WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Customer Churn Prediction using Machine Learning Classifiers

Urva Bhatnaggar [a], Mokshda Dubey [a], Udit Saran Bhatnagar [a], Aparna Pandey [b]*

[a] *Student, Department of Computer Science & Engineering, Bhilai Institute of Technology Raipur, Raipur, India*
[b] *Asst. Prof., Department of Computer Science & Engineering, Bhilai Institute of Technology Raipur, Raipur, India*
DOI : https://doi.org/10.55248/gengpi.5.1124.3425

ABSTRACT:

Customer churn prediction is an important part of business strategy because it enables companies to predict and retain at-risk customers. In this paper, the author explores the use of machine learning classifiers to predict customer churn on a dataset downloaded from Kaggle, which contains 7043 rows and 21 attributes. The methodology involves in-depth exploratory data analysis to understand the significance and relationship between the features and then proceeds to data preprocessing using one-hot encoding to handle categorical variables. SMOTEENN was used to balance the dataset with the problem of class imbalance; hence, the performance of the model improved considerably. The following machine learning classifiers have been used: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and CatBoost. Accuracy was used as the main performance metric for the models. Among these classifiers, the best performance was achieved by the CatBoost algorithm with the highest accuracy of 98.1%. This indicates that ensemble methods with balanced datasets are very effective for churn prediction tasks. In doing so, this work establishes the importance of using balanced data techniques and an algorithm to achieve the highest accuracy for prediction. From that aspect, the obtained insights have been useful to those using advanced machine learning approaches toward improving customer churn management within business organizations.

Keywords: Customer Churn, KNN, SVM, CatBoost, GridSearchCV, SMOTEENN

Introduction:

In the modern business environment, retaining customers has become essential for ensuring consistent profitability and achieving sustainable growth. Organizations across industries are increasingly recognizing the critical role of customer loyalty, as the cost of acquiring new customers is significantly higher compared to the expense of retaining current ones.

Customer churn, the phenomenon of customers discontinuing their association with a business, has therefore become a pressing challenge. The ability to predict and proactively mitigate churn is no longer optional but a strategic necessity for businesses seeking to maintain a competitive edge. Customer churn prediction is a data-driven approach that leverages analytical techniques and machine learning models to identify patterns indicative of a customer's likelihood to leave. With the advent of big data and advancements in computational technologies, businesses now have access to vast repositories of customer data, encompassing behavioral trends, transactional histories, and demographic information. Harnessing this wealth of information allows organizations to build predictive models capable of offering actionable insights into customer behavior, enabling the design of targeted retention strategies. High churn rates can adversely impact operational efficiency and resource allocation, leading to increased marketing expenditure and diminished workforce productivity. Furthermore, retaining existing customers is instrumental in enhancing customer lifetime value (CLV), fostering brand advocacy, and securing market share. For subscription-based industries, such as telecommunications, SaaS, and financial services, where recurring revenue forms the backbone of the business model, the ramifications of customer churn are particularly pronounced.

Predictive modeling thus acts as a powerful tool for identifying at-risk customers and prioritizing intervention efforts, translating into measurable business outcomes. From a business perspective, customer churn prediction aligns with the broader goals of customer relationship management (CRM). By integrating predictive insights into CRM workflows, companies can personalize customer engagement, optimize service delivery, and elevate the overall customer experience. For example, machine learning algorithms can segment the customer base into cohorts, each with distinct churn probabilities, allowing businesses to deploy tailored retention strategies. Such personalized approaches not only mitigate churn but also create opportunities for cross-selling and upselling, thereby maximizing revenue potential. Despite its potential, customer churn prediction presents significant challenges.

The quality and completeness of data are pivotal to building accurate models, as noisy or missing data can impair predictive performance. Moreover, customer behavior is inherently dynamic, influenced by external factors such as market trends, economic conditions, and competitor actions. Consequently, models must be continuously refined and updated to remain relevant. Ethical considerations, including data privacy and fairness, further complicate the deployment of predictive analytics in customer retention strategies, necessitating transparent and responsible practices.

