Customer Churn Prediction Analysis using Machine Learning Models

Rishabh Kumar

PG Student (Masters of Integrated Technology)
Department of Computer Science and Engineering
Noida Institute of Engineering and Technology
Greater Noida
Affiliated to AKTU Lucknow

---***---

Abstract - *Customer churn prediction is a critical aspect for businesses aiming to retain their client base in competitive markets. It is easier to retain a customer than to convert a new customer successfully This study applies advanced machine learning techniques to predict customer churn, leveraging a rich dataset with features including demographic information, service usage patterns, and customer account information. The analysis achieves robust prediction accuracy by employing methods such as SMOTE-Tomek resampling to address the class imbalance, and utilizing algorithms like LightGBM, Random Forest, Xgboost, etc. Feature importance techniques, including permutation importance and SHAP values, are employed to identify key factors influencing churn. This comprehensive approach provides actionable insights for targeted retention strategies, ultimately aiming to reduce churn rates and enhance customer satisfaction.*

Key Words**: SMOTE-Tomek, Xgboost, LightGBM, Random forest, SHAP values, Hyperparameter tuning.**

1. INTRODUCTION

The globalization and advancements in the telecommunication industry have significantly increased market competition by introducing numerous operators. To maximize profits, companies adopt strategies such as acquiring new customers, up-selling, and extending the retention period of existing customers. Among these, retaining current customers is the most cost-effective.

The primary goal of customer churn prediction is to develop strategies for customer retention. With growing market competition, the risk of churn rises, making it essential to track loyal customers. Churn prediction models aim to identify early signs of churn and forecast customers likely to leave, helping companies leverage their valuable databases to maintain customer loyalty and mitigate churn.

1.1 Exploratory Data Analysis.

- Performed EDA on churn datset to find out relationships bw various features that exist and conducted various statistical tests like t-test , anova test etc to confirm the relationships.

- Churn prediction decreases with tenure , people who spend more time with the company are likely to churn less.

- People who use fiber optics are likely to churn mostly.

- People having month-to-month contract prefer paying by Electronic Check mostly or mailed check. The reason might be short subscription cancellation process compared to automatic payment.

- People with no internet service are least likely to churn followed by people who have online security and at last comes people who have highest chances of churning are those people who don’t have any online security despite having internet services:

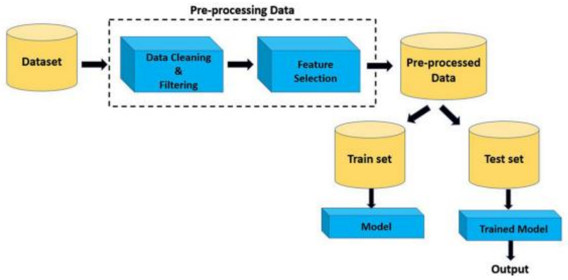


Fig1. System Architecture

1.2 Machine learning models

Random forest classifier:

- Random Forest is an ensemble learning method that builds multiple decision trees and merges them to get a more accurate and stable prediction. In churn prediction, it works by creating a large number of decision trees during training and outputting the class that is the mode of the classes of the individual trees.

- Its robustness to overfitting and ability to handle high-dimensional data make Random Forest a reliable algorithm for identifying churn patterns and predicting customer behavior.

Xgboost:

• XGBoost, or Extreme Gradient Boosting, is a highly efficient and scalable implementation of gradient boosting. In churn prediction, it constructs an ensemble of decision trees in a sequential manner, where each tree corrects errors made by its predecessors.

• Its ability to handle large datasets and optimize performance with techniques like regularization and parallel processing makes XGBoost a popular choice for identifying customers at risk of churn, offering high accuracy and robustness

LightGBM:

• LightGBM, or Light Gradient Boosting Machine, is designed for speed and efficiency. It uses histogram-based algorithms to bucket continuous feature values, significantly reducing memory usage and increasing training speed. In churn prediction, LightGBM excels with large datasets and high-dimensional data, providing fast and accurate predictions.

• Its ability to handle categorical features natively and its advanced optimization techniques make it an excellent choice for churn analysis.

SVM:

• Support Vector Machine is a supervised learning model that finds the optimal hyperplane which best separates the classes in the feature space. In churn prediction, SVM aims to classify customers into churners and non-churners by maximizing the margin between the two classes. We need a hyperplane which has largest margin .

• With kernels, SVM can handle non-linear relationships, making it suitable for complex datasets where the relationship between features and churn is not purely linear. Its strength lies in high-dimensional spaces and smaller datasets.

