used to define the interface, modules and data for a system to specified the demand to satisfy. System design is seen as the application of the system theory. The main thing of the design a system is to develop the system architecture by giving the data and information that is necessary for the implementation of a system.



A. Data collection:

Collect a comprehensive data set containing customer attributes such as demographics, service usage patterns, payment methods, and churn information.

B. Data pre-processing:

1. Check for and handle any missing or irrelevant data.
2. Transform categorical data into numerical values
3. .Normalize or standardize the data to ensure that all features have equal importance.

C. Model selection:

Select appropriate machine learning models that can handle the problem of predicting customer churn. In this project, decision tree, random forest, logistic regression, and SVM algorithms are chosen.

D. Training and validation:

1. Split the pre-processed data set into training and validation sets.
2. Train the selected models on the training set.
3. Validate the models' accuracy and reliability on the validation set.

E. Evaluation:

1. Evaluate the performance of each model using various metrics like accuracy, precision, recall, and F1 score.
2. Compare the performance of different models and select the best-performing model based on the evaluation metrics.

F. Final model:

1. Fine-tune the selected model to improve its performance further.
2. Train the final model on the entire pre-processed data set.
3. Test the final model's performance on a new, unseen data set to ensure its accuracy and reliability.

4. MOTIVATION OF THE PROJECT

The telecommunications industry is a highly competitive market, and customer churn poses a significant challenge for companies. Churn refers to the rate at which customers switch to competitors or terminate their services altogether. High churn rates can result in substantial revenue loss and negatively impact a company's reputation.

The motivation behind this project is to develop machine learning models to predict customer churn in the telecommunications industry. By accurately predicting churn, companies can take proactive steps to retain customers, improve customer satisfaction, and ultimately, increase revenue. The project's findings can help telecom companies gain valuable insights into customer behavior and develop targeted marketing and retention strategies to reduce churn rates.

5. ALGORITHMS

The algorithms which we are using for this project are Logistic regression, Random forest , Decision tree , support, vector machine

A. Logistic Regression:

It is a well-liked classification method utilised in artificial intelligence. It is a statistical technique used to examine a dataset in which an outcome is determined by one or more independent factors. Logistic regression can be used to categorise clients as either likely to churn or likely to stay based on their prior actions and demographics when predicting new mobile internet customers. Using a logistic function, the logistic regression model calculates the likelihood of the binary result (churn or no churn). The logistic function is an S-shaped curve that transfers each real-valued number to a probability between 0 and The model determines whether a new client is likely to churn or not by learning the link between the independent variables and the chance of churn

B. Support Vector Machine:

A well-liked supervised learning approach for classification and regression analysis is called Support Vector Machines (SVM). Based on the data that is currently available, SVM can be utilised in this project to determine if a new customer is likely to churn or not. SVM algorithm works by finding the best hyperplane that separates the data into different classes. The hyperplane is defined by the support vectors, which are the points closest to the hyperplane from each class. SVM tries to maximize the margin between the support vectors and the hyperplane, which helps in generalizing the model to new data.

C. Decision Tree:

A non-parametric approach known as a decision tree is utilised for both classification and regression problems. The process divides the data into subsets based on the most important features iteratively until a stopping requirement is satisfied. The end result is a structure that resembles a tree, with each internal node standing in for a feature, each branch for a decision rule, and each leaf node for a class label or numeric value. In the context of this project, a decision tree model can be used to identify the most important features that affect the churn rate of mobile internet customers. By examining the decision rules of the tree, we can understand the factors that are most closely associated with customer churn and develop strategies to mitigate the risk of customer attrition.

D. Random Forest:

An effective machine learning technique called Random Forest is used for both classification and regression problems. A number of decision trees are built using this ensemble learning technique, and they are then combined to get a final forecast. A Random Forest model could be trained to predict whether or not a new client will churn in the context of predicting new mobile internet subscribers utilising factors like customer demographics, service usage, and billing data. A final forecast is made by combining the predictions from all the decision trees created by the Random Forest method, each of which was created using a different random subset of features and data points.