This research paper explores the methodologies, applications, and implications of customer churn prediction in a corporate context. It emphasizes the practical significance of predictive analytics in enabling data-driven decision-making and mitigating churn-related risks. By analyzing real-world datasets and leveraging advanced machine learning techniques, this study seeks to provide actionable insights for academic and corporate stakeholders. The findings aim to contribute to the growing body of knowledge on customer retention strategies, offering a comprehensive understanding of how predictive models can drive sustainable business success.

Methodology:

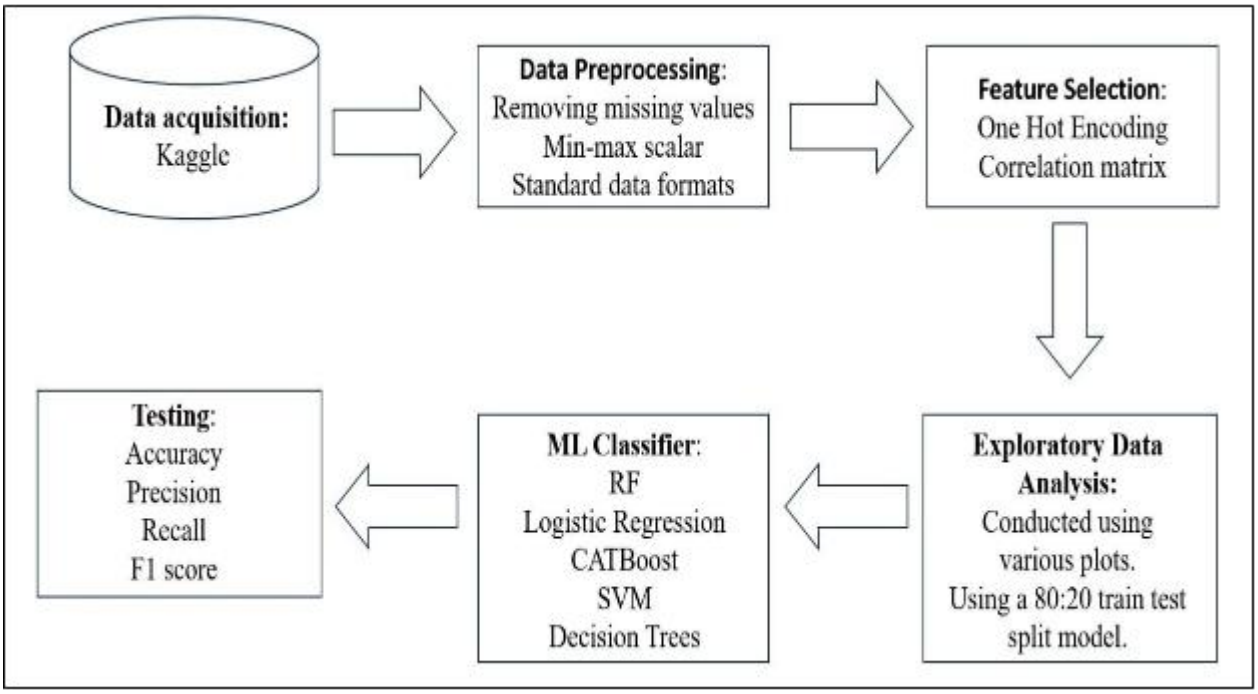

Fig 1 Workflow Diagram

Exploratory Data Analysis –

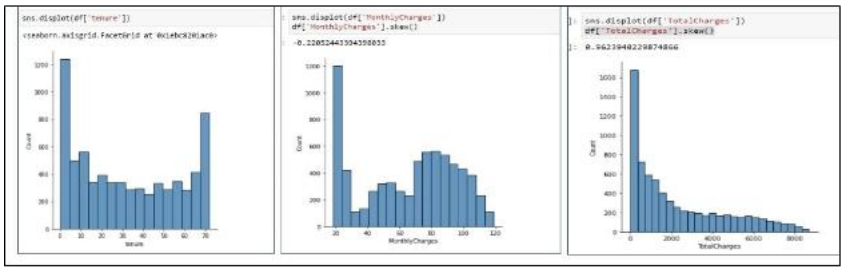

Fig 2 EDA, Numeric Features.