Research Methodology

1. Introduction

The objective of this study is to develop a predictive model for customer churn using machine learning techniques. The methodology involves several key steps: data preprocessing, feature engineering, dimensionality reduction using Principal Component Analysis (PCA), model training with various algorithms, hyperparameter tuning, and model evaluation.

2. Data Collection

The dataset used in this study consists of customer data, including demographic information, account information, and service usage patterns. The target variable is 'Churn,' indicating whether a customer has churned.

3. Data Preprocessing

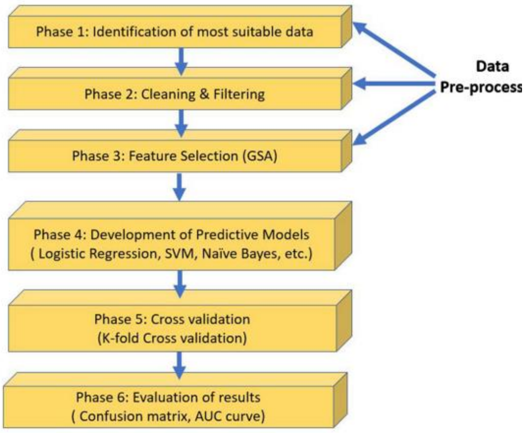

Phase 1: Identification of most suitable data
Phase 2: Cleaning & Filtering
Phase 3: Feature Selection (GSA)
Phase 4: Development of Predictive Models (Logistic Regression, SVM, Naïve Bayes, etc.)
Phase 5: Cross validation (K-fold Cross validation)
Phase 6: Evaluation of results (Confusion matrix, AUC curve)
Data Pre-process

Fig2.Phase model for developing a customer churn management framework.

Data preprocessing involved the following steps:

Handling Missing Values: Missing values were imputed using appropriate methods such as mean imputation for numerical features and mode imputation for categorical features.

Encoding Categorical Variables: Categorical variables were encoded using one-hot encoding to convert them into numerical form.

Feature Scaling: Features were standardized using StandardScaler from `sklearn preprocessing` to ensure that all features have a mean of 0 and a standard deviation of 1.

4. Dimensionality Reduction using PCA

To address the issue of high dimensionality and to improve the efficiency of the machine learning algorithms, PCA was applied:

Standardization: The dataset was standardized before applying PCA to ensure that each feature contributes equally to the principal components.

PCA Application: PCA was applied to reduce the dataset to a lower-dimensional space while retaining 95% of the variance. This step helps in noise reduction and improves the performance of the machine learning models.

5. Model Selection and Training

Several machine learning algorithms were considered for training the predictive model, including:

- Logistic Regression

- Random Forest Classifier

- Support Vector Machine (SVM)

- XGBoost Classifier

6. Hyperparameter Tuning

For each algorithm, hyperparameter tuning was performed using `GridSearchCV` to identify the optimal set of hyperparameters. This process involved defining a grid of hyperparameters and evaluating the model performance using cross-validation. The scoring metric used for evaluation was the area under the ROC curve (AUC).

7. Model Evaluation

Performance indicators

5.1.2 Recall

It is the ratio of real churners (i.e. True Positive), and is calculated under the following:

$$Recall = \frac{T_p}{T_p + F_n}$$

5.1.3 Precision

It is the ratio correct predicted churners, and is calculated under the following:

$$Precision = \frac{T_p}{T_p + F_p}$$

5.1.4 Accuracy

It is ration of number of all correct predictions, and is calculated under the following:

$$Accuracy = \frac{(T_p + T_n)}{(T_p + F_p + Tn + F_n)}$$

5.1.5 F-Measure

It is the harmonic average of precision and recall, and it is calculated under the following:

$$F - measure = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)}$$

A better combined precision and recall achieved by the classifier is implied due to a value closer to on

The performance of each model was evaluated on the test dataset using the following metrics:

Accuracy: The proportion of correctly classified instances.

AUC-ROC: The area under the receiver operating characteristic curve, which provides a measure of the model's ability to distinguish between classes.

Auc Curve Analysis:

To quantify the models performance on positive and negative classes of the test set,

AUC curve has been used. Higher the value of the AUC score, the better the model performs on both positive and negative classes. The obtained AUC scores of different predictive models which are used to predict the target variable , graphically represents the obtained AUC scores of Logistic Regression, Logis�tic Regression , Decision Trees,

(Extra Trees) , Random Forest, SVM Linear, Logistic Regression respectively.

