Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Implementation of Machine Learning Algorithms to Predicting Customer Churn for HRMS Software Vendors

Soniya Lalwani[1]
[1]*Department of Computer Science and Information Technology, Parul University Gujarat, India*
[1]*sonia.ccca@gmail.com*

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Introduction: In this competitive market it is very difficult to retain the customers, so it is very important for the HRMS software vendors to predict the customer churn to build the efficient customer retention strategies to work proactively while maintaining the profitability in the business.

Objectives: The objective of this research paper is to build a framework for customer churn prediction from the context of HRMS software vendors using the machine learning algorithms.

Methods: A customer data-set of proprietary HRMS software vendor is used to experiment various machine learning algorithms likes Decision Tree, Random Forest, Logistic Regression, Light GBM, XGBoost, two stack ensemble models are used one is Decision Tree as base model and Random Forest as meta model & another is Logistic Regression as base model and Support Vector Classifier as meta model. The dataset is pre-processed, categorical values are converted to numeric by label encoding technique, for class imbalance issue SMOTE technique is used and domain specific features are selected.Sentiment Analysis is used to read 'description' and 'solution' columns and analysed the sentiments in 1 & 0. Further to this feature engineering is performed and created the target variable from the dataset.

Results: As a result, stack ensemble model i,e Decision Tree as base model and Random Forest as meta model achieved high accuracy as compared to other machine learning models used individually.

Conclusions: This study provides a promising stack ensemble model to predict the customer churn for HRMS software vendors. This study provides a very practical and proactive approach to HRMS software vendors to build an efficient strategy to retain the customers. In future work explainable AI or deep learning techniques can be used to interpret the machine learning models working and to improve model performance.

Keywords: Customer Churn Prediction, HRMS Vendor, Machine Learning Algorithms

INTRODUCTION

In the world of competition and a competitive market, retaining a customer is a very challenging task. Especially when the customer has a choice and a lot of HRMS vendors are available to fulfil the requirements of the customer. It is truly said that retaining existing customers is more beneficial to companies as compared to acquiring new customers. One of the most critical problems is the situation when the customer discontinues the software subscription or switches to another vendor. Prediction of these customers who are going to churn shortly can help HRMS vendors to proactive take strategic action to retain the customer by improving the solution, the services or providing customers a hefty discount to ultimately increase the profitability of the organization.

Here comes the Machine learning, which is the best approach to predict customer churn by learning the hidden patterns of customer behaviour. In the past, all the HRMS vendors used to check software usage logs, or the feedback was collected to understand the customer behaviour, yet they were unable to identify exactly when and why the customer was going to end the software subscription contract. But machine learning models can handle simple and

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

complex data to understand the customer behaviour. From the context of HRMS, machine learning can be a positive and promising approach for predicting customer churn.

The objective of this paper is to examine and analyse the machine learning models used to predict customer churn for HRMS software vendors. This study aims to build an efficient model that ultimately helps HRMS software vendors to proactively identify churn and design effective customer retention strategies.

OBJECTIVES

The objective of this research paper is to build a framework for customer churn prediction from the context of HRMS software vendors using the machine learning algorithms.

Table 1: Classification Models used to predict customer churn are as follows:

Model Name	Abbreviation	Explanation
Logistic Regression	LR	It predicts outcomes in 0 or 1 based on features
Decision Tree	DT	It is like Branches of Tree predicts "Yes" and "No" based on feature values
Random Forest	RF	It combines Multiple Decision Tree Predictions and considers the average of all the Decision Tree predictions
Support Vector Machine	SVM	It separates the classes by the best hyperplane
K-Nearest Neighbors	KNN	The majority of labels are classified based on nearest neighbours
Naive Bayes	NB	It is a probability model based on Bayes' Theorem
Gradient Boosting	GB	In this technique the model is updated sequentially by correcting its previous iteration error.
XG Boost	XGB	It is an improved version of Gradient Boosting
Ada Boost	Ada Boost	In this technique, the miss classified instances are re-weighted to form a strong classifier by combining weak learners
K-Means Clustering	K-Means Clustering	In each cluster the variance is minimized in K clusters. Data is grouped into K Clusters.
Hierarchical Clustering	Hierarchical Clustering	The nested clusters are built by merging and splitting on distance metrics
Principal Component Analysis	PCA	In this technique, the features of the dataset are reduced and transformed into principal components
Artificial Neural Networks	ANN	It functions like the human brain, and through the neuron layer, complex patterns are learned
Convolution Neural Networks	CNN	The data is processed in the form of grid-like data
Recurrent Neural Network	RNN	Through feedback loops, the memory is maintained, and it is designed for sequence data.