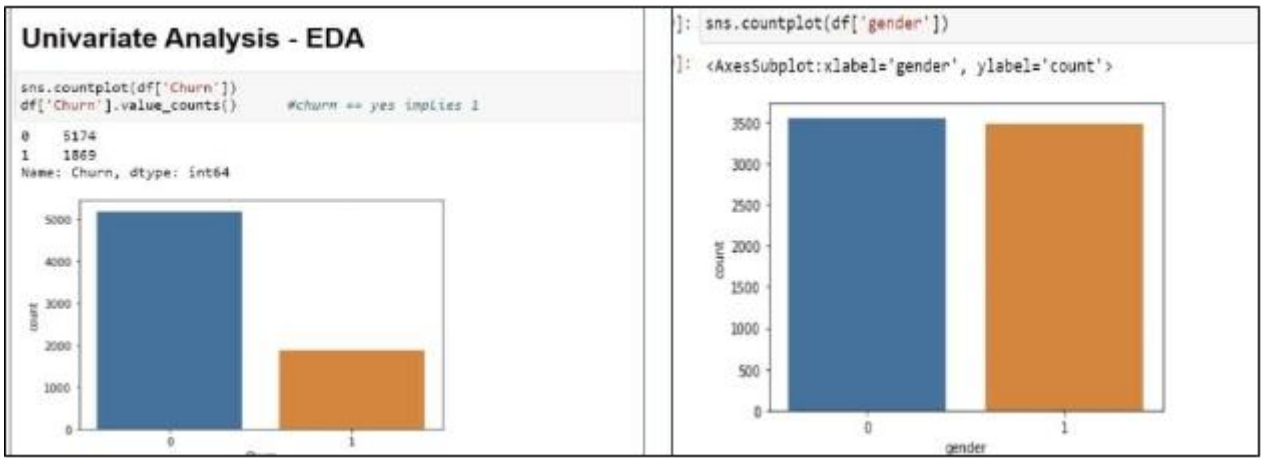

Fig 3 EDA, Univariate analysis.