S.no	Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
1	XGBoost	0.804	0.83	0.91	0.87	0.852
2	SVM	80.1	0.84	0.90	0.87	0.852
3	Logistic regression	0.79	0.76	0.89	0.87	0.851
4	Random Forest	0.72	0.91	0.67	0.77	0.82

Table1. Result analsyis.

XGBoost Classifier outperforms over other respective algorithms on the test set having an AUC score of 85%.

The XGBoost Classifier achieved the highest performance with an accuracy of 80.2% and an AUC score of 0.852 along with SVM (Same results).

8. Results

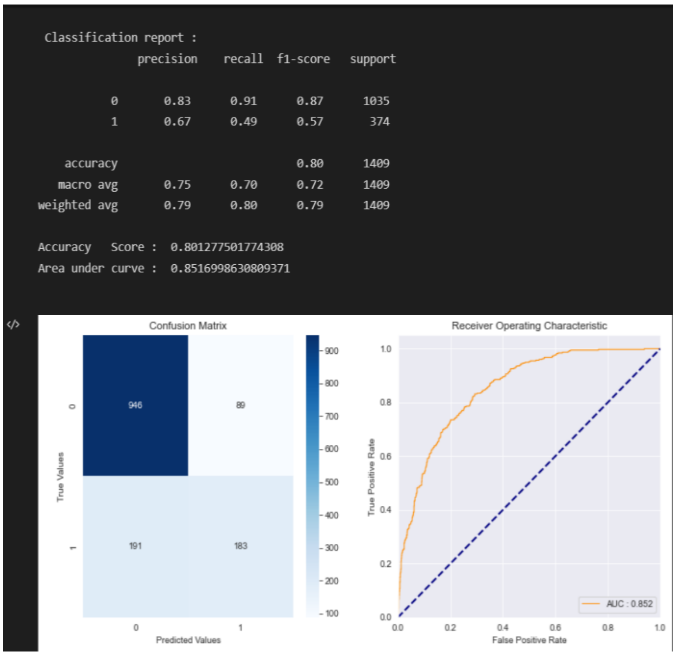
The results indicate that the XGBoost Classifier, with the application of PCA and hyperparameter tuning, outperformed the other algorithms. The key findings are summarized as follows:

Logistic Regression: Achieved an accuracy of 80.4% and an AUC of 0.851.

Random Forest Classifier: Achieved an accuracy of 79.5% and an AUC of 0.83.

SVM: Achieved an accuracy of 80.1% and an AUC of 0.852.

XGBoost Classifier: Achieved an accuracy of 80.2% and an AUC of 0.852, making it the best-performing model.



```
  Classification report :
               precision    recall  f1-score   support

           0       0.83      0.91      0.87      1035
           1       0.67      0.49      0.57       374

    accuracy                           0.80      1409
   macro avg       0.75      0.70      0.72      1409
weighted avg       0.79      0.80      0.79      1409

Accuracy   Score :  0.801277501774308
Area under curve :  0.8516998630809371
```

Fig 3. Results of model XGBoost.

AUC curves of all models analysis:

XgBoost:

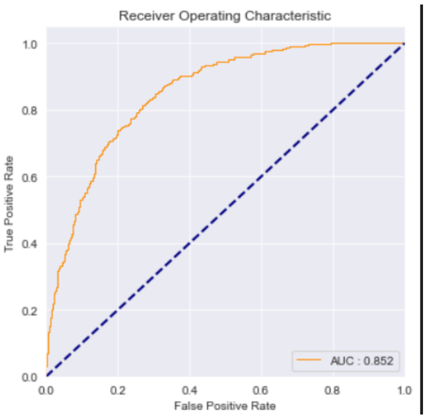


Figure 4.

LightGBM:

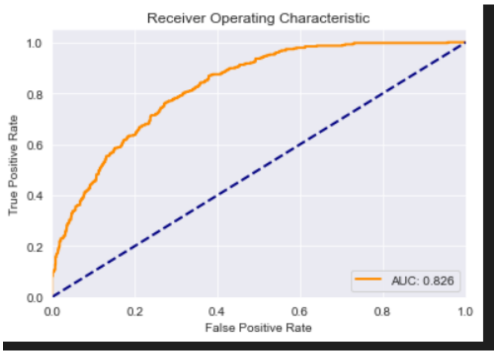


Figure 5.