LITERATURE REVIEW

Kumar, S. L. (2021) [1], the author provided the solution to the problem related to the banking industry. In the experiment author used famous Kaggle dataset of 10000 records with 14 features. Preprocessing techniques like Data cleaning, data imputation, Outliers handling, Data transformation and visualization were used. The author selected domain relevant features like credit score, geography, gender, age, tenure, balance, number of products, credit card usage, estimated salary, membership activity. Machine learning models like LR, DT, RF,KNN,Adaboost,Gradient Boosting, XGBoost were applied. Model validation done through Train/test Split and LOOCV. As a result RF achieved 87.22% accuracy. The papers provided very practical solution to predict the customer churn and very efficiently applied various machine learning models for comparison. But there is a room for more feature selection techniques

exploration and advance hyper parameter optimization methods. This paper supports my research, yet the paper is majorly focused on banking industry. I can adapt the methodology part and custom select the features relevant to HRMS customer churn prediction.

Kingawa, E. D., & Hailu, T. T. (2022) [2], presented the solution for Lion insurance company based in Ethiopia. The motor insurance dataset with 12007 records were experimented. In the preprocessing phase missing values were imputed using mean, mode and zeros. For feature scaling Min-Max normalization technique was applied. To address class imbalance author used SMOTE technique and for clustering K-means++ algorithm was used to label the data. To select most relevant features ExtraTreeClassifier was used. Features such as "Premium," "Carrying Capacity," and "Type of Body" were considered as important by the ExtraTreeClassifier. The author used machine learning models like Deep Neural Network (DNN), Random Forest, SVM, Naive Bayes, KNN. For Hyper parameter tuning Randomized search Optimization technique was used. As a result, DNN achieved 98.81% accuracy. In this study DNN outperformed but the dataset used was of one branch only that focused on motor insurance. This study lacks the generalization in terms of other insurance type and other domain like HRMS, etc. From this paper I can utilize the preprocessing, feature selection, hyper-parameter tuning techniques with model comparison. As the objective of this study and my research is to predict the customer churn using supervised learning algorithm.

Zimal, S., Shah, C., Borhude, S., Birajdar, A., & Patil, Prof. S. (2023) [3] ,the author aimed to solve the problem in the B2B subscription-based companies. The author experimented on financial administration subscription service dataset. Preprocessing techniques like noise removal, cleaning of missing values were used. And for class imbalance issue SMOTE and SMOTEENN were used. Categorical variables were converted to numerical using one-hot encoding technique. Relevant features were identified using domain knowledge. Machine learning models like RF, SVM, KNN were used. For hyper parameter tuning grid search and for cross validation 5-fold techniques were used. As a result, RF performed best. The paper uses various machine learning models but experiment on only one dataset. Cost benefit side remained unexplored. Also, paper does not discuss about the feature importance and scalability of the models. From this study I can use methodology like preprocessing, hyper-parameter tuning and class imbalance technique into my research. However I will need to select the HRMS domain specific features to further experiment.

Asfaw, T. (2023)[4],the author applied and experimented various machine learning models like Logistic Regression, Random Forest, Gradient Boosting, XG Boost, and Light GBM. In the preprocessing phase data was normalized using Min-Max scaler and for class imbalance issue SMOTE was utilized. Features were selected using p-values and chi-square values. As a result, Light GBM outperformed and achieved 98% accuracy. This study used robust approached, but it lacks generalization. From this study I can directly apply methods like SMOTE and can use Light GBM model to boost the prediction accuracy.

J, S., Gangadhar, Ch., Arora, R. K., Renjith, P. N., Bamini, J., & Chincholkar, Y. D. (2023) [5], The paper focuses on E-Commerce. In the preprocessing phase data normalization technique was applied. SVM model was used and for hyper parameter tuning cost and gamma values were used to increase the model accuracies achieved the high precision rate in detecting customer churn. This paper provides unique approach to use SVM for churn prediction, but various ML models can be experimented. Also, the paper is only focusing the E-commerce industry and there is a room to experiment on other domains like HRMS. From this paper I can use methodology and can apply it on HRMS customer dataset to get valuable insights.