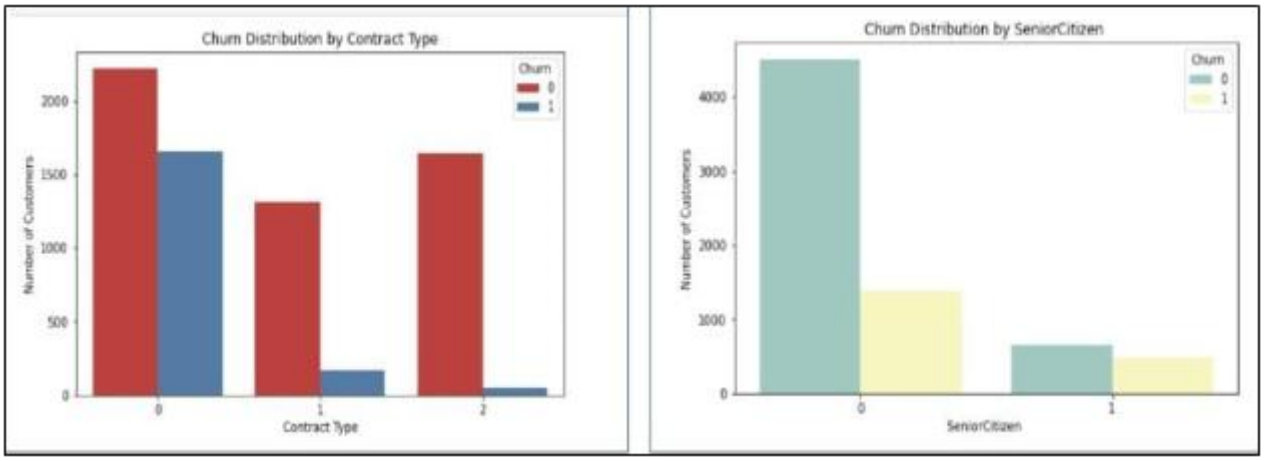

Fig 4 EDA, Multivariate analysis.

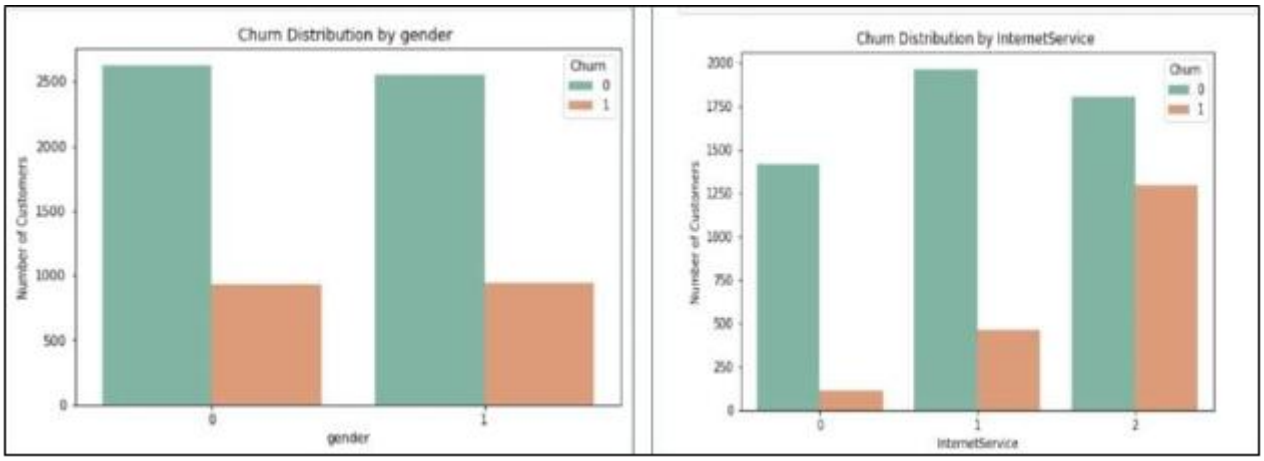

Fig 5 EDA, Multivariate analysis -

TECHNOLOGIES USED FOR A MODEL DEVELOPMENT:

A. Machine Learning Classifiers –

1. Decision Tree

The decision tree is a versatile and easily understood machine learning technique for applications, including regression and classification!!! Each leaf node in the created tree-like structure really represents the conclusion, which might be a class label for classification or a numerical value for regression???? with each internal node denoting a test or judgment on a characteristic [1][2]. By starting at the root node and choosing the most informative feature, the tree is constructed iteratively in order to separate the data into subsets that are as pure as possible regarding the target variable!!! The previously described process continues until an end condition is met, usually when the node reaches a particular depth or when it has a certain number of data points... A useful technique for clarifying the reasoning behind.

2. K-Nearest Neighbor

One of the simplest yet most important categorization methods in machine learning is K-Nearest Neighbors. It is heavily used in pattern recognition, data mining, and intrusion detection and is a member of the supervised learning domain. Since it is non-parametric that is, it does not make any underlying assumptions about the distribution of data—it is extensively applicable in real-life circumstances (in contrast to other algorithms like GMM, which assume a Gaussian distribution of the provided data) [3][5][6]. An attribute-based previous data set (also known as training data) is provided to us, allowing us to classify locations into groups. Because of its simplicity and ease of usage, the K-NN) algorithm is a popular and adaptable machine learning technique.