6. RESULTS

A. Logistic Regression:

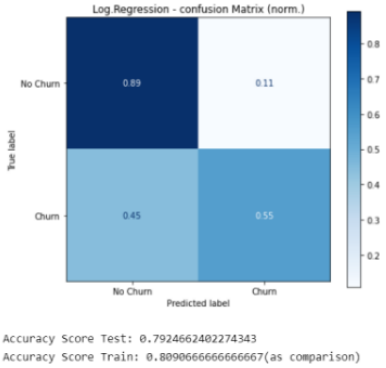


Fig1:accuracy for logistic regression from confusion matrix

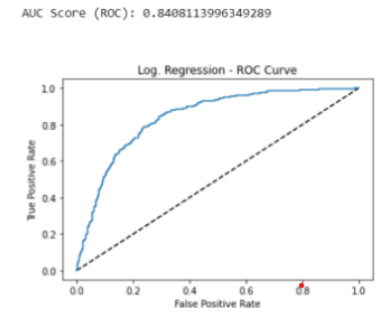


Fig2:AUC(ROC) Score for logistic regression algorithm

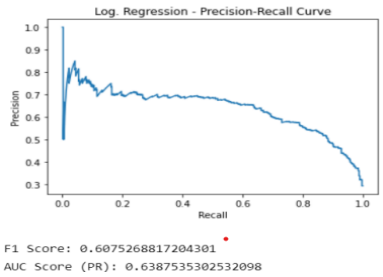


Fig3:F1 and AUC (PR) Score for logistic regression

B. Support Vector Machine:

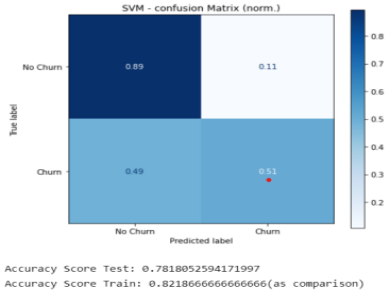


Fig4:accuracy for SVM from confusion matrix

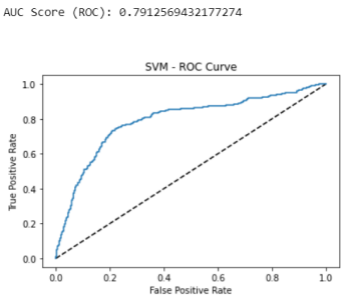


Fig5:AUC(ROC) Score for SVM algorithm

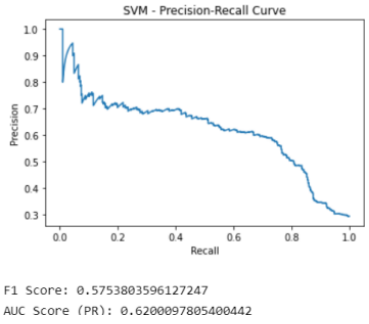


Fig6:F1 and AUC (PR) Score for SVM

C. *Decision Tree:*

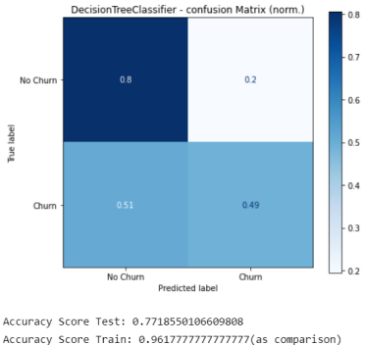


Fig7:accuracy for Decision Tree from confusion matrix

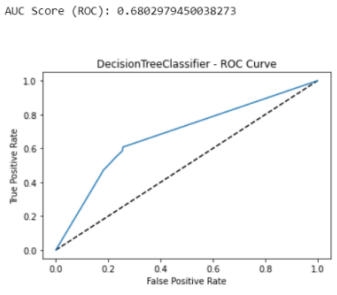


Fig8:AUC(ROC) Score for Decision Tree algorithm

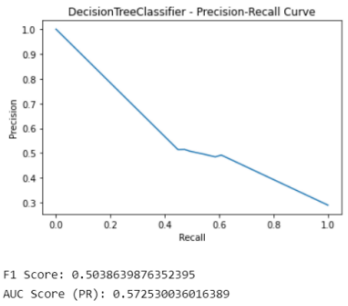


Fig9:F1 and AUC (PR) Score for Decision Tree

D. *Random Forest:*

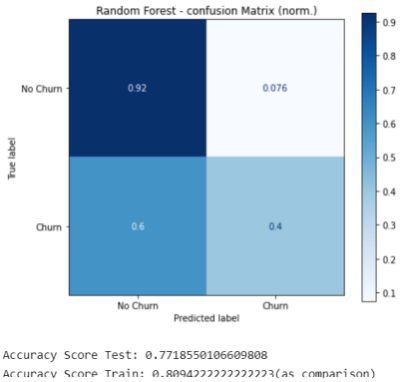


Fig10:accuracy for Random Forest from confusion matrix

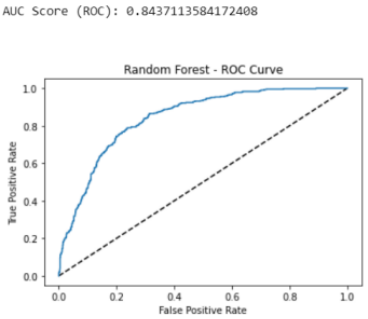


Fig11:AUC(ROC) Score for Random Forest algorithm

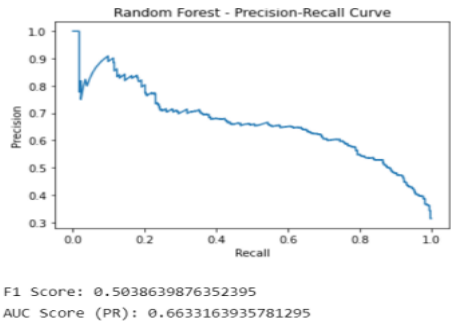


Fig12:F1 and AUC (PR) Score for Random Forest

MODEL	ACCURACY	AUC (ROC)	F1	AUC (PR)
LOGISTIC REGRESSION	79.25%	84.08%	60.75%	63.87 %
SVM	78.18%	79.12%	57.53%	62%
DECISION TREE	77.18%	68.02%	50.38%	57.25 %
RANDOM FOREST	77.18%	84.37%	50.38%	66.33 %

7. CONCLUSION

This project aimed to predict whether a customer would churn or not. The project utilized four different machine learning algorithms, namely logistic regression, SVM, random forest, and decision tree. The accuracy score, AUC score (ROC), F1 score, and AUC score (PR) were calculated for each algorithm to evaluate their performance.

Overall, the logistic regression algorithm performed the best in terms of accuracy and AUC score (ROC), making it the recommended algorithm for predicting customer churn in this context. However, it is important to note that there were tradeoffs between the different algorithms, such as the Random forest algorithm having a higher AUC score (PR) but a lower F1 score (PR). Additionally, the project identified potential outliers in the dataset that could be further investigated and addressed.

ACKNOWLEDGEMENT

We would especially want to thank PROF. MUTHU RAJA S for giving us the chance to present our project and report, as well as for assisting us at every step of the way. We also like to express our gratitude to VIT, Vellore, for giving us access to this platform.