Singh, P. P., Anik, F. I., Senapati, R., Sinha, A., Sakib, N., & Hossain, E. (2024) [6], focused on banking industry and used dataset of 10000 records from Kaggle. Preprocessing techniques like handling missing values, Data Normalization and removal of irrelevant features were included. For class imbalance issue SMOTE was used. Using feature engineering technique new variables like 'TenureByAge', 'BalanceSalaryRatio' & 'CreditScoreGivenAge' were introduced. Machine learning models like LR, RF, SVM, XGBoost were employed. For hyper parameter tuning Grid search used to optimize the models. As a result, RF outperformed. Indepth analysis for ML models were conducted but it was on limited dataset also it focused on banking industry on other domains like HRMS this experiment can be conducted to check the accuracy.

Maduna, M., Telukdarie, A., Munien, I., Onkonkwo, U., & Vermeulen, A. (2024) [7],focused on banking sector. The author experimented on banking dataset of 10k records and 14 features. In the preprocessing phase Missing data was

handled using ignoring tuples or manual imputation. For class imbalance issue SMOTE was used. Machine Learning models like DT, RF, SVM, KNN were employed. As a result, RF achieved highest accuracy of 87% percentage. In this paper clear methodology was used and various models were employed, and in-depth analysis was conducted for the results, but the study was limited to basic machine learning models only. From this study I can adapt the methodology and ML techniques, and I can choose the domain relevant feature for build the customer churn prediction model for HRMS vendors.

Poudel, S. S., Pokharel, S., & Timilsina, M. (2024) [8], the author has focused on telecom industry.The author has used Kaggle dataset 7043 records with 30 features. In the preprocessing technique author dropped the null values, removed duplicates and converted categorical variable to numeric using one-hot encoding technique. Author also created new features like Engagement Score, Service Utilization. Machine learning models like SVM, LR, RF, GBM, KNN were employed. As a result, GBM achieved higher accuracy of 81%. In this paper SHAP was for interpret-ability and in-depth comparison of models was demonstrated but the dataset was used of telecom industry only whereas other domains like HRMS can be explored. Furthermore advanced deep learning techniques can be explored.

R, A., Khan S, A., Murugan, H., & Ks, N. (2025) [9], focused on telecom industry. As the preprocessing stage feature selection and clustering were utilized. Out-of-Bag (OOB) error was employed for feature selection. Machine learning models like Kmeans & LR, SVM, NB, LSTM-RNN & Ensemble models were employed. As a result, ensemble models performed well. In this paper it was observed that there was a extensive use of clustering algorithms and more importance was provided to feature selection to improve the model accuracy, but the dataset used was domain specific. There is a room to experiment on HRMS Vendor Domain specific dataset. From this paper I can utilize the methodology, and the machine learning models on HRMS Vendor Dataset also I need to use the domain related features to build the prediction model accordingly.

Department of Information Technology from Matoshri Aasarabai Polytechnic., & Talele, A. (2025)[10], the author has aimed to provide solution for telecom industry. The author has used Kaggle dataset. In the preprocessing phase used techniques like Data cleaning, feature selection, and splitting into training and testing subsets. For feature engineering author has not mentioned anything specifically. Machine learning models like RF, DT, XGboost were employed. As a result, XGBoost outperformed. This paper uses strong predictive models but lacks discussion on model analysis. Also, this paper only focuses on telecom dataset. From this paper I can utilize the methodologies and models discussed for HRMS software customer churn prediction. Machine Learning models like Random Forest and XGBoost can help to analyse behavioural patterns in the customer churn dataset related to HRMS Software.

METHODS

Dataset: The dataset is of a proprietary HRMS platform usage data (n ≈ 12,268 records)

Features: The dataset was received in CSV format with 19 features and 12268 entries. The datatype of all the features was “Object”. The features were as mentioned below:

Index(['Sr.no', 'Date', 'Support Type', 'Issue Type', 'Product', 'Module', 'Module Form', 'Client Name', 'Contact Person Name', 'Contact Person-Number', 'Email ID', 'Description', 'Warranty status', 'Receive by', 'Status', 'If Forward then Assigen Name', 'Compilation date', 'Time Taken for solution in minutes', 'Resolution'], dtype='object')

Data Preprocessing: Imported necessary Libraries: Like Pandas,numpy,matplotlib, seaborn

Read Dataset: Using csv file

Sanity Check of the dataset was initiated: Checked for duplicates & checked for null values

Exploratory Data Analysis EDA: Explored and analysed the dataset using Scatter plot & Boxplot

Encoding: Label encoding for categorical variables

Categorial Features like 'Support Type', 'Issue Type', 'Product', 'Module', 'Module Form', 'Client Name', 'Contact Person Name', 'Warranty status', 'Receive by', 'Status', 'If Forward then Assign Name'] were converted into 64-bit signed integer.