$\sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$ Euclidean function (1)

$\sum_{k=0}^{n} |xi - yi|$ Manhattan function (2)

$(\sum_{i=1}^{k} (|xi - yi|)^q)^{1/q}$ Minkowski function (3)

3. Support Vector Classifier (SVM)

A supervised machine learning approach called Support Vector Machine (SVM) is utilized for regression as well as classification. Even yet, classification issues are the most appropriate use for regression problems. The SVM algorithm, which stands for Support Vector Machine, has a major objective of finding the super-duper best hyperplane of all. This dazzling hyperplane will conquer all other hyperplanes and emerge victorious like a superhero in a cape! And an N-dimensional space that may be used to divide data points into various feature space classes. The hyperplane attempts to maintain the largest feasible buffer between the nearest points of various classes. The number of features determines the hyperplane's dimension. The hyperplane is essentially a line if there are just 2 input characteristics. When considering the simplistic notion of a hyperplane, it can be described as a line! Whose existence arises due to the presence of solely two input characteristics. It is harder to envision when the number of features exceeds three [7].

4. Logistic Regression

Logistic regression stands out as a supervised machine learning algorithm primarily employed for binary classification tasks. It operates by utilizing a logistic function, also termed as a sigmoid function, which takes independent variables as input and generates a probability value ranging between 0 and 1. For instance, considering two classes, Class 0 and Class 1, if the logistic function's value for an input surpasses 0.5 (threshold value), the instance is assigned to Class 1; otherwise, it belongs to Class 0. The term "regression" is used because logistic regression extends from linear regression but is tailored specifically for classification tasks. Unlike linear regression, which yields continuous output, logistic regression predicts the probability of an instance belonging to a particular class or not.

B. Ensemble models

In machine learning, achieving high accuracy in predictions is crucial for establishing reliable and robust models. Ensemble learning serves as a supervised technique that integrates multiple models to create a more powerful and stable model. The core idea lies in combining the strengths of various models to mitigate overfitting and enhance adaptability to complex data patterns.

1. Random Forest

Random Forest is a robust ensemble learning algorithm used for classification and regression tasks. It operates by constructing multiple decision trees during training and combining their outputs to improve accuracy and reduce overfitting. Each tree in the forest is built using a random subset of the data and a random selection of features, introducing diversity among the trees.

The algorithm leverages bagging (Bootstrap Aggregating), where data samples are drawn with replacement to train individual trees. During prediction, classification tasks rely on majority voting, while regression tasks use the average of the outputs. This ensemble approach enhances the model's stability and resilience to noise in the data.

Random Forest is known for its ability to handle large datasets with high-dimensional features and its resistance to overfitting. Additionally, it provides feature importance rankings, aiding in understanding the contribution of each variable. Its versatility and effectiveness make it a popular choice in various domains.

2. CatBoost

CatBoost (Categorical Boosting) is a gradient boosting algorithm developed by Yandex, designed to handle categorical data efficiently. It is built on the principle of boosting, which combines multiple weak learners, typically decision trees, to create a robust predictive model. Unlike traditional gradient boosting methods, CatBoost incorporates innovative techniques to process categorical features directly, eliminating the need for extensive preprocessing like one-hot encoding.

A key feature of CatBoost is its use of ordered boosting, which prevents overfitting by introducing randomness into the selection of training data subsets. Additionally, it applies advanced handling of missing values and implements symmetric tree structures to optimize training speed and accuracy. The algorithm excels in reducing prediction bias and variance, making it suitable for complex datasets.

CatBoost is widely used for classification and regression tasks across industries due to its ability to work with mixed-type data, superior performance, and ease of use.