Logitsic Regression:

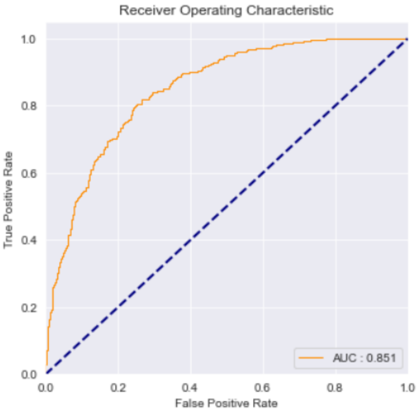


Figure 6.

Random Forest:

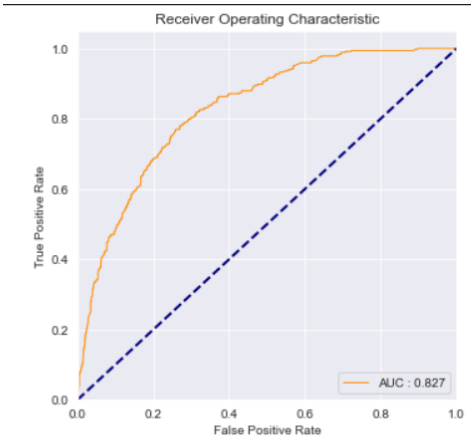


Figure 7.

3. CONCLUSIONS:

This study demonstrates the effectiveness of PCA for dimensionality reduction and the importance of hyperparameter tuning in improving model performance. The XGBoost Classifier emerged as the most accurate and reliable model for predicting customer churn in this dataset. Future work could explore additional feature engineering techniques and the use of ensemble methods to further enhance model performance.

Xgboost with hyperparameter tuning has best auc *r*esults which is 85.2% and here are the reasons for this:

• **Robustness to Noisy Data**: Churn prediction datasets often contain noisy and complex information. XGBoost's robustness helps it handle such noisy data effectively, enabling it to extract meaningful patterns even from intricate datasets.

• **Handling Non-linear Relationships**: Customer behavior, which influences churn, often exhibits non-linear relationships with various predictors. XGBoost's ability to capture complex interactions between predictors allows it to model these non-linear relationships accurately.

• **Feature Importance**: XGBoost provides insights into feature importance, indicating which factors have the most significant influence on churn prediction. This helps businesses understand the key drivers of churn and focus their efforts accordingly.

• **Ensemble Learning:** XGBoost is an ensemble learning method that combines the predictions of multiple individual models (trees) to produce a robust and accurate final prediction. This ensemble approach mitigates overfitting and improves generalization performance, crucial for reliable churn prediction.

REFERENCES

1. Customer churn prediction system: a machine learning approach Praveen Lalwani1 · Manas Kumar Mishra1 · Jasroop Singh Chadha1 · Pratyush Sethi1 Received: 19 June 2020 / Accepted: 12 January 2021 © The Author(s), under exclusive licence to Springer-Verlag GmbH, AT part of Springer Nature 2021.

2. T. Mitchell. Machine Learning. WCB/Mc Graw Hill, Boston, et al., 1997

3. Customer churn prediction in telecommunications Bingquan Huang ⇑ , Mohand Tahar Kechadi, Brian Buckley School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland.

4. Customer churn prediction system: a machine learning approach Praveen Lalwani1 · Manas Kumar Mishra1 · Jasroop Singh Chadha1 · Pratyush Sethi1 Received: 19 June 2020 / Accepted: 12 January 2021 © The Author(s), under exclusive licence to Springer-Verlag GmbH, AT part of Springer Nature 2021

5. Petrison LA, Blattberg RC, Wang P (1997) Database marketing: Past, present, and future. Journal of Direct Marketing 11(4):109–125

6.. Qureshi, S.A., Rehman, A.S., Qamar, A.M., Kamal, A., Rehman, A.: Telecommunication subscribers' churn prediction model using machine learning. In: Eighth International Conference on Digital Infor�mation Management (ICDIM 2013), pp. 131–136. IEEE (2013)

7. Radosavljevik D, van der Putten P, Larsen KK (2010) The impact of experimental setup in prepaid churn prediction for mobile telecommunications: What to predict, for whom and does the customer experience matter? Trans. MLDM 3(2):80–99

8. Rajamohamed R, Manokaran J (2018) Improved credit card churn prediction based on rough clustering and supervised learning techniques. Cluster Computing 21(1):65–77

9 Rodan A, Faris H, Alsakran J, Al-Kadi O (2014) A support vector machine approach for churn pre�diction in telecom industry. International journal on information 17(8):3961–3970

10 . Shaaban E, Helmy Y, Khedr A, Nasr M (2012) A proposed churn prediction model. International Journal of Engineering Research and Applications 2(4):693–697