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Object to Datetime conversion: Features like 'Date' & 'Compilation date' were converted in datetime64[ns]

Missing value treatments: The missing values in the dataset 221. For numerical features decided to use imputation with a mean strategy to treat the missing values, and for categorical features used the strategy. Used a simple imputer from scikit-learn to fill the missing values.

Outlier Treatments: Identified the outliers using box plots, scatter plots, and (IQR) Interquartile Range. There were three strategies for this dataset like Trimming/Removing outliers, Winsorizing meaning capping the values at certain percentiles, and Transformation e.g. log transformation. All the three strategies were used to treat the outliers.

Duplicate and Garbage Value Treatment: A few values in the dataset were identified as garbage. Replaced the garbage values to nan. Check for the duplicate values.

Normalization: Normalized the dataset using Min-Max Scaling & Standardization (Z Score)

Application of Sentiment Analysis: At the data exploration time it was found that the target variable was not provided in the data set so on 'Description' and 'Resolution' experimented the Roberta Sentiment Score (RSS). As the (RSS) was required to process 12000 plus records so using a batch processing function. The sentiment analysis 0 = positive and 1 = negative. The new features created in the dataset were 'Description_Sentiment_Label' & 'Resolution_Sentiment_Label'.

Created a 'Churn' variable: As in the dataset, there was no 'Churn' label or target variable missing. Created one column with the condition like if 'Description_Sentiment_Label' is 1 & 'Resolution_Sentiment_Label' is 1, meaning the churn is 1 or Negative and 'Description_Sentiment_Label' is 0 & 'Resolution_Sentiment_Label ', the churn is 0, meaning Positive.

Dropped irrelevant features: 'S.No',' Contact Person Name', 'Contact Person-Number',' Email ID', 'Warranty status'.

Furthermore, we analysed that 'Description' and 'Resolution' columns were used for sentiment analysis and the score was stored in 'Description_Sentiment_Label' & ' Resolution_Sentiment_Label' removed the original the features 'Description' & 'Resolution' from the dataset.

'Churn' label analysis: We examined the 'churn' feature and, to handle inaccuracy in the training dataset, we simply removed 'Description_Sentiment_Label' & 'Resolution_Sentiment_Label'.

Feature Engineering: Converted the 'Date 'and 'Compilation date' to year, month, day, weekday and weekends for further analysis and Converted 'Time Taken for solution in minutes' into minutes as in the dataset it was provided in object datatype.

Hyper Parameter Tuning: We used GridsearchCV for hyper parameter tuning and found the Best hyper parameters: {'C': 0.001, 'penalty': 'l1', 'solver': 'liblinear'} & Best score: 0.862237634775456

Machine Learning Model Implemented: Implemented RandomForestClassifier, SVC, KNeighborsClassifier, DecisionTreeClassifier and used StandardScaler for standardization. The 'Churn'feature is considered as a Target variable, so I have separated the dataset in to training and test by 80:20 ratio.

Table 2: Machine Learning Models Train and Test Accuracy are as follows:

Model	Training Accuracy	Test Accuracy
Logistic Regression	0.862238	0.855338
Random Forest	0.996943	0.854523
SVM	0.862645	0.855338
KNN	0.871510	0.839446

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Decision Tree	0.996943	0.767726
Light GBM	0.883839	0.859006
XGBoost	0.918178	0.852486

As identified, there was a class imbalance issue so handled it using SMOTE technique. Post implementing, the SMOTE all the classed were balanced. And then plotted the confusion matrix. Table 3: Checked if all the classes are balanced shown in the table are as follows:

	Predicted: 0	Predicted: 1
Actual: 0	1191	917
Actual: 1	865	1252

Table 4: Class Distribution for Churn was as follows:

0	10000
1	2000

Post application of SMOTE, the accuracy improved. I have used the stack or ensemble model strategy to build the model. Use a decision tree a e as a base model and Random Forest for Meta Model. Table 5: Machine Learning Models result metrics are as follows:

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.824852	0.813382	0.844119	0.828465
Random Forest	0.888521	0.905419	0.868210	0.886424