SMOTEENN –

SMOTEENN is a hybrid technique for handling imbalanced datasets that combines oversampling with SMOTE (Synthetic Minority Oversampling Technique) and undersampling with ENN (Edited Nearest Neighbors). SMOTE generates synthetic samples for the minority class by interpolating between existing samples and their nearest neighbors, balancing the dataset. ENN then removes noisy or borderline samples, particularly from the majority class, by eliminating instances misclassified by their K-nearest neighbors. This dual-step approach enhances dataset quality, reduces overfitting risk, and improves model performance on imbalanced data. However, it can be computationally expensive and requires parameter tuning for optimal results**.**

GridSearchCV –

For customer churn prediction, GridSearchCV optimizes hyperparameters to improve model performance. It systematically evaluates combinations of hyperparameters (e.g., learning rate, max depth, and regularization strength) for classifiers like Random Forest, XGBoost, or Logistic Regression. Using cross-validation, it trains and validates the model on different data folds, ensuring robust results.

For example, in churn prediction, GridSearchCV can fine-tune Random Forest's n_estimators and max_features or XGBoost's eta and max_depth. Once the optimal parameters are identified, the final model is trained on the full training set and evaluated on test data. This approach ensures better churn prediction accuracy and minimizes overfitting.

Objective:

1. Predict Churn Probability: Develop a machine learning model to predict the likelihood of a customer churning based on historical and real-time data.
2. Identify Key Drivers of Churn: Analyze customer behavior and identify factors contributing to churn, such as service usage, complaints, pricing, or competitor offerings.
3. Enhance Retention Strategies: Use the model to support targeted retention campaigns, focusing on customers at high risk of leaving.
4. Improve Business Revenue: Reduce revenue loss by proactively addressing churn and retaining valuable customers.
5. Optimize Resource Allocation: Prioritize resources and marketing efforts to maximize the return on investment (ROI) in customer retention initial

Results

The results of the machine learning classifiers K Nearest Neighbors, Random Forest, Decision Tree, Support Vector Classifier, Logistic Regression, and CatBoost for predicting customer churn are presented on both imbalanced and balanced datasets. The best accuracy of 80% was achieved in the case of the imbalanced dataset by CatBoost ensemble technique. This result indicates that CatBoost is robust to churn prediction in the case of skewed class distribution, where there are far more non-churned customers than churned ones. The performance of the models improved significantly when the dataset was balanced using the SMOTEENN technique, combining oversampling of the minority class with the editing of the nearest neighbors to reduce noise. hbhhggfvfvfcfcfrdcdrcdxdrxjkbasNJVFdufhuiwgfiesuegfiougfuiwGFMKDSUJGFSJUWGFYBYUHDGRFUYSHGFUYDHSEGFYH

On the balanced dataset, KNN performed best with an accuracy of 98.1%. Such a tremendous increase in accuracy further supports the critical role that balancing the dataset plays in enhancing model performance. ROC-AUC score and F1 score, the other two metrics measuring performance, were compared with the help of it as well. All the classifiers' results improved on the dataset balanced, and KNN always had the best scores of ROC-AUC and F1, that's, it could more sensitively distinguish between churned and non-churned customers in this case. Having been balanced in data, KNN distinguished better the inherent pattern, which was masked because of imbalance. In conclusion, balancing the dataset using SMOTEENN was instrumental in improving the performance of all classifiers, with KNN emerging as the top performer across accuracy, ROC-AUC, and F1 score metrics. BHJ

After applying **GridSearchCV** to optimize hyperparameters for algorithms like KNN, SVM, Logistic Regression, Random Forest (RF), Decision Tree (DT), and CatBoost, the results typically reveal the most suitable model for customer churn prediction.