REFERENCES

[1] Kayaalp, Fatih. (2017). Review of Customer Churn Analysis Studies in Telecommunications Industry. Karaelmas Science and Engineering Journal. 7. 696-705. 10.7212/zkufbd.v7i2.875.

[2] ZHAO, Ming-sheng & WU, Wei. (2018). Analysis and Prediction of Mobile Internet Users' Behavior Preferences. DEStech Transactions on Computer Science and Engineering. 10.12783/dtcse/CCNT2018/24780

[3] Alharbi, Y., & Alfares, H. K. (2018). Predicting customer churn in telecommunication industry using data mining techniques. Journal of Business Research, 89, 212-227. https://doi.org/10.1016/j.jbusres.2018.03.044

[4] Fu, Y., Liu, X., & Wu, T. (2019). Predicting customer churn in the telecommunications industry using machine learning algorithms. Journal of Business Research, 104, 291-303. https://doi.org/10.1016/j.jbusres.2019.06.034

[5] Huang, X., Li, Y., & Li, C. (2018). Predicting customer churn in telecommunications industry using data mining techniques: A case study. Journal of Intelligent & Fuzzy Systems, 35(6), 7049-7060. https://doi.org/10.3233/JIFS-179190

[6] Kim, S. M., & Han, J. H. (2019). A comparison of machine learning algorithms for customer churn prediction in the telecommunications industry. Journal of Industrial Engineering and Management Science, 2(1), 41-54. https://doi.org/10.1016/j.jiems.2019.03.001

[7] ang, Y., & Liu, K. (2018). Predicting customer churn in the telecommunication industry: A machine learning approach. Journal of Computational Science, 27, 312-321. https://doi.org/10.1016/j.jocs.2018.06.007

[8] Breiman, L.. (2001). Random forests, machine learning 45. Journal of Clinical Microbiology. 2. 199-228.

[9] [Jayaswal, Pretam & Prasad, Bakshi & Tomar, Divya & Agarwal, Sonali. (2016). An Ensemble Approach for Efficient Churn Prediction in Telecom Industry. International Journal of Database Theory and Application. 9. 211-232. 10.14257/ijdta.2016.9.8.21.

[10] Óskarsdóttir, María & Bravo, Cristián & Verbeke, Wouter & Sarraute, Carlos & Baesens, Bart & Vanthienen, Jan. (2020). Social Network Analytics for Churn Prediction in Telco: Model Building, Evaluation and Network Architecture.

[11] Li, Weilong & Zhou, Chujin. (2020). Customer churn prediction in telecom using big data analytics. IOP Conference Series: Materials