Further to this Logistic Regression was applied as base model and Support Vector Classifier as meta learner. Table 6: Ensemble Machine Learning Model Logistic Regression and SVC Accuracy are as follows:

Metric	Train	Test
Accuracy	0.573060	0.579408
Precisión	0.572031	0.578558
Recall	0.578399	0.591403
F1-Score	0.575197	0.584910

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

RESULTS & DISCUSSION

In this experiment, various machine learning algorithms like Decision Trees, Random Forest, Support Vector Machine, K-Nearest Neighbour, LightGBM, XGBoost were applied to predict customer churn. Even to improve the accuracy of the model, two-stack models were also applied, like the stack model 1 consists of a Decision Tree as a base model and Random Forest as a meta learner and in stack model 2 consists of Logistic Regression as a base model and SVC as a meta learner.

All the machine learning models were examined using classification metrics, which consist of Accuracy, Precision, Recall & F1-Score. As a result of the experiment stack model outperformed as compared to others and achieved 88% accuracy, 90% Precision, 86% Recall and 88% F1-Score. This study and the outcome of the study confirms that stack models is promising approach to predict customer churn for HRMS software vendors.

CONCLUSION & FUTURE WORK

The objective of this study is to develop an efficient and effective machine learning based solution for predicting customer churn for HRMS software Vendors. Almost all the supervised learning algorithms were applied but stack model like combining decision tree with random forest outperformed.

Further, SHAP or LIME can be implemented to understand the model's decision-making process, and this helps business owners build the optimum customer retention strategy. Deep learning models can be implemented to understand customer behaviour. The model can be experimented across different domain datasets apart from the HRMS software vendor test text generalization.

REFERENCES

[1] Kumar, S. L. (2021). Bank Customer Churn Prediction Using Machine Learning. International Journal for Research in Applied Science and Engineering Technology, 9(VIII), 727–732. https://doi.org/10.22214/ijraset.2021.37467

[2] Kingawa, E. D., & Hailu, T. T. (2022). Customer Churn Prediction Using Machine Learning Techniques: the case of Lion Insurance. Asian Journal of Basic Science & Research, 04(04), 60–73. https://doi.org/10.38177/AJBSR.2022.4407

[3] Zimal, S., Shah, C., Borhude, S., Birajdar, A., & Patil, Prof. S. (2023). Customer Churn Prediction Using Machine Learning. International Journal for Research in Applied Science and Engineering Technology, 11(2), 872–883. https://doi.org/10.22214/ijraset.2023.49142

[4] Asfaw, T. (2023). Customer churn prediction using machine-learning techniques in the case of commercial bank of Ethiopia. The Scientific Temper, 14(03), 618–624. https://doi.org/10.58414/SCIENTIFICTEMPER.2023.14.3.08

[5] J, S., Gangadhar, Ch., Arora, R. K., Renjith, P. N., Bamini, J., & Chincholkar, Y. D. (2023). E-commerce customer churn prevention using machine learning-based business intelligence strategy. Measurement: Sensors, 27, 100728. https://doi.org/10.1016/j.measen.2023.100728

[6] Singh, P. P., Anik, F. I., Senapati, R., Sinha, A., Sakib, N., & Hossain, E. (2024). Investigating customer churn in banking: a machine learning approach and visualization app for data science and management. Data Science and Management, 7(1), 7–16. https://doi.org/10.1016/j.dsm.2023.09.002

[7] Maduna, M., Telukdarie, A., Munien, I., Onkonkwo, U., & Vermeulen, A. (2024). Smart Customer Churn Management System Using Machine Learning. Procedia Computer Science, 237, 552–558. https://doi.org/10.1016/j.procs.2024.05.139

[8] Poudel, S. S., Pokharel, S., & Timilsina, M. (2024). Explaining customer churn prediction in telecom industry using tabular machine learning models. Machine Learning with Applications, 17, 100567. https://doi.org/10.1016/j.mlwa.2024.100567

[9] R, A., Khan S, A., Murugan, H., & Ks, N. (2025). Clustering Comparison of Customer Attrition Dataset using Machine Learning Algorithms. International Journal of Innovative Science and Research Technology, 3432–3436. https://doi.org/10.38124/ijisrt/24apr643

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[10] Department of Information Technology From Matoshri Aasarabai Polytechnic., & Talele, A. (2025). Customer Churn Prediction Using Machine Learning. INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, 09(02), 1–9. https://doi.org/10.55041/IJSREM41260