- **KNN**: Often performs moderately; sensitive to the number of neighbors (n_neighbors) and distance metrics, making it less effective for large datasets.
- **SVM**: Works well on high-dimensional data but may struggle with imbalanced datasets unless properly tuned (e.g., C and kernel parameters).
- **Logistic Regression**: Provides interpretable results with reasonable accuracy; effective when relationships are linear.
- **Random Forest**: Often outperforms others due to its ensemble nature and robustness to overfitting; key parameters include n_estimators and max_depth.
- **Decision Tree**: Simpler and interpretable but prone to overfitting without careful tuning (e.g., max_depth).
- **CatBoost**: Generally achieves the best performance, particularly on categorical-heavy datasets, due to its gradient-boosting mechanism.

The best model depends on dataset characteristics and evaluation metrics like precision, recall, or F1-score.

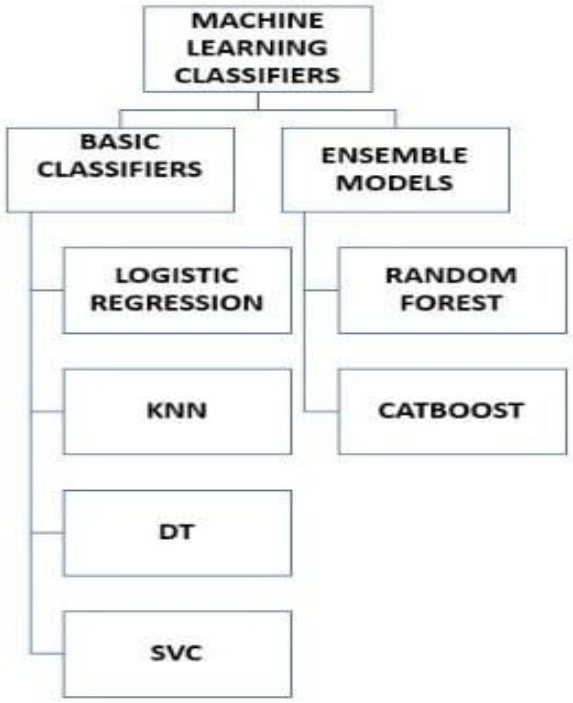

Fig 5 Block Diagram

Table 1 – Resulting accuracy of machine leaning classifier on imbalanced data.

S.NO.	MODEL NAME	ACCURACY	ROC-AUC SCORE	F1 SCORE
1.	LOGISTIC REGGRESSION	0.766 %	0.7800131926121	0.6511134676564
2.	K- NEAREST NEIGHBORS	0.791 %	0.6945923123518	0.5558912386707
3.	SUPPORT VECTOR CLASSIFIER	0.756 %	0.7660544336417	0.6348195329087
4.	RANDOM FOREST	0.799 %	0.6957781074505	0.5585023400936
5.	CATBOOST	0.802 %	0.7032072934098	0.5709876543209
6.	DECISION TREE	0.767 %	0.6663950289005	0.5089820359281

Table 2 – Resulting accuracy of machine leaning classifier on balanced data after applying SMOTEENN

S.NO.	MODEL NAME	ACCURACY	ROC-AUC SCORE	F1 SCORE
1.	LOGISTIC REGGRESSION	0.914 %	0.9136851615646	0.9191685912240
2.	K- NEAREST NEIGHBORS	0.981 %	0.9809258078231	0.9821567106284
3.	SUPPORT VECTOR CLASSIFIER	0.937 %	0.9372874149659	0.9397024275646
4.	RANDOM FOREST	0.976 %	0.9754570578231	0.9765990639625
5.	CATBOOST	0.959%	0.9588647959183	0.9612403100775
6.	DECISION TREE	0.95 %	0.9503401360544	0.9522317932655

Conclusion

The suggested study work has established a model for customer churn prediction that can be used to improve performance without incurring significant additional costs, as well as reducing prediction variation.

Customer churn prediction is an important tool for business organizations to identify at-risk customers in advance and retain them. This will lead to better customer relationships and lower revenue loss. Advanced machine learning techniques, balanced datasets, and robust feature engineering will greatly enhance the accuracy and effectiveness of the predictions. With this, the use of predictive models in retention strategies will reduce churn rates while enabling organizations to make optimum use of resources for sustainable long-term business gains.

Machine learning classifiers can be very powerful when it comes to predicting the accurately churned customer of any business, which enables business to proactively identify and alert the customers who would be at risk of moving away. It helps identify customers early to apply timely and targeted retention strategies that enhance customer loyalty while reducing churn rates. However, such models depend on the quality of data, careful selection of relevant features, and the choice of an appropriate classifier.

The problem of churn prediction models is class imbalance because there are far fewer churned customers than those who did not churn. This may lead to bias toward the majority class and reduces their ability to predict correctly. This is where SMOTEENN, Synthetic Minority Oversampling and Edited Nearest Neighbors comes into the play to alleviate such issues. SMOTEENN balances the dataset; therefore, the model learns a representative sample of both churned and non-churned customers, thus enhancing the generalization capability of the classifier, thereby enhancing the predictive accuracy.

Better performance metrics, which include precision and recall, ensue in case of well-balanced datasets, for which a higher value is an indicator of a good classifier to predict churning. Using balance churn prediction models, the resulting output is a reduction of churning while strengthening customer relationship and driving better business outputs. It is therefore critical to success in the long run.

References:

Research Papers:

1. 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
2. Lim, E. P., Chen, H., & Chen, G. (2013). Business Intelligence and Analytics. ACM Transactions on Management Information Systems, 3(4), 1–10. https://doi.org/10.1145/2407740.2407741
3. Munawar, H. S., Qayyum, S., Ullah, F., & Sepasgozar, S. (2020). Big Data and Its Applications in Smart Real Estate and the Disaster Management Life Cycle: A Systematic Analysis. Big Data and Cognitive Computing, 4(2), https://doi.org/10.3390/bdcc4020004
4. Bawack, R. E., Wamba, S. F., Carillo, K. D. A., & Akter, S. (2022). Artificial intelligence in E-Commerce: a bibliometric study and literature review. Electronic Markets, 32(1), 297–338. https://doi.org/10.1007/s12525-022-00537-z
5. Regions in Industrial Transition. (2019). In OECD regional development studies. https://doi.org/10.1787/c76ec2a1-en
6. Gulla, J. A., Svendsen, R. D., Zhang, L., Stenbom, A., & Frøland, J. (2021). Recommending news in traditional media companies. AI Magazine, 42(3), 55–69. https://doi.org/10.1609/aimag. v42i3.18146
7. Couper, P. (2015). A Student's Introduction to Geographical Thought: Theories, Philosophies, Methodologies. https://doi.org/10.4135/9781473910775
8. Petersen, J. A., Paulich, B. J., Khodakarami, F., Spyropoulou, S., & Kumar, V. (2022). Customer-based execution strategy in a global digital economy. International Journal of Research in Marketing, 39(2), 566–582. https://doi.org/10.1016/j.ijresmar.2021.09.010
9. Agrawal, S., Das, A., Gaikwad, A., & Dhage, S. (2018). Customer Churn Prediction Modelling Based on Behavioural Patterns Analysis using Deep Learning. https://doi.org/10.1109/icscee.2018.8538420
10. Gubela, R. M., Lessmann, S., & Jaroszewicz, S. (2020). Response transformation and profit decomposition for revenue uplift modeling. European Journal of Operational Research, 283(2), 647–661. https://doi.org/10.1016/j.ejor.2019.11.030
11. Rabbah, J., Ridouani, M., & Hassouni, L. (2022). A New Churn Prediction Model Based on Deep Insight Features Transformation for Convolution Neural Network Architecture and Stacknet. International Journal of Web-Based Learning and Teaching Technologies, 17(1), 1–18. https://doi.org/10.4018/ijwltt.300342